Multimodal Representation Learning for Medical Image Analysis

by

Ruizhi Liao

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

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2021

Author . . . .
Department of Electrical Engineering and Computer Science
Aug 26, 2021

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

Multimodal Representation Learning for Medical Image Analysis

by

Ruizhi Liao

Submitted to the
Department of Electrical Engineering and Computer Science
on Aug 26, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

My thesis develops machine learning methods that exploit multimodal clinical data to improve medical image analysis. Medical images capture rich information of a patient's physiological and disease status, central in clinical practice and research. Computational models, such as artificial neural networks, enable automatic and quantitative medical image analysis, which may offer timely diagnosis in low-resource settings, advance precision medicine, and facilitate large-scale clinical research.

Developing such image models demands large training data. Although digital medical images have become increasingly available, limited structured image labels for the image model training have remained a bottleneck. To overcome this challenge, I have built machine learning algorithms for medical image model development by exploiting other clinical data.

Clinical data is often multimodal, including images, text (e.g., radiology reports, clinical notes), and numerical signals (e.g., vital signs, laboratory measurements). These multimodal sources of information reflect different yet correlated manifestations of a subject's underlying physiological processes. I propose machine learning methods that take advantage of the correlations between medical images and other clinical data to yield accurate computer vision models. I use mutual information to capture the correlations and develop novel algorithms for multimodal representation learning by leveraging local data features. The experiments described in this thesis demonstrate the advances of the multimodal learning approaches in the application of chest x-ray analysis.

Thesis Supervisor: Polina Golland
Title: Henry Ellis Warren (1894) Professor of Electrical Engineering and Computer Science

Contents

List of Figures

List of Tables

Chapter 1

Introduction

My thesis proposes machine learning methods that exploit multimodal clinical data to improve medical image analysis. Medical images capture rich information of a patient's physiological processes and pathological conditions, central in clinical practice and research. For example, chest radiographs (CXRs) are used to identify acute and chronic cardiopulmonary conditions, to verify that devices such as pacemakers, central lines, and tubes are correctly positioned, and to assist in related medical workups. Clinical research also relies on medical images to study disease progression and treatment responses. For instance, brain magnetic resonance imaging (MRI) reveals the anatomic structure of the brain and is used to study the progression of Alzheimer's disease.

Computational models, such as artificial neural networks, enable automatic and quantitative medical image analysis, which may offer timely diagnosis in low-resource settings, advance disease phenotyping and precision medicine, and facilitate large-scale clinical studies. Developing such image models, however, requires large training data. Although digital medical images have become increasingly available, limited structured labels for the image model training have remained a bottleneck. For example, to learn a computer vision model that detects pneumonia in CXRs, labels that indicate the presence or absence of the infection are needed. Conventional ways of labeling natural images, such as crowdsourcing, are not suitable for medical image annotation, because interpreting medical images often requires years of professional training and extensive domain knowledge.

To overcome this challenge, I have developed machine learning algorithms for medical image model training by leveraging other clinical data. Clinical data is often multimodal, including images, text (e.g., radiology reports, clinical notes), and numerical signals (e.g., vital signs, laboratory measurements). These multimodal sources of information reflect different yet correlated manifestations of a subject's underlying physiological processes. For example, during routine clinical care, radiologists summarize their findings for a particular image in a free-text radiology report. As another example, B-type natriuretic peptide (BNP) is a hormone that can be measured in a laboratory test. A rising BNP level usually indicates heart failure exacerbation, which often manifests in CXRs as pulmonary edema.

The clinical data along with medical images have been increasingly and widely digitalized and archived in electronic health records (EHR) systems and picture archiving and communication systems (PACS). About 90% of the hospitals in the United States have adopted comprehensive EHR systems, a major advance from 10% in 2008 [2]. The treasure trove of archival healthcare data makes it possible to build computer vision models for medical image analysis without large-scale expensive expert annotations. Over my Ph.D., I have developed machine learning methodologies that take this possibility closer to becoming a reality.

1.1 Thesis Preview

Pulmonary edema is a manifestation of volume status in heart failure, sepsis, and renal failure. In this thesis, we have built baseline computer vision models for pulmonary edema quantification in CXRs [56, 35]. We have developed an image-text joint learning model to improve the accuracy of pulmonary edema assessment from CXRs by leveraging the rich information from free-text radiology reports [13]. We exploit mutual information (MI) to learn image representations jointly with text. Furthermore, we have demonstrated methods that capture the spatial structure in the images and sentence-level text features by maximizing MI to learn features that are *useful* for subsequent analysis of images [54]. Finally, we have proposed a neural network based mutual information estimator and demonstrated empirical advantages of our approach over the state of the art methods for estimating MI in synthetic and real image data. [55].

1.2 Background

Radiology reports capture radiologists' impressions of medical images in the form of unstructured text. While the images possess ground-truth information about the pathologies and disease status of a patient, manual annotation is often time intensive. Therefore, structured labels extracted from radiology reports using rule-based natural language labelers are commonly used as a proxy for ground-truth image labels [74, 38].

In the context of pulmonary edema assessment from chest radiographs, only limited numerical edema severity labels can be extracted from the corresponding reports, which limits the amount of labeled image data we can learn from. This presents a significant challenge for learning accurate image-based models for edema assessment.

I propose and demonstrates a semi-supervised learning algorithm to assess pulmonary edema. Limited ground truth labels are one of the most significant challenges in medical image analysis and many other machine learning applications in healthcare. It is of great practical interest to develop machine learning algorithms that take advantage of the entire data set to improve the performance of strictly supervised classification or regression methods. In this work, we develop a Bayesian model that learns probabilistic feature representations from the entire image set with limited labels for predicting edema severity.

Furthermore, to improve the performance of the image-based model and allow leveraging larger amount of multimodal training data, we make use of free-text reports to include rich information about radiographic findings and reasoning of pathology assessment.

1.3 Contributions

Quantifying pulmonary edema is more challenging than detection of pathology in chest x-ray images because grading of pulmonary edema severity. My PhD work demonstrated the first attempt to employ machine learning algorithms to automatically and quantitatively assess the severity of pulmonary edema from chest x-ray images. That investigation has developed a common, clinically meaningful machine learning task and evaluation framework with baseline performance metrics to benchmark future algorithmic developments in grading pulmonary edema severity from CXRs [56, 35].

Our image-text joint modeling approach was the first method to leverage the free-text radiology reports for improving the image model performance in the application of assessing pulmonary edema [13]. Our experimental results demonstrate that the joint representation learning framework improves the accuracy of edema severity estimates over a purely image-based model on a fully labeled subset of the data (supervised). Furthermore, I propose to exploit the image spatial structure and sentence-level text features with mutual information maximization to learn features that are *useful* for subsequent analysis of images [54]. In our experimental results, we demonstrate that the maximization of local MI yields the greatest improvement in the downstream image classification tasks.

Finally, we have proposed a neural network based mutual information estimator and demonstrated empirical advantages of our approach over the state of the art methods for estimating MI in synthetic and real image data. [55]. We show theoretically that our method and other variational approaches are equivalent when they achieve their optimum, while our method sidesteps the variational bound. Empirical results demonstrate high accuracy of our approach and the advantages of our estimator in the context of representation learning.

The contributions of my thesis have been published in the following papers:

- R. Liao, J. Rubin, G. Lam, S. Berkowitz, S. Dalal, W. Wells, S. Horng, P. Golland. Semi-supervised Learning for Quantification of Pulmonary Edema in

Chest X-Ray Images. *arXiv:1902.10785.* [56]

- S. Horng*, R. Liao*, X. Wang, S. Dalal, P. Golland, S. Berkowitz. Deep Learning to Quantify Pulmonary Edema in Chest Radiographs. *Radiology: Artificial Intelligence.* (* indicates equal contributions.) [35]

- G. Chauhan*, R. Liao*, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, P. Golland. Joint Modeling of Chest Radiographs and Radiology Reports for Pulmonary Edema Assessment. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020.* (* indicates equal contributions.) [13]

- R. Liao, D. Moyer, M. Cha, K. Quigley, S. Berkowitz, S. Horng, P. Golland, W. Wells. Multimodal Representation Learning via Maximization of Local Mutual Information.. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021.* [54]

- R. Liao, D. Moyer, P. Golland, W. Wells. DEMI: Discriminative Estimator of Mutual Information. *arXiv:2010.01766.* [55].

1.4 Roadmap

This thesis is organized as follows. In the next chapter, we describe the motivating clinical problem, the clinical data we have collected for model development, and the annotation processes. In chapter 3, we present baseline computer vision models for pulmonary edema assessment in chest radipgraphs. In chapter 4, we propose an image-text joint learning model that leverages the free-text radiology reports to improve the iamge model performance. In the following chapter, we further improve joint representation learning approach by associating the local features of images and text using mutual information. In chapter 6, we derive a simple and accurate estimator of mutual information. Finally, I discuss the future directions that my Ph.D. research has opened up.

Chapter 2

Motivating Problem, Data and Annotations

2.1 Motivating Clinical Problem

Heart failure (HF) is the leading cause of hospitalization in the US, with high readmission and mortality rates. About 6.5 million American adults live with chronic heart failure [10]. In the US alone, there are around 1.1 million emergency room (ER) visits and 1 million hospitalizations with HF as the primary cause every year. In addition, there are 4.1 million ER visits and 3.4 million hospitalizations with HF as a comorbid or contributing cause [39]. Most of the acute HF patients present with pulmonary edema (fluid overload in their lungs) and exhibit heterogeneous responses to treatment. This heterogeneity precludes effective treatment, which in turn leads to long hospital stays. Clinical practice guidelines recommend removal of all excess fluid prior to discharge; patients discharged with remaining fluid overload are more likely to be readmitted or die post-discharge [80]. Yet, accurate assessment of fluid status is recognized to be exceptionally challenging [64]. Close to 20% of the patients are readmitted within 30 days of discharge from the hospital and up to 30% within 3 months [81].

Clinical management decisions for patients with acutely decompensated heart failure are often based on grades of pulmonary edema severity, rather than its mere absence or presence. Clinicians often monitor changes in pulmonary edema severity to assess the efficacy of therapy. Accurate monitoring of pulmonary edema is essential when competing clinical priorities complicate clinical management. For example, a heart failure patient with a severe infection causing septic shock may have pulmonary edema driven both by volume overload due to heart failure and increased capillary permeability. This patient will likely be intravascularly depleted from their septic shock, but also total body volume overloaded, leading to pulmonary edema. The patient simultaneously needs both more fluid to optimize their hemodynamic function and less fluid to optimize their respiratory function. Often referred to as the ebb and flow of sepsis, patients need judicious fluid resuscitation early in their clinical course, and evacuation of fluid through diuresis later in their course [40, 63]. The accurate assessment of pulmonary edema is critical to maintaining this delicate fluid balance.

The assessment of patient fluid status is flawed and inaccurate in current clinical practice, which leads to poor clinical decision making for heart failure. In particular, bedside examination and evaluation based on symptoms are subjective and inconsistent. Laboratory tests of certain biomarkers (e.g., BNP or creatinine) are indirect and subject to confounding pathophysiological processes. Chest radiographs are commonly acquired to assess pulmonary edema. Unfortunately, the assessments of pulmonary edema severity based on chest x-ray images are inconsistent across practitioners (with a kappa inter-rater agreement of 0.67 among experienced radiologists) and even across different reads by the same practitioner [27]. Other surrogates for patient fluid status and response to treatment are either noisy (such as body weight and urine output) or require an invasive procedure (Swan-Ganz catheterization to measure pulmonary capillary wedge pressure).

Decompensated HF patients have heterogeneous responses to treatment [23], and that response is highly predictive of their clinical trajectory. Unfortunately, this response to treatment is poorly documented in the medical record, limiting the ability of researchers to discover important relationships between treatments and effects. Other surrogates for response to treatment such as urine output, total body fluid balance, and daily weights have been suggested, but are often not accurately and consistently measured.

Although improvement in dyspnea correlates with radiographic improvement, critically ill patients cannot provide this information and subjective information is not well quantified. The automatic and quantitative assessment for pulmonary edema severity will enable clinicians to make better treatment plans based on prior patient responses and will also enable clinical research studies that require quantitative phenotyping of patient status [12].

Beyond heart failure management, the quantification of pulmonary edema on chest radiographs is useful throughout clinical medicine. Pulmonary edema is a manifestation of volume status in sepsis and renal failure, just as in HF. Managing volume status is critical in the treatment of sepsis, but large-scale research has been limited due to lack of longitudinal data on volume status. Quantification of pulmonary edema in a chest radiograph could be used as a surrogate for volume status, which would rapidly advance research in sepsis and other disease processes where volume status is critical.

2.2 Data Collection

Chest radiographs are commonly performed to assess pulmonary edema [61]. The signs of pulmonary edema on chest radiographs have been known for over 50 years [28, 57]. The grading of pulmonary edema is based on well-known radiologic findings on chest radiographs [66, 91, 72, 86]. The symptom of dyspnea caused by pulmonary edema is the most common reason a patient with acute decompensated congestive heart failure (CHF) seeks care in the emergency department and is ultimately admitted to the hospital (89% of patients) [25, 37, 1]. Clinical management decisions for patients with acutely decompensated CHF are often based on grades of pulmonary edema severity, rather than its mere absence or presence. Clinicians often monitor changes in pulmonary edema severity to assess the efficacy of therapy. Accurate monitoring of pulmonary edema is essential when competing clinical priorities complicate clinical management. While we focus on patients with CHF within this study, the quantification of pulmonary edema on chest radiographs is useful throughout clinical medicine.

Large-scale and common datasets have been the catalyst for the rise of machine learning today [19]. In 2019, investigators released MIMIC-CXR, a large-scale publicly available chest radiograph dataset [44, 45, 26]. My PhD work described in this chapter builds upon that prior work by developing a common, clinically meaningful machine learning task and evaluation framework with baseline performance metrics to benchmark future algorithmic developments in grading pulmonary edema severity from chest radiographs. We developed image models using two common machine learning approaches: a semi-supervised learning model and a supervised learning model pre-trained on a large common image dataset.

2.3 Data and Label Extraction

This was a retrospective cohort study. This study was approved by the Beth Israel Deaconess Medical Center Committee on Clinical Investigation with a waiver of informed consent. We collected 369,071 chest radiographs and their associated radiology reports from 64,581 patients from the MIMIC-CXR chest radiograph dataset [44, 45, 26]. Each imaging study is associated with one or more images. We aimed to identify patients with CHF within the dataset to limit confounding labels from other disease processes. First, we limited our study to only frontal radiographs, excluding a total of 121,646 images. Of these frontal radiographs ($n = 247,425$), there were 17,857 images which were acquired during visits with an emergency department discharge diagnosis code consistent with CHF. In total, this resulted in 16,108 radiology reports and 1,916 patients that were included that had CHF. As part of a prior study [102], we manually reviewed patient charts and found this method of cohorting patients with CHF had 100% sensitivity and specificity. The other 62,665 patients were classified as non-CHF and data was used in the semi-supervised training model. An enrollment diagram is shown in Figure 2-1.

We extracted the pulmonary edema severity labels (“none”, “vascular congestion”,

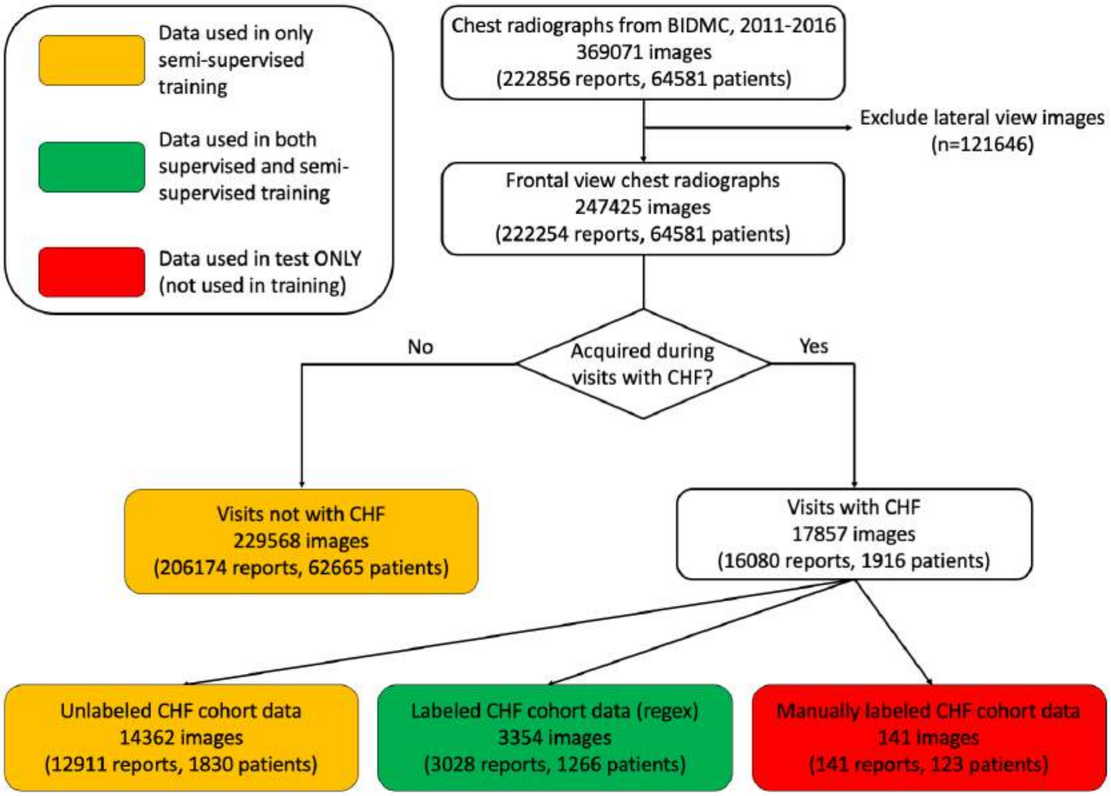

Data used in only semi-supervised training
Data used in both supervised and semi-supervised training
Data used in test ONLY (not used in training)
Chest radiographs from BIDMC, 2011-2016
369071 images
(222856 reports, 64581 patients)
Exclude lateral view images (n=121646)
Frontal view chest radiographs
247425 images
(222254 reports, 64581 patients)
Acquired during visits with CHF?
No
Yes
Visits not with CHF
229568 images
(206174 reports, 62665 patients)
Visits with CHF
17857 images
(16080 reports, 1916 patients)
Unlabeled CHF cohort data
14362 images
(12911 reports, 1830 patients)
Labeled CHF cohort data (regex)
3354 images
(3028 reports, 1266 patients)
Manually labeled CHF cohort data
141 images
(141 reports, 123 patients)

Figure 2-1: Cohort selection flowchart. A total of 369,071 chest radiographs and their associated radiology reports from 62,665 patients were collected. Images for this study were limited to frontal view radiographs (247,425). Of the 247,425 frontal view radiographs, 17,857 images were acquired during visits with a diagnosis consistent with congestive heart failure (CHF). In the CHF cohort, we were able to label 3,028 radiology reports and thus 3354 frontal view radiographs from 1,266 patients, using regular expressions (regex) on the reports. We also curated a test set of 141 radiographs that were manually labeled by radiologists (from the 650 unlabeled radiographs from patients with CHF). BIDMC = Beth Israel Deaconess Medical Center.

"interstitial edema", and "alveolar edema") from the reports using regular expressions with negation detection. The extracted labels were numerically coded as follows: 0, none; 1, vascular congestion; 2, interstitial edema; and 3, alveolar edema (Table 2.1). Examples of the grades are shown in Figure 2-2. We were able to label 3,028 radiology reports and thus 3,354 frontal view radiographs from 1,266 patients (Figure 2-1). Among the 1,266 patients, 1,180 patients still have some of their reports unlabeled. The other 650 patients with CHF had no labeled reports.

To validate our label extraction in radiology reports, we randomly selected 485

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

Table 2.1: Validation of regex keyword terms. The accuracy (positive predictive value) of the regular expression results for levels 0-3 based on the expert review results are 90.74%, 80.61%, 95.24%, and 90.91%, respectively. The total number of reports from all the keywords is more than 485 because some reports contain more than one keywords.

labeled reports. A board-certified radiologist (SB, 5 years of experience, interventional radiology) and two domain experts then manually labeled the 200 reports, blinded from our label extraction results. We report the accuracy for each category and each keyword.

