Multimodal Representation Learning via Maximization of Local Mutual Information

Citation: Liao, Ruizhi, Moyer, Daniel, Cha, Miriam, Quigley, Keegan, Berkowitz, Seth et al. 2021. "Multimodal Representation Learning via Maximization of Local Mutual Information." MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION - MICCAI 2021, PT II, 12902.

As Published: 10.1007/978-3-030-87196-3_26

Publisher: Springer International Publishing

Persistent URL: https://hdl.handle.net/1721.1/143584

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Multimodal Representation Learning via Maximization of Local Mutual Information

Ruizhi Liao[1], Daniel Moyer[1], Miriam Cha[2], Keegan Quigley[2],
Seth Berkowitz[3], Steven Horng[3], Polina Golland[1], and William M. Wells[1,4]

[1] CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA
[2] MIT Lincoln Laboratory, Lexington, MA, USA
[3] Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
[4] Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Abstract. We propose and demonstrate a representation learning approach by maximizing the mutual information between local features of images and text. The goal of this approach is to learn *useful* image representations by taking advantage of the rich information contained in the free text that describes the findings in the image. Our method trains image and text encoders by encouraging the resulting representations to exhibit high local mutual information. We make use of recent advances in mutual information estimation with neural network discriminators. We argue that the sum of local mutual information is typically a lower bound on the global mutual information. Our experimental results in the downstream image classification tasks demonstrate the advantages of using local features for image-text representation learning. Our code is available at: https://github.com/RayRuizhiLiao/mutual_info_img_txt.

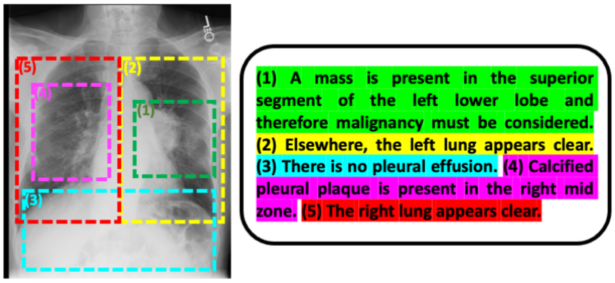
Keywords: Multimodal representation learning · Local feature representations · Mutual information maximization.

1 Introduction

We present a novel approach for image-text representation learning by maximizing the mutual information between local features of the images and the text. In the context of medical imaging, the images could be, for example, radiographs and the text could be radiology reports that capture radiologists' impressions of the images. A large number of such image-text pairs are generated in the clinical workflow every day [7, 13]. Jointly learning from images and raw text can support a leap in the quality of medical vision models by taking advantage of existing expert descriptions of the images.

Learning to extract *useful* feature representations from training data is an essential objective of a deep learning model. The definition of *usefulness* is task-specific [3,5,25]. In this work, we aim to learn image representations that improve classification tasks, such as pathology detection, by making use of the rich information contained in the raw text that describe the findings in the image.

We exploit mutual information (MI) to learn useful image representations jointly with text. MI quantifies statistical dependencies between two random



(1) A mass is present in the superior segment of the left lower lobe and therefore malignancy must be considered. (2) Elsewhere, the left lung appears clear. (3) There is no pleural effusion. (4) Calcified pleural plaque is present in the right mid zone. (5) The right lung appears clear.

Fig. 1. An example image-text pair (a chest radiograph and its associated radiology report). Each sentence describes the image findings in a particular region of the image. This figure is best viewed in color.

variables. Prior work has estimated and optimized MI across images for image registration [20,29], and MI between images and image features for unsupervised learning [6,10,23]. Since the text usually describes image findings that are relevant for downstream image classification tasks, it is sensible to encourage the image and text representations to exhibit high MI.

