

# Image Classification with Consistent Supporting Evidence

by

Peiqi Wang

B.S., University of Toronto (2019)

Submitted to the Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author .....  
Department of Electrical Engineering and Computer Science  
January 26, 2022

Certified by .....  
Polina Golland  
Henry Ellis Warren (1894) 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



# Image Classification with Consistent Supporting Evidence

by

Peiqi Wang

Submitted to the Department of Electrical Engineering and Computer Science  
on January 26, 2022, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Computer Science and Engineering

## Abstract

Adoption of machine learning models in healthcare requires end users' trust in the system. Models that provide additional supportive evidence for their predictions promise to facilitate adoption. We define consistent evidence to be both compatible and sufficient with respect to model predictions. We propose measures of model inconsistency and regularizers that promote more consistent evidence. We demonstrate our ideas in the context of edema severity grading from chest radiographs. We demonstrate empirically that consistent models provide competitive performance while supporting interpretation.

Thesis Supervisor: Polina Golland

Title: Henry Ellis Warren (1894) Professor of Electrical Engineering and Computer Science



## Acknowledgments

I would like to thank my advisor Polina Golland for her excellent guidance to this work. Polina has given me tremendous support and provided me with a stimulating environment for learning. Her sensitivity to good research problems and seasoned feedbacks have been invaluable to my project and crucial to my personal growth as a researcher. Her ability to work through problems with clarity is something that I will always aspire to. I also want to thank Polina for being consistently available and ungrudgingly kind along the way. I am truly grateful to have her as my advisor.

I would like to thank members of the Medical Vision group: Danielle Pace, Maz Abulnaga, Ruizhi Liao, Nalini Singh, Clinton Wang, Razvan Marinescu, and Daniel Moyer for providing a productive and fun research environment. I am thankful for the countless delightful conversations as well as the various kinds of help I have received. I would also like to thank collaborators, Seth Berkowitz and Steven Horng, for their cheerful attitude and clinical insights.

I want to thank many of my friends for bringing joy to my every day life, for their companionship during work and play, and for the inspirations and criticisms that have influenced me for the better.

Finally, I owe a great deal of debt to my parents, for their love and sacrifice. They have been an unfaltering source of my optimism and strength.



# Contents

<b>1</b>	<b>Introduction</b>	<b>15</b>
<b>2</b>	<b>Related Work</b>	<b>19</b>
2.1	Interpretable Machine Learning . . . . .	19
2.2	Logical Constraints . . . . .	20
<b>3</b>	<b>Classification with Consistent Evidence</b>	<b>21</b>
3.1	Consistent Evidence . . . . .	22
3.2	Edema Severity Grading Application . . . . .	23
3.3	Measuring Inconsistency . . . . .	25
3.4	Consistency Regularization . . . . .	26
3.5	Optimization . . . . .	28
3.6	Connection with Semantic Loss . . . . .	29
<b>4</b>	<b>Experiments and Results</b>	<b>31</b>
4.1	Implementation Details . . . . .	31
4.2	Data . . . . .	32
4.3	Experiments . . . . .	32
<b>5</b>	<b>Conclusions</b>	<b>39</b>
<b>A</b>	<b>Appendix</b>	<b>43</b>
A.1	Truth Table for Consistency Constraints . . . . .	43
A.2	Corresponding Table for Figure 4-4 and 4-6 . . . . .	44

A.3 Comparing Different Regularizers . . . . . 45



# List of Figures

1-1	Our model provides prediction of disease stage $y$ and supporting evidence. We use $\mathcal{I}(c)$ to denote the set of evidence labels detected in the image that directly support disease stage $c$ . We show examples of inconsistent evidence highlighted in red produced by the baseline model (left column). Our proposed regularizer corrects these mistakes so that predicted evidence label becomes compatible (top right) and sufficient (bottom right). . . . .	15
3-1	Findings that directly support a particular severity level. . . . .	23
3-2	Examples of consistent and inconsistent evidence. . . . .	24
4-1	Model inconsistency across different values of $\hat{y}$ evaluated on the test set $\hat{\mathcal{D}}$ for a naively trained model, i.e., $\omega_1 = \omega_2 = 0$ . The majority of inconsistent evidence comes from incompatible evidence associated with small values of the predicted task label $\hat{y}$ . Roughly, one out of four image will yield incompatible evidence. . . . .	33
4-2	The effect of varying strength of regularizations during training on model consistency. Here $\mathcal{R}_1, \mathcal{R}_2$ is short hand for $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. The proposed regularizers $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$ encourage the model to provide more compatible evidence (left) or more sufficient evidence (middle), respectively. The application of regularizers at the same time encourages the model to provide more consistent evidence (right). . .	34

4-3	Correctly and incorrectly classified test images with supporting evidence given by a consistent model ( $\omega_1 = \omega_2 = 8$ ). We use $\mathcal{I}(c)$ to denote the set of evidence labels detected in the image that directly support disease stage $c$ . . . . .	35
4-4	The effect of regularization on model inconsistency (left 2) and performance of predicting task label $y$ (middle 2) and evidence labels $z$ (right 2). Here $\mathcal{R}_1, \mathcal{R}_2$ is short hand for $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. When regularizing for both $\mathcal{R}_1, \mathcal{R}_2$ , e.g., down the diagonals of the matrix, we notice dramatic decrease in model inconsistency and competitive performance for the regularized model. . . . .	36
4-5	The effect of varying strength of regularizations using semantic loss during training on model consistency. Here $\mathcal{R}_1, \mathcal{R}_2$ is short hand for $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. It is not possible to keep both $\mathcal{R}_1(\hat{\mathcal{D}})$ and $\mathcal{R}_2(\hat{\mathcal{D}})$ down to small values at the same time. . . . .	37
4-6	The effect of regularization on model inconsistency and performance of predicting task label $y$ and evidence label $z$ , where the models predict $y$ via the $z$ bottleneck. * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. We observe similar pattern in inconsistency and performance when compared to that of models that predict $(y, z)$ jointly. . . . .	38
A-1	The effect of varying strength of regularization using $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$ during training on model consistency. . . . .	46
A-2	The effect of varying strength of regularization using $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$ on model inconsistency (left 2) and performance of predicting task label $y$ (middle 2) and evidence labels $z$ (right 2). . . . .	46
A-3	The effect of varying strength of regularization using $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$ during training on model consistency. . . . .	47

A-4	The effect of varying strength of regularization using $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$ on model inconsistency (left 2) and performance of predicting task label $y$ (middle 2) and evidence labels $z$ (right 2). . . . .	47
A-5	The effect of varying strength of regularization using $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$ during training on model consistency. . . . .	48
A-6	The effect of varying strength of regularization using $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$ on model inconsistency (left 2) and performance of predicting task label $y$ (middle 2) and evidence labels $z$ (right 2). . . . .	48
A-7	The effect of varying strength of regularization using $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$ during training on model consistency. . . . .	49
A-8	The effect of varying strength of regularization using $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$ on model inconsistency (left 2) and performance of predicting task label $y$ (middle 2) and evidence labels $z$ (right 2). . . . .	49
A-9	The effect of varying strength of regularization using $L^\alpha(x, \theta)$ during training on model consistency. . . . .	50



# List of Tables

- A.1 A list of satisfying instantiations in truth table of consistency constraints  $\alpha$  for the edema severity grading task. For brevity, we do not specify values that  $z_1, \dots, z_3$  take for the  $y = 2$  partition. This means  $z_1, \dots, z_3$  are unconstrained, and there are  $2^3$  possible instantiations for any given row within the  $y = 2$  partition. Similarly, there are  $2^5$  possible instantiations for any given row within the  $y = 3$  partition. . . . . 43
  
- A.2 The effect of regularization on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ , where the models predict  $(y, z)$  jointly. \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds. We notice dramatic decrease in model inconsistency and competitive performance for the regularized model. . . . . 44
  
- A.3 The effect of regularization on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ , where the models predict  $y$  via the  $z$  bottleneck. \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds. We observe similar pattern in inconsistency and performance when compared to that of models that predict  $(y, z)$  jointly. . . . . 44

A.4	The effect of regularization using $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$ on model inconsistency and performance of predicting task label $y$ and evidence label $z$ . * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. . . . .	46
A.5	The effect of regularization using $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$ on model inconsistency and performance of predicting task label $y$ and evidence label $z$ . * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. . . . .	47
A.6	The effect of regularization using $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$ on model inconsistency and performance of predicting task label $y$ and evidence label $z$ . * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. . . . .	48
A.7	The effect of regularization using $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$ on model inconsistency and performance of predicting task label $y$ and evidence label $z$ . * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. . . . .	49
A.8	The effect of regularization using $L^\alpha(x, \theta)$ on model inconsistency and performance of predicting task label $y$ and evidence label $z$ . * stands for baseline model that is trained to predict task label $y$ . We compute average numbers with $\pm 2$ standard deviations over random seeds. . .	50

# Chapter 1

## Introduction


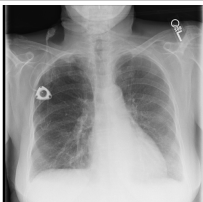
Input	Unregularized	Regularized
	Prediction: $\hat{y} = 0$ Evidence: <ul style="list-style-type: none"><li>• <math>\mathcal{I}(1)</math> : NULL</li><li>• <math>\mathcal{I}(2)</math> : septal lines, interstitial abnormality</li><li>• <math>\mathcal{I}(3)</math> : NULL</li></ul> <b>Inconsistent</b>	Prediction: $\hat{y} = 0$ Evidence: <ul style="list-style-type: none"><li>• <math>\mathcal{I}(1)</math> : NULL</li><li>• <math>\mathcal{I}(2)</math> : NULL</li><li>• <math>\mathcal{I}(3)</math> : NULL</li></ul> Consistent
	Prediction: $\hat{y} = 2$ Evidence: <ul style="list-style-type: none"><li>• <math>\mathcal{I}(1)</math> : hilar congestion</li><li>• <math>\mathcal{I}(2)</math> : NULL</li><li>• <math>\mathcal{I}(3)</math> : NULL</li></ul> <b>Inconsistent</b>	Prediction: $\hat{y} = 2$ Evidence: <ul style="list-style-type: none"><li>• <math>\mathcal{I}(1)</math> : hilar congestion</li><li>• <math>\mathcal{I}(2)</math> : interstitial abnormality</li><li>• <math>\mathcal{I}(3)</math> : NULL</li></ul> Consistent

Figure 1-1: Our model provides prediction of disease stage  $y$  and supporting evidence. We use  $\mathcal{I}(c)$  to denote the set of evidence labels detected in the image that directly support disease stage  $c$ . We show examples of inconsistent evidence highlighted in red produced by the baseline model (left column). Our proposed regularizer corrects these mistakes so that predicted evidence label becomes compatible (top right) and sufficient (bottom right).

Identifying radiological findings and inferring disease stages from medical images is common in clinical practice. Many models make predictions without explaining the conclusion. In contrast, human experts often provide specific explanation based on prior knowledge of human physiology to support their image-based diagnosis. We aim to build models that are transparent in the reasoning process, at an appropriate level of understanding consumable by end users, e.g., clinicians. What additional

information should a machine learning model provide to gain the trust of its end users? We propose a solution motivated by an example of how radiologists themselves operate.