We had three senior radiology residents and one attending radiologist manually label a set of 141 frontal view radiographs from 123 patients (from the unlabeled dataset of 650 patients with CHF), which had no patient overlap with the report labeled set (Figure 2-3). These images were set aside as our test set. Each radiologist assessed the images independently and we report their inter-rater agreement (Fleiss' Kappa).

We performed a modified Delphi consensus process to develop a gold standard image label. We had 3 senior radiology residents and 1 attending radiologist manually label a set of 141 frontal view chest radiographs from 123 patients. The three residents labeled the images independently. If the three residents had exactly the same pulmonary edema severity of an image, then a consensus label is assigned. If only two out of the three residents agreed on the edema severity, then an attending radiologist reviewer was added. If a majority of the reviewers (three out of four) now agreed, then a consensus label is assigned. If no consensus was reached, then the four radiologists discussed their interpretations in a round-robin process, and then again voted anonymously on their edema severity levels. If a majority of the votes was reached, then a consensus label is assigned. If no consensus was reached, then another round-robin discussion is performed with another anonymous vote. This process is then repeated one additional time, and if no consensus is reached, then the image is labelled as no consensus. The flowchart of the consensus process is shown in the Figure 2-3.

To understand how many and how frequently chest radiographs have been taken on our CHF cohort and non-CHF cohort, we calculated the number of images from each patient in our dataset and plotted the histograms of the numbers for the CHF cohort and for the non-CHF cohort, shown in Figure 2-4. We also showed, in Figure 2-4, the distributions of time intervals between two consecutive chest radiographs taken on a patient with CHF.

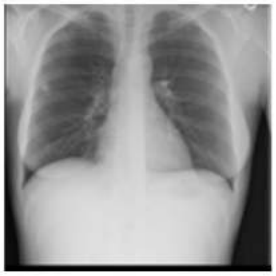
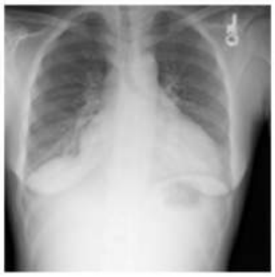
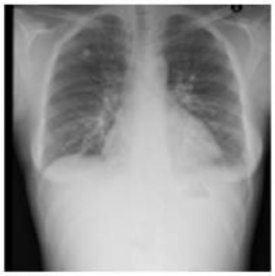
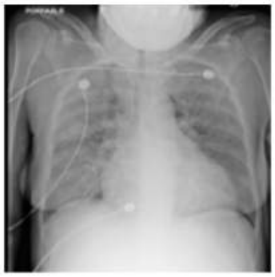
Pulmonary Edema Severity Level Label	Pathophysiology	Representative Chest Radiograph	Common Radiographic Findings	Pulmonary Capillary Wedge Pressure
0	None			
1	Vascular congestion		cephalization, pulmonary vascular congestion, hilar vascular indistinctness	13-18 mm Hg
2	Interstitial edema		increased interstitial markings, Kerley B lines, peribronchial cuffing	18-25 mm Hg
3	Alveolar edema		bilateral, symmetric, airspace opacities radiating centrally from the hila pleural effusion	>25 mm Hg

Figure 2-2: Representative images and radiographic findings of each pulmonary edema severity level.

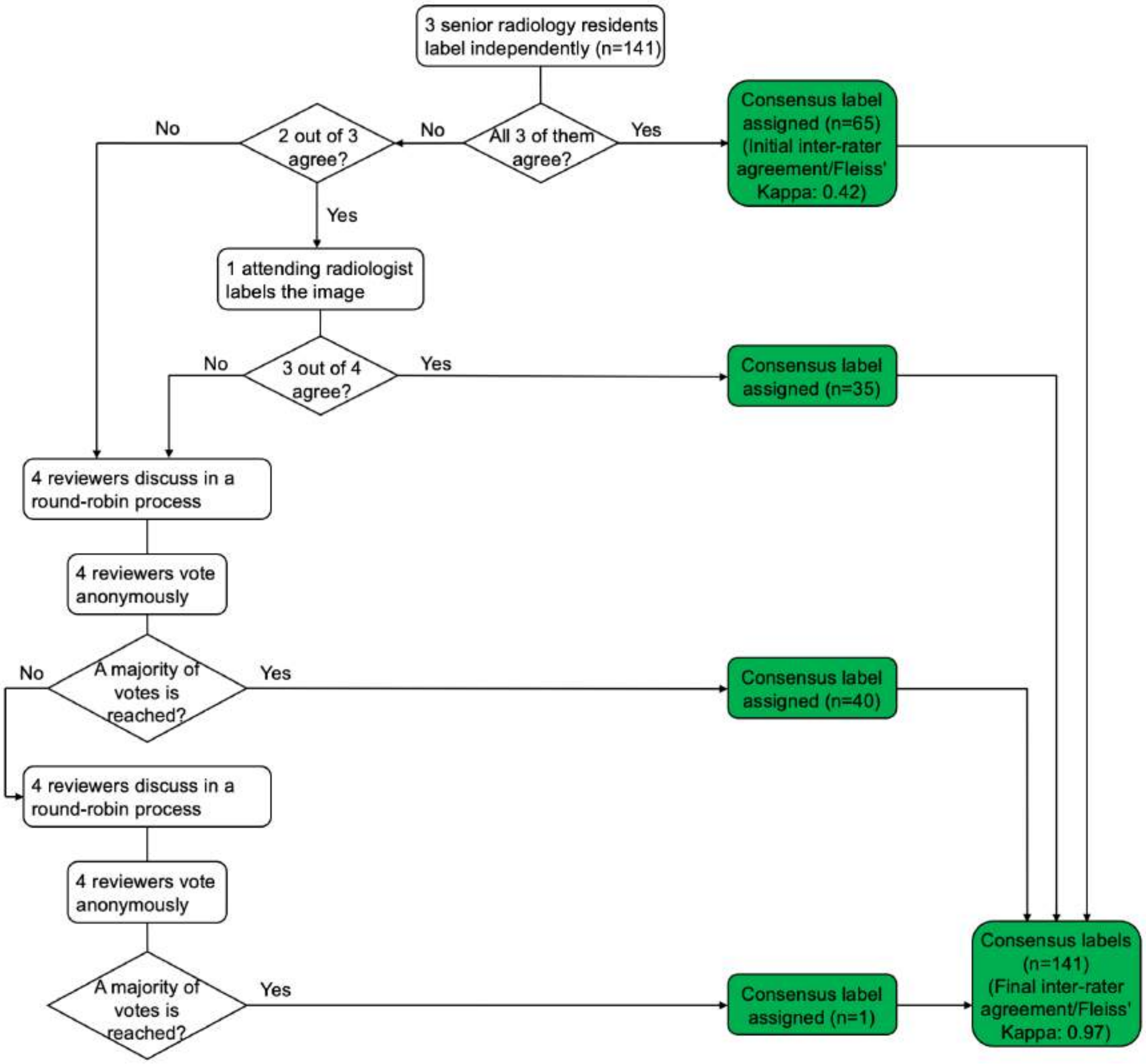

3 senior radiology residents label independently (n=141)
All 3 of them agree?
Yes
No
Consensus label assigned (n=65) (Initial inter-rater agreement/Fleiss' Kappa: 0.42)
2 out of 3 agree?
No
Yes
1 attending radiologist labels the image
3 out of 4 agree?
No
Yes
Consensus label assigned (n=35)
4 reviewers discuss in a round-robin process
4 reviewers vote anonymously
A majority of votes is reached?
No
Yes
Consensus label assigned (n=40)
4 reviewers discuss in a round-robin process
4 reviewers vote anonymously
A majority of votes is reached?
Yes
Consensus label assigned (n=1)
Consensus labels (n=141) (Final inter-rater agreement/Fleiss' Kappa: 0.97)

Figure 2-3: The flowchart of our consensus image labeling process. The initial labels independently provided by the 3 senior radiology residents against the final consensus labels have quadratic-weighted Kappa values of 0.83, 0.74, and 0.72. The predictions from the semi-supervised learning model and the pre-trained supervised learning model against the final consensus labels have quadratic-weighted Kappa values of 0.70 and 0.41.

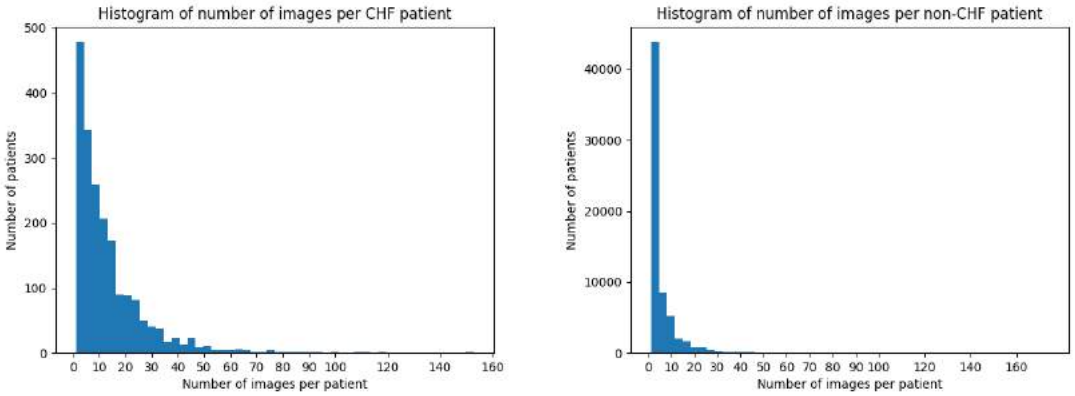


Figure 2-4: Chest radiograph distributions. Histograms of the number of images per CHF patient and per non-CHF patient. On average, 13.78 chest radiographs were taken per CHF patient and 5.43 chest radiographs were taken per non-CHF patient in our dataset. The median number of chest radiographs taken per CHF patient is 9 (ranging from 1 to 153) and per non-CHF patient is 3 (ranging from 1 to 174).

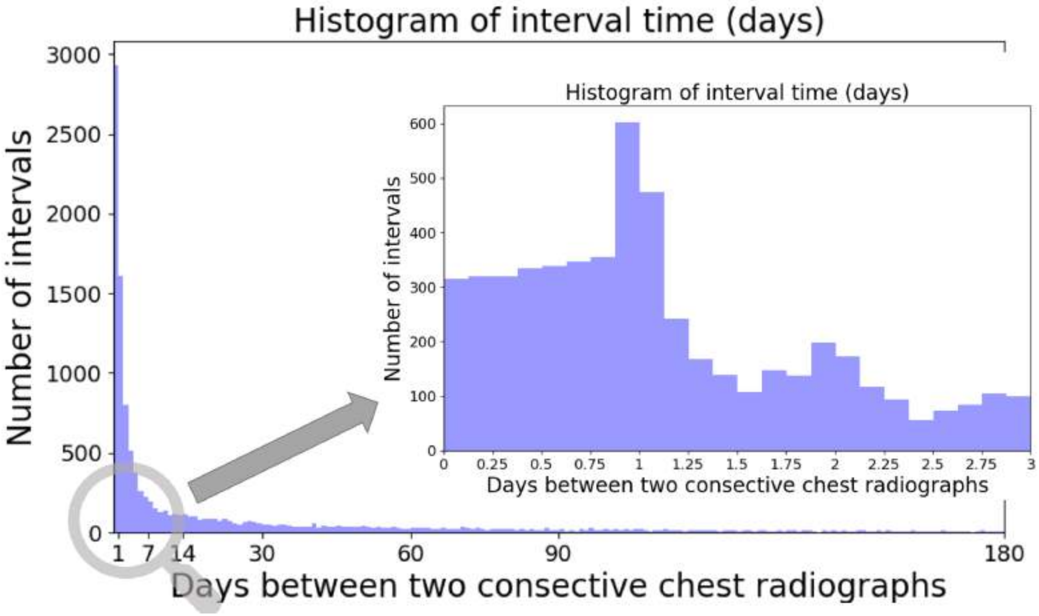


Figure 2-5: Chest radiograph distributions. Distributions of time intervals between serial chest radiographs in CHF cohort. The x-axis is in log scale. The mean interval time between each two consecutive chest radiographs of the same CHF patient is 71.34 days. The median interval time between each two consecutive chest radiographs of the same CHF patient is 7.09 days (ranging from 180 minutes to 1545.84 days). 21.53% of the interval times for CHF patients are within 1 day and 66.08% are within 30 days.

2.4 Summary

In summary, the MIMIC-CXR dataset that consists of 377,110 chest radiographs with free-text radiology reports offers a tremendous opportunity to develop computer vision models for chest radiographs analysis, such as pulmonary edema assessment. We have curated the labels of pulmonary edema severity grades extracted from the MIMIC-CXR dataset through three different means:

- Regular expression (regex) from radiology reports. Regex was able to label 6710 radiology reports.

- Expert labeling from radiology reports. A board-certified radiologist and two domain experts have read 485 radiology reports and give pulmonary edema severity grades based on the reports.

- Consensus labeling from chest radiographs. Three senior radiology residents and one attending radiologist have labeled 141 chest radiographs. This label set is the highest-quality among the three sets, and we recommend holding it out for testing.

The curated labels are publicly available on PhysioNet [53], aiming to support the algorithmic development of pulmonary edema assessment from chest x-ray images and benchmark its performance. For the experiments performed in this thesis, we used regex labels for model training, and used expert labels (from reports) and consensus labels (from images) for model evaluation.

Chapter 3

Image-based Models for Edema Assessment

In order to establish a baseline performance benchmark for this clinical machine learning task and to address the challenge of limited pulmonary edema labels, we developed a computer vision model in a semi-supervised learning fashion using a variational autoencoder. This approach aims to address the challenge of limited pulmonary edema labels. The semi-supervised model takes advantage of the chest radiographs without pulmonary edema severity labels, which includes approximately 220,000 images (from individuals with and without CHF) and is domain specific. As another baseline benchmark, we also built and evaluated a pre-trained image model.

3.1 Methods

Let $x \in \mathbb{R}^{n \times n}$ be a 2D x-ray image and $y \in \{0, 1, 2, 3\}$ be the corresponding edema severity label. Our dataset includes a set of N images $\mathbf{x} = \{x_i\}_{i=1}^{N}$ with the first N_{L} images annotated with severity labels $\mathbf{y} = \{y_i\}_{i=1}^{N_{\mathrm{L}}}$. Here, we derive a learning algorithm that constructs a compact probabilistic feature representation z that is learned from all images and used to predict pulmonary edema severity. Fig. 3-1 illustrates the Bayesian model and the inference algorithm.

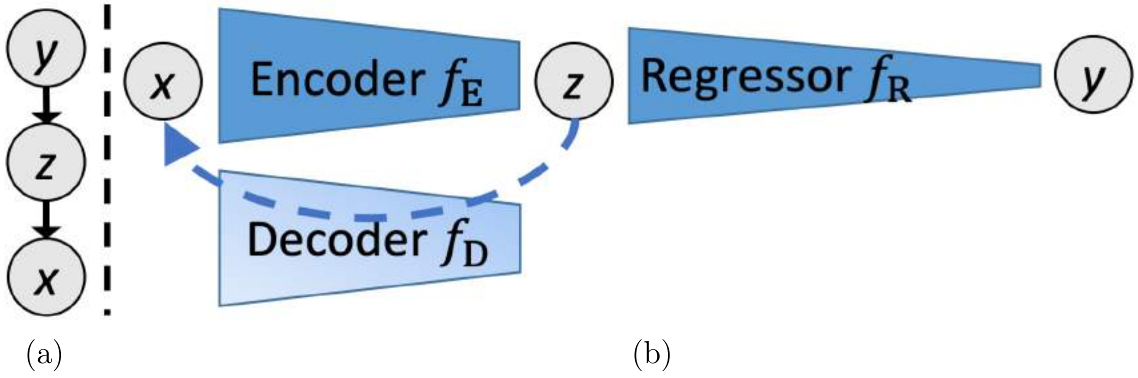


Figure 3-1: Graphical model and inference algorithm. (a): Probabilistic graphical model, where x represents chest x-ray image, z represents latent feature representation, and y represents pulmonary edema severity. (b): Our computational model. We use neural networks for implementing the encoder, decoder, and regressor. The dashed line (decoder) is used in training only. The network architecture is provided in the supplementary material.

3.1.1 Learning

The learning algorithm maximizes the log probability of the data with respect to parameters θ:

$$\log p(\mathbf{x}, \mathbf{y}; \theta) = \sum_{i=1}^{N_{\mathrm{L}}} \log p(\mathrm{x}_i, \mathrm{y}_i; \theta) + \sum_{i=N_{\mathrm{L}}+1}^{N} \log p(\mathrm{x}_i; \theta). \tag{3.1}$$

We model z as a continuous latent variable with a prior distribution $p(z)$, which generates images and predicts pulmonary edema severity. Unlike [48] that constructs a separate encoder $q(z|x, y)$ for each value of discrete label y, we use a single encoder $q(z|x)$ to capture image structure relevant to labels. Distribution $q(z|x)$ serves as a variational approximation for $p(z|x, y)$ for the lower bound:

$$\begin{aligned}
\mathcal{L}_1(\theta; \mathrm{x}_i, \mathrm{y}_i) &= \log p(\mathrm{x}_i, \mathrm{y}_i; \theta) - D_{KL}(q(\mathrm{z}_i|\mathrm{x}_i; \theta)||p(\mathrm{z}_i|\mathrm{x}_i, \mathrm{y}_i)), \\
&= \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta)} \left[\log p(\mathrm{x}_i, \mathrm{y}_i; \theta) + \log p(\mathrm{z}_i|\mathrm{x}_i, \mathrm{y}_i) - \log q(\mathrm{z}_i|\mathrm{x}_i; \theta) \right] \\
&= \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta)} \left[\log p(\mathrm{x}_i, \mathrm{y}_i|\mathrm{z}_i; \theta) + \log p(\mathrm{z}_i) - \log q(\mathrm{z}_i|\mathrm{x}_i; \theta) \right] \\
&= \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta)} \left[\log p(\mathrm{x}_i, \mathrm{y}_i|\mathrm{z}_i; \theta) \right] - D_{KL}(q(\mathrm{z}_i|\mathrm{x}_i; \theta)||p(\mathrm{z}_i)).
\end{aligned}$$

We assume that x, z, and y form a Markov chain, i.e., $y \perp\!\!\!\perp x \mid z$, and therefore

$$\begin{aligned}\mathcal{L}_1(\theta; \mathrm{x}_i, \mathrm{y}_i) =& \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})}\big[\log p(\mathrm{x}_i|\mathrm{z}_i;\theta_\mathrm{D})\big] + \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})}\big[\log p(\mathrm{y}_i|\mathrm{z}_i;\theta_\mathrm{R})\big] \\ &- D_{KL}(q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})||p(\mathrm{z}_i)),\end{aligned} \tag{3.2}$$

where θ_E are the parameters of the encoder, θ_D are the parameters of the decoder, and θ_R are the parameters of the regressor. Similarly, we have a variational lower bound for $\log p(\mathrm{x}_i;\theta)$:

$$\mathcal{L}_2(\theta;\mathrm{x}_i) = \mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})}\big[\log p(\mathrm{x}_i|\mathrm{z}_i;\theta_\mathrm{D})\big] - D_{KL}(q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})||p(\mathrm{z}_i)). \tag{3.3}$$

By substituting Eq. (3.2) and Eq. (3.3) into Eq. (3.1), we obtain a lower bound for the log probability of the data and aim to minimize the negative lower bound:

$$\begin{aligned}\mathcal{J}(\theta;\mathbf{x},\mathbf{y}) =& -\sum_{i=1}^{N_\mathrm{L}}\mathcal{L}_1(\theta;\mathrm{x}_i,\mathrm{y}_i) - \sum_{i=N_\mathrm{L}+1}^{N}\mathcal{L}_2(\theta;\mathrm{x}_i) \\ =& \sum_{i=1}^{N} D_{KL}(q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})||p(\mathrm{z}_i)) - \sum_{i=1}^{N_\mathrm{L}}\mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})}\big[\log p(\mathrm{y}_i|\mathrm{z}_i;\theta_\mathrm{R})\big] \\ &- \sum_{i=1}^{N}\mathbb{E}_{q(\mathrm{z}_i|\mathrm{x}_i;\theta_\mathrm{E})}\big[\log p(\mathrm{x}_i|\mathrm{z}_i;\theta_\mathrm{D})\big].\end{aligned} \tag{3.4}$$

3.1.2 Latent Variable Prior

We let the latent variable prior $p(z)$ be a multivariate normal distribution, which serves to regularize the latent representation of images.