We propose to learn an image encoder and a text encoder by maximizing the MI of their resulting image and text representations. Moreover, we estimate and optimize the MI between local image features and sentence-level text representations. Fig. 1 shows an example image-text pair, where the image is a chest radiograph and the document is the associated radiology report [13]. Each sentence in the report describes a local region in the image. A sentence is usually a minimal and complete semantic unit [24,32]. The findings described in that semantic unit are usually captured in a local region of the image [8].

Prior work in image-text joint learning has leveraged image-based text generation as an auxiliary task during the image model training [21,27,31], or has blended image and text features for downstream inference tasks [22]. Other work has leveraged contrastive learning, an approach to maximize a lower bound on MI to learn image and text representations jointly [4,32]. To the best of our knowledge, this work represents the first attempt to exploit the image spatial structure and sentence-level text features with MI maximization to learn image and text representations that are *useful* for subsequent analysis of images. In our experimental results, we demonstrate that the maximization of local MI yields the greatest improvement in the downstream image classification tasks.

This paper is organized as follows. In Section 2, we derive our approach for image-text representation learning by maximizing local MI. Section 3 discusses the theoretical motivation behind local mutual information. This is followed by empirical evaluation in Section 4, where we describe the implementation details of our algorithms in application to chest radiographs and radiology reports.

2 Methods

Let x^{I} be an image and x^{R} be the associated free text such as a radiology report or a pathology report that describes findings in the image. The objective is to learn useful latent image representations $z^{\mathrm{I}}(x^{\mathrm{I}})$ and text representations $z^{\mathrm{R}}(x^{\mathrm{R}})$ from image-text data $\mathcal{X} = \{\mathrm{x}_j\}_{j=1}^N$, where $\mathrm{x}_j = (\mathrm{x}_j^{\mathrm{I}}, \mathrm{x}_j^{\mathrm{R}})$. We construct an image encoder and a text encoder parameterized by $\theta_{\mathrm{E}}^{\mathrm{I}}$ and $\theta_{\mathrm{E}}^{\mathrm{R}}$, respectively, to generate the representations $z^{\mathrm{I}}(x^{\mathrm{I}}; \theta_{\mathrm{E}}^{\mathrm{I}})$ and $z^{\mathrm{R}}(x^{\mathrm{R}}; \theta_{\mathrm{E}}^{\mathrm{R}})$.

Mutual Information Maximization. We seek such image and text encoders and learn their representations by maximizing MI between the image representation and the text representation:

$$I(z^{\mathrm{I}}, z^{\mathrm{R}}) \triangleq \mathbb{E}_{p(z^{\mathrm{I}}, z^{\mathrm{R}})} \left[\log \frac{p(z^{\mathrm{I}}, z^{\mathrm{R}})}{p(z^{\mathrm{I}}) p(z^{\mathrm{R}})} \right]. \tag{1}$$

We employ MI as a statistical measure that captures dependency between images and text in the joint representation space. Maximizing MI between image and text representations is equivalent to maximizing the difference of the entropy and the conditional entropy of image representation given text: $I(z^{\mathrm{I}}, z^{\mathrm{R}}) = H(z^{\mathrm{I}}) - H(z^{\mathrm{I}} | z^{\mathrm{R}})$. This criterion encourages the model to learn feature representations where the information from one modality reduces the entropy of the other data modality, which is a better choice than solely minimizing the conditional entropy, where the image encoder could generate identical features for all data to achieve the conditional entropy minimum.

Stochastic Optimization of MI. Estimating mutual information between high-dimensional continuous variables from finite data samples is challenging. We leverage the recent advances that employ neural network discriminators for MI estimation and maximization [2, 18, 23, 26]. The key idea is to construct a discriminator $f(\mathrm{z}_i^{\mathrm{I}}, \mathrm{z}_j^{\mathrm{R}}; \theta_{\mathrm{D}})$, parameterized by θ_{D}, that estimates the likelihood (or the likelihood ratio) of whether a sample pair $(\mathrm{z}_i^{\mathrm{I}}, \mathrm{z}_j^{\mathrm{R}})$ is sampled from the joint distribution $p(z^{\mathrm{I}}, z^{\mathrm{R}})$ or from the product of marginals $p(z^{\mathrm{I}})p(z^{\mathrm{R}})$. The discriminator is commonly found by maximizing the lower bound of the MI approximated by the likelihood ratio in Eq. (1) [2, 23].