Radiological *findings* are concepts determined as useful by radiologists. Findings include image features, pathological states, and observations about the underlying physiology [12]. The radiologists aggregate the findings to provide an overall interpretation of the image. They support the eventual diagnosis by providing an account of the identified findings based on prior knowledge of relationships between findings and the patient’s physiological state. We propose and demonstrate an approach that re-capitulates the reasoning process of domain experts. In addition to primary predictions, the model provides supporting evidence, i.e., findings, deemed useful by the end users.

It is critical that predictions and supporting evidence are consistent with each other. In practice, radiologists cannot draw their conclusions based on incompatible evidence, nor could they support their conclusions with insufficient evidence. Similarly, end users will question the credibility of a model when its predictions and accompanying evidence are incompatible or insufficient.

In this thesis, we build explainable models that supplement their predictions with consistent supporting evidence, illustrated in Figure 1-1. We define measures of inconsistency between the model’s primary output and its supporting evidence and propose simple regularizers that encourage the classifier to be more consistent. We demonstrate that we can train consistent models without loss in performance in the context of pathology grading from a chest radiograph. We show our method is fairly flexible and agnostic to how predicted task label and evidence labels are computed.

In Chapter 2, we discuss previous work on interpretable machine learning as well as methodologies that represent and enforce logical statements. In Chapter 3, we formulate the consistent evidence problem, motivated by the edema severity grading application. In addition, we discuss our approach to measure and enforce model consistency via regularization. We compare our approach to semantic loss, a previous method that enforce logical constraints. In Chapter 4, we conduct experiments that



demonstrate the effectiveness of our proposed method. We show that our method is more flexible than semantic loss and can be integrated with an alternative method that computes predictions and evidence. Finally in Chapter 5, we conclude with a summary of our contributions, interpretations of our findings, and their implications for practical use and future work.



# Chapter 2

## Related Work

### 2.1 Interpretable Machine Learning

While model interpretability is an important topic in machine learning, few methods take the end users' needs into account. For example, some method localize image regions important for a prediction [24, 30], but fail to express what properties of the image region are associated with the model output. Others aim to use a simpler model [4] or to approximate the behavior of a complex model with a simpler one [23]. While effective for handling low dimensional tabular data where the covariates are physically meaningful attributes, such methods are less useful for extremely high dimensional imaging data. Our work provides clinically meaningful supporting evidence useful to end users of the system, rather than support the developers' understanding of how the model reaches its decision.

Another approach is to train a classifier whose predictions rely on higher-level concepts. Unsupervised methods can make for a more interpretable model for general purpose tasks but cannot take advantage of strong domain knowledge ubiquitous in healthcare [1]. Alternatively, concept bottleneck models that learn concepts with supervision have been applied to arthritis grading [18], retinal disease classification [7], and other applications [3, 21]. This strategy relies on the appropriate choice of the concepts to maintain good performance.

Alternatively, some prior methods focus on learning a mapping to the product

space of the task label and supporting evidence [6, 14], with application to text classification [28, 29]. In contrast, we learn a structured output where known relationships between predictions and evidence are enforced. We inject domain specific knowledge and require our model to provide supporting evidence that is clinically feasible under a specific prediction.

For exposition, we assume predictions and evidence are separate outputs of the model. Our proposed method is independent of how the task label and evidence labels are computed, and applies equally to models which make predictions via an evidence bottleneck and to models which outputs predictions and evidence jointly.

A recently demonstrated unsupervised strategy that requires a forward model that relates supporting evidence to a subset of the input features is also relevant [22]. This method can be difficult to implement in our radiograph grading task as it assumes knowledge of an accurate forward model, from supporting evidence to a high dimensional image, which is infeasible for most medical imaging problems.

## 2.2 Logical Constraints

There are multiple ways to represent symbolic constraints. For example, trees have been used to express subsumption relationships between attributes, e.g., hierarchical annotation of medical images [9]. Hierarchical multi-label learning aims to enforce such constraints [2, 11, 25, 27]. Unfortunately, trees are overly restrictive and cannot express consistency constraints that are important in our application.

Alternatively, boolean statements can be quite expressive in representing logical constraints. Logical constraints on model outputs can be enforced by replacing logical operators with their subdifferentiable fuzzy t-norms [8, 19] or through the use of specialized loss functions [26]. Our approach to representing and enforcing logical constraints is easier to interpret, simpler to implement, and more flexible.

# Chapter 3

## Classification with Consistent Evidence

In this chapter, we define consistency of evidence and introduce measures of inconsistency. We provide an example application to ground our definitions. We construct novel loss functions that encourage consistency and discuss optimization that arises when training classifiers with consistent supporting evidence.

Let  $x, y$ , and  $z = (z_1, \dots, z_K)$  be random variables representing an image, a  $C$ -class task label, and  $K$  binary evidence labels. Let  $\mathcal{D}_t$  be a data set that includes pairs  $(x, y)$  and  $(x, z_k)$  for  $k = 1, \dots, K$ . In this work, we do not assume full tuples  $(x, y, z_1, \dots, z_K)$  are available. Learning joint predictors from available sub-tuples is an interesting direction that is outside the scope of this thesis. Moreover, we allow the same image  $x$  to be included as part of several different pairs in the data set  $\mathcal{D}_t$ .

We use the training data set  $\mathcal{D}_t$  to build probabilistic classifiers  $p(y | x; \theta)$  and  $p(z_k | x; \theta)$  for  $k = 1, \dots, K$ . The maximum a posteriori (MAP) estimates of the task label  $y$  and evidence labels  $z$  are obtained via

$$\hat{y} = \arg \max_{c \in [C]} p(y = c | x; \theta), \quad (3.1)$$

$$\hat{z}_k = \arg \max_{a \in \{-1, +1\}} p(z_k = a | x; \theta), \quad k = 1, \dots, K. \quad (3.2)$$

We use  $\hat{y}(x)$  to express dependence of the predicted label  $\hat{y}$  on the input image  $x$ .

### 3.1 Consistent Evidence

We assume that domain experts provide domain specific knowledge in the form of logical constraints between the task label  $y$  and the evidence labels  $z$ . We identify two major logical constraints that are important in our application, specifically that supporting evidence should be compatible and sufficient with the task label.

Let  $\mathcal{I}_1 : [C] \rightarrow \mathcal{P}([K])$  be the indexing function for evidence that is incompatible with a particular value of task label, where we use  $\mathcal{P}(\cdot)$  to denote the power set. Specifically, if evidence labels  $\{z_{i_1}, \dots, z_{i_M}\}$  are incompatible with task label  $y = c$ , then  $\mathcal{I}_1(c) = \{i_1, \dots, i_M\}$ . Let  $\mathcal{I}_2 : [C] \rightarrow \mathcal{P}([K])$  be the indexing function for evidence that directly supports a particular value of task label. We assume that  $\mathcal{I}_1$  and  $\mathcal{I}_2$  are provided by domain experts.

**Definition 1.** (*Consistent Evidence*) *The task label  $y \in [C]$  and the evidence label vector  $z = (z_1, \dots, z_K) \in \{-1, +1\}^K$  are consistent if*

$$\forall k \in \mathcal{I}_1(y) : z_k = -1, \tag{3.3}$$

$$\exists k \in \mathcal{I}_2(y) : z_k = +1. \tag{3.4}$$

The first criterion specifies that no evidence is incompatible with the task label  $y$ . The second criterion specifies that there should be at least one direct evidence label present that supports the task label  $y$ .

In reality, perfectly consistent evidence may not be necessary or possible. For example, domain experts often specify constraints with a notion of uncertainty, e.g.,  $\mathcal{I}_1(y)$  is incompatible with  $y$  most of the time except for occasional corner cases. In addition, certain direct evidence might be so rare that it becomes impossible to include it in  $\mathcal{I}_2(y)$ . Therefore, it is perfectly sensible that there is no direct evidence present in some cases, if we have not included the corresponding evidence label in the construction. This motivates us to consider these constraints in probabilistic terms.

Definition 1 is a specification over the values that random variables can take. The same definition applies to the true data distribution  $\{y, z\}$  and to the predicted distribution  $\{\hat{y}, \hat{z}\}$ . In practice, we construct training data  $\{y, z\}$  to be perfectly consistent and demonstrate a training method that encourages the model outputs  $\{\hat{y}, \hat{z}\}$  to be consistent as well.

Note that we are not restricted to predicting the findings in  $\cup_{c \in [C]} \mathcal{I}_2(c)$ . If some findings provide useful information but are not directly supportive, they can still be included in the set of evidence labels.

## 3.2 Edema Severity Grading Application

This section illustrates the construction of indexing functions  $\mathcal{I}_1$  and  $\mathcal{I}_2$  for the pulmonary edema grading task that motivated our work.

Pulmonary edema is defined as an abnormal accumulation of fluid in the lungs. Higher hydrostatic pressure in the vasculature causes more severe symptoms. Typically, radiologists grade the severity of edema based on findings that are typical of the most severe stage of pulmonary edema [12].

We use a categorization that identifies four edema severity levels, in order of increasing severity: *no edema* (0), *mild edema* (1), *moderate edema* (2), and *severe edema* (3) [20, 15]. The edema severity grading task involves assigning a severity level  $y \in \{0, 1, 2, 3\}$  to a test image. In this task, there are 4 classes, i.e.,  $C = 4$ .

Severity $y$	Findings $\mathcal{I}_2(y)$
0 (none)	-
1 (mild)	vascular congestion hilar congestion peribronchial cuffing
2 (moderate)	septal lines interstitial abnormality
3 (severe)	air bronchograms parenchymal opacity

Figure 3-1: Findings that directly support a particular severity level.

Severity $y$	Evidence $z$
1	peribronchial cuffing
2	vascular congestion septal lines interstitial abnormality

(a) Examples of consistent evidence.

Severity $y$	Evidence $z$
1	hilar congestion septal lines
2	vascular congestion interstitial abnormality air bronchograms
1	–
3	septal lines

(b) Examples of inconsistent evidence. First two examples are incompatible. The latter two examples are insufficient.

Figure 3-2: Examples of consistent and inconsistent evidence.

In our work, we identify  $K = 7$  supporting evidence labels deemed useful by clinicians, as shown in Table 3-1. They are canonical radiological manifestation of the underlying pathology. End users expect presence of these findings to be indicative of a specific edema severity level.

As an example, radiologists grade an image as *moderate edema* if they observe *septal lines* (short parallel lines at the periphery of the lung) or *interstitial abnormality* (excess fluids in the supporting tissue within the lung). Note that presence of evidence from a lower value of edema severity is not inconsistent. For example, radiologists may at the same time observe presence of *vascular congestion* (enlargement of pulmonary veins) and *septal lines* in a moderate edema case.

In the severity grading task, we consider an evidence label as incompatible if its presence directly supports a higher level severity level. Thus define

$$\mathcal{I}_1(c) = \bigcup_{c' > c} \mathcal{I}_2(c'). \quad (3.5)$$

As an example, a model that grades an image as *moderate edema* should not use *air bronchograms* (opacification of alveoli) as supporting evidence.

We consider evidence as insufficient when no direct evidence for edema severity grading is present. As an example, a model which grades an image as *severe edema*



cannot rely on *septal lines* only to support its prediction.

Tables 3-2a and 3-2b illustrate further examples of consistent and inconsistent evidence, respectively.