3.1.3 Latent Representation

We apply the reparameterization trick used in [49]. Conditioned on image x_i, the latent representation becomes a multivariate Gaussian variable, $\mathrm{z}_i|\mathrm{x}_i \sim \mathcal{N}(\mathrm{z}_i;\mu_i,\Lambda_i)$, where μ_i is a D-dimensional vector $[\mu_{ik}]_{k=1}^{D}$ and Λ_i is a diagonal covariance matrix represented by its diagonal elements as $[\lambda_{ik}^2]_{k=1}^{D}$. Thus, the first term in Eq. (3.4) becomes:

$$\mathcal{J}_{KL}(\theta_{\mathrm{E}}; \mathrm{x}_i) = -\frac{1}{2}\sum_{k=1}^{D}\left(\log \lambda_{ik}^2 - \mu_{ik}^2 - \lambda_{ik}^2\right) + \text{const.} \tag{3.5}$$

We implement the encoder as a neural network $f_{\mathrm{E}}(x; \theta_{\mathrm{E}})$ that estimates the mean and the variance of $z|x$. Samples of z can be readily generated from this estimated Gaussian distribution. We use one sample per image for training the model.

3.1.4 Ordinal Regression

In radiology reports, pulmonary edema severity is categorized into four groups: no/mild/ moderate/severe. Our goal is to assess the severity of pulmonary edema as a continuous quantity. We employ ordinal representation to capture the ordering of the categorical labels. We use a 3-bit representation $\mathrm{y}_i = [\mathrm{y}_{ij}]_{j=1}^3$ for the four severity levels. The three bits represent the probability of any edema, of moderate or severe edema, and of severe edema respectively (i.e., “no" is $[0,0,0]$, “mild" is $[1,0,0]$, “moderate" is $[1,1,0]$, and “severe" is $[1,1,1]$). This encoding yields probabilistic output, i.e., both the estimate of the edema severity and also uncertainty in the estimate. The three bits are assumed to be conditionally independent given the image:

$$p(\mathrm{y}_i|\mathrm{z}_i; \theta_{\mathrm{R}}) = \prod_{j=1}^{3} f_{\mathrm{R}}^j(\mathrm{z}_i; \theta_{\mathrm{R}})^{\mathrm{y}_{ij}} \left(1 - f_{\mathrm{R}}^j(\mathrm{z}_i; \theta_{\mathrm{R}})\right)^{1-\mathrm{y}_{ij}},$$

where y_{ij} is a binary label and $f_{\mathrm{R}}^j(\mathrm{z}_i; \theta_{\mathrm{R}})$ is interpreted as the conditional probability $p(\mathrm{y}_{ij} = 1|\mathrm{z}_i)$. $f_{\mathrm{R}}(\cdot)$ is implemented as a neural network. The second term in Eq. (3.4) becomes the cross entropy:

$$\mathcal{J}_{\mathrm{R}}(\theta_{\mathrm{E}}, \theta_{\mathrm{R}}; \mathrm{y}_i, \mathrm{z}_i) = -\sum_{j=1}^{3} \mathrm{y}_{ij} \log f_{\mathrm{R}}^j(\mathrm{z}_i; \theta_{\mathrm{R}}) - \sum_{j=1}^{3}(1 - \mathrm{y}_{ij}) \log\left(1 - f_{\mathrm{R}}^j(\mathrm{z}_i; \theta_{\mathrm{R}})\right). \tag{3.6}$$

3.1.5 Decoding

We assume that image pixels are conditionally independent (Gaussian) given the latent representation. Thus, the third term in Eq. (3.4) becomes:

$$\begin{aligned}\mathcal{J}_{\mathrm{D}}(\theta_{\mathrm{E}}, \theta_{\mathrm{D}}; \mathrm{x}_i, \mathrm{z}_i) &= -\log \mathcal{N}(\mathrm{x}_i; f_{\mathrm{D}}(\mathrm{z}_i; \theta_{\mathrm{D}}), \Sigma_i) \\ &= \frac{1}{2}(\mathrm{x}_i - f_{\mathrm{D}}(\mathrm{z}_i; \theta_{\mathrm{D}}))^T \Sigma_i^{-1} (\mathrm{x}_i - f_{\mathrm{D}}(\mathrm{z}_i; \theta_{\mathrm{D}})) + \text{const.},\end{aligned} \tag{3.7}$$

where $f_{\mathrm{D}}(\cdot)$ is a neural network decoder that generates an image implied by the latent representation z, and Σ_i is a diagonal covariance matrix.

3.1.6 Loss Function

Combining Eq. (3.5), Eq. (3.6) and Eq. (3.7), we obtain the loss function for training our model:

$$\begin{aligned}\mathcal{J}(\theta_{\mathrm{E}}, \theta_{\mathrm{R}}, \theta_{\mathrm{D}}; \mathbf{x}, \mathbf{y}) =& \sum_{i=1}^{N} \mathcal{J}_{KL}(\theta_{\mathrm{E}}; \mathrm{x}_i) + \sum_{i=1}^{N_{\mathrm{L}}} \mathcal{J}_{\mathrm{R}}(\theta_{\mathrm{E}}, \theta_{\mathrm{R}}; \mathrm{y}_i, \mathrm{z}_i) + \sum_{i=1}^{N} \mathcal{J}_{\mathrm{D}}(\theta_{\mathrm{E}}, \theta_{\mathrm{D}}; \mathrm{x}_i, \mathrm{z}_i) \\ =& -\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{D} \left(\log \lambda_{ik}^2 - \mu_{ik}^2 - \lambda_{ik}^2 \right) \\ & - \sum_{i=1}^{N_{\mathrm{L}}} \left(\sum_{j=1}^{3} \mathrm{y}_{ij} \log f_{\mathrm{R}}^{j}(\mathrm{z}_i; \theta_{\mathrm{R}}) + \sum_{j=1}^{3} (1 - \mathrm{y}_{ij}) \log \left(1 - f_{\mathrm{R}}^{j}(\mathrm{z}_i; \theta_{\mathrm{R}}) \right) \right) \\ & + \frac{1}{2} \sum_{i=1}^{N} \left(\mathrm{x}_i - f_{\mathrm{D}}(\mathrm{z}_i; \theta_{\mathrm{D}}) \right)^T \Sigma_i^{-1} (\mathrm{x}_i - f_{\mathrm{D}}(\mathrm{z}_i; \theta_{\mathrm{D}})) \,.\end{aligned} \tag{3.8}$$

We employ the stochastic gradient-based optimization procedure Adam [47] to minimize the loss function. Our training procedure is outlined in the supplementary materiel. The pulmonary edema severity category extracted from radiology reports is a discrete approximation of the actual continuous severity level. To capture this, we compute the expected severity:

$$\hat{y} = 0 \times (1 - \hat{y}_1) + 1 \times (\hat{y}_1 - \hat{y}_2) + 2 \times (\hat{y}_2 - \hat{y}_3) + 3 \times \hat{y}_3 = \hat{y}_1 + \hat{y}_2 + \hat{y}_3.$$

3.2 Implementation Details

The size of the chest x-ray images in our dataset varies and is around 3000×3000 pixels. We randomly rotate and translate the images (differently at each epoch) on the fly during training and crop them to 2048×2048 pixels as part of data augmentation. We maintain the original image resolution to preserve subtle differences between different levels of pulmonary edema severity.

The encoder is implemented as a series of residual blocks [32]. The decoder is implemented as a series of transposed convolutional layers, to build an output image of the same size as the input image (2048×2048). The regressor is implemented as a series of residual blocks with an averaging pooling layer followed by two fully connected layers. The regressor output $\hat{y}$ has 3 channels. The latent representation z has a size of 128×128. During training, one sample is drawn from z per image. The KL-loss (Eq. (3.5)) and the image reconstruction error (Eq. (3.7)) in the loss function are divided by the latent feature size and the image size respectively. The variances in Eq. (3.7) are set to 10, which gives a weight of 0.1 to the image reconstruction error. The learning rate for the Adam optimizer training is 0.001 and the minibatch size is 4. The model is trained on a training dataset and evaluated on a separate validation dataset every few epochs during training. The model checkpoint with the lowest error on the validation dataset is used for testing. The neural network architecture is provided in the supplementary material.

As another baseline model, we started with a neural network that had been pre-trained to recognize common images (e.g., cats and dogs) and then further tuned it to recognize the specific image features of chest radiographs for assessing pulmonary edema. Specifically, we use the densely connected convolutional neural networks (DenseNet) and the model is pre-trained on ImageNet. The DenseNet has four dense blocks, which consist of 6, 12, 24, 16 convolutional layers respectively. The four dense blocks are concatenated with a 2-by-2 averaging pooling layer between each two consecutive dense blocks. We keep the first three pre-trained dense blocks for low-level image feature extraction, followed by one global average pooling layer, one dropout layer and two fully connected layers. We then re-trained this model on our labeled chest radiographs. We also use data augmentation by random image translation, rotation, and cropping to a size of 512 × 512 (for adjusting the image size in the ImageNet) during training in order to improve the robustness of the model.

3.3 Experiments

3.3.1 Receiver Operating Characteristics Curve Analysis

The receiver operating characteristic (ROC) curves of the two models on the test set are shown in Figure3-2. As expected, both models perform well on the task of distinguishing images between level 0 and level 3 and on the task of classifying between level 3 and the rest. The AUC for differentiating alveolar edema (score 3) from no edema (score 0) was 0.99 and 0.87 for semi-supervised and pre-trained models, respectively. Performance of the algorithm was inversely related to the difficulty in categorizing milder states of pulmonary edema (shown as the AUC for the semi-supervised and pretrained model, respectively, for differentiating the following categories): 2 versus 0, 0.88 and 0.81; 1 versus 0, 0.79 and 0.66; 3 versus 1, 0.93 and 0.82; 2 versus 1, 0.69 and 0.73); 3 versus 2, 0.88 and 0.63.

3.3.2 Confusion Matrix Analysis

We computed a confusion matrix for each of the models on the test set (Figure 3-3). Each image was placed in a cell by the true severity level from consensus score and the predicted severity level from the image model. In each cell, we reported the fraction of the predicted severity level in the actual severity level. Both models performed better in predicting level 0 and level 3 compared to predicting level 1 and level 2.

3.3.3 Predicted Edema Severity in Bar Charts

We plotted bar charts of predicted edema severity versus true edema severity on the test set (Figure 3-4). Both plots show the linear trend of predicted edema severity with

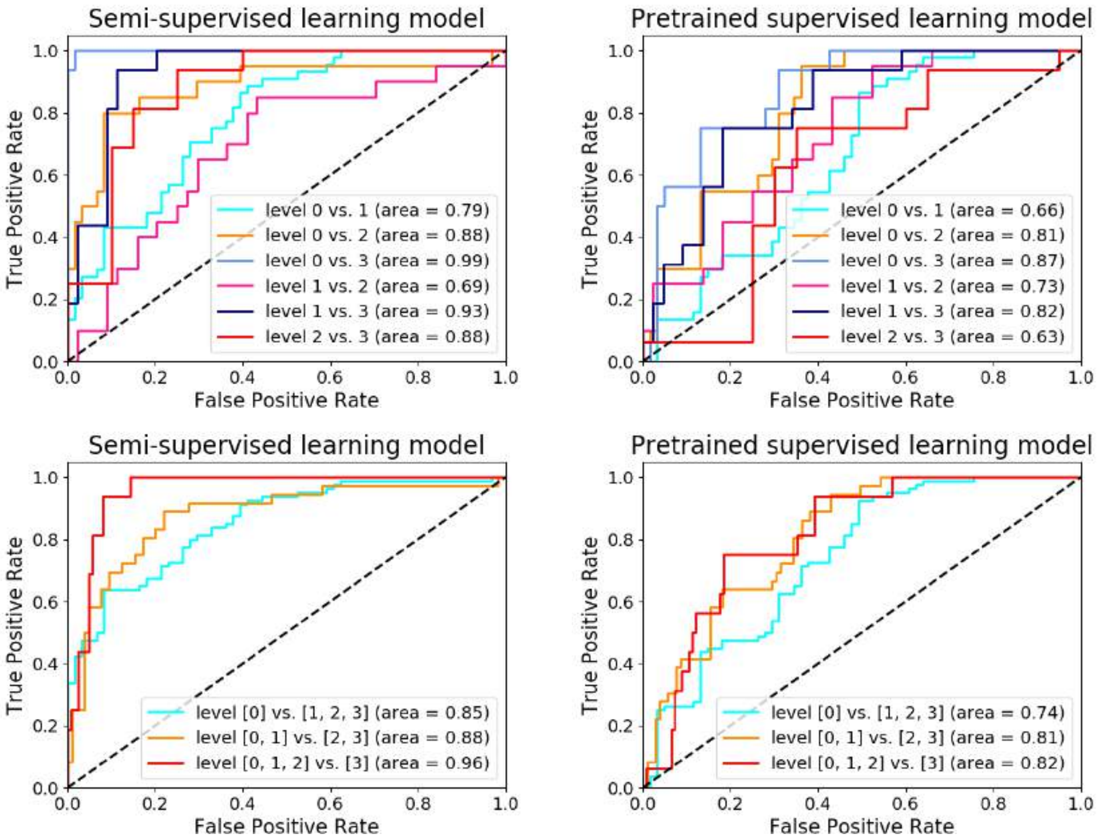


Figure 3-2: Receiver operating characteristic (ROC) curves of the semi-supervised learning model and the pre-trained supervised learning model. All the curves are based on the predictions of the test set. ROC curves for six pairwise comparisons (**top**). ROC curves for three dichotomized severity comparisons (**bottom**). All the curves are based on the predictions of the test set.

ground truth edema severity. Overlap of error bars graphically depicts the challenges in discriminating less severe stages of pulmonary edema. Pulmonary edema severity exists on a continuous spectrum and future work on this will be discussed in the following section.

3.3.4 Model Interpretation

We used Grad-CAM to visualize the regions in a radiograph that are important for the model prediction. (Figure 3-5) demonstrates two sample images from the two models.

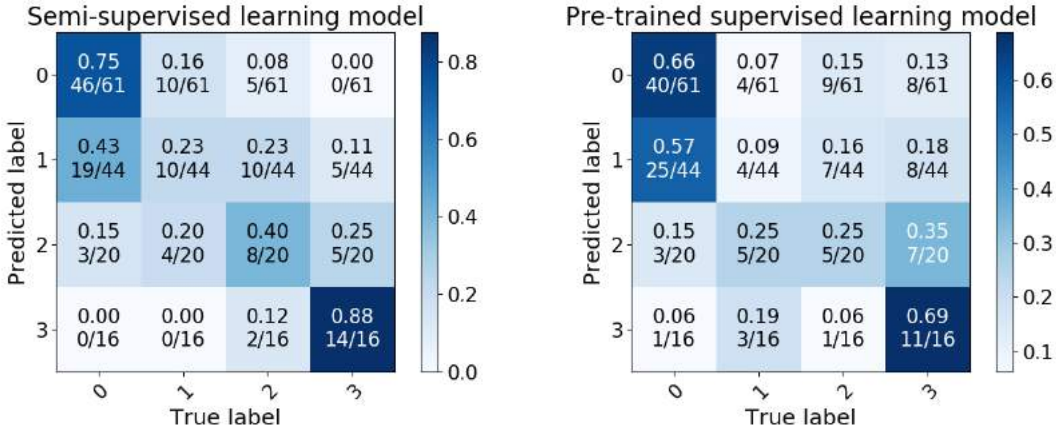


Figure 3-3: Confusion matrices from the semi-supervised learning model and the pre-trained supervised learning model. The denominator of each fraction number is the number of images that the algorithm predicts of the corresponding row, and the numerator is the number of images that belongs to the corresponding column. The quadratic-weighted Kappa values of the semi-supervised learning model and the pre-trained supervised learning model are 0.70 and 0.41. All the results are based on the predictions of the test set.

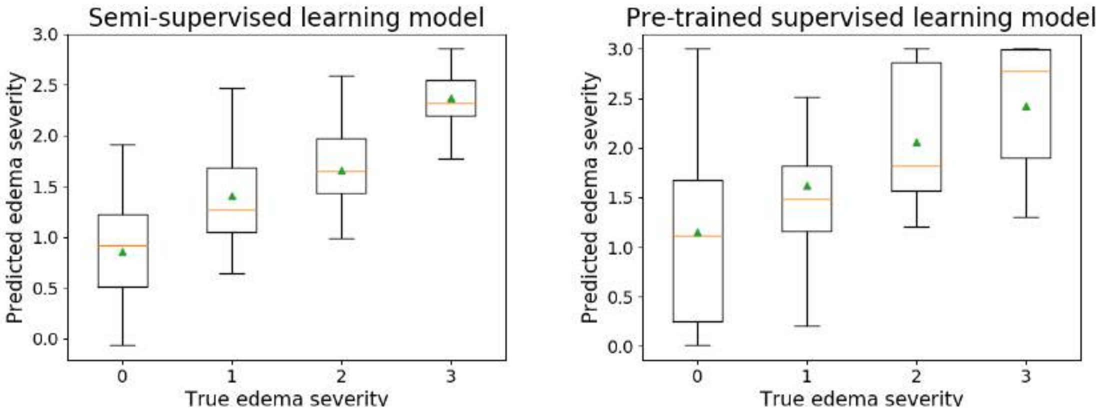


Figure 3-4: Predicted edema severity scores versus true edema severity labels from the semi-supervised learning model and the pre-trained supervised learning model. The box extends from the lower to upper quartile values of the distribution, with the orange line at the median and the green triangle at the mean. The whiskers extend from the box to show the range of the data. All the results are based on the predictions of the test set.

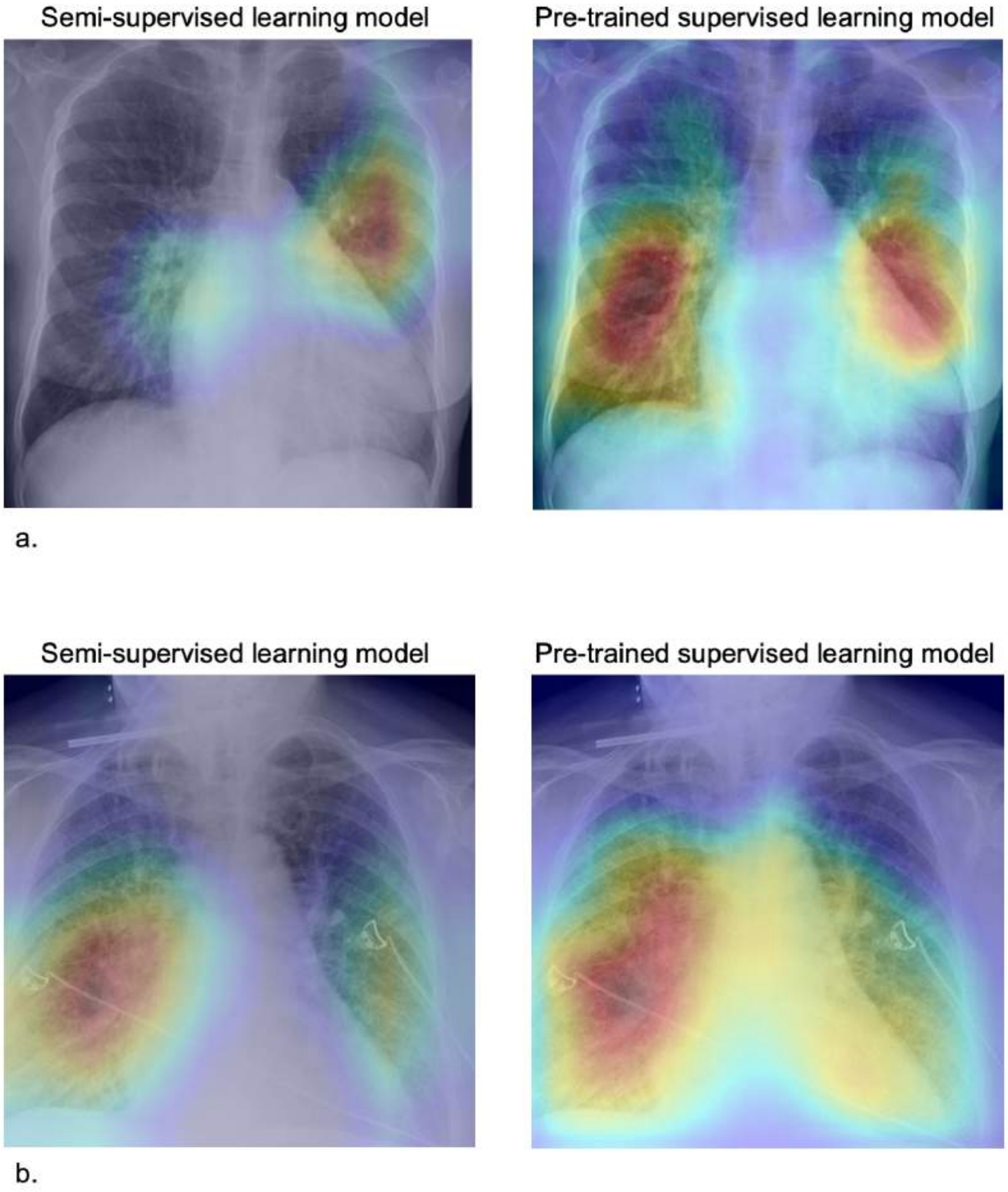

Semi-supervised learning model
Pre-trained supervised learning model
a.
Semi-supervised learning model
Pre-trained supervised learning model
b.

