GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition

Shih-Cheng Huang* Liyue Shen* Matthew P. Lungren Serena Yeung

Stanford University

Abstract

In recent years, the growing utilization of medical imaging is placing an increasing burden on radiologists. Deep learning provides a promising solution for automatic medical image analysis and clinical decision support. However, large-scale manually labeled datasets required for training deep neural networks are difficult and expensive to obtain for medical images. The purpose of this work is to develop label-efficient multimodal medical imaging representations by leveraging radiology reports. We propose an attention-based framework for learning global and local representations by contrasting image sub-regions and words in the paired report. In addition, we propose methods to leverage the learned representations for various downstream medical image recognition tasks with limited labels. Our results demonstrate high-performance and label-efficiency for image-text retrieval, classification (finetuning and zero-shot settings), and segmentation on different datasets.

1. Introduction

Advancements in medical imaging technologies have revolutionized healthcare practices and improved patient outcome. However, the growing number of imaging studies in recent years places an ever-increasing burden on radiologists, impacting the quality and speed of clinical decision making. While deep learning and computer vision provide a promising solution for automating medical image analysis, annotating medical imaging datasets requires domain expertise and is cost-prohibitive at scale. Therefore, the task of building effective medical imaging models is hindered by the lack of large-scale manually labeled datasets.

To address this problem, a natural solution is to leverage the corresponding medical reports that contain detailed descriptions of the medical conditions observed by radiologists. Several recent works utilize these medical reports to provide supervision signals and learn multimodal representations by maximising mutual information between the global representations of the paired image and report [13, 3, 41, 40]. However, pathology usually occupies only small proportions of the medical image, making it difficult to effectively represent these subtle yet crucial visual cues using global representations alone. This motivates a need for learning localized features to capture fine-grained semantics in the image in addition to global representations. While the idea of learning local representations has been explored in several other contexts for natural images [7, 27, 25, 4], including image-text retrieval and text-to-image generation, these works typically require pre-trained object detection models to extract localized image features, which are not readily available for medical images.

In this work, we focus on jointly learning global and local representations for medical images using the cor-

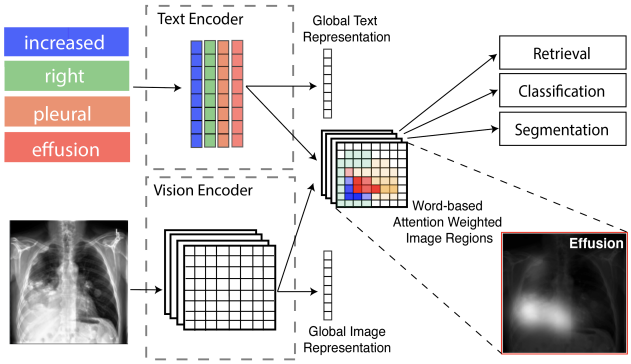

Figure 1: Our multimodal global-local representation learning framework (GLoRIA) extracts features through the image and text encoders, and learns global and localized representations by contrasting attention-weighted image sub-regions and words in the reports. The learned global-local representations are utilized to obtain label-efficient models for various downstream tasks including image-text retrieval, classification (fine-tuning and zero-shot settings) and segmentation.

*Equal Contribution
Correspondence: mschuang@stanford.edu
Emails: {mschuang,liyues,mlungren,syyeung}@stanford.edu
Code: https://github.com/marshuang80/gloria

responding radiology reports. Specifically, we introduce **GLoRIA**: a framework for learning **G**lobal-**Lo**cal **R**epresentations for **I**mages using **A**ttenion mechanism by contrasting image sub-regions and words in the paired report. Instead of relying on pretrained object detectors, we learn attention weights that emphasize significant image sub-regions for a particular word to create context-aware local image representations (Fig. 1). Due to the lengthy nature of medical reports, we introduce a self-attention-based image-text joint representation learning model, which is capable of multi-sentence reasoning. Furthermore, we propose a token aggregation strategy to handle abbreviations and typos common in medical reports.

We demonstrate the generalizability of our learned representations for data-efficient image-text retrieval, classification and segmentation. We conduct experiments and evaluate our methods on three different datasets: CheXpert [16], RSNA Pneumonia [32] and SIIM Pneumothorax. Utilizing both global and local representations for image-text retrieval is non-trivial due to the difficulty in incorporating multiple representations for each image-text pair. Therefore, we introduce a similarity aggregation strategy to leverage signals from both global and local representations for retrieval. Furthermore, our localized image representations are generated using attention weights that rely on words to provide context. Thus, to leverage localized representations for classification, we generate possible textual descriptions of the severity, sub-type and location for each medical condition category. This allows us to frame the image classification task by measuring the image-text similarity and enables zero-shot classification using the learned global-local representations. Finally, experimental results on various tasks and datasets show that our GLoRIA achieves good performance with limited labels and consistently outperforms other methods in previous works.

