Significantly Improving Zero-Shot X-ray Pathology Classification via Fine-tuning Pre-trained Image-Text Encoders

Jongseong Jang, Daeun Kyung, Seung Hwan Kim, Honglak Lee, Kyunghoon Bae, **Edward Choi**

Jeeyoung Kim

University of Ulsan College of Medicine,

Asan Medical Center

77imjee@gmail.com

2023.02.22 Wed

Introduction

- In the medical imaging domain, **qualified experts are required** to create labeled samples.
- Recent studies tackled this problem with some success by taking advantage of the **powerful pre-trained models** trained on large-scale general domain data
- **CLIP**
  - self-supervisingly trained via contrastive learning on approximately 400M image-text pairs
  - demonstrated impressive zero-shot classifications that outperformed in lots of datasets
- **CheXzero**
  - used CLIP’s generalization capability by fine-tuning it with over than 200,000 pairs of CXR and reports
  - perform zero-shot classification and successfully matching the performance of board-certificated radiologists
  - it is possible to perform pathology classification **even without having an expert-annotated training set**

Introduction

- Previous zero-shot approach **fails to properly address the multi-labeled nature of medical image-report pairs.**

- Their report must at least share the semantics of Lung Opacity and Edema – ***partial positive***

- CheXzero : uses the vanilla contrastive learning, which regards partial positive pairs as completely negative pairs
- MedCLIP : another zero-shot pathology classification study, partially addresses this issue but requires expert labels to do so

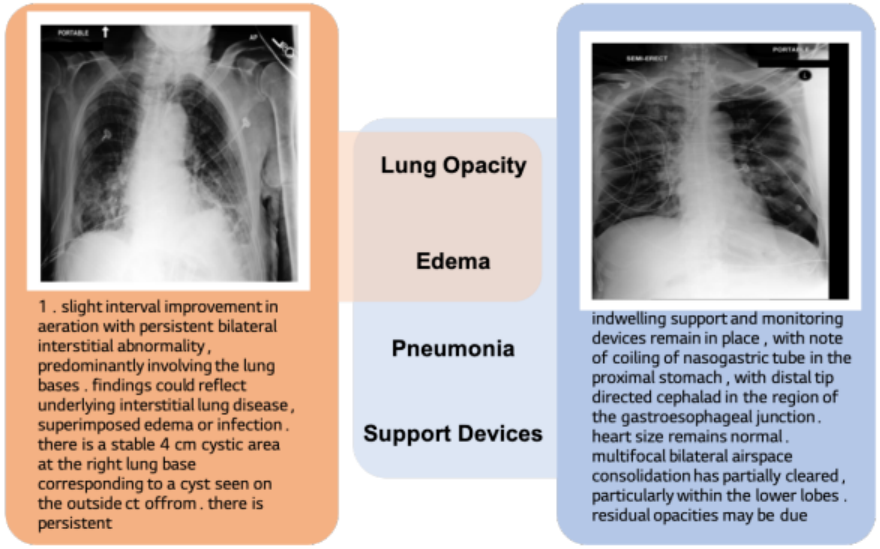


Figure 1. Example of partial positiveness between medical data. Two image-text pairs share two classes of disease; *lung opacity* and *edema*. Existing image-text contrastive learning frameworks deal other pair as perfectly negative one.