Figure 3-5: Grad-CAM heatmaps that highlight important regions for the model prediction. (**a**) A sample radiograph that is labeled as “vascular congestion” (level 1). (**b**) A sample radiograph that is labeled as “alveolar edema” (level 3).

3.4 Summary

In this chapter, we demonstrated the first attempt to employ machine learning algorithms to automatically and quantitatively assess the severity of pulmonary edema from chest x-ray images. Our results suggest that granular information about a patient's status captured in medical images can be extracted by machine learning algorithms, which promises to enable clinicians to deliver better care by quantitatively summarizing an individual patient's clinical trajectory, for example response to different treatments. This work also promises to enable clinical research studies that require quantitative phenotyping of patient status.

We demonstrated a computer vision model augmented with a VAE trained on a large image dataset with a limited number of labeled images. Our results suggest that it is difficult for a generative model to learn distinct data clusters for the labels that rely on subtle image features. In contrast, learning compact feature representations jointly from images and limited labels can help inform prediction by capturing structure shared by the image distribution and the conditional distribution of labels given images.

The semi-supervised approach learns from all the radiographs in the training set. The pre-trained image model learns from a large common image set and the labeled radiographs. Both approaches aim to address the challenge of limited labels extracted from the radiology reports. Both approaches have similar performance statistically in terms of AUC on most pairwise classification comparisons (seven out of nine). On the other two comparisons (two out of nine), the semi-supervised approach outperforms the pre-trained approach. The semi-supervised approach may give better results because it has learned from approximately 220,000 chest radiographs and is thus tailored to the image feature extraction of chest radiographs.

In the following chapters, we address the challenge of limited numerical labels by leveraging the rich information contained in the free-text radiology reports.

Chapter 4

Joint Image-text Modeling

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

4.1 Prior Work

The ability of neural networks to learn effective feature representations from images and text has catalyzed the recent surge of interest in joint image-text modeling. In supervised learning, tasks such as image captioning have leveraged a recurrent visual attention mechanism using recurrent neural networks (RNNs) to improve captioning performance [99]. The TieNet used this attention-based text embedding framework for pathology detection from chest radiographs [96], which was further improved by introducing a global topic vector and transfer learning [100]. A similar image-text embedding setup has been employed for chest radiograph (image) annotations [68]. In unsupervised learning, training a joint global embedding space for visual object discovery has recently been shown to capture relevant structure [29]. All of these models used RNNs for encoding text features. More recently, transformers such as the bidirectional-encoder-representations-from-transformers (BERT) model [20] have shown the ability to capture richer contextualized word representations using self-attention and have advanced the state-of-the-art in nearly every language processing task compared to variants of RNNs.

Prior work in image-text modeling can be grouped into three broad categories: i) Visual question answering (e.g., VQA [92]), where inference is performed on an image-text pair. Since text is not available at inference time in our setup, the model must estimate the severity exclusively from the image. After joint training is completed, we decouple the image model from the text model. ii) Report generation or image captioning [41]), where the model produces a report from the input image. In contrast, our goal is to generate a numerical measure of the edema severity from an image. We show that using available radiology reports during training improves accuracy of the image model at inference time. iii) Pathology detection (aka image classification). Prior work has used CNN-RNN architectures to either build models for classification of image-text pairs [67, 68] or to indirectly improve classification accuracy by jointly training an image classifier and a text decoder for report generation [100, 96]. Instead, we encode the text during training using the BERT model that has advanced the state-of-the-art on almost every NLP task over RNNs. Rather than generating reports from images as a way to regularize the image model, we build a joint embedding space for images and text during training and train the image and text classifiers jointly based on the representations in the embedding space. The ranking-based criterion we use to associate matching image and text descriptors was proposed and used to learn a joint embedding space of audio-visual data [30, 29]. Ours was the first attempt to jointly train a CNN-based image encoder and a BERT-based text encoder, and to enable decoupling of the two modalities at inference time in clinical applications.

We incorporate free-text information associated with the images by including them during the model training process. The setup proposed in my thesis uses a series of residual blocks [32] to encode the image representation and uses the BERT model to encode the text representation, similar to CNN-RNN based TieNet [96]. We use the radiology reports during training only, to improve the image-based model's performance. This is in contrast to visual question answering [6, 59, 5], where inference is performed on an image-text pair, and image/video captioning [99, 75, 92, 41], where the model generates text from the input image.

4.2 Methods

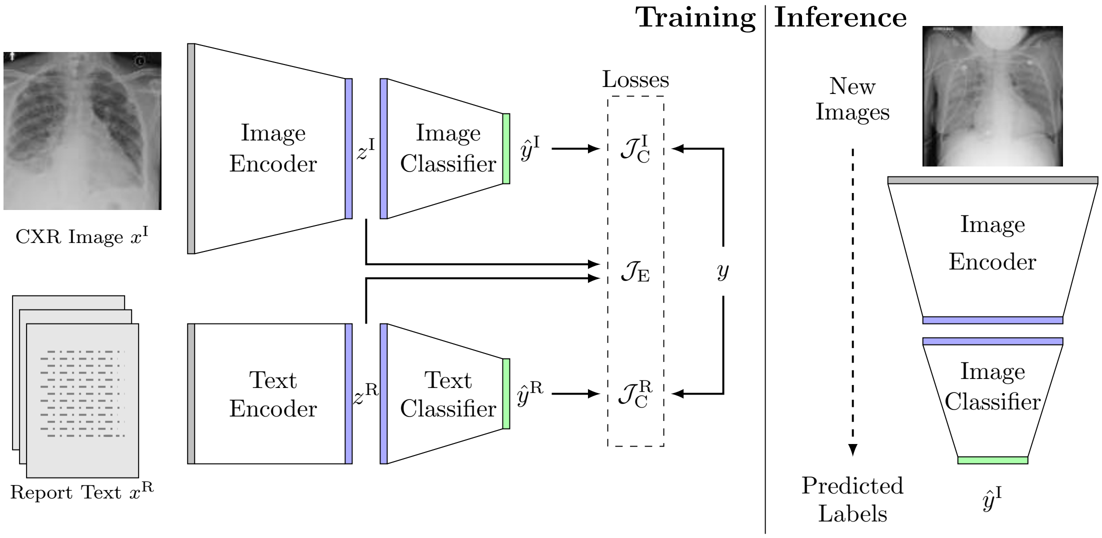


Figure 4-1: The architecture of our joint model, along with an example chest radiograph x^{I} and its associated radiology report x^{R}. At training time, the model predicts the edema severity level from images and text through their respective encoders and classifiers, and compares the predictions with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

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

1). Two classifiers are trained to classify the severity labels independently from the image features and from the text features. This setup enables us to decouple the image classification and the text classification at inference time. Learning the two representations jointly at training time improves the performance of the image model.

4.2.1 Joint Representation Learning

We apply a ranking-based criterion [14, 30] for training the image encoder and the text encoder parameterized by $\theta_{\mathrm{E}}^{\mathrm{I}}$ and $\theta_{\mathrm{E}}^{\mathrm{R}}$ respectively, to learn image and text feature representations $z^{\mathrm{I}}(x^{\mathrm{I}};\theta_{\mathrm{E}}^{\mathrm{I}})$ and $z^{\mathrm{R}}(x^{\mathrm{R}};\theta_{\mathrm{E}}^{\mathrm{R}})$. Specifically, given an image-text pair $(\mathrm{x}_j^{\mathrm{I}}, \mathrm{x}_j^{\mathrm{R}})$, we randomly select an impostor image $\mathrm{x}_{s(j)}^{\mathrm{I}}$ and an impostor report $\mathrm{x}_{s(j)}^{\mathrm{R}}$ from X. This selection is generated at the beginning of each training epoch. Map $s(j)$ produces a random permutation of $\{1, 2, ..., N\}$.

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

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

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

4.2.2 Classification

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

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

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

4.2.3 Loss Function

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

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

4.3 Implementation Details

The image encoder is implemented as a series of residual blocks [32], the text encoder is a BERT model that uses the beginner `[CLS]` token's hidden unit size of 768 and maximum sequence length of 320 [20]. The image encoder is trained from a random initialization, while the BERT model is fine-tuned during the training of the joint model. The BERT model parameters are initialized using pre-trained weights on scientific text [9]. The image features and the text features are represented as

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

We employ the stochastic gradient-based optimization procedure AdamW [98] to minimize the loss in Eq. (4.3) and use a warm-up linear scheduler [93] for the learning rate. The model is trained on all the image-text pairs by optimizing the first term in Eq. (4.3) for 10 epochs and then trained on the labeled image-text pairs by optimizing Eq. (4.3) for 50 epochs. The mini-batch size is 4. We use dot product as the similarity metric in Eq. (4.1). The dataset is split into training and test sets. All the hyper-parameters are selected based on the results from 5-fold cross validation within the training set.

4.4 Experiments

4.4.1 Data Preprocessing

The size of the chest radiographs varies and is around 3000×3000 pixels. We randomly translate and rotate the images on the fly during training and crop them to 2048×2048 pixels as part of data augmentation. We maintain the original image resolution to capture the subtle differences in the images between different levels of pulmonary edema severity. For the radiology reports, we extract the *impressions*, *findings*, *conclusion* and *recommendation* sections. If none of these sections are present in the report, we use the *final report* section. We perform tokenization of the text using ScispaCy [70] before providing it to the BERT tokenizer.

4.4.2 Model Evaluation

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

- **image-only**: An image-only model with the same architecture as the image stream in our joint model. We trained the image model in isolation on the 6,212 labeled images.

- A joint image-text model trained on the 6,212 labeled image-text pairs only. We compare two alternatives to the joint representation learning loss:
  - **ranking-dot**, **ranking-l2**, **ranking-cosine**: the ranking based criterion in Eq. (4.1) with $\mathrm{Sim}(z^{\mathrm{I}}, z^{\mathrm{R}})$ defined as one of the dot product ${z^{\mathrm{I}}}^{\top} z^{\mathrm{R}}$, the reciprocal of euclidean distance $-\|z^{\mathrm{I}} - z^{\mathrm{R}}\|$, and the cosine similarity $\frac{{z^{\mathrm{I}}}^{\top} z^{\mathrm{R}}}{\|z^{\mathrm{I}}\| . \|z^{\mathrm{R}}\|}$;
  - **dot**, **l2**, **cosine**: direct minimization on the similarity metrics without the ranking based criterion.
- **ranking-dot-semi**: A joint image-text model trained on the 6,212 labeled and the 240K unlabeled image-text pairs in a semi-supervised fashion, using the ranking based criterion with dot product in Eq. (4.1). Dot product is selected for the ranking-based loss based on cross-validation experiments on the supervised data comparing ranking-dot, ranking-l2, ranking-cosine, dot, l2, and cosine.

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

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

4.5 Results

Table 4.1 reports the performance statistics for all similarity measures. The findings are consistent with our cross-validation results: the ranking based criterion offers significant improvement when it is combined with the dot product as the similarity metric.

Method	AUC (0 *vs* 1,2,3)	AUC (0,1 *vs* 2,3)	AUC (0,1,2 *vs* 3)	macro-F1
`l2`	0.78	0.76	0.83	0.42
`ranking-l2`	0.77	0.75	0.80	0.43
`cosine`	0.77	0.75	0.81	0.44
`ranking-cosine`	0.77	0.72	0.83	0.41
`dot`	0.65	0.63	0.61	0.15
ranking-dot	**0.80**	**0.78**	**0.87**	**0.45**

Table 4.1: Performance statistics for all similarity measures.

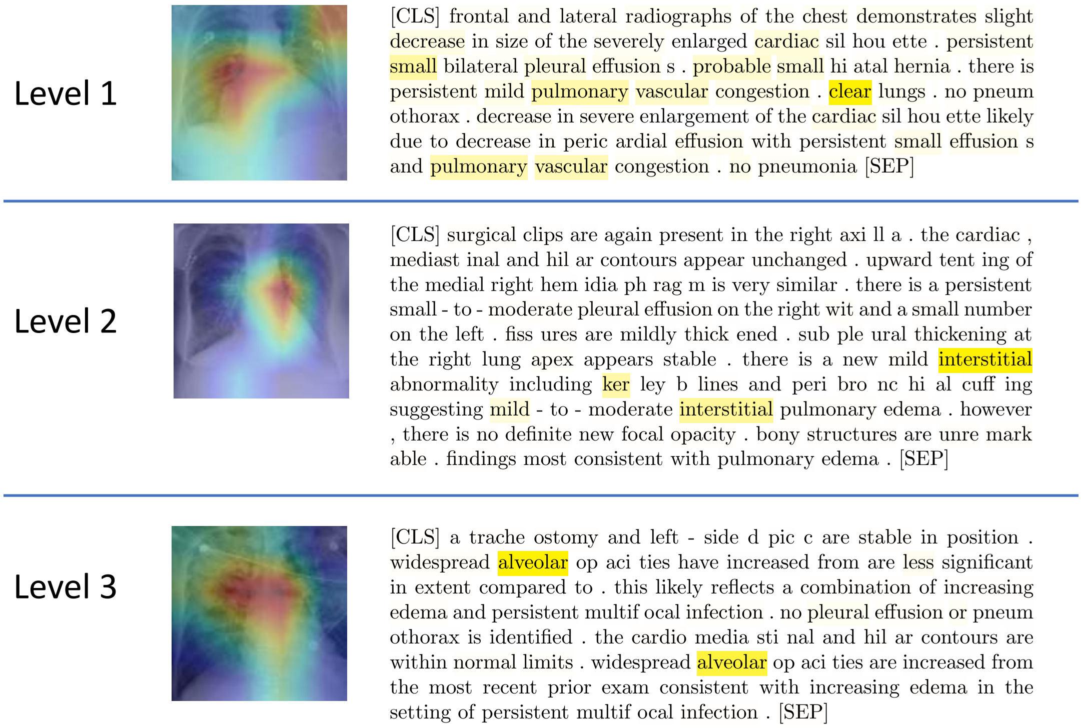
Table 4.2 reports the performance of the optimized baseline model (`image-only`) and two variants of the joint model (`ranking-dot` and `ranking-dot-semi`). We observe that when the joint model learns from the large number of unlabeled image-text pairs, it achieves the best performance. The unsupervised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

Method	AUC (0 *vs* 1,2,3)	AUC (0,1 *vs* 2,3)	AUC (0,1,2 *vs* 3)	macro-F1
`image-only`	0.74	0.73	0.78	0.43
`ranking-dot`	0.80	0.78	0.87	0.45
ranking-dot-semi	**0.82**	**0.81**	**0.90**	**0.51**

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

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

As a by-product, our approach provides the possibility of interpreting model classification using text. While a method like Grad-CAM [83] can be used to localize regions in the image that are “important" to the model prediction, it does not identify the relevant characteristics of the radiographs, such as texture. By leveraging the image-text embedding association, we visualize the heatmap of text attention corresponding to the last layer of the `[CLS]` token in the BERT model. This heatmap indicates report tokens that are important to our model prediction. As shown in Fig. 4-2, we use Grad-CAM [83] to localize relevant image regions and the highlighted words (radiographic findings, anatomical structures, etc.) from the text embedding to explain the model's decision making.



[CLS] frontal and lateral radiographs of the chest demonstrates slight decrease in size of the severely enlarged cardiac sil hou ette . persistent small bilateral pleural effusion s . probable small hi atal hernia . there is persistent mild pulmonary vascular congestion . clear lungs . no pneum othorax . decrease in severe enlargement of the cardiac sil hou ette likely due to decrease in peric ardial effusion with persistent small effusion s and pulmonary vascular congestion . no pneumonia [SEP]

[CLS] surgical clips are again present in the right axi ll a . the cardiac , mediast inal and hil ar contours appear unchanged . upward tent ing of the medial right hem idia ph rag m is very similar . there is a persistent small - to - moderate pleural effusion on the right wit and a small number on the left . fiss ures are mildly thick ened . sub ple ural thickening at the right lung apex appears stable . there is a new mild interstitial abnormality including ker ley b lines and peri bro nc hi al cuff ing suggesting mild - to - moderate interstitial pulmonary edema . however , there is no definite new focal opacity . bony structures are unre mark able . findings most consistent with pulmonary edema . [SEP]

[CLS] a trache ostomy and left - side d pic c are stable in position . widespread alveolar op aci ties have increased from are less significant in extent compared to . this likely reflects a combination of increasing edema and persistent multif ocal infection . no pleural effusion or pneum othorax is identified . the cardio media sti nal and hil ar contours are within normal limits . widespread alveolar op aci ties are increased from the most recent prior exam consistent with increasing edema in the setting of persistent multif ocal infection . [SEP]

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

4.6 Summary

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

The joint model visualization suggests the possibility of using the text to semantically explain the image model, which represents a promising direction for the future investigation. In the next chapter, we leverage the correspondences of report sentences and local image regions to further improve the joint representation learning.

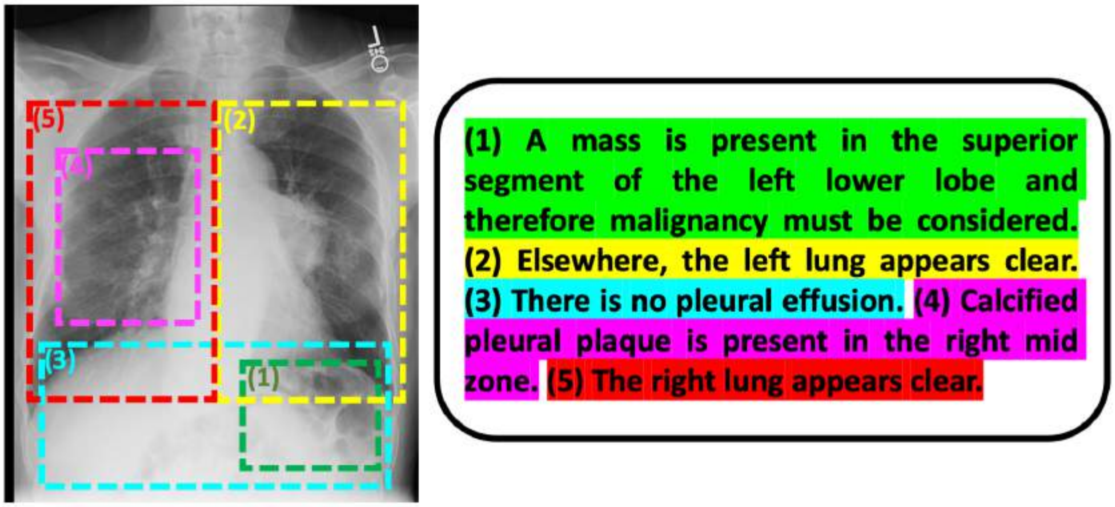
Chapter 5

Mutual Information for Representation Learning

In this chapter, 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.

