Learning to improve clinical decisions and AI safety by leveraging structure

by

Geeticka Chauhan

B.S., Florida International University (2017)
S.M., Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2024

Authored by: Geeticka Chauhan
Department of Electrical Engineering and Computer Science
June 30, 2024

Certified by: Peter Szolovits
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Learning to improve clinical decisions and AI safety by leveraging structure

by
Geeticka Chauhan

Submitted to the Department of Electrical Engineering and Computer Science
on June 30, 2024, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The availability of large collections of digitized healthcare data along with the increasing power of computation has allowed machine learning (ML) for healthcare to become one of the key applied research domains in ML. ML for health has great potential in providing clinical decision-making support that can improve quality of care and reduce healthcare spending by easing clinical operations. However, the successful development of ML models in healthcare is contingent on data that is complex, noisy, heterogeneous, limited in labels and highly sensitive. In this thesis, we leverage the unique structure present in medical data along with the availability of external knowledge to guide model predictions. Additionally, we develop differentially private (DP) training techniques using gradient structure to mitigate privacy leakage.

In this thesis, we develop methods on different medical modalities such as multivariate physiological signals of ICU patients, patient discharge summaries, biomedical scientific articles, radiology reports, chest radiography imaging and spoken utterances. We tackle tasks such as forecasting patient states, relationship extraction, disease prediction, medical report generation and differentially private model training. We begin the thesis by offering open source data processing and modeling frameworks, move towards improved interpretability of model predictions to develop clinician trust and finally investigate differentially private ML techniques to protect user data.

First, we show that the use of aggregated feature representations based on clinical knowledge offers model robustness against evolving hospital systems. Second, we leverage external knowledge in the form of clinical concept extraction to significantly improve relationship extraction. Third, we leverage the rich information from reports associated with chest radiographs to develop highly accurate disease severity prediction models using contrastive learning. Fourth, we showcase that the report generation task offers competitive disease prediction capabilities, label efficiency and improved interpretability. Finally, we introduce novel methods for improved privacy-utility-compute tradeoffs for DP pre-training of large speech models. We highlight DP as an important component of model safety, necessitating its development in conjunction with AI safety approaches that will be pertinent in healthcare and beyond.

Thesis Supervisor: Peter Szolovits
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Ever since I was introduced to natural language processing as an idealistic junior in undergrad, I have dreamt of working in the area of artificial intelligence and machine learning. Being able to pursue that dream at MIT has been nothing short of incredible. Completing the PhD marks an incredible personal milestone for me; not just due to the opportunity to contribute to science, but because of the ability to work with fantastic mentors and collaborators.

First and foremost, I would like to thank my research advisor Prof. Peter Szolovits for his generous intellectual support throughout my journey. Pete encouraged me to collaborate extensively, and push my boundaries towards pursuing new research directions. This mindset helped broaden my skillset and form meaningful collaborations within and outside my research group. Though I faced imposter syndrome on many occasions, Pete supported me in pursuing my ideas by sharing his own personal journey through research. Pete always asked astute questions that helped solidify the motivations of my research, and ensured that the directions I pursued would have measurable impact in clinical scenarios. Pete has a kind sternness I really appreciated throughout my journey, and I can't thank him enough for teaching me how to be a good researcher, collaborator and communicator. Few advisors are able to cut through bureaucracy as effectively as Pete, and I feel grateful to have had so much intellectual freedom under Pete's guidance. Thank you immensely, Pete!

I would also like to thank my committee members, Profs. Polina Golland and Marzyeh Ghassemi. I thoroughly enjoyed the discussions they led during our committee meetings, and for making me think deeply about my research vision and story. I first met Polina when we started collaborating on the multimodal disease prediction project, and I was immediately impressed by how effectively she led interdisciplinary meetings between clinical collaborators and computer scientists. Polina knows the magic sauce needed to clearly communicate computer science jargon to top radiologists in the country. Not only did I learn about the field of medical vision through Polina, but also about how to communicate research findings meaningfully to clinical stakeholders. I am so grateful to Polina for being generous with her time, and for making me feel cared for as a student. Marzyeh was the first senior postdoc I worked with during my journey at MIT, and I was struck by her genuine enthusiasm and passion for machine learning for health. During our early collaborations, Marzyeh brought her contagious enthusiasm for our field, and taught me how to clearly communicate research findings. I felt fueled towards pursuing a solid research path after working with Marzyeh. After Marzyeh started as a faculty member at MIT, I feel grateful to have had the chance to discuss and learn from her insights in the interdisciplinary field of machine learning for health. Particularly, her recommendations and thoughts on Differential Privacy have criticially impacted my vision for the future.

I would like to thank my academic advisor, Prof. Charles Leiserson, for his immeasurable insights every semester before registration day. Charles has a strong belief in accumulating technical knowledge and improve soft skills every semester to holistically benefit growth during the PhD, and his advice inspired me to keep growing my knowledge throughout the degree program. Thanks for the great discussions, Charles!

When I was an undergraduate student, I was lucky to meet Prof. Mark Finlayson who introduced the vast discipline of natural language processing to me. Mark truly embodies the persona of an excellent educator, and is able to expand students' minds even when they have little background in machine learning. The fundamentals I learned in Mark's classes made me a better computer scientist and researcher, and having the opportunity to conduct research under his guidance was transformative to my journey. Mark is generous with his support, and still shares his valuable feedback with me when I need it. Thank you for supporting me since the beginning, Mark!

I would like to thank my internship mentors at Google: Guolong Su, Vincent Perot, Om Thakkar and Abhradeep Guha Thakurta. Thank you for believing in me, Guolong and Vincent, and for your guidance through my very first internship at Google! Thank you for introducing me to the exciting field of Differential Privacy, Om and Abhradeep, and for teaching me so many valuable skills as a researcher.

I would like to thank Prof. Leslie Kolodziejski and Janet Fischer from the MIT EECS grad office, for offering emotional support during my journey and for answering my numerous questions every time. Members of the clinical decision making group (MEDG) have been highly impactful to my research over the years, including Di Jin, Matthew McDermott, Harry Hsu, Willie Boag, Emily Alsentzer, Elena Sergeeva, Wei-Hung Weng, Tristan Naumann and Eric Lehman. Finally, I really appreciate my research collaborators for sharing their incredible expertise: Ray Liao, Keegan Quigley, Miriam Cha, Dr. Steven Horng, Dr. Seth Berkowitz, Labiba Jahan, Rumen Dangovski, Charlotte Loh, Steve Chien, Arun Narayanan, Virat Shejwalkar, Lun Wang and Matt Li.

I would not have been able to start my PhD without my amazing undergrad experiences at Florida International University (FIU). I would like to thank FIU for their generous funding that allowed me to pursue a bachelors in my dream major in a new country. Thank you also to the Honors College and Profs. Juan Carlos Espinosa and John William Bailly for expanding my mind beyond my major, and teaching me how to be a holistic student. I always told people that I was lucky to have the feeling of being part of a small department in a large university: the best of both worlds. The School of Computer Science at FIU offered this perfect balance, and I am grateful to have learned from fantastic educators at the institute, including Profs. Jai Navlakha, Giri Narasimhan, Raju Rangaswami, Christine Lisetti, Mark Finlayson, Prabakar Nagarajan, Michael Robinson, Tiana Solis, Scott Graham, Susan Gorman, Gregory Shaw and Geoffrey Smith. Thank you for the fantastic opportunity to pursue computer science and learn from the wonderful community in Miami, FIU!

During this journey, I have been incredibly lucky to have the support of both my parents. They have created an excellent environment full of learning and love, and have given me freedom to pursue all my interests at a time in India when it was unusual to give girls academic freedom. I am lucky to have feminist parents, who taught me the value of hard work, dedication and mental strength. I am also grateful to have been surrounded by so much family growing up: my Chachu and Chachi, Nikki, grandparents and sister. Thank you for showing me the value of togetherness and fun. I would also like to thank my Cuban family in Miami, for their love when I was lonely in a new country. I hold you warmly in my heart.

Last but not the least, thank you to my friends who have supported me over the years: Emellyn and Almita, Madelin, Diana, Mia and Gia, Neil, Kruthika and Bella, Tooba, Monica, Asmita, Linda, Olivia, Himanshu, Krishna and Dinesh, Tejas, Kunal and Simba, Hsin-Yu, Iqra and John, Mahnoor, Mohammad, Niharika and Sakshi. I'd like to give a special shout out to Kruthika, Emellyn, Himanshu, Neil, Krishna and Tooba for their support this past year. I am also grateful to the Sidney Pacific Graduate Community for giving me a sense of home at MIT.

The research included in this thesis was generously funded by the following organizations: Frederick and Barbara Cronin, Wistron Corporation, Bayer, the Canada Research Chair Program, the Canadian Institute for Advanced Research, IBM Corporation, Microsoft Research, the National Institutes of Mental Health, the National Library of Medicine, the Natural Sciences and Engineering Research Council of Canada, the National Science Foundation, Philips Research, the U.S. Air Force, the National Institute of Biomedical Imaging and Bioengineering, IBM Watson, the National Human Genome Research Institute, and the Office of Naval Research.

The journey of completing a PhD involves various new challenges and growth areas, and I have been lucky to have had the support of my friends, family and my academic community. Truly, thank you for believing in me and my ideas. MIT has been one of the best places that I could have ended up for pursuing moonshot ideas, and meet like minded people to support my growth. Going forward, I hope to continue pursuing ideas that create impact and contribute back to society.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The wide adoption of technologies such as electronic health record (EHR) systems and medical imaging techniques by healthcare organizations have led to the availability of large collections of digitized healthcare data [39, 287]. This, combined with the increasing power of computation has made machine learning (ML) for healthcare one of the key applied research domains in ML [287, 112, 211]. Applying modern ML methods to observational health data holds the potential to improve healthcare in many ways for offering clinical decision support, such as recommending better patient treatments, improving hospital operations and answering fundamental scientific questions [88].

ML for health has several useful applications in diagnosing diseases [74, 100, 49], forecasting patient states [201], building treatment plans [236, 151] and reducing healthcare spending by easing clinical operations [324, 208, 116]. There are also varied models applied to different types of healthcare data, such as lab measurements [232, 212], claims data [232, 67], clinical text narratives [26, 306], medical images [100, 74, 16, 179] and waveform signals [159].

1.1 Challenges and Opportunities with medical data

Despite the vast potential utility of ML for health, working with healthcare data offers a unique set of challenges. This is largely due to its high dimensionality relative to the dataset size, presence of very noisy labels, disparate and sporadically recorded modalities and suffering from a high rate of missingness [87, 43, 304]. Data heterogeneity is also a common feature, with highly varied sources such as vital signs, genomics, audio, demographics, laboratory tests, clinical narratives and imaging from modalities such as chest X-ray or MRI [88, 287]. Limited data availability is another feature, as a result of the expense of data collection and lack of consistent expert annotation from medical professionals [279, 282]. This may result in related issues such as class imbalance [141, 158, 185], commonly observed in fields such as dermatology and cancer detection [183, 60]. Another well-known class imbalance scenario exists with rare disease classification where the vast majority of patients might not have the disease under consideration [335]. Finally, the sensitive nature of healthcare data

necessitates re-thinking techniques for data sharing and model development, while respecting privacy of patients [77, 182].

1.1.1 Common techniques for modeling medical data

Pre-processing is often utilized for cleaning up messy data, whereas techniques like representation learning are used for modeling data without explicit feature engineering. Hence, pre-processing and representation learning go hand in hand in the clinical ML space. Additionally, techniques that address the privacy leakage from ML models are becoming increasingly relevant in healthcare and beyond as the requirement for accessing larger datasets increases with increasing model sizes.

Pre-processing

This is the act of cleaning up data in a manner that removes noise and leaves it in a form suitable for feeding to an ML model [146, 93]. Several different pre-processing measures are taken for clinical data, such as data cleansing, feature tuning, feature transformation, feature extraction and feature selection. Data cleansing is the act of removing or correcting records with corrupted or invalid values. Feature tuning involves operations such as scaling numerical values, imputing missing values and clipping outliers. Feature transformation involves operations such as conversion of numerical features to categorical features and conversion of categorical features to a numerical representations like one-hot encoding. Feature extraction involves reduction of high dimensional features to low-dimensions using techniques like PCA. Feature selection involves choosing a subset of input features that are most relevant and statistically significant for good performance of a model.

Additionally, specific modalities require different pre-processing techniques for effective machine learning model development. For example, text data involves techniques like stemming and lemmatization (for general domain data) and named entity recognition based replacement (for clinical data) [40]. Image processing involves techniques such as clipping, resizing, cropping and filtering. EHR datasets involve careful patient cohort selection for consistent vital ranges, outlier removal, duplicate removal and feature aggregation [299, 24].

Representation Learning

This involves a set of techniques to learn useful features without the labor intensive work of manual feature selection known as feature engineering [20]. These techniques involve traditional unsupervised learning methods such as principal component analysis and matrix factorization, as well as newer approaches with large-scale pre-training involving self-attention and transformer models [146, 230, 64, 142]. Other interesting and relevant approaches in representation learning involve multiple modalities as data sources [280, 307, 41]. Representation learning has been particularly useful in the healthcare space for modeling high dimensional noisy data such as by Suresh

et al. [280] to model demographics, time-varying variables like vital signs and labs, and clinical narrative notes.

Representation learning encompasses several newer deep learning techniques, including those involving convolutional neural networks (CNN), long short-term memory networks (LSTM) as well as the transformer architecture [154, 155, 81, 220, 113, 293]. What distinguished these networks from those reliant on feature engineering was that they relied on weaker inductive biases on the type of learning function (e.g., linear versus non-linear) but stronger inductive biases on the relationships between components of the network, inspired by the domain of interest. Battaglia et al. [15] describe inductive biases of a learning algorithm as being relational vs non-relational.

Non-relational inductive biases are assumptions involving components of a model such as L2 regularization, which prioritizes solutions where model parameters have small values to minimize overfitting. Examples of other non-relational inductive biases are activation non-linearities, dropout and data augmentation that impose constraints on the trajectory and outcome of learning. *Relational* inductive biases impose constraints on relationships and interactions among entities in a learning process. For example, the design choice of stacking multiple layers in deep learning can be seen as imposing the relational bias of hierarchical processing, where computations are performed in stages to result in increasingly long range interactions among information in the input signal.

Relational inductive biases imposed on the building blocks themselves distinguish major deep learning architectures of today. For example, convolutional layers force locality relationships between grid elements such as pixels and the reuse of the same rule across localities in the input (known as local and translational invariance). These biases effectively help in processing natural image data due to high covariance within local neighborhoods, and statistics being mostly stationary across an image. Another example is recurrent layers, where there is a markov dependence between one step's hidden state on the previous hidden state and the current input. This rule is reused every step, reflecting temporal invariance.

Transformer architectures introduced by Vaswani et al. [293] have the weakest relational biases, by introducing the self-attention mechanism to allow automatic learning of relationships between the building blocks (or hidden units) of the model. They are universal architectures, showing promise in several domains such as language, vision and genomics data [64, 68, 51, 97]. They have quickly become the gold-standard in deep learning, including forming the basis for the popular ChatGPT framework [219]. In the healthcare space, transformer models have formed the basis for Med-PaLM and MedPaLM M models that solve various multi-modal tasks such as medical question answering, radiology report generation, image interpretation and genomic variant calling [271, 289].

Data Privacy

ML models are becoming increasingly prone to training data leakage as seen in [270, 32, 33, 35]. These works have demonstrated several attacks such as membership inference and reconstruction attacks over popular models. The consequences of train-

ing data memorization in these models can be especially catastrophic in healthcare settings, such as causing increased discrimination and bias against patients [25].

One of the most robust and popular techniques for mitigating privacy leakage is differential privacy (DP) [70]. By adding a calibrated amount of noise to introduce randomness to a mechanism over a dataset (such as a query or an ML model), this technique ensures that an attacker cannot make conclusions about whether any particular data was used to produce the result, even while having access to the mechanism and arbitrary external side information. DP effectively protects against popular attacks such as reconstruction and membership inference attacks. Most recently, researchers have been actively exploring methods to ensure DP during the model training stage, through techniques such as Differentially Private Stochastic Gradient Descent (DP-SGD) [1] and Private Aggregation of Teacher Ensembles (PATE) [223].

1.1.2 Challenges and Trade-offs

Compared to classical approaches that relied on feature engineering, deep neural networks offer low bias and high variance due to their large model capacity [17]. This might make them the perfect network from a performance perspective, but the high model capacity also makes them very *data hungry*. The latest transformer architectures rely on a new paradigm introduced by Devlin et al. [64] of pre-training the model on a large amount of unlabeled external data and fine-tuning on labeled in-domain data for downstream classification. This makes transformer models even more data hungry than prior deep learning approaches. The original BERT model, of size 110 or 340 million parameters for BERT-base and BERT-large respectively, relied on a pre-trained corpus of approximately 3500 million words. While many of the latest models don't reveal the source of their pre-training data anymore, it is estimated that they are trained on even larger amount of data than BERT. For example, the latest Whisper model from OpenAI has been trained on approximately 680,000 hours of audio [248] and the GPT-3 model from OpenAI has been trained on approximately 500 billion tokens [27].

In the context of healthcare, strict compliance laws and limited data availability can hinder the development of high capacity neural network models. While variants like Med-PaLM and Med-PaLM M [271, 289] have pushed the healthcare space forward, there are still several tasks that remain outside the scope of high capacity model development. Healthcare data is also noisy enough that the performance gap between classical and neural network approaches is minimized [26]. Therefore, depending on the size of the dataset and labels, along with its noisiness, machine learning practitioners have to make careful decisions not to introduce an unnecessarily high capacity model when a simpler one like logistic regression might be sufficient. Boag et al. [26] advise against adopting a one-size-fits-all approach in modeling healthcare data, and Battaglia et al. [15] recommend a key path forward for modern AI being approaches that integrate classical and modern end-to-end deep learning.

1.2 Goals and Contribution

In this thesis, we develop approaches that leverage the unique structure of medical data along with available external knowledge; which offer robustness, label-efficiency, interpretability, high performance on downstream tasks and data privacy. We believe the best way forward for ML in health is to marry popular end-to-end approaches with those leveraging explicit structure and feature engineering to counteract the trade-offs between high performance and needing large amounts of clean data for training. An increasingly important aspect of ML model development in healthcare and beyond is addressing data privacy compliance issues. We develop representation learning methods to tackle four main challenges:

1. Medical data consists of **heterogeneous features**, necessitating *smarter feature grouping*. We demonstrate these challenges on electronic health record (EHR) lab and vitals data, by sharing open source pre-processing techniques over a public EHR dataset and developing an effective and robust machine learning (ML) model for this dataset using smart grouping of features based on clinical knowledge. This work is described in the papers [299, 212].

2. Medical data is **noisy**, making *smart feature selection* necessary. We specifically target medical text, requiring pre-processing techniques based on external knowledge such as tokenization, punctuation and digit removal and named entity recognition (NER) based token replacement. By providing an open-source reproducible pre-processing, modeling and training pipeline, we explore and report sources of performance variability to facilitate fair comparison between different modeling approaches developed for the downstream task of relation extraction. The framework and analysis is done in [40], with some of the modeling groundwork laid out in [131].

3. Medical data is **expensive to label**, making medical machine learning a natural application area for *semi- and self- supervised* approaches. We apply approaches leveraging paired radiology report data to capture ground truth radiologist narrative in conjunction with medical vision models for improved performance and interpretability. We first introduce a foundational approach using contrastive learning to create implicit structure from paired and unpaired chest x ray images and radiology reports [41]. We continue building on this direction by developing approaches using radiology report generation as a pre-training task, for improved robustness towards learning from even fewer labels [242, 243, 244]. We make comparisons between contrastive learning and radiology report generation pre-training tasks, and introduce new methods for improved interpretability and interaction between the radiologist and the machine learning model.

4. Medical data is **highly sensitive**, with strict *compliance* and *data privacy* regulations as detailed in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [56]. Through a collaborative project with Google researchers, we develop differentially private pre-training approaches in the general domain

[42]. We improve upon existing methods by providing privacy guarantees over the pre-training data and by tackling the challenging area of pre-training a model from scratch using differential privacy for the general domain. These approaches have direct applicability to the healthcare space, whose specific investigation we leave to future work.

In this thesis, we highlight how leveraging the structure of healthcare data with end-to-end deep learning approaches includes just the right amount of inductive biases towards improved performance, interpretability, robustness and label-efficiency. We also highlight data privacy as an important aspect of AI safety, and discuss how it will be an integral part of the ML model development process of the future. The work done in this thesis reflects important dataset, modeling and privacy considerations for practical development of machine learning for healthcare models.

1.3 Publications

This thesis primarily relates with the following publications:

- [299] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of ACM CHIL*, pages 222–235, 2020

- [212] Bret Nestor, Matthew McDermott, Geeticka Chauhan, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. *arXiv preprint arXiv:1811.12583*, 2018

- [40] Geeticka Chauhan, Matthew B.A. McDermott, and Peter Szolovits. REflex: Flexible framework for relation extraction in multiple domains. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 30–47, Florence, Italy, August 2019. Association for Computational Linguistics

- [131] Di Jin, Franck Dernoncourt, Elena Sergeeva, Matthew McDermott, and Geeticka Chauhan. MIT-MEDG at SemEval-2018 task 7: Semantic relation classification via convolution neural network. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 798–804, New Orleans, Louisiana, June 2018. Association for Computational Linguistics

- [41] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *In MICCAI International Conference*, pages 529–539. Springer, 2020

- [242] Keegan Quigley, Miriam Cha, Ruizhi Liao, Geeticka Chauhan, Steven Horng, Seth Berkowitz, and Polina Golland. Radtex: Learning efficient radiograph representations from text reports. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 22–31. Springer, 2022

- [243] Keegan Quigley, Miriam Cha, Josh Barua, Geeticka Chauhan, Seth Berkowitz, Steven Horng, and Polina Golland. Bidirectional captioning for clinically accurate and interpretable models. *arXiv preprint arXiv:2310.19635*, 2023

- [244] Keegan Quigley, Miriam Cha, Josh Barua, Geeticka Chauhan, Steven Horng, Seth Berkowitz, and Polina Golland. Improving medical visual representations through radiology report generation. *In anonymous review at conference*, 2024

- [42] Geeticka Chauhan, Steve Chien, Om Thakkar, Abhradeep Guha Thakurtha, and Arun Narayanan. Training Large ASR Encoders with Differential Privacy. *In anonymous review at conference*, 2024

Throughout my academic career, I have also explored other works that are not directly related to this thesis, including:

- [132] Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3):e1548, 2022

- [133] Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, 2021

- [128] Labiba Jahan, Geeticka Chauhan, and Mark A Finlayson. A new approach to animacy detection. In *Proceedings of the 27th International COLING*, 2018

- [36] Triana Carmenate, Peeraya Inyim, Nupoor Pachekar, Geeticka Chauhan, Leonardo Bobadilla, Mostafa Batouli, and Ali Mostafavi. Modeling occupant-building-appliance interaction for energy waste analysis. *Procedia Engineering*, 145:42–49, 2016

- [127] Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. Building on word animacy to determine coreference chain animacy in cultural narratives. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, pages 198–203, 2017

I have also had the pleasure of serving as a co-organizer for the NLP in Open Source Software Workshop at EMNLP 2020 [225] and EMNLP 2023 [283], along with the Women in Machine Learning Workshop at NeurIPS 2021.

Part I

Modeling heterogeneous and noisy medical data

Chapter 2

Developing feature robustness among heterogeneous EHR data

Abstract

In this thesis, we develop machine learning methods that leverage the unique structure of medical data along with available external knowledge. This chapter primarily highlights two works [212, 299] that discuss important considerations and contributions toward the deployability effort in machine learning (ML) for health models. The work [212] formed the foundational basis for the extended work [213].

Two key pre-requisites for deployability of ML for health models are 1) model generalizabilty: a model's ability to sustain performance over time as care practices, database systems and population demographics evolve and 2) reproducibility: developing pre-processing and modeling strategies that are easy to understand and open-source. In ML for healthcare, we often rely on de-identified datasets with randomly shifted calendar dates and experimental practices that are time agnostic. As a result, care records are often assigned to train or test sets without regard for the actual dates of care, which hurts the generalizability of developed models. Additionally, despite the wide availability of MIMIC-III [136], Electronic Health Record (EHR) data offers a steep learning curve for beginner researchers in the machine learning for healthcare space. This is counter-productive to the goals of data availability and realizing the potential of ML for healthcare, as the barrier to entry leads to duplicated efforts and slower modeling advances due to worse reproducibility [137, 197].

The contributions of the two works are as follows: 1) we create novel clinically aggregated features for the open source electronic health record (EHR) dataset called MIMIC-III [136] and perform detailed ablations to show robustness in area under the curve (AUC) performance across changing hospital systems over time, and 2) we make our feature aggregation, along with our pre-processing of the MIMIC-III data into an open-source framework that others in the community can openly use.

2.1 Introduction

The shift towards electronic health records (EHR) in modern healthcare systems has enabled the secondary use of these records for machine learning model development for mortality risk [103], sepsis treatment [250] and many other promising applications [202, 173, 251]. Most recently, there has been a push towards deployment of machine learning in healthcare.

With the Food and Drug Administration (FDA) being increasingly interested in regulating AI for health [79], it is even more important to deploy ML in a safe and generalizable manner. Some examples of ML in health already in deployment are drug discovery methods [276] being used by several pharmacological companies such as Pfizer and BioNtech [194, 226, 206]. However, the deployability of machine learning has several concerns. In the case of drug discovery, this issue is commonly related to lack of relevant data for training [206]. In this chapter, we target two issues hindering the deployability of ML in health models for EHR data: 1) model generalizability over time, and 2) reproducibility of data pre-processing and modeling techniques.

2.1.1 Challenge 1: Model Generalizability

Due to the sensitive nature of patient information, EHR data is typically de-identified in order to reduce risk to patients prior to its use in research. A well-known example of publicly-available, de-identified EHR data is the MIMIC-III database [136], which contains information about intensive care unit (ICU) patients from the Beth Israel Deaconess Medical Center (BIDMC).

A crucial step in de-identification is obscuring calendar dates related to care. In the MIMIC-III dataset, dates are shifted into the future between the years "2100 and 2200" by a consistent random offset for each patient [136]. As a result of the de-identification procedure, modeling in a temporally consistent manner becomes challenging (e.g. training on historical data and evaluating on future data) and several related works [103, 237, 50] use time-agnostic evaluation protocols. These time-agnostic protocols do not account for a significant source of error that would affect models during true deployment: evolution of care practices and the resultant concept drift.

The concept drift [341] caused by shifts in the calendar date are important to account for due to their ability to induce significant differences in clinical data [251, 153]. These are reflected in MIMIC-III through an EHR system update from Carevue[1] to Metavision[2] in the Beth Israel Deaconess Medical Center in 2008 [136]. This shift caused fundamental changes in the way every clinical measurement was recorded in the EHR (yielding entirely new database tables with new variable names). A recent update of MIMIC-IV [140] provides the additional `anchor_year_group` column to allow analyses which incorporate changes in medical practice over time. Future work can corroborate the results from our study with this newly created dataset.

[1] `https://mimic.physionet.org/mimicdata/carevue/`

[2] `https://mimic.physionet.org/mimicdata/metavision/`

Core challenge Developing time-agnostic models on de-identified EHR data can therefore hurt model generalizability because the data used to train the model would be significantly different from that used to test it in deployment, as a result of concept drift. Even without hospital system changes, random date shifts as part of MIMIC might cause models to unintentionally train with data generated from newer care practice than they are tested on, which could result in unexpected model outputs from the machine learning model.

2.1.2 Challenge 2: Reproducibility

To realize the potential of applying machine learning to observational health data, several efforts had been made to make healthcare data widely available to credentialed researchers with human subjects training. An effort in this direction is the publicly available healthcare dataset, the Medical Information Mart for Intensive Care (MIMIC-III) [136].

Despite the wide availability of MIMIC-III, Electronic Health Record (EHR) data offers a steep learning curve for beginner researchers in the machine learning for healthcare space. This is counter-productive to the goals of data availability and realizing the potential of machine learning for healthcare, as the barrier of entry would lead to duplicated efforts and slower modeling advances due to worse reproducibility.

The extent and dangers of lack of reproducibility are highlighted in [137, 197]. Johnson et al. [137] highlight the lack of reproducibility about the patient cohorts along with lack of consistent reporting of model design and methodology as core issues affecting reproducibility in machine learning for critical care. McDermott et al. [197] find a lack of technical, statistical and conceptual replicability through a broader study across the machine learning for healthcare community.

Core challenge It is clear that the primary difficulties of working with EHR data rest in the complexity of the data and the myriad choices that must be made to extract a clinically-relevant cohort for analysis. These same difficulties hinder the reproducibility of studies that apply machine learning to MIMIC-III data, because researchers develop code independently to extract and preprocess task-appropriate cohorts. The majority of papers do not share code used to extract study-specific data [137], resulting in expensive yet redundant efforts to build upon existing work and creating the potential for hard-to-explain differences in results.

2.1.3 Our Contributions

Model generalizability Towards the model generalizability challenge, we introduce a *clinically aggregated* feature representation to improve model robustness across changing hospital systems. We find that the choice of input representation substantially impacts how robust a model is to changing care practices. Specifically, we make the following contributions:

1. We show that models using raw, non-featurized data representations are unable to generalize well across large dataset shifts as exemplified by the 2008 EHR system update within MIMIC.

2. We introduce novel clinically aggregated representations by analyzing related data items across varying EHR tables such as lab results and bedside measurements that reduce overall data missingness and the presence of duplicate measurements. Our effort condensed close to 13000 measurements into 100 groups.

3. Compared to raw feature representations, our clinically aggregated representations nearly eliminate all performance deterioration across temporal shifts by reducing deterioration by close to 5-fold for in-ICU mortality prediction (0.29 versus 0.06 AUROC) and close to 3-fold for long length-of-stay (0.1 versus 0.03 AUROC).

Reproducibility Towards the reproducibility challenge, we introduce `MIMIC-Extract`, an open source pipeline to streamline data pre-processing of MIMIC-IIIv1.4, including unit conversion, outlier handling and feature selection. We intend this pipeline to serve as a foundation for both benchmarking the state-of-the-art and enabling progress on new research tasks. We advance the field with three primary contributions:

1. Robust representations of labs and vitals time series with standardized units, outlier correction and clinical meaningfulness

2. Clinically meaningful interventions and outcomes such as providing hourly-observed treatment signals for blood pressure management and outcomes such as mortality and length of stay

3. Our pipeline with a focus on usability, reproducibility and extensibility. For example, our patient selection criteria can be easily adjusted, requiring changes to only keyword arguments rather than source code and thus preventing the user from making potentially confusing pre-processing choices.

2.2 Related Works

2.2.1 Feature Robustness across time-varying changes in EHR

The standard de-identification process for electronic health record (EHR) datasets like MIMIC-III [136] make it challenging to analyze the data in a temporally consistent way. As a result, the wide literature on MIMIC-III [103, 237, 50] use time-agnostic evaluation protocols that do not account for evolution of care practices and the resultant concept drift [341, 251, 153]. The closest methods have considered automated mapping of clinical data elements with mapping tools [92] or learned vector space embeddings [251]. However, it is unknown whether these methods can withstand fine-grained time-varying changes in EHR. To the best of our knowledge, researchers

have not yet assessed how robust state-of-the-art models trained on MIMIC-III are to temporal drift. In this work, we use a Limited Data Use Agreement allowing restricted access to the underlying calendar year of each event within MIMIC-III to perform such an assessment. We examine how a popular model architecture generalizes to unseen future-only data through our proposed *clinically aggregated* feature representation and different time-aware training regimes. We also demonstrate how models using raw, non-featurized data representations, as advocated by deep learning ICU prediction systems such as [237], are universally unable to generalise well across large dataset shifts as exemplified by the 2008 system switch within MIMIC-III.

2.2.2 Public EHR pre-processing frameworks

We intend `MIMIC-Extract` to serve as a foundation for both benchmarking the state-of-the-art and enabling progress on new research tasks. Several other recent works have developed, in parallel, extraction pipelines and prediction benchmark tasks for MIMIC-III data [103, 237, 284]. However, compared to these, we advance the field by introducing robust representations of labs and vitals time series, clinically meaningful interventions and outcomes and a focus on usability, reproducibility and extensibility. Our pipeline has been used as the foundation for reproducing many recent machine learning studies of MIMIC-III data [84, 85, 309, 86, 196, 213, 250, 280, 212].

2.3 Efficacy of clinically aggregated representations towards model generalizability

We focus on two binary prediction tasks, mortality and long length-of-stay, which are commonly studied for applying machine learning to the MIMIC-III critical care setting. In Figure 2-1, we describe the full prediction pipeline of our method.

2.3.1 Data and processing

We use MIMIC III, a public dataset with EHR data from over 58,900 hospital admissions of nearly 38,600 adults at Beth Israel Deaconess Medical Center from 2001 to 2012 [136]. Within the MIMIC-III dataset, each patient may be admitted on multiple occasions to the hospital, and may be transferred to and from the intensive care unit (ICU) multiple times. We choose to focus on a patient's first exposure to the ICU (by far the most common case), avoiding the complications of those that transfer multiple times. We thus extract a targeted cohort of patient EHR data corresponding to the *first* ICU visit. We include only ICU stays that lasted at least 36 hours. We also focus on non-paediatric cases by requiring all patients to be over 15 years old. These criteria, which broadly follow prior work [86, 280, 196], result in a cohort of 21,877 unique ICU stays.

We use the first 24 hours of data for each patient, and collect physiological measurements into hourly buckets via averaging. Several works have focused on imputation methods for healthcare data [43, 288, 317, 130, 5, 177]. We use simple imputation

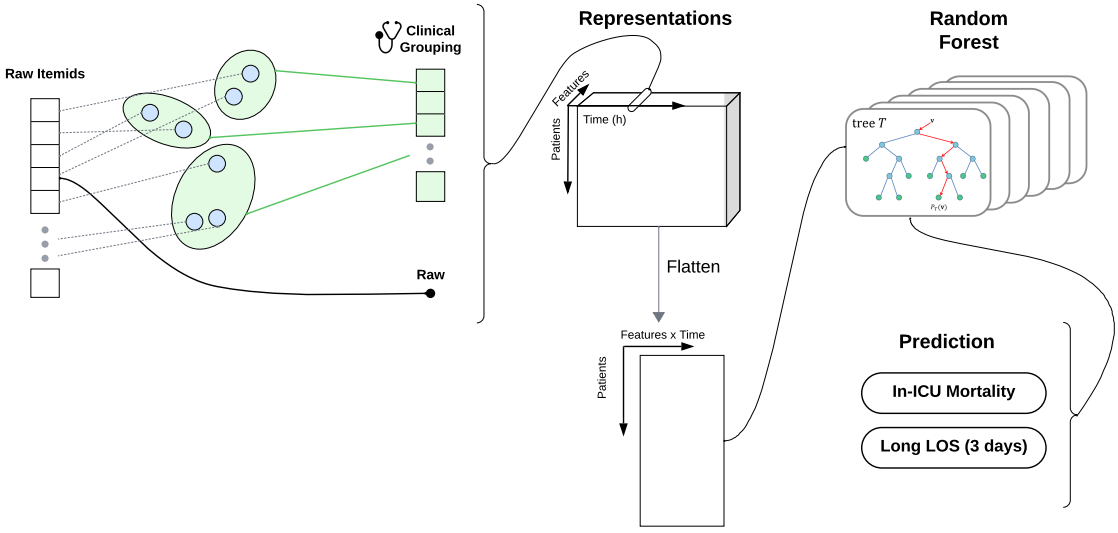


Figure 2-1: The full experimental pipeline. We provide code for reproduction of experiments with the assumption that researchers have obtained the limited-use years data mapping for patient identifiers. Figure inspired by [213].

to assign three sub-features to the data: the imputed (forward-filled) measurement of the feature, a binary indicator of whether or not that feature was observed at that time, and the number of hours since the feature was last observed [43].