We train the discriminator $f(\mathrm{z}_i^{\mathrm{I}}, \mathrm{z}_j^{\mathrm{R}}; \theta_{\mathrm{D}})$ jointly with image and text encoders $z^{\mathrm{I}}(x^{\mathrm{I}}; \theta_{\mathrm{E}}^{\mathrm{I}})$ and $z^{\mathrm{R}}(x^{\mathrm{R}}; \theta_{\mathrm{E}}^{\mathrm{R}})$ via MI maximization:

$$\hat{\theta}_{\mathrm{E}}^{\mathrm{I}}, \hat{\theta}_{\mathrm{E}}^{\mathrm{R}}, \hat{\theta}_{\mathrm{D}} = \arg \max_{\theta_{\mathrm{E}}^{\mathrm{I}}, \theta_{\mathrm{E}}^{\mathrm{R}}, \theta_{\mathrm{D}}} \hat{I}(z^{\mathrm{I}}(x^{\mathrm{I}}; \theta_{\mathrm{E}}^{\mathrm{I}}), z^{\mathrm{R}}(x^{\mathrm{R}}; \theta_{\mathrm{E}}^{\mathrm{R}}); \theta_{\mathrm{D}}), \tag{2}$$

where $\hat{I}(z^{\mathrm{I}}, z^{\mathrm{R}}; \theta_{\mathrm{D}})$ is a lower bound on $I(z^{\mathrm{I}}, z^{\mathrm{R}})$. We consider two MI lower bounds: Mutual Information Neural Estimation (MINE) [2] and Contrastive Predictive Coding (CPC) [23]. In our experiments, we empirically show that our method is not sensitive to the choice of the lower bound. MINE estimates the MI lower bound by approximating the log likelihood ratio in Eq. (1), using the Donsker-Varadhan (DV) variational formula of the KL divergence between the joint distribution and the product of the marginals, which yields the lower bound

$$\hat{I}^{(\mathrm{MINE})}_{\theta_{\mathrm{E}}^{\mathrm{I}}, \theta_{\mathrm{E}}^{\mathrm{R}}, \theta_{\mathrm{D}}}(z^{\mathrm{I}}, z^{\mathrm{R}}) = \mathbb{E}_{p(z^{\mathrm{I}}, z^{\mathrm{R}})} \left[f(z^{\mathrm{I}}, z^{\mathrm{R}}; \theta_{\mathrm{D}}) \right] - \log \mathbb{E}_{p(z^{\mathrm{I}}) p(z^{\mathrm{R}})} \left[e^{f(z^{\mathrm{I}}, z^{\mathrm{R}}; \theta_{\mathrm{D}})} \right]. \tag{3}$$

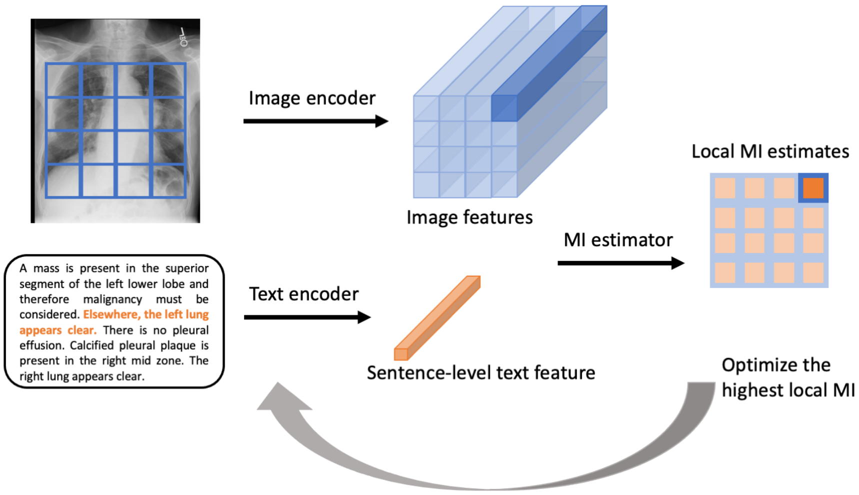
CPC computes the MI lower bound by approximating the likelihood of an image-text feature pair being sampled from the joint distribution over the product of marginals, which leads to the objective function