5.1 Motivation

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 [15, 79, 11]. In this thesis, I aim to learn image representations that improve classification tasks, such as pathology detection, by making use of the rich information contained



(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.

Figure 5-1: 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.

in the raw text that describe the findings in the image.

I exploit mutual information (MI) to learn useful image representations jointly with text. MI quantifies statistical dependencies between two random variables. Prior work has estimated and optimized MI across images for image registration [97, 60], and MI between images and image features for unsupervised learning [17, 73, 34]. 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.

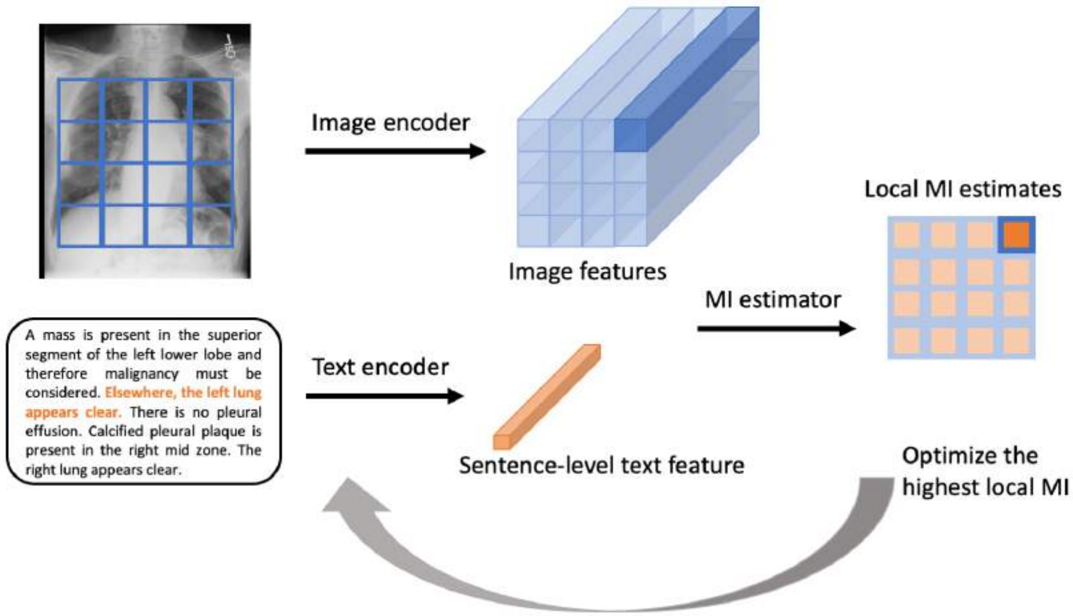
Furthermore, my thesis proposes to exploit the image spatial structure and sentence-level text features with mutual information maximization to learn image and text representations that are *useful* for subsequent analysis of images. Fig. 5-1 shows an example image-text pair, where the image is a chest radiograph and the document is the associated radiology report [42]. Each sentence in the report describes a local region in the image. A sentence is usually a minimal and complete semantic unit [101, 78]. The findings described in that semantic unit are usually captured in a local region of the image [29].

Prior work in image-text joint learning has leveraged image-based text generation as an auxiliary task during the image model training [96, 100, 67], or has blended image and text features for downstream inference tasks [68]. Other work has leveraged contrastive learning, an approach to maximize a lower bound on mutual information (MI) to learn image and text representations jointly [101, 13]. In the experimental results, I demonstrate that the maximization of local MI yields the greatest improvement in the downstream image classification tasks.

Furthermore, my thesis investigated the advantages of using local features for image-text representation learning and we compared local MI with global MI while keeping the architecture and the downstream task the same across all approaches. Our work was the first empirical demonstration and theoretical analysis of advantages offered by local MI for modeling joint image-text structure. Representation learning is an active research area with contrastive and MI-based approaches leading the field [34, 101]. Any state of the art representation learning framework, such as CNN-RNN joint embedding, can be readily improved by employing local MI in its loss function and the feature selection as we explain in the following sections.

5.2 Methods

Let x^{I} be an image, 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 $z^{\mathrm{R}}(x^{\mathrm{R}}; \theta_{\mathrm{E}}^{\mathrm{R}})$.



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.

Figure 5-2: 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.

5.2.1 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:

$$I(z^{\mathrm{I}}, z^{\mathrm{R}}) \triangleq \mathbb{E}_{p(z^{\mathrm{I}}, z^{\mathrm{R}})} \left[\log \frac{p(z^{\mathrm{I}}, z^{\mathrm{R}})}{p(z^{\mathrm{I}}) p(z^{\mathrm{R}})} \right]. \tag{5.1}$$

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 compared to solely minimizing the conditional entropy, where the image encoder could generate identical features for all data to achieve the conditional entropy minimum.

5.2.2 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 [8, 73, 55, 87]. The essence of those methodologies is to construct a discriminator $f(\mathrm{z}^{\mathrm{I}}_i, \mathrm{z}^{\mathrm{R}}_j; \theta_{\mathrm{D}})$, parameterized by θ_{D}, that estimates the likelihood (or the likelihood ratio), given a sample pair $(\mathrm{z}^{\mathrm{I}}_i, \mathrm{z}^{\mathrm{R}}_j)$, of whether or not this pair 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 as the lower bound of the MI by approximating the likelihood ratio in Eq. (5.1) [8, 73].

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

$$\hat{\theta}^{\mathrm{I}}_{\mathrm{E}}, \hat{\theta}^{\mathrm{R}}_{\mathrm{E}}, \hat{\theta}_{\mathrm{D}} = \arg \max_{\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}, \theta_{\mathrm{D}}} \hat{I}(z^{\mathrm{I}}(x^{\mathrm{I}}; \theta^{\mathrm{I}}_{\mathrm{E}}), z^{\mathrm{R}}(x^{\mathrm{R}}; \theta^{\mathrm{R}}_{\mathrm{E}}); \theta_{\mathrm{D}}) \leq I(z^{\mathrm{I}}, z^{\mathrm{R}}). \tag{5.2}$$

We consider two MI lower bounds: Mutual Information Neural Estimation (MINE) [8] and Contrastive Predictive Coding (CPC) [73]. 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. (5.1), using the Donsker-Varadhan (DV) variational formula of the KL divergence between the joint distribution and the product of the marginals. Employing MINE yields the lower bound

$$\hat{I}^{(\mathrm{MINE})}_{\theta^{\mathrm{I}}_{\mathrm{E}}, \theta^{\mathrm{R}}_{\mathrm{E}}, \theta_{\mathrm{D}}}(z^{\mathrm{I}}, z^{\mathrm{R}}) = \mathbb{E}_{p(z^{\mathrm{I}}, z^{\mathrm{R}})} \left[f(z^{\mathrm{I}}, z^{\mathrm{R}}; \theta_{\mathrm{D}}) \right] - \log \mathbb{E}_{p(z^{\mathrm{I}})p(z^{\mathrm{R}})} \left[e^{f(z^{\mathrm{I}}, z^{\mathrm{R}}; \theta_{\mathrm{D}})} \right]. \tag{5.3}$$

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.

CPC leads to objective function

$$\hat{I}^{(\mathrm{CPC})}_{\theta^{\mathrm{I}}_{\mathrm{E}},\theta^{\mathrm{R}}_{\mathrm{E}},\theta_{\mathrm{D}}}(z^{\mathrm{I}},z^{\mathrm{R}}) = \mathbb{E}_{p(z^{\mathrm{I}},z^{\mathrm{R}})}\left[f(z^{\mathrm{I}},z^{\mathrm{R}};\theta_{\mathrm{D}})\right] - \mathbb{E}_{p(z^{\mathrm{I}})}\mathbb{E}_{p(z^{\mathrm{R}})}\left[\log \sum_{\hat{\mathrm{z}}^{\mathrm{R}}_j \in z^{\mathrm{R}}} e^{f(z^{\mathrm{I}},\hat{\mathrm{z}}^{\mathrm{R}}_j;\theta_{\mathrm{D}})}\right]. \quad (5.4)$$

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.

5.2.3 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, as shown in Fig. 5-2. We train the image and text encoders, as well as the MI discriminator from all the image-text data:

$$\hat{\theta}^{\mathrm{I}}_{\mathrm{E}}, \hat{\theta}^{\mathrm{R}}_{\mathrm{E}}, \hat{\theta}_{\mathrm{D}} = \arg \max_{\theta^{\mathrm{I}}_{\mathrm{E}},\theta^{\mathrm{R}}_{\mathrm{E}},\theta_{\mathrm{D}}} \sum_j \sum_m \max_n \hat{I}(\mathrm{z}^{\mathrm{I}}_{j,(n)}, \mathrm{z}^{\mathrm{R}}_{j,(m)}), \quad (5.5)$$

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

5.3 Implementation Details

Chest radiographs are downsampled to 256×256. We use a 5-block resnet [32] as the image encoder in the local MI approach and the image feature representation z^{I} is 16×512 (4×4×512) feature vectors, as shown in Fig. 5-3. 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 [4] as the text encoder for both report-level and sentence-level feature extraction, architecture detailed in Fig. 5-4. 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 1024→512→1 multilayer perceptron. The image feature and the text feature are concatenated before fed into 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 [98] optimizer is employed for the BERT encoder and the Adam [47] 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.

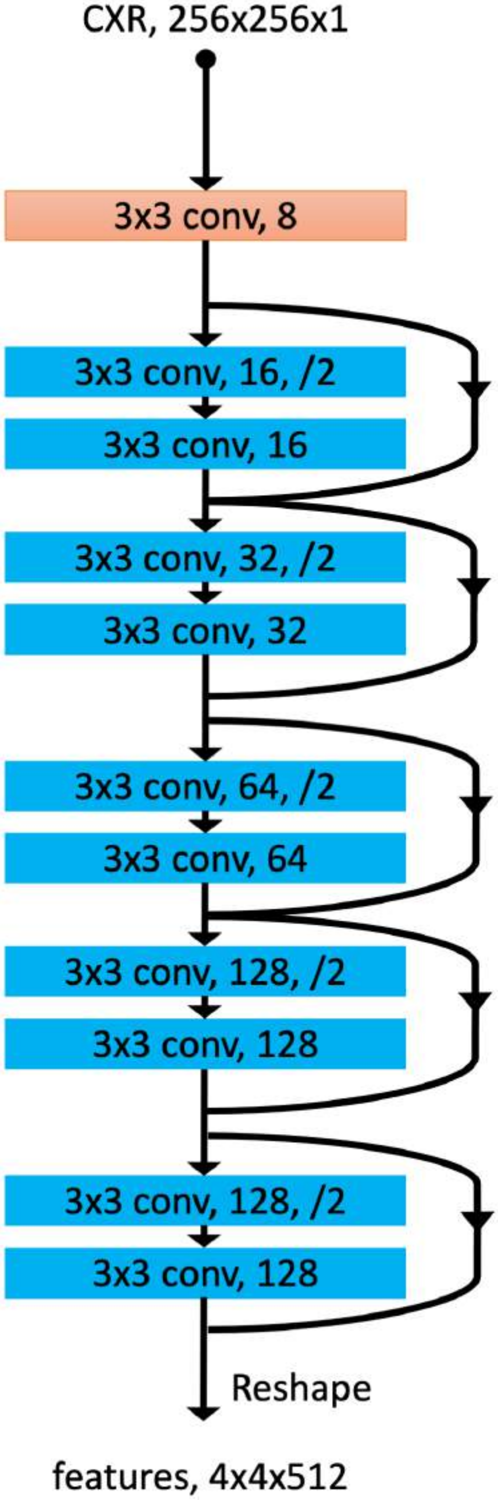


Figure 5-3: Top: Image encoder using residual neural network. Each residual block includes 2 convolutional layers.

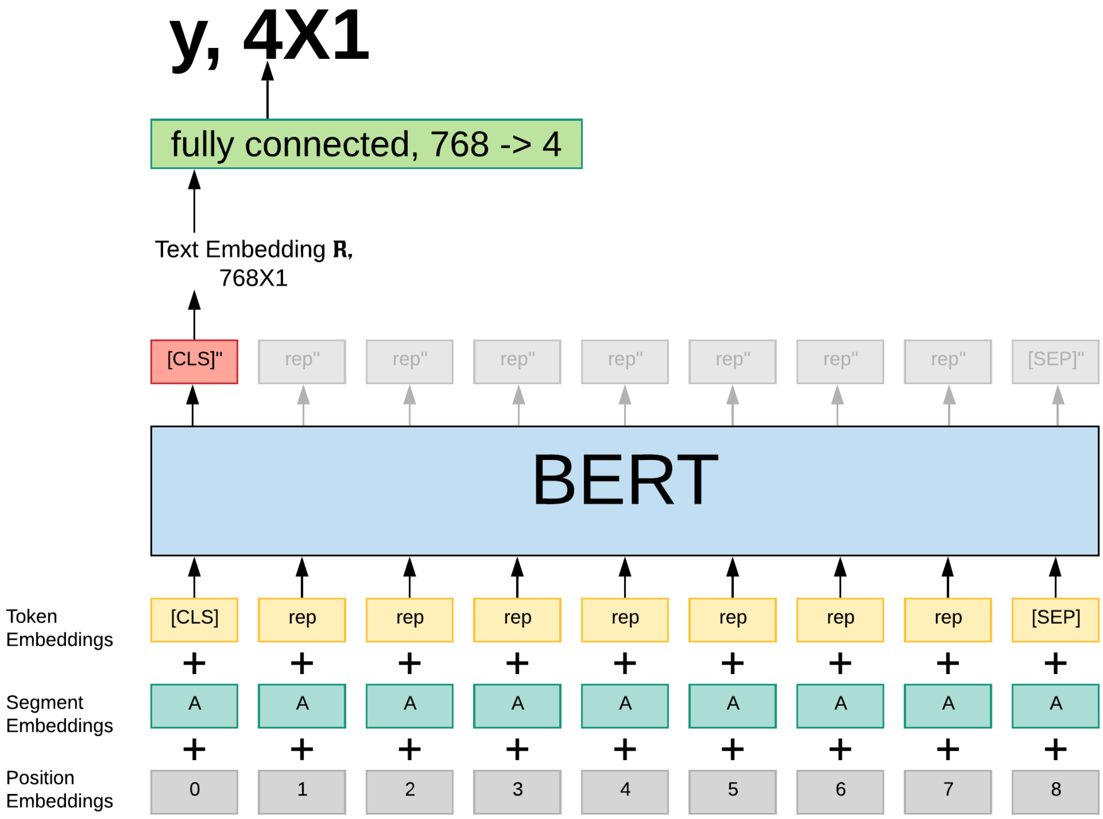


Figure 5-4: Text encoder using the BERT model. A full radiology report is encoded between `[CLS]` and `[SEP]` tokens; `rep` is the text associated with the report. As is standard in the literature, we use the hidden representation corresponding to the `[CLS]` token marked as `[CLS]` as the text embedding. Maximum input sequence length is set to 320.

5.4 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 5-5, each local image region x_n^{I} has a hidden variable H_n that specifies the physiological processes and disease status in that region. This image region x_n^{I} 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).

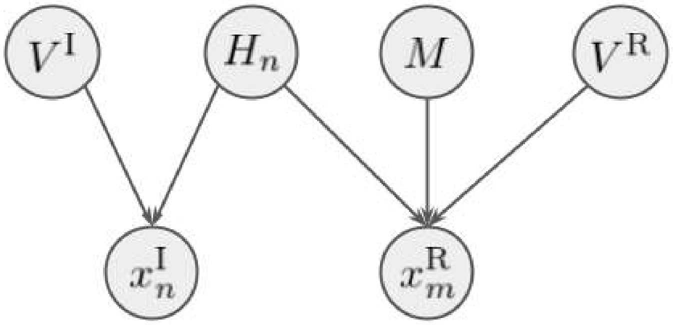


Figure 5-5: A conjectured generative model that describes how paired chest radiograph and radiology report are constructed and the underlying structural assumptions.

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 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 a concatenation of z_n^{I} and $z_n^{\bar{\mathrm{I}}}$, where the z_n^{I} is the feature of the local image region generated from H_n and $z_n^{\bar{\mathrm{I}}}$ is the rest of the image feature. Applying the chain rule of mutual information, we have:

$$\begin{aligned} I(z^{\mathrm{I}}, H_n) &= I(z_n^{\mathrm{I}}, H_n) + I(z_n^{\bar{\mathrm{I}}}, H_n | z^{\mathrm{I}}) && (5.6) \\ &\geq I(z_n^{\mathrm{I}}, H_n). && (5.7) \end{aligned}$$

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

5.5 Experiments

5.5.1 Data and Model Evaluation

We demonstrate our approach on the MIMIC-CXR dataset v2.0 [42] 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 [38, 44, 43]. Note that there are 14 findings available in the repository [43]. We only train and evaluate 9 out of the 14 pathologies, where 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 [35, 53, 56]. The severity level ranges from 0 to 3 with a high score indicating high risk.

The two test sets provided in those two publicly available label repositories are used to evaluate our methods [43, 53]. The patients that are in either of the two repositories' test sets are excluded from our model training. Table 5.1 summarizes the size of the (labeled) training data and test data.

5.5.2 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

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

Table 5.1: The number of images in the (labeled) training sets and the test sets.

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 the fully supervised image model uses.

We compare our MI maximization approach on local features with the global MI maximization approach. We test both MINE [8] and CPC [73] 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 [43, 53].

- **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. (5.5).

    - **encoder-frozen**, **encoder-tuned**: The resulting image encoder followed by a classifier is *re-trained.*

At the image model training or *re-training* time, all variants are trained on the same training sets. No image from the test set patients is ever seen by the models at any training phase. 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.

5.6 Results

In Table 5.2 and Table 5.3, we present the area under the receiver operating characteristic curve (AUC) values for the variants of our algorithms on the **EdemaSeverity** ordinary classification task and the **Pathology9** binary classification tasks. For most classification tasks, the local MI approach with encoder tuning performs the best and has significantly improved the performance of 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.

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 5.4 that shows that under the certain structural assumption, the sum of local MI values is the lower bound to the global MI. Finally, we compared the ranking based joint modeling approach in the last chapter with the global mutual information approach proposed in this chapter on the **EdemaSeverity** ordinary classification task. Both approaches achieve comparable performance.

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-rank	**tuned**	0.82		0.83		0.91	
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 5.2: The AUCs on the **EdemaSeverity** ordinal classification task.

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 5.3: The AUCs on the **Pathology9** binary classification tasks.

5.7 Summary

In this chapter, 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 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) [34] 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 showed that the local MI approach offers the greatest improvement to the downstream image classification tasks.

Both MI estimators (MINE and CPC) that utilize variational bounds yield similar results. In the next chapter, we discuss the challenges of estimating MI and propose a discriminator based MI estimator that does not rely the variational lower bounds of MI.

Chapter 6

DEMI: Discriminative Estimator of Mutual Information

Estimating mutual information between continuous random variables is often intractable and extremely challenging for high-dimensional data. Recent progress has leveraged neural networks to optimize variational lower bounds on mutual information. Although showing promise for this difficult problem, the variational methods have been theoretically and empirically proven to have serious statistical limitations: 1) many methods struggle to produce accurate estimates when the underlying mutual information is either low or high; 2) the resulting estimators may suffer from high variance. Our approach is based on training a classifier that provides the probability that a data sample pair is drawn from the joint distribution rather than from the product of its marginal distributions. Moreover, we establish a direct connection between mutual information and the average log odds estimate produced by the classifier on a test set, leading to a simple and accurate estimator of mutual information. We show theoretically that our method and other variational approaches are equivalent when they achieve their optimum, while our method sidesteps the variational bound. Empirical results demonstrate high accuracy of our approach and the advantages of our estimator in the context of representation learning.

6.1 Prior Work

Mutual information (MI) measures the information that two random variables share. MI quantifies the statistical dependency — linear and non-linear — between two variables. This property has made MI a crucial measure in machine learning. In particular, recent work in unsupervised representation learning has built on optimizing MI between latent representations and observations [17, 103, 73, 34, 90, 3, 94]. Maximization of MI has long been a default method for multi-modality image registration [60], especially in medical applications [97], though in most work the dimensionality of the random variables is very low. Here, coordinate transformations on images are varied to maximize their MI.