2.3.2 Methods

Data Representations For each patient, we investigate the impact of two common data representation strategies on tasks.

1. **Raw ItemID**: The Raw `ItemID` representation is the simplest; we include all selected 181 labs and vitals, each identified via a unique `ItemID` code in the MIMIC database. Due to the `ItemID`s being explicitly connected to the underlying EHR software, the MIMIC-III shift in 2008 from Carevue to Metavision caused several old features to not be used anymore. Additionally, some vitals such as "Mean Arterial Blood Pressure" (`ItemID 6702`) spontaneously increased their frequency of recording in 2004 (Figure from Appendix A i.e. A-1). This resulted in highly sparse representations full of missing values before imputation.

2. **Clinical Aggregations**: We use expert knowledge to manually define groupings of `ItemID`s that span the discrepancies between Carevue and Metavision, such as by grouping `ItemID` values "Heart Rate" under CareVue (211) and MetaVision (220045). The groupings also gather together `ItemID`s which measure the same biophysical quantity merely through different means, such as aggregating MetaVision `ItemID`s 225664 ("Glucose finger stick"), 220621 ("Glucose (serum)"), and 226537 ("Glucose (whole blood)") into one unified category

for glucose blood sugar. The resulting representation groups all 13000 raw `ItemID`s into close to 100 clinically meaningful categories, and yields a dataset with a rate of 78.25% missingness before imputation compared with the raw `ItemID` representations that have a $> 90.6\%$ rate of missingness. The detailed aggregation table is shown in Appendix A section A.1.

Model We use a random forest (RF) classifier for all tasks with simple imputation to handle missing data, similar to the implementation in [43]. RF classifiers are non-linear, defined using bagged decision trees and are often competitive baseline methods. RF implementation in SciKit Learn's `RandomForestClassifier` class is used. Before feeding the 24-hour time-series data to the model, they are flattened along the time dimension as seen in Figure 2-1.

Evaluation Tasks We test on two common baseline clinical machine learning tasks: mortality prediction and long length-of-stay (LOS) predicted that have been commonly used as prediction targets in past works [103, 237]. The **In-ICU** mortality task is defined by patient death within the ICU. The **long LOS** task is realized as a classification task by splitting patients between high LOS (≥ 3 days) and low LOS (< 3 days), where 3 days happens to be the median LOS for our patient cohort.

Experiments In addition to measuring the year-agnostic performance of our model (i.e. the way models are typically run, with no knowledge of the admission year), we report results on three training paradigms that are reflective of mechanisms that can be applied on historical data before deploying a clinical model. The four different paradigms are:

1. **Year-agnostic** Training and testing on randomly shuffled data, with no knowledge of the year of care.

2. **2001-2002** Training on data from 2001–2002 only, and then testing on all future years, reflecting the situation where a practitioner trains a model on historical data but doesn't update the training set containing more recent data.

3. **Prior Year** Training on the data from the prior year *only*, e.g., data from 2005 will be used to train a model that is tested on data from 2006. This reflects the situation where a model is deployed and updated with yearly frequency by training on only the prior year's data.

4. **Full history** Training on *all* prior data, e.g., data from 2001–2005 will be used to train a model that is tested on data from 2006. This reflects the situation where a model is deployed and updated yearly by training on all data ever observed.

2.3.3 Results

We present the average AUROC per representation across all years for both In-ICU mortality prediction and long LOS tasks in Figure 2-2. This performance deterioration across years of training for 2001-2002, Prior Year and Full History training regimes are reflective of mechanisms that can be applied on historical data before deploying a clinical model. Additionally, we report 2 set of results: 1) Year-agnostic training regime in table 2.1, 2) Full history training regime in table 2.2, to compare the robustness of our proposed clinically aggregated representations across historically typical training regimes versus the true year-averaged *Full history* training regime.

Year-agnostic Results We report the year-agnostic AUROC scores for both in-ICU mortality and long LOS tasks across 5 x 2 fold cross validation splits [65] in table 2.1. This is a representation of typical machine learning model performance when trained on electronic health records.

Task	Average AUROC for Random Splits	
	Raw ItemID	Clinical
in-ICU mortality	0.82 ± 0.02	0.86 ± 0.02
long LOS	0.70 ± 0.00	0.71 ± 0.01

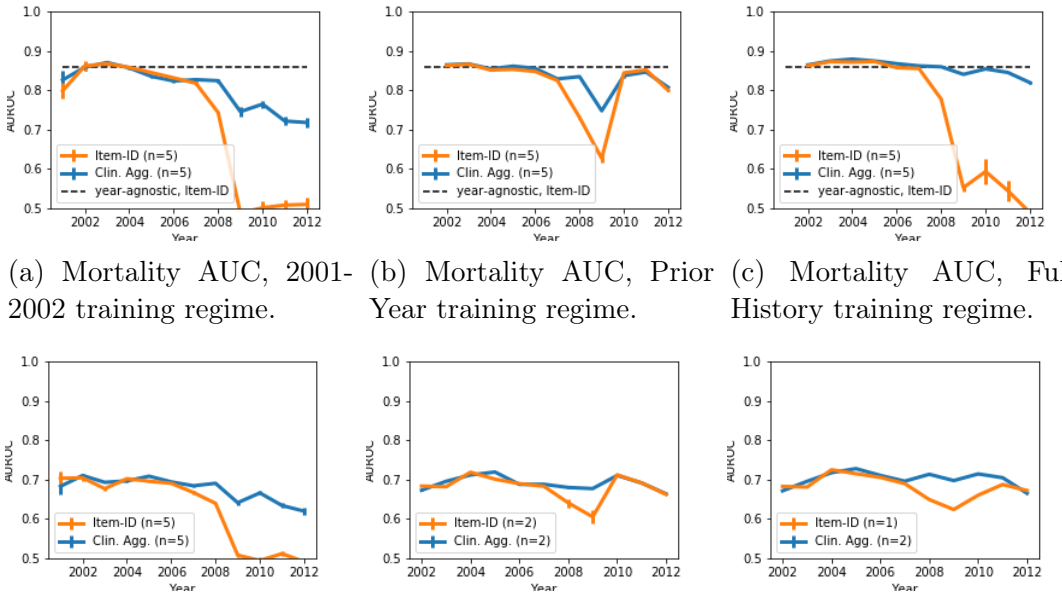
Table 2.1: The in-ICU and long LOS model performance when trained in a year-agnostic fashion. The AUROC (mean $\pm$ std) is reported and results are consistent with those reported in [213]

Full History Results We compute the AUROC over each unseen year from 2003 onward for reporting the average and standard deviation using the *Full History* training regime. Between each of these results, we also report the maximum drop of AUROC observed between the first year of evaluation and subsequent years' performance from 2003 onward. We report results in table 2.2. This table shows that the clinical representation tends to improve the overall performance and decreases the magnitude of performance deterioration during non-stationary healthcare practice.

Performance deterioration across years of training In figure 2-2 we report the AUROC across the 2001-2002, prior year and full history training paradigms over time. Overall, we note that the Clinical Aggregate representation is much more robust to the performance degradation over time observed under the `Item-ID` representation. This resembles the findings in [92], though our clinically determined groupings appear to offer a lower drop in performance across the shift in practice than their learned representations.

Task	Average AUROC		Max AUROC Drop	
	Raw ItemID	Clinical	Raw ItemID	Clinical
In-ICU Mortality	0.76 ± 0.13	$\mathbf{0.85 \pm 0.02}$	0.29	**0.06**
long LOS	0.67 ± 0.04	$\mathbf{0.68 \pm 0.03}$	0.10	**0.03**

Table 2.2: A comparison of the a) average ($\pm$ standard deviation) AUROC over each unseen year from 2003 onward, and b) max loss observed between the first year of evaluation and subsequent years' performance from 2003 onward for the *Full history* training regime. **Bold** indicates best performance. Bigger is better for averages, while smaller is better for maximum loss and standard deviation. Results consistent with those reported in [213].



(a) Mortality AUC, 2001-2002 training regime.

(b) Mortality AUC, Prior Year training regime.

(c) Mortality AUC, Full History training regime.

(d) LOS AUC, 2001-2002 training regime.

(e) LOS AUC, Prior Year training regime.

(f) LOS AUC, Full History training regime.

Figure 2-2: Performance of RF classifier using Raw `ItemID` and `Clinically Aggregated` representations on In-ICU mortality (top) and long LOS prediction (bottom). Error bars indicate $\pm$ standard error.

2.3.4 Discussion

In this study, we find consistent evidence of the benefits of our proposed *Clinical Aggregate* representations where we manually group related `ItemID`s based on clinical knowledge. There are several interesting observations:

1. Clinical Aggregate Representations are more robust than raw `ItemID` representations: Models trained on the raw `ItemID` representation suffer a rapid performance decrease in 2008, after the EHR change, as observed in figure 2-2. The clinically aggregated representation maintains much more consistent per-

formance across years for the RF classifier, and the continuing work in [213] shows that this finding holds for multiple model architectures.

2. Year-agnostic training overstates performance, especially for raw `ItemID` representation: We replicate the year-agnostic training and test practice common to most reporting in machine learning papers, and found that this method creates an unrealistic upper bound to model performance, *especially* on the raw representation. For example, RF models report a year-agnostic mortality AUROC of 0.82 ± 0.02 (5 x 2 fold CV splits [65]) in table 2.1, as compared to their true year-averaged AUROC under the raw representation of 0.76 ± 0.13 in table 2.2. Under the *Clinical Aggregate* representation, in contrast, RF reports a year-agnostic mortality AUROC of 0.86 ± 0.02, in comparison to the true year-averaged AUROC of 0.85 ± 0.02.

3. Models Saturate Quickly on Mortality Prediction, Impacting Generalisation: By profiling the changes in model performance over time, we find evidence to suggest that both of the tasks considered (each of which are commonly studied) require relatively few years of aggregated data to saturate in prediction quality. Looking in figure 2-2 under the *Full History* training regime and the *Clinical Aggregates* representation, model performance is very steady from the beginning of the training period where only one year of data is used. This issue has been further studied in detail for the in-ICU mortality prediction task in [212].

2.4 An overview of `MIMIC-Extract` towards reproducibility in ML for health

In this section, we summarize `MIMIC-Extract`, an open source pipeline to extract, preprocess, and represent data from MIMIC- III v1.4, including static demographic information available at admission, vitals and laboratory measurements, intervention signals, and static outcomes such as length-of-stay or mortality. Figure 2-3 gives a visual summary of the data we extract from the observed records of an individual patient stay available in MIMIC-III. Our principled approach yields a comprehensive cohort of time-series data that is well-suited for several clinically-meaningful prediction tasks, several of which we profile here, while simultaneously providing flexibility in cohort selection and variable selection.

We intend this pipeline to serve as a foundation for both benchmarking the state-of-the-art and enabling progress on new research tasks. Our pipeline has been used as the foundation for reproducing many recent machine learning studies of MIMIC-III data [84, 85, 309, 86, 196, 213, 250, 280, 212].

2.4.1 Data Pipeline Overview

Figure 2-4 summarizes the data extraction and processing steps involved in MIMIC-Extract. From the MIMIC relational database, SQL query results are processed to

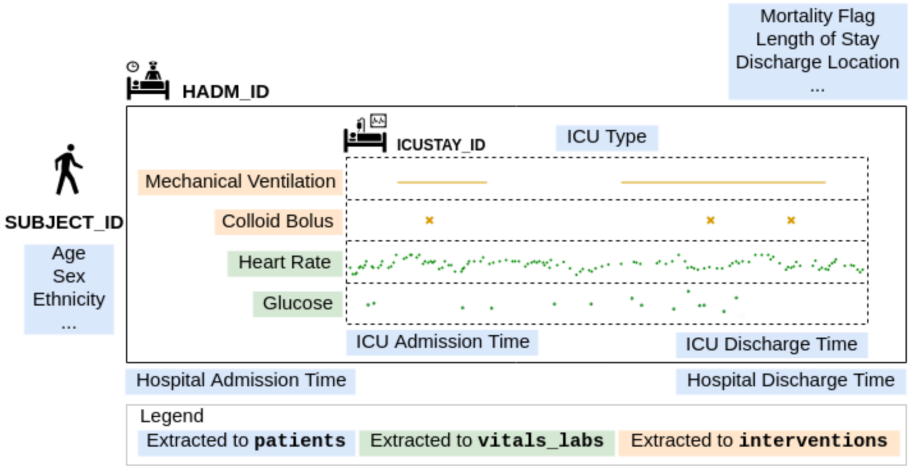


Figure 2-3: Example data produced by `MIMIC-Extract` to summarize a single subject's stay in the intensive care unit(ICU). Time evolves on the x-axis, and all extracted time series are discretized into hourly buckets. Mechanical Ventilation is an example intervention with multi-hour continuous duration. Colloid bolus is an example of an intermittent fluids intervention. All interventions are recorded as binary indicators at each hour. Heart Rate is an example of a frequent vital sign. Glucose is an example of an infrequent lab measurement.

generate four output tables. These tables maintain the time series nature of clinical data and also provide an aggregated featurization of the cohort selected.

1. **Cohort Selection**: Our proposed pipeline includes all patient ICU stays in the MIMIC-III database that meet the following criteria: the subject is an adult (age of at least 15 at time of admission), the stay is the first known ICU admission for the subject, and the total duration of the stay is at least 12 hours and less than 10 days. This cohort selection is consistent with many previous papers using MIMIC-III [84, 85, 309, 86, 196, 213, 250, 280, 212].

2. **Variable Selection**: By default, our extraction code extracts various static demographic variables such as age, ethnicity etc., along with static outcomes such as in-ICU mortality, in-hospital mortality, and the patient's total ICU length-of-stay (LOS), in hours. Our pipeline presents values for static variables as they originally appear in MIMIC-III raw data with no additional outlier removal. For time-varying vitals and labs, our extraction code extracts 104 clinically aggregated time-series variables (listed in Appendix A section A.1) related to vital signs (e.g., heart rate) and laboratory test results (e.g., white blood cell counts).

3. **Unit Conversion and Outlier Detection**: Our data pipeline standardizes measurements into consistent units, including weight into kilograms, height into centimeters, and temperature into degrees Celsius. This process is easily extensible if any additional unit-classes are added by downstream users. To handle

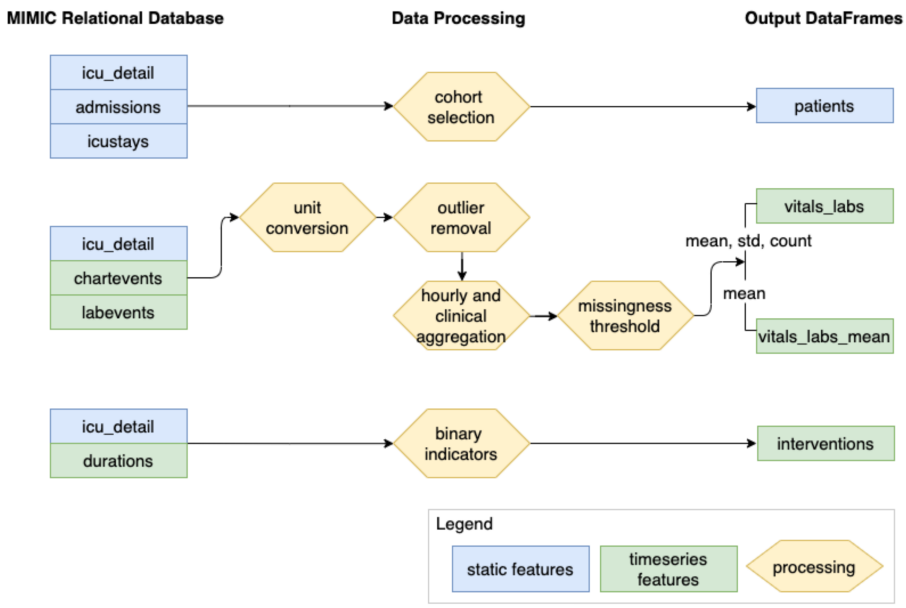


Figure 2-4: MIMIC-Extract Overview: First, a cohort is created that meets our selection criteria. Static demographic variables and ICU stay information for patients in the cohort are extracted and stored in patients. Next, labs and vitals for patients in the cohort are extracted and stored in vital_labs and vitals_labs_mean. By default, only labs and vitals that are missing less frequently than a predefined threshold are extracted and outlier values are filtered based on physiological valid ranges. Finally, hourly intervention time series for the same patients are extracted and stored in interventions.

outliers, we make use of a list of clinically reasonable variable ranges provided in the source code repository of [103]. We mark raw observed values as missing if they fall outside these ranges. Additionally, each variable is associated with more refined upper and lower thresholds for defining the physiologically valid range of measurements. Any non-outlier value that falls outside the physiologically valid range is replaced with the nearest valid value.

4. **Hourly Aggregation**: To obtain a denser representation for each laboratory measurement and recorded vital sign, that is easier to reason about and readily applied to modern machine learning methods for time-series that expect discretized time representations, we aggregate the observations from each ICU stay's time-series into hourly buckets.

5. **Semantic Grouping of Raw Features into Clinical Aggregates**: We make our clinical aggregations described in section 2.3.2 available as part of the `MIMIC-Extract` pipeline.

6. **Time-varying treatment labels**: Our code extracts hourly binary indicators of when (if ever) common treatments were provided to each patient over time.

We include device treatments such as mechanical ventilation, as well as drug treatments such as vasopressors and fluid boluses.

7. **Extensibility of data pipeline**: While MIMIC-Extract promotes reproducibility by providing a default cohort for common benchmark tasks, it is also able to extract data tailored to specific research questions. We offer several modifications to our framework such as keywords to change cohort selection, configurable variable grouping and outlier resource files, embedded SQL queries in the extraction code and extending the pipeline to extract variables such as prescriptions or caregiver notes.

8. **Output cohort characterization**: Our pipeline produces a cohort of 34,472 patients by default with diverse demographic and admission coverage, as summarized in table 2.3.

		Gender		Total
		F	M	
Ethnicity	Asian	370	472	842 (2%)
	Hispanic	448	689	1,137 (3%)
	Black	1,448	1,219	2,667 (8%)
	Other	2,061	3,122	5,183 (15%)
	White	10,651	13,992	24,643 (71%)
Age	<30	748	1,084	1,832 (5%)
	31-50	2,212	3,277	5,489 (16%)
	51-70	4,888	8,054	12,942 (38%)
	>70	7,130	7,079	14,209 (41%)
Insurance Type	Self Pay	125	352	477 (1%)
	Government	402	648	1,050 (3%)
	Medicaid	1,186	1,596	2,782 (8%)
	Private	4,415	7,431	11,846 (34%)
	Medicare	8,850	9,467	18,317 (53%)
Admission Type	Urgent	409	528	937 (3%)
	Elective	2,282	3,423	5,705 (17%)
	Emergency	12,287	15,543	27,830 (81%)
First Careunit	TSICU	1,777	2,725	4,502 (13%)
	CCU	2,185	3,008	5,193 (15%)
	SICU	2,678	2,842	5,520 (16%)
	CSRU	2,326	4,724	7,050 (20%)
	MICU	6,012	6,195	12,207 (35%)
Total		14,978 (43%)	19,494 (57%)	34,472 (100%)

Table 2.3: Default Cohort Summary by Static Demographic and Admission Variables

2.4.2 Benchmark Tasks and Models

Tasks We support the following benchmark tasks:

1. **Mortality prediction and length-of-stay (LOS)**: Risk prediction tasks like mortality and long LOS predictions are highlighted as benchmark tasks in section 2.3.2. We consider several varieties of these tasks, including in-ICU mortality, in-hospital mortality, LOS > 3 days prediction, and LOS > 7 days prediction. For all tasks, we use clinically grouped time-varying labs and vitals features alone to predict these targets as binary classification task.

2. **Clinical Intervention Prediction**: We target two interventions, mechanical intervention and vasopressors, which are common among prior work targeting clinical intervention prediction [309, 86, 280]. Well-executed intervention prediction can alert caregivers about administrating effective treatments while avoiding unnecessary harms and costs [309, 86]. In a high-paced ICU, such decision-support systems could be a fail-safe against catastrophic errors. We extract clinically aggregated outputs (as described in Section 2.3.2) over a sliding window of size 6 hours as input features, then predict intervention onset/offset within a 4 hour prediction window offset from the input window by a 6 hour gap window. For each intervention at each prediction window, there are 4 possible outcomes:

   Onset: When the intervention begins off and is turned on.

   Stay On: When the intervention begins on and stays on.

   Wean: When the intervention begins on and is stopped.

   Stay Off: When the intervention begins off and stays off.

Data Preprocessing Time-varying lab and vital data are preprocessed by adjusting their values to be mean centered and scaled to unit variance. Then, missing data was imputed using a variant of the "Simple Imputation" scheme outlined in [43]. For the clinical intervention prediction task, we additionally include 5 static variables (gender, age bucket, ethnicity, ICU type, and admission type) and time-of-day as additional features.

Models For mortality and LOS tasks, we profiled logistic regression (LR), random forest (RF), and gated recurrent unit with delay (GRU-D) [43] models. For clinical intervention prediction tasks, we profile LR, RF, convolutional neural network (CNN) models, and Long Short-Term Memory (LSTM) models.

Results For mortality and LOS tasks, we report results in table 2.4. Our AUROCs are very much in line with the literature for these tasks, fully allowing reproducibility of models for practitioners interested in extending our work. One interesting observation is that random forest models often have poor F1 scores, even while maintaining competitive AUPRC scores. This may indicate that these models are more sensitive

to the initial choice of threshold than are other models. Similarly, GRU-D often displays stronger performance under the AUPRC metric than the AUROC metric relative to other models, which likely speaks in its favor here given the strong rates of class imbalance in these tasks.

For clinical intervention prediction tasks, we report results in table 2.5. We find that CNN and LSTM models perform very similarly to prior studies. This is notable given we *do not* include notes, whereas many prior studies do [280]. RF models perform surprisingly well, outperforming CNN and LSTM models and prior results reported in the literature.

Task	Model	AUROC	AUPRC	Accuracy	F1
In-ICU Mortality	LR	88.7	46.4	93.4%	38.4
	RF	89.7	49.8	93.3%	12.6
	GRU-D	89.1	50.9	94.0%	43.1
In-Hospital Mortality	LR	85.6	49.1	91.1%	42.1
	RF	86.7	53.1	90.7%	19.6
	GRU-D	87.6	53.2	91.7%	44.8
LOS > 3 Days	LR	71.6	65.1	68.6%	59.4
	RF	73.6	68.5	69.5%	59.5
	GRU-D	73.3	68.5	68.3%	62.2
LOS > 7 Days	LR	72.4	18.5	91.9%	7.2
	RF	76.4	19.5	92.3%	0.0
	GRU-D	71.0	17.9	91.2%	10.7

Table 2.4: Performance Results on In-ICU Mortality, In-Hospital Mortality, > 3 Day LOS, and > 7 Day LOS. (Note that due to their additional computational overhead, GRU-D models were undersampled during hyperparameter turning as compared to LR and RF models.)

2.4.3 Discussion

We intend `MIMIC-Extract` to serve as a foundation for both benchmarking the state-of-the-art and enabling progress on new research tasks. Compared to other relevant frameworks [103, 237, 284], we provide the only pipeline that generates a generic cohort that can be directly read in a simple format like Pandas DataFrame. Ours is also the only pipeline that uses clinical aggregation, unit conversion and outlier detection on a large set of raw MIMIC-III data. Ours is also the only work demonstrating an intervention prediction task through predicting the onset, offset, stay on, and stay off of mechanical ventilation and vasopressors. This task requires the model to handle the decisions needed in a real ICU where subjects may go on and off treatments throughout their stay using most recently observed data. We also offer easy extensibility to allow `MIMIC-Extract` to be used with other common predictive tasks such

	RF		LR		CNN		LSTM	
	Vent.	Vaso.	Vent.	Vaso.	Vent.	Vaso.	Vent.	Vaso.
Onset (AUROC)	87.1	71.6	71.9	68.4	72.2	69.4	70.1	71.9
Wean (AUROC)	94.0	94.2	93.2	93.9	93.9	94.0	93.1	93.9
Stay On (AUROC)	98.5	98.5	98.4	98.2	98.6	98.4	98.3	98.3
Stay Off (AUROC)	99.0	98.3	98.3	98.5	98.4	98.1	98.4	98.1
Macro AUROC	94.6	90.7	90.4	89.8	90.8	90.0	90.0	90.1
Accuracy	79.7	83.8	78.5	72.9	61.8	77.6	84.3	82.6
Macro F1	48.1	48.9	47.7	45.1	44.4	44.4	50.1	48.1
Macro AUPRC	42.7	42.0	43.1	40.2	42.4	38.9	44.4	41.7

Table 2.5: Performance Results on Mechanical Ventilation and Vasopressor Prediction

as ICD-9 group classification or acute respiratory failure (ARF).

Our pipeline, despite its many benefits, also has some limitations due to design choices more relevant to our benchmark tasks. We exclude features such as prescriptions, certain labs and vitals, various treatments/interventions, and notes that might be beneficial in prediction of other common tasks. However, many of these features can be externally extracted and joined to our pipeline's output. In addition, the time-series coarsening into hourly buckets can also be limiting for certain tasks such as care delivery [4]. Similarly, our clinical groupings, while offering high task performance, are also manually curated and limit the extensibility of the pipeline to new labs and vitals.

2.5 Conclusion

Realizing the potential of machine learning (ML) for healthcare towards deployment is a multi-faceted effort, made possible by the availability of vast amount of digitized electronic health record (EHR) data such as MIMIC-III [136]. Several advances in the academic ML for health community have either reached advanced stages pre-deployment, or have already been deployed by large biotech companies. Some examples include the MedKnowts effort [209, 94] towards smarter autocomplete based EHR and drug discovery methods [276] being used by several pharmacological companies such as Pfizer and BioNtech [194, 226, 206].

In this chapter, we target two important issues hindering the deployability of ML in health models for EHR data: 1) model generalizability, and 2) reproducibility of data pre-processing and modeling techniques. Towards the model generalizability challenge, we demonstrate how standard year-agnostic models developed on EHR data fail to generalize over time and across changing hospital systems. We introduce robust clinically aggregate feature representations that nearly eliminate performance deterioration across temporal shifts on two key prediction tasks: in-ICU mortality prediction

and long length-of-stay. We analyze standard year-agnostic training scheme versus the year-aware training scheme to confirm the robustness of our method, along with presenting detailed year-by-year model performance spanning the Carevue to Metavision system change in 2008 for MIMIC-III. Towards the reproducibility challenge, we introduce `MIMIC-Extract`, an open source pipeline to streamline data pre-processing of MIMIC-IIIv1.4, including unit conversion, outlier handling and feature selection (including our clinical aggregate feature representations). This pipeline offers robust representations of labs and vitals time series, clinically meaningful interventions and outcomes and easy usability and extensibility. Our pipeline has served as the basis for several ML studies of MIMIC-III data [84, 85, 309, 86, 196, 213, 250, 280, 212].

This thesis develops methods that leverage the unique structure of medical data along with available external knowledge to advance ML for healthcare. The methods introduced in this chapter streamline the development of ML models for heterogeneous EHR data, reducing the barrier of entry in this technically challenging area. Our novel clinically aggregated embeddings are developed by grouping related measurements, using knowledge from healthcare and statistics. Additionally, `MIMIC-Extract` offers several benefits such as cohort selection, outlier handling and standard models developed on EHR data, making it straightforward for practitioners to pre-process MIMIC-III data and develop ML models with it. The findings of this chapter emphasize that developing features using clinical knowledge increases the robustness and performance capabilities of ML for health models. We believe the best way forward for machine learning in health is to marry popular end-to-end approaches with those leveraging explicit structure and feature engineering to counteract the trade-offs between high performance and needing large amounts of clean data for training, in line with recommendations from Battaglia et al. [15].

Chapter 3

Streamlining relation extraction for noisy medical text

Abstract

In this thesis, we develop approaches that bring together feature representations leveraging external knowledge with end-to-end machine learning methods for modeling complex medical data. This chapter is based on two works [40, 131], the second of which is featured in my S.M thesis.

An important consideration towards advancing state-of-the-art in ML for health is the reproducibility of studies to allow for compounding and transparent scientific improvements. We tackle reproducibility challenges in the field of relation extraction (RE), an important task allowing for automatic extraction of relational knowledge from scientific and medical literature and forming an important component of Natural Language Understanding (NLU). After conducting a thorough literature review (as of the publication), we find that reproducibility is hindered due to 1) many experiments in the field not being described precisely enough, and 2) many papers failing to report ablation studies that would highlight the relative contributions of their various combined techniques, to clearly highlight the techniques that offer the most improvement. As a result, there is a lack of consensus in the field on techniques that will generalize to novel tasks, datasets and contexts.

Our main contributions are in: 1) introduction of a highly accurate RE model that can effectively extract relational information from scientific abstracts, 2) development of a unifying and extendable framework for RE (known as `REflex`) that allows for easy exploration into the missing ablation studies and identifies best design practices for accessibility to new researchers. Additionally, the systematic exploration of modeling, pre-processing and training methodologies using our framework reveal that choices of pre-processing using external knowledge are a large contributor to performance and that omission of such information can further hinder fair comparison.

3.1 Introduction

Relation Extraction (RE) has gained a lot of interest from the community with the introduction of the Semeval tasks from 2007 by Girju et al. [90] and 2010 by Hendrickx et al. [111]. The field is a subset of information extraction (IE) with the goal of finding semantic relationships between concepts in a given sentence, and is an important component of Natural Language Understanding (NLU). Applications include automatic knowledge base creation, question answering, as well as analysis of unstructured text data. Since the introduction of RE tasks in the general and medical domains, many researchers have explored the performance of different neural network architectures on the datasets.

However, the findings as of our published papers suggest that progress in RE is hampered by reproducibility issues as well as the difficulty in assessing which techniques in the literature will generalize to novel tasks, datasets and contexts. This chapter introduces `REflex`, an open source unifying framework for RE, that allows researchers to perform various modeling and model-complementing explorations on a new dataset of their choice. This chapter also highlights RE methods that are effective at extracting relationships from scientific article abstracts, which form the basis of our technical contributions in `REflex`.

3.1.1 Our Contributions

Given the lack of detailed evaluation studies in RE, it is difficult to assess the causes of large variability of results, which makes a *fair comparison* of models a difficult task. An open-source unifying framework enabling the comparison of various training methodologies, pre-processing, modeling techniques and evaluation metrics would help add clarity to what techniques add true performance and generalize best. The contributions of this work is as follows:

1. A quantitative literature review highlighting the extent of the reproducibility issue in RE to motivate this study

2. An exploration into the best modeling approaches for extracting relations in scientific article abstracts, forming the modeling basis of our framework

3. An open-source unifying framework known as `REflex`[1], that is extendable to new datasets.

4. Exploration of modeling and model-complementing (training methodologies and pre-processing) techniques on 3 popular RE datasets, along with a discussion of the implications of different evaluation metrics, particularly for the medical settings. Insights from our exploration allow us to provide recommendations for future research in the RE area.

[1] code available at https://github.com/geetickachauhan/relation-extraction

3.2 Brief introduction to the relation extraction task

Relation Extraction (RE) is a popular task in Natural Language Processing (NLP) research, and the goal of RE is to find semantic relationships between entities in a document. A relation is defined as a function $t = r(e_1, e_2, ...e_n)$ where e_i are entities in a predefined relation r in a document D. More commonly, the community considers binary relations of the form *father-of(Manuel Blum, Avrim Blum)*. Relation Classification (RC) is a subset of RE that involves distinguishing between relation types as opposed to detecting whether a relation exists between entities.

This task has been commonly applied in the general as well as biomedical domains. In particular, Ravichandran and Hovy [252] employ the use of relational patterns for answering factoid questions related to topics such as *birthdate*, *location* and *definition*. Zhang et al. [331] apply a neural model to the slot-filling task (an alias for relation classification rather than extraction), which assists in populating knowledge bases. They predict varied relations such as *spouse*, *siblings* and *title*.

In the biomedical domain, Liu et al. [184] extracted protein-protein interactions using a feature-based approach with a support vector machine (SVM) classifier. The relation they try to discover is a tertiary relation between a protein, organism and a location. In a sentence like, *Exoenzyme S is an extracellular product of Pseudomonas aeruginosa*, they predict the existence of the *Protein-Organism-Location* relation between *Exoenzyme S*, *Pseudomonas aeruginosa* and *extracellular*. In the biomedical domain, relation extraction can have important applications such as assisting in drug discovery and in detection of cancerous genes [12]. In particular, drug-drug interaction extraction [266] is useful in allowing for automatic identification of drug interactions, in order to reduce the time spent by health care professionals in reviewing the medical literature. This detection is also an important research area in patient safety as the interactions can have life threatening effects.

3.3 The extent of the reproducibility crisis in relation extraction

Notes and updates since paper publishing The paper associated with this study was published in 2019, making our quantitative literature review current as of that year. Since then, there has been substantial growth in the field of natural language processing (NLP) through the introduction of transformers by Vaswani et al. [293]. The BERT [64] paper introduced a streamlined paradigm of pre-training and fine-tuning to all NLP tasks, including relation extraction. In this chapter, major methods introduced before 2019 (and right around the introduction of the BERT paper) are referenced. Since the advent of transformers, few-shot learning and the use of task agnostic models is more common [27] than our suggested approach of relation-extraction specific models.

This section consists of two types of literature reviews: a quantitative one providing evidence into the problems hindering progress in RE and a methods one, introducing the modeling and evaluation techniques commonly used in the field. The

quantitative review provides evidence of the reproducibililty issue, while the methods review summarizes the common techniques and open-source frameworks already existing in the field.

3.3.1 Quantitative Literature Review

To motivate the problems hampering progress in RE, we performed a systematic search process as of February 2019 by looking at the *cited by* list on Google Scholar (roughly ordered by number of citations) of 3 dataset papers: Hendrickx et al. [111] (`semeval`), Segura-Bedmar et al. [266] (`ddi`) and Uzuner et al. [290] (`i2b2`). These are the datasets forming the primary reproducibility study, and the scientific abstracts dataset by Buscaldi et al. [29] was primarily used for exploring the RE models that form the basis of the reproducibility study[2]. We skimmed through the first 40 papers for the `semeval` paper, 110 papers for the `ddi` paper and 578 papers for the `i2b2` paper, looking specifically for neural relation extraction papers that used neural network architectures.