### 3.3 Measuring Inconsistency

We quantify the inconsistency probabilistically based on Definition 1. First, we define a measure of incompatibility as the probability that there is an incompatible evidence label

$$\mathbb{P} \left[ \bigcup_{k \in \mathcal{I}_1(y)} \{z_k = +1\} \right]. \quad (3.6)$$

To facilitate computation, we upper bound this probability using union bound by

$$\mathcal{R}_1^+(y, z) = \sum_{k \in \mathcal{I}_1(y)} \mathbb{P}[z_k = +1]. \quad (3.7)$$

We provide an estimate of incompatibility over data set  $\mathcal{D}$  by taking expectation over its empirical distribution

$$\mathcal{R}_1^+(\mathcal{D}) = \mathbb{E}_{(y,z) \sim \mathcal{D}} \left[ \sum_{k \in \mathcal{I}_1(y)} \mathbb{1}[z_k = +1] \right], \quad (3.8)$$

where we have replaced  $\mathbb{P}[z_k = +1]$  with  $\mathbb{1}[z_k = +1]$  since  $z_k$  is binary valued. Intuitively,  $\mathcal{R}_1^+(\mathcal{D})$  is the average count of evidence labels incompatible with the task label.

Equation 3.8 depends on the size of  $\mathcal{I}_1(y)$ , implying that the amount of incompatibility is dependent on the sample point. We can eliminate the dependency on the size of  $\mathcal{I}_1(y)$  by defining

$$\mathcal{R}_1(\mathcal{D}) = \mathbb{E}_{(y,z) \sim \mathcal{D}} [\mathcal{R}_1(y, z)] \quad \text{where} \quad \mathcal{R}_1(y, z) = \max_{k \in \mathcal{I}_1(y)} \mathbb{1}[z_k = +1] \quad (3.9)$$

In our application, we care more about existence of incompatible evidence over the degree to which there is incompatible evidence. For this reason, we prefer Equation 3.9 as a measure of incompatible evidence.

Similarly, we define a measure of insufficiency as the probability that there is no sufficient evidence

$$\mathbb{P} \left[ \bigcap_{k \in \mathcal{I}_2(y)} \{z_k = -1\} \right], \quad (3.10)$$

which leads to an upper bound

$$\mathcal{R}_2(y, z) = \min_{k \in \mathcal{I}_2(y)} \mathbb{P} [z_k = -1] \quad (3.11)$$

and its empirical estimate

$$\mathcal{R}_2(\mathcal{D}) = \mathbb{E}_{(y,z) \sim \mathcal{D}} \left[ \min_{k \in \mathcal{I}_2(y)} [1 - \mathbb{1} [z_k = +1]] \right] \quad (3.12)$$

$$= 1 - \mathbb{E}_{(y,z) \sim \mathcal{D}} \left[ \max_{k \in \mathcal{I}_2(y)} \mathbb{1} [z_k = +1] \right]. \quad (3.13)$$

Note that  $\mathcal{R}_2(\mathcal{D})$  is the average count of absence of direct evidence.

Now we can provide an upper bound on probability of inconsistent evidence  $\mathcal{R}(y, z) = \mathcal{R}_1(y, z) + \mathcal{R}_2(y, z)$  and its empirical estimate  $\mathcal{R}(\mathcal{D}) = \mathcal{R}_1(\mathcal{D}) + \mathcal{R}_2(\mathcal{D})$ .

### 3.4 Consistency Regularization

Models trained naively to predict labels  $y$  and  $z$  jointly are not guaranteed to be consistent. Here, we provide regularizers that encourage supporting evidence to be more consistent.

Observe that Equations 3.7 and 3.11 are upper bounds on the true probability of model being inconsistent. We can simply use these upper bounds, or modification thereof, as regularizers. We opt to use cross entropy to create regularizers consistent with the classification framework.

To penalize incompatibility, we define

$$\overline{\mathcal{R}}_1^+(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \sum_{k \in \mathcal{I}_1(\hat{y}(x))} \ln p(z_k = -1 \mid x; \theta) \right]. \quad (3.14)$$

Intuitively,  $\overline{\mathcal{R}}_1^+(\theta)$  penalizes evidence probability that is incompatible with the predicted task label. Including  $\overline{\mathcal{R}}_1^+(\theta)$  in the loss function is equivalent to supplying pseudo negative samples for evidence obtained from the predicted task label  $\hat{y}$ .

Instead of penalizing incompatibility with respect to MAP estimate of the task label  $\hat{y}(x)$ , we can penalize incompatibility for each value of task label weighted by the posterior probability, i.e.,

$$\mathcal{R}_1^+(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \sum_{c \in [C]} \sum_{k \in \mathcal{I}_1(c)} p(y = c \mid x; \theta) \ln p(z_k = -1 \mid x; \theta) \right]. \quad (3.15)$$

In contrast to Equation 3.14 where gradients cannot flow through  $\hat{y}(x)$  due to the arg max operator, Equation 3.15 provides a softer regularizer that affects the predictions of both the task and the evidence labels.

Similar to Equation 3.9, we can penalize incompatibility agnostic to the size of  $\mathcal{I}_1(y)$  by defining

$$\overline{\mathcal{R}}_1(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \ln \left[ 1 - \max_{k \in \mathcal{I}_1(\hat{y}(x))} p(z_k = +1 \mid x; \theta) \right] \right], \quad (3.16)$$

$$\mathcal{R}_1(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \sum_{c \in [C]} p(y = c \mid x; \theta) \ln \left[ 1 - \max_{k \in \mathcal{I}_1(c)} p(z_k = +1 \mid x; \theta) \right] \right]. \quad (3.17)$$

Intuitively,  $\overline{\mathcal{R}}_1(\theta)$  is equivalent to supplying pseudo negative samples for evidence that is predicted to be the most incompatible with the predicted task label  $\hat{y}$ .

Similarly, we define

$$\overline{\mathcal{R}}_2(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \ln \max_{k \in \mathcal{I}_2(\hat{y}(x))} p(z_k = +1 \mid x; \theta) \right]. \quad (3.18)$$

Intuitively,  $\overline{\mathcal{R}}_2(\theta)$  encourages presence of *some* evidence to the support predicted task

label. Including  $\overline{\mathcal{R}}_2(\theta)$  in the loss function is equivalent to supplying pseudo positive samples obtained from the predicted task label  $\hat{y}$ .

Similar to Equation 3.17, we can penalize insufficiency using posterior probability as weights,

$$\mathcal{R}_2(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[ \sum_{c \in [C]} p(y = c \mid x; \theta) \ln \max_{k \in \mathcal{I}_2(c)} p(z_k = +1 \mid x; \theta) \right]. \quad (3.19)$$

In the main text, we focus on regularizers  $\mathcal{R}_1(\theta)$  and  $\mathcal{R}_2(\theta)$ . We supply additional results based on other proposed regularizers in Appendix A.3.

### 3.5 Optimization

There are different approaches to train the classifiers to predict task label  $y$  and evidence labels  $z_1, \dots, z_K$ . In our experiments, we apply deep multitask learning for joint predictions of  $y, z_1, \dots, z_K$ . In particular, we parameterize  $p(y \mid x; \theta)$  and  $p(z_k \mid x; \theta)$  for  $k = 1, \dots, K$  with neural network  $f(x; \theta)$  and assume function  $f$  outputs logits over  $K + 1$  marginals.

Given a classification loss function  $L(\cdot, \cdot)$ , the objective is the empirical risk,

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_t} [L(y, f(x; \theta))] + \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{(x,z_k) \sim \mathcal{D}_t} [L(z_k, f(x; \theta))]. \quad (3.20)$$

We add consistency regularization to multitask classification loss, which yields a regularized empirical risk minimization problem

$$\min_{\theta} \mathcal{L}(\theta) + \omega_1 \mathcal{R}_1(\theta) + \omega_2 \mathcal{R}_2(\theta), \quad (3.21)$$

where  $\omega_1, \omega_2 \in \mathbb{R}^+$  are coefficients that control the degree of regularization.

### 3.6 Connection with Semantic Loss

In this section, we show that our proposed soft regularizer upper bounds the average semantic loss [26] for the consistent evidence constraints.

As briefly mentioned in Chapter 2, semantic loss function is an alternative method to enforce logical constraints over neural network outputs [26]. Given consistency constraints  $\alpha$  over variables  $(y, z)$  detailed in Definition 1 and marginal probabilities of network’s output, the semantic loss is

$$L^\alpha(x, \theta) = -\ln p(\alpha \mid x; \theta), \quad (3.22)$$

where the probability that the constraints are satisfied is defined as

$$p(\alpha \mid x; \theta) = \sum_{(y,z) \models \alpha} \left[ \prod_{c \in [C]} p(y = c \mid x; \theta)^{[y=c]} \cdot \prod_{k \in [K]} \prod_{a \in \{-1, +1\}} p(z_k = a \mid x; \theta)^{[z_k=a]} \right]. \quad (3.23)$$

Here, we use  $(y, z) \models \alpha$  to denote all instantiations  $(y, z)$  that satisfy the constraints  $\alpha$ . We use a natural generalization of the original formulation of the semantic loss function to account for variable  $y$  taking non-binary values.

Semantic loss is invariant to how the constraints are specified because Equation 3.23 is a sum over probabilities of all satisfying instantiations over the truth table of  $\alpha$ . Table A.1 provides an example of a truth table of  $\alpha$  for the edema severity grading task. By looking at the truth table, we can simplify the expression considerably,

$$p(\alpha \mid x; \theta) = \sum_{c \in [C]} \left[ p(y = c \mid x; \theta) \prod_{k \in \mathcal{I}_1(c)} p(z_k = -1 \mid x; \theta) \left[ 1 - \prod_{k \in \mathcal{I}_2(c)} p(z_k = -1 \mid x; \theta) \right] \right]. \quad (3.24)$$

The term involving  $\mathcal{I}_1$  measures incompatibility while the term involving  $\mathcal{I}_2$  measures

insufficiency of the network’s output probabilities.

We demonstrate that soft consistency regularizer in Equation 3.15 and 3.19 for a single input image  $x$

$$\bar{\mathcal{R}}^+(x, \theta) = - \sum_{c \in [C]} p(y = c | x; \theta) \ln \left[ \prod_{k \in \mathcal{I}_1(c)} p(z_k = -1 | x; \theta) \max_{k \in \mathcal{I}_2(c)} p(z_k = +1 | x; \theta) \right] \quad (3.25)$$

upper bounds the semantic loss

$$\bar{\mathcal{R}}^+(x, \theta) \geq - \sum_{c \in [C]} p(y = c | x; \theta) \ln \left[ \prod_{k \in \mathcal{I}_1(c)} p(z_k = -1 | x; \theta) \left[ 1 - \prod_{k \in \mathcal{I}_2(c)} p(z_k = -1 | x; \theta) \right] \right] \quad (3.26)$$

$$\geq L^\alpha(x, \theta). \quad (3.27)$$

The first inequality follows from the fact that  $\max_i p_i \leq 1 - \prod_i (1 - p_i)$  for  $(p_1, \dots, p_n)$  satisfying  $0 \leq p_i \leq 1$ . The second inequality is implied by Jensen’s inequality.