Estimating MI from finite data samples has been challenging and is intractable for most continuous probabilistic distributions. Traditional MI estimators [88, 18, 50, 24] do not scale well to modern machine learning problems with high-dimensional data. This impediment has motivated the construction of variational bounds for MI [71, 7]; in recent years this has led to maximization procedures that use deep learning architectures to parameterize the space of functions, exploiting the expressive power of neural networks [87, 8, 73, 69].

Unfortunately, optimizing lower bounds on MI has serious statistical limitations. Specifically, prior work has shown that any high-confidence distribution-free lower bound cannot exceed $O(\log N)$, where N is the number of samples [65]. This implies that if the underlying MI is high, it cannot be accurately and reliably estimated by variational methods like MINE [8]. Song *et al.* further categorized the state-of-the-art variational methods into "generative" and "discriminative" approaches, depending on whether they estimate the probability densities or the density ratios [87]. They showed that the "generative" approaches perform poorly when the underlying MI is small and "discriminative" approaches perform poorly when MI is large; moreover, certain approaches like MINE [8] are prone to high variances.

We propose a simple discriminative approach that avoids the limitations of previous discriminative methods that are based on variational bounds. Instead of estimating density or attempting to predict one data variable from another, our method estimates the likelihood that a sample is drawn from the joint distribution versus the product of marginal distributions. A similar classifier-based approach was used in [58] for "two sample testing" – hypothesis tests about whether two samples are from the same distribution or not. If the two distributions are the joint and product of the marginals, then the test is for independence. A generalization of this work was used in [84] to test for conditional independence. We show that accurate performance on this classification task provides an estimate of the log odds. This can greatly simplify the MI estimation task in comparison with generative approaches: estimating a single likelihood ratio may be easier than estimating three distributions (the joint and the two marginals). Moreover, classification tasks are generally amicable to deep learning, while density estimation remains challenging in many cases. Our approach avoids the estimation of the partition function, which induces large variance in most discriminative methods [87]. Our empirical results bear out these conceptual advantages.

Our approach, as well as other sampling-based methods such as MINE, uses the given joint/paired data with derived "unpaired" data that captures the product of the marginal distributions. The unpaired data can be synthesized via permutations or resampling of the paired data. This construction, which synthesizes unpaired data and then defines a metric to encourage paired data points to map closer than the unpaired data in the latent space, has previously been used in other machine learning applications, such as audio-video and image-text joint representation learning [30, 13]. Recent contrastive learning approaches [89, 33, 16, 31] further leverage a machine learning model to differentiate paired and unpaired data mostly in the context of unsupervised representation learning. [85] used paired and unpaired data in conjunction with a classifier-based loss function for patch-based image registration.

This chapter is organized as follows. In Section 6.2, we derive our approach to estimating MI. Section 6.3 discusses connections to related approaches, including MINE. This is followed by empirical evaluation in Section 6.4. Our experimental results on synthetic and real image data demonstrate the advantages of the proposed discriminative classification-based MI estimator, which has higher accuracy than the state-of-the-art variational approaches and a good bias/variance tradeoff.

6.2 Methods

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be two random variables generated by joint distribution $p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. Mutual Information (MI)

$$I(x;y) \triangleq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{6.1}$$

is a measure of dependence between x and y. Let $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$ be a set of n independent identically distributed (i.i.d.) samples from $p(x,y)$. The law of large numbers implies

$$\hat{I}_p(\mathcal{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \to I(x;y) \quad \text{as } n \to \infty, \tag{6.2}$$

which suggests a simple estimation strategy via sampling. Unfortunately, the joint distribution $p(x,y)$ is often unknown and therefore the estimate in Eq. (6.2) cannot be explicitly computed. Here we develop an approach to accurately approximating the estimate $\hat{I}_p(\mathcal{D})$ based on discriminative learning.

In our development, we will find it convenient to define a Bernoulli random variable $z \in \{0,1\}$ and to “lift" the distribution $p(x,y)$ to the product space $\mathcal{X} \times \mathcal{Y} \times \{0,1\}$. We thus define a family of distributions parametrized by $\alpha \in (0,1)$ as follows:

$$p_*(x, y|z = 1; \alpha) = p(x,y), \tag{6.3}$$

$$p_*(x, y|z = 0; \alpha) = p(x)p(y), \tag{6.4}$$

$$p_*(z = 1; \alpha) = 1 - p_*(z = 0; \alpha) = \alpha. \tag{6.5}$$

Using Bayes' rule, we obtain

$$\frac{p_*(z=1|x,y)}{p_*(z=0|x,y)} = \frac{p_*(x,y,z=1)}{p_*(x,y,z=0)} = \frac{p_*(x,y|z=1)\,p_*(z=1)}{p_*(x,y|z=0)\,p_*(z=0)} = \frac{p(x,y)}{p(x)p(y)} \cdot \frac{\alpha}{1-\alpha}, \tag{6.6}$$

which implies that the estimate in (6.2) can be alternatively expressed as

$$\hat{I}_p = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_*(z=1|x_i,y_i)}{p_*(z=0|x_i,y_i)} - \log \frac{\alpha}{1-\alpha} \tag{6.7}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \text{logit}\left[p_*(z=1|x_i,y_i)\right] - \text{logit}[\alpha], \tag{6.8}$$

where $\text{logit}[u] \stackrel{\Delta}{=} \log \frac{u}{1-u}$ is the log-odds function.

Our key idea is to approximate the latent posterior distribution $p_*(z=1|x,y)$ by a classifier that is trained to distinguish between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$ as described below.

6.2.1 Training Set Construction

We assume that we have access to a large collection $\hat{\mathcal{D}}$ of i.i.d. samples (x,y) from $p(x,y)$ and define $\hat{p}(x,y;\hat{\mathcal{D}})$, $\hat{p}(x;\hat{\mathcal{D}})$, and $\hat{p}(y;\hat{\mathcal{D}})$ to be the empirical joint and marginal distributions respectively induced by data set $\hat{\mathcal{D}}$.

Each sample is generated independently of all others as follows. First, a value $z^j \in \{0,1\}$ is sampled from the prior distribution $p_*(z)$ in (6.5). If $z^j = 1$, then a pair (x^j, y^j) is sampled randomly from the empirical joint distribution $\hat{p}(x,y;\hat{\mathcal{D}})$; otherwise value x^j is sampled randomly from the empirical marginal distribution $\hat{p}(x;\hat{\mathcal{D}})$ and value y^j is sampled randomly from the empirical marginal distribution $\hat{p}(y;\hat{\mathcal{D}})$, independently from x^j. This sampling is easy to implement as it simply samples an element from a set of unique values in the original collection $\hat{\mathcal{D}}$ with frequencies adjusted to account for repeated appearances of the same value.

It is straightforward to verify that any individual sample in the training set $\mathcal{T}$ is generated from distribution $p_*(x,y,z)$ up to the sampling of $\hat{\mathcal{D}}$. Where $\hat{\mathcal{D}}$ is small, multiple samples may not be jointly from $\hat{\mathcal{D}}$ but from some idiosyncratic subset; however, the empirical distribution induced by the set $\mathcal{T}$ converges to $p_*(x, y, z)$ as the size of available data $\hat{\mathcal{D}}$ and the size m of the training set $\mathcal{T}$ becomes large.

6.2.2 Classifier Training for Mutual Information Estimation

Let $q(z = 1|x, y; \theta, \mathcal{T})$ be a (binary) classifier parameterized by θ and derived from the training set $\mathcal{T}$. If $q(z = 1|x, y; \theta, \mathcal{T})$ accurately approximates the posterior distribution $p_*(z = 1|x, y; \alpha)$, then we can use this classifier q instead of $p_*(z = 1|x, y; \alpha)$ in (6.8) to estimate MI.

We follow the widely used maximum likelihood approach to estimating the classifier's parameters θ and form the cross-entropy loss function

$$\ell(\theta; \mathcal{T}) = -\frac{1}{m}\sum_{j=1}^{m} \log q(z^j|x^j, y^j; \theta, \mathcal{T}) \tag{6.9}$$

$$= -\frac{1}{m}\sum_{j=1}^{m} z^j \log q(z^j = 1|x^j, y^j; \theta, \mathcal{T}) + (1 - z^j)\log(1 - q(z^j = 1|x^j, y^j; \theta, \mathcal{T})) \tag{6.10}$$

to be minimized to determine the optimal value of parameters $\hat{\theta}$. Once the optimization is completed, we form the estimate

$$\hat{I}_q(\mathcal{D}, \mathcal{T}) = \frac{1}{n}\sum_{i=1}^{n} \text{logit}\left[q(z = 1|x_i, y_i; \hat{\theta}, \mathcal{T})\right] - \text{logit}[\alpha] \tag{6.11}$$

that approximates the estimate in (6.8). Note that the estimate is computed using the data set $\mathcal{D}$, which is distinct from the training set $\mathcal{T}$.

6.2.3 Asymptotic Analysis

As the size of available data $\hat{\mathcal{D}}$ and the size m of the training set $\mathcal{T}$ increase to infinity, the law of large numbers implies

$$\ell(\theta;\mathcal{T}) \to \mathbb{E}_{p_*(x,y,z)}\left[\log q(z|x,y;\theta,\mathcal{T})\right], \tag{6.12}$$

and therefore

$$\hat{\theta} \overset{\Delta}{=} \arg\min_{\theta} \ell(\theta;\mathcal{T}) \to \arg\max_{\theta} \mathbb{E}_{p_*(x,y,z)}\left[\log q(z|x,y;\theta,\mathcal{T})\right]. \tag{6.13}$$

Thus, when the model capacity of the family $q(z|x,y;\theta)$ is large enough to include the original distribution $p_*(z|x,y)$, Gibb's inequality implies

$$q(z|x,y;\hat{\theta},\mathcal{T}) \to p_*(z|x,y) \quad \text{and} \quad \hat{I}_q(\mathcal{D},\mathcal{T}) \to I(x;y) \tag{6.14}$$

as both the training data and testing data grow.

6.3 Connections to Other Mutual Information Estimators

MINE and SMILE [8] introduced the Mutual Information Neural Estimation (MINE) method, wherein they proposed learning a neural network $f(x,y;\theta)$ that maximizes the objective function $J(f) = \mathbb{E}_{p(x,y)}\left[f(x,y;\theta)\right] - \log \mathbb{E}_{p(x)p(y)}\left[e^{f(x,y;\theta)}\right]$, which is the Donsker-Varadhan (DV) lower bound for the Kullback–Leibler (KL) divergence. For analysis purposes, we define $\hat{q}(x,y;\theta) \overset{\Delta}{=} \frac{1}{Z}e^{f(x,y;\theta)}p(x)p(y)$, where $Z = \mathbb{E}_{p(x)p(y)}\left[e^{f(x,y;\theta)}\right]$. By substituting into the definition of $J(\cdot)$ and invoking Gibb's inequality, we obtain

$$\begin{aligned} J(f) &= \mathbb{E}_{p(x,y)}\left[\log \hat{q}(x,y;\theta)\right] - \mathbb{E}_{p(x,y)}\left[\log p(x)p(y)\right] && (6.15) \\ &\leq \mathbb{E}_{p(x,y)}\left[\log p(x,y)\right] - \mathbb{E}_{p(x,y)}\left[\log p(x)p(y)\right] = I(x;y), && (6.16) \end{aligned}$$

with equality if and only if $\hat{q}(x, y; \theta) \equiv p(x, y)$, i.e.,

$$f(x, y) = \log \frac{p(x, y)}{p(x)p(y)} + C, \tag{6.17}$$

where C is a constant that is absorbed into the partition function Z. Thus the objective function is a lower bound on MI and is maximized when the unspecified “statistics network" $f(x, y)$ is the log likelihood ratio of the joint distribution and the product of the marginals.

[87] introduced the Smoothed Mutual Information Lower Bound Estimator (SMILE) approach which is a modification of the MINE estimator. To alleviate the high variance of $f(x, y)$ in practice, the tilting factor $e^{f(x,y)}$ is constrained to the interval $[e^{-\tau}, e^{\tau}]$, for a tuned hyper-parameter τ. As $\tau \to \infty$, SMILE estimates converge to those produced by MINE.

The log likelihood ratio of the joint versus the marginals, which the $f(x, y)$ network from both these methods approximates, is the optimal classifier function for the task defined on our training set $\mathcal{T}$ above. Our parameterization of this ratio makes use of a classifier and the logit transformation. While analytically equivalent, the MINE and SMILE optimization procedures must instead search over ratio functions directly, optimizing $f(x, y) \approx p(x, y)/p(x)p(y)$ itself. Our experimental results demonstrate the advantage of using our estimator in (6.11).

CPC [73] proposed a contrastive predictive coding (CPC) method that also maximizes a lower bound

$$J(f) = \mathbb{E}_{p(x,y)} \left[\frac{1}{N} \sum_{i=1}^{N} \log \frac{f(x_i, y_i; \theta)}{\frac{1}{N} \sum_{j=1}^{N} f(x_i, y_j; \theta)} \right] + \log N \leq I(x; y), \tag{6.18}$$

where $f(x, y; \theta)$ is a neural network and N is the batch size. CPC is not capable of estimating high underlying MI accurately– it is constrained by their batch size N, and this constraint scales logarithmically. In our approach, we do not estimate the likelihood ratio directly, instead we construct an auxiliary variable and “lift" the joint distribution, where we leverage the power of a discriminative neural network classifier. The logit transformation of our classifier response is used to approximate the log likelihood ratio in Eq. (6.1).

CCMI [69] recently proposed a classifier based (conditional) MI estimator (CCMI). The classifier $g(x, y; \theta)$ is trained on paired and unpaired sample pairs to yield the posterior probability that the joint distribution $p(x, y)$ (rather than the product of marginals $p(x)p(y)$) generated a sample pair (x, y). Unlike DEMI, the CCMI estimator

$$\hat{I}(x, y) = \mathbb{E}_{p(x,y)} [\text{logit} [g(x, y)]] - \log \mathbb{E}_{p(x)p(y)} \left[\frac{1 - g(x, y)}{g(x, y)} \right] \tag{6.19}$$

still relies on a variational lower bound in [8]. The first term above employs paired sample pairs and is identical to our estimator in Eq. (6.11) for $g(x, y) \overset{\Delta}{=} q(z = 1|x, y; \hat{\theta}, \mathcal{T})$ and $\alpha = 0.5$. The second term depends on the unpaired samples and is asymptotically zero. Thus CCMI and DEMI are asymptotically equivalent, but for finite sample sizes CCMI is prone to higher error than DEMI, as we demonstrate empirically later in the paper.

6.4 Experiments

We employ two setups widely used in prior work [87, 8, 76, 34] to evaluate the proposed estimator and to compare it to the state of the art approaches for estimating MI. In particular, we directly evaluate the accuracy of the resulting estimate in synthetic examples where the true value of MI can be analytically derived and also compare the methods' performance in a representation learning task where the goal is to maximize MI.

Additional experiments that investigate self-consistency and long-run training behavior are reported in Appendices 6.4.3 and 6.4.4 respectively.

6.4.1 MI Estimation

Experimental Design

We sample jointly Gaussian variables x and y with known correlation and thus known MI values, which enables us to measure the accuracy of MI estimators when trained on this data. We vary the dimensionality of x and y (20-d, 50-d, and 100-d), the underlying true MI, and the size of the training set (32K, 80K, and 160K) in order to characterize the relative behaviors of different MI estimators. In an additional experiment, we employ an element-wise cubic transformation ($y_i \mapsto y_i^3$) to generate non-linear dependencies in the data. Since deterministic transformations of x and y preserve MI, we can still access ground truth values of MI in this setup. We generate a different set of 10240 samples held out for testing/estimating MI given each training set. We generate 10 independently drawn training and test sets for each two correlated Gaussian variables.

We assess the following estimators in this experiment:

- **DEMI**, the proposed method, with three settings of the parameter $\alpha \in \{0.25, 0.5, 0.75\}$ in Eq. (6.5).
- **SMILE** [87], with three settings of the clipping parameter $\tau \in \{1.0, 5.0, \infty\}$. The $\tau = \infty$ case (i.e., no clipping) is equivalent to the **MINE** [8] objective.
- **InfoNCE** [73], the method used for contrastive predictive coding (CPC).
- **CCMI** [69].
- A generative model (**GM**), i.e., directly approximating $\log p(x, y)$ and marginals $\log p(x)$ and $\log p(y)$ using a flow network. We note that it is difficult to make comparable parameterizations between **GM**-flow networks and the rest of the methods, and that additionally because the "base" flow distribution is a Gaussian, these networks have a structural advantage for our synthetic tests. They are, in a sense, correctly specified for the Gaussian case, which probably would not happen in real data.

Each estimator uses the same neural network architecture: a multi-layer perceptron with an initial concatenation layer for the x and y inputs, then two fully connected layers with ReLU activations, then a single output. This final layer uses a linear output for **MINE**, **SMILE**, **InfoNCE**, and **CCMI**, and a logistic output for **DEMI**. We use 256 hidden units for each of the fully connected layers. For the **GM** flow network we use the RealNVP scheme [21], which includes a "transformation block" of two 256-unit fully connected layers, each with ReLU activations. This network outputs two parameters, a scale and a shift, both element-wise. This transformation block is repeated three times.

We train each MI estimator for 20 epochs and with the mini-batch size of 64. We employ the Adam optimizer with learning rate parameter 0.0005. The architecture choices above and the optimization settings are comparable with [87].

Results

Figure 6-1 reports the MI estimation error $(I(x, y) - \hat{I}(x, y))$ versus the true underlying MI for the experiments with joint Gaussians and joint Gaussians with a cubic transformation, when the size of the training data size is 160K. The results of **DEMI** with three different settings of α are very close. In Figure 6-1, we only show **DEMI** ($\alpha = 0.5$).

For all experiments, **InfoNCE** substantially underestimated MI. This is due to its log-batchsize ($\log N$) maximum, which saturates quickly relative to the actual mutual information in these regimes. The limited training data setup leads to increased errors of **CCMI**.

Overall, for Gaussian variables, the **GM** method performed very well. This is somewhat expected, as its base distribution for the flow network is itself a Gaussian. This trend begins to fall off at higher MI values for the 100-d case. For the cubic case, however, the **GM** method performs quite poorly, perhaps due to the increased model flexibility required for the transformed distribution.