Our contribution can be summarized as follows: (1) We propose GLoRIA: a framework for jointly learning multimodal global and local representations of medical images by contrasting attention weighted image regions with words in the paired reports and (2) we demonstrate the label-efficiency of our framework by evaluating the learned multimodal global-local representations on image-text retrieval, classification (finetune and zero-shot) and segmentation tasks with limited labels.

2. Related Work

2.1. Utilizing radiology reports for medical images

To leverage information from radiology reports, a number of previous work explore methods for extracting labels from reports via natural language processing (NLP) as a surrogate for manual annotations [16, 35, 18]. Although these approaches can be scaled to generate labels for large datasets, the extracted labels are noisy and often limit the model's performance. Furthermore, these efforts disregard the rich and detailed descriptions originally contained in the reports during the process of label extraction.

Deep learning models that utilize both text and image data as inputs have drawn more attention in recent years. These methods extract knowledge from both the image and corresponding report by leveraging attention mechanisms or image-text transformers [28]. However, some of these approaches require radiology reports as inputs for inference, making them less applicable in contemporaneous model deployment in practice. Other studies have developed methods that avoid the need of text reports during inference [41, 3], but they still require large-scale manual annotations during training. In addition, very few prior works investigates methods for learning localized features for multimodal data which is crucial for medical images.

In contrast, image-text joint representation learning strategies typically do not require manual annotations during training and can be used to fine-tune for downstream tasks using only one of the modalities. For instance, [13] uses an unsupervised adversarial training and showed promising results for image-text retrieval. [40] maximizes mutual information across modalities through contrastive learning and evaluate on retrieval and classification tasks. However, these studies considered only global representations, which can be limiting since medical conditions often occupy a small proportion of the entire medical image. Our work builds on top of these prior works by jointly learning both global and local multimodal representations for medical images by leveraging medical reports.

2.2. Localized image-text representation learning

Image-text joint representation learning has been studied extensively for tasks such as VisualQA [2, 11, 14, 39], image captioning [22, 36, 21, 20], and image-text retrieval [5, 38, 8, 9, 15, 38, 34]. Recent studies have achieved progress by utilizing localized representations through stacked attention [25], semantic ordering [15] and graph convolutional neural networks [27, 7]. Most of these works rely on object detection models that are pretrained using natural image datasets to extract image region features. While object detectors are effective for natural images, direct transfer to medical image datasets is limited by the domain gap between medical and natural images. Furthermore, few previous works have applied the learned representations to tasks beyond image-text retrieval.

Other works explore learning localized representation without relying on pretrained object detection model, but only demonstrated effectiveness for specific natural image tasks. [38] proposed to use a ranking loss function to learn both global and localized representations for image-text retrieval. [37] learned attention weights to unify image

regions and word representations for fine-grained text-to-image generation. However, medical reports often contain typographical errors, as well as long-range context dependencies, which introduce unique challenges not common in natural image-caption datasets. We address these challenges by utilizing a self-attention based model that are effective for multi-sentence reasoning and propose a token aggregation strategy.

2.3. Zero-shot classification

Since zero-shot learning was introduced [24], many studies have investigated methods to classify images without training labels [23, 33, 10, 26, 6, 31, 29]. One possible solution is to leverage information from other modalities [23, 33, 10, 26]. Recent efforts introduce strategies for learning visual representations using text data as supervision [6, 31, 29]. However, these methods only learn global representations of images, which can be limiting when applied to medical image recognition tasks due to the high inter-class similarities among the medical images that are distinguishable only by very subtle visual cues. In contrast, our work jointly learns global and local representations, which can provide complementary information from both the full image and the critical local region of interest.

3. Method

The goal of this work is to jointly learn global and local multimodal representations of medical images by leveraging medical reports for various downstream tasks where manual annotations are limited. Specifically, we observe that pathologies present in medical imaging examinations often occupy a small proportion of the image and only correspond to certain key words in medical reports. Motivated by this, we propose an attention-based framework for multimodal representation learning by contrasting image sub-regions to words in the corresponding report. Our method generates context-aware local representations of images by learning attention weights that emphasize significant image sub-regions for a particular word. Here we first describe in Sec 3.1 the image and text encoders we use to extract features from each modality. In Sec. 3.2, we formalize our multimodal global-local representation learning objective. Finally, in Sec. 3.3, we present strategies for utilizing both global and local representations for label-efficient and zero-shot learning in various downs-stream tasks.