Upon applying this filtering procedure, we found 22 papers for `semeval` (+ 4 papers that were not in the search list, but were cited in section 3.3.2), 15 papers for `ddi` (+ 2 papers from the section 3.3.2) and 12 papers for `i2b2`. There was an overlap of 2 papers in the `semeval` and `ddi` list, but since they were being applied to the biomedical tasks, we decided to move them to the `ddi` list. Finally, there were 24 papers for `semeval`, 17 papers for `ddi` and 12 papers for `i2b2`. For the final list of papers, please refer to Appendix B. In total, there are 53 relevant neural RE papers discussed in the following subsections, filtered from a total of 728 papers.

Reproducibility

Reproducibility is important for validating previous work and building upon it [78]. Lack of reproducibility can be attributed to many factors such as difficulty in availability of source code [122] and omission of sources of variability such as hyperparameter details [52].

Only 16 out of the 53 relevant papers had released their source code. In the `semeval` list, only 6 out of 24 total papers had source code available. This number was 6 out of 17 for `ddi` and 4 out of 12 for `i2b2`. Additionally, much of this code was lacking in modularity to be easily extendable to new datasets. In many cases, the process of reproducing the paper results was also unclear and lack of proper documentation made this more difficult.

Models were more frequently evaluated on only one dataset. However, papers in the general domain often evaluated their models on a larger number of datasets than the biomedical domain. In `semeval`, an average of 1.75 datasets were evaluated, with 8 papers being evaluated on more than 1 dataset. Out of these papers, one was evaluated on 6 datasets and the others were evaluated on 7 datasets. Only one of these papers had source code available, which was not mentioned in the paper

[2]Note that the dataset by Buscaldi et al. [29] was new and ours was one of the first studies participating in the original task challenge

and was found by additional search on Google. In `ddi`, 1.23 datasets were evaluated on average with 4 papers being evaluated on 2 datasets. For `i2b2`, 1.42 datasets were evaluated on average, and this number was driven by one paper evaluated on 5 datasets whose source code was not publicly available.

Most papers from the list mentioned some hyperparameter details. However, the list was often incomplete, and the common missing hyperparameters were *number of epochs*, *batch size* and whether a random initialization seed was set for the model or the random functions used in the code. Some papers that used the early stop mechanism were missing information about the size and criterion of the early stop evaluation data. Papers also failed to mention if a specific hyperparameter search strategy like grid search, manual search, or random search was performed [21].

Ablation Studies

Ablation studies are important in understanding the sources of variation in results as well as which parts of the model drive performance. While 20 of the 24 papers in the `semeval` list performed ablation studies, very few from the `ddi` and `i2b2` list performed them. 7 of 17 papers performed an ablation study in `ddi` and 3 of 12 papers did so for `i2b2`. In ablation studies and other reported experiments, key details related to pre-processing were missing, which we found critical in our experiments.

3.3.2 Methods Literature Review

Survey of Modeling Techniques

Given the popularity of neural relation extraction in the recent years, there is an abundance of papers that apply similar techniques to different datasets. Despite neural relation extraction existing since 2012, the biomedical domain saw a less rapid application of these models as compared to the general domain, as seen in the following subsections. And even though these papers investigated different neural network architectures for this task, no studies were published that explored the extent of improvement offered by non-modeling techniques such as pre-processing, evaluation techniques and hyperparameter tuning techniques for RE.

General domain Relation extraction over the general purpose domain has seen rapid progress in recent years with the introduction of the SemEval 2007 and 2010 tasks on relation classification between pairs of nominals, as well as 2018 task on relation extraction and classification in scientific papers [90, 111, 82].

The submissions to the 2007 and 2010 tasks involved the use of varied classification models such as Naive Bayes, k-nearest neighbor (k-NN), Maximum Entropy (MaxEnt) and SVM classifiers. Neural Network (NN) applications to NLP were only made popular in 2011 by Collobert et al.. In 2012, Socher et al. [274] applied a matrix-vector based recursive neural network (MVRNN) using a syntactic parse tree feature to improve performance for the SemEval 2010

dataset on top of existing non-neural techniques. However, Zeng et al. [322] were the first to apply a model not based on semantic features. They applied a Convolutional Neural Network (CNN) architecture with novel position-based features to the same task and achieved a better performance than MVRNN.

Since then, the field of neural relation extraction saw many advances. In 2015, Zeng et al. [323] introduced a distant supervision technique for relation classification using a multi-pooling approach over CNNs. In the same year, various other CNN based approaches were introduced [265] (`CRCNN` model used in our experiments) and [311] and so were Recurrent Neural Network (RNN) based approaches [328, 72, 312]. 2016 saw even more complex models and better performance on relation classification [205, 297, 30, 313]. Finally, additional methods explored different architectures beyond the standard RNN and CNN, by using graph convolutions over dependency trees of the sentences [333]. Another method reduced relation extraction to answering simple reading comprehension questions [160].

Biomedical Domain Advances in the biomedical domain have been inspired from techniques in the general domain, but have happened at a slower pace. There exist relation extraction challenges in this domain as well, including the drug-drug interaction extraction task known as DDI Extraction [266] (`ddi`) and the relation classification task on clinical notes [290].

For both challenges, participants submitted non-neural models, but there has been considerable work on these datasets since their respective years. Despite the many modeling techniques proposed by researchers working in the general purpose domain, most papers built on top of the idea of using CNN with position-based embeddings from [322]. Even for tasks involving relation extraction from scientific abstracts [11, 82], this modeling technique seems to be a common baseline [156, 259, 131]. The first time a neural model was applied to `ddi` was in 2016 by Liu et al. [181]. The model involved dataset specific pre-processing on top of the CNN with the position features model proposed by Zeng et al.. Around the same time and with similar performance, Zhao et al. [336] introduced a syntax CNN method, that made use of word embeddings based on the syntactic parse of the sentence on top of position embeddings as well as grammatical features.

In the same vein as the multi-pooling approach by Zeng et al. [323], Luo et al. [191] proposed a segmented CNN approach with position embeddings in 2017 by dividing the sentence into 5 parts based on the position of the entities. Similarly, He et al. [108] applied a multi-pooling architecture on top of a CNN with position embeddings and a loss function with a category-level constraint matrix. The same authors also explored a unified CNN-RNN architecture in [107]. Similar to the shortest dependency path idea by Xu et al. [312], Li et al. [162] used an RNN along the shortest dependency path along with character-based convolutions to extract relations in two common biomedical datasets. Finally, another paper discussing the improvement that character embeddings

can provide for a biomedical dataset is Nguyen and Verspoor [215]. Character embeddings is a popular idea employed previously in the relation extraction domain [156].

Prior frameworks and studies

Existing open source frameworks Literature review suggests that the field of relation extraction would benefit from an open source, extendable and transparent framework. While there do exist frameworks for RE such as Björne and Salakoski [22] and Kang et al. [147], they are based on a support vector machine (SVM). There does not exist a generalizable neural network-based framework for this field. In terms of existing products, Amazon released the Amazon Comprehend Medical API, allowing relation extraction for clinical notes, but this is more of a black-box model, which is not as beneficial to the research community.

Existing evaluation studies Even though there is a gap between the general and medical relation extraction domains at the moment, more mainstream research is now being applied to medical datasets. Another study by Mandya et al. [195] employed a combined LSTM-CNN model for cross-sentence relation extraction to `semeval` and a biomedical dataset with the aim to show state of the art performance on the biomedical domain. Additionally Zhao et al. [336] provided a detailed ablation study on the effect of performance provided by negative instance filtering, which is a pre-processing technique specific to `ddi`, as well as the modeling techniques that they choose.

Outside the relation extraction domain, impact of non-modeling techniques is being studied, with Reimers and Gurevych [253] reporting the effect of different hyperparameters in the performance for the named entity recognition (NER) task. The same authors also studied the effects of random initialization seeds on the performance of models in [255], with the conclusion that comparing score distributions of two models is much more impactful than simply comparing one evaluation score. Additionally, Crane [58] discussed similar problems for the question-answering field.

The effects of pre-processing for sentiment analysis and text categorization were tested in [31]. Addressing the replication and reproduction issue for NER and Wordnet:Similarity tasks is an older work by Fokkens et al. [78]. They spoke about the impact of different non-modeling techniques such as preprocessing, experimental setup, versioning, system output and system variation for these tasks and conclude that these categories are important to explore in order to maintain reproducibility of results. Another paper aiming to understand the text processing capabilities of CNN filters is Jacovi et al. [124].

RE would benefit from such studies to understand the true source of performance gains in results. Current studies in RE are local in nature in that they simply focus on the improvement offered by modeling techniques rather than

those provided by non-modeling techniques, such as a range of pre-processing techniques.

3.4 Relation extraction on scientific abstracts

We formed the modeling basis for our reproducibility study via a thorough exploration into recurrent versus convolutional neural network architectures for relation extraction on scientific abstracts. In this section, we will briefly explain the task and highlight key findings that led us to use the convolutional architecture for our reproducibility study.

3.4.1 The task: Semeval 2018 task 7

SemEval 2018 Task 7 [82] focuses on relation classification and extraction on a corpus of 350 scientific paper abstracts consisting of 1228 and 1248 annotated sentences for subtasks 1.1 and 1.2, respectively. There are six possible relations: *USAGE*, *RESULT*, *MODEL-FEATURE*, *PART_WHOLE*, *TOPIC*, and *COMPARE*.

Given this data, our task is to take an example sentence, as well as the left and right entities within that sentence, and an indicator as to whether the relation is reversed, and predict the relation type for that sentence. In subtasks 1.1 and 1.2, all presented sentences have a relation. The difference between the two subtasks is that subtask 1.1 deals with relation extraction on clean data where entity occurences are manually annotated whereas subtask 1.2 deals with relation extraction on noisy data with entity occurences being automatically annotated. Relations are manually annotated in both subtasks.

3.4.2 Methods

Pre-processing

Data was tokenized using the SpaCy tokenizer[3]. Part of speech (POS) tags were extracted using SpaCy, while lemmas and hypernyms were extracted via WordNet [200], inspired by Rink and Harabagiu [256].

Initial Experiments

We tested several machine learning methods on these data, including a logistic regression classifier over `tf-idf` features extracted from words, lemmas, hypernyms, and POS. Additionally, we tested deep random forests with multi-grain sequence scanning over word embeddings sequences [340] and LSTM with attention [339] over word or lemma or hypernym embeddings plus character sequence embeddings, with position indicators. Lastly, we tested a CNN model over these data, using word or lemma embeddings, position embeddings, and a variant of negative sampling. After optimizing all model configurations and doing preliminary hyperparameter optimization via

[3]https://github.com/explosion/spaCy

Model	Acc. (%)
SVM	64.0 ± 5.3
LR	65.3 ± 4.3
DEEP RF	63.1 ± 4.1
LSTM	61.4 ± 5.5
CNN	$\mathbf{66.3 \pm 4.4}$

Table 3.1: Comparison of best performance of different model types in our initial experimentation.

automatic grid search, early comparisons between the differing model classes yielded the results in Table 3.1. These results were measured in accuracy over 15-fold cross validation on the 1.1 train set. Given these initial results, we focused principally on the CNN model.

CNN model details

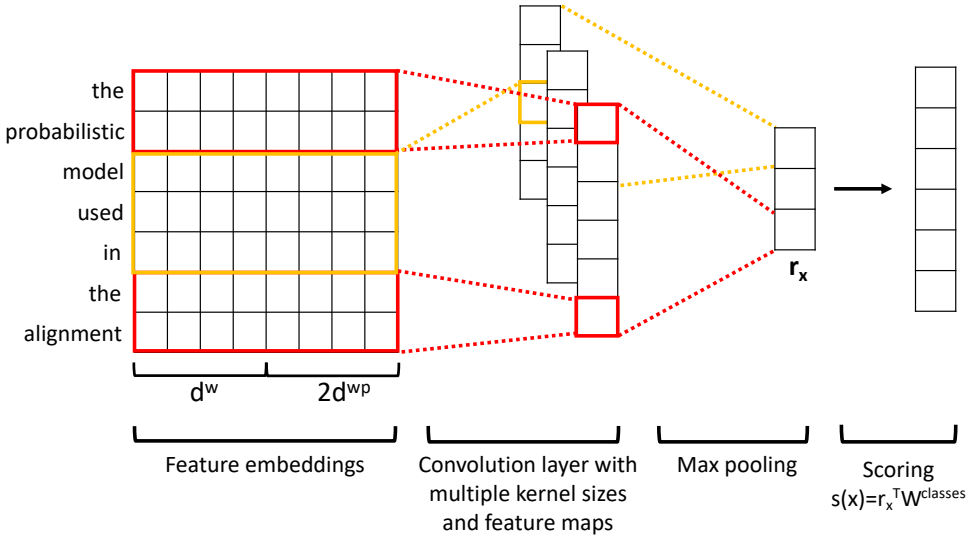


Figure 3-1: Illustration of CNN model architecture. The entities *probabilistic model* and *alignment* have the *USAGE* relation between each other, which the model is expected to predict as its objective.

Figure 3-1 presents the architecture of the CNN model. The model first takes the tokenized sentence, as well as the targeted entities, and transforms it to a sequence of continuous embedding vectors. These vectors contain word embeddings (d^w) and two set of word position embeddings (d^{wp}), corresponding to the relative distance of the other sentence words from the left entity and the right entity. Next, the model uses a convolution layer followed by max pooling to transform the embedded sentence to

Symbol	Name	Value
d^w	Word Embed. Size	50
d^{wp}	Pos. Embed. Size	42
d^c	Convolution Units	900
k	Convolution Kernel	2,3,4
m^+	Correct Label Margin	2.2
m^-	Incorrect Label Margin	0.7
γ	Penalty Scale Factor	3.1
λ	Learning Rate	0.0008
β	L2 Regularization	0.01
d	Dropout Ratio	0.5

Table 3.2: CNN model final hyperparameters.

a fixed-size representation of the whole sentence (r^x). Finally, the score is computed for each relation class via a linear transformation (s^x). The overall system is trained end-to-end via a cross entropy loss augmented with a variant of negative sampling to increase the score of the correct label while decreasing the score of the incorrect ones.

3.4.3 Results

We first optimized the CNN model hyperparameters via random search with 90 samples; final hyperparameters are shown in Table 3.2.

Beyond traditional hyperparameter optimization, a number of modifications with this model achieved performance gains during our final stages of experimentation, as determined by cross validation over either the 1.1 or 1.2 data. We detail the types of these changes below, then show the performance results obtained on the *test set* (not the cross validation results which motivated their use in our system) in Table 3.3.

Merged Training Sets Merging the 1.1 & 1.2 training datasets as a new training set had a large impact on the macro F1 score of our models. Both training datasets are relatively small, containing only approximately 1200 examples. Merging the 1.1 and 1.2 training sets helps equalize class imbalance and expand the dataset size, at the cost of introducing a biased distribution of relation types for either class alone.

Reversal Indicator Features Each entity pair was given the information whether the relation of it is reversed or not. We added this binary feature, which proved useful.

Custom ACL Embeddings Specializing our word vector embeddings pre-training source to an ACL-specific corpus [245] offered notable gains.

Context words We explored using a context window of varying sizes around the entity-enclosed text within the sentence. Our pre-submission cross validation

Condition	1.1 (%)	1.2 (%)
1.1 Train Set	49.0 ± 1.2	N/A
1.2 Train Set	N/A	66.5 ± 3.2
Merged Train Sets	68.5 ± 3.8	74.4 ± 3.2
Reversed Feature	69.0 ± 1.2	78.0 ± 3.6
ACL Embeddings	71.7 ± 0.7	80.5 ± 1.5
Context Words	71.3 ± 1.0	82.5 ± 1.6
Ensemble	**72.7**	**85.0**

Table 3.3: CNN Improvements over a series of modifications. Each row includes the modifications of the previous rows. All numbers are macro-F1 scores on test set after 10 runs in the form of {average}±{standard deviation} (the "Ensemble" row lacks deviation numbers as it, being a variance reduction technique, does not have the same sources of variation as the other models). We report ±20 context words here, which was found to be optimal in post-submission experimentation, but our submitted models used ±50 context words, which was preferred under initial cross validation.

experiments suggested a context window of ±50 words was optimal, but post-submission evaluation on the provided test set yielded better results with a ±20 word window. Empirically, the number of context words to be included needs to be optimized on the specific dataset.

Ensembling We trained 50 copies of our network, using different random initializations and dev sets (for early stopping), then averaged their scores for prediction. This reduced variance of our predictions and improved performance.

3.4.4 Summary

The CNN model, along with the hyperparameters found in this subsection, formed the modeling basis for our `REflex` framework. While the ablation studies presented here are specific to the Semeval 2018 task 7 [82], many of the hyperparameter-related findings are generalizable to our framework.

3.5 REflex: A flexible framework for relation extraction in multiple domains

To support our reproducibility research in relation extraction, we introduce a unifying framework seen in figure 3-2 that allows us to perform systematic explorations into modeling, pre-processing and training methodologies for filling in the knowledge gaps around unreported ablation studies and understand the methods' true sources of performance improvements. We also intend for our framework to be an easy-to-use reference for beginner researchers in the field.

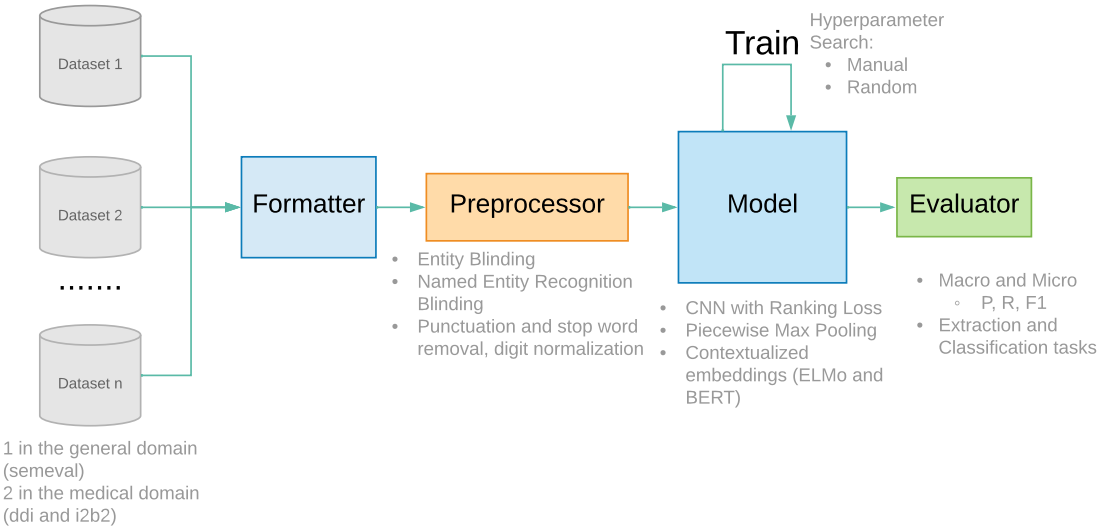


Figure 3-2: Systematic exploration framework. Each dataset results computed separately.

The framework breaks up various parts of processing into different stages, allowing for modular addition of components in the future. First, a formatter converts the raw dataset into a common input format accepted by the preprocessor, and the preprocessed dataset is then fed to the model. The model then performs the training after which evaluator performs evaluation on the test set (or development set for cross validation). With this framework, we explore various pre-processing, modeling and model-complementing strategies on 3 datasets we refer to in the rest of this chapter as `semeval`, `ddi` and `i2b2`. Our exploration, tasks and datasets are described below.

3.5.1 Datasets and Tasks

semeval Semeval 2010 task 8 [111] consists of 8000 training sentences as well as 2,717 test sentences for the multi-way classification of semantic relations between pairs of nominals. There are a total of 19 relations (where 18 relations consist of taking directionality into account), with an *Other* class which is considered noisy, with annotators classifying this class if no fit was found in the other classes. The official evaluation reported **macro-F1** scores and did not count the *Other* class in calculations. Inter-annotator agreement for this dataset is between 60% and 95%.

ddi Semeval 2013 task 9.2 [266], commonly known in the literature as drug-drug interaction (DDI) extraction, consists of 1,017 texts with 18,491 pharmacological substances and 5,021 drug-drug interactions from PubMed articles in the pharmacological literature. A total of 5 relations are present with a *None* class indicating no interaction between the drug pairs. The official evaluation reported **macro-F1** scores for classification, along with a detection macro-F1. While classification was a multi-class classification task, detection converted the problem into a binary clas-

sification between non-*None* classes and *None* classes. The challenge task dataset was developed from two separate annotated sources, the first with an inter-annotator agreement greater than 80%, and the second with agreement between 55% and 72%.

i2b2 i2b2/VA 2010 relations [290] consists of discharge summaries from Partners Healthcare and the MIMIC II Database [261]. They released 394 training reports, 477 test reports and 877 unannotated reports for this purpose. After the challenge, only a part of the data was publicly released for research and the dataset consists of 8 non-*None* relations in three categories: Medical *Problem* - *Problem*, *Problem* - *Test* and *Problem* - *Treatment* relations. There were also *None* relations present in each of the three categories. The official evaluation reported **micro-F1** scores and did not count the *None* class in calculations.

3.5.2 Pre-processing methods

Various pre-processing methods are tested after performing simple tokenization and lower-casing of the words: entity blinding used by Liu et al. [181], commonly applied **stop-word and punctuation removal**, digit normalization applied for `ddi` in [336], and named entity recognition related replacement (this is known as NER blinding in this work). We used the spaCy framework[4] to perform tokenization as well as to identify punctuations and digits.

Stop word removal is a common technique in Natural Language Processing (NLP) to remove commonly used words such as *the* and *is* in order to simplify the sentence. The technique was first coined in Luhn [188], and was commonly used in Information Retrieval (IR) to make the processing of natural language queries faster and more accurate.

Digit normalization refers to the replacement of all decimals and integers in the sentence by the word *number*. Instead of using regular expressions to search for decimals and digits, we used spaCy's `like_num` argument which identifies decimals and digits as well as language specific words like *ten* or *hundred*.

Entity blinding and **NER blinding** are similar concept blinding techniques where the first is performed based on gold standard annotations, while the second is performed by running NER on the original sentence. We replace the words in the sentence matching the entity or named entity span with the target label and use those for training and testing.

Entity labels for `semeval` were not annotated with type information, whereas `ddi` identified drugs and `i2b2` identified medical problems, tests and treatments. Therefore, entity labels for `semeval` were *ENTITY*, for `ddi` were *DRUG* and for `i2b2` were *PROBLEM*, *TREATMENT* and *TEST*. In this work, we use *fine-grained concept type* to refer to the presence of more than one concept type, as in the the case of `i2b2`.

NER labels for `semeval` consisted of those provided by the large english model by spaCy and provided standard types such as *PERSON* and *ORGANIZATION*,

[4]https://github.com/explosion/spaCy

whereas those for the medical datasets was provided by the scispacy medium size model[5] and did not provide types. In this case, blinding consisted of replacing the words in the sentence by *Entity*.

As an example of blinding, consider the sentence in figure 3-3 with its entity blinded version fed to the model.

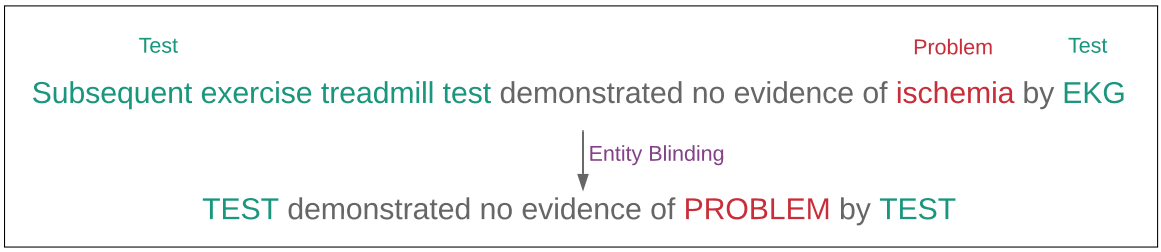


Figure 3-3: Result of entity blinding for a sentence in the `i2b2` dataset

3.5.3 Model

We employ the baseline model described in section 3.4.2, which is based upon [322] and [265]. Our model is a convolutional neural network (CNN) with position embeddings and a **ranking loss** with negative sampling (referred to as `CRCNN` in this work. The model is initialized with pre-trained word embeddings based on the domain it is applied to: for the general domain dataset, the model is initialized with senna embeddings by Collobert et al. [53], whereas for the medical domain (biomedical and clinical) the model is initialized with the `PubMed-PMC-wikipedia` embeddings released by Pyssalo et al. [238]. Many perturbations on top of `CRCNN` model are tested, such as **piecewise max-pooling**, as suggested by Zeng et al. [323] and **ELMo embeddings** Peters et al. [229]. To compare different featurizations of contextualized embeddings, we also employ the embeddings generated by the **BERT** model (rather than using them in the standard fine-tuning approach).

The fine-tuning approach, which tends to be computationally expensive, has been thoroughly explored for multiple tasks, including medical relation extraction by Lee et al. [157], but the approach of using contextualized embeddings had not been explored in the literature as much at the time of publication. We chose to explore different ways of incorporating the BERT contextualized embeddings for researchers that wanted to utilize a less computationally intensive technique, while still aiming for performance gains for their task.

Because ELMo provides token level embeddings, they were concatenated with the word and position embeddings from `CRCNN` before the convolution phase. According to the terminology used in section 3.4.2, new feature embeddings were generated by concatenating the word embeddings, word position embeddings as well as the ELMo embeddings on a word-by-word basis.

BERT, in contrast, provides word-piece level as well as sentence level embeddings. The word-piece level embeddings were concatenated similar to ELMo (known

[5]https://allenai.github.io/scispacy/

as BERT-tokens) after the individual word pieces were averaged to form one word embedding. For example, if BERT split the word "playing" to generate embeddings for "play" and "##ing," we averaged the embeddings for the two word pieces to form one word embedding for "playing." The sentence level embeddings were concatenated with the fixed size sentence representation, known as r_x in section 3.4.2, which is output after convolution of word and position embeddings (known as BERT-CLS).

3.5.4 Training Methodologies

Two types of hyperparameter tuning were explored: **manual tuning** and **random search** [21].

Hyperparameter	**Values**
epoch	{50,100,150,200}
lr decay	[1e-3, 1e-4, 1e-5]
sgd momentum	{T, F}
early stop	{T, F}
pos embed	{10, 50, 80, 100}
filter dimension	{50, 150}
filter size	2-3-4, 3-4-5
batch size	{70, 30}

Table 3.4: Hyperparameters explored for the first pass of manual search. lr decay means learning rate decay at [60, 120] epochs, pos embed refers to the position embedding size.

Hyperparameter	**Distributions**
epoch	uniform(70, 300)
lr	{constant, decay}
lr init	uniform(1e-5, 0.001)
filter size	2-3, 2-3-4, 2-3-4-5 3-4-5, 3-4-5-6
early stop	{T, F}
batch size	uniform(30, 70)

Table 3.5: Hyperparameter distributions for random search. Those written in {} are picked with equal probabilities. The learning rate (lr) was uniformly initialized, and decayed from 0.001 to the lr init value (used as a post decay value in this scenario) at half of the number of epochs. If early stop was true, patience was set to a fifth of the number of epochs. We ran 100-120 experiments for each dataset to search for optimal hyperparameters.

Evaluating on 3 datasets meant that we needed to identify a default list of hyperparameters by tuning on one of the datasets before identification of the hyperparameter list for the other two. We chose `semeval` for initial tuning due to its larger literature and because the `CRCNN` model was originally evaluated on this dataset. We started

with reference hyperparameters listed in Zeng et al. [322] and Santos et al. [265] and identified default hyperparameters after tuning on a dev set randomly sampled from the training data of the `semeval` dataset. These default hyperparameters[6] were used as starting points for manual tuning on the medical datasets as well as random search for all datasets.

We perform manual tuning on a subset of the hyperparameters, mentioned in table 3.4. In order to avoid overfitting in cross validation pointed out by Cawley and Talbot [37], we perform a nested cross validation procedure, keeping a dev fold for hyperparameter tuning and a held out fold for score reporting.

On these dev folds, we perform paired t-tests for each of the perturbations to the parameters listed in table 3.4. The first pass involves changing one hyperparameter per experiment and noting the ones that cause a statistically significant improvement, which helps in identification of a narrower list of hyperparameters to tune on. We further refine the hyperparameter values in our second pass by testing on values similar to those that were leading to statistically significant improvements in the first pass. For example, if we noticed that lower epoch values were helpful in the first pass, we tested them in combination with the other optimal hyperparameter values (from first pass) in the second pass.

For each of the datasets, we tuned based on their official challenge evaluation metrics listed in section 3.5.1. `ddi` and `i2b2` involved 5-fold nested cross validation, whereas `semeval` involved 10-fold cross validation.

Random search was performed based on the official evaluation metrics for each dataset, on a fixed dev set randomly sampled from the training data. Distributions used for the search are listed in table 3.5.

3.5.5 Reported metrics

We report results on the official metrics for each of the challenge tasks, as described in section 3.5.1. The official challenge problems for all datasets compared models based on multi-class classification, but for the medical datasets, we were also interested in the changes in model performance if the task was treated as a binary classification problem. This was based on the rationale that in the drug literature, for example, pharmacologists would not want to sacrifice the ability to identify a potentially life threatening drug interaction pair, even if the type of the drug pair is not known. Therefore, we report results for the multi-class as well as the binary classification scenario. For clarity, let us refer to them in the rest of the chapter as *classification* and *detection* respectively.

Detection results were obtained using our evaluation scripts by treating existing relations as one class, ignoring the types outputted by the model. The other class in this task was the *None* or *Other* class, representing non-existing relations. Note that we did not re-train the model for this task.

[6] listed in source code

3.5.6 Result 1: Pre-processing causes large variations in performance, and often goes unreported in the literature

Often, papers fail to mention the importance of pre-processing in performance improvements. Experiments in table 3.6 reveal that they can cause larger variations in performance than modeling.

We applied pre-processing changes with the `CRCNN` model with default hyperparameters for `semeval` and manual hyperparameters for the medical datasets. All comparisons are performed against the original pre-processing technique, which involved using the original dataset sentences in training and test. Further details of this analysis are located in the appendices of my S.M. thesis [7].

Preprocess \ Dataset	`semeval`	`ddi`		`i2b2`	
		Class	Detect	Class	Detect
Original	**81.55**	65.53	81.74	59.75	83.17
	80.85 (1.31)	82.23 (0.32)	88.40 (0.48)	70.10 (0.85)	86.45 (0.58)
Entity Blinding	72.73	**67.02**	**82.37**	**68.76**	**84.37**
	71.31 (1.14)	83.56 (2.05)•	89.45 (1.05)•	76.59 (1.07)	88.41 (0.37)
Punct and Digit	81.23	63.41	80.49	58.85	81.96
	80.95 (1.21)•	80.44 (1.77)	87.52 (0.98)	69.37 (1.43)•	85.82 (0.43)
Punct, Digit and Stop	72.92	55.87	76.57	56.19	80.47
	71.61 (1.25)	78.52 (1.99)	85.65 (1.21)	68.14 (2.05)•	84.84 (0.77)
NER Blinding	81.63	57.22	79.03	50.41	81.61
	80.85 (1.07)•	78.06 (1.45)	86.79 (0.65)	66.26 (2.44)	86.72 (0.57)•

Table 3.6: Preprocessing techniques with `CRCNN` model. Row labels Original = simple tokenization and lower casing of words, Punct = punctuation removal, Digit = digit removal and Stop = stop word removal. Test set results at the top with cross validated results (average with standard deviation) below. All cross validated results are statistically significant compared to Original preprocessing ($p < 0.05$) using a paired t-test except those marked with a •

Punctuation and Digits are important in the biomedical domain

Removal of punctuation and digits (`punct`) hurts *classification* and *detection* performance for the `ddi` dataset, which is a biomedical dataset. On the other hand, performance on `i2b2` is worse only for the detection task. Statistical significance is not found for the other tasks and datasets.

This indicates that punctuation and digits are more important for the `ddi` dataset and that they are important only for the detection of relations for `i2b2`. To further investigate which of punctuation and digit normalization was the larger contributor in worse performance, we looked at examples where misclassifications were occuring.

[7] `https://dspace.mit.edu/handle/1721.1/122694`

Stop words are important in relation extraction settings

Removal of punctuation, digits and stop words (`stop`) is hurting performance more than `punct` (statistically significant for `ddi` and `semeval` with $p < 0.005$). This effect is less drastic for `i2b2`: `stop` is not statistically significantly worse than `punct` for *classification* task, but is significantly worse with $p = 0.015$ for the *detection* task. This indicates that stop words are important for relation extraction.

Fine-grained concept types could be helpful in general because of their ability to simplify the sentence

The availability of fine-grained concept types is likely to boost performance in relation extraction settings. The `i2b2` dataset provided fine-grained concept types in the form of medical problem, test and treatments. Entity blinding causes almost 9% improvement in *classification* performance and 1% improvement in *detection* performance. In contrast, `ddi` only provided gold standard annotations for drug types in the sentence, and while this does not cause statistically significant improvements for cross validation, it does improve test set classification performance by about 1.5% and detection performance by 1%. For these medical datasets, NER blinding consisted of replacing the detected named entities by *Entity* because named entity types were not available (more details in section 3.5.2). Due to the coarse-grained nature of the entities, it hurts *classification* performance significantly, and *detection* performance a little.

Entity blinding hurts performance for `semeval`, possibly due to the coarse grain nature of the replacement and the entity bias [332]. The replacement loses associations between the entity mentions and relation types, which reduces performance. While a finer-grain replacement in this setting (NER blinding) does not cause a statistically significant change in performance, it has been shown to be a helpful feature by [274]. To recall, entity blinding involved replacement of entity words by *Entity*, while NER blinding involved replacing named entities in the sentence with labels such as *ORGANIZATION* and *PERSON* (more details in section 3.5.2).

Reasonable performance is maintained on the *Detection* task

For the medical datasets, while *classification* performance varies highly with different pre-processing techniques, *detection* is relatively unaffected. In a setting where one cares more about detection of relationships rather than multi-class classification, one would be able to get away with using non-complicated pre-processing techniques to maintain reasonable performance.

3.5.7 Result 2: Reporting on one test set score is problematic due to split bias

All 3 datasets evaluate models based on one score on the test set, which is common practice for NLP challenges. Reporting one score as opposed to a distribution of scores has been shown to be problematic by Reimers and Gurevych [254] for sequence

tagging. Crane [58] discuss similar problems for question-answering. Our experiments show that even if you keep the same random initialization seed (all our experiments have a fixed random initialization seed), split bias can be another source of variation in scores.

Significance testing of some cross validated results reveals no significance even when the test set result improves in performance. This is particularly concerning for `ddi` where entity blinding (called drug blinding in the literature) is used as a standard pre-processing technique without ablation studies demonstrating its effectiveness. Results suggest the contrary: entity blinding seems to help test set performance for `ddi` in table 3.6, but shows no statistical significance. Table 3.10 further demonstrates that using this in conjunction with other techniques results in test score variations despite being statistically insignificant.

No statistical significance is seen even when the test set result worsens in performance for BERT-CLS in table 3.7 where it hurts test set performance on `ddi` but is not statistically significant when cross validation is performed.

3.5.8 Result 3: Debunking the effects of common modeling techniques

In table 3.7, we tested the generalizability of the commonly used piecewise pooling technique proposed in [323], a variant of which was applied in the model by Luo et al. for `i2b2`. We also tested the improvements offered by different featurizations of contextualized embeddings, which had not been explored much for relation extraction at the time of publication.