$$\hat{I}^{(\mathrm{CPC})}_{\theta^{\mathrm{I}}_{\mathrm{E}},\theta^{\mathrm{R}}_{\mathrm{E}},\theta_{\mathrm{D}}}(z^{\mathrm{I}},z^{\mathrm{R}}) = \mathbb{E}_{p(z^{\mathrm{I}},z^{\mathrm{R}})}\left[f(z^{\mathrm{I}},z^{\mathrm{R}};\theta_{\mathrm{D}})\right] - \mathbb{E}_{p(z^{\mathrm{I}})p(z^{\mathrm{R}})}\left[\log\sum_{\hat{z}^{\mathrm{R}}_j\in z^{\mathrm{R}}} e^{f(z^{\mathrm{I}},\hat{z}^{\mathrm{R}}_j;\theta_{\mathrm{D}})}\right]. \quad (4)$$

Both methods sample from the matched image-text pairs and from shuffled pairs (to approximate the product of marginals), and train the discriminator to differentiate between these two types of sample pairs.

Local MI Maximization. We propose to maximize MI between local features of images and sentence-level features from text. Given a sentence-level feature in the text, we estimate the MI values between all local image features and this sentence, select the image feature with the highest MI, and maximize the MI between that image feature and the sentence feature (Fig. 2). We train the image and text encoders, as well as the MI discriminator based on all the image-text data:

$$\hat{\theta}^{\mathrm{I}}_{\mathrm{E}},\hat{\theta}^{\mathrm{R}}_{\mathrm{E}},\hat{\theta}_{\mathrm{D}} = \arg\max_{\theta^{\mathrm{I}}_{\mathrm{E}},\theta^{\mathrm{R}}_{\mathrm{E}},\theta_{\mathrm{D}}} \sum_{j}\sum_{m} \max_{n} \hat{I}(\mathrm{z}^{\mathrm{I}}_{j,(n)},\mathrm{z}^{\mathrm{R}}_{j,(m)}), \quad (5)$$



A mass is present in the superior segment of the left lower lobe and therefore malignancy must be considered. Elsewhere, the left lung appears clear. There is no pleural effusion. Calcified pleural plaque is present in the right mid zone. The right lung appears clear.

Fig. 2. Local MI Maximization. First, we randomly select a sentence in the text and encode the sentence into a sentence-level feature. The corresponding image is encoded into a M×M×D feature block. We estimate the MI values between all local image features and the sentence feature. Note that the MI estimation needs shuffled image-text data, which is not illustrated in this diagram. We select the local image feature with the highest MI and update the image encoder, text encoder, and the MI discriminator such that the local MI between that image feature and the sentence feature is maximized.

where $\mathrm{z}^{\mathrm{I}}_{j,(n)}$ is the n-th local feature in image $\mathrm{x}^{\mathrm{I}}_{j}$, and $\mathrm{z}^{\mathrm{R}}_{j,(m)}$ is the m-th sentence feature in text $\mathrm{x}^{\mathrm{R}}_{j}$. We use this *one-way* maximum, because in image captioning, every sentence was written to describe some finding in the corresponding image. In contrast, not every region in the image has a related sentence in the text that describes it.