Considering this connection, we expect the two approaches to behave similarly in enforcing consistency constraints. However, there are crucial differences between them. In its original form, semantic loss weighs the two constraints equally, whereas our proposed regularizers can prioritize one constraints over the other by tuning the corresponding coefficients.

# Chapter 4

## Experiments and Results

### 4.1 Implementation Details

We use residual networks to parameterize our probabilistic classifiers [13]. The network is modified to output a  $(C + K)$ -dimensional vector representing the posterior marginal probabilities for  $y, z_1, \dots, z_K$ .

We use weighted cross entropy loss as  $L(\cdot, \cdot)$  to handle class imbalances. We employ the Adam optimizer with a constant learning rate of  $2 \cdot 10^{-4}$  with mini-batch size of 32 for stochastic optimization of network parameters [17]. Each gradient update involves random sampling of a label (task or evidence), assembling a mini-batch of data corresponding to the sampled label, computing the objective function, and updating parameters with backpropagated gradients. This approach enables us to learn even if some labels are missing for some images.

We normalize images to zero mean and unit variance and resize them to 224x224 pixels. We apply random image augmentations to images, e.g., crop, horizontal flip, brightness and contrast variations, to alleviate model overfitting.

We implement Equation 3.16, Equation 3.17, Equation 3.18, and Equation 3.19 by substituting the max operator with a soft maximum operator, i.e.,  $\text{LSE}_{i \in [n]}(x_i) = \log \sum_i \exp(x_i)$ . This way, we enforce sufficiency of evidence by upscaling probabilities of direct evidence that are larger to begin with.

To make fair comparison with the semantic loss function, we use the one-hot

representation for the task label  $y$  via the softmax function, instead of enforcing the exactly-one representation with an additional constraint [26].

We use exactly the same network architecture, data augmentation, and optimization parameters to isolate the impact of the proposed regularization on consistency and performance.

We compute mean and standard deviation statistics for inconsistency and test prediction from 3 runs with different random seed.

## 4.2 Data

We use a subset of 238,086 frontal-view chest X-ray from the MIMIC-CXR data set [16]. We split the data set into training (217,016), validation (10,445), and test (10,625) sets randomly. The performance of predicted evidence is computed over this test set. There is no patient overlap between training, validation and test sets.

Edema severity labels are extracted from associated reports by searching for keywords that are indicative of a specific disease stage. The 7,802 labeled image/report pairs are split into training (6,656), validation (648), and test (498) set. The test set was corrected for keyword matching errors by an expert radiologist, as detailed in prior work [5]. We use  $\hat{\mathcal{D}}$  to denote this test set that includes images and predicted labels  $(x, \hat{y}, \hat{z})$ . All subsequent evaluations of model consistency and performance is computed using  $\hat{\mathcal{D}}$ .

## 4.3 Experiments

### Model Inconsistency

We examine model inconsistency overall and over partitions of data with respect to values of predicted label  $\hat{y}$ . The sum of model inconsistency over the partitions gives the quantities  $\mathcal{R}_1(\hat{\mathcal{D}})$  in Equation 3.9 and  $\mathcal{R}_2(\hat{\mathcal{D}})$  in Equation 3.13.

Figure 4-1 reports model inconsistency over partitions of  $\hat{y}$  for a model that is trained without consistency regularization, i.e.,  $\omega_1 = \omega_2 = 0$ . We observe that  $\mathcal{R}_1(\hat{\mathcal{D}})$



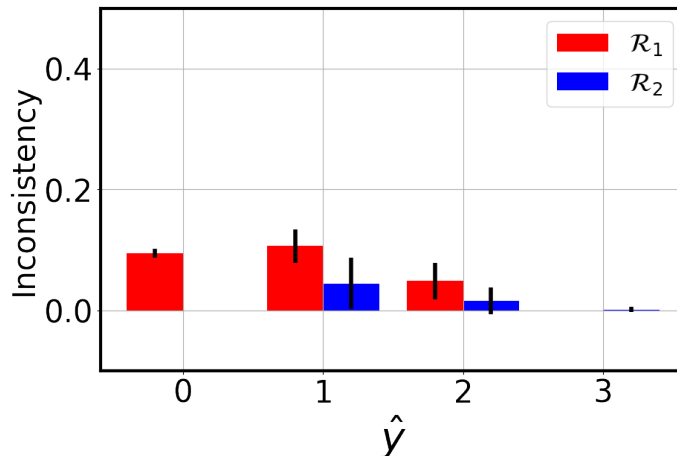


Figure 4-1: Model inconsistency across different values of  $\hat{y}$  evaluated on the test set  $\hat{\mathcal{D}}$  for a naively trained model, i.e.,  $\omega_1 = \omega_2 = 0$ . The majority of inconsistent evidence comes from incompatible evidence associated with small values of the predicted task label  $\hat{y}$ . Roughly, one out of four image will yield incompatible evidence.

is typically larger than  $\mathcal{R}_2(\hat{\mathcal{D}})$  due to the fact that compatibility is an intrinsically harder constraint than sufficiency for our task. We also observe a downward trend in values of  $\mathcal{R}_1(\hat{\mathcal{D}})$  with increasing values for  $\hat{y}$ . This is reasonable as there are many ways to create conflicting evidence for a small value of  $\hat{y}$ , while there is no way to provide conflicting evidence when  $\hat{y} = 3$ .

## Consistency Regularization

To demonstrate that the proposed regularization promotes model consistency, we vary values of  $\omega_1, \omega_2$  in the objective function and train multiple models. We select the most accurate model on the validation set and compute inconsistency on the test set  $\hat{\mathcal{D}}$ .

Figure 4-2 demonstrates the effects of regularization on model consistency. We observe that the regularizers  $\mathcal{R}_1(\theta)$  and  $\mathcal{R}_2(\theta)$  are effective in reducing the respective intended model inconsistency, indicated by a reduction of  $\mathcal{R}_1(\hat{\mathcal{D}})$  in Figure 4-2a and  $\mathcal{R}_2(\hat{\mathcal{D}})$  in Figure 4-2b respectively. Additionally, we observe that penalizing  $\mathcal{R}_1(\hat{\mathcal{D}})$  inadvertently makes  $\mathcal{R}_2(\hat{\mathcal{D}})$  larger and vice versa. This makes intuitive sense, since a model that is more likely to predict absence of evidence will be (i) less likely to

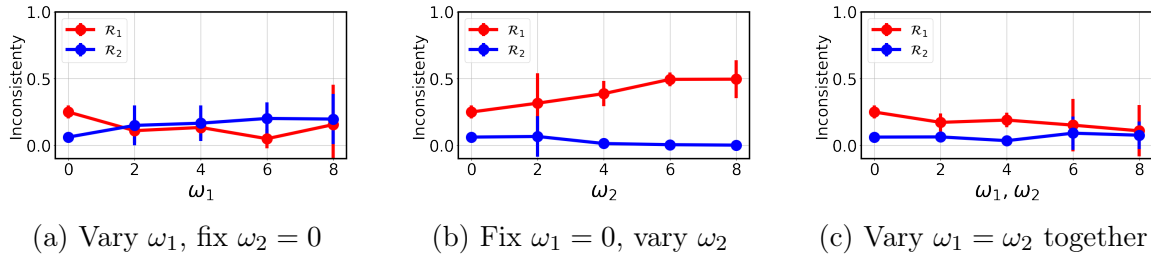


Figure 4-2: The effect of varying strength of regularizations during training on model consistency. Here  $\mathcal{R}_1, \mathcal{R}_2$  is short hand for  $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. The proposed regularizers  $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$  encourage the model to provide more compatible evidence (left) or more sufficient evidence (middle), respectively. The application of regularizers at the same time encourages the model to provide more consistent evidence (right).

provide incompatible evidence and (ii) less likely to provide some direct evidence. We observe that we can reduce both types of inconsistency by regularizing with both loss terms, as shown in Figure 4-2c.

It important to note that even though  $\mathcal{R}_2(\hat{\mathcal{D}})$  is relatively small in models trained with  $\omega_1 = 0$ , regularizing with  $\mathcal{R}_2(\hat{\mathcal{D}})$  is necessary as we want to avoid situations in Figure 4-2a where  $\mathcal{R}_2(\hat{\mathcal{D}})$  becomes intolerably large.

## Interpretability

Figure 4-3 illustrates how a consistent model (trained with  $\omega_1 = \omega_2 = 8$ ) provides supporting evidence for randomly sampled test images. We provide correctly and incorrectly classified test images for each severity level. We observe that the regularized model provides consistent evidence in all 8 examples, even in cases where model prediction of task label is not correct.

How does providing consistent supporting evidence build trust in the model? We note that supporting findings are already described in radiological reports and can be easily mined for training and verified in an image. When the supporting evidence is clearly correct, it builds additional trust in the predicted task label. When the consistent but wrong evidence is presented, it is easy to see in the image and helps the end users understand why the main task label is wrong. Our method avoids confusions that arise from model providing inconsistent evidence.

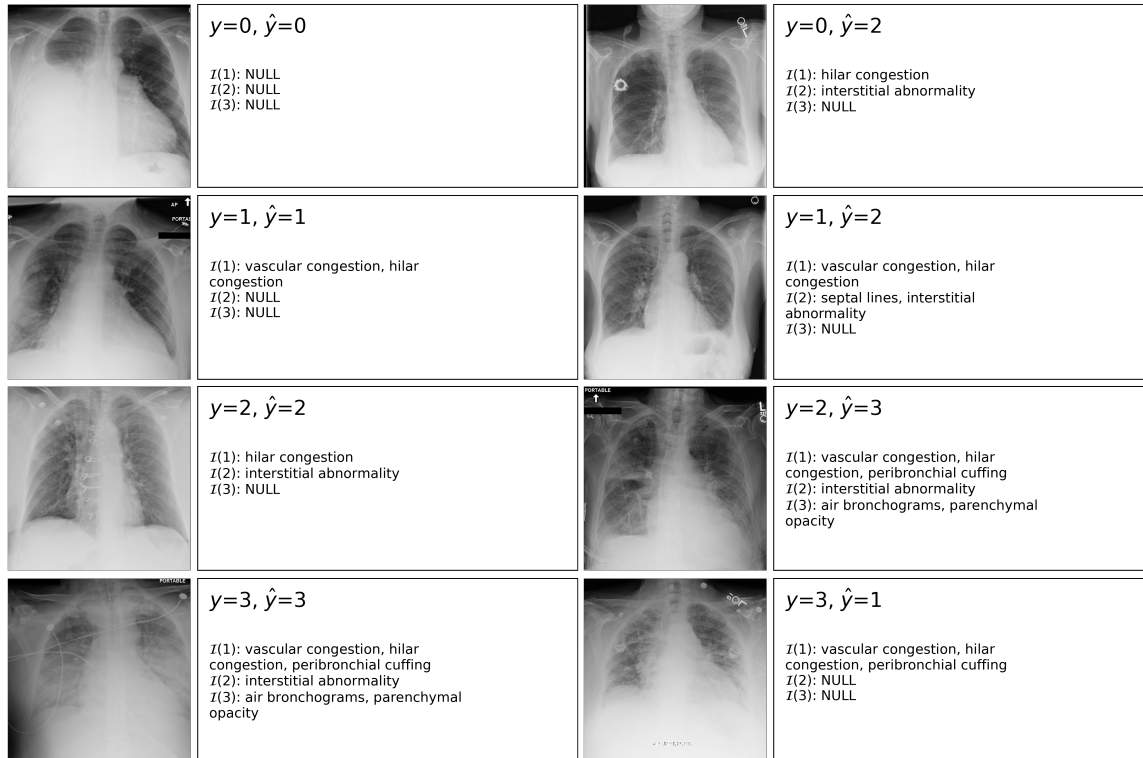


Figure 4-3: Correctly and incorrectly classified test images with supporting evidence given by a consistent model ( $\omega_1 = \omega_2 = 8$ ). We use  $\mathcal{I}(c)$  to denote the set of evidence labels detected in the image that directly support disease stage  $c$ .