Modeling \ Dataset	`semeval`	`ddi`		`i2b2`	
		Class	Detect	Class	Detect
CRCNN	81.55 80.85 (1.31)	65.53 82.23 (0.32)	81.74 88.40 (0.48)	59.75 70.10 (0.85)	83.17 86.45 (0.58)
Piecewise pool	81.59 80.55 (0.99)•	63.01 81.99 (0.38)•	80.62 88.47 (0.48)•	60.85 73.79 (0.97)	83.69 89.29 (0.61)
BERT-tokens	85.67 85.63 (0.83)	**71.97** 85.35 (0.53)	**86.53** 90.70 (0.46)	63.11 72.06 (1.36)	**84.91** 87.57 (0.75)
BERT-CLS	82.42 80.83 (1.18)•	61.3 82.71 (0.68)•	79.63 88.35 (0.77)•	56.79 67.37 (1.08)	81.91 85.43 (0.36)
ELMo	**85.89** 84.79 (1.08)	66.63 84.53 (0.96)	83.05 90.11 (0.56)	**63.18** 72.53 (0.80)	84.54 87.81 (0.34)

Table 3.7: Modeling techniques with original preprocessing. Test set results at the top with cross validated results (average with standard deviation) below. All cross validated results are statistically significant compared to `CRCNN` model ($p < 0.05$) using a paired t-test except those marked with a •. In terms of statistical significance, comparing contextualized embeddings with each other reveals that BERT-tokens is equivalent to ELMo for `i2b2`, but for `semeval` BERT-tokens is better than ELMo and for `ddi` BERT-tokens is better than ELMo only for detection.

Modeling changes were applied with the original pre-processing technique for the `CRCNN` model with default hyperparameters for `semeval` and manual hyperparameters for the medical datasets. All comparisons are performed with the baseline performance of the `CRCNN` model.

Piecewise pooling is not a generalizable technique

While piecewise pooling helps `i2b2` by 1%, it hurts test set performance on `ddi` and doesn't affect performance on `semeval`. It may be intuitive to split pooling by entity location, but this technique is not experimentally found to be generalizable to other datasets.

Contextualized embeddings should be featurized correctly in CNN models

Contextualized embeddings generally boost performance, but they should be concatenated with the word embeddings before the convolution stage. ELMo and BERT-tokens boosted performance significantly for all datasets, but BERT-CLS hurt performance for the medical datasets. While BERT-CLS boosted test set performance for `semeval`, this was not found to be a statistically significant difference for cross validation. Note that ELMo was featurized similarly to BERT-tokens and featurization details are present in section 3.5.3.

This indicates that the technique of featurizing the contextualized embeddings matters for a CNN architecture. Concatenating the contextualized embeddings with the word embeddings keeps a tighter coupling, which is helpful for relation extraction where the word level associations are essential in predicting the relation type.

3.5.9 Result 4: Exploring different hyperparameter tuning methods

Bergstra and Bengio [21] show the superiority of random search over grid search in terms of faster convergence, but leave to future work automating the procedure of manual tuning, i.e., sequential optimization. Bayesian optimization strategies [273] could help with this but often require expert knowledge for correct application. We tested how manual tuning, requiring less expert knowledge than Bayesian optimization, would compare to the random search strategy in table 3.8.

Manual search outperformed random search

Tables in appendix C demonstrate that random search reduces the variability of results and converges to better performance than the default hyperparameters. Additionally, manual search outperformed random search for both `i2b2` and `ddi` corpus. Both methods present different challenges for barrier of entry.

Manual search is often criticized for the high barrier of entry [21]. Knowledge about which hyperparameters are more important in specific contexts can make this search faster and provide improved results. Our proposed two-pass method helps in

Hyperparam Tuning \ Dataset	semeval	ddi		i2b2	
		Class	Detect	Class	Detect
Default	81.55	62.55	80.29	55.15	81.98
	80.85 (1.31)	81.62 (1.35)	87.76 (1.03)	67.28 (1.83)	86.57 (0.58)
Manual Search	-	**65.53**	**81.74**	**59.75**	**83.17**
		82.23 (0.32)•	88.40 (0.48)•	70.10 (0.85)	86.45 (0.58)•
Random Search	**82.2**	62.29	79.04	55.0	80.77
	81.10 (1.26)•	75.43 (1.48)	83.54 (0.60)	60.66 (1.43)	82.73 (0.49)

Table 3.8: Hyperparameter tuning methods with original preprocessing and fixed CRCNN model. Test set results at the top with cross validated results (average with standard deviation) below. All cross validated results are statistically significant compared to Default with $p < 0.05$ except those marked with a •. Note that hyperparameter tuning can involve much higher performance variation depending on the distribution of the data. Therefore, even though there is no statistical significance in the manual search case for the held out fold in the ddi dataset, there was statistical significance for the dev fold which drove those set of hyperparameters. For both ddi and i2b2 datasets, manual search is better than random search with $p < 0.05$.

developing intuition on the important hyperparameters by changing each hyperparameter in isolation to test the statistical significance of the performance difference. By further changing the narrow list of hyperparameters found from the first pass, convergence to better results is found in the second pass.

Random search, on the other hand, can be complicated because one needs to pick the right distributions for the hyperparameters and the right search space. A larger search space and sub-optimal distributions run into the possibility of running too many experiments in a hyperparameter space leading to lower performance. Ideally, random search should run enough experiments in the vicinity of the global maxima to converge to it faster. Additional findings related to result distributions for random search is present in appendix C.

3.5.10 Result 5: Comparison to state of the art methods

In order to compare results with state-of-the-art results for each dataset at the time of publication, we ran additional experiments to test combinations of techniques from the previous results sections that showed the most improvements. These are listed in tables 3.9 and 3.10.

The best *classification* test set results found are listed in table 3.11. Note that we do not compare the *extraction* task for datasets other than ddi because the official challenges only compared classification results. Even though the official challenge did not rank models based on the *detection* task, several papers in the ddi literature mention these results.

We report results in table 3.11 to perform a comparison to state-of-the-art approaches consistent with the current method, and show why this leads to unfair comparisons. This is not only because of the problem of split bias highlighted in section

Technique \ Task	Classification	Detection
E + ent	70.46 77.70(1.26)	86.17 89.36 (0.50)
B + ent	70.56 76.72 (1.04)	85.66 88.63 (0.33)
E + piece + ent	70.62 79.41 (0.53)	86.14 90.37 (0.44)
B + piece + ent	**71.01** 79.51 (1.09)	**86.26** 90.34 (0.53)
piece + ent	69.73 78.12 (1.10)	85.44 89.74 (0.44)
E + piece	63.19 74.76 (0.68)	84.92 89.90 (0.37)
B + piece	63.23 74.67 (0.89)	85.45 89.61 (0.68)

Table 3.9: Additional experiments for `i2b2`. E = ELMo, B = BERT-tokens, ent = entity blinding, piece = piecewise pooling. All results are statistically significant compared to BERT-tokens and ELMo models respectively from table 3.7 and piece + ent row is statistically significant compared to piecewise pool model as well as entity blinding model. These are all statistically significantly better than the `CRCNN` model from table 3.7. All $p < 0.05$.

3.5.7, but also because different models are using different pre-processing techniques, which are critical sources of variation in results. The issue is more pronounced for the medical datasets, where omission of ablation studies is common as seen in section 3.3.1.

Wang et al. [297] report a result of 88% on `semeval` and do not provide any public source code for replication purposes. Despite being below the state of the art range, `REflex` provides the best performing publicly available model for this dataset.

Zheng et al. [337] report the best result on `ddi` (77.3%) but perform negative instance filtering, which is a highly specific pre-processing technique that does not fit with the flexible nature of `REflex`. This technique also makes the data smaller, but the paper is unclear about whether they apply this technique to shorten the test set as well. Unfortunately, the source code is not publicly available to answer these questions. Additionally, cutting out sentences from the training as well as test data would make the prediction task a lot easier and impractical to use in real-world settings due to its highly specific nature.

Zhao et al. [336] already show that negative instance filtering causes a 4.1% improvement in test set performance. If our model were to use this pre-processing technique, it would reach the state-of-the-art range in the *classification* task. On the other hand, the *detection* results **outperform** this model by 2.53%.

Sahu et al. [262] (code unavailable) report a state of the art result of 71.16% on

Technique \ Task	Classification	Detection
E + ent	68.69 86.25 (1.54)	83.72 91.35 (0.90)
B + ent	**70.66** 85.79 (1.54)	**85.35** 91.26 (0.63)

Table 3.10: Additional experiments for `ddi`. E = ELMo, B = BERT-tokens, ent = entity blinding. Results are not statistically significant compared to BERT-tokens and ELMo models respectively from table 3.7 and not from each other either.

Dataset	Result	Technique
`semeval`	85.89	ELMo
`ddi`	71.97, 86.53	BERT-tokens
`i2b2`	71.01	BERT-tokens + piece + ent

Table 3.11: Best test set *classification* results for all datasets, except `ddi` where *detection* results are mentioned after the classification results. piece = Piecewise pooling, ent = entity blinding. Result corresponds to F1 scores, macro for `semeval` and `ddi`, but micro for `i2b2`.

`i2b2`, which the results in table 3.11 are able to match. Note that [257] report a result of 73.7% with a support vector machine, but they used a larger version of the dataset. After the official challenge, only a subset of the data was publicly available, so comparing against this number would not be fair.

Comparison against these numbers demonstrates that `REflex` is the only open-source framework, capable of achieving performance in the state of the art ranges for all 3 datasets we evaluate on. Therefore, `REflex` can be used as a strong baseline model in future relation extraction studies.

3.6 Conclusion

Towards the goal of advancing the state-of-the art in Machine Learning (ML) for health, reproducibility is a key component in allowing for compounding and transparent scientific improvements. Chapter 2 touched on advancing reproducibility for ML models developed on electronic health records by introducing the `MIMIC-Extract` framework. In this chapter, we dive deeper into the natural language processing (NLP) subfield of relation extraction, commonly applied in the medical domain towards automated knowledge base generation and natural language understanding.

We perform a detailed quantitative literature review to showcase the extent of the reproducibility crisis in relation extraction. We find a lack of consensus on generalizable techniques in the field, making it difficult to perform systematic comparisons of methods and determining the true sources of performance gains to use for future works in the field. To support reproducibility for future research in relation extrac-

tion, we introduce `REflex`, a unifying framework that we apply on three highly used datasets from the general, biomedical and clinical domains with the ability to be extendable to new datasets. Our key findings reveal that: 1) Pre-processing can have a strong effect on performance, sometimes more than modeling techniques, as is the case for the `i2b2` dataset. The use of external knowledge in the form of concept types is highly beneficial, perhaps revealing semantic information that is helpful for better predictions. 2) Reporting on one test set score, as is commonly done in the literature, can be problematic due to split bias, and a cross validation approach with significance tests may help ease this issue. 3) Contextualized embeddings are generally helpful, but selecting the right featurizing technique is important depending on the model used. For convolutional neural network (CNN) models, concatenating them with the word embeddings prior to convolution is most beneficial. 4) Selecting the right hyperparameters for a dataset is highly impactful to performance. We suggest an initial manual hyperparameter search based on cross validation significance tests for those who are pressed for time. Random search is a reasonable automated option, but requires more experience for picking the right search space and the right distributions for the hyperparameters.

Through our study, the most surprising and unreported ablation study showcased the drastic effects of pre-processing in the biomedical and medical domains, with the entity blinding technique based on external knowledge to be causing close to 9% improvement in the micro-F1 score for the `i2b2` dataset, and close to 2% for the macro-F1 score with the `ddi` dataset. In task challenges, these numbers can be the distinction between a winning system versus other contributed systems. In the clinical setting, the higher accuracy of relation extraction can have great impacts on efforts in automated knowledge base generation, search or natural language understanding.

Though the results of our study are based on task-specific models around the advent of the Bidirectional Encoder Representations from Transformers (BERT) [64] paper, our recommendations associated with the importance of good pre-processing and generalizing evaluation metrics reporting beyond limited data splits remain relevant and useful. Particularly, as the field moves towards task-agnostic and higher capacity models, the importance of data quality and pre-processing will remain paramount.

Part II

Handling low-label and multi-modal scenarios

Chapter 4

Multimodal representation learning for disease severity prediction

Abstract

In this thesis, we focus on methods that leverage clinical structure to improve representation learning capabilities in machine learning for health. This chapter highlights a work [41] focused on developing methods that take advantage of the rich semantic information in radiology reports to support medical vision model capabilities.

Tackling the disease severity classification task requires deep semantic understanding of the underlying data distribution. Our work focuses on pulmonary edema severity quantification, a crucial step towards effective management of patient fluid status for acute congestive heart failure (CHF) patients. While large publicly available datasets of chest radiographs and free-text radiology reports exist, only limited labels are present due to 1) expensive and time consuming annotation efforts and 2) bronze-standard labels based on keyword matching from radiology reports. This forms a significant challenge in learning accurate models for image classification.

We propose and demonstrate a novel machine learning algorithm at the time of publication, that assesses pulmonary edema severity from chest radiographs, while taking advantage of the rich information present in the radiology reports. We develop a neural network model that is pre-trained on both images and free-text to infer pulmonary edema severity directly from chest radiographs, using contrastive learning. Our experimental results suggest that the joint image-text representation learning improves the performance of pulmonary edema assessment by an average of 10% AUC, when comparing with a supervised model trained on images only. We also show the benefits of using text to explain the image classification by the joint model. At the time of publication, our approach was the first to leverage free-text radiology reports for improving the image model performance in this application. Our code[1] and disease severity labels [170] are available for public use.

[1] `https://github.com/RayRuizhiLiao/joint_chestxray`

4.1 Introduction

Medicine is inherently multimodal. Clinicians provide thorough care by inspecting data from a variety of modalities such as clinical notes, laboratory tests, vital signs and observations, medical images, and more. Similar to choices made in clinical practice [54, 221, 66], models that are able to leverage the multi-modal structure of clinical data can take advantage of multiple views of the same patient for comprehensively assessing their condition.

This chapter focuses on the assessment of pulmonary edema disease severity. Pulmonary edema is the most common reason patients with acute congestive heart failure (CHF) seek care in hospitals [89, 121, 2]. The treatment success in acute CHF cases depends crucially on effective management of patient fluid status, which in turn requires pulmonary edema quantification, rather than detection of its mere absence or presence. This quantification takes the form of a numeric assessment of the degree of pulmonary edema, ranging from 0 (absent) to 3 (severe), as seen in Section 4.4.1.

Chest radiographs are commonly acquired to assess pulmonary edema in routine clinical practice. Radiology reports capture radiologists' impressions of the edema severity in the form of unstructured text. Reports often contain mentions of radiographic findings which could be associated with confounding disease processes, and there also remains a lack of standardized reporting of edema quantification. While the chest radiographs in theory possess ground-truth information about the disease, manually labeling them is often a complex and time intensive (and therefore expensive) task. Therefore, labels extracted from reports are used as a proxy for ground-truth image labels. Only limited numerical edema severity labels can be extracted from the reports, which limits the amount of labeled image data we can learn from. These challenges presents a significant barrier to learning accurate image-based models for edema assessment.

4.2 Our Contributions

To improve the performance of the image-based model and allow leveraging larger amounts of training data, we make use of free-text reports to include rich information about radiographic findings and capture radiologists' reasoning of pathology assessment. We incorporate free-text information associated with the images by including them during our training process. Our contributions are as follows:

1. We propose a neural network model that jointly learns from images and free-text to quantify pulmonary edema severity from images (chest radiographs)

2. We are the first to apply a novel contrastive learning objective towards the challenge of disease severity prediction via multi-modal representation learning using chest radiographs and radiology reports

3. Compared to prior work in the image-text domain that fuses image and text features [19], our approach allows decoupling the two modalities during inference to construct an accurate image-based model.

4. Our approach is successfully able to improve upon the image-only baseline by an AUROC of approximately 10%, bringing our machine learning (ML) for healthcare approach much closer to what is expected in hospital clinical efficacy.

At training time, the model learns from a large number of chest radiographs and their associated radiology reports, with a limited number of numerical edema severity labels. At inference time, the model computes edema severity given the input image. While the model can also make predictions from reports, our main interest is to leverage free-text information during training to improve the accuracy of image-based inference.

4.3 Prior Work

Prior work in assessing pulmonary edema severity from chest radiographs has focused on using image data only [171]. To the best of our knowledge and at the time of publication, ours is the first method to leverage the free-text radiology reports for improving the image model performance in this application. Our experimental results demonstrate that the joint representation learning framework improves the accuracy of edema severity estimates over a purely image-based model on a fully labeled subset of the data (supervised). The joint learning framework uses a ranking-based criterion [104, 44], allowing for training the model on a larger dataset of unlabeled images and reports.

The ability of neural networks to learn effective feature representations from images and text has catalyzed the recent surge of interest in joint image-text modeling. In supervised learning, tasks such as image captioning have leveraged a recurrent visual attention mechanism using recurrent neural networks (RNNs) to improve captioning performance [310]. TieNet used this attention-based text embedding framework for pathology detection from chest radiographs [302], which was further improved by introducing a global topic vector and transfer learning [314]. A similar image-text embedding setup has been employed for chest radiograph (image) annotations [207]. In unsupervised learning, training a joint global embedding space for visual object discovery has recently been shown to capture relevant structure [105]. All of these models used RNNs for encoding text features.

More recently, transformers such as the BERT model [64] have shown the ability to capture richer contextualized word representations using self-attention and have advanced the state-of-the-art in nearly every language processing task compared to variants of RNNs. Our setup, while similar to [302] and [105], uses a series of residual blocks [110] to encode the image representation and uses the BERT model to encode the text representation. We use the radiology reports during training only, to improve the image-based model's performance. This is in contrast to visual question answering [10, 186, 8], where inference is performed on an image-text pair, and image/video captioning [310, 233, 291, 134], where the model generates text from the input image.

Current State of the Field Like many sub-fields of machine learning, there was an explosion of follow up work in medical vision-language modeling (MVLM). Major

follow up works to our paper include ConVIRT [334], which formed the basis for the popular CLIP [247] architecture. The architectures were highly similar to our approach, with the major distinction offered by the loss function based on the InfoNCE loss [218]. Following the interest in multi-modal large language models (MLLM), works such as PALM-E [69] and MedPALM M [289] introduced the multi-modal learning task as a language generation task by treating the image representations as a prior. MedPALM M demonstrated strong zero-shot generalization capabilities by training on paired multi-modal data.

4.4 Dataset

For training and evaluating our model, we use the MIMIC-CXR dataset v2.0 [138], consisting of 377,110 chest radiographs associated with 227,835 radiology reports. The data was collected in routine clinical practice, and each report is associated with one or more images. We limited our study to 247,425 frontal-view radiographs.

4.4.1 Regex Labeling

We extracted pulmonary edema severity labels from the associated radiology reports using regular expressions (regex) with negation detection [38]. This quantification takes the form of a numeric assessment of the degree of pulmonary edema, ranging from 0 (absent) to 3 (severe), associated with different radiographic findings. The keywords of each severity level ("none"=0, "vascular congestion"=1, "interstitial edema"=2, and "alveolar edema"=3) are summarized in Appendix D. In order to limit confounding keywords from other disease processes, we limited the label extraction to patients with congestive heart failure (CHF) based on their ED ICD-9 diagnosis code in the MIMIC dataset [91]. Cohort selection by diagnosis code for CHF was previously validated by manual chart review. This resulted in 16,108 radiology reports. Regex labeling yielded labels for 6,710 reports associated with 6,743 frontal-view images[2]. Hence, our dataset includes 247,425 image-text pairs, 6,743 of which are of CHF patients with edema severity labels. Note that some reports are associated with more than one image, so one report may appear in more than one image-text pair.

4.5 Methods

Let x^{I} be a 2D chest radiograph, x^{R} be the free-text in a radiology report, and $y \in \{0, 1, 2, 3\}$ be the corresponding edema severity label. Our dataset includes a set of N image-text pairs $\mathrm{X} = \{\mathrm{x}_j\}_{j=1}^{N}$, where $\mathrm{x}_j = (\mathrm{x}_j^{\mathrm{I}}, \mathrm{x}_j^{\mathrm{R}})$. The first N_{L} image-text pairs are annotated with severity labels $\mathrm{Y} = \{\mathrm{y}_j\}_{j=1}^{N_{\mathrm{L}}}$. Here we train a joint model that constructs an image-text embedding space, where an image encoder and a text encoder are used to extract image features and text features separately (Fig. 4-1). Two classifiers are trained to classify the severity labels independently from the

[2]The numbers of images of the four severity levels are 2883, 1511, 1709, and 640 respectively.

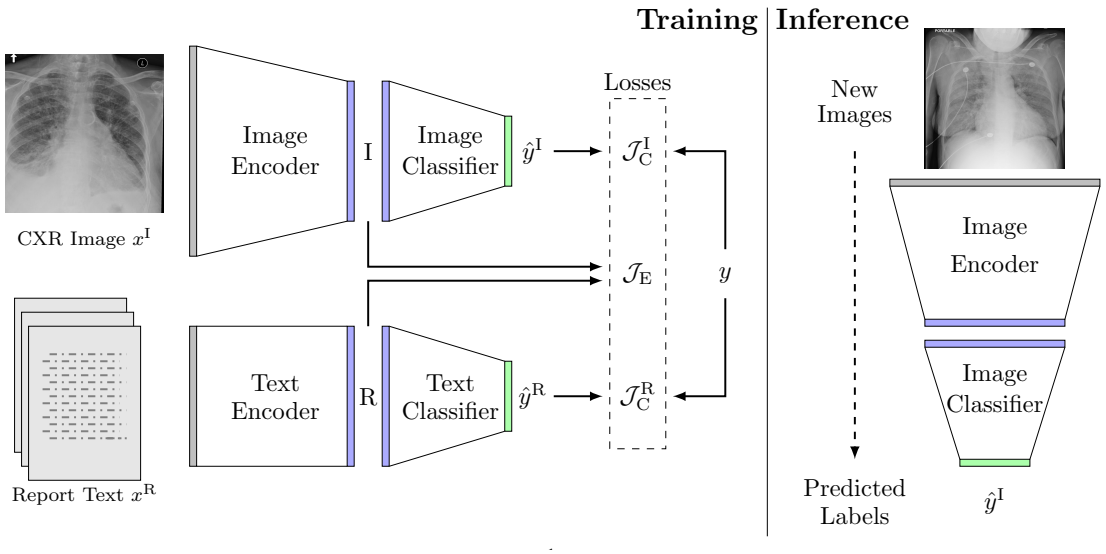


Figure 4-1: The architecture of our joint model, along with an example chest radiograph x^{I} and its associated radiology report x^{R}. At training time, the model predicts the edema severity level from images and text through their respective encoders and classifiers, and compares the predictions with the labels. The joint embedding loss $\mathcal{J}_{\mathrm{E}}$ associates image embeddings I with text embeddings R in the joint embedding space. At inference time, the image stream and the text stream are decoupled and only the image stream is used. Given a new chest radiograph (image), the image encoder and classifier compute its edema severity level.

image features and from the text features. This setup enables us to decouple the image classification and the text classification at inference time. Learning the two representations jointly at training time improves the performance of the image model.

4.5.1 Joint Representation Learning

We apply a ranking-based criterion [44, 104] for training the image encoder and the text encoder parameterized by $\theta_{\mathrm{E}}^{\mathrm{I}}$ and $\theta_{\mathrm{E}}^{\mathrm{R}}$ respectively, to learn image and text feature representations $I(x^{\mathrm{I}}; \theta_{\mathrm{E}}^{\mathrm{I}})$ and $R(x^{\mathrm{R}}; \theta_{\mathrm{E}}^{\mathrm{R}})$. Specifically, given an image-text pair $(\mathrm{x}_j^{\mathrm{I}}, \mathrm{x}_j^{\mathrm{R}})$, we randomly select an impostor image $\mathrm{x}_{s(j)}^{\mathrm{I}}$ and an impostor report $\mathrm{x}_{s(j)}^{\mathrm{R}}$ from X. This selection is generated at the beginning of each training epoch. Map $s(j)$ produces a random permutation of $\{1, 2, ..., N\}$.

We encourage the feature representations between a matched pair $(\mathrm{I}_j, \mathrm{R}_j)$ to be "closer" than those between mismatched pairs $(\mathrm{I}_{s(j)}, \mathrm{R}_j)$ and $(\mathrm{I}_j, \mathrm{R}_{s(j)})$ in the joint embedding space. Direct minimization of the distance between I and R could end up pushing the image and text features into a small cluster in the embedding space. Instead we encourage matched image-text features to be close while spreading out all feature representations in the embedding space for downstream classification by

constructing an appropriate loss function:

$$\begin{aligned}\mathcal{J}_{\mathrm{E}}(\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}; \mathrm{x}_j, \mathrm{x}_{s(j)}) =& \max(0, \mathrm{Sim}(\mathrm{I}_j, \mathrm{R}_{s(j)}) - \mathrm{Sim}(\mathrm{I}_j, \mathrm{R}_j) + \eta) \\ &+ \max(0, \mathrm{Sim}(\mathrm{I}_{s(j)}, \mathrm{R}_j) - \mathrm{Sim}(\mathrm{I}_j, \mathrm{R}_j) + \eta),\end{aligned} \tag{4.1}$$

where $\mathrm{Sim}(\cdot, \cdot)$ is the similarity measurement of two feature representations (explained further in Section 4.5.4) in the joint embedding space and η is a margin parameter that is set to $|\mathrm{y}_j - \mathrm{y}_{s(j)}|$ when both $j \leqslant N_{\mathrm{L}}$ and $s(j) \leqslant N_{\mathrm{L}}$; otherwise, $\eta = 0.5$. The margin is determined by the difference due to the mismatch, if both labels are known; otherwise the margin is a constant.

4.5.2 Classification

We employ two fully connected layers (with the same neural network architecture) on the joint embedding space to assess edema severity from the image and the report respectively. For simplicity, we treat the problem as multi-class classification, i.e., the classifiers' outputs $\hat{y}^{\mathrm{I}}(\mathrm{I}; \theta^{\mathrm{I}}_{\mathrm{C}})$ and $\hat{y}^{\mathrm{R}}(\mathrm{R}; \theta^{\mathrm{R}}_{\mathrm{C}})$ are encoded as one-hot 4-dimensional vectors. We use cross entropy as the loss function for training the classifiers and the encoders on the labeled data:

$$\begin{aligned}\mathcal{J}_{\mathrm{C}}(\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}, \theta^{\mathrm{I}}_{\mathrm{C}}, \theta^{\mathrm{R}}_{\mathrm{C}}; \mathrm{x}_j, \mathrm{y}_j) = &- \sum_{i=0}^{3} \mathrm{y}_{ji} \log \hat{y_i}^{\mathrm{I}}(\mathrm{I}_j(x^{\mathrm{I}}_j; \theta^{\mathrm{I}}_{\mathrm{E}}); \theta^{\mathrm{I}}_{\mathrm{C}}) \\ &- \sum_{i=0}^{3} \mathrm{y}_{ji} \log \hat{y_i}^{\mathrm{R}}(\mathrm{R}_j(x^{\mathrm{R}}_j; \theta^{\mathrm{R}}_{\mathrm{E}}); \theta^{\mathrm{R}}_{\mathrm{C}}),\end{aligned} \tag{4.2}$$

i.e., minimizing the cross entropy also affects the encoder parameters.

4.5.3 Loss Function

Combining Eq. (4.1) and Eq. (4.2), we obtain the loss function for training the joint model:

$$\mathcal{J}(\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}, \theta^{\mathrm{I}}_{\mathrm{C}}, \theta^{\mathrm{R}}_{\mathrm{C}}; \mathrm{X}, \mathrm{Y}) = \sum_{j=1}^{N} \mathcal{J}_{\mathrm{E}}(\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}; \mathrm{x}_j, \mathrm{x}_{s(j)}) + \sum_{j=1}^{N_{\mathrm{L}}} \mathcal{J}_{\mathrm{C}}(\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}, \theta^{\mathrm{I}}_{\mathrm{C}}, \theta^{\mathrm{R}}_{\mathrm{C}}; \mathrm{x}_j, \mathrm{y}_j). \tag{4.3}$$

4.5.4 Implementation Details

The image encoder is implemented as a series of residual blocks [110], the text encoder is a BERT model that uses the beginning `[CLS]` token's hidden unit size of 768 and maximum sequence length of 320 [64]. The image encoder is trained from a random initialization, while the BERT model is fine-tuned during the training of the joint model. The BERT model parameters are initialized using pre-trained weights on scientific text [18]. The image features and the text features are represented as

768-dimensional vectors in the joint embedding space. The two classifiers are both 768-to-4 fully connected layers. The neural network architecture is provided in the supplementary material.

We employ the stochastic gradient-based optimization procedure AdamW [308] to minimize the loss in Eq. (4.3) and use a warm-up linear scheduler [292] for the learning rate. The model is trained on all the image-text pairs by optimizing the first term in Eq. (4.3) for 10 epochs and then trained on the labeled image-text pairs by optimizing Eq. (4.3) for 50 epochs. The mini-batch size is 4. We use dot product as the similarity metric in Eq. (4.1). The dataset is split into training and test sets. All the hyper-parameters are selected based on the results from 5-fold cross validation within the training set.

4.6 Experimental Results

4.6.1 Data Preprocessing

The size of the chest radiographs varies and is around 3000×3000 pixels. We randomly translate and rotate the images on the fly during training and crop them to 2048×2048 pixels as part of data augmentation. We maintain the original image resolution to capture the subtle differences in the images between different levels of pulmonary edema severity. For the radiology reports, we extract the *impressions*, *findings*, *conclusion* and *recommendation* sections. If none of these sections are present in the report, we use the *final report* section. We perform tokenization of the text using ScispaCy [214] before providing it to the BERT tokenizer.

4.6.2 Expert Labeling

For evaluating our model, we randomly selected 531 labeled image-text pairs (corresponding to 485 reports) for expert annotation. A board-certified radiologist and two domain experts reviewed and corrected the regex labels of the reports. We use the expert labels for model testing. The overall accuracy of the regex labels (positive predictive value compared against the expert labels) is 89%. The other 6,212 labeled image-text pairs and around 240K unlabeled image-text pairs were used for training. There is no patient overlap between the training set and the test set.

4.6.3 Model Evaluation

We evaluated variants of our model and training regimes as follows:

- **image-only**: An image-only model with the same architecture as the image stream in our joint model. We trained the image model in isolation on the 6,212 labeled images.
- A joint image-text model trained on the 6,212 labeled image-text pairs only. We compare two alternatives to the joint representation learning loss:

  - **ranking-dot**, **ranking-l2**, **ranking-cosine**: the ranking based criterion in Eq. (4.1) with $\text{Sim}(I, R)$ defined as one of the dot product $I^\top R$, the reciprocal of euclidean distance $-\|I - R\|$, and the cosine similarity $\frac{I^\top R}{\|I\|.\|R\|}$;
  - **dot**, **l2**, **cosine**: direct minimization on the similarity metrics without the ranking based criterion.

- **ranking-dot-semi**: A joint image-text model trained on the 6,212 labeled and the 240K unlabeled image-text pairs in a semi-supervised fashion, using the ranking based criterion with dot product in Eq. (4.1). Dot product is selected for the ranking-based loss based on cross-validation experiments on the supervised data comparing ranking-dot, ranking-l2, ranking-cosine, dot, l2, and cosine.

All reported results are compared against the expert labels in the test set. The image portion of the joint model is decoupled for testing, and the reported results are predicted from images only. To optimize the baseline performance, we performed a separate hyper-parameter search for the `image-only` model using 5-fold cross validation (while holding out the test set).

We use the area under the ROC (AUC) and macro-averaged F1-scores (macro-F1) for our model evaluation. We dichotomize the severity levels and report 3 comparisons (0 *vs* 1,2,3; 0,1 *vs* 2,3; and 0,1,2 *vs* 3), since these 4 classes are ordinal (e.g., $\mathbb{P}(\text{severity} = 0 \text{ or } 1) = \hat{y}^{\text{I}}_0 + \hat{y}^{\text{I}}_1$, $\mathbb{P}(\text{severity} = 2 \text{ or } 3) = \hat{y}^{\text{I}}_2 + \hat{y}^{\text{I}}_3$).

4.6.4 Results

Method	AUC (0v123)	AUC (01v23)	AUC (012v3)	macro-F1
`l2`	0.78	0.76	0.83	0.42
`ranking-l2`	0.77	0.75	0.80	0.43
`cosine`	0.77	0.75	0.81	0.44
`ranking-cosine`	0.77	0.72	0.83	0.41
`dot`	0.65	0.63	0.61	0.15
ranking-dot	**0.80**	**0.78**	**0.87**	**0.45**

Table 4.1: Performance statistics for all similarity measures.

Method	AUC (0v123)	AUC (01v23)	AUC (012v3)	macro-F1
`image-only`	0.74	0.73	0.78	0.43
`ranking-dot`	0.80	0.78	0.87	0.45
ranking-dot-semi	**0.82**	**0.81**	**0.90**	**0.51**

Table 4.2: Performance statistics for the two variants of our joint model and the baseline image model.

Table 4.1 reports the performance statistics for all similarity measures. The findings are consistent with our cross-validation results: the ranking based criterion offers

significant improvement when it is combined with the dot product as the similarity metric.

Table 4.2 reports the performance of the optimized baseline model (`image-only`) and two variants of the joint model (`ranking-dot` and `ranking-dot-semi`). We observe that when the joint model learns from the large number of unlabeled image-text pairs, it achieves the best performance. The unsupervised learning minimizes the ranking-based loss in Eq. (4.1), which does not depend on availability of labels.

It is not surprising that the model is better at differentiating the severity level 3 than other severity categories, because level 3 has the most distinctive radiographic features in the images.

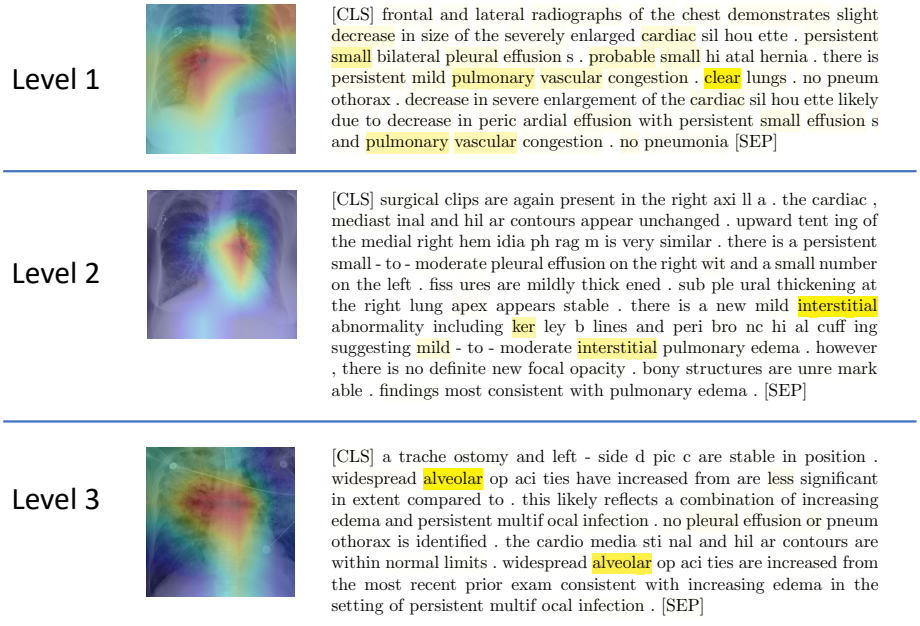


Figure 4-2: Joint model visualization. Top to bottom: (Level 1) The highlight of the Grad-CAM image is centered around the right hilar region, which is consistent with findings in pulmonary vascular congestion as shown in the report. (Level 2) The highlight of the Grad-CAM image is centered around the left hilar region which shows radiating interstitial markings as confirmed by the report heatmap. (Level 3) Grad-CAM highlights bilateral alveolar opacities radiating out from the hila and sparing the outer lungs. This pattern is classically described as “batwing" pulmonary edema mentioned in the report. The report text is presented in the form of sub-word tokenization performed by the BERT model, starting the report with a `[CLS]` token and ending with a `[SEP]`.