3 Generative Model and Motivation

To provide further insight into the theoretical motivation behind local mutual information, we describe a conjectured generative model for how paired chest radiograph and radiology report are constructed. As shown in Fig 3, each local image region x^{I}_{n} has a hidden variable H_n that specifies the physiological processes and disease status in that region. This image region x^{I}_{n} is generated by the hidden variable H_n and another random variable V^{I} that is independent of H_n (e.g., the image acquisition protocol). The *corresponding* sentence in the radiology report is generated by first choosing the sentence index m (mapping from the image region index n via M, i.e., $m = f(n; M)$) and then generated as a function of H_n and another random variable V^{R} that is independent of H (e.g., the radiologist's training background).

The task we are interested in is to predict the hidden disease statuses $\{H_n\}$ given an image x^{I}. Therefore, it is sensible to learn an image feature representation z^{I} that has high mutual information with $\{H_n\}$, i.e., $\sum_n I(z^{\mathrm{I}}, H_n)$. z^{I} is the concatenation of z^{I}_{n} and $z^{\bar{\mathrm{I}}}_{n}$, where the z^{I}_{n} is the feature of the local image region generated from H_n and z^{I}_{n} is the rest of the image features. Applying the chain rule of mutual information, we have:

$$I(z^{\mathrm{I}}, H_n) = I(z^{\mathrm{I}}_{n}, H_n) + I(z^{\bar{\mathrm{I}}}_{n}, H_n | z^{\mathrm{I}}) \tag{6}$$

$$\geq I(z^{\mathrm{I}}_{n}, H_n). \tag{7}$$

Since $I(z^{\mathrm{I}}_{n}, H_n)$ is a lower bound to $I(z^{\mathrm{I}}, H_n)$, we maximize $I(z^{\mathrm{I}}_{n}, H_n)$. The challenge of learning such image feature representations is that we have limited labels

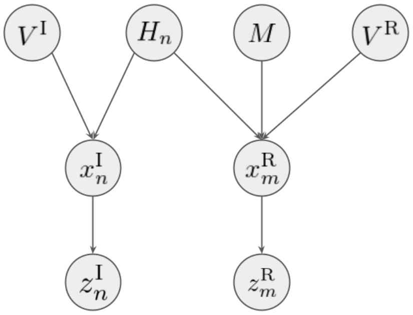


Fig. 3. A conjectured generative model that describes how paired chest radiograph and radiology report are constructed and the underlying structural assumptions.

for disease status. However, both the local image region and the *corresponding* sentence in the report are generated by the same hidden disease status. Assuming V^{I} and V^{R} are independent, maximizing $I(z_n^{\mathrm{I}}, z_m^{\mathrm{R}})$ will likely lead to high $I(z_n^{\mathrm{I}}, H_n)$, because H_n is the only source of information shared by z_n^{I} and z_m^{R}. Here we do the index mapping by selecting the sentence in the report that has the highest mutual information with z_n^{I}.

Therefore, conjecturing this generative model by making structural (conditional independence) assumptions of the image and report data results in our proposed local mutual information maximization approach. The local MI optimization is usually an easier task given its lower dimension and more training samples to discover useful representations. The utility of our strategy is supported by our experimental results.

4 Experiments

Data and Model Evaluation. We demonstrate our approach on the MIMIC-CXR dataset v2.0 [13] that includes around 250K frontal-view chest radiographs with their associated radiology reports. We evaluate our representation learning methods on two downstream classification tasks:

- **Pathology9**. Detecting 9 pathologies from the chest radiographs against the labels that were extracted from the corresponding radiology reports using a radiology report labeler CheXpert [12,14,15]. Note that there are 14 findings available in the repository [14]. We only train and evaluate 9 for which there are more than around 100 images available in the test set.
- **EdemaSeverity**. Assessing pulmonary edema severity from chest radiographs against the labels that were annotated by radiologists on the images [11,17,19,28]. The severity level ranges from 0 to 3 with a higher score indicating higher risk.