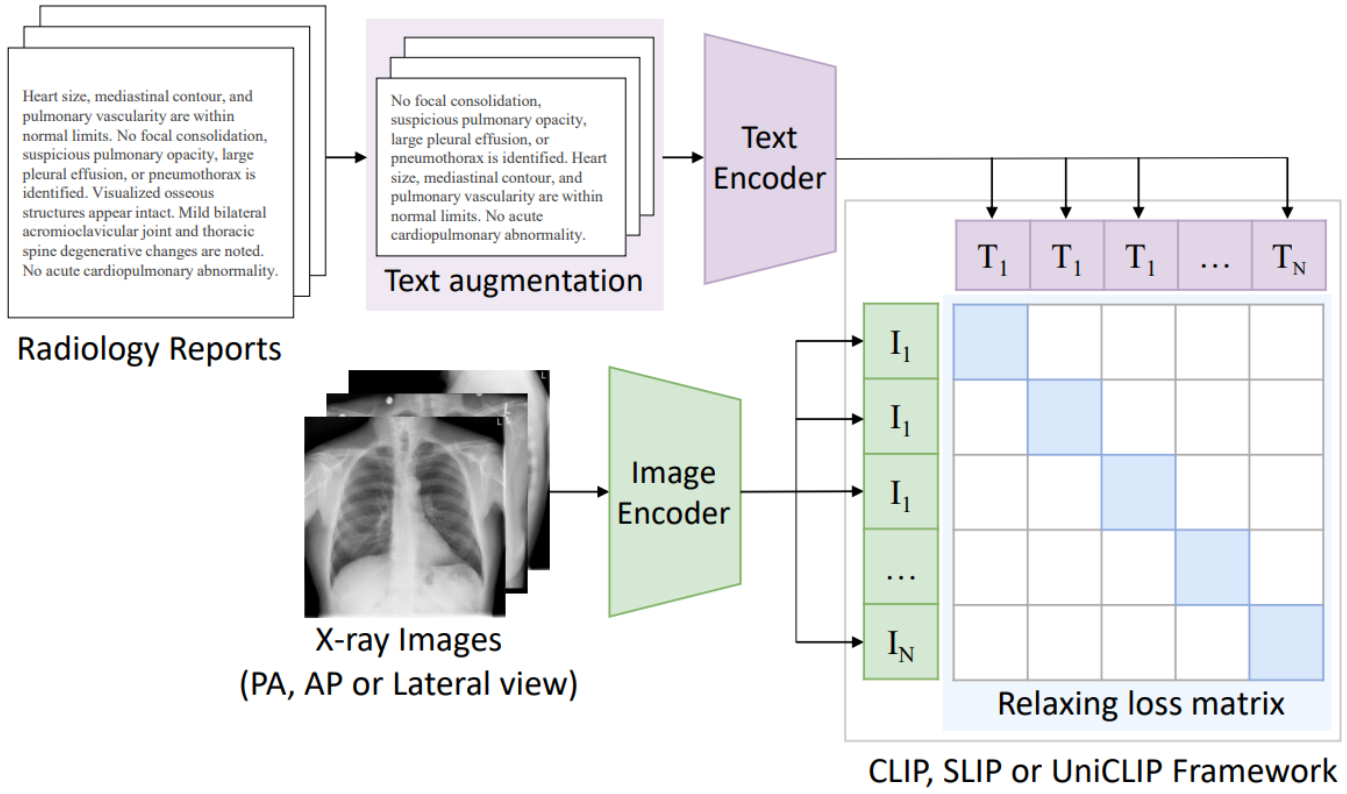
Contribution

1. We propose **a new fine-tuning strategy for zero-shot pathology classification from unannotated X-ray images**.
2. Using **four different CXR datasets** and **three different pre-trained models**, we show that the proposed method consistently **improve zero-shot pathology classification performance** that average AUROC increase 5.77%
3. Especially for the **CheXpert dataset, fine-tuning CLIP with our method sometimes outperformed board-certified radiologists marginally** in detecting five prominent disease with 0.619 vs 0.625 F1 scores and 0.530 vs 0.433 in MCC

Methods

1. **Random text augmentation**

2. **Loss for relaxed image-text agreement**



(a) Contrastive fine-tuning

Methods – Random text augmentation

- Medical reports
  - consist of meaningful content for diagnosis, and most of the sentences are **short and formal**
  - **text augmentation in the token-level can be harmful** to keep the correct clinical meaning of reports

- **Random text augmentation**
  - randomly sub-sample ***n*** sentences from the report for every training epoch and feed them to the contrastive learning framework
  - suppose the number of sentences in the report is ***m***, image could match with ${}_mC_n$ positive samples