For 20-d Gaussian variables, **MINE** and **SMILE** with both parameter settings overestimated MI in comparison to **DEMI**, which provided estimates that were fairly close to the ground truth values. Appendix 6.4.4 further investigates this behavior. For the 50-d joint Gaussian case, **DEMI** again produced accurate estimates of MI, while **MINE** and **SMILE** underestimated MI substantially. For the 100-d joint Gaussian case, all approaches underestimated MI, with **DEMI** and **CCMI** perform-

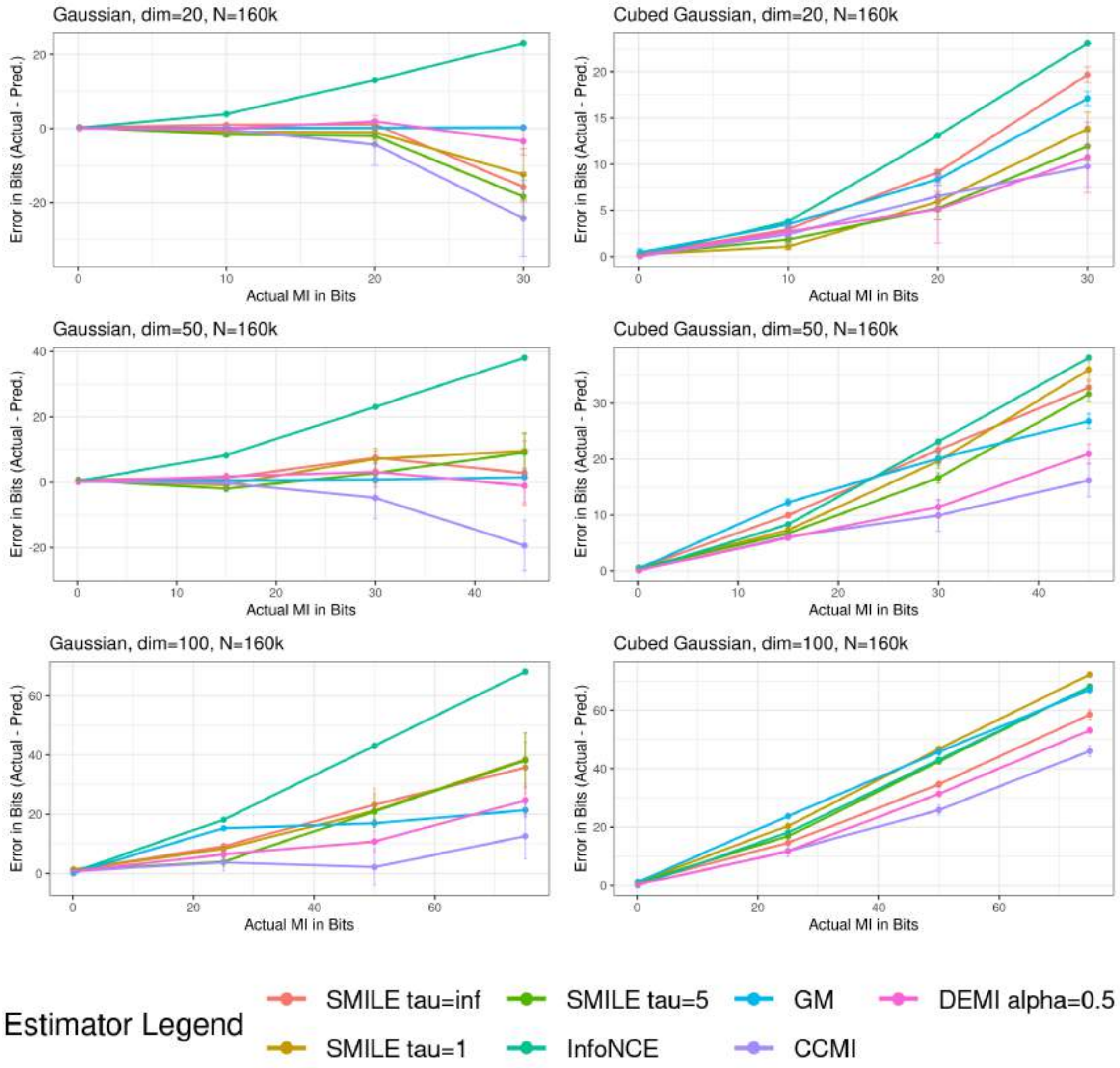


Figure 6-1: Mutual information estimation between multivariate Gaussian variables (**left column**) and between multivariate Gaussian variables with a cubic transformation (**right column**). **Closer to Zero is better**. The estimation error $(I(x, y) - \hat{I}(x, y))$ versus the true underlying MI are reported. These estimates are based on training data size of 160K. We only show **DEMI** ($\alpha = 0.5$) since the results of the other two parameter settings are very close.

ing best.

For 20-d joint Gaussians with a cubic transformation, all approaches underestimated MI, **SMILE** ($\tau = 5$) and **DEMI** performed best. For the 50-d and 100-d cases, all approaches understimated MI, with **DEMI** performing the best.

In summary, **DEMI** performed best or very similar to the best baseline in all the experiments. It further was not sensitive to the setting of its parameter α. Its performance relative to the other MI estimators held up with the training data size decreased.

6.4.2 Representation Learning

Experimental Design

Our second experiment demonstrates the viability of **DEMI** as the differentiable loss estimate in a representation learning task. Specifically, we train an encoder on CIFAR10 and CIFAR100 [51] data sets using the Deep InfoMax [34] criterion. Deep InfoMax learns representations by maximizing mutual information between local features of an input and the output of an encoder, and by matching the representations to a prior distribution. To evaluate the effectiveness of different MI estimators, we only include the MI maximization as the representation learning objective, without prior matching. We compare **DEMI** with **MINE**, **SMILE**, **InfoNCE**, and **JSD** [34] for MI estimation and maximization as required by Deep InfoMax.

As discussed in [34], evaluation of the quality of a representation is case-driven and relies on various proxies. We use *classification* as a proxy to evaluate the representations, i.e., we use Deep InfoMax to train an encoder and learn representations from a data set without class labels, and then we freeze the weights of the encoder and train a small fully-connected neural network classifier using the representation as input. We use the *classification* accuracy as a performance proxy to the representation learning and thus the MI estimators. We build two separate classifiers on the last convolutional layer (conv(256,4,4)) and the following fully connected layer (fc(1024)) for *classification* evaluation of the representations, similar to the setup in [34]. The size of the input images is 32 × 32 and the encoder has the same architecture as the one in [34].

Results

Table 6.1 reports the top 1 classification accuracy of CIFAR10 and CIFAR100. **DEMI** is comparable to **InfoNCE** in 3 out of 4 tasks and outperforms the other MI estimators by a significant margin. When the encoder is allowed to train with the class labels, it becomes a fully-supervised task. We also report its classification accuracy as reference. The classification accuracy based on the representations learned by Deep InfoMax with **DEMI** is close to or even surpasses the fully-supervised case. Note that the Deep InfoMax objective we use here does not include prior distribution matching to regularize the encoder.

Table 6.1: Top 1 *classification* accuracy as a proxy to representation learning (Deep InfoMax without prior matching) performance with different MI estimators.

	CIFAR10		CIFAR100	
	conv(256,4,4)	fc(1024)	conv(256,4,4)	fc(1024)
Fully-supervised	75.39		42.27	
MINE	71.36	66.30	42.52	37.23
SMILE ($\tau = 5.0$)	71.88	66.92	42.74	37.48
SMILE ($\tau = 1.0$)	71.12	66.22	42.13	37.10
JSD	72.79	67.94	42.78	37.76
InfoNCE	73.73	69.77	44.91	39.95
DEMI	**74.09**	**70.16**	**45.59**	**42.02**

6.4.3 Self-consistency Tests

Experimental Design

We assess and compare the MI estimators using the self-consistency tests proposed in [87]. We perform the tests on MNIST images [52]. The self-consistency tests examine some important properties that a “useful" MI estimator should have, because optimizing MI is more important for many downstream machine learning applications than estimating the exact value of MI.

The self-consistency tests examine: 1) capability of detecting independence, 2) monotonicity with data processing, 3) and additivity. We thus perform the following experiments, where the MNIST image set induces a data distribution and each MNIST image is a random variable that follows this data distribution:

- **MI estimation between one MNIST image and one row-masked image.** Given an MNIST image X, we mask out the bottom rows and leave the top t rows of the image, which creates $Y = h(X; t)$. The estimated MI $\hat{I}(X, Y)$ should be equal or very close to zero, if X and Y are independent. In this context, $\hat{I}(X, Y)$ should be close to 0 when t is small and be non-decreasing with t. We normalize this measurement to the final value at $t = 28$ (the last row), which should be the maximum information.

- **MI estimation between two identical MNIST images and two row-masked images.** Given an MNIST image X, we create two row-masked images: $Y_1 = h(X; t_1)$ and $Y_2 = h(X; t_2)$, where $t_1 = t_2 + 3$. Since additional data processing should not increase mutual information, $\hat{I}([X, X], [Y_1, Y_2])/\hat{I}(X, Y_1)$ should be close to 1.

- **MI estimation between two MNIST images and two row-masked images.** We randomly select two MNIST images and concatenate them: $[X_1, X_2]$, and mask the same number of rows on them: $[h(X_1; t), h(X_2; t)] = [Y_1, Y_2]$. $\hat{I}([X_1, X_2], [Y_1, Y_2])/\hat{I}(X_1, Y_1)$ should be close to 2.

We have 60k MNIST images or concatnated images for training and a test set of 10k images. We train each MI estimator for 100 epochs and set the mini-batch size to 64. For all methods, we concatentate inputs, then convolve with a 5×5 kernel with stride 2 and 64 output channels, then apply a fully connected layer with 1024 hidden units, which then maps to a single output. ReLU is applied after all but the last layer.

Results

In Figure 6-2 we plot the results of the three self-consistency metrics for each method. In general most methods perform well for the first measurement (monotonicity) with the exception of SMILE ($\tau = \infty$), which exhibits charateristicly high variance. Other settings of SMILE and DEMI both are relatively well behaved, though overall InfoNCE performs best. For the second metric ("data processing"), all methods perform well, again aside from SMILE ($\tau = \infty$) variance. In the third metric, SMILE $\tau = 1$ also exhibits a large bump in the center (where optimal should be constant 2 overall), but both InfoNCE and DEMI converge to 1 overall, and no method performs optimally. In general InfoNCE performs best across between the first two measures, but DEMI and SMILE ($\tau = 1$) also do well in two of three.

6.4.4 Long-run Training Behavior of SMILE

As shown in Section 6.4, SMILE somewhat overestimates the MI for the 20 dimensional Gaussian case in high MI regimes (~ 30 Nats or more). This did not occur at lower MI conditions or in higher dimensions.

Further investigation showed this problem to increase as training went on; to illustrate this, we set up a new experiment on the 20 dimensional Gaussian case. We ran each setting of SMILE ($\tau = 1, 5, \infty$) for 100000 training steps with batch size 64, drawing samples directly from the generating distributions. This means that the training set has effectively a very large size. We did this for three ground-truth MI values of 10,20, and 30. For comparison we also run the proposed method through the same.

This setup exactly mirrors the experiment in [87] Figure 1 in Section 6.1 of that paper, and uses their provided code and generation method, except that we replace their step-wise increasing MI schedule with a constant 10, 20, or 30 nat generator, and we run the experiment longer.

The curves for the first row of Figure 6-3 show good performance with relatively stable long-term behavior, particularly for $\tau = 1$. The curves in the third row of Figure

6-3 on the otherhand suggest that for certain distribution/domain combinations, even though SMILE and MINE are based on a lower bound of MI, they can both grossly overestimate it. This may be as [65] suggests due in part to a sensitivity of the estimate of $-\ln \mathbb{E}[e^{f(x,y)}]$ to outliers. The proposed method eventually overestimates as well in both the 20 and 30 nat cases, but does not have the strongly divergent

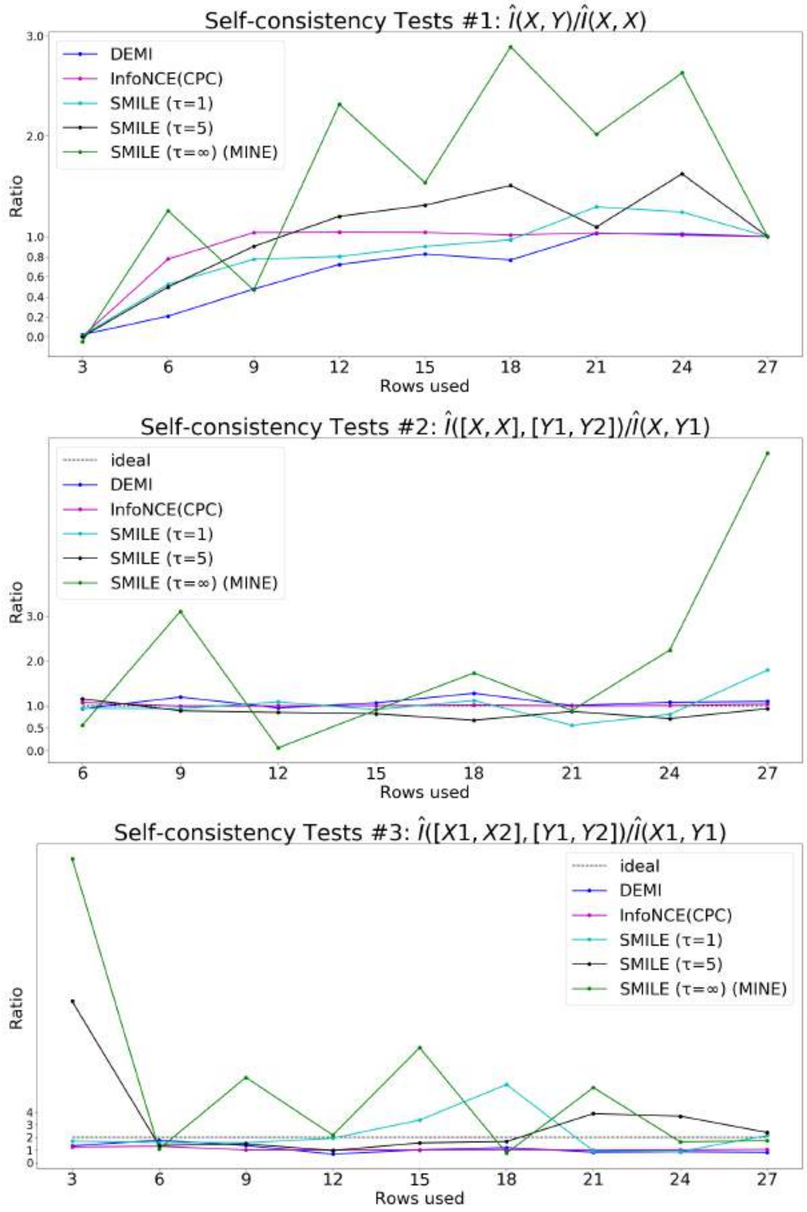


Figure 6-2: Results of the three self-consistency tests for SMILE ($\tau = 1, 5, \infty$), InfoNCE, and DEMI. "Monotonicity" is **top**, "data processing" is **middle**, and "additivity" is **bottom**.

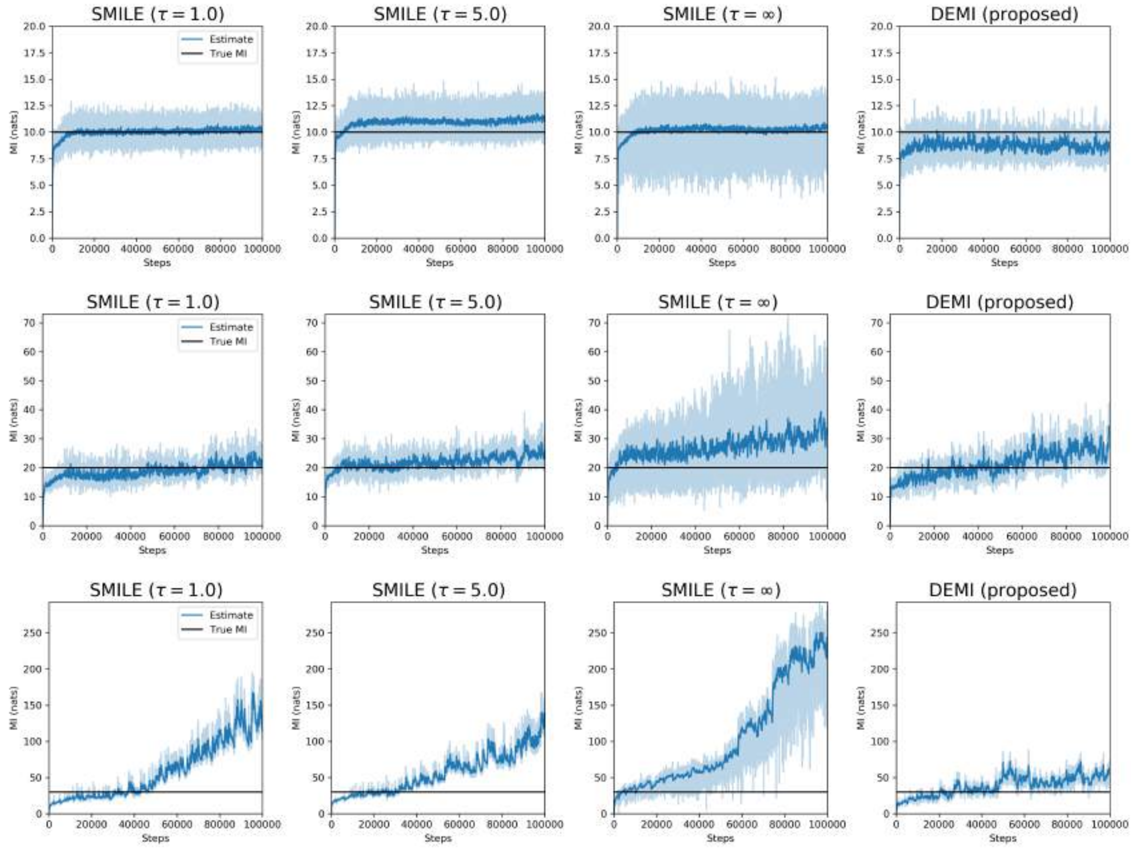


Figure 6-3: Long-run behavior of SMILE and DEMI for 10 (**top row**), 20 (**middle row**), and 30 (**bottom row**) Nats. Analytically SMILE converges to the MINE objective for $\tau \to \infty$. Smoothed trajectories are plotted in bold, exact trajectories are the semi-translucent curve, and the actual Mutual information is the black constant line.

behavior exhibited by SMILE (seen particularly strongly in $\tau = \infty$ settings).

6.5 Summary

In this chatper, we described a simple approach for estimating MI from joint data that is based on a neural network classifier that is trained to distinguish whether a sample pair is drawn from the joint distribution or the product of its marginals. The resulting estimator is the average over joint data of the logit transform of the classifier responses. Theoretically, the estimator converges to MI when the data sizes grow to infinity and the neural network capacity is large enough to contain the corresponding true conditional probability.

The accuracy of our estimator is governed by the ability of the classifier to predict the true posterior probability of items in the test set, which in turn depends on (i) the number of training sample pairs and (ii) the capacity of the neural network used in training. Thus the quality of our estimates is subject to the classical issues of model capacity and overfitting in deep learning. We leave theoretical analysis (which is closely related to the classifier's convergence to the true separating boundary for a general hypothesis class) for future work.

Although DEMI outperforms other estimators in terms of bias of estimating the value of mutual information. The challenges of applying our estimator in mutual information optimization are two folds: 1) In our experiments of synthetic data from Gaussian variables, we have access to an infinite number of samples. We constructed a validation set to select hyper-parameters of the DEMI classifier (e.g., learning rate, number of training epochs). When using DEMi to optimize mutual information from very limited training samples, we will not be able to select hyper-parameters using a holdout validation set. 2) When using DEMI to maximize mutual information of two variables generated from encoders or generators we would like to train, we need to optimize both the encoders and the DEMI classifier. Every time the encoders are updated, the resulting underlying variables are updated as well. The DEMI classifier has to learn from new samples from the updated variables to update its mutual information estimation. This primal-dual optimization framework may significantly prolong the convergence time.

We discussed close connections between our approach and the lower bound approaches of MINE and SMILE and InfoNCE(CPC). Unlike the difference-of-entropies (DoE) estimator described in [65], our approach does not make use of assumed distributions. We also demonstrate empirical advantages of our approach over the state of the art methods for estimating MI in synthetic and real image data. Given its simplicity and promising performance, we believe that DEMI is a good candidate for use in research that optimizes MI for representation learning.

Chapter 7

Discussion

7.1 Pulmonary Edema Assessment

In this thesis, we have developed multimodal learning algorithms to build an image model that assesses pulmonary edema from chest radiographs. The results of these algorithms provide a performance benchmark for future work. We have shown that it is feasible to automatically classify four levels of pulmonary edema on chest radiographs. Understandably, the performance of the algorithm mirrors the challenge of distinguishing these disease states for radiologists. The differentiation of alveolar edema from no pulmonary edema (level 3 vs 0) is an easier task than distinguishing interstitial edema from pulmonary vascular congestion (level 2 vs 1). Even among radiologists, there is substantial variability in the assessment of pulmonary edema. More machine learning approaches should be explored for this clinical task in future work.