The test sets provided in MIMIC-CXR with CheXpert labels [14] and with edema severity labels [17] are used to evaluate our methods. The patients that are in either of the two test sets are excluded from the model training. Table 1 summarizes the size of the (labeled) training data and test data.

–	Support Devices	Cardiomegaly	Consolidation	Edema	Lung Opacity
training	76,492	65,129	20,074	56,203	58,105
test	286	404	95	373	318
–	Pleural Effusion	Pneumonia	Pneumothorax	Atelectasis	Edema Severity
training	86,871	43,951	56,472	50,416	7,066
test	451	195	191	262	141

Table 1. The number of images in the (labeled) training sets and the test sets.

Experimental Design. Our goal is to learn representations that are useful for downstream classification tasks. Therefore, we use a fully supervised image model trained on the chest radiographs with available training labels as our benchmark. We compare two ways to use our image representations when *re-training* the image classifier: 1) freezing the image encoder; 2) fine-tuning the image encoder. In either case, the image encoder followed by a classifier is trained on the same training set that is used to train the fully supervised image model.

We compare our MI maximization approach on local features with the global MI maximization. We test both MINE [2] and CPC [23] as MI estimators. To summarize, we evaluate the variants of our model and training regimes as follows:

- **image-only-supervised**: An image-only model trained on the training data provided in [14, 17].
- **global-mi-mine**, **global-mi-cpc**: Representation learning on the chest radiographs and the radiology reports using global MI maximization.
  - **encoder-frozen**, **encoder-tuned**: Once representation learning is completed, the image encoder followed by a classifier is *re-trained* on the labeled training image data, with the encoder frozen or fine-tuned.
- **local-mi-mine**, **local-mi-cpc**: Representation learning using local MI maximization in Eq. (5).
  - **encoder-frozen**, **encoder-tuned**: The resulting image encoder followed by a classifier is *re-trained*, with the encoder frozen or fine-tuned.

At the image model training or *re-training* time, all variants are trained on the same training sets. Note that the **local-mi** approach makes use of lower level image features. To make the **encoder-frozen** experiments comparable between **local-mi** and **global-mi**, we only freeze the same lower level feature extractor in both encoders.

Implementation Details. Chest radiographs are downsampled to 256×256. We use a 5-block resnet [9] as the image encoder in the local MI approach and the image feature representation z^{I} is 16×512 (4×4×512) feature vectors. We use a 6-block resnet as the image encoder for the global MI maximization, where the image representation z^{I} from this encoder is a 768-dimensional feature vector. We use the clinical BERT model [1] as the text encoder for both report-level and sentence-level feature extraction. The `[CLS]` token is used as the text feature z^{R}, which is a 768-dimensional vector. The MI discriminator for both MINE and CPC is a $1280 \to 1024 \to 512 \to 1$ multilayer perceptron to estimate local MI and a $1536 \to 1024 \to 512 \to 1$ multilayer perceptron to estimate global MI. The image feature and the text feature are concatenated to construct the input for the discriminator for MI estimation. The image models in all training variants at the image training or *re-training* time have the same architecture (6-block resnet followed by a fully connected layer).

The AdamW [30] optimizer is employed for the BERT encoder and the Adam [16] optimizer is used for the other parts of the model. The initial learning rate is $5 \cdot 10^{-4}$. The representation learning phase is trained for 5 epochs and the image model *re-training* phase is trained for 50 epochs. The fully supervised image model is trained for 100 epochs. Data augmentation including random rotation, translation, and cropping is performed on the images during training.