3.1. Image and text encoding

Given a paired input $[x_v, x_t]$, where x_v denotes an image and x_t is the corresponding report, we use an image encoder E_v and a text encoder E_t to extract global and local features from each modality. The global features contain the semantic information that summarizes the image and report. The local image features capture the semantics in the image sub-regions, while the local text features are word-level embeddings. These global and local features are used to learn multimodal representations using our framework, and the encoders are trained jointly with our representation learning objective. We then apply the learned representations to downstream image recognition tasks, such as retrieval, classification and segmentation.

3.1.1 Image encoding

To construct the image encoder E_v, we use the ResNet-50 architecture [12] as the backbone to extract features from the image. The global image features $f_g \in \mathbb{R}^C$ are extracted from the final adaptive average pooling layer of the ResNet-50 model, where C denotes the feature dimension. We extract the local image features from an intermediate convolution layer and vectorize to get the C-dimentional features for each of M image sub-regions: $f_l \in \mathbb{R}^{C \times M}$.

3.1.2 Text encoding

Medical reports typically consist of long paragraphs and require reasoning across multiple sentences. Therefore, we utilize a self-attention based language model for learning long-range semantic dependencies in medical reports. In particular, we use the BioClinicalBERT [1] model, pretrained with medical texts from the MIMIC III dataset [19] as our text encoder E_t to obtain clinical-aware text embeddings. We further employ word-piece tokenization to minimize the out-of-vocabulary embeddings for abbreviations and typographical errors which are common in medical reports. For a medical report with W words, each word is tokenized to n_i sub-words. The tokenizer would generate a total of $N = \sum_{i=0}^{W} n_i$ word piece embeddings as the input to the text encoder. The text encoder extracts features for each word piece, respectively. Thus, the local text features output from the text encoder can be denoted as $g_l \in \mathbb{R}^{K \times N}$ where K is the dimension of each word-piece feature. The global text feature is defined as the aggregation of all the word-piece features $g_g = \sum_{i=0}^{N} g_{li}$.

3.2. Global and local representation learning

An overview of our representation learning framework is shown in Figure 2. In addition to training image encoder E_v and text encoder E_t for feature extraction, we also learn global representation functions (denoted as R_{vg}, R_{tg} for image and text features respectively) and local representation functions (denoted as R_{vl}, R_{tl} for image and text features) to project the image and text features to a multimodal semantic space, where representations from true pairs of image and text are in close proximity. The overall representation learning objectives contain: 1) global contrastive loss that learns to relate the entire image to the paired report, and 2) local contrastive loss that learns fine-grained alignments

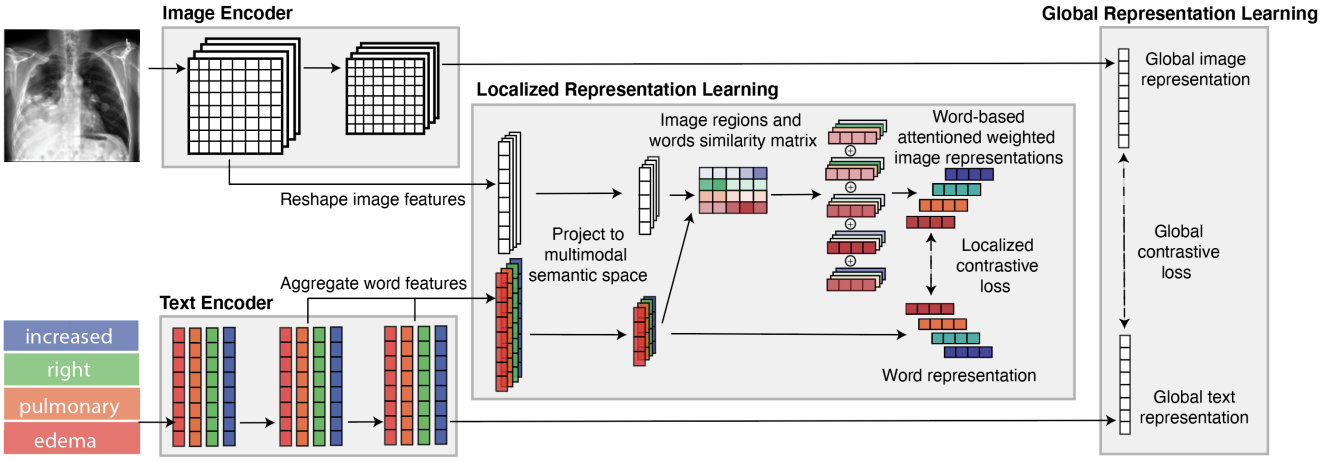

Figure 2: Overview of the proposed multimodal global-local representation learning framework (GLoRIA). Given a pair of medical image and report, we first use the image encoder and text encoder to extract image and text features respectively. The global image-text representations are learned through the global constrastive loss. For learning local representations, we compute the similarity matrix based on the image sub-region features and word-level features to generate attention-weighted image representations. The local contrastive objective is based on the attention-weighted image representations and the corresponding word representations. The overall representation learning framework is trained end-to-end by jointly optimizing both local and global contrastive losses.

between image sub-regions and word pieces. Through simultaneously training with both global and local losses, the model is able to learn better global and local representations using complementary mutual information.