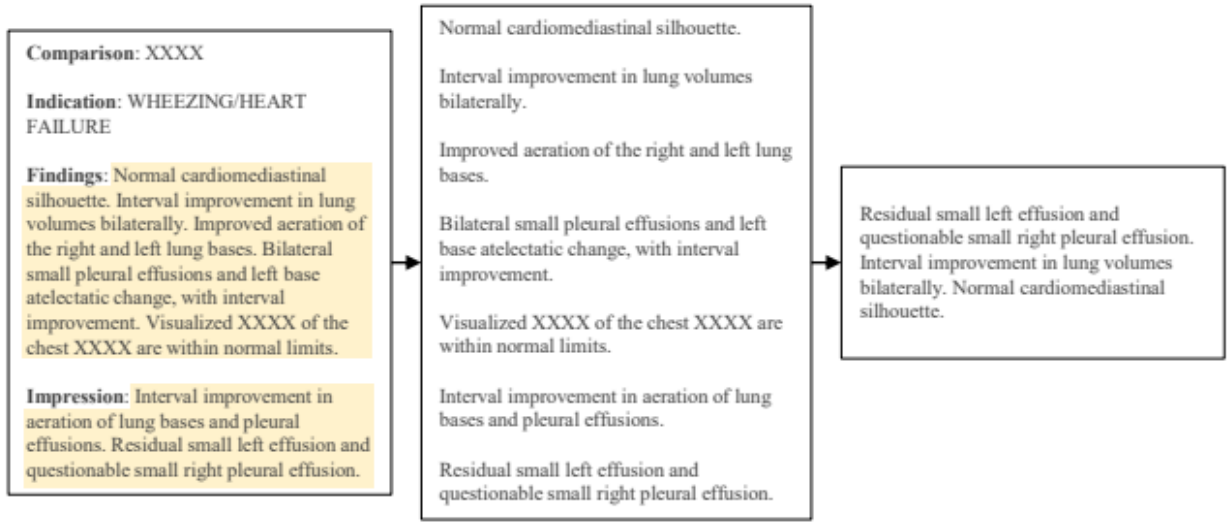
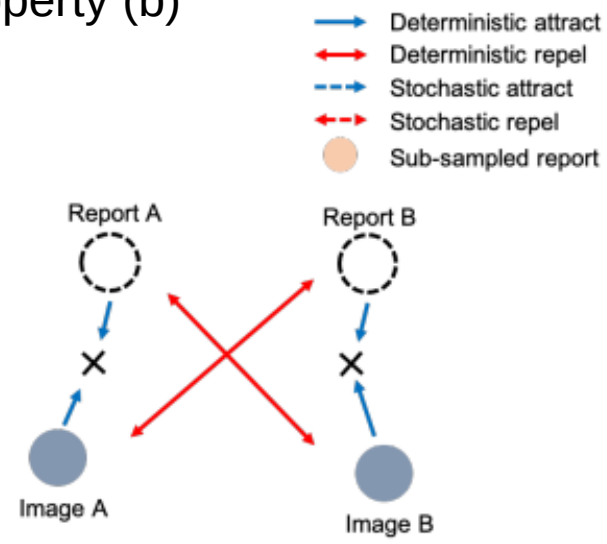


Figure 3. Example of text augmentation. Given original free-form report (Leff), we extract only "Findings" and "Impression" section, and split them into sentences (Middle). n sentences are randomly selected to make new positive text pairs (Right).

Methods – Random text augmentation

- When using the **whole report**, negative pairs deterministically repel each other even if there is shareable information (a)
- When using **random text augmentation**, an image can stochastically attract various positive texts while the chances to repel the negative text of the same pathology can decreased due to stochastic property (b)



(a) Use of whole report leads deterministic agreement between image and text.

(b) Random text augmentation lets a image stochastically match to sub-sampled reports.

Methods – Loss for relaxed image-text agreement

- CLIP (InfoNCE loss)

$$l_i^{v \to u} = -\log \frac{\exp\left(\mathbf{sim}(v_i, u_i)/\tau\right)}{\sum_{k=1}^{N} \exp\left(\mathbf{sim}(v_i, u_k)/\tau\right)}$$

- *v, u* are normalized vector from image and text encoders
- (v_i, u_i) is a positive pair, **sim** is a function to calculate similarity between vectors
- τ means learnable temperature, *N* is the mini-batch size
- $l_i^{v \to u}$ means the InfoNCE loss from image to text