4.6.5 Joint Model Visualization

As a by-product, our approach provides the possibility of interpreting model classification using text. While a method like Grad-CAM [267] can be used to localize regions

in the image that are “important" to the model prediction, it does not identify the relevant characteristics of the radiographs, such as texture. By leveraging the image-text embedding association, we visualize the heatmap of text attention corresponding to the last layer of the `[CLS]` token in the BERT model. This heatmap indicates report tokens that are important to our model prediction. As shown in Fig. 4-2, we use Grad-CAM [267] to localize relevant image regions and the highlighted words (radiographic findings, anatomical structures, etc.) from the text embedding to explain the model's decision making.

4.7 Conclusion

In this chapter, we presented a neural network model that jointly learns from images and text to assess pulmonary edema severity from chest radiographs. The joint image-text representation learning framework incorporates the rich information present in the free-text radiology reports and significantly improves the performance of edema assessment compared to learning from images alone. Moreover, our experimental results show that joint representation learning benefits from the large amount of unlabeled image-text data.

This approach offered the first way in the research community to jointly learn image and text representations using contrastive learning, and allowed a way to take advantage of large amount of paired image-text data towards better downstream classification with fewer labels. Our approach, though innovative, could have several limitations. The latent image-text representations being contrasted after full processing through the ResNet and BERT models might only offer regularization at a higher semantic level, whereas clinical findings are often specific, located within small regions of the image and mentioned in few words inside the radiology report. Works such as GLoRIA [120] improved on this by introducing contrastive connections at earlier model layers, but subsequent works in autoregressive multimodal radiology report generation showcased further improvements in label efficiency and interpretability. Discussions of such works are covered in Chapter 5.

Medical Vision-Language Modeling (MVLM) since the publishing of our paper has exploded, with the latest work MedPALM M [289] able to showcase zero-shot generalization towards the tasks of medical question answering, radiology report generation, image interpretation and genomic variant calling.

While the contributions of recent works is impressive, there remains significant work to be done in directions such as accurate disease progression modeling and validating the factuality of generated reports. Some works have proposed the use of retrieval augmented generation [161] for improving performance in rare disease scenarios [102]. They also discuss challenges such as data privacy, the need for design of better evaluation metrics for report generation, reduction of hallucinations and overcoming catastrophic forgetting. Mainly, the adoption of MVLMs in clinical scenarios is also dependent on a healthy collaboration between machine learning experts and clinical stakeholders involving several parameters such as trust, alignment with clinical needs and ethical deployment [102].

Chapter 5

Label Efficiency and Interpretability via autoregressive multi-modal report generation

Abstract

This thesis investigates methods that take advantage of clinical structure to improve representation learning capabilities. Following our work in Chapter 4, we improve upon multi-modal representation learning by tackling challenges associated with label efficiency and interpretability. This chapter is based on the work [242, 243, 244].

Automated medical image analysis remains a challenging task due to limited availability of annotated data, whose creation is time-consuming and necessitates expert knowledge, making fully supervised modeling approaches impractical. Self-supervised learning techniques such as contrastive learning [41, 334, 286] demonstrated substantial improvements in image representation, but don't offer direct interpretability for model decisions. Multi-modal Large Language Models (LLMs) such as [289] have recently demonstrated zero-shot and language generation capabilities. Encouraging the model to generate accurate captions during training could enable interpretable outputs to accompany downstream decisions in radiologist workflows.

In this chapter, we introduce RadTex, a multi-modal autoregressive captioning approach to learn *efficient* radiograph representations that require fewer labels for training compared with image encoders learned via supervised pre-training and contrastive learning. To assess downstream task quality, we first pre-train the image encoder on a captioning objective, and then fine-tune the learned encoder over the downstream task labels. Though generating quality radiology reports is not a direct objective of this work, we showcase how our pre-training method encourages the model to capture fine-grained semantics and provide interpretable outputs that align with clinical workflows. Additionally, we introduce a novel approach for probing our model with textual prompts, highlighting the potential of bicaptioning pre-training for delivering interactive models to support radiologist workflows.

5.1 Introduction

Clinical practice involves a high degree of multi-modality and requires interpretability to develop trust in proposed patient assessments. To support automated medical image analysis, multi-modal representation learning methods have emerged as a new technique to take advantage of the multiple views of the same patient for comprehensive and efficient assessment of their conditions. Contrastive learning approaches, described in Chapter 4, demonstrated substantial improvements in image representation quality compared with supervised learning.

Contrastive learning approaches, though powerful, only encode high level semantics due to the similarity induced in later layers of the image and text encoders. Attempts at improving these limitations were proposed in works such as GLoRIA [120], but assessing interpretability of these techniques still remained challenging. Large Language Models (LLMs) have received widespread attention for their zero-shot capabilities on question-answering, comprehension and reasoning tasks [27, 150]. Generative decoder-only and encoder-decoder models use next token prediction and masked language modeling tasks to learn complex semantic meaning from unlabeled training sequences, and scaling these models up to billions of parameters has led to impressive results [246, 249]. LLMs also demonstrate the ability to encode clinical knowledge [271], showcasing their usefulness in the healthcare domain.

We hypothesize that using generative token-level modeling to process radiology reports may encourage fine-grained medical semantics to be encoded, like relations between pathologies or subregion contents. Though numerous appraoches have adapted contrastive learning to address this challenge, pretraining with next-token prediction tasks similar to [320] may allow inherently capturing these semantics.

5.2 Our Contributions

We introduce RadTex, a simple pretraining approach that utilizes bidirectional captioning (*bicaptioning*) to learn chest radiograph representations. Bidirectional captioning refers to pre-training using forward and backward captioning to more accurately capture sentence semantics. Our contributions are as follows:

1. Our framework has significant label-efficient contributions. RadTex pre-training outperforms ImageNet and in-domain supervised pre-training methods when fewer than 1000 labeled examples are available. When labeled data is reduced to 100 examples, decreases of only 0.05 AUC and 0.01 AUPRC are observed.

2. Our analysis into the amount of image-text paired data necessary for pre-training amounts to at least 100k examples to build transferrable representations. This has long-term implications for aligned data access.

3. We showcase the competitiveness of bicaptioning pre-training over contrastive learning, while continuing to beat prior works in the domain on label efficiency.

4. RadTex is inherently interpretable due to the bicaptioning generation objective used in pre-training. Minimal modifications are made to generate high-quality radiology reports, and the clinical efficacy benefits of prompting approaches using our model are introduced for enhanced radiologist-AI interaction.

5.3 Related Work

Our work is closest to the general domain bicaptioning framework VirTex by Desai and Johnson [62] who demonstrate improved label efficiency compared with supervised pre-training and contrastive learning approaches. The domain RadTex and VirTex operates in is closest to the subfields of medical vision-language modeling (MVLM) and radiology report generation (RRG).

5.3.1 Medical Vision-Language Modeling

Before self-supervised deep learning approaches were introduced, successful training of visual backbones such as ResNet and Vision Transformers [110, 68] required large amounts of annotations. Limited labeled examples and time-consuming annotation in the medical domain limited progress of such approaches until the release of large-scale weakly annotated datasets [123, 138] and the development of self-supervised vision-language modeling.

Work from Chapter 4 [41] effectively encoded radiographs and radiology reports using a contrastive loss function to pre-train an effective visual encoder. ConVIRT [334] similarly proposed adapting the InfoNCE loss [218] to vision-language tasks for learning visual representations. CLIP [247] was directly inspired by ConVIRT, and quickly became state-of-the-art for representation learning for natural images and text. CheXzero [286] adapted this technique back to radiology, after replacing the vision encoder with a vision transformer [68]. Simultaneously, there were efforts to improve contrastive pre-training by incorporating both local semantics from image patches and subwords with global representations [120, 296].

Transformer-based [293] encoder-decoder architectures have recently gained attention in MVLM. Popular approaches in the natural domain [187, 163, 338, 62, 320] have given way to radiology-specific applications that take advantage of encoder-decoder or decoder models. Bannur et al. [13] augmented their contrastive learning pre-training with a transformer decoder, and Yan and Pei [315] used masked language modeling (MLM) alongside other objectives to train an encoder-decoder but didn't make direct comparisons with contrastive approaches for MVLP. PRIOR [46] took inspiration from CoCa [320], but assumed report sentences to be non-sequential & modeled their reconstruction with a prototype bank of sentences, rather than token-level decoding.

Transformer-decoder based autoregressive models have recently gained popularity [27], and MedPaLM M [289] recently showed impressive results by combining representations from a vision transformer into the autoregressive text-based objective. Through their billion parameter model, they demonstrated zero-shot capabilities across a variety of tasks in language, imaging and genomics.

5.3.2 Radiology Report Generation

The task of radiology report generation is inspired by the general domain of image captioning [294, 277], but captioning in the radiology domain is associated with unique challenges. Radiology reports are typically longer than natural image captions, and require stronger semantic consistency to clinical concepts. Some authors have proposed using hybrid retrieval-generation models or templates [169, 231, 73].

Neural text generation, which allows for free-text generation from an image, does not require retrieval from a bank of templates or existing reports and is a viable alternative to contrastive zero-shot classification. Jing et al. [135] and TieNet [302] used LSTM-based generation, while $\mathcal{M}^2$ Trans [204] and R2Gen [45] utilize transformers to generate relevant clinical reports. Large language models have also been used to improve report generation, with CvT212DistilGPT2 [217] fine-tuning a distilled LLM on MIMIC-CXR, and generalist model MedPaLM M demonstrating report generation with zero- and few-shot prompting [289].

5.4 Methods

RadTex employs the bidirectional captioning (*bicaptioning*) pre-training approach similar to the general domain framework by Desai and Johnson [62]. The objective of our work is to improve medical image representations and simultaneously generate interpretable radiology reports. Figure 5-1 provides a visual representation of RadTex capabilities.

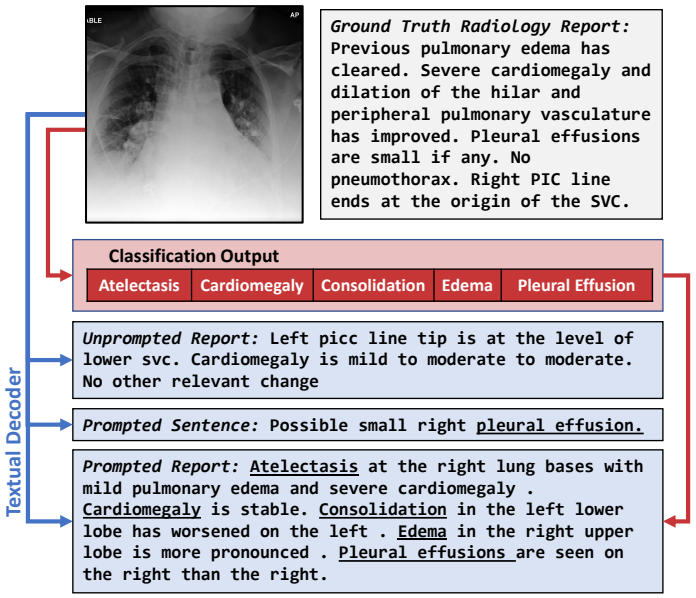


Figure 5-1: Overview of RadTex interpretable outputs.

Given an image, RadTex provides predictive classification labels and a detailed radiology report (I.e., unprompted report). Additionally, RadTex can take advantage

of prompting capabilities for transformer decoders to provide more accurate explanations in either the sentence or multi-sentence form. The generation mechanism is flexible, allowing for the output in sentence or full report form.

The full model architecture is presented in Figure 5-2.

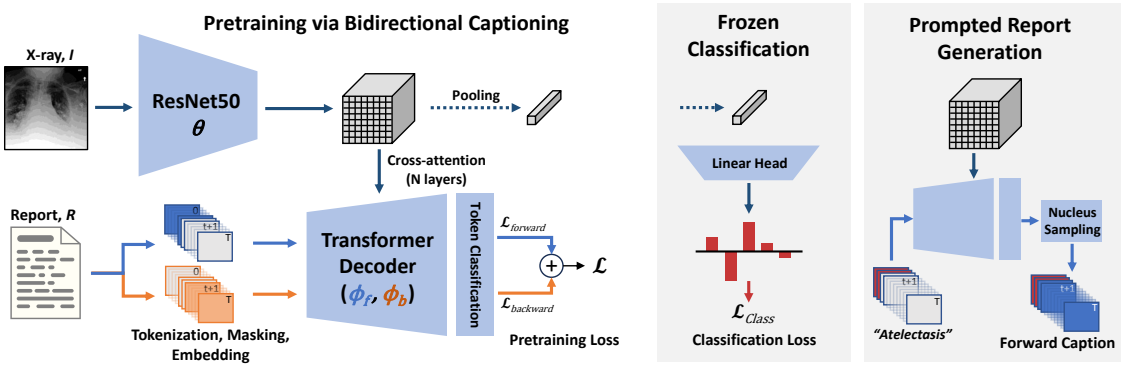


Figure 5-2: Overview of RadTex architecutre, pre-training, classification experiments and report generation in the *Prompted* setting. Report generation (right) does not require any additional training following pre-training. The ResNet50 and Transformer Decoder are both frozen for downstream tasks.

5.4.1 Pre-training

An image encoder and textual decoder are jointly trained via bidirectional image captioning. A CXR Image I and a paired radiology report R are transformed into sequences. We use ResNet50 [110], denoted θ, to extract visual features from I after which a linear projection is applied to create a sequence of spatial image features x_{vis}. R is processed via tokenizer ϕ_{tok} into tokens $w = \{w_0, ..., w_{T+1}\}$ where $w_i \in D$ I.e., the vocabulary of the tokenizer. The entire input sequence is encoded by a learned word and positional embedding into a text sequence x_{text}.

Both x_{vis} and x_{text} are processed via the transformer decoder, attending to x_{text} via masked multi-head self attention and x_{vis} via cross attention. Implementation for the transformer decoder follows that for VirTex [62]. The transformer decoder architecture is duplicated for the backward captioning, and weights across both are not shared. Input sequence masks when predicting the t-th token w_t are applied to $w_{i \geq t}$ and $w_{i \leq t}$ for forward and backward transformers, respectively. The output from each of the forward and backward transformers are passed through a shared linear layer to predict the token-wise log probabilities. We compute cross-entropy loss on these logits, minimizing negative log-likelihood of selecting the correct tokens.

5.4.2 Explainable Report Generation

The pre-trained model can generate reports up to the maximum sequence length, and uses autoregressive captioning to iteratively build upon an initial sequence by adding one token at a time. At each step, token logits are computed using the frozen

visual encoder and textual decoder, based on the input image I and the existing token sequence. A token is then sampled from the computed logits using a strategy like nucleus sampling [115], which focuses on generating tokens that are more coherent and contextually appropriate.

Prior to report generation, a chosen string can be inserted as an additional input to the transformer decoder, conditioning the rest of the generation on the input string (known as *prompting*). We refer to the report from the standard generation scenario as the *unprompted* report. We propose two prompting strategies: 1) *Prompted* captioning begins with adding a word or phrase (typically a pathology of interest), prompting the model to run forward captioning given the starting sequence; and 2) *Iterative Prompted* captioning which follows the *Prompted* procedure to generate a report, and then treats that report as a prompt for the backward captioning step. The *Iterative Prompted* setting allows the ability to generate additional tokens, and might provide the model with capabilities to more completely detail a particular finding.

5.5 Datasets

5.5.1 Pre-training

MS-COCO [178] is a natural images and paired captions dataset, whose official 2017 split of 118K image-caption pairs are used for pre-training.

MIMIC-CXR [138, 139] contains 377,110 Chest X ray images with paired radiology reports and CheXpert pathology labels. For pre-training, *Findings* and *Impression* sections of reports and both frontal and lateral MIMIC-CXR images are used. Reports are pre-processed to remove references to prior studies in the training and validation sets (train/val/test: 368,960/2991/5159).

5.5.2 Fine-tuning

CheXpert [123] includes 224,316 chest radiographs from Stanford Hospital. Following the official split, our experiments focus on five competition pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

Pathology9 Liao et al. [172] derived Pathology9 by running the CheXpert labeler from [123] on the MIMIC-CXR radiology reports [138] to derive 14 pathology labels. Further, they filtered the dataset to only include 9 of the 14 original pathology labels with over 100 test examples.

Edema Severity [170] is derived from MIMIC-CXR, grading 7,390 radiographs for pulmonary edema severity on a 0-3 scale (0: none, 1: vascular congestion, 2: interstitial edema, 3: alveolar edema). While most labels derive from regular expressions (regex), the test set ($n = 141$) uses consensus labels from radiologists. This test set is unseen during MIMIC-CXR pretraining.

RSNA Pneumonia [269] comprises approximately 30,000 frontal radiographs from NIH CXR-8 [301], labeled for pneumonia presence.

COVIDx [227] is an updated version of the multinational COVIDx-CXR3 data which we use. The dataset contains COVID and non-COVID labels for 30,386 images. We report on their official, class balanced, 400-image test set, and use 5% of the training set for validation set, split by patient identifier.

5.6 Results

5.6.1 Result 1: Label Efficiency

In this experiment, we investigate RadTex's effectiveness towards learning visual representations during pre-training and its downstream transfer efficency compared to other models with the same encoder architecture: randomly initialized ResNet50 [110], or pre-trained with ImageNet (IN-Pretrained) [61] or ChestX-ray14 (CXR14-Pretrained) [117, 301]. For evaluating these set of experiments, we pre-trained the model on MIMIC-CXR dataset [138, 139], and fine-tuned on both the Pathology9 [172] and Edema Severity tasks [170], while carefully ensuring that the test sets were not seen during pre-training.

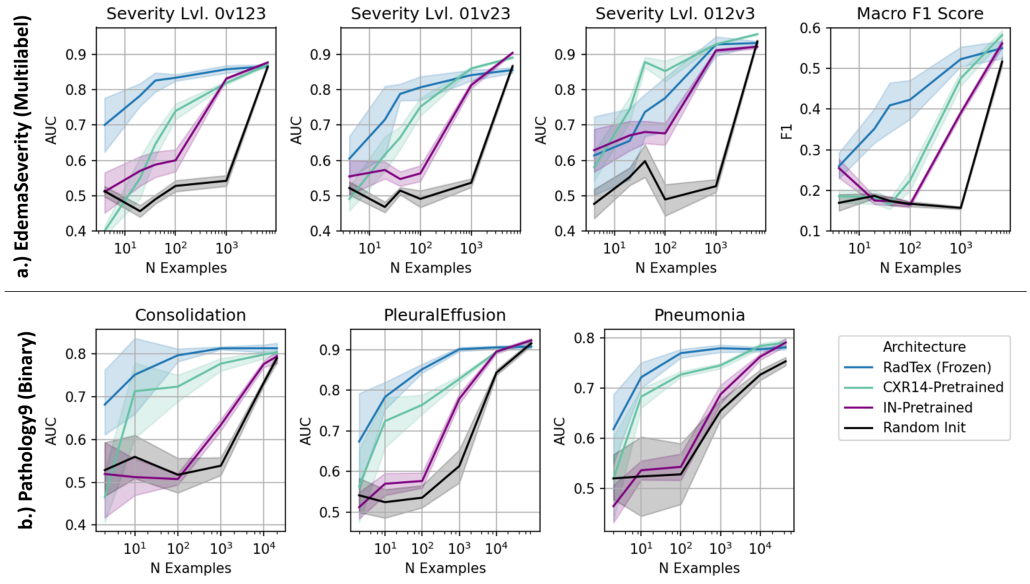


Figure 5-3: AUC with a varying amount of labeled training images (N) from a) EdemaSeverity and b) Pathology9. We compare frozen RadTex to other initializations, as unfrozen RadTex results were similar. Mean AUC from five trials and 95% confidence interval is shown. Macro F1 score is reported for EdemaSeverity.

Figure 5-3 compares area under the receiver operating characteristic curve (shortened as AUC) [199] for RadTex, CXR14-Pretrained ResNet, IN-Pretrained and Random init. All comparisons are for ResNet models pre-trained under different paradigms:

1) MIMIC-CXR based autoregressive captioning (RadTex where frozen refers to freezing the visual encoder and training the linear projection layer for classification), 2) CXR14 based supervised pre-training, 3) ImageNet based supervised pre-training and 4) random initialization. In each experiment, the ResNet models are fine-tuned on the exact same datasets for fair comparisons over the label efficiency of different pre-training schemes. We train all models on randomly sampled portions of the downstream fine-tuning dataset to investigate the relationship between labeled dataset size and performance.

For both downstream datasets we consider, RadTex trains visual representations that transfer to downstream tasks with much better efficiency than other models. When 1000 or fewer examples are available, RadTex matches or outperforms ImageNet and CXR14 pre-trained models. Crucially for Pathology9, we found that RadTex performance on 100 labeled examples (AUC: 0.752) was nearly as good as the top performing model with 100% labeled data (AUC: 0.801), where ∼100 is a reasonably sized dataset to ask physicians to annotate.

The major benefit of RadTex is in taking advantage of a multi-view approach using paired radiographs and reports, reducing the amount of downstream labeled data necessary for fine-tuning. In settings where the multi-view data is abundantly available and cheaply collected, the costs of having clinicians labeling downstream diseases will be significantly reduced, if RadTex is used.

5.6.2 Result 2: Pre-training data necessity

While it is intuitive that pre-training with more examples produces better visual representations, our next experiment reveals quantitative evidence that justifies the importance of having sufficient image-text aligned data towards efficient multi-modal learning. In order to conduct a brief study, we pre-train RadTex on varying fractions of MIMIC-CXR radiographs and reports, and fine-tune the learned visual encoder over fewer labels to simulate low-label settings.

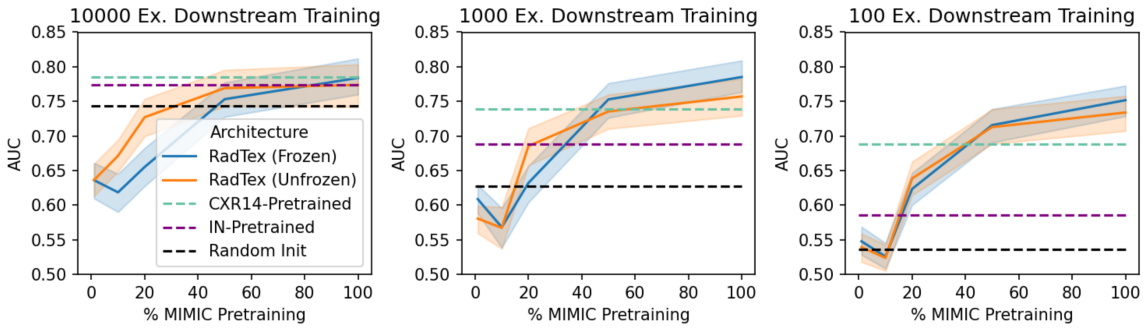


Figure 5-4: Averaged Pathology9 AUCs after training on 10K, 1K and 100 downstream examples vs. pretraining dataset (MIMIC-CXR) size.

In Figure 5-4, we capture the effect of reducing the size of the pre-training set by transferring models to 10K, 1K and 100 example downstream fine-tuning on Pathol-

ogy9. We observe that pre-training with a dataset size 50% of the MIMIC-CXR corpus degrades downstream AUCs only slightly compared to the full pre-training set (100 example Pathology9 drops by 0.03 AUC), yet pre-training with smaller datasets yields significantly worse performance, as the model seems unable to "recover" from poor initialization. Noting that 50% MIMIC pre-training (∼125k examples) represents an inflection point in RadTex downstream performance in Figure 5-4, and that VirTex similarly used a pre-training dataset with 118k examples, we recommend that a pre-training dataset of at least 100k multi-modal examples be available for those looking to apply RadTex to additional radiographic modalilities or imaging regions.

5.6.3 Result 3: Comparison to Contrastive Learning

We evaluate the quality of our learned visual encoder on four downstream classification tasks: CheXpert, COVIDx, RSNA pneumonia detection and Pulmonary Edema Severity. All visual encoders are compared in the frozen linear classification setup, where a randomly initialized linear classification head is added to a pre-trained encoder, and only the linear head is trained on available data.

We first showcase the results in Figure 5-5, where RadTex is pre-trained over MIMIC-CXR and MS-COCO (called RadTex/C+M), and other comparative models are visual encoders from CheXzero (ViT-B/32) [286], BiomedCLIP (ViT-B/16) [327], OpenAI's CLIP model (ResNet50) [247] and ResNet50 which is randomly initialized (called Supervised) and tuned end-to-end [110]. ViT-B refers to visual transformer base (ViT-Base) introduced by Dosovitskiy et al. [68] with 12 layers and the suffix "/16 " and "/32 " refers to the patch size of either 16×16 or 32×32 pixels.

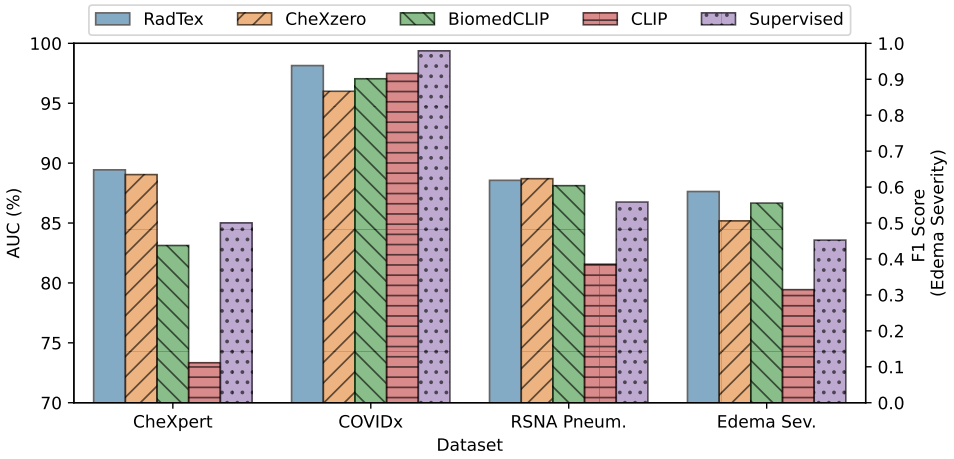


Figure 5-5: Bar plot showing linear classification results. RadTex is competitive with CheXzero and other methods across multiple downstream classification tasks. RadTex results are for RadTex/C+M pretraining. Each model's visual backbone is frozen and a linear layer is trained in three separate trials. We display mean results over three random trials.

In this experiment, we were interested in evaluating the quality of the learned

visual representation by directly comparing the ability to transfer models for downstream classification using a learned linear head. CheXzero, initialized from CLIP loss and pre-trained on MIMIC-CXR, is considered the closest contrastive comparison in terms of the visual encoder size and pre-training datasets. BiomedCLIP is trained on a crawled set of 15 million image-text pairs from PubMed and serves as an alternative data source when in-domain data might be scarce. CLIP is trained on 400 million image-text pairs and serves as a natural domain baseline. Supervised represents a randomly initialized ResNet50 trained in a supervised fashion on the available labels for each task.

Figure 5-5 shows that RadTex outperforms frozen encoder benchmarks in three out of four tasks, except for the RSNA Pneumonia challenge. Notably, the COVIDx task involves binary classification on a condition not existing in the pre-training data. RadTex still achieves the highest score among the pre-trained models, which demonstrates its generalizability in a zero-shot fashion. While Supervised exceeds performance for all frozen visual encoders on COVIDx, an additional end-to-end fine-tuning experiment for all encoders showcased that RadTex achieves the highest score across all baselines.

		Visual	CheXpert AUC (%)		
		Enc.	1%	10%	100%
	Random Init*	RN50	$71.9_{\pm 2.6}$	$82.3_{\pm 3.5}$	$85.0_{\pm 4.7}$
Contrastive	OpenAI CLIP	RN50	$59.8_{\pm 0.4}$	$72.4_{\pm 0.6}$	$73.3_{\pm 0.2}$
	ConVIRT	RN50	85.9	86.8	87.3
	BiomedCLIP	ViT-B	$77.2_{\pm 0.9}$	$82.1_{\pm 2.6}$	$83.1_{\pm 0.5}$
	GLoRIA	RN50	86.6	87.8	88.1
	MGCA	RN50	87.6	88.0	88.2
	CheXzero	ViT-B	$88.9_{\pm 0.6}$	$89.1_{\pm 0.2}$	$89.0_{\pm 0.4}$
Cap.	VirTex/C+M	RN50	$86.6_{\pm 0.8}$	$86.7_{\pm 0.7}$	$87.3_{\pm 0.2}$
	RadTex/M	RN50	$88.4_{\pm 0.2}$	$89.2_{\pm 0.5}$	$89.0_{\pm 0.4}$
	RadTex/C+M	RN50	$\mathbf{89.2_{\pm 0.4}}$	$\mathbf{89.6_{\pm 0.1}}$	$\mathbf{89.4_{\pm 0.1}}$

Table 5.1: CheXpert competition linear classification results with variable amounts of downstream fine-tuning data. $\text{Mean}_{\pm \text{SD}}$ AUC across 3 trials presented. Random initialization is fine-tuned end-to-end, while other models are frozen and only a linear head is trained. ConVIRT, GLoRIA, and MGCA results from [334], [120], and [296], respectively. RadTex/C+M (bottom) denotes pretraining on both COCO and MIMIC CXR datasets.

In addition to comparing the frozen encoder performance for RadTex versus popular contrastive learning approaches, we assessed the performance of models trained on limited labeled data as well. In Table 5.1, we additionally perform direct comparisons against contrastive methods like GLoRIA [120] and MGCA [296] that are based on ResNet50 encoder like ours. Despite their complex semantic encoding methods taking

into account cross modal similarity between patches of the images and subwords in the text, they do not achieve CheXzero's performance level, which is likely due to the power of the visual encoders in both cases (CheXzero's ViT has 88M params whereas ResNet50 is 24M params only). Despite the limitation of these methods, RadTex's ResNet50 is competitive with CheXzero's ViT, exceeding its performance at all levels when COCO and MIMIC-CXR are used for pre-training, speaking to the effectiveness of bidirectional captioning pre-training.

Recently, billion-parameter foundation models have presented a promising avenue for zero-shot performance across a range of clinical tasks. MedPaLM M [289], reports few classification results and does not make their code or model weights available. However, on the CheXpert competition pathologies in the MIMIC-CXR dataset (further subset of Pathology9), they report a macro-AUC of 79.09% in Table 2 of their paper. To the best of our ability, we replicated their classification experiment with a frozen RadTex encoder, finding a macro-AUC of 84.07%, suggesting that much smaller domain-specific models may outperform larger, generalist ones in specialized tasks.

5.6.4 Result 4: Radiology Report Generation Quality

Method	Textual Similarity		Clinical Efficacy		
	BLEU-2	BERTScore†	CheXpert macro-F1	CheXbert† Cosine Sim.	RadGraph† F1
Random Report Retrieval	0.089	0.213	0.177	0.166	0.048
$\mathcal{M}^2$ Trans [204]	–	0.227	0.304	0.268	0.110
R2Gen [45]	**0.218**	0.186	0.276	0.204	0.057
CXR-RePaiR[73]	0.069	0.191	0.256	**0.379**	0.091
MedPaLM M	—	—	**0.398**	—	**0.267**
RadTex Unprompted	<u>0.100</u>	0.261	0.289	0.259	0.096
RadTex Prompted	0.069	0.262	<u>0.349</u>	<u>0.336</u>	0.098
RadTex Iterative Prompt	0.082	<u>**0.271**</u>	<u>0.349</u>	0.333	<u>0.112</u>

Table 5.2: Comparison of radiology report captioning techniques on a range of metrics. BLEU and BERTScore represent measures of textual similarity, without clinical awareness. CheXpert macro-F1, CheXbert, and RadGraph F1 scores represent measures of clinical efficacy. Other model scores are drawn from existing literature, and we follow their setups as described when comparing RadTex results. †Following [319] and using only *Impression* section for ground truth. For BLEU-2 and CheXpert F1 scores, R2Gen and CXR-RePaiR compare to ground truths with both *Findings* and *Impression*, while $\mathcal{M}^2$ Trans uses just *Findings*. Key: **Best Result**, <u>Best RadTex Result</u>

Though generating high-quality captions is not a direct goal of our research, RadTex's inherent ability to perform radiology report generation provides interpretability and insight into the representations learned during pre-training. We investigate captioning in *Unprompted*, *Prompted* and *Iterative Prompted* settings, as described in Section 5.4.2, with RadTex/C+M pre-training.

We measure both the textual similarity and clinical efficacy for reporting caption quality. BLEU-2 [224] and BERTScore [329] are common natural language generation metrics for syntactic and semantic similarity, but fail to describe the diagnostic accuracy of reports. To measure clinical efficacy, we also report CheXpert [123] macro-F1 scores over 14 labeled pathologies, CheXbert [272] embedding cosine similarities between prediction and ground truth, and RadGraph F1 [129], which converts reports into knowledge graphs and computes overlap with ground truth.

We generate captions for the 5,159 MIMIC-CXR official test set radiographs. Table 5.2 compares RadTex to other radiology report generation (RRG) methods: $\mathcal{M}^2$ Trans [57, 204], R2Gen [45] the two top-performing transformer-based methods, and CXR-RePaiR [73] is cited as a report retrieval technique. Boag et al. [23] noted that randomly retrieved reports score surprisingly well on these metrics, and these results are included as a baseline. 82B parameter MedPaLM M is also included as a baseline, although Tu et al. [289] added the *Indication* section to model context for RRG.

In Table 5.2, RadTex Unprompted, which was optimized through ablations for downstream visual performance, performs comparably to other transformer-based RRG methods: it falls between performances of R2Gen and $\mathcal{M}^2$ Trans on clinical efficacy metrics, and beats both on the BERTScore textual similarity metric. Through prompting, RadTex exceeds performance of both transformer-based models on CE metrics, but still falls well short of generalist model MedPaLM M.