3.2.1 Multimodal embedding framework

For each input image, we use the image encoder E_v detailed in Sec. 3.1 to extract both global and local features. Next, we train the global and local image representation learning functions R_{vg} and R_{vl} to transform the global and local image features to representations in multimodal feature space: $v_g = R_{vg}(E_v(x_v))$ and $v_l = R_{vl}(E_v(x_v))$. The global image representation $v_g \in \mathbb{R}^D$ is a single D-dimensional vector while local image representation $v_l \in \mathbb{R}^{D \times M}$ consists of D-dimensional vectors for all M image regions.

As aforementioned, we overcome challenges from abbreviations and typographical errors common in medical reports by using tokenization to represent words as word-piece embeddings. However, we want to learn the correspondence of visual semantics to specific words instead of word-pieces for precise multi-modal representations, particularly for medical terms. For instance, instead of finding visual signals for each word-pieces [*"Car", "dio", "mega", "ly"*], it is important to understand the direct correspondence of the term *"Cardiomegaly"* to the image sub-regions that contain an enlarged heart. Therefore, we aggregate by averaging the word-piece features encoded by the text encoder to obtain word-level features. The aggregated word-level features are then projected to representations in D dimensional multimodal feature space using global and local representation learning functions denoted as R_{tg} and R_{tl} respectively: $t_g = R_{tg}(E_t(x_t))$ and $t_l = R_{tl}(E_t(x_t))$.

3.2.2 Global contrastive loss

Since medical reports contain detailed descriptions of the observations for the corresponding medical images, the paired image and report are expected to have similar semantic information in the multimodal feature space. Thus, the first learning objective is to maximize the alignment between the true pairs of image and text versus random pairs by using the global representations. To achieve this, we follow [40, 29] to use contrastive loss functions for maximizing the posterior probability of the global image representation v_{gi} given its corresponding text representation t_{gi}. Therefore, the global objective is formulated as minimizing the negative log posterior probability:

$$L_g^{(v|t)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gi}, t_{gk}\rangle/\tau_1)}\right) \quad (1)$$

where $\tau_1 \in \mathbb{R}$ is a scaling temperature parameter, and $\langle v_{gi}, t_{gi}\rangle$ represents the cosine similarity between the global image representation v_{gi} and global text features t_{gi}.

Similarly, due to the mutual correlation between the image and text pairs, we also maximize the posterior probability of the text given its corresponding image. In this way, it is ensured that the image-text correlation is asymmetric to either modality.

$$L_g^{(t|v)} = \sum_{i=1}^{N} -\log\left(\frac{\exp(\langle v_{gi}, t_{gi}\rangle/\tau_1)}{\sum_{k=1}^{N}\exp(\langle v_{gk}, t_{gi}\rangle/\tau_1)}\right) \quad (2)$$

3.2.3 Attention weighted image representation

While the global contrastive loss constrains alignment between the entire image and text, relying only on global representations can be limiting for medical image recognition. In contrast to natural images, the regions of interest for medical images are indicated by very subtle visual cues, and can easily be underrepresented using global feature alone. To bypass the reliance of pretrained object detection models for extracting image features for sub-regions, we instead learn attentions that weigh different image sub-regions based on their significance for a given word. By contrasting the attention weighted image representations to the corresponding word embedding, the attention weights are learned as part of our local representation objective.

To generate the word-based attention weighted image representation, we first compute the dot-product similarity between all combinations of local text and image features:

$$s = v_l^T t_l \tag{3}$$

$s \in \mathbb{R}^{M \times W}$ indicates the similarity matrix between W words and M image sub-regions. Thus, $s_{i,j}$ corresponds to the similarity between the word i in the text and sub-region j in the image. We normalize the similarities for each sub-region to ensure comparable similarities across the image regions.

For every word in the report, we compute an attention weighted image representation c_i based on its similarities to all the image sub-regions. The attention weight a_{ij} is the normalized similarity for a word across all image-regions:

$$a_{ij} = \frac{\exp(s_{ij}/\tau_2)}{\sum_{k=1}^{M} exp(s_{ik}/\tau_2)} \tag{4}$$

where $\tau_2 \in \mathbb{R}$ is a temperature parameter.