Crucially, evidence labels should not only be consistent, but also correct. To this point, we reported performance of evidence detection in Section 4.3. Our proposed regularizers offers a complementary tool to help end users understand why the model erred. Our method can be integrated with technologies, e.g., RCNN, GradCAM [10, 24], that provide localization, i.e., confirmation that the model is focusing on the correct regions in the image.

## Performance-Consistency Tradeoff

Next, we show that we can achieve good model consistency without compromising predictive performance. We vary  $\omega_1, \omega_2$  together in the objective function and train multiple models. We select for the most accurate model on the validation set for subsequent evaluations.

Figure 4-4 demonstrates that we can ensure satisfactory model consistency. At the

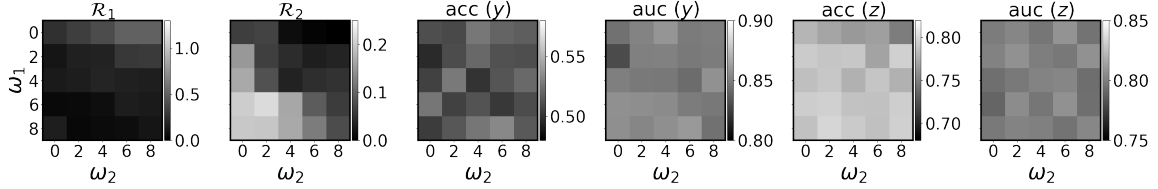


Figure 4-4: The effect of regularization on model inconsistency (left 2) and performance of predicting task label  $y$  (middle 2) and evidence labels  $z$  (right 2). Here  $\mathcal{R}_1, \mathcal{R}_2$  is short hand for  $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. When regularizing for both  $\mathcal{R}_1, \mathcal{R}_2$ , e.g., down the diagonals of the matrix, we notice dramatic decrease in model inconsistency and competitive performance for the regularized model.

same time, the regularized model achieves similar performance on the severity grading task as well as the evidence prediction tasks. The improvement in performance can be attributed empirically to fact that heavily regularized models over-fit less.

From additional results on other types of regularizers in Appendix A.3, we do observe a significant drop in the average performance of the model for predicting evidence when the strength of regularization is high. The drop in predicted evidence performance is tolerable if we consider that the model rarely provides inconsistent supporting evidence. Consistency versus predictive performance is a trade-off only if we want extremely consistent models.

Figure 4-4 reinforces previous observation that penalizing  $\mathcal{R}_1(\hat{\mathcal{D}})$  makes  $\mathcal{R}_2(\hat{\mathcal{D}})$  higher and vice versa, when  $\omega_i$  that is held constant is different from 0, for  $i = 1, 2$ . Figure 4-2 shows the first column, the first row, and the diagonal slices of the grid in left two sub-figures in Figure 4-4. We refer the reader to Table A.2 for detailed statistics of inconsistency and performance along the diagonal slice of the grid.

## Comparison with Semantic Loss

We show that semantic loss is effective in promoting model consistency but inflexible. We vary the weight  $\omega$  of the following regularized problem

$$\min_{\theta} \mathcal{L}(\theta) + \omega \mathbb{E}_{x \sim \mathcal{D}} [L^{\alpha}(x, \theta)]. \quad (4.1)$$

We select the most accurate model on the validation set for subsequent evaluations.

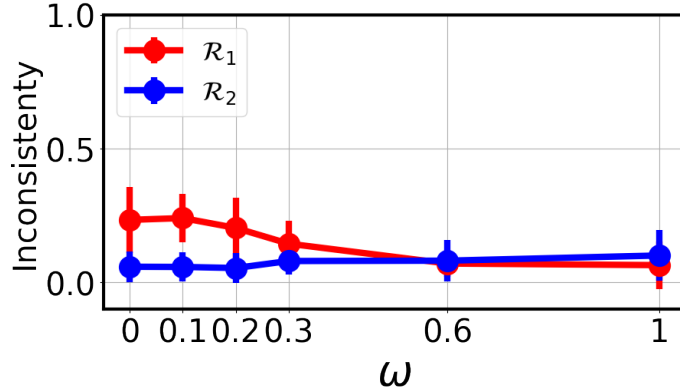


Figure 4-5: The effect of varying strength of regularizations using semantic loss during training on model consistency. Here  $\mathcal{R}_1, \mathcal{R}_2$  is short hand for  $\mathcal{R}_1(\hat{\mathcal{D}}), \mathcal{R}_2(\hat{\mathcal{D}})$ , respectively. It is not possible to keep both  $\mathcal{R}_1(\hat{\mathcal{D}})$  and  $\mathcal{R}_2(\hat{\mathcal{D}})$  down to small values at the same time.

Figure 4-5 demonstrates effect of semantic loss on model consistency. It is not possible to keep both incompatibility and insufficiency down to small values at the same time. This makes sense as the terms corresponding to the two constraints are equally weighted in Equation 3.24. Semantic loss lacks the flexibility to enforce multiple constraints with differing relevance. In contrast, our method is flexible because we take into account each constraint separately.

When using semantic loss, we notice that performance on the severity grading task is compromised in Table A.8. In contrast, our method does not suffer from performance loss.

## Bottleneck Architecture

Finally, we demonstrate that the concept bottleneck network [18] has similar behavior under consistency regularization, when compared with networks that predict  $(y, z)$  jointly. For fair comparison with previous experiments, we train the bottleneck network with the same objective function given in Equation 3.21. This optimization problem corresponds to the joint bottleneck approach [18]. We vary  $\omega_1, \omega_2$  together in the objective function and train multiple models. We select for the most accurate model on the validation set for subsequent evaluations.

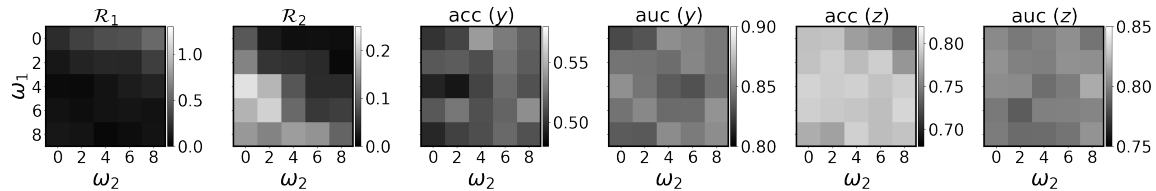


Figure 4-6: The effect of regularization on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ , where the models predict  $y$  via the  $z$  bottleneck. \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds. We observe similar pattern in inconsistency and performance when compared to that of models that predict  $(y, z)$  jointly.

Figure 4-6 demonstrates that concept bottleneck is amenable to consistency regularization. We observe that we can achieve good model consistency and maintain predictive performance on all tasks. Under consistency regularization, the bottleneck network behaves similarly to models that predict  $(y, z)$  jointly. We refer the reader to Table A.3 for detailed statistics of inconsistency and performance along the diagonal slice of the grid.

# Chapter 5

## Conclusions

In this thesis, we argued for supplementing model predictions with supporting evidence that is deemed useful by the end users. We defined a notion of consistent evidence via incorporating domain specific constraints. Then, we proposed ways to measure and enforce such constraints during model training. We evaluated our method on the pulmonary edema severity grading task, which provides a grounding for our consistent evidence framework. We demonstrated that consistent models remain competitive on the severity grading task as well as evidence prediction tasks.

Motivated by the edema severity grading task, we identified incompatible and insufficient evidence as a source of confusion for end users. Our definition of consistent evidence is represented as logical constraints, which can then be integrated into optimization. Other tasks in healthcare can benefit from representing the desired requirements as logical constraints and enforcing these constraints with established methodologies. Our work serves as a stepping stone for extending this approach to other applications.

We found that there exists a set of models with varying degree of consistency and similar predictive performance, under moderate consistency regularization. This finding is in contrast with the impression that consistency and performance may force a trade-off. Practically, we can always enforce consistency for the sake of interpretation without worrying too much that doing so may hurt classification performance.

Initially, we found it cumbersome to select the weights to the regularizers, since

the regularizers scale differently. One must adjust these weights accordingly to ensure that the two types of inconsistencies be approximately equal. To address this issue, we came up with regularizers that involve taking maximums for the compatibility and sufficiency constraints. The regularizers take values in the same interval, and therefore it is easier to strike a balance between them. To make hyperparameter search easier, we encourage the usage of regularizers that scale similarly.

Our proposed consistency regularization is agnostic to the choice of how  $(y, z)$  are predicted from the input image  $x$ . In particular, we found that models which predict  $(y, z)$  jointly as outputs and models which predict  $y$  via a bottleneck of evidence  $z$  behave similarly under consistency regularization. This finding is surprising, since the choice of evidence  $z$  as bottleneck variables limits the amount of information that can be used to classify  $y$ . This is not a concern for models that predict  $(y, z)$  jointly. Empirically, models that predict  $(y, z)$  jointly offer slightly better performance [18]. In contrast, our experiments do not exhibit such discrepancy in performance, likely because of how our data set is generated. Specifically, we determine  $z$  from clinical texts and compute  $y$  as a deterministic function of  $z$ . To conclude, our method is fairly generic and can be applied regardless of how  $(y, z)$  are computed, subject to certain condition on the data set.

We found that our proposed regularizers are more flexible than semantic loss. In particular, users can tune the relative importance of logical constraints for our proposed regularizers while this is not possible for semantic loss. For future work, we are motivated to generalize semantic loss to account for the relative importance of the logical constraints. In addition, we showed that one of our proposed regularizers upper bounds semantic loss for the consistency constraints. These loss functions have similar formulation, and therefore it is unsurprising that both are capable of enforcing consistency. It will be interesting to understand how these loss functions are different.

Our work does not provide a robust strategy to compare different methods effectively. Conceptually, end users specify a fixed inconsistency budget and will want to use a model with the best performance under budget constraint. To determine the best approach, we want to be able to compare the performance of models trained using



different regularizers that have exhausted a fixed inconsistency budget. In practice, we select models based on accuracy instead of consistency, making such comparison difficult. One strategy is to pre-specify the inconsistency budget and consider the consistency constraints as constraints of the optimization problem. One can then solve the constrained optimization problem via Lagrangian relaxation. This represents a promising future direction of research.