Method	*Re-train* Encoder?	Level 0 vs 1,2,3		Level 0,1 vs 2,3		Level 0,1,2 vs 3	
–	–	CPC	MINE	CPC	MINE	CPC	MINE
image-only	N/A	0.80		0.71		0.90	
global-mi	**frozen**	0.81	0.83	0.77	0.78	0.93	0.89
global-mi	**tuned**	0.81	0.82	0.79	0.81	0.93	0.93
local-mi	**frozen**	0.77	0.76	0.72	0.76	0.75	0.86
local-mi	**tuned**	**0.87**	0.83	0.83	**0.85**	**0.97**	0.93

Table 2. The AUCs on the **EdemaSeverity** ordinal regression task. The average AUC score of **tuned local-mi** is 0.88 (±0.05); The average AUC score of **tuned global-mi** is 0.85 (±0.06).

Method	*Re-train* Encoder?	Atelectasis		Cardiomegaly		Consolidation	
–	–	CPC	MINE	CPC	MINE	CPC	MINE
image-only	N/A	0.76		0.71		0.78	
global-mi	**frozen**	0.65	0.63	0.79	0.79	0.67	0.65
global-mi	**tuned**	0.74	0.77	0.81	0.81	0.81	0.82
local-mi	**frozen**	0.74	0.61	0.73	0.77	0.65	0.65
local-mi	**tuned**	0.73	**0.86**	0.82	**0.84**	**0.83**	**0.83**
–	–	Edema		Lung Opacity		Pleural Effusion	
–	–	CPC	MINE	CPC	MINE	CPC	MINE
image-only	N/A	**0.89**		0.86		0.69	
global-mi	**frozen**	0.81	0.81	0.69	0.68	0.74	0.74
global-mi	**tuned**	0.87	0.88	0.83	0.84	0.90	0.90
local-mi	**frozen**	0.78	0.80	0.66	0.69	0.69	0.72
local-mi	**tuned**	**0.89**	**0.89**	0.82	**0.88**	**0.92**	**0.92**
–	–	Pneumonia		Pneumothorax		Support Devices	
–	–	CPC	MINE	CPC	MINE	CPC	MINE
image-only	N/A	0.75		0.65		0.72	
global-mi	**frozen**	0.71	0.70	0.65	0.66	0.70	0.68
global-mi	**tuned**	0.75	0.76	0.75	0.77	0.77	0.79
local-mi	**frozen**	0.61	0.66	0.70	0.67	0.72	0.74
local-mi	**tuned**	0.78	**0.79**	**0.79**	0.76	**0.87**	0.81

Table 3. The AUCs on the **Pathology9** binary classification tasks. The average AUC score of **tuned local-mi** is 0.84 (±0.05); The average AUC score of **tuned global-mi** is 0.81 (±0.05).

Results. In Table 2 and Table 3, we present the area under the receiver operating characteristic curve (AUC) statistics for the variants of our algorithms on the **EdemaSeverity** classification task and the **Pathology9** binary classification tasks. For most classification tasks, the local MI approach with encoder tuning performs the best and significantly improves the performance over solely supervised learning on labeled images. The local MI approach brings in noteworthy improvement compared to global MI. Both CPC and MINE perform similar in most tasks. Remarkably, the classification results from the frozen encoders approach the fully supervised learning results in many tasks, suggesting that the unsupervised learning captures useful features for image classification tasks even before supervision is provided.

The local MI offers substantial improvement in performance when the features are fine-tuned with the downstream model, while its performance is comparable with global MI if the features are frozen for the subsequent classification. In our experiments, training jointly with the downstream classifier (fine-tuning) typically improves performance of all tasks, with greater benefits for local MI. This suggests that local MI yields more flexible representations that adjust better for the downstream task. Our results are also supported by the analysis in Section 3 that shows certain structural assumptions lead to the local MI approach, which is easier to discover useful representations due to its lower dimension and more training samples.