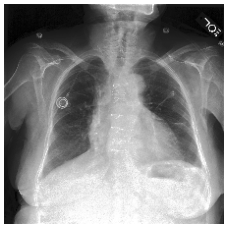
CXR Image	Ground Truth	RadTex Generated Report
	Right port catheter tip ends in the mid svc. Sternal wires are intact and midline. There are small bilateral pleural effusions, slightly larger on the left than on the right. The cardiac silhouette is moderately enlarged. There is mild engorgement of the pulmonary vasculature. There has been improvement in the previously noted pulmonary edema with minimal residual edema. There is plate-like atelectasis seen in the left base. There is no consolidation or pneumothorax. Number. Small bilateral pleural effusions. Number. Improvement in pulmonary edema.	Left-sided central venous catheter terminates in the mid svc. Overall cardiac size is top normal. Aortic knob is noted. Patient is status post median sternotomy and intact median sternotomy with intact median sternotomy wires. Mediastinal contours are stable in appearance. Residual right pleural effusion with airspace opacity seen on the left is present, likely atelectasis. Atelectasis is seen at the left lung base. No pneumothorax. If there is a small left pleural effusion.

Table 5.3: RadTex/C+M captioning on a test-set CXR, comparing radiologist-written Ground Truth and *Unprompted* Report. Agreement of generated report with ground truth is highlighted. Key: GT Agreement GT Disagreement Irrelevant Info

Beyond quantitative evaluations, Table 5.3 presents an example report generated by RadTex. The unprompted report successfully identified a catheter positioned in the mid superior vena cava (svc) and a sternotomy. It also accurately detected pleural effusion on both sides and noted atelectasis while confirming the absence of pneumothorax. However, it incorrectly determined the side of the catheter and inaccurately represented the cardiac silhouette. We shade the generated words based on their agreement with the Ground Truth.

5.7 Conclusion

In this chapter, we further improved upon multi-modal contrastive learning approaches by introducing a bidirectional captioning-based pre-training strategy for interpretable medical image analysis. RadTex not only yields competitive performance against contrastive learning methods, but also exhibits label efficiency and inherent interpretability in the form of radiology report generation. Additionally, we introduce a flexible prompting mechanism for pathologies of interest, demonstrating its potential in improving clinical efficacy.

Traditional contrastive learning approaches primarily capture high-level global semantics of images and text, missing fine-grained understanding needed for diagnostic tasks [296]. RadTex uses an autoregressive captioning objective which encourages cross attention between tokens in the report and the image representation, which could more effectively utilize the dense semantics in the image and text data necessary for effective diagnostics. This could be observed from RadTex's superior performance on COVIDx and EdemaSeverity tasks. The COVID-19 virus and its symptoms were identified three years after MIMIC-CXR was released [138] (the pre-training data for RadTex), and it is possible that the high performance of RadTex on COVIDx dataset can be attributed to its superior encoding of radiographic findings irrespective of the disease type. On the EdemaSeverity task, radiologists consider the differentiation between vascular congestion (level 1) and interstitial edema (level 2) as more challenging compared with the differentiation between interstitial (level 2) and alveolar edema (level 3). We believe RadTex's strong differentiation of the former further strengthens the hypothesis of our approach being beneficial towards capturing the fine-grained semantics of radiographic findings in the images.

Our work offers a step in improvement of Medical Vision-Language Modeling (MVLM). In Chapter 4, we introduced the first technique in the research community to jointly learn image and text representations using contrastive learning, which offered encoding high-level semantic similarities in the modalities. This chapter further improved upon contrastive learning techniques through improved encoding of radiographic findings, offering label efficiency and interpretability as advantages. We also showcased how smaller, more specialized models like ours can offer improved performance compared to more generalist models created for zero-shot classification.

Future work can be inspired by the comparison of our domain-specific model to a larger general domain one such as MedPaLM M [289], and move towards model sparsity in larger general domain models [98]. Retrieval augmented generation (RAG) [161] is another aspect of research that can benefit the more accurate generation of radiology reports and mitigate hallucination issues. Another key finding of our work was around the necessity of significant amount of image-text paired data (100k+) to train multi-modal approaches. Data is currency for machine learning model development; and there needs to be a careful balance between regulation, accurate model training, data privacy and clinical utility for the adoption of MVLM approaches [102].

Part III

Mitigating privacy leakage

Chapter 6

Differentially Private pre-training of language model-like approaches

Abstract

This thesis describes methods that leverage structure towards improved clinical decisions and artificial intelligence (AI) safety in the form of data privacy. While data becomes currency in machine learning (ML), compliance issues with respect to accessing sensitive data become paramount. This chapter is based on the work [42], focusing on differentially private (DP) pre-training of large models.

As we progressed through the structured prediction approaches in this thesis, we explored themes of robustness and low data availability for inspiring our modeling decisions. The scaling laws of large language models (LLMs) [148, 114] have shown the importance of having larger datasets along with increasing model sizes to maintain strong performance over a variety of tasks. While the necessity of training on larger datasets increases in ML, so do the regulation and compliance issues needed to access these datasets. With the interest in public deployment of large pre-trained models, there is a rising concern for unintended memorization and leakage of sensitive data points from the training data.

In this chapter, we apply DP pre-training to a large state-of-the art Conformer-based encoder, and study its performance on a downstream automatic speech recognition (ASR) task assuming the fine-tuning data is public. This work is the first to apply DP to self-supervised learning (SSL) for ASR, investigating the DP noise tolerance and introduces a novel variant of model pruning called *gradient-based layer freezing*, that provides strong improvements in privacy-utility-compute trade-offs. The training set-up we introduce is agnostic to the downstream task, and has important applications towards the safe and effective deployment of large ML models in sensitive spaces such as healthcare. We also discuss the challenges of enforcing DP in the ML training pipeline, and suggest future directions which could enhance deployment efforts in this space to protect data privacy and safety.

6.1 Introduction

One of the key findings in Chapter 5 was around the necessity of large amounts of image-text paired data for training large medical vision language models (MVLMs). Similar trends around the data hungry nature of large language models (LLMs) have been observed as the model size is increased to improve performance on downstream tasks [148, 114]. Popular models are pre-trained on millions or billions of tokens, such as BERT [64] over 3500 million tokens and GPT-3 [27] over approximately 500 billion tokens. A popular graphic showcasing the exponential growth of model parameter size over the years in seen in 6-1. It is known from the scaling laws papers [148, 114] that efficient training of such models requires equivalent scaling of dataset size.

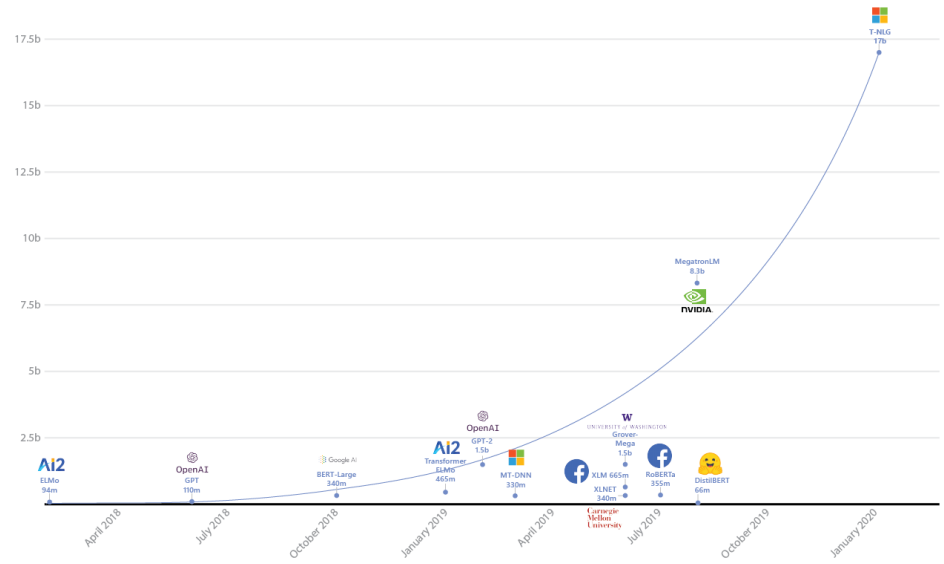


Figure 6-1: Growth of model parameters as time has passed. Latest state of the art models are now in the trillion parameter range. Figure borrowed from [258] and based on [264].

Popular machine learning (ML) models are often released as modifiable checkpoints after being pre-trained on vast amount of crawled data. However, it is well-known that ML models can leak sensitive information about their training dataset, even when the data is kept private. This has been extensively discussed by works such as [270, 32, 33, 35] in natural language processing (NLP) and computer vision (CV) and later extended to the speech domain by [7, 125, 298, 126]. Further, Zhang et al. [325] studied the influence of rare memorized examples from the training data over the model predictions and Carlini et al. [34] showed that larger models tend to have stronger memorization over the training data.

The memorization of training examples presents high risks for sensitive domains such as healthcare, and breaks compliance with existing laws such as Health Insurance Portability and Accountability Act of 1996 (HIPAA) [56] in the United States, General Data Protection Regulation (GDPR) [75] in European Union and the California

Consumer Privacy Act (CCPA) [55].

So far in the thesis, we have focused more on questions surrounding the modeling assumptions for medical data that is heterogeneous, noisy and consisting of few labels. An important aspect we have overlooked so far is around the sensitive nature of the data, after accepting that the de-identification procedure will remove all sensitive attributes associated with patients. Popular data scrubbing techniques still leave data vulnerable to linking attacks, as observed by Narayanan and Shmatikov [210] over de-identified Netflix movie ratings. Boag et al. [25] highlight different scenarios for huge number of privacy risks in the healthcare space.

Differential Privacy (DP) [70] is one way to combat the privacy leakage issue, by providing theoretical guarantees about the limits of influence of any individual training point towards the final model, thereby preventing an attacker from confidently inferring whether any particular sample was used for training. This technique provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks such as differencing, linkage, reconstruction and membership inference attacks [71].

There are several stages of the model lifecycle at which DP can be applied [235], and this chapter focuses on applying DP during the model training stage. In this set-up, the training data is kept private, the model is trained using a noise additive technique such as DP-SGD [1] and the model can be released publicly along with its parameter weights. Due to the post-processing property of DP, any modifications to the released model hold the same theoretical guarantees over the training data. In the healthcare space, this will be highly beneficial to deployment efforts where the data must be kept private, and the model is allowed to be released with strong privacy guarantees as seen in Figure 6-2.

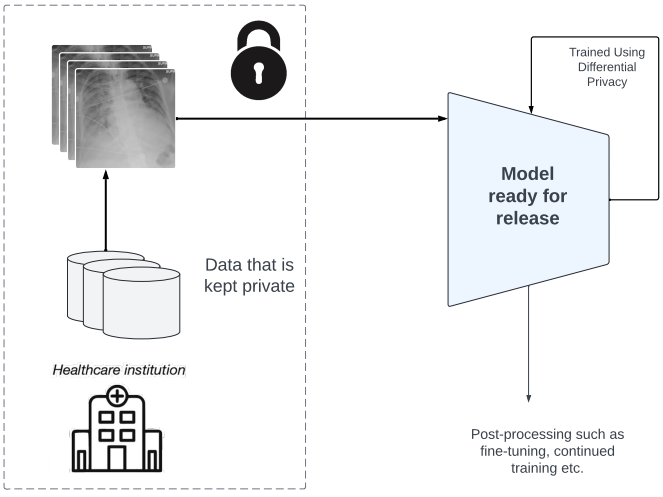


Figure 6-2: Training data is kept private and the model is ready to be publicly released after differentially private training. Any post-processing to the model maintains the same theoretical guarantees over the original data. Some reference elements borrowed from freepik.com and [275].

In this chapter, we focus on the differentially private pre-training regime and consider the speech domain in our study. The methods we introduce are universally applicable to multiple domains, and also have important applications in the healthcare space. Towards the end of the chapter, we comment on future directions applicable to healthcare, and also highlight DP as a universal technique to maintain compliance to important data privacy laws across the globe, making it an important component of artificial intelligence (AI) safety approaches.

6.1.1 Our Contributions

This chapter explores methods focused on DP pre-training of automatic speech recognition (ASR) encoders for mitigating the privacy leakage from trained encoders. The contribution of our work is as follows:

1. We are the first to evaluate the DP noise tolerance in the semi-supervised learning (SSL) setting of a large automatic speech recognition (ASR) encoder model.

2. We introduce a new variant of model pruning called *gradient-based layer freezing* where we determine the model layers to freeze from a square of gradient analysis.

3. Collectively, our recommended approach yields strong improvements in utility (Word Error Rate, i.e., WER in our case), while guaranteeing strong privacy of $\epsilon = 10$; compared to our baseline of applying DP with the same ϵ without our modeling improvements. Our approach yields a LibriSpeech test-clean/other WER (%) of 3.78/8.41 with (10, 1e-9)-DP for extrapolation towards low dataset scales, and 2.81/5.89 with (10, 7.9e-11)-DP for extrapolation towards high scales.

6.2 Differential Privacy

Differential Privacy (DP) [70] is widely considered a gold standard for bounding and quantifying the privacy leakage of sensitive data when performing learning tasks. Intuitively, DP prevents an adversary from confidently making any conclusions about whether any particular data was used in training a model, even while having access to the model and arbitrary external side information. The formal definition of DP depends on the notion of neighboring datasets: we will refer to a pair of datasets $D, D' \in \mathcal{D}$ as neighbors if D' can be obtained from D by adding or removing one data sample.

Definition 1 ((ϵ, δ)-DP) *A (randomized) algorithm $\mathcal{A} : \mathcal{D} \to \Theta$ is (ϵ, δ)-differentially private if for all pairs of neighboring datasets $D, D' \in \mathcal{D}$, and for any $S \subseteq \Theta$ we have,*

$$P[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \cdot P[\mathcal{A}(D') \in S] + \delta. \tag{6.1}$$

6.2.1 Practical considerations for DP

Typical recommendations for ϵ and δ are to be as small as possible, as ϵ is the multiplicative factor between the probabilities of the two neighboring datasets and δ is the additive scalar which controls the strength of the relaxation from the stricter ϵ-DP definition [70]. The general recommendation in the literature is to choose $\delta \ll \frac{1}{n}$ where n is the number of records in the dataset [71]. Ponomareva et al. [235] recommend different tiers for ϵ values going from strong formal guarantees to reasonable and weak guarantees, where Tier1 $:= \epsilon \leq 1$, Tier2 $:= \epsilon \leq 10$ and Tier 3 $:= \epsilon > 10$.

Differential Privacy also satisfies 2 important properties known as composition and invariance to post-processing [235, 71]. Sequential composition deals with the scenario of applying multiple DP mechanisms to the same dataset, allowing for accumulation of the ϵ and δ of each mechanism. Parallel composition assumes that the dataset is partitioned into mutually disjoint subsets upon which each mechanism is applied and the combined mechanism holds a maximum of each of the considered ϵ and δ. Finally, the invariance to post-processing property guarantees that any data-idependent transformation to a DP mechanism is guaranteed to remain differentially private with the same privacy guarantees. The sequential composition and invarariance to post-processing properties are pivotal to differentially private training of neural networks, from the use of differentially private gradients in each iteration, updating the model based on DP gradients and continuing this process for a fixed number of training steps [235].

6.2.2 Training models with DP

Typically, differentially private training is performed using variants of Differentially Private Stochastic Gradient Descent (DP-SGD) [1], where the main distinctions from non-private training are the clipping of per-example gradients and the addition of spherical Gaussian noise, as illustrated (for our ASR pre-training scenario) by Figure 6-3. Note that the magnitude of Gaussian noise (called noise multiplier) is directly correlated with the value of ϵ, calculated using the chosen privacy accounting technique such as the one by Abadi et al. [1]. This is implemented as a modification to the gradient computation during the optimization step by computing per-example gradients [278], clipping to limit their per-sample sensitivity, and the addition of calibrated Gaussian noise. Therefore, DP training is relatively independent of the exact choice of optimizer. For our experiments, we rely on the Adam optimizer with DP modifications for example-level DP. Training with DP incurs several challenges as a result of clipping and addition of noise, commonly characterized as privacy-utility-compute trade-offs (truncated as trade-offs in this chapter).

One challenge associated with DP training is the stringent trade-offs for training large models. A straightforward technique to increase privacy is adding more noise, but that negatively affects the model performance or utility. One method to mitigate this is to increase batch size [145, 228], but this comes at the cost of increasing compute which can be expensive. Recently, works in language modeling and vision have demonstrated the utility of their DP methods being close to their non-private

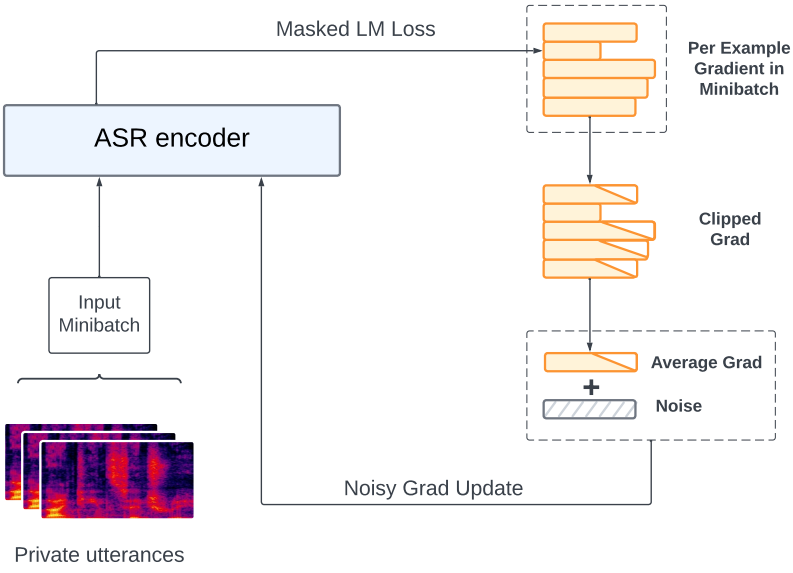


Figure 6-3: The Differentially Private pre-training method for ASR encoder involving clipping per-example gradients from the minibatch, and addition of calibrated gaussian noise. Gradients with norms below clip value are not clipped, as shown above. Once private pre-training of the ASR encoder is done, fine-tuning is done publicly after attaching an ASR decoder and using CTC loss [99, 96]

baselines [318, 28]. Most works are focused on improving trade-offs for the fine-tuning regime, with positive effects seen for parameter efficient techniques such as LoRA [118] that mitigate the issue of growing magnitude of DP noise as the model size increases [14, 318].

This chapter is focused on DP during the pre-training regime (see Figure 6-3), where the challenges associated with adding larger DP noise as a result of large model sizes for full model training remain. Recently, a work in language modeling has successfully reduced the gap between private DP pre-training and the non-private baseline through the use of private tokenization and higher compute [234]. More recently, Pelikan et al. [228] discuss improving trade-offs in the setting of DP Federated Learning (FL) in ASR by using per-layer clipping [198].

6.3 Related Works

Many works [14, 145, 167] have shown that the trade-offs are substantial for training large neural networks with state-of-the-art techniques like DP-SGD [14, 1]. Consequently, there has been work [1, 9, 152, 59] on pre-training using public data for improving the utility of DP-SGD. A recent work [228] has considered DP training for ASR models, but focusing on the Federated Learning (FL) regime. Additionally, many works [166, 318, 28] have focused on privately fine-tuning neural networks (focusing largely on vision and language models (LMs)) after pre-training using public data to improve the trade-offs for DP-SGD. While it is common in literature to treat

pre-training data as public, modern large model pre-training can involve sensitive data that is susceptible to be memorized and potentially leaked. There is only one recent work [234] that studies DP pre-training for LMs, and demonstrates that such models can be fine-tuned to high accuracies on downstream tasks. Related to modifications on bounding sensitivity within a training step, [303] have considered the role of gradient clipping and suggest model pruning as a strategy to improve the trade-offs.

6.4 Datasets

LibriLight (LL) Containing over 60k hours of audio derived from audiobooks, LL [144] is one of the largest freely-available corpora of speech. We follow the pre-training scheme defined in the BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ) paper [48], and utilize the "unlab-60k" subset of the dataset which doesn't contain alignment to the language transcriptions. Only small versions of the dataset (around 10 hours) are aligned with transcriptions using a phonemizer[1] so do not count as gold standard.

LibriSpeech (LS) Another dataset of 960 hours of transcribed audio, [222] consists of several rounds of careful transcription to obtain supervision and is a gold standard dataset used in various supervised automatic speech recognition (ASR) tasks. For evaluation, results are reported on both the test-clean and test-other data partitions, which refer to audiobook recordings without background noise (the ideal scenario) and those with background noise respectively. Fine-tuning using this dataset also follows the procedure for BEST-RQ [48].

6.5 Methods

6.5.1 BEST-RQ pre-training method and the ASR task

BERT-based Speech pre-Training with Random-projection Quantizer (BEST-RQ) [48] is a self-supervised learning (SSL) method that introduces the pre-training and fine-tuning paradigm to the automatic speech recognition (ASR) task. During the pre-training stage, the model learns to predict masked speech signals represented as discrete tokens with a random-projection quantizer. Pre-training involves masking speech signals and feeding them to the ASR encoder, which learns to predict the masked region using a masked language modeling (MLM) style loss similar to the BERT model [64]. Fine-tuning involves attaching an ASR decoder based on an LSTM transducer [95], supervised training using the aligned transcript and optimization using the CTC loss [96].

While BEST-RQ defines a pre-training and fine-tuning paradigm for the ASR task, the architecture is based on the Conformer [99] model. This defines a convolution-augmented transformer model whose main building blocks include multi-head self-

[1] https://gitlab.coml.lscp.ens.fr/mbernard/phonemizer

attention and convolution modules stacked across many layers (24 in our case). The Conformer is considered state-of-the-art for ASR tasks, and forms the basis for major improvements in this direction.

6.5.2 Our Experimental Set-up

For our model, we choose the 300M variant [298] of the state-of-the art ASR model architecture, Conformer XL [330]. The encoder is pre-trained on LibriLight (LL) [144] for 1M steps using self-supervised learning via BEST-RQ [48]. Fine-tuning is done for 60k steps post attaching an additional projection layer on the encoder, using the LibriSpeech (LS) [222] dataset. Hyperparameter details and model architecture follow the BEST-RQ paper [48], and official dataset splits were used for training, validation and hyperparameter tuning. Pre-training takes $\sim$ 1 week on Dragonfish TPUs with 8x8 topology, fine-tuning takes $\leq$ 1 day and original batch size was set at 512.

In this paper, we apply DP to the pre-training stage of our model (with LL), and assume that the fine-tuning dataset (LS) for the downstream ASR task is public. Utility is reported as test-clean/other WER on the LS dataset. We use the updated moments accountant [1, 203] for calculating our privacy guarantees. We report experiments with different DP noise multipliers in the range [1e-4, 1e-2], since we find that noise multipliers beyond 1e-2 lead to divergence (more details in Section 6.6.1).

Since the trade-offs with large model training can be significant, we follow the extrapolation strategy similar to recent works [145, 228]. We extrapolate the (ϵ, δ)-DP assuming the training dynamic remains unchanged upon linearly scaling minibatch size and noise multiplier (to maintain the expected signal-to-noise ratio for the gradient update) along with scaling the dataset size (for improved privacy accounting). Thus, we experiment with adding different DP noise multipliers and map the value of ϵ using the moments accountant, while the batch size, noise multiplier, and the pre-training dataset size are scaled up by the same factor as seen in Figure 6-4. We can see the positive effects of the scale-up factor on ϵ, thus significantly improving the privacy guarantees. In table 6.1, we report the precise scale-up factors for noise multipliers we consider in this paper to achieve a DP of $\epsilon = 10$ at $\delta = n^{-1.1}$, where n is the scaled-up dataset size. Note that according to recent work [235], such a level of DP can be classified in the "Tier 2: Reasonable privacy guarantees".

Table 6.1: Extrapolation factor for linearly scaling-up noise multiplier, batch size and dataset size needed for each used noise multiplier value to get DP $\epsilon = 10$ at $\delta = n^{-1.1}$, where n is the scaled-up dataset size.

Noise multiplier	1e-4	5e-4	1e-3	5e-3	1e-2
Scale-up	5450	1070	530	105	52

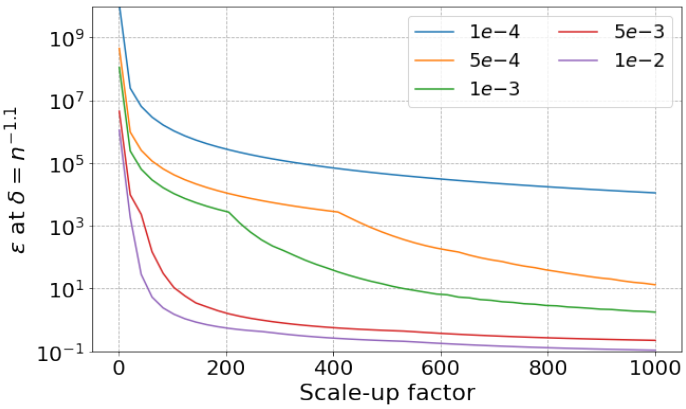


Figure 6-4: Extrapolating the noise multiplier linearly with batch size and dataset size to maintain the signal-to-noise ratio and improve privacy accounting.

6.6 Results

In Section 6.6.1, we introduce the baseline of (non-privately) pre-training the BEST-RQ 300M model. We detail preliminary modeling changes required to comply with DP training and include an analysis of the amount of DP noise tolerable for minimal performance regression of the model. Next, in Section 6.6.2 following [228], we incorporate per layer clipping for improved utility and noise tolerance. Lastly, we introduce our gradient-based layer freezing strategy. Our results denote a synergy between per-layer clipping and our model pruning technique, based on the compounding improvements we observe in model quality (summary of results in Table 6.3).

6.6.1 Noise tolerance of the BEST-RQ model

We establish non-private baselines for the BEST-RQ 300M model, and analyze the degree of noise tolerated for minimal utility regression. As is typical for DP training, we replace batch normalization with group normalization to effectively limit per-sample contributions and avoid mixing of batch statistics across samples [235]. After experimentation, we find the best setting of group normalization to have input rank of 3, number of groups as 1 and group norm epsilon as 1e-4, resulting in a test-clean/other WER (%) of 2.17/4.23 post fine-tuning on LibriSpeech.

Then, we experiment with choices for per-example clipping bounds, and find the bound 1.5 to be clipping almost all samples during training while providing minimal loss in performance, resulting in a WER of 2.21/4.29. We refer to this as the **non-private lower bound** result. Thus, the non-private baseline we report for BEST-RQ consists of group normalization and per-example clipping, to offer a direct comparison to the level of additive noise in our experiments. Our results for the non-private baseline, and for differing level of DP noise are reported in Table 6.2.

Note that the performance of the model with fine-tuning from random initialization (no pre-training) is a WER of 4.43/ 11.23, which is the upper bound for effectively measuring the positive effects of pre-training. We refer to this result as the **no pre-**

Table 6.2: Noise tolerance of the BEST-RQ 300M model. Our no pre-train upper bound is WER of 4.43/11.23. Above noise multiplier 1e-2, the model diverges into WER of 100.

Noise multiplier	test-clean/other WER
0	2.21/4.29
1e-4	2.24/4.51
5e-4	2.57/5.98
1e-3	3.54/8.31
5e-3	10.98/22.62
1e-2	15.38/29.62

train upper bound, which is effectively the same as not applying BEST-RQ style pre-training to ConformerXL and just doing supervised training on the Librispeech dataset. As can be seen from Table 6.2, we start seeing significant regressions (greater than 10% relative) for noise multiplier 5e-4, where the standard extrapolation technique achieves DP $\epsilon = 10$ only at a practically prohibitive scale-up factor of 1070 (Table 6.1). For reference, extrapolation factor for DP $\epsilon = 10$ from noise multiplier 1e-2 is as low as 52, though with the current approach we get WER of 15.38/29.62 which is higher than the no pre-train upper bound. Our focus in the rest of the paper is to improve trade-offs for the settings with larger noise multipliers in the range [5e-4, 1e-2].

6.6.2 Improving the noise tolerance

Recently, there have been increasing efforts towards improving trade-offs associated with DP training and we consider these efforts heavily in our experimentation. Ganesh et al. [83] show the importance of public pre-training for private model training, especially with an in-domain public checkpoint. Pelikan et al. [228] revive per-layer clipping and show improvements for DP in the supervised training setting of FL for ASR. A couple of recent works [192, 303] have shed light on the benefits of model pruning for DP training, by minimizing the negative effects of compounding noise affected by the model dimensionality. Thus, in order to bridge the utility gap with the non-private pre-trained baseline, we consider the following three improvements: warm-starting (WS) using public data, per-layer clipping, and our novel method of gradient-based layer freezing. Table 6.3 summarizes the compounding improvements on the three considered techniques.

Table 6.3: Final noise tolerance WERs for BEST-RQ 300M model with our considered improvements. If we observe divergence (mainly for higher noise multipliers), we report results on fine-tuning with an early 200k step pre-trained checkpoint instead.

Noise	**Public WS**	**+PerLayerClip**	**+LayerFreeze**
5e-4	3.82/7.65	2.78/5.9	**2.67/5.74**
1e-3	4.03/8.62	2.85/6.02	**2.81/5.89**
5e-3	6.34/13.88	3.78/8.09	**3.19/7.17**
1e-2	8.16/17.42	100/100	**3.78/8.41**

Warm-starting using in-domain public data (Public WS)

For our experiments, in line with prior works [6, 83] on using in-domain public data for warmstarting DP training, we treat a random 1% partition of the LibriLight (LL) train dataset as a substitute for a small amount of in-domain public data being available. Further, for improved trade-offs, we conduct the DP pre-training on the entire LL train dataset (i.e., samples in the 1% public partition are incorporated into the private training dataset, providing a marginal improvement in the privacy accounting). Fine-tuning with LibriSpeech (LS) after only (non-private) pre-training with 1% LL yields a WER of 3.88/8.94. Note that this is better than our no pre-train upper bound of 4.43/11.23, but still substantially worse than the non-private lower bound of 2.21/4.29, validating the assumption about only a small amount on in-distribution public data being available in practical scenarios.

We present the results with public warmstart in the second column in Table 6.3, and compared to the random initialization results in Table 6.2, we observe a slight regression for smaller noise multipliers $\{5\text{e-}4, 1\text{e-}3\}$, whereas a significant improvement for the higher noise multipliers $\{5\text{e-}3, 1\text{e-}2\}$.

Per-layer clipping

There are two commonly-used variants of per-layer clipping [198, 228], denoted by the *uniform* variant (which splits the clipping bound equally amongst all layers), and the *dim* variant (which splits the clipping bound proportional to each layer's dimension). We conducted experiments using both the variants, and but found the *dim* variant to be outperforming the uniform one (similar to results seen in [228]).

We present the results for adding per-layer clipping for DP pre-training, post public warmstarting, in the third column in Table 6.3. While we observed the model diverging for the highest noise multiplier of 1e-2, we notice significant improvements in model quality for all other considered values of noise multiplier, corroborating the observation in [228] regarding the usefulness of per-layer clipping in the ASR domain.

Gradient-based layer freezing (LayerFreeze)

For reducing the dimensionality of DP training, some recent works [192, 303] propose starting from a pruned model that is initialized from a publicly pre-trained checkpoint. In this work, we devise a novel one-shot variant of model pruning called **Gradient-based Layer Freezing** (Algorithm 1), where instead of removing or freezing individual parameters based on their magnitudes, we freeze them layer-wise based on the normalized squared $\boldsymbol{\ell}_2$ norm of their gradients observed throughout the public warmstarting phase. After this operation, we continue DP pre-training with the pruned model and entire LL dataset.

Algorithm 1 Gradient-based Layer Freezing (LayerFreeze)

Input: *Model F with params $\theta \in \mathbb{R}^M$, num layers $\{i\}_{i=1}^{L}$, Loss fn $\mathcal{L}(\theta)$ over minibatch, Num iterations T, Optimizer* `opt`*, Grad* `update()` *fn, total params per layer* `dim` *() fn, top params $p\%$,* `freeze_top_layers` *whether to freeze top layers or the rest*

$\mathbf{u_o} \leftarrow 0$ ▷ Init sq grad vector with M dim
for $t \in [T]$ **do**
  $\mathbf{g_t} \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t)$
  $\mathbf{u_t} \leftarrow \mathbf{u_{t-1}} + \mathbf{g_t^2}$ ▷ Accumulate sq grad
  $\theta_t \leftarrow \texttt{update}(\texttt{opt}, \mathbf{g_t}, \theta_{t-1})$
for $i \in [L]$ **do**
  $\mathbf{g_{layer_i}} \leftarrow \sum_{\text{param}\in i} \mathbf{u_T}[\texttt{param}]/\texttt{dim}(i)$
$top\_par \leftarrow p \cdot M$ ▷ Top num of params
$top\_layers \leftarrow [\,]$
for $layer_i$, _ in `sorted` (`enumerate` ($\mathbf{g_{layer}}$)) **do**
  if $\sum \texttt{dim}(top\_layers + [layer_i]) \leq top\_par$ **then**
    $top\_layers.append(layer_i)$
  else *break* ▷ Sort by grad & get layers till cut off $p\%$
if `freeze_top_layers` is `True` **then**
  $fr\_layers \leftarrow top\_layers$
else $fr\_layers \leftarrow \{i\}_{i=1}^{L} - top\_layers$
return Model F with frozen layers fr_layers

Once the norms of the per-layer gradients until our public warmstarting checkpoint are accumulated, we focus on $p\%$ of the model parameters, consisting of layers with the highest normalized accumulated squared gradient norm. We perform tuning experiments by freezing layers associated with either these $p\%$ parameter, or the remaining $1 - p\%$ parameters. P is treated as a hyperparameter, explored in the range $\{0.015\%, 10\%]$ as seen in Figure 6-5. We consistently find that DP pre-training benefits from freezing layers with the top $p\%$ parameters, where the best case is when $p = 1\%$. We report the results of using LayerFreeze, along with per-layer clipping and public warmstarting, in the fourth column in Table 6.3. It is important to note that LayerFreeze provides significant improvements in model quality in all the considered settings.

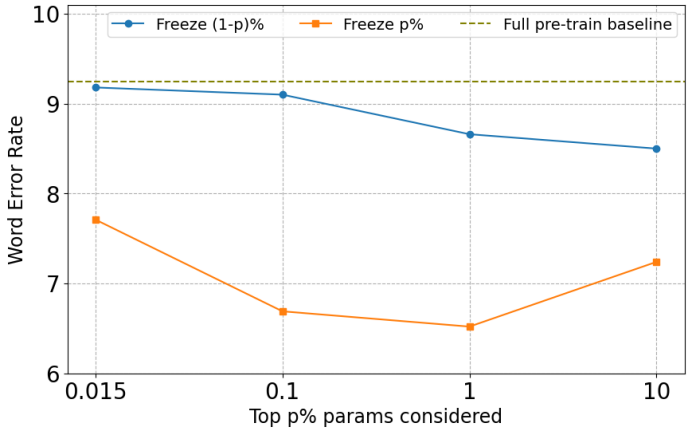


Figure 6-5: Performance from tuning our LayerFreeze with different percentage of frozen parameters, while keeping the DP noise multiplier constant at 1e-3. Along x-axis, we use p to refer to the % of parameters consisting of layers with the highest accumulated gradient norms. We run experiments with freezing either the p% parameters, or the remaining $(1-p)$%. Saving on compute, fine-tuning is done using an early pre-train checkpoint of 200k, assuming that the same conclusions hold for 1M.