Methods – Loss for relaxed image-text agreement

- In the medical domain a report and image for different patients are likely to have **shareable information** about pathologies
  - If naively using the InfoNCE of CLIP method, it's hard to share the information with other pairs
  - we cannot easily explore the additional positive sample which shares the same semantics without label information

$$\mathbf{sim}(v_i, u_j) = \begin{cases} \frac{1}{1+\exp(-\alpha(v_i \cdot u_j - t))} & , \text{if } i = j \text{ and } v_i \cdot u_j \geq t \\ v_i \cdot u_j / 2t & , \text{else if } i = j \text{ and } t > v_i \cdot u_j \geq 0 \\ v_i \cdot u_j & , \text{otherwise} \end{cases}$$

- $v_i \cdot u_j$ means cosine similarity between vectors
- α is slope coefficient of Sigmoid function
- $t\ (0 < t < 1)$ is threshold from which similarity level we regard a pair as "positive"

→ For a positive pair, this can make image-text agreement relaxed so that it makes room to agree with other semantics from different samples

Datasets

- **MIMIC-CXR** for fine-tuning, extract “Findings” or “Impression” section. If the section is not extracted by the rule-base method, we use the last paragraph of those samples
- **CheXpert :** validation set for model selection and the test set for evaluation
- **PadChest** : use test set to evaluate zero-shot classification, image-text alignment
- **Open-I chest X-ray dataset** : use for evaluating image-text similarity and the labelled one for zero-shot classification

		# of Images	# of Reports	# of classes
Fine-tuning	MIMIC-CXR [13]	377,110	227,835	-
	CheXpert$_{valid}$ [12]	234	200	-
Evaluation	CheXpert$_{test}$ [12]	668	500	14
	Open-i [1]	7,470	3,955	14
	PadChest [3]	39,053	26,413	61
	VinDr-CXR [22]	3,000	-	20

Table 1. The statistics of used datasets.

Experiments

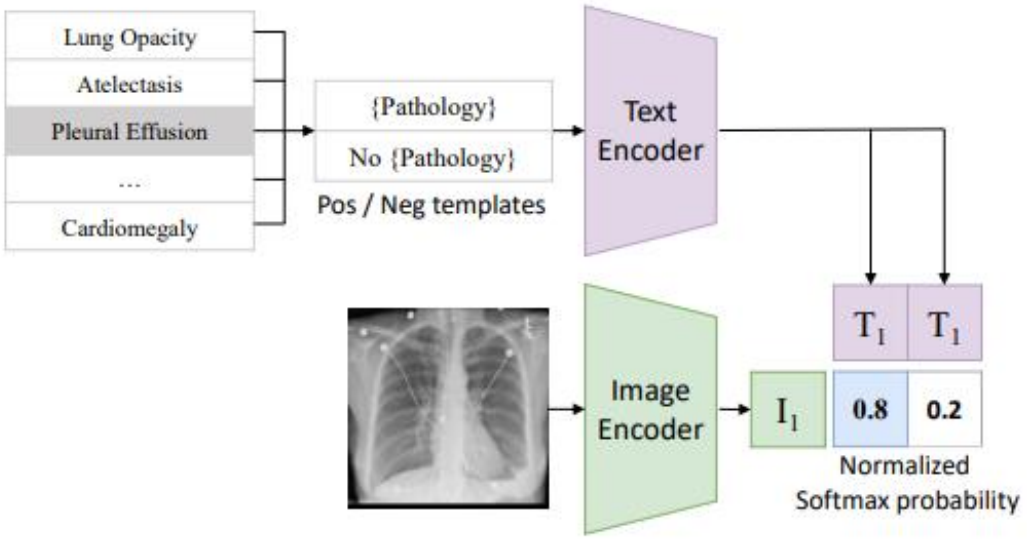
- Baseline
  - **CLIP** : multi-modal contrastive learning framework pre-trained on 400M image-text pairs. CLIP uses InfoNCE loss to maximize mutual information of latent vectors from image and text encoders
  - **SLIP** : multi-task learning framework that enhances the representation quality by combining image self-supervised learning with CLIP pre-training
  - **UniCLIP** : unified framework for visual-language pre-training that improves data efficiency by integrating inter and intra domain pairs' contrastive loss into a single universal space
- text encoder : BERT (base) model
- image encoder : ViT-B/16