5 Conclusion

In this paper, we proposed a multimodal representation learning framework for images and text by maximizing the mutual information between their local features. The advantages of the local MI approach are tri-fold: 1) better fit to image-text structure: each sentence is typically a minimal and complete semantic unit that describes a local image region (Fig. 1) and therefore learning at the level of sentences and local image regions is more efficient than learning global descriptors; 2) better optimization landscape: the dimensionality of the representation is lower and every training image-report pair provides more samples of image-text descriptor pairs; 3) better representation fit to downstream tasks: as demonstrated in prior work, image classification usually relies on local features (e.g., pleural effusion detection based on the appearance of the region below the lungs) [10] and thus by learning local representations local MI improves classification performance.

By encouraging sentence-level features in the text to exhibit high MI with local image features, the image encoder learns to extract *useful* feature representations for subsequent image analysis. We provided further insight into local MI by showing that, under a Markov condition, maximizing local MI is equivalent to maximizing global MI. Our experimental results demonstrate that the local MI approach offers the greatest improvement for the downstream image classification tasks, and is not sensitive to the choice of the MI estimator.

Acknowledgments. This work was supported in part by NIH NIBIB NAC P41EB015902, Wistron, IBM Watson, MIT Deshpande Center, MIT J-Clinic, MIT Lincoln Lab, and US Air Force.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D.: MINE: Mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
3. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. In: International Conference on Machine Learning. pp. 517–526. PMLR (2017)
4. Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–539. Springer (2020)
5. Chen, R.T., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. arXiv preprint arXiv:1802.04942 (2018)
6. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems. pp. 2172–2180 (2016)
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
8. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 649–665 (2018)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
11. Horng, S., Liao, R., Wang, X., Dalal, S., Golland, P., Berkowitz, S.J.: Deep learning to quantify pulmonary edema in chest radiographs. Radiology: Artificial Intelligence p. e190228 (2021)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031 (2019)
13. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 1–8 (2019)
14. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). PhysioNet. https://doi.org/10.13026/8360-t248. (2019)
15. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG,

a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)

16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Liao, R., Chauhan, G., Golland, P., Berkowitz, S.J., Horng, S.: Pulmonary edema severity grades based on MIMIC-CXR (version 1.0.1). PhysioNet. https://doi.org/10.13026/rz5p-rc64. (2021)
18. Liao, R., Moyer, D., Golland, P., Wells, W.M.: Demi: Discriminative estimator of mutual information. arXiv preprint arXiv:2010.01766 (2020)
19. Liao, R., Rubin, J., Lam, G., Berkowitz, S., Dalal, S., Wells, W., Horng, S., Golland, P.: Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. arXiv preprint arXiv:1902.10785 (2019)
20. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE transactions on Medical Imaging **16**(2), 187–198 (1997)
21. Moradi, M., Guo, Y., Gur, Y., Negahdar, M., Syeda-Mahmood, T.: A cross-modality neural network transform for semi-automatic medical image annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 300–307. Springer (2016)
22. Moradi, M., Madani, A., Gur, Y., Guo, Y., Syeda-Mahmood, T.: Bimodal network architectures for automatic generation of image annotation from text. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 449–456. Springer (2018)
23. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
24. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
25. Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M.: Disentangling factors of variation for facial expression recognition. In: European Conference on Computer Vision. pp. 808–822. Springer (2012)
26. Song, J., Ermon, S.: Understanding the limitations of variational mutual information estimators. In: International Conference on Learning Representations (2019)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: 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. pp. 9049–9058 (2018)
28. Wang, X., Schwab, E., Rubin, J., Klassen, P., Liao, R., Berkowitz, S., Golland, P., Horng, S., Dalal, S.: Pulmonary edema severity estimation in chest radiographs using deep learning. In: International Conference on Medical Imaging with Deep Learning–Extended Abstract Track (2019)
29. Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R.: Multi-modal volume registration by maximization of mutual information. Medical image analysis **1**(1), 35–51 (1996)
30. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. ArXiv pp. arXiv–1910 (2019)
31. Xue, Y., Huang, X.: Improved disease classification in chest x-rays with transferred features from report generation. In: International Conference on Information Processing in Medical Imaging. pp. 125–138. Springer (2019)

32. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)