In summary, we obtain LibriSpeech WERs of 3.78/8.41 with (10, 1e-9)-DP for LibriLight with an extrapolation factor of 52 (low dataset scaling regime), and 2.81/5.89 with (10, 7.9e-11)-DP for LibriLight with an extrapolation factor of 530 (high dataset scaling regime).

6.7 Conclusion

We introduce DP to SSL for ASR, and a novel variant of model pruning called gradient-based layer freezing. Our technique improves the trade-offs for DP ASR pre-training, over improvements from public warmstarting and per-layer clipping. Overall, we demonstrate a DP training method that improves utility significantly while maintaining robust privacy guarantees under various extrapolation factors. Though our work provides a way to pre-train ASR encoders with strong DP guarantees, the extrapolations required to reach those guarantees can be limiting in some practical regimes. Improving computation trade-offs that we incur for reaching strong DP guarantees is an interesting direction we leave for future investigation.

DP offers robust privacy guarantees, but has significant challenges in deployment due to the stark privacy-utility-compute tradeoffs. Existing works have successfully been able to reduce these tradeoffs for fine-tuning by assuming access to a large public dataset for pre-training, but our research closes the loop on private model creation by introducing DP to the pre-training stage. Our work, through the use of *gradient-based layer freezing* is able to significantly improve tradeoffs for DP pre-training, but further work is necessary to close the gap between non-private and private pre-training. Interesting directions to be explored in this space include thorough comparisons to model pruning techniques such as iterative magnitude pruning (IMP) [80]. Improve-

ments to gradient clipping techniques such as adaptive per-layer clipping [109] can potentially reduce tradeoffs.

Recent works have opened up a new area of exploration in the space of DP for Large Language Models (LLMs). Researchers have analyzed specific structured data leakage from models such as personally identifiable information (PII) [189], and developed differentially private in-context learning approaches [285]. With the rapid progress in development of DP approaches for neural networks, several open questions have emerged along the lines of memorization, leakage and mitigation. However, we are still far from an easy to communicate language of privacy leakage that can be understood beyond the circle of privacy experts [25].

Future work in DP should focus on easy to interpret measures for privacy leakage beyond only ϵ and δ values, and quantify leakage in a structural fashion (e.g. based on the type of information leaked) as opposed to only a record-level or group-level measure. In the healthcare space, this would further reduce the gap between compliance experts and privacy researchers. An added layer of complexity is introduced by DP for ML, on top of the existing communication gap between clinicians and ML experts on which technologies must be prioritized for clinical deployment.

Chapter 7

Conclusion and Discussion

In this thesis, we explore themes related to leveraging structure for improving clinical decisions and artificial intelligence (AI) safety via differentially private (DP) model training. We look at clinically aggregated representations towards model robustness across changing hospital systems, clinical concept extraction to improve relation extraction, leveraging paired radiology reports associated with chest radiographs for predictive pathology prediction capacity & report generation, and gradient-based parameter selection for improved privacy-utility-compute tradeoffs in differentially private model development.

In Chapter 2, we touch on key ideas of model generalizability and reproducibility, to support research and deployment efforts in machine learning (ML) for health. We introduce clinically aggregated feature representations to support ML models' ability to sustain performance over time as care practices, hospital database systems and population demographics evolve. We provide our feature aggregation, along with a full pre-processing pipeline for a popular electronic health record (EHR) dataset i.e., MIMIC-III [136] into an open-source framework known as MIMIC-Extract. MIMIC-Extract, since its creation, has served as the basis for several ML studies and supported reproducibility efforts in ML for health.

Chapter 3 further delves into challenges for reproducibility in a popular biomedical natural language processing (NLP) task known as relation extraction. We showcased the extent of the reproducibility challenge and found that in addition to many experiments not being described precisely enough, ablation studies that clearly delineate performance contributions of individual techniques were often missing. This lack of consensus over effective techniques made it harder for the field to determine ideas that generalized to novel tasks, datasets and contexts. Our main contributions in this domain are of the introduction of a highly accurate, unifying and extendable framework for the relation extraction task that mitigates the reproducibility challenge in the domain. One of the surprising findings of our work is related to choices of pre-processing using external knowledge in the form of clinical concept extraction as a large contributor to performance, even more than choices of modeling.

Chapter 4 and 5 introduce new frameworks for accurate classification for medical vision-based disease severity tasks. Chapter 4 first introduces the idea of leveraging free-text radiology reports associated with chest radiographs to improve on the task

of pulmonary edema severity detection. By taking advantage of the rich information present in the radiology reports, our framework is the first to use contrastive learning for representation learning towards highly accurate disease severity prediction. Chapter 5 further builds upon this idea by introducing a new framework based on multi-modal autoregressive captioning to learn efficient visual representations that require fewer labels for training compared with image encoders learned via supervised pre-training and contrastive learning objectives. Our framework also enables interpretable outputs through the ability to generate accurate and prompt-based radiology reports at inference.

Finally, Chapter 6 discusses differentially private (DP) model training as an important component of artificial intelligence (AI) safety. We apply DP pre-training to a large state-of-the-art model in the speech domain that is optimized using a language modeling-style objective. Ours is the first method to apply DP to self-supervised learning (SSL) for automatic speech recognition (ASR) and to introduce a novel model pruning variant based on gradient structure. We highlight DP training as an important technique in preventing data leakage from the training data, and underscore its importance towards deployability efforts in ML for health as the field moves towards adopting larger and more data hungry models.

This thesis flows from themes of adoptability, reproducibility and robustness questions towards multi-modality, label efficiency and interpretability in machine learning for health. Finally, as the field moves towards adopting models requiring training of larger datasets containing potentially sensitive attributes, we discuss techniques that prevent privacy leakage from training data with provable theoretical guarantees.

7.1 Future Work

Multi-modality As this thesis also demonstrates in its flow, there is a movement within the ML for health field towards clinical decision support by using multi-modal information. Already, we are moving beyond supervised learning algorithms towards those using self-supervised learning. As seen in Chapters 4 and 5, the usage of multi-view data in the form of radiology reports and chest x-ray images improves accuracy as well as label-efficiency of models. Recent efforts such as Med-PALM and Med-PaLM M [271, 289] have demonstrated zero-shot capabilities of large language models (LLMs) and vision-language models (VLMs) pre-trained using large amounts of general domain data and instruction fine-tuned on medical data.

Notably, Med-Gemini [260] incorporates the popular mixture-of-experts (MoE) architecture [76, 268] to efficiently scale and reason over longer contexts at inference time. While large zero-shot reasoning models are becoming popular, Med-Gemini aims to introduce a family of models that are optimized for individual application-specific scenarios such as medical question-answering, summarization, multi-modal visual question answering and video question-answering. Following the advances seen in general domain LLMs and multi-modal vision-language models (VLMs), I see the ML for health field moving from models that operate over single modality towards multiple modalities (including video understanding). By design, LLM-style models

are interactive and use prompting along with their long context windows to capture a more natural flow of conversation. This will have major advantages for the clinical space, by allowing doctors to interact with the models in natural language, allow for self-correction via prompting and develop better trust through multi-turn conversation. These models can also serve as educational tools, supporting early-career medical professionals in being able to ask questions about past cases and test their knowledge interactively.

Model sparsity Recent research in LLMs has demonstrated the benefits of model sparsity and pruning towards specialized task performance [98, 80, 63]. This research is essential towards ensuring the cost of inference can be reduced and making it more feasible to deploy these models in real-world settings. Additionally, this research helps in understanding how a network uses its parameters [63], notably observing that larger models might not be efficiently leveraging the parameters in the deeper layers. Connecting to findings observed in Differentially Private (DP) training as seen in Chapter 6, model sparsity can help in reducing the effective DP noise applied to the model and improve model utility significantly. Between benefits observed from model sparsity in non-DP settings as well as DP settings, there is an opportunity to better understand the learning trajectory and loss landscapes of ML model training [303]. This has overarching benefits throughout the ML community in understanding the optimal model sparsity necessary for specialized task performance, which is a direction already starting to be pursued by [260] in the medical domain through the sparsely-gated MoE layer [268]. Sparsity will continue to be a key area to better control model capacity for optimal utility, and will help the field propose more computationally efficient models that don't sacrifice downstream task performance.

AI safety In Chapter 6, we discussed differential privacy (DP) as an important component of artificial intelligence (AI) safety. Although DP works by offering guarantees at the sentence level against membership inference and reconstruction attacks, the privacy protection this technique offers against leakage of personal identifiable information (PII) is unknown. Lukas et al. [189] address the issue of measuring structured privacy leakage in the form of PII leakage, and observe that despite DP offering strong theoretical guarantees, PII leakage is still not completely eliminated. More work in the DP space needs to address the issue of structured information leakage, along with hierarchical protection for more sensitive vs. less sensitive attributes.

In addition, recent work in the alignment domain of LLMs has demonstrated harmfulness by fine-tuning with small amounts of adversarial examples and degradation of safety through benign fine-tuning [239], indicating the brittleness of safety alignment methods [305]. Understanding how information is distributed throughout the parameters of a model is critical in moving the field towards mitigating such attacks. This also connects back to the line of research where the learning trajectory of LLMs is investigated further, and where safety fine-tuning approaches can be measured in conjunction for their level of brittleness. If we know parameters vital for safety guardrails in the model as suggested in [305], providers could allow only for

localized fine-tuning where the safety-critical model regions are left untouched.

A definition of safety that not only encompasses harmful actions by the model but also memorization and privacy metrics is essential if the field wants to develop truly harmless models. Similar to the privacy-utility trade-offs observed in the DP literature, the safety literature has observed a safety-utility trade-off [239], which must be measured and defined transparently. Parallel to how the DP literature considers privacy as a budget to be consumed while training a model, safety could be similarly defined using the notion of a budget. Significant work also remains to be done in safety theory to move beyond only reporting on empirical metrics over small number of examples (e.g., a curated dataset of 330 examples in [239]). Further research is also needed on safety tuning for LLMs in healthcare; Han et al. [101] take a step in this direction by highlighting the safety gaps in medical LLMs.

What's next ML for health has seen rapid progress in the last few years as a key applied area for ML, LLMs and multi-modal models. Research areas of AI safety, privacy, model robustness, computational and label efficiency, usage of external information, importance of data pre-processing and the availability of multi-modal data will remain paramount in the ML for health domain.

In this thesis, we have taken small steps in the directions of reproducibility, model robustness, clinical concept usage as external knowledge, label efficiency, multi-modality and differential privacy. Areas that continue to push the state-of-the-art in model performance will continue being relevant, such as the usage of multi-modal approaches to push forward performance benchmarks. In the future, I see the field moving beyond just optimizing performance benchmarks towards AI safety and privacy, which will play a crucial role in training models with large amounts of sensitive healthcare data. Theoretical and practical definitions of safety and privacy will be important for developing healthy collaborations with clinical & policy stakeholders.

Appendices

Appendix A

Clinically aggregated features and concept drift

In Chapter 2, we discuss raw feature extraction from the MIMIC-III dataset [136], which often includes missing values. Over time, changes in frequency of data collection (in Figure A-1) can lead to highly sparse feature representations. Additionally, changes in measurements (in Figure A-2) over time leads to accuracy loss for machine learning models developed using raw feature representations.

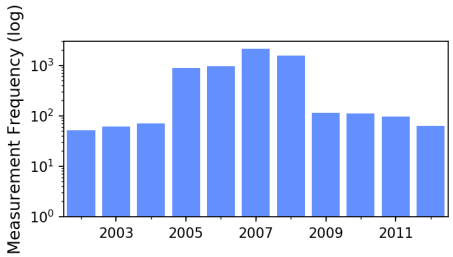


Figure A-1: The frequency of data collection can change in clinical practice. Shown is an example of the collection frequency for Mean Arterial Blood Pressure. Figure borrowed from [213].

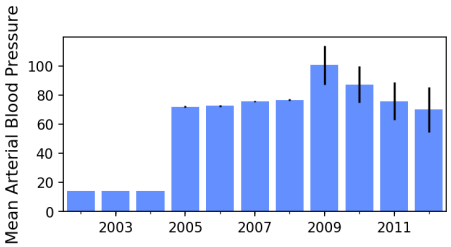


Figure A-2: The measured values of data can shift in clinical practice. Figure borrowed from [213].

Our solution for mitigation of this accuracy loss is our clinically aggregated feature representations, presented in Section A.1.

A.1 Clinically Aggregated Feature Set

Grouping	low	high	strict	avg	std	pres.	pres. cv	pres. mv	ItemID	Table	DB	avg	std	pres.
alanine aminotransferase	1.7E-04	1.9E-04	3.9E-04	282.3	916.4	2.0E-02	1.5E-02	2.5E-02	769	chartevents	cv	335.1	984.9	6.7E-03
									220644	chartevents	mv	281.1	907.0	6.7E-03
albumin	0.0E+00	1.9E-05	0.0E+00	3.1	0.7	1.3E-02	1.1E-02	1.5E-02	772	chartevents	cv	2.9	0.6	4.9E-03
									1521	chartevents	cv	3.0	0.6	3.8E-03
									227456	chartevents	mv	3.1	0.7	4.1E-03
alkaline phosphate	4.1E-04	1.6E-05	7.5E-05	122.3	143.6	1.9E-02	1.5E-02	2.4E-02	773	chartevents	cv	126.7	157.6	6.5E-03
									225612	chartevents	mv	120.1	146.5	6.5E-03
anion gap	1.7E-04	2.1E-05	1.4E-05	13.7	4.0	8.3E-02	6.4E-02	1.0E-01	227073	chartevents	mv	13.3	3.8	2.9E-02
asparate aminotransferase	2.1E-04	4.2E-05	4.7E-05	348.1	1239.6	2.0E-02	1.5E-02	2.5E-02	770	chartevents	cv	404.3	1299.0	6.7E-03
									220587	chartevents	mv	347.4	1239.9	6.7E-03
bicarbonate	0.0E+00	0.0E+00	6.1E-06	24.2	4.7	8.8E-02	6.8E-02	1.0E-01	227443	chartevents	mv	24.4	4.7	2.9E-02
bilirubin	1.2E-03	4.1E-05	2.8E-05	2.6	5.2	2.0E-02	1.5E-02	2.5E-02	225690	chartevents	mv	2.7	5.2	6.7E-03
									848	chartevents	cv	3.2	6.4	6.3E-03
									1538	chartevents	cv	3.3	6.4	5.0E-03
									225651	chartevents	mv	3.0	4.3	5.7E-04
									803	chartevents	cv	3.2	4.9	5.8E-04
blood urea nitrogen	0.0E+00	2.1E-05	0.0E+00	26.2	21.8	8.8E-02	6.9E-02	1.1E-01	781	chartevents	cv	26.6	22.1	3.5E-02
									1162	chartevents	cv	26.4	22.0	2.7E-02
									225624	chartevents	mv	25.9	21.4	3.0E-02
calcium	NAN	NAN	NAN	8.3	1.9	7.0E-02	5.1E-02	8.6E-02	786	chartevents	cv	8.3	0.8	2.7E-02
									1522	chartevents	cv	8.3	0.8	2.2E-02
									225625	chartevents	mv	8.3	2.9	2.5E-02
calcium ionized	NAN	NAN	NAN	1.3	5.1	5.1E-02	4.9E-02	4.7E-02	816	chartevents	cv	1.5	7.3	2.6E-02
									225667	chartevents	mv	1.1	0.9	1.5E-02
cardiac index	NAN	NAN	NAN	2.9	0.8	3.5E-02	5.8E-02	2.6E-04	116	chartevents	cv	2.9	0.8	3.5E-02
cardiac output fick	NAN	NAN	NAN	5.7	2.0	7.3E-03	1.2E-02	6.6E-05	89	chartevents	cv	5.7	2.0	7.3E-03
cardiac output thermodilution	NAN	NAN	NAN	5.7	1.9	3.0E-02	5.0E-02	2.1E-04	90	chartevents	cv	5.7	1.9	3.0E-02
central venous pressure	NAN	NAN	NAN	11.6	16.1	2.0E-01	2.2E-01	1.3E-01	220074	chartevents	mv	13.7	27.8	6.0E-02
									113	chartevents	cv	10.6	5.4	1.4E-01
chloride	1.1E-05	1.7E-06	0.0E+00	105.2	6.3	9.6E-02	7.3E-02	1.2E-01	220602	chartevents	mv	105.3	6.4	3.1E-02
									1523	chartevents	cv	105.8	6.2	2.7E-02
									226536	chartevents	mv	106.2	5.9	2.9E-03
									788	chartevents	cv	105.6	6.1	3.4E-02
cholesterol	4.2E-04	9.4E-05	1.9E-04	161.9	51.3	1.8E-03	1.6E-03	1.7E-03	1524	chartevents	cv	159.8	47.4	5.5E-04
									789	chartevents	cv	161.7	49.1	7.4E-04
									220603	chartevents	mv	160.1	52.0	4.9E-04

Grouping	low	high	strict	avg	std	pres.	pres. cv	pres. mv	ItemID	Table	DB	avg	std	pres.
co2	NAN	NAN	NAN	24.1	4.8	3.4E-02	5.6E-02	3.5E-05	787	chartevents	cv	24.1	4.8	3.4E-02
co2 (etco2, pco2, etc.)	NAN	NAN	NAN	25.2	5.3	8.2E-02	8.1E-02	7.5E-02	777	chartevents	cv	25.2	5.3	4.2E-02
									225698	chartevents	mv	25.2	5.2	2.4E-02
									857	chartevents	cv	25.8	8.2	4.6E-04
									223679	chartevents	mv	25.3	7.8	7.0E-04
creatinine	1.4E-05	0.0E+00	8.5E-05	1.4	1.5	8.9E-02	7.0E-02	1.1E-01	791	chartevents	cv	1.4	1.5	3.5E-02
									1525	chartevents	cv	1.4	1.5	2.7E-02
									220615	chartevents	mv	1.4	1.4	3.0E-02
diastolic blood pressure	0.0E+00	0.0E+00	1.3E-05	60.9	14.1	8.7E-01	8.2E-01	7.7E-01	224643	chartevents	mv	66.0	16.2	1.6E-04
									220180	chartevents	mv	63.6	15.0	2.5E-01
									8441	chartevents	cv	59.0	14.8	2.8E-01
									8555	chartevents	cv	57.4	12.6	2.7E-03
									220051	chartevents	mv	60.4	13.4	1.3E-01
									8368	chartevents	cv	60.2	13.5	2.5E-01
									8440	chartevents	cv	62.1	14.5	4.5E-04
									225310	chartevents	mv	59.3	14.4	1.1E-02
fibrinogen	NAN	NAN	NAN	295.6	175.3	9.4E-03	8.0E-03	1.1E-02	227468	chartevents	mv	288.7	177.5	2.9E-03
									806	chartevents	cv	300.6	178.5	3.3E-03
									1528	chartevents	cv	298.5	177.5	2.9E-03
fraction inspired oxygen	4.6E-05	0.0E+00	2.1E-03	0.5	0.2	4.5E-02	3.0E-03	9.1E-02	189	chartevents	cv	0.6	0.2	1.7E-03
									223835	chartevents	mv	0.5	0.2	4.3E-02
									727	chartevents	cv	0.4	0.2	7.2E-05
fraction inspired oxygen set	NAN	NAN	NAN	0.5	0.2	6.5E-02	1.1E-01	5.2E-04	190	chartevents	cv	0.5	0.2	6.5E-02
glascow coma scale total	NAN	NAN	NAN	12.5	3.6	1.7E-01	2.8E-01	1.5E-03	198	chartevents	cv	12.5	3.6	1.7E-01
glucose	2.0E-04	0.0E+00	3.1E-06	140.5	57.2	2.3E-01	2.2E-01	2.2E-01	220621	chartevents	mv	136.2	61.5	2.9E-02
									226537	chartevents	mv	131.8	42.0	1.6E-02
									1529	chartevents	cv	132.8	52.1	5.1E-02
									807	chartevents	cv	144.5	57.8	6.2E-02
									811	chartevents	cv	135.8	53.8	6.8E-02
									225664	chartevents	mv	150.2	60.7	4.6E-02
heart rate	0.0E+00	0.0E+00	6.7E-07	85.0	17.3	9.0E-01	8.4E-01	8.1E-01	211	chartevents	cv	85.1	17.1	5.2E-01
									220045	chartevents	mv	84.8	17.4	3.8E-01
height	0.0E+00	8.8E-05	1.8E-04	168.8	13.8	3.5E-03	2.1E-05	7.3E-03	1394	chartevents	cv	167.6	18.0	9.1E-07
									226707	chartevents	mv	168.8	13.8	3.4E-03
									226730	chartevents	mv	168.8	13.9	3.4E-03
hematocrit	0.0E+00	0.0E+00	3.8E-06	31.0	5.4	1.2E-01	1.0E-01	1.3E-01	813	chartevents	cv	30.7	4.9	5.1E-02
									220545	chartevents	mv	30.6	5.2	3.5E-02
hemoglobin	0.0E+00	1.0E-06	6.0E-06	10.6	1.9	9.3E-02	8.0E-02	1.0E-01	814	chartevents	cv	10.6	1.7	3.4E-02
									220228	chartevents	mv	10.5	1.9	2.8E-02
lactic acid	NAN	NAN	NAN	2.7	2.8	2.4E-02	2.0E-02	2.5E-02	1531	chartevents	cv	2.9	3.2	1.0E-02
									818	chartevents	cv	3.0	3.3	1.2E-02

Grouping	low	high	strict	avg	std	pres.	pres. cv	pres. mv	ItemID	Table	DB	avg	std	pres.
									225668	chartevents	mv	2.5	2.3	1.3E-02
magnesium	0.0E+00	5.9E-06	4.6E-05	2.1	0.4	8.2E-02	6.4E-02	9.6E-02	821	chartevents	cv	2.0	0.4	3.5E-02
									1532	chartevents	cv	2.1	0.4	2.7E-02
									220635	chartevents	mv	2.1	0.4	2.8E-02
mean blood pressure	2.2E-04	2.0E-05	6.1E-05	79.4	15.5	8.6E-01	8.1E-01	7.8E-01	224	chartevents	cv	81.5	14.5	7.4E-03
									220181	chartevents	mv	77.2	15.0	2.5E-01
									456	chartevents	cv	78.3	14.7	2.8E-01
									224322	chartevents	mv	79.6	15.2	2.1E-03
									6702	chartevents	cv	76.9	13.5	2.7E-03
									225312	chartevents	mv	79.7	18.7	1.1E-02
									220052	chartevents	mv	81.1	18.6	1.3E-01
									52	chartevents	cv	81.5	16.8	2.5E-01
oxygen saturation	0.0E+00	6.6E-07	3.6E-06	96.7	3.6	8.6E-01	7.9E-01	7.9E-01	646	chartevents	cv	97.0	3.5	4.8E-01
									220277	chartevents	mv	96.7	3.2	3.7E-01
									834	chartevents	cv	96.7	3.4	1.6E-02
									220227	chartevents	mv	96.1	4.1	7.3E-03
partial pressure of carbon dioxide	0.0E+00	2.0E-06	2.0E-06	41.2	9.6	8.2E-02	8.1E-02	7.5E-02	226062	chartevents	mv	44.6	13.4	7.0E-04
									778	chartevents	cv	40.8	9.1	4.2E-02
									220235	chartevents	mv	41.2	9.4	2.4E-02
partial pressure of oxygen	2.8E-04	0.0E+00	0.0E+00	145.8	84.9	4.2E-02	6.8E-02	5.9E-04	779	chartevents	cv	145.8	84.9	4.2E-02
partial thromboplastin time	4.0E-04	0.0E+00	5.2E-06	41.2	24.6	6.2E-02	5.3E-02	6.9E-02	1533	chartevents	cv	41.4	24.3	2.0E-02
									825	chartevents	cv	42.1	24.9	2.6E-02
									227466	chartevents	mv	42.0	25.0	2.0E-02
peak inspiratory pressure	6.9E-02	3.2E-03	5.7E-03	22.8	6.6	5.2E-02	4.2E-02	5.6E-02	535	chartevents	cv	25.3	6.1	2.5E-02
									224695	chartevents	mv	20.4	6.1	2.7E-02
ph	0.0E+00	2.2E-06	4.5E-05	7.4	0.1	9.1E-02	9.0E-02	8.2E-02	220274	chartevents	mv	7.4	0.1	1.7E-03
									860	chartevents	cv	7.4	0.1	1.7E-03
									780	chartevents	cv	7.4	0.1	4.5E-02
									223830	chartevents	mv	7.4	0.1	2.5E-02
									1126	chartevents	cv	7.4	0.1	4.2E-02
phosphorous	NAN	NAN	NAN	3.5	1.4	5.2E-02	4.5E-02	5.2E-02	1534	chartevents	cv	3.5	1.4	2.2E-02
									827	chartevents	cv	3.5	1.5	2.8E-02
									225677	chartevents	mv	3.4	1.4	2.5E-02
plateau pressure	NAN	NAN	NAN	20.2	5.7	3.1E-02	3.3E-02	2.3E-02	543	chartevents	cv	20.7	6.0	2.0E-02
									224696	chartevents	mv	19.3	5.0	1.1E-02
platelets	0.0E+00	2.3E-06	2.3E-06	205.0	113.4	8.5E-02	6.9E-02	9.8E-02	828	chartevents	cv	196.8	108.9	3.5E-02
									227457	chartevents	mv	203.9	113.8	2.8E-02
positive end-expiratory pressure	0.0E+00	1.6E-04	2.1E-04	7.2	3.5	1.6E-02	1.2E-02	2.0E-02	224700	chartevents	mv	7.4	3.5	6.5E-03
positive end-expiratory pressure set	NAN	NAN	NAN	6.2	2.9	7.4E-02	7.2E-02	6.4E-02	506	chartevents	cv	6.2	2.9	4.4E-02
									220339	chartevents	mv	6.2	2.9	3.0E-02

Grouping	low	high	strict	avg	std	pres.	pres. cv	pres. mv	ItemID	Table	DB	avg	std	pres.
post void residual	NAN	NAN	NAN	205.6	135.0	1.4E-03	2.3E-03	0.0E+00	512	chartevents	cv	205.6	135.0	1.4E-03
potassium	0.0E+00	9.7E-06	1.6E-05	4.1	0.6	1.1E-01	1.1E-01	9.6E-02	1535	chartevents	cv	4.1	0.6	4.3E-02
									829	chartevents	cv	4.1	0.6	5.7E-02
									227464	chartevents	mv	4.2	0.7	1.1E-02
potassium serum	NAN	NAN	NAN	4.1	0.9	3.2E-02	2.2E-03	6.5E-02	227442	chartevents	mv	4.1	0.9	3.2E-02
prothrombin time inr	NAN	NAN	NAN	1.5	1.2	5.9E-02	5.0E-02	6.6E-02	227467	chartevents	mv	1.5	0.9	1.9E-02
									1530	chartevents	cv	1.5	1.2	1.9E-02
									815	chartevents	cv	1.5	1.3	2.4E-02
prothrombin time pt	NAN	NAN	NAN	16.0	7.0	5.9E-02	5.0E-02	6.6E-02	1286	chartevents	cv	15.6	5.8	1.9E-02
									227465	chartevents	mv	16.5	7.5	1.9E-02
									824	chartevents	cv	15.4	5.4	2.4E-02
pulmonary artery pressure mean	NAN	NAN	NAN	29.6	9.3	3.1E-02	5.0E-02	1.1E-04	491	chartevents	cv	29.6	9.3	3.1E-02
pulmonary artery pressure systolic	NAN	NAN	NAN	38.1	12.3	9.6E-02	1.2E-01	4.9E-02	492	chartevents	cv	38.4	12.6	7.3E-02
									220059	chartevents	mv	37.0	11.1	2.3E-02
pulmonary capillary wedge pressure	NAN	NAN	NAN	17.1	7.2	2.7E-03	4.4E-03	3.8E-06	504	chartevents	cv	17.1	7.2	2.7E-03
red blood cell count	NAN	NAN	NAN	3.5	0.7	6.4E-02	6.4E-02	6.0E-02	833	chartevents	cv	3.5	0.6	3.2E-02
respiratory rate	0.0E+00	3.6E-07	2.4E-06	19.1	5.7	8.8E-01	8.2E-01	8.0E-01	224690	chartevents	mv	18.9	5.7	2.1E-02
									618	chartevents	cv	19.4	5.7	5.0E-01
									220210	chartevents	mv	19.2	5.5	3.8E-01
									224689	chartevents	mv	9.5	10.8	2.6E-02
									614	chartevents	cv	2.8	5.0	2.7E-02
									651	chartevents	cv	22.3	7.3	3.4E-03
									224422	chartevents	mv	20.8	6.9	2.5E-03
									615	chartevents	cv	18.6	6.2	4.3E-02
respiratory rate set	NAN	NAN	NAN	15.6	8.0	4.6E-02	4.7E-02	3.7E-02	224688	chartevents	mv	16.8	10.7	1.8E-02
									619	chartevents	cv	14.9	5.7	2.9E-02
sodium	1.6E-05	1.6E-06	0.0E+00	138.6	5.3	1.0E-01	8.0E-02	1.2E-01	220645	chartevents	mv	138.8	5.4	3.1E-02
									837	chartevents	cv	138.9	5.1	3.9E-02
									1536	chartevents	cv	138.9	5.1	3.0E-02
									226534	chartevents	mv	136.2	5.1	4.2E-03
systemic vascular resistance	NAN	NAN	NAN	996.6	354.7	3.4E-02	5.6E-02	2.5E-04	626	chartevents	cv	996.6	354.7	3.4E-02
systolic blood pressure	0.0E+00	0.0E+00	1.1E-06	121.8	22.0	8.7E-01	8.2E-01	7.8E-01	227243	chartevents	mv	123.7	26.5	1.4E-04
									224167	chartevents	mv	120.2	27.6	1.7E-04
									442	chartevents	cv	120.3	24.7	4.7E-04
									220179	chartevents	mv	121.1	21.4	2.5E-01
									455	chartevents	cv	121.2	22.0	2.8E-01
									6701	chartevents	cv	109.3	22.3	2.7E-03
									225309	chartevents	mv	115.3	23.7	1.1E-02
									220050	chartevents	mv	121.5	22.3	1.3E-01
									51	chartevents	cv	122.5	24.4	2.5E-01
temperature	8.0E-06	0.0E+00	2.7E-04	37.0	0.8	2.9E-01	3.0E-01	2.3E-01	676	chartevents	cv	37.2	0.7	5.9E-02

Grouping	low	high	strict	avg	std	pres.	pres. cv	pres. mv	ItemID	Table	DB	avg	std	pres.
									677	chartevents	cv	36.9	0.8	1.3E-01
									223762	chartevents	mv	37.1	0.9	1.3E-02
									678	chartevents	cv	36.9	0.8	1.3E-01
									223761	chartevents	mv	36.8	0.7	9.7E-02
									679	chartevents	cv	37.2	0.7	5.9E-02
tidal volume observed	NAN	NAN	NAN	541.5	737.0	6.4E-02	5.6E-02	6.3E-02	224685	chartevents	mv	504.8	2208.9	2.7E-02
									682	chartevents	cv	575.2	124.7	2.8E-02
tidal volume set	NAN	NAN	NAN	535.8	121.2	4.3E-02	4.3E-02	3.6E-02	224684	chartevents	mv	489.4	88.5	1.7E-02
									683	chartevents	cv	565.3	129.7	2.7E-02
tidal volume spontaneous	NAN	NAN	NAN	489.4	2207.5	3.2E-02	3.0E-02	2.9E-02	684	chartevents	cv	457.2	194.1	1.8E-02
									224686	chartevents	mv	531.9	3355.6	1.4E-02
total protein	NAN	NAN	NAN	5.7	1.1	1.9E-04	3.1E-04	3.8E-06	849	chartevents	cv	5.7	1.1	1.9E-04
									1539	chartevents	cv	5.7	1.1	1.5E-04
troponin-i	0.0E+00	4.1E-04	0.0E+00	7.6	10.7	9.2E-04	1.5E-03	9.1E-05	851	chartevents	cv	7.7	10.7	6.2E-04
troponin-t	2.1E-05	4.7E-04	1.7E-04	0.9	2.2	1.3E-02	8.8E-03	1.7E-02	227429	chartevents	mv	0.7	1.9	5.5E-03
venous pvo2	NAN	NAN	NAN	43.9	14.9	4.3E-04	7.0E-04	1.7E-05	859	chartevents	cv	43.9	14.9	4.3E-04
weight	0.0E+00	0.0E+00	2.1E-04	83.1	23.4	2.8E-02	1.3E-02	4.2E-02	763	chartevents	cv	84.3	23.0	8.1E-03
									224639	chartevents	mv	86.3	23.7	6.9E-03
									226512	chartevents	mv	80.8	22.5	6.6E-03
									226531	chartevents	mv	80.7	23.4	1.1E-02
white blood cell count	0.0E+00	0.0E+00	2.5E-06	11.9	10.0	8.1E-02	6.5E-02	9.6E-02	220546	chartevents	mv	11.6	9.7	2.7E-02
									861	chartevents	cv	12.2	10.0	3.2E-02
									1127	chartevents	cv	12.2	10.2	3.1E-02
									1542	chartevents	cv	12.2	10.4	2.5E-02

Appendix B

Quantitative literature review for relation extraction

Following Chapter 3, this section describes the quantitative literature review conducted as of February 2019, to assess the reproducibility issue in the relation extraction (RE) domain. The literature review is found in Table B.1, with the following column key:

cite = number of papers that cited the paper
code = whether code was publicly available (y for yes and • for no)
ablation = whether an ablation study was performed
hyperparam = whether hyperparameter details were mentioned
cross val = whether cross validation details were mentioned
word-embed = whether word embeddings were used and mentioned
datasets = number of datasets evaluated on

paper	cite	code	ablation	hyperparam	cross val	word-embed	datasets
Socher et al. [274]	890	y	•	y	•	y	2
Zeng et al. [322]	477	•	y	y	y	y	1
Santos et al. [265]	220	•	y	y	y	y	1
Nguyen and Verspoor [215]	146	•	y	y	y	•	2
Miwa and Bansal [205]	175	•	y	y	y	•	3
Li and Jurafsky [164]	107	y	y	y	•	y	6
Xu et al. [311]	108	•	y	y	•	y	1
Wang et al. [297]	102	•	y	•	•	y	1
Hashimoto et al. [106]	64	•	y	y	•	y	1
Zhang and Wang [326]	68	•	y	•	y	y	2
Vu et al. [295]	57	•	y	y	•	y	1
Yin et al. [316]	116	•	n	y	•	•	7
Yu et al. [321]	45	y	y	y	y	y	1
Xu et al. [313]	54	y	y	y	•	•	1
Zhang et al. [328]	51	•	•	•	•	y	1
Nguyen and Grishman [216]	42	•	y	y	•	y	2
Qin et al. [240]	39	•	•	y	y	y	1
Cai et al. [30]	44	•	y	y	•	y	1
Sahu et al. [262]	32	•	y	y	y	y	1
Adel et al. [3]	29	y	y	•	•	y	1
Zeng et al. [323]	190	•	y	y	•	y	1
Xu et al. [312]	171	•	y	y	•	y	1
Zhang et al. [333]	3	•	y	y	•	y	2

Paper	cite	code	ablation	hyperparam	cross val	word-embed	datasets
Levy et al. [160]	20	y	y	y	•	y	1
Liu et al. [181]	48	•	•	y	•	y	1
Zhao et al. [336]	41	y	y	y	•	y	1
Ebrahimi and Dou [72]	30	•	•	•	•	•	2
Li et al. [162]	27	y	y	y	y	y	2
Quan et al. [241]	23	y	•	y	y	y	2
Sahu and Anand [263]	13	y	y	y	•	y	1
Liu et al. [180]	9	•	•	y	•	y	1
Lim and Kang [175]	4	•	•	•	•	•	1
Zheng et al. [337]	12	•	y	y	y	y	1
Wang et al. [300]	5	n	y	y	•	y	1
Lim et al. [176]	1	y	y	y	y	y	2
Kavuluru et al. [149]	8	•	•	y	•	•	1
Huang et al. [119]	4	•	•	y	•	y	1
Juan Hou and Ceesay [143]	1	•	•	•	•	y	1
Lim and Kang [174]	4	y	•	y	•	y	1
Rotsztejn et al. [259]	2	•	•	y	y	y	1
Jin et al. [131]	0	•	y	y	y	y	1
Sahu et al. [262]	31	•	y	y	y	y	1
Luo [190]	21	•	•	y	•	y	1
Lv et al. [193]	15	•	•	•	•	•	1
Jin et al. [131]	14	•	y	y	•	y	1
Chikka and Karlapalem [47]	1	y	•	y	•	•	1

Paper	**cite**	**code**	**ablation**	**hyperparam**	**cross val**	**word-embed**	**datasets**
Li et al. [168]	0	y	•	y	y	y	1
Li et al. [165]	0	•	•	•	•	•	5
Suster et al. [281]	0	y	•	y	•	y	1
Luo et al. [191]	16	y	•	y	•	y	1
He et al. [108]	2	•	•	y	•	y	1
He et al. [107]	0	•	•	y	y	y	2
Nguyen and Verspoor [215]	1	•	y	y	•	y	1

Table B.1: Quantitative Literature Review

Appendix C

Random Search result distributions for relation extraction

Following the discussion in Section 3.5.9 about random search for hyperparameter tuning, the exact number of experiments run on each dataset differed due to variability in the availability of computation time. A total of `107` experiments were run for `semeval`, `104` for `ddi` and `134` for `i2b2`. Statistics for performance on the randomly sampled dev set are present in tables C.1, C.2 and C.3.