# Appendix A

## Appendix

### A.1 Truth Table for Consistency Constraints

# Rows	$y$	$z_1$	$z_2$	$z_3$	$z_4$	$z_5$	$z_6$	$z_7$	$p(\alpha   x; \theta)$ over partitions w.r.t. $y$	
1	0	-1	-1	-1	-1	-1	-1	-1	$p(y = 0) \prod_{k=1}^K p(z_k = -1)$	
7	1	-1	-1	1					$p(y = 1) [1 - \prod_{k=1}^3 p(z_k = -1)] \prod_{k=4}^7 p(z_k = -1)$	
		-1	1	-1						
		1	-1	-1						
		-1	1	1	-1	-1	-1	-1		
		1	1	-1						
24	2								$p(y = 2) [1 - \prod_{k=4}^5 p(z_k = -1)] \prod_{k=6}^7 p(z_k = -1)$	
					-1	1				
					1	-1	-1	-1		
96	3								$p(y = 3) [1 - \prod_{k=6}^7 p(z_k = -1)]$	
								-1		1
								1		-1
								1	1	

Table A.1: A list of satisfying instantiations in truth table of consistency constraints  $\alpha$  for the edema severity grading task. For brevity, we do not specify values that  $z_1, \dots, z_3$  take for the  $y = 2$  partition. This means  $z_1, \dots, z_3$  are unconstrained, and there are  $2^3$  possible instantiations for any given row within the  $y = 2$  partition. Similarly, there are  $2^5$  possible instantiations for any given row within the  $y = 3$  partition.

## A.2 Corresponding Table for Figure 4-4 and 4-6

$\omega_1, \omega_2$	*	0.0,0.0	2.0,2.0	4.0,4.0	6.0,6.0	8.0,8.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	0.249 $\pm$ 0.049	0.171 $\pm$ 0.067	0.189 $\pm$ 0.055	0.151 $\pm$ 0.196	0.108 $\pm$ 0.192
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	0.061 $\pm$ 0.033	0.062 $\pm$ 0.037	0.034 $\pm$ 0.026	0.090 $\pm$ 0.125	0.075 $\pm$ 0.103
acc ( $y$ )	0.514 $\pm$ 0.107	0.511 $\pm$ 0.043	0.510 $\pm$ 0.028	0.500 $\pm$ 0.078	0.502 $\pm$ 0.007	0.515 $\pm$ 0.013
auc ( $y$ )	0.841 $\pm$ 0.018	0.845 $\pm$ 0.042	0.850 $\pm$ 0.013	0.847 $\pm$ 0.021	0.849 $\pm$ 0.025	0.845 $\pm$ 0.017
acc (vascular congestion)	-	0.786 $\pm$ 0.028	0.780 $\pm$ 0.036	0.776 $\pm$ 0.063	0.780 $\pm$ 0.024	0.779 $\pm$ 0.005
acc (hilar congestion)	-	0.766 $\pm$ 0.047	0.790 $\pm$ 0.035	0.769 $\pm$ 0.070	0.788 $\pm$ 0.054	0.791 $\pm$ 0.046
acc (peribronchial cuffing)	-	0.813 $\pm$ 0.029	0.815 $\pm$ 0.029	0.814 $\pm$ 0.048	0.807 $\pm$ 0.130	0.831 $\pm$ 0.045
acc (septal lines)	-	0.846 $\pm$ 0.079	0.875 $\pm$ 0.092	0.817 $\pm$ 0.047	0.847 $\pm$ 0.109	0.852 $\pm$ 0.132
acc (interstitial abnormality)	-	0.655 $\pm$ 0.009	0.649 $\pm$ 0.026	0.642 $\pm$ 0.013	0.658 $\pm$ 0.017	0.636 $\pm$ 0.041
acc (air bronchograms)	-	0.872 $\pm$ 0.017	0.863 $\pm$ 0.109	0.874 $\pm$ 0.010	0.866 $\pm$ 0.052	0.899 $\pm$ 0.031
acc (parenchymal opacity)	-	0.720 $\pm$ 0.019	0.745 $\pm$ 0.065	0.747 $\pm$ 0.014	0.755 $\pm$ 0.019	0.770 $\pm$ 0.031
acc ( $z$ )	-	0.779	0.788	0.777	0.786	0.794

Table A.2: The effect of regularization on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ , where the models predict  $(y, z)$  jointly. \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds. We notice dramatic decrease in model inconsistency and competitive performance for the regularized model.

$\omega_1, \omega_2$	*	0.0,0.0	2.0,2.0	4.0,4.0	6.0,6.0	8.0,8.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	0.229 $\pm$ 0.056	0.178 $\pm$ 0.078	0.100 $\pm$ 0.040	0.107 $\pm$ 0.047	0.100 $\pm$ 0.036
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	0.086 $\pm$ 0.062	0.052 $\pm$ 0.064	0.084 $\pm$ 0.073	0.055 $\pm$ 0.071	0.098 $\pm$ 0.050
acc ( $y$ )	0.514 $\pm$ 0.107	0.500 $\pm$ 0.084	0.516 $\pm$ 0.051	0.507 $\pm$ 0.036	0.519 $\pm$ 0.042	0.512 $\pm$ 0.055
auc ( $y$ )	0.841 $\pm$ 0.018	0.829 $\pm$ 0.045	0.845 $\pm$ 0.039	0.836 $\pm$ 0.044	0.842 $\pm$ 0.032	0.855 $\pm$ 0.023
acc (vascular congestion)	-	0.785 $\pm$ 0.025	0.792 $\pm$ 0.023	0.774 $\pm$ 0.012	0.776 $\pm$ 0.042	0.793 $\pm$ 0.031
acc (hilar congestion)	-	0.798 $\pm$ 0.039	0.806 $\pm$ 0.074	0.801 $\pm$ 0.013	0.793 $\pm$ 0.084	0.776 $\pm$ 0.028
acc (peribronchial cuffing)	-	0.815 $\pm$ 0.051	0.832 $\pm$ 0.011	0.819 $\pm$ 0.011	0.811 $\pm$ 0.051	0.826 $\pm$ 0.038
acc (septal lines)	-	0.857 $\pm$ 0.004	0.856 $\pm$ 0.061	0.861 $\pm$ 0.073	0.838 $\pm$ 0.122	0.830 $\pm$ 0.084
acc (interstitial abnormality)	-	0.644 $\pm$ 0.012	0.641 $\pm$ 0.029	0.646 $\pm$ 0.016	0.647 $\pm$ 0.011	0.640 $\pm$ 0.024
acc (air bronchograms)	-	0.869 $\pm$ 0.026	0.869 $\pm$ 0.070	0.892 $\pm$ 0.024	0.880 $\pm$ 0.091	0.884 $\pm$ 0.020
acc (parenchymal opacity)	-	0.731 $\pm$ 0.013	0.751 $\pm$ 0.031	0.764 $\pm$ 0.008	0.745 $\pm$ 0.058	0.759 $\pm$ 0.028
acc ( $z$ )	-	0.785	0.792	0.794	0.784	0.787

Table A.3: The effect of regularization on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ , where the models predict  $y$  via the  $z$  bottleneck. \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds. We observe similar pattern in inconsistency and performance when compared to that of models that predict  $(y, z)$  jointly.

### A.3 Comparing Different Regularizers

We show model inconsistency and performance in subsequent sub-sections where the model is trained using the following list of regularizers

1.  $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$  ( $\mathcal{R}$  sum, hard)
2.  $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$  ( $\mathcal{R}$  sum, soft)
3.  $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$  ( $\mathcal{R}$  max, hard)
4.  $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$  ( $\mathcal{R}$  max, soft)
5.  $L^\alpha(x, \theta)$  (semantic loss)

We found that using the regularizer for compatibility that is agnostic to the size of  $\mathcal{I}_1(y)$ , i.e.,  $\overline{\mathcal{R}}_1(\theta)$  and  $\mathcal{R}_1(\theta)$  is beneficial for parameter tuning. The coefficients  $\omega_1, \omega_2$  can be set on the same scale to achieve a similar level of model compatibility and sufficiency.

When using hard regularizers, i.e.,  $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$ , we observe a drop in the average performance of the model for predicting evidence. This effect is less obvious for soft regularizers.

It is difficult to compare models trained using different regularizers. Ideally, we want to compare model performance at a fixed inconsistency budget. Since we select models using performance metrics, it is almost impossible to select the correct weights  $\omega_1, \omega_2$  to ensure a fixed inconsistency.

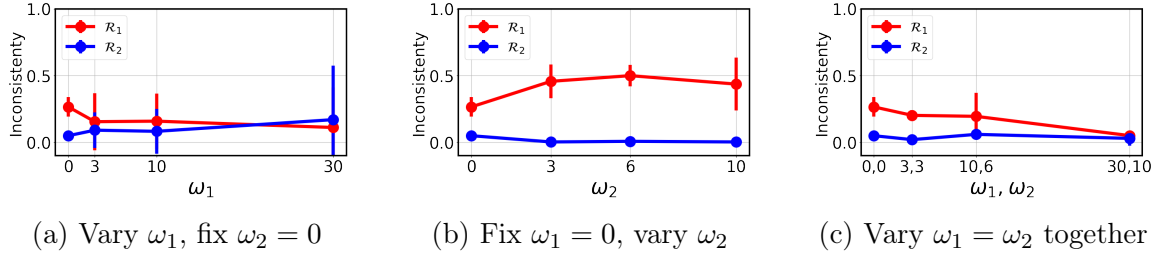
$\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$  ( $\mathcal{R}$  sum, hard)


Figure A-1: The effect of varying strength of regularization using  $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$  during training on model consistency.

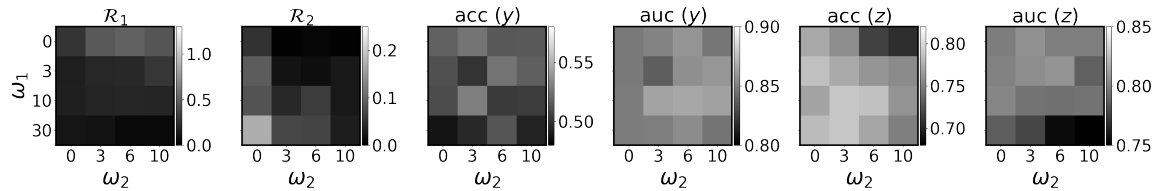


Figure A-2: The effect of varying strength of regularization using  $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$  on model inconsistency (left 2) and performance of predicting task label  $y$  (middle 2) and evidence labels  $z$  (right 2).

$\omega_1, \omega_2$	*	0.0,0.0	3.0,3.0	10.0,6.0	30.0,10.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	$0.265 \pm 0.074$	$0.201 \pm 0.038$	$0.194 \pm 0.175$	$0.050 \pm 0.015$
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	$0.049 \pm 0.026$	$0.019 \pm 0.026$	$0.059 \pm 0.036$	$0.029 \pm 0.055$
acc ( $y$ )	$0.514 \pm 0.107$	$0.518 \pm 0.053$	$0.500 \pm 0.040$	$0.503 \pm 0.040$	$0.494 \pm 0.053$
auc ( $y$ )	$0.841 \pm 0.018$	$0.848 \pm 0.034$	$0.837 \pm 0.028$	$0.866 \pm 0.011$	$0.846 \pm 0.022$
acc (vascular congestion)	-	$0.781 \pm 0.028$	$0.791 \pm 0.007$	$0.780 \pm 0.030$	$0.749 \pm 0.058$
acc (hilar congestion)	-	$0.752 \pm 0.064$	$0.773 \pm 0.067$	$0.777 \pm 0.027$	$0.703 \pm 0.063$
acc (peribronchial cuffing)	-	$0.815 \pm 0.013$	$0.795 \pm 0.076$	$0.815 \pm 0.045$	$0.805 \pm 0.004$
acc (septal lines)	-	$0.846 \pm 0.083$	$0.785 \pm 0.083$	$0.851 \pm 0.084$	$0.826 \pm 0.009$
acc (interstitial abnormality)	-	$0.637 \pm 0.004$	$0.663 \pm 0.039$	$0.628 \pm 0.014$	$0.607 \pm 0.010$
acc (air bronchograms)	-	$0.862 \pm 0.083$	$0.860 \pm 0.029$	$0.891 \pm 0.031$	$0.827 \pm 0.115$
acc (parenchymal opacity)	-	$0.714 \pm 0.041$	$0.745 \pm 0.048$	$0.757 \pm 0.025$	$0.732 \pm 0.033$
acc ( $z$ )	-	0.772	0.773	0.786	0.750

Table A.4: The effect of regularization using  $\overline{\mathcal{R}}_1^+(\theta), \overline{\mathcal{R}}_2(\theta)$  on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ . \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds.