Experiments – Implementation Details

- Augmentation : **random crop with large ratio** - reduce the intensity of augmentation to avoid information loss (CLIP - resized crop, SLIP, UniCLIP – color jitter, random crop)
- random text augmentation – **3** sentences are sampled from each report
- relaxation loss threshold is **0.5**, and the slope coefficient is fixed at **10**
- set the image resolution to **224 x 224**

Zero-shot classification

- Evaluation



(b) Zero-shot pathology classification

- construct positive and negative prompts such as “Atelectasis” and “No Atelectasis” for each pathology
- calculate the softmax probability of each pathology through the cosine similarity between the target image embedding and the embedding of the positive/negative prompt

Image-Text alignment

- In medical report, each sentence generally contains different label information
- We calculate the cosine similarity between image embedding and report embedding for each X-ray-Report pair for report-level evaluation
- We also calculate the average similarity between the image and sentence embedding for all sentences in the corresponding report to measure sentence-level alignment

Results – Zero-shot classification

	CheXpert		Open-i		PadChest	VinDr-CXR
	avg. 5 AUC	avg. total AUC	avg. 5 AUC	avg. total AUC	avg. total AUC	avg. total AUC
CLIP	0.870 ± 0.006	0.771 ± 0.001	0.785 ± 0.030	0.691 ± 0.010	0.685 ± 0.026	0.801 ± 0.020
CLIP w/ ours	**0.885 ± 0.007**	**0.789 ± 0.017**	**0.808 ± 0.008**	**0.708 ± 0.013**	**0.725 ± 0.014**	**0.842 ± 0.007**
SLIP	0.831 ± 0.008	0.722 ± 0.008	0.760 ± 0.008	0.666 ± 0.004	0.684 ± 0.004	0.799 ± 0.018
SLIP w/ ours	**0.884 ± 0.009**	**0.779 ± 0.020**	**0.810 ± 0.016**	**0.719 ± 0.018**	**0.741 ± 0.005**	**0.831 ± 0.009**
UniCLIP	0.851 ± 0.008	0.741 ± 0.037	0.744 ± 0.032	0.643 ± 0.027	0.684 ± 0.009	0.759 ± 0.019
UniCLIP w/ ours	**0.888 ± 0.006**	**0.802 ± 0.018**	**0.788 ± 0.016**	**0.694 ± 0.006**	**0.717 ± 0.011**	**0.843 ± 0.016**
CheXzero	0.844 ± 0.008	0.733 ± 0.008	0.771 ± 0.006	0.684 ± 0.006	0.702 ± 0.011	0.771 ± 0.024
MedCLIP*	0.880	0.829	0.794	0.728	0.728	0.829

∗: use label information defined in CheXpert for training

Table 2. Zero-shot classification performance on three different CLIP-based frameworks on the four different datasets (mean ± std). Best performance is in **bold**. For the CheXpert and Open-i datasets, the average AUC of five CheXpert competition pathologies and a total of 13 pathologies are reported. For the PadChest and VinDr-CXR dataset, an average of 61 pathologies ($n > 100$) and 20 pathologies ($n > 10$) is reported.

Results – Zero-shot classification

- CheXpert

	mean AUC	Mean F1	Mean MCC
Radiologists (mean) [27]	-	0.619	0.530
CLIP	0.870	0.578	0.423
CLIP w/ ours	0.885	0.603	0.443
$\text{CLIP}_{\text{ensemble}}$	0.887	0.613	0.519
$\text{CLIP}_{\text{ensemble}}$ w/ ours	**0.893**	**0.614**	**0.525**
SLIP	0.831	0.578	0.423
SLIP w/ ours	0.884	0.608	0.424
$\text{SLIP}_{\text{ensemble}}$	0.845	0.564	0.456
$\text{SLIP}_{\text{ensemble}}$ w/ ours	**0.893**	**0.625**	**0.537**
UniCLIP	0.851	0.561	0.458
UniCLIP w/ ours	0.888	0.610	0.522
$\text{UniCLIP}_{\text{ensemble}}$	0.881	0.614	0.524
$\text{UniCLIP}_{\text{ensemble}}$ w/ ours	**0.900**	**0.623**	**0.544**
CheXzero	0.844	0.559	0.417
$\text{CheXzero}_{\text{ensemble}}$	0.857	0.576	0.480
MedCLIP*	0.880	0.625	0.537