Statistic \ Search Subset	**All**	**Top 10%**
Mean	76.83	80.87
Stddev	9.93	0.31
Median	79.42	80.74
Max	81.37	81.37
Min	4.73	80.54
Range	76.64	0.83

Table C.1: Random Search experiment statistics for `semeval`. The two columns All and Top 10% determine the subset of the results statistics are gathered for. All = distribution of Macro-F1 scores over 107 experiments, top 10% = distribution over top 10% of the results.

These tables demonstrate that random search reduces the variability of results and converges to better performance than the default hyperparameters.

Statistic \ Search Subset	**All**	**Top 10%**
Mean	80.24	82.08
Stddev	1.63	0.25
Median	80.45	82.04
Max	82.57	82.57
Min	71.21	81.74
Range	11.36	0.83

Table C.2: Random Search experiment statistics for `ddi`. The two columns All and Top 10% determine the subset of the results statistics are gathered for. All = distribution of Macro-F1 scores over 104 experiments, top 10% = distribution over top 10% of the results.

Statistic \ Search Subset	**All**	**Top 10%**
Mean	69.61	72.19
Stddev	1.54	0.39
Median	69.78	72.13
Max	72.86	72.86
Min	62.92	71.64
Range	9.94	1.22

Table C.3: Random Search experiment statistics for `i2b2`. The two columns All and Top 10% determine the subset of the results statistics are gathered for. All = distribution of Macro-F1 scores over 134 experiments, top 10% = distribution over top 10% of the results.

Appendix D

Supplementary work for multimodal representation learning

Figure D-1: t-SNE visualization in 2 dimensions for image embeddings in the joint model (Chapter 4) the embeddings in the image-only model. We can observe a clearer separation between the disease categories via our joint modeling technique.

Edema severity	Regex keyword terms	Number of reports	Accuracy
"Overall"	N/A	485	89.69%
Level 0 – none (n=216)	(no) pulmonary edema	222	88.74%
	(no) vascular congestion	43	100.00%
	(no) fluid overload	4	100.00%
	(no) acute cardiopulmonary process	115	98.27%
Level 1 – vascular congestion (n=98)	cephalization	17	94.12%
	pulmonary vascular congestion	96	98.96%
	hilar engorgement	3	100.00%
	vascular plethora	13	100.00%
	pulmonary vascular prominence	1	100.00%
	pulmonary vascular engorgement	8	87.50%
Level 2 – interstitial edema (n=105)	interstitial opacities	30	73.33%
	kerley	13	100.00%
	interstitial edema	92	94.57%
	interstitial thickening	6	66.67%
	interstitial pulmonary edema	21	100.00%
	interstitial marking	19	68.42%
	interstitial abnormality	10	70.00%
	interstitial abnormalities	2	100.00%
	interstitial process	2	100.00%
Level 3 – alveolar edema (n=66)	alveolar infiltrates	10	100.00%
	severe pulmonary edema	58	98.28%
	perihilar infiltrates	1	100.00%
	hilar infiltrates	1	100.00%
	parenchymal opacities	6	16.67%
	alveolar opacities	7	100.00%
	ill defined opacities	1	100.00%
	ill-defined opacities	1	0.00%
	patchy opacities	10	10.00%

Table D.1: Validation of regex keyword terms. The accuracy (positive predictive value) of the regular expression results for levels 0-3 based on the expert review results are 90.74%, 80.61%, 95.24%, and 90.91%, respectively. The total number of reports from all the keywords is more than 485 because some reports contain more than one keywords.

In Chapter 4, we report final joint model results with the dot product similarity metric. We experimented with 3 different similarity metrics with or without the contrastive loss. Without the contrastive loss, the joint loss would be equivalent to minimizing the distance between the paired image and text and not considering any imposters. In table D.2, we can see that the best performance is offered by `ranking-dot`, which refers to the contrastive loss applied with negative sampling and dot product used as a similarity metric.

Metrics	`dot`	`l2`	`cosine`	`ranking-dot`	`ranking-l2`	`ranking-cosine`
AUC(0)	0.65	0.78	0.77	0.80	0.77	0.77
AUC(1)	0.55	0.62	0.61	0.64	0.63	0.62
AUC(2)	0.57	0.66	0.63	0.68	0.62	0.62
AUC(3)	0.61	0.83	0.81	0.87	0.81	0.83
AUC(0v1)	0.61	0.71	0.69	0.73	0.71	0.71
AUC(0v2)	0.65	0.78	0.77	0.79	0.76	0.74
AUC(0v3)	0.72	0.93	0.91	0.93	0.92	0.90
AUC(1v2)	0.50	0.60	0.56	0.58	0.58	0.55
AUC(1v3)	0.51	0.80	0.79	0.83	0.77	0.77
AUC(2v3)	0.52	0.70	0.68	0.78	0.66	0.72
MSE	1.18	0.85	0.87	0.76	0.87	0.91
Macro-F1	0.15	0.42	0.44	0.45	0.43	0.41
Accuracy	0.44	0.49	0.49	0.51	0.48	0.47

Table D.2: Initial experiments to assess the model performance with different similarity metrics applied for the joint loss, with or without considering the negative samples.

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *The SIGSAC conference on computer and communications security*, 2016.

[2] Kirkwood F Adams Jr, Gregg C Fonarow, Charles L Emerman, Thierry H LeJemtel, Maria Rosa Costanzo, William T Abraham, Robert L Berkowitz, Marie Galvao, Darlene P Horton, ADHERE Scientific Advisory Committee, Investigators, et al. Characteristics and outcomes of patients hospitalized for heart failure in the united states: rationale, design, and preliminary observations from the first 100,000 cases in the acute decompensated heart failure national registry (adhere). *American heart journal*, 149(2):209–216, 2005.

[3] Heike Adel, Benjamin Roth, and Hinrich Schütze. Comparing convolutional neural networks to traditional models for slot filling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 828–838. Association for Computational Linguistics, 2016.

[4] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361, 2018.

[5] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, 361(k1479), 2018. URL `https://www.bmj.com/content/bmj/361/bmj.k1479.full.pdf`.

[6] Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Vinith M. Suriyakumar, Om Thakkar, and Abhradeep Thakurta. Public data-assisted mirror descent for private model training. In *ICML*, 2022.

[7] Ehsan Amid, Om Dipakbhai Thakkar, Arun Narayanan, Rajiv Mathews, and Francoise Beaufays. Extracting Targeted Training Data from ASR Models, and How to Mitigate It. In *Interspeech*, 2022.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image

captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[9] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

[10] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[11] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555. Association for Computational Linguistics, 2017.

[12] Nguyen Bach and Sameer Badaskar. A survey on relation extraction. "`http://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction-Slides.pdf`", 2007. [Online; accessed 30-Dec-2018].

[13] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

[14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Annual Symposium on Foundations of Computer Science*, 2014.

[15] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[16] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[17] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[18] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, 2019.

[19] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[20] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[21] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[22] Jari Björne and Tapio Salakoski. Tees 2.2: biomedical event extraction for diverse corpora. *BMC bioinformatics*, 16(16):S4, 2015.

[23] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *Machine Learning for Health Workshop*, pages 126–140. PMLR, 2020.

[24] William Boag, Mercy Oladipo, and Peter Szolovits. Ehr safari: Data is contextual. In *Machine Learning for Healthcare Conference*, pages 391–408. PMLR, 2022.

[25] William Boag, Geeticka Chauhan, Peter Szolovits, and Weitzner Daniel. Grounding Privacy Notions for Differential Privacy Trade-offs in Healthcare. *In anonymous review at conference*, 2024.

[26] Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26, 2018.

[27] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[28] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term only fine-tuning of foundation models. *arXiv preprint arXiv:2210.00036*, 2022.

[29] Davide Buscaldi, Anne-Kathrin Schumann, Behrang Qasemizadeh, Haïfa Zargayouna, and Thierry Charnois. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *International Workshop on Semantic Evaluation (SemEval-2018)*, pages 679–688, 2017.

[30] Rui Cai, Xiaodong Zhang, and Houfeng Wang. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 756–765, 2016.

[31] Jose Camacho-Collados and Mohammad Taher Pilehvar. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 40–46, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[32] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, 2019.

[33] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

[34] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=TatRHT_1cK`.

[35] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

[36] Triana Carmenate, Peeraya Inyim, Nupoor Pachekar, Geeticka Chauhan, Leonardo Bobadilla, Mostafa Batouli, and Ali Mostafavi. Modeling occupant-building-appliance interaction for energy waste analysis. *Procedia Engineering*, 145:42–49, 2016.

[37] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[38] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[39] Dustin Charles, Meghan Gabriel, Talisha Searcy, et al. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2014. *ONC data brief*, 23(4), 2015.

[40] Geeticka Chauhan, Matthew B.A. McDermott, and Peter Szolovits. REflex: Flexible framework for relation extraction in multiple domains. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 30–47, Florence, Italy, August 2019. Association for Computational Linguistics.

[41] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *In MICCAI International Conference*, pages 529–539. Springer, 2020.

[42] Geeticka Chauhan, Steve Chien, Om Thakkar, Abhradeep Guha Thakurtha, and Arun Narayanan. Training Large ASR Encoders with Differential Privacy. *In anonymous review at conference*, 2024.

[43] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1), 2018. URL `http://www.nature.com/articles/s41598-018-24271-9`.

[44] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010.

[45] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020.

[46] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. PRIOR: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21361–21371, 2023.

[47] Veera Raghavendra Chikka and Kamalakar Karlapalem. A hybrid deep learning approach for medical relation extraction. *CoRR*, 2018.

[48] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, 2022.

[49] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

[50] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.

[51] Sanghyuk Roy Choi and Minhyeok Lee. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology*, 12(7): 1033, 2023.

[52] Marc Claesen and Bart De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

[53] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[54] Nneka I Comfere, Margot S Peters, Sarah Jenkins, Kandace Lackore, Kathleen Yost, and Jon Tilburt. Dermatopathologists' concerns and challenges with clinical information in the skin biopsy requisition form: a mixed-methods study. *Journal of cutaneous pathology*, 42(5):333–345, 2015.

[55] United States Congress. California Consumer Privacy Act of 2018 (CCPA) – Section 3, Title 1.81.5 added to Part 4 of Division 3 of the California Civil Code. `https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5`, 2018.

[56] United States Congress. Health insurance portability and accountability act of 1996 (HIPAA) – pub. l. no. 104-191 (1996). `https://www.govinfo.gov/app/details/PLAW-104publ191`, 2022.

[57] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.

[58] Matt Crane. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association of Computational Linguistics*, 6:241–252, 2018.

[59] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[60] Kevin De Angeli, Shang Gao, Ioana Danciu, Eric B Durbin, Xiao-Cheng Wu, Antoinette Stroup, Jennifer Doherty, Stephen Schwartz, Charles Wiggins, Mark Damesyn, et al. Class imbalance in out-of-distribution datasets: Improving the robustness of the textcnn for the classification of rare cancer types. *Journal of biomedical informatics*, 125:103957, 2022.

[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[62] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021.

[63] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRa: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[65] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[66] Dan Doherty, Kathleen J Millen, and A James Barkovich. Midbrain and hindbrain malformations: advances in clinical diagnosis, imaging, and genetics. *The Lancet Neurology*, 12(4):381–393, 2013.

[67] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

[68] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[69] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023.

[70] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, 2006.

[71] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407, 2014.

[72] Javid Ebrahimi and Dejing Dou. Chain based rnn for relation classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1244–1249, 2015.

[73] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model. In *Proceedings of Machine Learning for Health*, 2021.

[74] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[75] European Parliament and Council of the European Union. General Data Protection Regulation (GDPR) – Regulation (EU) 2016/679 of the European Parliament and of the Council. URL `https://data.europa.eu/eli/reg/2016/679/oj`.

[76] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[77] Joseph Ficek, Wei Wang, Henian Chen, Getachew Dagne, and Ellen Daley. Differential privacy in health research: A scoping review. *Journal of the American Medical Informatics Association*, 28(10):2269–2276, 2021.

[78] Antske Fokkens, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1691–1701, 2013.

[79] Food, Drug Administration, et al. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). *web entry*, 2019.

[80] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019.

[81] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[82] Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, 2018.

[83] Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *ICML*, 2023.

[84] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2014.

[85] Marzyeh Ghassemi, Marco Pimentel, Tristan Naumann, Thomas Brennan, David Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[86] Marzyeh Ghassemi, Mike Wu, Michael C Hughes, Peter Szolovits, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. *AMIA Summits on Translational Science Proceedings*, 2017:82, 2017.

[87] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4):e157–e159, 2019.

[88] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[89] Mihai Gheorghiade, Ferenc Follath, Piotr Ponikowski, Jeffrey H Barsuk, John EA Blair, John G Cleland, Kenneth Dickstein, Mark H Drazner, Gregg C Fonarow, Tiny Jaarsma, et al. Assessing and grading congestion in acute heart failure: a scientific statement from the acute heart failure committee of the heart failure association of the european society of cardiology and endorsed by the european society of intensive care medicine. *European journal of heart failure*, 12(5):423–433, 2010.

[90] Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics, 2007.

[91] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[92] Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting clinical outcomes across changing electronic health record systems. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1497–1505, 2017.

[93] Google Cloud Architecture Center. Data preprocessing for ml: options and recommendations. `https://cloud.google.com/architecture/data-preprocessing-for-ml-with-tf-transform-pt1`, 2022. Accessed: 2022-10-24.

[94] Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, and David Sontag. Fast, structured clinical documentation via contextual autocomplete. In *Machine Learning for Healthcare Conference*, pages 842–870. PMLR, 2020.

[95] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[96] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *The International Conference on Machine Learning (ICML)*, 2006.

[97] Daria Grechishnikova. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific reports*, 11(1):321, 2021.

[98] Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.

[99] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech*, 2020.

[100] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[101] Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. Towards safe and aligned large language models for medicine. *arXiv preprint arXiv:2403.03744*, 2024.

[102] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*, 2024.

[103] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[104] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. *Advances in Neural Information Processing Systems*, 29, 2016.

[105] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018.

[106] Kazuma Hashimoto, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. Simple customization of recursive neural networks for semantic relation classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1372–1376, 2013.

[107] Bin He, Yi Guan, and Rui Dai. Convolutional gated recurrent units for medical relation classification. *CoRR*, abs/1807.11082, 2018.

[108] Bin He, Yi Guan, and Rui Dai. Classifying medical relations in clinical text via convolutional neural networks. *Artificial intelligence in medicine*, 2018.

[109] Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *The Eleventh International Conference on Learning Representations*, 2023.

[110] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[111] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics, 2009.

[112] Geoffrey Hinton. Deep learning—a technology with the potential to transform health care. *Jama*, 320(11):1101–1102, 2018.

[113] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

[114] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[115] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[116] Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708, 2017.

[117] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*, pages 3–13. Springer, 2021.

[118] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[119] Degen Huang, Zhenchao Jiang, Li Zou, and Lishuang Li. Drug-drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. *Information Sciences*, 415, 06 2017.

[120] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[121] Sharon Ann Hunt, William T Abraham, Marshall H Chin, Arthur M Feldman, Gary S Francis, Theodore G Ganiats, Mariell Jessup, Marvin A Konstam, Donna M Mancini, Keith Michl, et al. 2009 focused update incorporated into the acc/aha 2005 guidelines for the diagnosis and management of heart failure in adults: a report of the american college of cardiology foundation/american heart association task force on practice guidelines developed in collaboration with the international society for heart and lung transplantation. *Journal of the American College of Cardiology*, 53(15):e1–e90, 2009.

[122] Darrel C Ince, Leslie Hatton, and John Graham-Cumming. The case for open computer programs. *Nature*, 482(7386):485, 2012.

[123] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.

[124] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, 2018. Association for Computational Linguistics.

[125] Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *The International Conference on Learning Representations (ICLR)*, 2023.

[126] Matthew Jagielski, Om Thakkar, and Lun Wang. Noise masking attacks and defenses for pretrained speech models. In *ICASSP*, 2024.

[127] Labiba Jahan, Geeticka Chauhan, and Mark Finlayson. Building on word animacy to determine coreference chain animacy in cultural narratives. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 13, pages 198–203, 2017.

[128] Labiba Jahan, Geeticka Chauhan, and Mark A Finlayson. A new approach to animacy detection. In *Proceedings of the 27th International COLING*, 2018.

[129] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[130] Kristel J.M. Janssen, A. Rogier T. Donders, Frank E. Harrell, Yvonne Vergouwe, Qingxia Chen, Diederick E. Grobbee, and Karel G.M. Moons. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, 63(7):721–727, 2010.

[131] Di Jin, Franck Dernoncourt, Elena Sergeeva, Matthew McDermott, and Geeticka Chauhan. MIT-MEDG at SemEval-2018 task 7: Semantic relation classification via convolution neural network. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 798–804, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[132] Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3):e1548, 2022.

[133] Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, 2021.

[134] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[135] Baoyu Jing, Pengtao Xie, and Eric Xing. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[136] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[137] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine learning for healthcare conference*, pages 361–376. PMLR, 2017.

[138] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[139] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[140] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[141] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

[142] Rohit Joshi and Peter Szolovits. Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1276. American Medical Informatics Association, 2012.

[143] Wen Juan Hou and Bamfa Ceesay. Extraction of drug-drug interaction using neural embedding. *Journal of Bioinformatics and Computational Biology*, 16, 2018.

[144] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP*, 2020.

[145] Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *ICML*, 2021.

[146] Myeongsu Kang and Jing Tian. Machine learning: Data pre-processing. *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, pages 111–130, 2018.

[147] Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. Eliie: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071, 2017.

[148] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[149] Ramakanth Kavuluru, Anthony Rios, and Tung Tran. Extracting drug-drug interactions with word and character-level recurrent neural networks. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 5–12, 2017.

[150] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[151] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[152] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

[153] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[154] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL `https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf`.

[155] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

[156] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 978–984. Association for Computational Linguistics, 2017.

[157] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

[158] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):1–30, 2018.

[159] Eric P Lehman, Rahul G Krishnan, Xiaopeng Zhao, Roger G Mark, and H Lehman Li-Wei. Representation learning approaches to detect false arrhythmia alarms from ecg dynamics. In *Machine learning for healthcare conference*, pages 571–586. PMLR, 2018.

[160] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342. Association for Computational Linguistics, 2017.

[161] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[162] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):198, 2017.

[163] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[164] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*, 2015.

[165] Q. Li, Z. Yang, L. Luo, L. Wang, Y. Zhang, H. Lin, J. Wang, L. Yang, K. Xu, and Y. Zhang. A multi-task learning based approach to biomedical entity relation extraction. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 680–682, 2018.

[166] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

[167] Xuechen Li, Daogao Liu, Tatsunori B Hashimoto, Huseyin A Inan, Janardhan Kulkarni, Yin-Tat Lee, and Abhradeep Guha Thakurta. When does differentially private learning not suffer in high dimensions? *NeurIPS*, 2022.

[168] Yifu Li, Ran Jin, and Yuan Luo. Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (seg-gcrns). *Journal of the American Medical Informatics Association*, 26(3):262–268, 2018.

[169] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *NeurIPS*, 2018.

[170] R Liao, G Chauhan, P Golland, S Berkowitz, and S Horng. Pulmonary edema severity grades based on mimic-cxr (version 1.0. 1). *PhysioNet*, 2021.

[171] Ruizhi Liao, Jonathan Rubin, Grace Lam, Seth Berkowitz, Sandeep Dalal, William Wells, Steven Horng, and Polina Golland. Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. *arXiv preprint arXiv:1902.10785*, 2019.

[172] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 273–283. Springer, 2021.

[173] Bryan Lim and Mihaela van der Schaar. Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pages 137–160. PMLR, 2018.

[174] Sangrak Lim and Jaewoo Kang. Chemical–gene relation extraction using recursive neural network. In *Database*, 2018.

[175] Sangrak Lim and Jaewoo Kang. Drug drug interaction extraction from the literature using a recursive neural network. In *PloS one*, 2018.

[176] Sangrak Lim, Kyubum Lee, and Jaewoo Kang. Drug drug interaction extraction from the literature using a recursive neural network. *Plos one*, 13:1–17, 2018.

[177] Jau-Huei Lin and Peter J Haug. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of biomedical informatics*, 41(1):1–14, 2008.

[178] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[179] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[180] Shengyu Liu, Kai Chen, Qingcai Chen, and Buzhou Tang. Dependency-based convolutional neural network for drug-drug interaction extraction. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1074–1080, 2016.

[181] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016, 2016.

[182] WeiKang Liu, Yanchun Zhang, Hong Yang, and Qinxue Meng. A survey on differential privacy for medical data analysis. *Annals of Data Science*, 11(2): 733–747, 2024.

[183] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.

[184] Yudong Liu, Zhongmin Shi, and Anoop Sarkar. Exploiting rich syntactic information for relation extraction from biomedical articles. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 97–100. Association for Computational Linguistics, 2007.

[185] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.

[186] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[187] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[188] HP Luhn. 11 keyword-in-context index for technical literature (kwic index). *Readings in automatic language processing*, 1:159, 1966.

[189] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE, 2023.

[190] Yuan Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 72, 07 2017.

[191] Yuan Luo, Yu Cheng, Özlem Uzuner, Peter Szolovits, and Justin Starren. Segment convolutional neural networks (seg-cnns) for classifying relations in clinical notes. *Journal of the American Medical Informatics Association*, 25(1):93–98, 2017.

[192] Zelun Luo, Daniel J Wu, Ehsan Adeli, and Li Fei-Fei. Scalable differential privacy with sparse network finetuning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[193] Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. Clinical relation extraction with deep learning. In *International Journal of Hybrid Information Technology*, 2016.

[194] Kit-Kay Mak and Mallikarjuna Rao Pichika. Artificial intelligence in drug development: present status and future prospects. *Drug discovery today*, 24(3): 773–780, 2019.

[195] Angrosh Mandya, Danushka Bollegala, Frans Coenen, and Katie Atkinson. Combining long short term memory and convolutional neural network for cross-sentence n-ary relation extraction. *arXiv preprint arXiv:1811.00845*, 2018.

[196] Matthew McDermott, Tom Yan, Tristan Naumann, Nathan Hunt, Harini Suresh, Peter Szolovits, and Marzyeh Ghassemi. Semi-supervised biomedical translation with cycle wasserstein regression gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[197] Matthew BA McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586): eabb1655, 2021.

[198] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *ICLR*, 2018.

[199] Francisco Melo. *Area under the ROC Curve*, pages 38–39. Springer New York, New York, NY, 2013. ISBN 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_209. URL `https://doi.org/10.1007/978-1-4419-9863-7_209`.

[200] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[201] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[202] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[203] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[204] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[205] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116. Association for Computational Linguistics, 2016.

[206] Marissa Mock, Suzanne Edavettal, Christopher Langmead, and Alan Russell. Ai can help to speed up drug discovery—but only if we give it the right data. *Nature*, 621(7979):467–470, 2023.

[207] Mehdi Moradi, Ali Madani, Yaniv Gur, Yufan Guo, and Tanveer Syeda-Mahmood. Bimodal network architectures for automatic generation of image annotation from text. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 449–456. Springer, 2018.

[208] Mohammad Amin Morid, Kensaku Kawamoto, Travis Ault, Josette Dorius, and Samir Abdelrahman. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1312. American Medical Informatics Association, 2017.

[209] Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, and David R Karger. Medknowts: Unified documentation and information retrieval for electronic health records. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 1169–1183, 2021.

[210] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[211] C David Naylor. On the prospects for a (deep) learning health care system. *Jama*, 320(11):1099–1100, 2018.

[212] Bret Nestor, Matthew McDermott, Geeticka Chauhan, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. *arXiv preprint arXiv:1811.12583*, 2018.

[213] Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.

[214] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL `https://www.aclweb.org/anthology/W19-5034`.

[215] Dat Quoc Nguyen and Karin Verspoor. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. *arXiv preprint arXiv:1805.10586*, 2018.

[216] Thien Huu Nguyen and Ralph Grishman. Combining neural networks and log-linear models to improve relation extraction. *CoRR*, abs/1511.05926, 2015.

[217] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023.

[218] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[219] OpenAI. Gpt-4 technical report, 2023.

[220] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[221] Kerem Ozturk, Zuzan Cayci, Jason Gotlib, Cem Akin, Tracy I George, and Celalettin Ustun. Non-hematologic diagnosis of systemic mastocytosis: collaboration of radiology and pathology. *Blood Reviews*, 45:100693, 2021.

[222] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *ICASSP*, 2015.

[223] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkZB1XbRZ`.

[224] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[225] Eunjeong L Park, Masato Hagiwara, Dmitrijs Milajevs, Nelson F Liu, Geeticka Chauhan, and Liling Tan. Proceedings of second workshop for nlp open source software (nlp-oss). In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020.

[226] Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, and Rakesh K Tekade. Artificial intelligence in drug discovery and development. *Drug discovery today*, 26(1):80, 2021.

[227] Maya Pavlova, Tia Tuinstra, Hossein Aboutalebi, Andy Zhao, Hayden Gunraj, and Alexander Wong. Covidx cxr-3: a large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2206.03671*, 2022.

[228] Martin Pelikan, Sheikh Shams Azam, Vitaly Feldman, Jan Silovsky, Kunal Talwar, Tatiana Likhomanenko, et al. Federated learning with differential privacy for end-to-end speech recognition. *arXiv preprint arXiv:2310.00098*, 2023.

[229] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018.

[230] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

[231] Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. Clinically Correct Report Generation from Chest X-Rays Using Templates. In *Machine Learning in Medical Imaging*. Springer International Publishing, 2021.

[232] Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58:156–165, 2015.

[233] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5781–5789, 2017.

[234] Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics (ACL)*, 2022.

[235] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 2023.

[236] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

[237] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[238] Sampo Pyssalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43, 2013.

[239] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2023.

[240] Pengda Qin, Weiran Xu, and Jun Guo. An empirical convolutional neural network approach for semantic relation classification. *Neurocomput.*, 190(C): 1–9, May 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2015.12.091. URL `https://doi.org/10.1016/j.neucom.2015.12.091`.

[241] Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. Multichannel convolutional neural network for biological relation extraction. In *BioMed research international*, 2016.

[242] Keegan Quigley, Miriam Cha, Ruizhi Liao, Geeticka Chauhan, Steven Horng, Seth Berkowitz, and Polina Golland. Radtex: Learning efficient radiograph representations from text reports. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 22–31. Springer, 2022.

[243] Keegan Quigley, Miriam Cha, Josh Barua, Geeticka Chauhan, Seth Berkowitz, Steven Horng, and Polina Golland. Bidirectional captioning for clinically accurate and interpretable models. *arXiv preprint arXiv:2310.19635*, 2023.

[244] Keegan Quigley, Miriam Cha, Josh Barua, Geeticka Chauhan, Steven Horng, Seth Berkowitz, and Polina Golland. Improving medical visual representations through radiology report generation. *In anonymous review at conference*, 2024.

[245] DragomirR. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26, 2013. ISSN 1574-020X. doi: 10.1007/s10579-012-9211-2. URL `http://dx.doi.org/10.1007/s10579-012-9211-2`.

[246] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[247] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[248] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[249] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[250] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017.

[251] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1 (1):18, 2018.

[252] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 41–47. Association for Computational Linguistics, 2002.

[253] Nils Reimers and Iryna Gurevych. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*, 2017.

[254] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1035. URL https://www.aclweb.org/anthology/D17-1035.

[255] Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1035. URL http://aclweb.org/anthology/D17-1035.

[256] Bryan Rink and Sanda Harabagiu. UTD: Classifying semantic relations by combining lexical and semantic resources. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1057.

[257] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[258] Corby Rosset. Turing-NLG: A 17-billion-parameter language model by microsoft, 2020. URL https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/. Accessed on May 8, 2024.

[259] Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. Eth-ds3lab at semeval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 689–696. Association for Computational Linguistics, 2018.

[260] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[261] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[262] Sunil Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. Relation extraction from clinical texts using domain invariant convolutional neural network. In *Proceedings of the 15th Workshop on Biomedical*

Natural Language Processing, pages 206–215. Association for Computational Linguistics, 2016.

[263] Sunil Kumar Sahu and Ashish Anand. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of Biomedical Informatics*, 86:15 – 24, 2018. ISSN 1532-0464.

[264] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[265] Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*, 2015.

[266] Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 341–350, 2013.

[267] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[268] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2016.

[269] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[270] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[271] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023.

[272] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations

for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020.

[273] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[274] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[275] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[276] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pages 20503–20521. PMLR, 2022.

[277] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2023.

[278] Pranav Subramani, Nicholas Vadivelu, and Gautam Kamath. Enabling fast differentially private sgd via just-in-time compilation and vectorization. *NeurIPS*, 2021.

[279] Haozhan Sun, Chenchen Xu, and Hanna Suominen. Analyzing the granularity and cost of annotation in clinical sequence labeling. *arXiv preprint arXiv:2108.09913*, 2021.

[280] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337. PMLR, 2017.

[281] Simon Suster, Madhumita Sushil, and Walter Daelemans. Revisiting neural relation classification in clinical notes with external information. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 22–28. Association for Computational Linguistics, 2018.

[282] Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26, 2023.

[283] Liling Tan, Dmitrijs Milajevs, Geeticka Chauhan, Jeremy Gwinnup, and Elijah Rippeth. Proceedings of the 3rd workshop for nlp open source software (nlp-oss 2023). In *Proceedings of Third Workshop for NLP Open Source Software (NLP-OSS)*, 2023.

[284] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

[285] Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=oZtt0pRnOl`.

[286] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

[287] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[288] Volker Tresp and Thomas Briegel. A Solution for Missing Data in Recurrent Neural Networks with an Application to Blood Glucose Prediction. In *Advances in Neural Information Processing Systems*, 1997.

[289] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[290] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[291] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590, 2017.

[292] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

[293] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[294] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[295] Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 534–539. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1065. URL `http://aclweb.org/anthology/N16-1065`.

[296] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35: 33536–33549, 2022.

[297] Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, volume 1, pages 1298–1307, 2016.

[298] Lun Wang, Om Thakkar, and Rajiv Mathews. Unintended memorization in large asr models, and how to mitigate it. In *ICASSP*, 2024.

[299] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of ACM CHIL*, pages 222–235, 2020.

[300] Wei Wang, Xi Yang, Canqun Yang, Xiao-Wei Guo, Xiang Zhang, and Chengkun Wu. Dependency-based long short term memory network for drug-drug interaction extraction. *BMC Bioinformatics*, 18, 2017.

[301] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *CVPR*, 2017.

[302] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[303] Lauren Watson, Eric Gan, Mohan Dantam, Baharan Mirzasoleiman, and Rik Sarkar. Inference and interference: The role of clipping, pruning and loss landscapes in differentially private stochastic gradient descent. *arXiv preprint arXiv:2311.06839*, 2023.

[304] Griffin M Weber, Kenneth D Mandl, and Isaac S Kohane. Finding the missing link for big biomedical data. *Jama*, 311(24):2479–2480, 2014.

[305] Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *ICLR Workshop on Understanding of Foundation Models*, 2024.

[306] Wei-Hung Weng, Kavishwar B Wagholikar, Alexa T McCray, Peter Szolovits, and Henry C Chueh. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making*, 17(1):1–13, 2017.

[307] Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3121–3131, 2019.

[308] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[309] Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association*, 24(3):488–495, 2017.

[310] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[311] Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*, 2015.

[312] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1785–1794, 2015.

[313] Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with

data augmentation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1461–1470. The COLING 2016 Organizing Committee, 2016.

[314] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *International Conference on Information Processing in Medical Imaging*, pages 125–138. Springer, 2019.

[315] Bin Yan and Mingtao Pei. Clinical-Bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

[316] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *CoRR*, abs/1702.01923, 2017.

[317] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *International Conference on Machine Learning*, 2018. URL `http://proceedings.mlr.press/v80/yoon18a.html`.

[318] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *ICLR*, 2022.

[319] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

[320] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[321] Mo Yu, Matthew Gormley, and Mark Dredze. Factor-based compositional embedding models. In *NIPS Workshop on Learning Semantics*, pages 95–101, 2014.

[322] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, 2014.

[323] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1753–1762, 2015.

[324] Ana Cecilia Zenteno, Tim Carnes, Retsef Levi, Bethany J Daily, and Peter F Dunn. Systematic or block allocation at a large academic medical center. *Annals of surgery*, 264(6):973–981, 2016.

[325] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

[326] Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*, 2015.

[327] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[328] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, 2015.

[329] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[330] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[331] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, 2017.

[332] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[333] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215. Association for Computational Linguistics, 2018.

[334] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from

paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[335] Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of healthcare engineering*, 2018, 2018.

[336] Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453, 2016.

[337] Wei Zheng, Hongfei Lin, Ling Luo, Zhehuan Zhao, Zhengguang Li, Yijia Zhang, Zhihao Yang, and Jian Wang. An attention-based effective neural model for drug-drug interactions extraction. In *BMC Bioinformatics*, 2017.

[338] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[339] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212, 2016.

[340] Zhi-Hua Zhou and Ji Feng. Deep forest. *National science review*, 6(1):74–86, 2019.

[341] Indre Zliobaite. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*, 41, 2010.