Our work expands on prior studies by employing machine learning algorithms to automatically and quantitatively assess the severity of pulmonary edema from chest radiographs. Prior work has shown the ability of convolutional neural networks to detect pulmonary edema among several other pathologies that may be visualized in chest radiographs [95, 22, 77]. Neural networks have been validated in large datasets to achieve expert level identification of findings in chest radiographs [62]. Their AUCs in detecting the presence of pulmonary edema range from 0.83 to 0.88. By treating pulmonary edema as a single pathology, it is difficult to draw direct comparison to our work which considers pulmonary edema as a spectrum of findings. A conservative comparison would be to compare prior work to our model's ability to distinguish no edema and pulmonary vascular congestion from interstitial and alveolar edema (levels 0,1 vs 2,3) which have AUCs of 0.81 (pre-trained) and 0.88 (semi-supervised). Although their test sets are based on labels extracted from radiology reports, our test set labels are annotated and reached consensus on by four radiologists. Others have trained neural networks on B-type natriuretic peptide values to produce a quantitative assessment of congestive heart failure [82]. However, B-type natriuretic peptide increases non-linearly with worsening CHF, and exhibits marked inter-patient variability. A B-type natriuretic peptide of 1000 in one patient could represent an acute exacerbation, while being the baseline for another patient, making B-type natriuretic peptide a poor surrogate outcome measure for acute pulmonary edema. The grading of pulmonary edema severity relies on much more subtle radiological findings (image features). The clinical management of patients with pulmonary edema requires comparisons of serial exams and understanding serial trends. Accurate, reproducible, and rapid quantification of pulmonary edema is of paramount value to clinicians caring for these patients.

There were limitations in the data curation. Extracting labels from clinical radiology reports allowed us to quickly obtain a reasonable amount of labelled data, but is inferior to data labelled for a specific purpose. Not only is there poor inter-reader agreement among radiologists for pulmonary edema detection [27], but radiologists may use different languages to describe a similar pathophysiologic state. In future work, we will explore joint modeling of chest radiographs and radiology reports and aim to mitigate the bias introduced by simply employing regular expressions.

Pulmonary edema exists on a continuous spectrum of severity. By discretizing our data into four classes, we have potentially lost valuable information and contaminated the categories. The category of severe edema in our dataset contains all images containing alveolar edema, even though this varies wildly in clinical practice. In practice, it is challenging to quantify pulmonary edema at a more granular level.

Comparisons between images are easier and more reproducible. Future work could leverage pairs of images to quantify edema on a continuous scale.

The diagnosis of pulmonary edema is often challenging due to the possibility of other competing diagnoses that have overlapping radiographic findings. For example, multifocal pneumonia can be confused with alveolar pulmonary edema, and chronic interstitial edema can be misinterpreted as interstitial pulmonary edema. In order to minimize this bias, we restrict our labeled data to a cohort of patients diagnosed with CHF. In this work, we purposely ignore image findings such as cardiomegaly and pleural effusions that are correlated with pulmonary edema and often used by radiologists when making the diagnosis. In future work, we plan to leverage multi-task training to jointly learn these associated features. By incorporating multiple image observations in the model training, an algorithm would approximate the clinical gestalt that a radiologist has when considering the etiology of pulmonary opacities. By separating the features of pulmonary edema from features that are associated with CHF, however, our model is not biased against detecting non-cardiogenic pulmonary edema.

Lastly, we compare our results only to the chest radiograph rather than some other reference standard of pulmonary edema. In clinical practice, the chest radiograph is usually considered the reference standard to measure pulmonary edema. Pulmonary capillary wedge pressure might be more accurate, but is extremely invasive, and performed only on a small fraction of patients, so would be impractical to be used as a reference standard.

7.2 Image Model Interpretability

In this thesis, we take advantage of rich multimodal data available during training of the machine learning models to improve the inference accuracy when the model is applied to a new chest x-ray image during patient assessment. Our multimodal learning framework generates image and text embeddings that are used by respective classifiers to score edema severity from either an image or a radiology report.

The embeddings from matching image-text pairs are encouraged to be close while mismatched pairs are encouraged to yield embedding vectors that are far from each other.

The key methodological innovation in this thesis is introducing measures of local dependencies between image and text features in the context of medical images and associated radiology reports. Rather than require that the embedding vectors represent the entire image or radiology report, we build representations of local image or sentence features and optimize them for downstream prediction. We estimate and maximize the mutual information between the representations that correspond to the same finding. Because the embedding task requires no explicit labels, it can be learned from the entire image-text collection, not only the images that could be labeled with the edema severity, vastly increasing the amount of data available and, in turn, improving the quality of the resulting image-based predictions.

The image-text joint learning framework demonstrated in this thesis offers a novel pathway for model explanation by associating local image features with sentence-level text representations in the embedding space. When interpreting model predictions, pairs of an image region and a sentence that contribute the most to such prediction can be presented to explain the underlying pathology using natural language.

Such image model interpretation can be learned from the already existing clinical data archive. In clinical practice, for instance, radiologists look for the radiographic features in the images, infer the underlying pathologies and conditions of the patients, and document their thought process in the radiology reports. When a computer vision model is trained from the images and the structured labels, the thought process is *neglected*. Jointly learning from the images and the text that describe the images opens a door to reproduce the domain experts' feature extraction process and manifest it using natural language.

Maximization of mutual information between local image features and sentence-level text features is a theoretic sounding approach to learning the image interpretation documented in the text. MI measures the statistical dependencies – both linear and nonlinear – between two variables. Given a large number of image-text pairs, the image regions and the corresponding sentences can statistically be captured by an MI estimator and should exhibit high MI values. A sentence that has highest MI value with a given image region is likely to be associated with each other. This association can also be adjusted by the specific downstream computer vision tasks.

7.3 Joint Learning between Images and Tabular Data

This thesis has developed methodologies for multimodal representation learning and has demonstrated its advantages in image-text joint learning. In addition to images and text, clinical data also includes tabular data, such as vital signs, laboratory tests, hospitalization history. These multimodal sources of information reflect different yet correlated manifestations of a subject's underlying physiological processes.

The machine learning methods proposed in this thesis can naturally make use of the correlations between medical images and tabular data to improve computer vision models and/or to predict clinical events.

In the context of heart failure patients, tabular data such as blood tests that indicate heart and kidney functions, body weight, and oxygen saturation, implies the severity of underlying pathophysiology condition. The severity of pulmonary edema is a manifestation of this underlying condition. Jointly learning an image model with the features extracted from those usually longitudinal tabular data can potentially yield task-specific image representations that are more indicative of the prediction task. Therefore, joint learning with other data modalities holds promise to more accurate computer vision models.

Joint learning between images and certain adverse clinical events can further make use of multimodal machine learning to recognize patterns that have not yet been picked up by domain experts. For example, a patient's readmission or mortality may reflect the severity of certain disease or symptoms that we would like our computer vision models to learn. The grounding of the machine learning tasks in those clinical events can further shape the embedding space for images and other data to accurately reflect the related condition status of the patient.

7.4 Clinical Implications

Automatic and quantitative phenotyping from medical images, such as computer vision models we have developed in this thesis to grade pulmonary edema from chest radiographs, have two major implications for clinical practice and research: 1) offering timely impression of imaging tests for physicians; 2) enabling clinical research or decision making that is based on series of images. In this section, we discuss one potential clinical use case that my thesis may lead to and future directions.

Most CHF hospitalizations are driven by symptoms from fluid overload [36]. Clinical practice guidelines recommend removal of all excess fluid prior to discharge; patients discharged with remaining fluid overload are more likely to be readmitted or die post-discharge [80]. Our computer vision models for pulmonary edema assessment in chest radiographs can potentially assist physicians in discharge decision making and post-discharge transitional care planning.

Our preliminary retrospective analysis suggests that there is underutilized information in chest radiographs that is indicative of readmission risk. Our analysis is based on clinical records of 431 CHF patients' hospital stays that had chest x-ray taken within 24 hours before discharge, based on the MIMIC-IV and MIMIC-CXR datasets [46]. Of these 431 images, 47 pre-discharge x-ray images were labeled by our computer vision model as containing more than mild edema (39 patients with interstitial edema and 8 patients with alveolar edema), as shown in Table 7.1. Of these, 18 interstitial edema patients (46.15%) and 5 alveolar edema patients (62.50%) were readmitted within 30 days of discharge. Of the remaining 384 patients, 137 (35.68%) were readmitted, close to and below the overall 30-day readmission rate of 37.93%. Thus the presence of moderate edema increases the chances of readmission from 38% to 46% and the presence of severe edema further increases the risk of readmission to 63%. We note that the readmission event in this preliminary analysis includes both readmission for in-patient hospitalization and readmission for ER observation. Therefore, the overall readmission rate here (about 37%) is higher than the national average (about 20%) which only considers readmission for in-patient hospitalization.

Pulmonary edema severity estimate	0: None	1 (mild): Cardio-vascular congestion	2 (moderate): Interstitial edema	3 (severe): Alveolar edema	Overall
30-day *readmission* rate	30.43% (21/69)	36.83% (116/315)	46.15% (18/39)	62.50% (5/8)	37.93% (160/431)

Table 7.1: The 30-day readmission rates in different pulmonary edema severity groups based on the last pre-discharge chest x-ray image.

Since all the 431 images were taken within 24 hours before discharge, an application based on the information extracted using our chest x-ray model may improve discharge decision making or post-discharge pathway planning. To investigate the feasibility and challenges of developing this type of application and shed light on future directions, we looked into those 8 patients whose pre-discharge chest radiographs were labeled as level 3: alveolar edema by our model and who were discharged from the hospital within 24 hours after the x-ray was taken. We looked into their hospitalization records, the discharge notes of the index admissions, and the radiology reports for the 8 images. Table 7.2 summarizes our key findings.

Patients 1-3 were all admitted due to CHF exacerbation with kidney disease. Their corresponding radiology reports all mentioned pulmonary edema, except for patient 2, somewhat ambiguous on pulmonary edema severity. According to the discharge notes, all three patients were treated with aggressive diuresis during the index hospital stays. In particular, patient 2 was discharged with this condition, because a cardiologist was consulted who advised it will be difficult to medically optimize this patient's CHF given severe mitral regurgitation. Patients 1 and 3 were both discharged to their home instead of other transitional care facilities. These findings on the retrospective clinical records indicate that the clinical teams might have underestimated the remaining fluid overload of those two patients at discharge.

Patients 4 and 5 were both admitted to the hospital due to medical conditions other than CHF exacerbation. "Increased opacification" mentioned in the radiology report of patient 5 was likely to describe pneumonia. In the discharge note of patient 5, the clinical team discussed that it was difficult to determine if there was

Patient	Primary reason for index admission	Impression from radiology reports on the pre-discharge CXRs		Discharge location	Readmit in (days)
–	–	Relevant terms in the reports	Severity estimate	–	–
1	CHF exacerbation with CKD	Pulmonary vascular redistribution	1,2,3	Home	4
2	CHF exacerbation with CKD	Moderate-to-severe pulmonary edema	2,3	Long-term acute care	12
3	CHF exacerbation with CKD	Pulmonary edema	1,2,3	Home	16
4	Cardioembolic stroke	Pulmonary vascular congestion	1	Rehab	27
5	Pneumonia	Increased opacification	N/A	Home	28
6	Tracheostomy tube dysfunction	Moderate pulmonary edema	2,3	Unknown	33
7	Left femoral neck fracture	Mild-to-moderate pulmonary edema	2,3	Nursing facility	57
8	CHF exacerbation	Elevated pulmonary venous pressure	N/A	Nursing facility	146

Table 7.2: The clinical course of the patients whose pre-discharge chest radiographs were labeled as level 3: alveolar edema by our image model.

new pneumonia or whether the air space findings were unresolved findings consistent with recently treated pneumonia, without mentioning fluid overload. Patient 4 was discharged with an increased dosage of diuretics, showing that the clinical team was aware of the patient's fluid overload.

Patients 1-5 were all readmitted to the hospital within 30 days of discharge. The potentially underutilized information in chest radiographs may be extracted by machine learning models and help clinical teams optimize discharge decision making (e.g., timing of discharging a patient) and discharge planning (e.g., post-discharge medication planning, transitional care planning). Our current computer vision model has been developed and trained to assess the severity of pulmonary edema from a single x-ray image in order to match radiologists' assessment. Future work could train a machine learning model that directly predicts the readmission risk using retrospective clinical records. Furthermore, comparing a CHF patient's conditions to their baseline status is important to clinical decision making. Future work could also train machine learning models based longitudinal chest x-ray images. Lastly, pulmonary edema, manifesting in chest radiographs, is one of the many conditions caused by CHF. Predicting readmissions from a series of chest x-ray images and other clinical data of a patient could characterize the patient's risk holistically and may have a greater clinical impact.

7.5 Summary

My thesis develops machine learning methods that exploit multimodal clinical data to improve medical image analysis. Clinical data is often multimodal, including images, text (e.g., radiology reports, clinical notes), and numerical signals (e.g., vital signs, lab results). This thesis proposes machine learning methods that make use of the correlations between medical images and free-text reports to yield accurate computer vision models. My thesis demonstrates the advances of this multimodal learning approach in the application of chest radiograph analysis.

Bibliography

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

[2] Julia Adler-Milstein, A Jay Holmgren, Peter Kralovec, Chantal Worzala, Talisha Searcy, and Vaishali Patel. Electronic health record adoption in us hospitals: the emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association*, 24(6):1142–1148, 2017.

[3] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168. PMLR, 2018.

[4] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

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

[7] David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, page None, 2003.

[8] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

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

[10] Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, Sarah D De Ferranti, James Floyd, Myriam Fornage, Cathleen Gillespie, et al. Heart disease and stroke statistics—2017 update: a report from the american heart association. *circulation*, 135(10):e146–e603, 2017.

[11] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pages 517–526. PMLR, 2017.

[12] Simon Chakko, David Woska, Humberto Martinez, Eduardo De Marchena, Laurie Futterman, Kenneth M Kessler, and Robert J Myerburg. Clinical, radiographic, and hemodynamic correlations in chronic congestive heart failure: conflicting results may lead to inappropriate care. *The American journal of medicine*, 90(1):353–359, 1991.

[13] 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.

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

[15] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[17] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[18] Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

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

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[21] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[22] Jared A Dunnmon, Darvin Yi, Curtis P Langlotz, Christopher Ré, Daniel L Rubin, and Matthew P Lungren. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2):537–544, 2019.

[23] Gary S Francis, Rebecca Cogswell, and Thenappan Thenappan. The heterogeneity of heart failure: will enhanced phenotyping be necessary for future clinical trial success?, 2014.

[24] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pages 277–286, 2015.

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

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

[27] Matthias Hammon, Peter Dankerl, Heinz Leonhard Voit-Höhne, Martin Sandmair, Ferdinand Josef Kammerer, Michael Uder, and Rolf Janka. Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach. *BMC anesthesiology*, 14(1):1–9, 2014.

[28] Margaret O Harrison, Patrick J Conte, and ER Heitzman. Radiological detection of clinically occult cardiac failure following myocardial infarction. *The British journal of radiology*, 44(520):265–272, 1971.

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

[30] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

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

[33] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[34] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[35] 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*, page e190228, 2021.

[36] Brian A Houston, Rohan J Kalathiya, Daniel A Kim, and Sammy Zakaria. Volume overload in heart failure: an evidence-based review of strategies for treatment and prevention. In *Mayo Clinic Proceedings*, volume 90, pages 1247–1261. Elsevier, 2015.

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

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

[39] Sandra L Jackson, Xin Tong, Raymond J King, Fleetwood Loustalot, Yuling Hong, and Matthew D Ritchey. National burden of heart failure events in the united states, 2006 to 2014. *Circulation: Heart Failure*, 11(12):e004873, 2018.

[40] Anja K Jaehne and Emanuel P Rivers. Early liberal fluid therapy for sepsis patients is not harmful: Hydrophobia is unwarranted but drink responsibly. *Critical care medicine*, 44(12):2263, 2016.

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

[42] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 2019.

[43] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.0.0). *PhysioNet. https://doi.org/10.13026/8360-t248.*, 2019.

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

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

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

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[49] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[50] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[51] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[52] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[53] Ruizhi Liao, Geeticka Chauhan, Polina Golland, Seth J Berkowitz, and Steven Horng. Pulmonary edema severity grades based on MIMIC-CXR (version 1.0.1). *PhysioNet. https://doi.org/10.13026/rz5p-rc64.*, 2021.

[54] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. *arXiv preprint arXiv:2103.04537*, 2021.

[55] Ruizhi Liao, Daniel Moyer, Polina Golland, and William M Wells. Demi: Discriminative estimator of mutual information. *arXiv preprint arXiv:2010.01766*, 2020.

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

[57] R Bruce Logue, James V Rogers Jr, and Brit B Gay Jr. Subtle roentgenographic signs of left heart failure. *American Heart Journal*, 65(4):464–473, 1963.

[58] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

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

[60] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[61] Hooman Mahdyoon, Roger Klein, William Eyler, Jeffrey B Lakier, SC Chakko, and Mihai Gheorghiade. Radiographic pulmonary congestion in end-stage congestive heart failure. *The American journal of cardiology*, 63(9):625–627, 1989.

[62] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.

[63] Manu LNG Malbrain, Niels Van Regenmortel, Bernd Saugel, Brecht De Tavernier, Pieter-Jan Van Gaal, Olivier Joannes-Boyau, Jean-Louis Teboul, Todd W Rice, Monty Mythen, and Xavier Monnet. Principles of fluid management and stewardship in septic shock: it is time to consider the four d's and the four phases of fluid therapy. *Annals of intensive care*, 8(1):1–16, 2018.

[64] Anna M Maw, Brian P Lucas, Brenda E Sirovich, and Nilam J Soni. Discharge-ready volume status in acute decompensated heart failure: a survey of hospitalists. *Journal of Community Hospital Internal Medicine Perspectives*, 10(3):199–203, 2020.

[65] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884, 2020.

[66] ENC Milne. Correlation of physiologic findings with chest roentgenology. 1973.

[67] Mehdi Moradi, Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood. A cross-modality neural network transform for semi-automatic medical image annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–307. Springer, 2016.

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

[69] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pages 1083–1093. PMLR, 2020.

[70] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[71] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[72] William H Noble and DJ Sieniewicz. Radiological changes in controlled hypervolaemic pulmonary oedema in dogs. *Canadian Anaesthetists' Society Journal*, 22(2):171–185, 1975.

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

[74] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

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

[76] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180, 2019.

[77] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[78] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[79] Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *European Conference on Computer Vision*, pages 808–822. Springer, 2012.

[80] Jorge Rubio-Gracia, Biniyam G Demissei, Jozine M Ter Maaten, John G Cleland, Christopher M O'Connor, Marco Metra, Piotr Ponikowski, John R Teerlink, Gad Cotter, Beth A Davison, et al. Prevalence, predictors and clinical outcome of residual congestion in acute decompensated heart failure. *International journal of cardiology*, 258:185–191, 2018.

[81] Marc D Samsky, Andrew P Ambrosy, Erik Youngson, Li Liang, Padma Kaul, Adrian F Hernandez, Eric D Peterson, and Finlay A McAlister. Trends in readmissions and length of stay for patients hospitalized with heart failure in canada and the united states. *JAMA cardiology*, 4(5):444–453, 2019.

[82] Jarrel CY Seah, Jennifer SN Tang, Andy Kitchen, Frank Gaillard, and Andrew F Dixon. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology*, 290(2):514–522, 2019.

[83] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[84] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in neural information processing systems*, pages 2951–2961, 2017.

[85] Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pages 10–18. Springer, 2016.

[86] PD Snashall, Suzanne J Keyes, Barbara M Morgan, RJ McAnulty, PF Mitchell-Heggs, JM McIvor, and KA Howlett. The radiographic detection of acute pulmonary oedema. a comparison of radiographic appearances, densitometry and lung water in dogs. *The British journal of radiology*, 54(640):277–288, 1981.

[87] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

[88] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008.

[89] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[90] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[91] JOSEPH M VAN DE WATER, Jaen-Min Sheh, NICHOLAS E O'CONNOR, IAN T MILLER, and ERIC NC MILNE. Pulmonary extravascular water volume: measurement and significance in critically ill patients. *Journal of Trauma and Acute Care Surgery*, 10(6):440–449, 1970.

[92] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590, 2017.

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[94] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pages 577–585, 2014.

[95] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[96] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[97] William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996.

[98] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[99] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[100] Yuan Xue and Xiaolei Huang. Improved disease classification in chest x-rays with transferred features from report generation. In *International Conference on Information Processing in Medical Imaging*, pages 125–138. Springer, 2019.

[101] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[102] Claire Y Zhao, Minnan Xu-Wilson, Sandeep R Gangireddy, and Steven Horng. Predicting disposition decision, mortality, and readmission for acute heart failure patients in the emergency department using vital sign, laboratory, echocardiographic, and other clinical data. *Circulation*, 138(Suppl_1):A14287–A14287, 2018.

[103] Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, 2018.