The context-aware image representation c_i is an attention-weighted sum of all the image sub-region features based on the sub-region's similarity to the given word:

$$c_i = \sum_{j=0}^{M} a_{ij} v_j \tag{5}$$

3.2.4 Local contrastive loss

In order to learn the attention weights introduced in previous section, we need a localized objective for training. Here, we set up a contrastive objective for learning localized multimodal representations. Specifically, we use a localized feature matching function Z to aggregate the similarities between all W word features t_i and their corresponding attention weighted image features c_i.

$$Z(x_t, x_v) = \log(\sum_{i=1}^{W} \exp(\langle c_i, t_i \rangle / \tau_3))^{\tau_3} \tag{6}$$

where $\tau_3 \in \mathbb{R}$ is another scaling factor while x_v and x_t are local features for an image and report.

Since the matching function captures the similarity between the attention-weighted image features and word-level text features, the local contrastive loss can be defined as the posterior probability based on the matching function $Z(x_t, x_v)$. This way, the local contrastive loss aims to maximize the posterior probability of the attention-weighted image region representations given the word representations:

$$L_l^{(v|t)} = \sum_{i=1}^{N} -\log(\frac{\exp(Z(x_{vi}, x_{ti})/\tau_2)}{\sum_{k=1}^{N} \exp(Z(x_{vi}, x_{tk})/\tau_2)}) \tag{7}$$

Similarly, to ensure that the multi-modal representations are asymmetric to either of the input modalities, we also minimize:

$$L_l^{(t|v)} = \sum_{i=1}^{N} -\log(\frac{\exp(Z(x_{vi}, x_{ti})/\tau_2)}{\sum_{k=1}^{N} \exp(Z(x_{vk}, x_{ti})/\tau_2)}) \tag{8}$$

3.2.5 Total loss

The final training objective for our representation learning framework contains both global $L_g^{(t|v)} + L_g^{(v|t)}$ and local $L_l^{(t|v)} + L_l^{(v|t)}$ contrastive losses. By jointly optimizing global and local objectives, both losses can mutually complement each other for learning better global and local representations simultaneously.

$$L = L_g^{(t|v)} + L_g^{(v|t)} + L_l^{(t|v)} + L_l^{(v|t)} \tag{9}$$

3.3. Utilizing global and local representations

After the multimodal representation learning stage, the learned representations can be used for different downstream tasks, including retrieval, classification and segmentation. Existing studies typically fine-tune task-specific models based on the learned global representations for different downstream tasks. However, these approaches do not take advantage of the local features learned through our framework. Jointly utilizing global and local representations for downstream tasks such as image-text retrieval is non-trivial because it requires incorporating multiple representations for each image and text pair. We therefore propose an aggregation strategy to consider both global and local image-text similarities as shown in Fig. 3

Furthermore, our localized image representations are generated using attention weights, which rely on the words to provide context. Since image classification datasets typically do not provide context words, we generate possible textual descriptions of the severity, sub-type and location for the medical condition we are predicting to represent

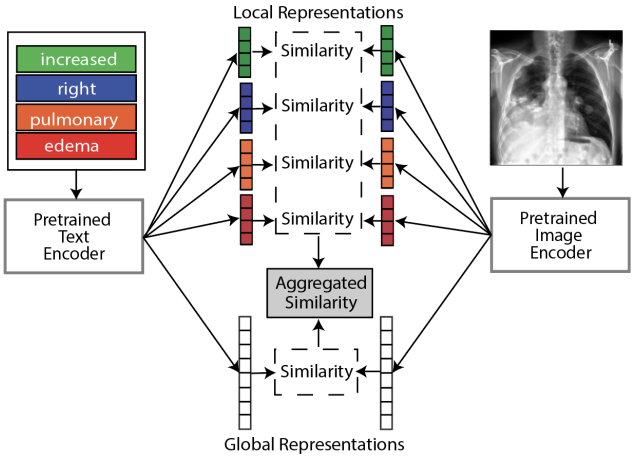

Figure 3: Hybrid global and local image-text similarity. After the feature extraction through image and text representation project, the global similarity is calculated based on global image and text representations. The local similarities are computed using on the word-based attention-weighted image representations and the corresponding word representations. The final image-text similarity is obtained by averaging the global and local similarities.

each class. This allows us to frame the image classification as an image-text similarity task, and enables zero-shot classification using the learned global-local representations.