*: use label information defined in CheXpert for training

Results – Image-text alignment

	Open-i		PadChest	
	sentence-level	report-level	sentence-level	report-level
CheXzero	0.319 ± 0.015	0.387 ± 0.015	0.232 ± 0.079	0.258 ± 0.081
CLIP	0.275 ± 0.053	0.407 ± 0.007	0.102 ± 0.099	0.135 ± 0.111
CLIP w/ ours	0.658 ± 0.005	0.693 ± 0.005	0.572 ± 0.021	0.579 ± 0.015
SLIP	0.163 ± 0.021	0.287 ± 0.031	0.007 ± 0.022	0.018 ± 0.014
SLIP w/ ours	0.434 ± 0.034	0.489 ± 0.035	0.512 ± 0.008	0.500 ± 0.009
UniCLIP	0.209 ± 0.013	0.318 ± 0.016	0.096 ± 0.010	0.118 ± 0.011
UniCLIP w/ ours	0.451 ± 0.038	0.543 ± 0.024	0.396 ± 0.060	0.418 ± 0.068
MedCLIP*	0.044	0.047	0.040	0.047

Table 4. Image-text similarity for Open-i and PadChest datasets (mean ± std). We report both sentence-level similarity and report-level similarity for each dataset.

Results – Ablation study

	Text aug	Relax loss	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	avg. 5 AUC	avg. total AUC
CLIP			0.828 ± 0.016	0.845 ± 0.012	0.895 ± 0.007	0.861 ± 0.016	0.878 ± 0.020	0.861 ± 0.002	0.787 ± 0.044
		✓	0.858 ± 0.009	0.815 ± 0.013	0.889 ± 0.023	0.875 ± 0.022	**0.890 ± 0.006**	0.865 ± 0.002	0.794 ± 0.045
	✓		**0.864 ± 0.004**	**0.867 ± 0.008**	0.904 ± 0.015	**0.903 ± 0.003**	0.877 ± 0.021	**0.883 ± 0.005**	0.820 ± 0.001
	✓	✓	0.859 ± 0.006	0.840 ± 0.016	**0.913 ± 0.010**	0.881 ± 0.009	0.869 ± 0.012	0.872 ± 0.005	**0.824 ± 0.003**
SLIP			0.760 ± 0.046	0.788 ± 0.008	0.877 ± 0.028	0.847 ± 0.021	**0.902 ± 0.010**	0.835 ± 0.010	0.791 ± 0.015
		✓	0.826 ± 0.010	0.828 ± 0.010	0.899 ± 0.006	0.873 ± 0.013	0.899 ± 0.010	0.865 ± 0.005	**0.828 ± 0.003**
	✓		0.830 ± 0.022	0.837 ± 0.021	0.864 ± 0.042	0.899 ± 0.008	0.876 ± 0.013	0.861 ± 0.004	0.800 ± 0.015
	✓	✓	**0.845 ± 0.014**	**0.855 ± 0.018**	**0.904 ± 0.014**	**0.905 ± 0.016**	0.868 ± 0.009	**0.875 ± 0.006**	0.819 ± 0.028
UniCLIP			0.774 ± 0.060	0.833 ± 0.012	0.869 ± 0.016	0.856 ± 0.033	0.882 ± 0.020	0.843 ± 0.010	0.750 ± 0.016
		✓	0.840 ± 0.009	0.818 ± 0.021	0.870 ± 0.012	0.869 ± 0.014	**0.897 ± 0.021**	0.859 ± 0.008	0.808 ± 0.023
	✓		0.860 ± 0.012	**0.862 ± 0.010**	0.900 ± 0.019	0.879 ± 0.005	0.887 ± 0.012	0.877 ± 0.009	0.790 ± 0.020
	✓	✓	**0.870 ± 0.007**	0.844 ± 0.007	**0.911 ± 0.002**	**0.886 ± 0.014**	0.882 ± 0.017	**0.878 ± 0.001**	**0.825 ± 0.028**

∗: use label information defined in CheXpert for training

Table 5. Ablation study of different components of our proposed method on the CheXpert validation set (mean ± std). We report the zero-shot classification performance on three different CLIP-based frameworks. Best performance is in **bold** and second best is in underline for each framework. For the CheXpert dataset, the AUC for each of the five competition pathologies, the average value of these five pathologies, and the total of 13 pathologies of this dataset are reported.