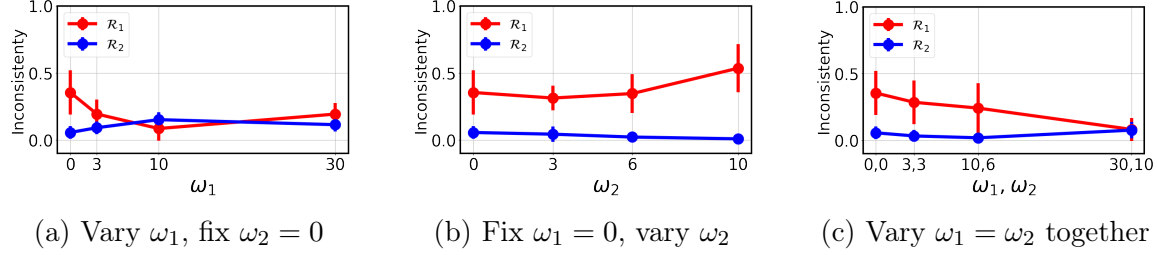
$\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$  ( $\mathcal{R}$  sum, soft)


Figure A-3: The effect of varying strength of regularization using  $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$  during training on model consistency.

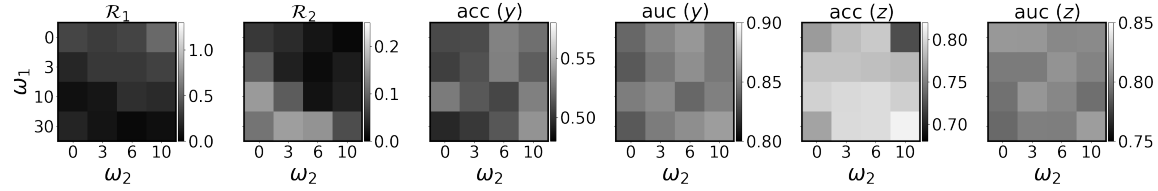


Figure A-4: The effect of varying strength of regularization using  $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$  on model inconsistency (left 2) and performance of predicting task label  $y$  (middle 2) and evidence labels  $z$  (right 2).

$\omega_1, \omega_2$	*	0.0,0.0	3.0,3.0	10.0,6.0	30.0,10.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	$0.354 \pm 0.165$	$0.284 \pm 0.163$	$0.241 \pm 0.187$	$0.081 \pm 0.086$
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	$0.056 \pm 0.047$	$0.032 \pm 0.045$	$0.018 \pm 0.033$	$0.076 \pm 0.061$
acc ( $y$ )	$0.514 \pm 0.107$	$0.509 \pm 0.049$	$0.512 \pm 0.021$	$0.507 \pm 0.048$	$0.538 \pm 0.049$
auc ( $y$ )	$0.841 \pm 0.018$	$0.840 \pm 0.016$	$0.847 \pm 0.013$	$0.840 \pm 0.028$	$0.862 \pm 0.012$
acc (vascular congestion)	-	$0.783 \pm 0.056$	$0.784 \pm 0.046$	$0.796 \pm 0.018$	$0.782 \pm 0.030$
acc (hilar congestion)	-	$0.775 \pm 0.066$	$0.761 \pm 0.092$	$0.802 \pm 0.017$	$0.826 \pm 0.024$
acc (peribronchial cuffing)	-	$0.769 \pm 0.130$	$0.795 \pm 0.074$	$0.818 \pm 0.037$	$0.843 \pm 0.026$
acc (septal lines)	-	$0.820 \pm 0.155$	$0.895 \pm 0.035$	$0.872 \pm 0.082$	$0.898 \pm 0.034$
acc (interstitial abnormality)	-	$0.648 \pm 0.029$	$0.632 \pm 0.049$	$0.661 \pm 0.011$	$0.650 \pm 0.041$
acc (air bronchograms)	-	$0.846 \pm 0.105$	$0.894 \pm 0.054$	$0.891 \pm 0.022$	$0.910 \pm 0.028$
acc (parenchymal opacity)	-	$0.714 \pm 0.060$	$0.747 \pm 0.055$	$0.765 \pm 0.026$	$0.777 \pm 0.007$
acc ( $z$ )	-	0.765	0.787	0.801	0.812

Table A.5: The effect of regularization using  $\mathcal{R}_1^+(\theta), \mathcal{R}_2(\theta)$  on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ . \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds.

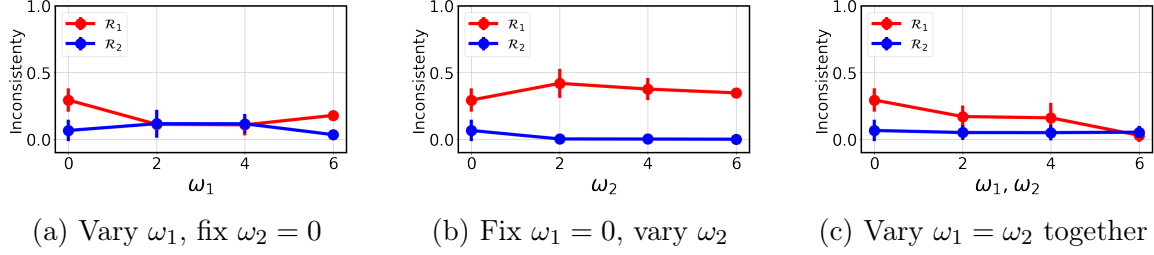
$\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$  ( $\mathcal{R}$  max, hard)


Figure A-5: The effect of varying strength of regularization using  $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$  during training on model consistency.

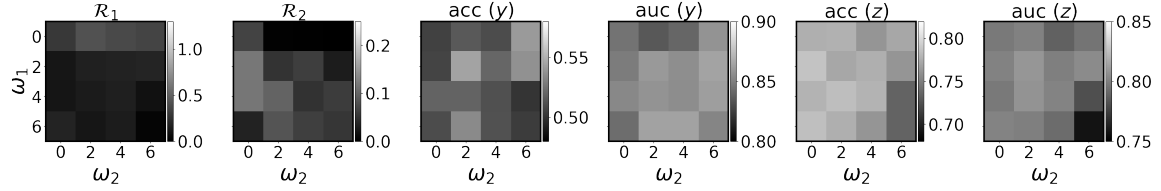


Figure A-6: The effect of varying strength of regularization using  $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$  on model inconsistency (left 2) and performance of predicting task label  $y$  (middle 2) and evidence labels  $z$  (right 2).

$\omega_1, \omega_2$	*	0.0,0.0	2.0,2.0	4.0,4.0	6.0,6.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	$0.294 \pm 0.088$	$0.170 \pm 0.082$	$0.161 \pm 0.111$	$0.029 \pm 0.048$
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	$0.066 \pm 0.080$	$0.050 \pm 0.056$	$0.050 \pm 0.057$	$0.053 \pm 0.047$
acc ( $y$ )	$0.514 \pm 0.107$	$0.506 \pm 0.007$	$0.544 \pm 0.050$	$0.511 \pm 0.020$	$0.504 \pm 0.049$
auc ( $y$ )	$0.841 \pm 0.018$	$0.845 \pm 0.006$	$0.860 \pm 0.016$	$0.855 \pm 0.021$	$0.852 \pm 0.015$
acc (vascular congestion)	-	$0.775 \pm 0.047$	$0.774 \pm 0.042$	$0.786 \pm 0.008$	$0.772 \pm 0.010$
acc (hilar congestion)	-	$0.755 \pm 0.104$	$0.798 \pm 0.039$	$0.808 \pm 0.011$	$0.667 \pm 0.102$
acc (peribronchial cuffing)	-	$0.803 \pm 0.040$	$0.802 \pm 0.013$	$0.820 \pm 0.018$	$0.805 \pm 0.074$
acc (septal lines)	-	$0.844 \pm 0.043$	$0.792 \pm 0.061$	$0.779 \pm 0.050$	$0.700 \pm 0.064$
acc (interstitial abnormality)	-	$0.647 \pm 0.018$	$0.652 \pm 0.022$	$0.637 \pm 0.050$	$0.621 \pm 0.006$
acc (air bronchograms)	-	$0.892 \pm 0.048$	$0.845 \pm 0.039$	$0.872 \pm 0.044$	$0.829 \pm 0.026$
acc (parenchymal opacity)	-	$0.718 \pm 0.040$	$0.737 \pm 0.009$	$0.745 \pm 0.031$	$0.746 \pm 0.013$
acc ( $z$ )	-	0.776	0.771	0.778	0.734

Table A.6: The effect of regularization using  $\overline{\mathcal{R}}_1(\theta), \overline{\mathcal{R}}_2(\theta)$  on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ . \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds.



## $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$ ( $\mathcal{R}$ max, soft)

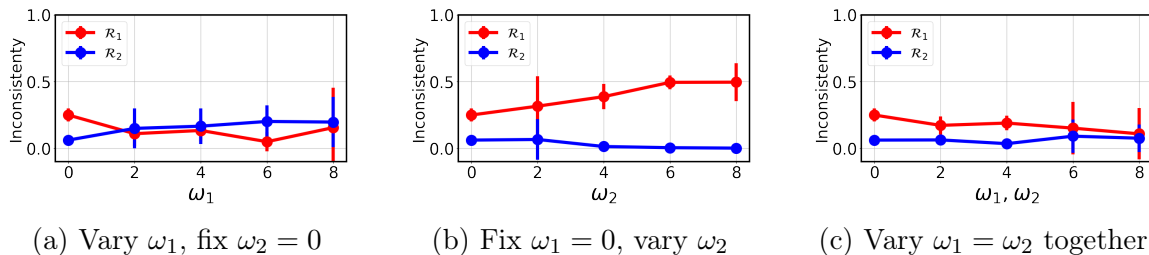


Figure A-7: The effect of varying strength of regularization using  $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$  during training on model consistency.