3.3.1 Image-text retrieval

In the image-text retrieval task, a query image is used as the input to retrieve the closet matching text based on the similarities between their representations. Formally, given a query image x_v and a collection of candidate texts X_t, we extract global image and text representations v_g, t_g by using their respective encoders and representation learning function. Then the target sentence is retrieved by finding the highest similarity score: $\text{argmax}_k\, S(v_g, t_{gk})$. Note that $S(v_g, t_{gk})$ can be any similarity between the query image v_g and candidate sentence k. This formulation, however, only compares similarities between the global representations of query and candidates. For medical images, key words in the entire report often correspond to only a small proportion of the image, the fine-grained alignment between word and image regions is needed to improve retrieval performance. Thus, we propose to leverage both the global and local features for a more accurate retrieval. We use the attention-driven image-text matching-score $Z(t_{li}, v_{li})$ defined in Eq. 6 as the similarity metric for the local representations. In this way, the localized similarity between the query image and candidate sentences can be calculated base on the context-aware local representations. Finally, the image-text retrieval task is completed based on the aggregated image-text similarity metric by averaging the global and local similarities as shown in Fig. 3.

3.3.2 Zero-shot image classification

In zero-shot classification, we take an image x_v as input and aim at predicting the corresponding label $y = C(x_v)$ even though the classifier C is not explicitly trained with class labels y. Inspired by [29], we convert the classification classes into textual captions and frame the image classification task as measuring the image-text similarity. Specifically, we consult a radiologist to utilize the medical domain knowledge to generate reasonable texts to describe the possible sub-types, severities, and locations for each of the medical conditions in the classification categories. In this way, we generate such textual prompts to represent each classification class by randomly combining the possible words for sub-types, severities and locations. Next, the generated candidate prompts Y_t for all n classes are projected to multimodal embedding space using our pretrained representation learning functions: $t_g = R_{tg}(E_t(Y_t))$ and $t_l = R_{tl}(E_t(Y_t))$. Similarity, we obtain the global and local representations for the input image $v_g = R_{vg}(E_v(x_v))$, $v_l = R_{vl}(E_v(x_v))$. Therefore, the input image is classified by finding the class prompts with the highest average similarity according to global and local representation: $\text{argmax}_i[\frac{1}{2}(S(t_{gi}, v_g) + Z(t_{li}, v_l))]$.

4. Experimental Results

To validate the effectiveness of our representation learning framework, we conduct experiments using the learned global and local representations for image-text retrieval, image classification (fine-tune & zero-shot) and segmentation. We compare our method to several state-of-the-art image-text joint embedding methods and show that our method achieves better results consistently on 3 different datasets.

4.1. Datasets

CheXpert [16]. We use the CheXpert dataset to train our representation learning framework and evaluate for classification tasks. The CheXpert dataset contains a total of 224,316 chest radiographs from 65,240 patients, where each radiograph is paired with the corresponding radiology reports. Each radiograph is labeled for the presence of 14 total medical observations. In our experiment, we focus on investigating the frontal chest radiographs with 191229 image-text pairs. Following the experiments setting in [40], we hold out the expert-labeled validation set, containing 202 images, as the test set since the official test set is not publicly available. Therefore, we randomly sample 5,000 images from the training data for validation.

CheXpert 5x200. The chest radiographs in the original CheXpert dataset are multi-labeled to account for the simultaneous presence of multiple medical observations. Since our zero-shot classification and retrieval is based on finding the most similar target, having multiple possible label for

Method	Prec@5	Prec@10	Prec@100
DSVE [8]	40.64	32.77	24.74
VSE++ [9]	44.28	36.81	26.89
ConVIRT [40]	66.98	63.06	49.03
GLoRIA (Ours) - global only	67.02	64.68	49.55
GLoRIA (Ours) - local only	68.22	64.58	48.17
GLoRIA (Ours)	**69.24**	**67.22**	**53.78**

Table 1: Results of image-text retrieval on the CheXpert 5x200 dataset. The top K Precision metrics are reported for $K = 5, 10, 100$. Ours method achieves the best performance by incorporating both global and local representations.

a target can cause confounding results between categories. Therefore, following the setting in [40], we use the partial data from CheXpert to create the CheXpert 5x200 dataset, which includes 200 exclusively positive images for each of the CheXpert competition tasks: *Atelectasis, Cardiomegaly, Edema, Pleural, Effsion*. In this dataset, each image contains positive labels for only one specific condition.

RSNA Pneumonia [32]. To evaluate the generalizability of our pretrained representation framework for classification on external datasets, we use the RSNA Pneumonia dataset containing 30k frontal view chest radiographs labeled either as "healthy" or "peumothorax positive". The train/valid/test split each constitutes 70%/30%/30% of the dataset.

SIIM Pneumothorax. We use the SIIM Pneumothorax dataset to evaluate the learned representations' capability for segmentation. This dataset contains a total of 12047 chest radiographs, each paired with manually annotated segmentation masks for pneumothorax. The train/valid/test split respectively constitutes 70%/30%/30% of the dataset.