Results – Ablation study

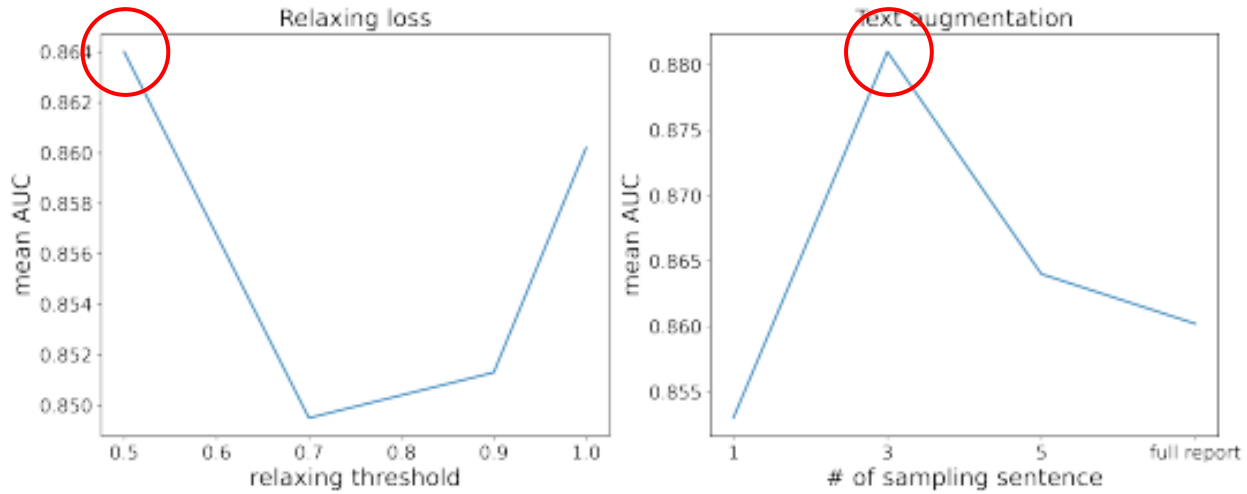


Figure 5. Effect of hyper-parameters. (Left) Relation between the relaxing threshold and mean AUC. (Right) Relation between the number of sampling sentences of text augmentation and mean AUC. Both evaluate the performance of zero-shot classification on the CheXpert validation set.

Conclusion

- In this paper, we propose a **simple yet effective fine-tuning strategy for medical image-text multi-modal learning**, which consists of two ways, **random sentence sampling, and loss relaxation**
- Our fine-tuning strategy improves zero-shot pathology classification performance in all three CLIP-based frameworks
- In addition, our proposed method significantly outperformed board-certified radiologists without additional label information

Collaborators

MIRL Medical Imaging Intelligent Reality Lab

Radiology

Joon Beom Seo, SangMin Lee[A,B]
Dong Hyun, Yang, Hyung Jin Won
Ho Sung Kim, Seung Chai Jung
Ji Eun Park, So Jung Lee
Jeong Hyun Lee, Gilsun Hong

Pathology

Hyunjeong Go, Gyuheon Choi
Gyungyub Gong, Dong Eun Song

Cardiology

Jaekwan Song, Jongmin Song
Young-Hak Kim

Anesthesiology

Sung-Hoon Kim, Eun Ho Lee

Neurology

Dong-Wha Kang, Chongsik Lee
Jaehong Lee, Sangbeom Jun
Misun Kwon, Beomjun Kim, Sun Kwon, Eun-Jae Lee

Surgery

Beom Seok Ko, JongHun Jeong
Songchuk Kim, Tae-Yon Sung

Gastroenterology

Jeongsik Byeon, Kang Mo Kim, Do-hoon Kim

Emergency Medicine

Dong-Woo Seo

Pulmonology and Critical Care Medicine

Yoen-mok Oh, Sei Won Lee, Jin-won Huh