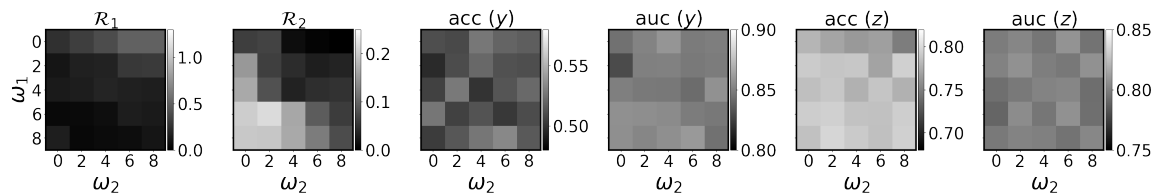


Figure A-8: The effect of varying strength of regularization using  $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$  on model inconsistency (left 2) and performance of predicting task label  $y$  (middle 2) and evidence labels  $z$  (right 2).

$\omega_1, \omega_2$	*	0.0,0.0	2.0,2.0	4.0,4.0	6.0,6.0	8.0,8.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	$0.249 \pm 0.049$	$0.171 \pm 0.067$	$0.189 \pm 0.055$	$0.151 \pm 0.196$	$0.108 \pm 0.192$
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	$0.061 \pm 0.033$	$0.062 \pm 0.037$	$0.034 \pm 0.026$	$0.090 \pm 0.125$	$0.075 \pm 0.103$
acc ( $y$ )	$0.514 \pm 0.107$	$0.511 \pm 0.043$	$0.510 \pm 0.028$	$0.500 \pm 0.078$	$0.502 \pm 0.007$	$0.515 \pm 0.013$
auc ( $y$ )	$0.841 \pm 0.018$	$0.845 \pm 0.042$	$0.850 \pm 0.013$	$0.847 \pm 0.021$	$0.849 \pm 0.025$	$0.845 \pm 0.017$
acc (vascular congestion)	-	$0.786 \pm 0.028$	$0.780 \pm 0.036$	$0.776 \pm 0.063$	$0.780 \pm 0.024$	$0.779 \pm 0.005$
acc (hilar congestion)	-	$0.766 \pm 0.047$	$0.790 \pm 0.035$	$0.769 \pm 0.070$	$0.788 \pm 0.054$	$0.791 \pm 0.046$
acc (peribronchial cuffing)	-	$0.813 \pm 0.029$	$0.815 \pm 0.029$	$0.814 \pm 0.048$	$0.807 \pm 0.130$	$0.831 \pm 0.045$
acc (septal lines)	-	$0.846 \pm 0.079$	$0.875 \pm 0.092$	$0.817 \pm 0.047$	$0.847 \pm 0.109$	$0.852 \pm 0.132$
acc (interstitial abnormality)	-	$0.655 \pm 0.009$	$0.649 \pm 0.026$	$0.642 \pm 0.013$	$0.658 \pm 0.017$	$0.636 \pm 0.041$
acc (air bronchograms)	-	$0.872 \pm 0.017$	$0.863 \pm 0.109$	$0.874 \pm 0.010$	$0.866 \pm 0.052$	$0.899 \pm 0.031$
acc (parenchymal opacity)	-	$0.720 \pm 0.019$	$0.745 \pm 0.065$	$0.747 \pm 0.014$	$0.755 \pm 0.019$	$0.770 \pm 0.031$
acc ( $z$ )	-	0.779	0.788	0.777	0.786	0.794

Table A.7: The effect of regularization using  $\mathcal{R}_1(\theta), \mathcal{R}_2(\theta)$  on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ . \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds.

## $L^\alpha(x, \theta)$ (semantic loss)

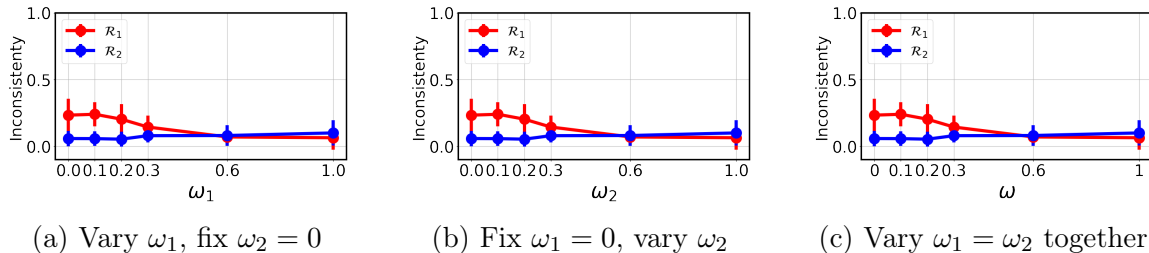


Figure A-9: The effect of varying strength of regularization using  $L^\alpha(x, \theta)$  during training on model consistency.

$\omega_1, \omega_2$	*	0.0,0.0	0.1,0.1	0.2,0.2	0.3,0.3	0.6,0.6	1.0,1.0
$\mathcal{R}_1(\hat{\mathcal{D}})$	-	$0.233 \pm 0.123$	$0.240 \pm 0.090$	$0.203 \pm 0.113$	$0.144 \pm 0.085$	$0.070 \pm 0.051$	$0.064 \pm 0.089$
$\mathcal{R}_2(\hat{\mathcal{D}})$	-	$0.058 \pm 0.059$	$0.057 \pm 0.056$	$0.053 \pm 0.057$	$0.079 \pm 0.051$	$0.080 \pm 0.078$	$0.100 \pm 0.095$
acc ( $y$ )	$0.514 \pm 0.107$	$0.501 \pm 0.040$	$0.513 \pm 0.035$	$0.502 \pm 0.042$	$0.503 \pm 0.059$	$0.503 \pm 0.030$	$0.474 \pm 0.066$
auc ( $y$ )	$0.841 \pm 0.018$	$0.838 \pm 0.006$	$0.829 \pm 0.029$	$0.831 \pm 0.010$	$0.829 \pm 0.042$	$0.811 \pm 0.033$	$0.798 \pm 0.032$
acc (vascular congestion)	-	$0.794 \pm 0.015$	$0.768 \pm 0.044$	$0.797 \pm 0.004$	$0.770 \pm 0.045$	$0.768 \pm 0.051$	$0.767 \pm 0.034$
acc (hilar congestion)	-	$0.778 \pm 0.036$	$0.773 \pm 0.039$	$0.814 \pm 0.058$	$0.808 \pm 0.019$	$0.787 \pm 0.074$	$0.797 \pm 0.041$
acc (peribronchial cuffing)	-	$0.819 \pm 0.017$	$0.795 \pm 0.030$	$0.817 \pm 0.013$	$0.843 \pm 0.036$	$0.813 \pm 0.073$	$0.848 \pm 0.006$
acc (septal lines)	-	$0.865 \pm 0.075$	$0.851 \pm 0.115$	$0.824 \pm 0.107$	$0.859 \pm 0.077$	$0.884 \pm 0.076$	$0.872 \pm 0.069$
acc (interstitial abnormality)	-	$0.651 \pm 0.012$	$0.644 \pm 0.054$	$0.649 \pm 0.009$	$0.635 \pm 0.016$	$0.651 \pm 0.023$	$0.660 \pm 0.023$
acc (air bronchograms)	-	$0.884 \pm 0.049$	$0.863 \pm 0.004$	$0.870 \pm 0.053$	$0.862 \pm 0.073$	$0.887 \pm 0.046$	$0.894 \pm 0.037$
acc (parenchymal opacity)	-	$0.729 \pm 0.024$	$0.729 \pm 0.063$	$0.725 \pm 0.032$	$0.726 \pm 0.037$	$0.759 \pm 0.037$	$0.763 \pm 0.051$
acc ( $z$ )	-	0.789	0.775	0.785	0.786	0.793	0.800

Table A.8: The effect of regularization using  $L^\alpha(x, \theta)$  on model inconsistency and performance of predicting task label  $y$  and evidence label  $z$ . \* stands for baseline model that is trained to predict task label  $y$ . We compute average numbers with  $\pm 2$  standard deviations over random seeds.

# Bibliography

- [1] David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [2] Wei Bi and James T. Kwok. Multi-label classification on tree- and DAG-structured hierarchies. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11*, pages 17–24, Madison, WI, USA, June 2011. Omnipress.
- [3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Semantic Bottleneck for Computer Vision Tasks. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, Lecture Notes in Computer Science, pages 695–712, Cham, 2019. Springer International Publishing.
- [4] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 1721–1730, Sydney, NSW, Australia, 2015. ACM Press.
- [5] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020.
- [6] Noel C. F. Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei, and Aleksandra Mojsilović. Teaching AI to Explain its Decisions Using Embeddings and Multi-Task Learning. *ICML Workshop on Human in the Loop Learning*, June 2019.
- [7] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa

- Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, September 2018.
- [8] Michelangelo Diligenti, Soumali Roychowdhury, and Marco Gori. Integrating Prior Knowledge into Deep Learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 920–923, December 2017.
- [9] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10-11):2436–2449, October 2011.
- [10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, June 2014.
- [11] Eleonora Giunchiglia and Thomas Lukasiewicz. Coherent hierarchical multi-label classification networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9662–9673. Curran Associates, Inc., 2020.
- [12] Thomas Gluecker, Patrizio Capasso, Pierre Schnyder, François Gudinchet, Marie-Denise Schaller, Jean-Pierre Revelly, René Chiolero, Peter Vock, and Stéphan Wicky. Clinical and Radiologic Features of Pulmonary Edema. *Radiographics*, 19(6):1507–1531, November 1999.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.
- [14] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. TED: Teaching AI to Explain its Decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, Honolulu HI USA, January 2019. ACM.
- [15] Steven Horng, Ruizhi Liao, Xin Wang, Sandeep Dalal, Polina Golland, and Seth J. Berkowitz. Deep Learning to Quantify Pulmonary Edema in Chest Radiographs. *Radiology: Artificial Intelligence*, 3(2):e190228, March 2021.
- [16] Alistair E. W. 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):317, December 2019.

- [17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2015.
- [18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In *ICML*, January 2020.
- [19] Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. A Logic-Driven Framework for Consistency of Neural Models. 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 3924–3935, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [20] Ruizhi Liao, Jonathan E. Rubin, Grace Lam, Seth J. Berkowitz, Sandeep Dalal, William Wells, Steven Horng, and Polina Golland. Semi-supervised Learning for Quantification of Pulmonary Edema in Chest X-Ray Images. *ArXiv*, abs/1902.10785, 2019.
- [21] Max Losch, Mario Fritz, and Bernt Schiele. Interpretability Beyond Classification Output: Semantic Bottleneck Networks. *arXiv:1907.10882 [cs]*, July 2019.
- [22] Aniruddh Raghu, John Gutttag, Katherine Young, Eugene Pomerantsev, Adrian V. Dalca, and Collin M. Stultz. Learning to predict with supporting evidence: Applications to clinical risk prediction. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 95–104, Virtual Event USA, April 2021. ACM.
- [23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery.
- [24] 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. *International Journal of Computer Vision*, 2016.
- [25] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical Multi-Label Classification Networks. In *International Conference on Machine Learning*, pages 5075–5084. PMLR, July 2018.
- [26] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5502–5511. PMLR, July 2018.

- [27] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2740–2748, Santiago, Chile, December 2015. IEEE.
- [28] Omar Zaidan, Jason Eisner, and Christine Piatko. Using “Annotator Rationales” to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics.
- [29] Ye Zhang, Iain Marshall, and Byron C. Wallace. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas, November 2016. Association for Computational Linguistics.
- [30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Las Vegas, NV, USA, June 2016. IEEE.