4.2. Baselines

We compare our method with other state-of-the-art multi-modal representation learning method. Within the same medical image domain, we compare our work to **ConVIRT** [40] which has shown state-of-the-art performance for image-text retrieval and classification by contrasting only global representations of image and report pairs. Since the codebase for ConVIRT is not publicly released, we implement the method according to the description in [40]. In addition, we also compare our method with other multi-modal representation learning methods proposed for natural image tasks. Most state-art-the art methods require pretrained object detection model for local feature extraction, which is not applicable for medical images. Thus, we focus on comparing ours method with **DSVE** [8], which shows localization capabilities without using object detectors. We also compare our method to **VSE++** [9] which achieves the best performance for image-text retrieval by using only global representations.

	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
Random	56.1	62.6	65.7	58.9	69.4	74.1
ImageNet	74.4	79.1	81.4	74.9	74.5	76.3
DSVE [8]	50.1	51.0	51.5	49.7	52.1	57.8
VSE++ [9]	50.3	51.2	524	49.4	57.2	67.9
ConVIRT [40]	85.9	86.8	87.3	77.4	80.1	81.3
GLoRIA (Ours)	**86.6**	**87.8**	**88.1**	**86.1**	**88.0**	**88.6**

Table 2: Results of fine-tuned image classification (AUROC score) on CheXpert and RSNA test sets based on different portion of training data: 1%, 10%, 100%.

4.3. Image-text retrieval

First, we use CheXpert 5x200 dataset to evaluate the effectiveness of our representation learning framework for image-text retrieval. Given an image as input query, we retrieve the target reports by computing the similarity between the query image and all candidate reports using the learned representations. We use the Precision@K metric to calculate the precision in the top K retrieved reports by checking if the selected report belongs to the same category as the query image.

Based on the results presented in Table 1, our model achieves comparable performance with ConVIRT when only global representations are used. This is expected since we use the same global contrastive loss as ConVIRT to train our global representations. While we find our approach to achieve slightly better results using localized representation alone, our best retrieval results are based on leveraging both the local and global representations, outperforming all the baselines by a large extent. This indicates that the global-local representations learned in our method efficiently provide complementary semantic information.

4.4. Classification

We further evaluate the learned representations on an image classification task in two different settings. For **supervised classification**, we train a linear classifier on top of the pretrained image encoder using different amounts of training data (1%, 10% or 100%) to evaluate the data-efficiency of the global image representations. For **zero-shot classification**, we employ the approach described in Sec. 3.3.2 to evaluate the effectiveness of our learned representations for classification without additional labels for fine-tuning.

In Table 2, we show the classification results for CheXpert and RSNA datasets across different percentage of training data. To account for the variance in results from randomly sampling training data, we averaged results from five independent runs. We use the area under the ROC curve (AUROC) as our evaluation metric. Our method outperforms the other representation learning methods on both datasets. It is also worth noting that our method trained with only 1% of the data consistently outperform imagenet initilized models with 100% data for training. This indi-

CheXpert	Acc.	Sens.	Spec.	PPV	NPV	F1
100%	0.57	**0.83**	0.80	0.51	**0.95**	0.63
10%	0.55	0.76	0.82	0.51	0.92	0.61
1%	0.47	0.68	0.85	0.53	0.91	0.59
Zero-shot	**0.61**	0.70	**0.91**	**0.65**	0.92	**0.67**
RSNA	Acc	Sen	Spe	PPV	NPV	F1
100%	**0.79**	0.87	0.76	**0.52**	0.95	**0.65**
10%	0.78	0.78	**0.79**	0.52	0.92	0.63
1%	0.72	0.82	0.69	0.44	0.93	0.57
Zero-shot	0.70	**0.89**	0.65	0.43	**0.95**	0.58

Table 3: Results of zero-shot image classification on the CheXpert 5x200 and RSNA datasets. Note that representation learning framework is trained using CheXpert. We compare classification results with different amounts of training data for comparison.

cates that simultaneously training both global and local contrastive objectives can also help to learn better global representations for label-efficient classification.

Although DSVE and VSE++ demonstrate effective representation learning for image-text retrieval, directly application for medical image datasets does not show comparable results. These methods only focus on minimizing the distance the representations of true image and text pairs, without contrasting with other samples. Therefore, when applied to medical images where inter-class visual similarities are high, these methods can easily overfit by learning irrelevant patient/case specific visual cues.

For zero-shot clasification, we use the CheXpert 5x200 dataset for 5 class classification and RSNA Pneumonia dataset for binary classification. We present zero-shot classification results in Table. 3. On the CheXpert dataset, our zero-shot classifier is able to achieve better F1 score as compared to classification models fine-tuned with training labels. Although we only use the CheXpert dataset to train the representation learning framework, the performance on the RSNA datasets is still comparable with supervised models fine-tuning with 1% of training data.

4.5. Segmentation

We also demonstrate the effectiveness of our representation learning framework for segmentation. Specifically, we adopt the UNet [30] architecture for segmentation and initialize the encoder portion of the model with weights from our pretrained image encoder E_v. We compare our method with random, imagenet and ConVIRT initialization. In Table 4, we report Dice scores and evaluate the data-efficiency of each method by using 1%, 10% or 100% data for training. We show that the learned representations using our framework are effective for segmentation task when limited segmentation masks are available for training.

	Pneumothorax Segmentation		
Initialization Method	1%	10%	100%
Random	0.090	0.286	0.543
ImageNet	0.102	0.355	**0.635**
ConVIRT [40]	0.250	0.432	0.599
GLoRIA (Ours)	**0.358**	**0.469**	0.634

Table 4: Results of image segmentation (Dice score) on SIIM dataset with different portion of training data: 1%, 10%, 100%.

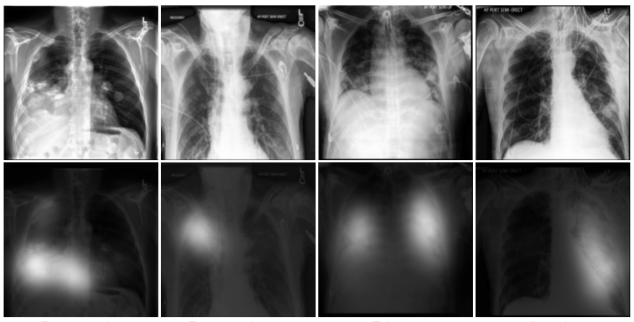

Figure 4: Examples of frontal radiographs of the chest (top) with corresponding attention weights for the given word (below).

4.6. Visualization of attention weights

We visualize the attention weights (See Eq. 4), which are trained as part of our representation learning framework, to qualitatively evaluate our method. While attention is not explanation, well-trained attention weights should correctly identify significant image regions that correspond to a particular word [17]. We reshape the attention weights to match the input image size and overlay the attention map on the original image for visualization. Fig. 4 demonstrates our attention model is able correctly identify significant image-regions for a given word. For instance, the attention based on the word "Pneumonia" Fig. 4a (bottom) correctly localize regions of the right lower lobe containing heterogenous consolidative opacities indicative of pneumonia. Similarly, the attention weights for "Pneumothorax" shown in Fig. 4b (bottom) correctly highlights luncency in the right lung apex that suggests penumothorax Fig. 4b (top). We show similar results for "Edema" and "Opacity" in Fig. 4c and Fig. 4d.

5. Conclusion

We propose a multimodal global-local representation learning framework for medical images by leveraging radiology reports. Specifically, the representations are learned by contrasting attention-weighted image sub-regions and words in the reports. Experimental results demonstrate data-efficiency and zero-shot capability of learned representations for various downstream tasks on different datatsets including retrieval, classification and segmentation.

References

[1] 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.

[2] 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.

[3] 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. *arXiv:2008.09884 [cs]*, Aug. 2020. arXiv: 2008.09884.

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663, 2020.

[5] Jingjing Chen, Lei Pang, and Chong-Wah Ngo. Cross-modal recipe retrieval: How to cook this dish? In *International Conference on Multimedia Modeling*, pages 588–600. Springer, 2017.

[6] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[7] Haiwen Diao, Ying Zhang, Lin Mafdevlin2018bert, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*, 2021.

[8] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2018.

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[10] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013.

[11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[12] 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.

[13] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised Multimodal Representation Learning across Medical Images and Reports. *arXiv:1811.08615 [cs]*, Nov. 2018. arXiv: 1811.08615.

[14] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017.

[15] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[16] 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. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[17] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[18] 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.

[19] 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.

[20] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018.

[21] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[23] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[24] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[26] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017.

[27] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4654–4662, 2019.

[28] Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1999–2004. IEEE, 2020.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[31] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. *arXiv preprint arXiv:2008.01392*, 2020.

[32] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[33] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *arXiv preprint arXiv:1301.3666*, 2013.

[34] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019.

[35] 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.

[36] 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. PMLR, 2015.

[37] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[38] Quanzeng You, Zhengyou Zhang, and Jiebo Luo. End-to-end convolutional semantic embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5735–5744, 2018.

[39] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018.

[40] 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.

[41] Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang. Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 320–328, Cham, 2017. Springer International Publishing.