MedCLIP : Contrastive Learning from Unpaired Medical Images and Text

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, Jimeng Sun

Medical Imaging & Intelligent Reality Lab.
Convergence Medicine/Radiology,
University of Ulsan College of Medicine
Asan Medical Center

Jihoon Jung

2023.03.16

Abtract

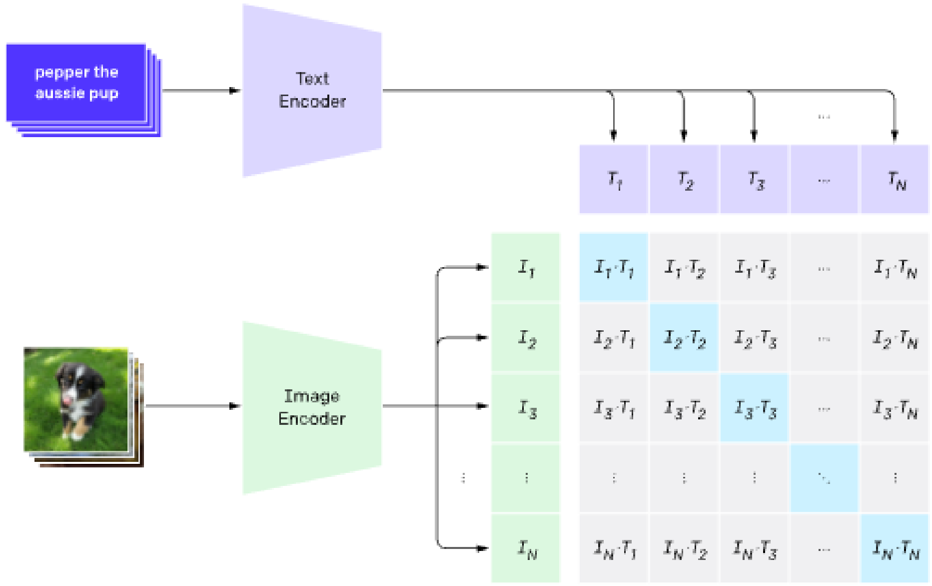
- Existing **vision-text** contrastive learning like **CLIP** (Radford et al., 2021) aims **to match the paired image and caption embeddings** while pushing others apart.
- Previous method limitation in medical domain
  1. medical image-text datasets are orders of magnitude below the general images and captions from the internet.
  2. **previous methods encounter many false negatives**, i.e., images and reports from separate patients probably carry the same semantics but are wrongly treated as negatives.
- Novelty
  1. we **decouple images and texts** for multimodal contrastive learning.
  2. We also propose to replace the **InfoNCE loss with semantic matching loss based on medical knowledge to eliminate false negatives** in contrastive learning.

Introduction

- The issues with adopting the CLIP model for the medical domain.
  1. CLIP's (Radford et al., 2021) **data-hungry nature** :

     CLIP is trained on a dataset of 400M image-text pairs collected from the internet
  2. **Specificity** of medical images and reports: compared to general domains :

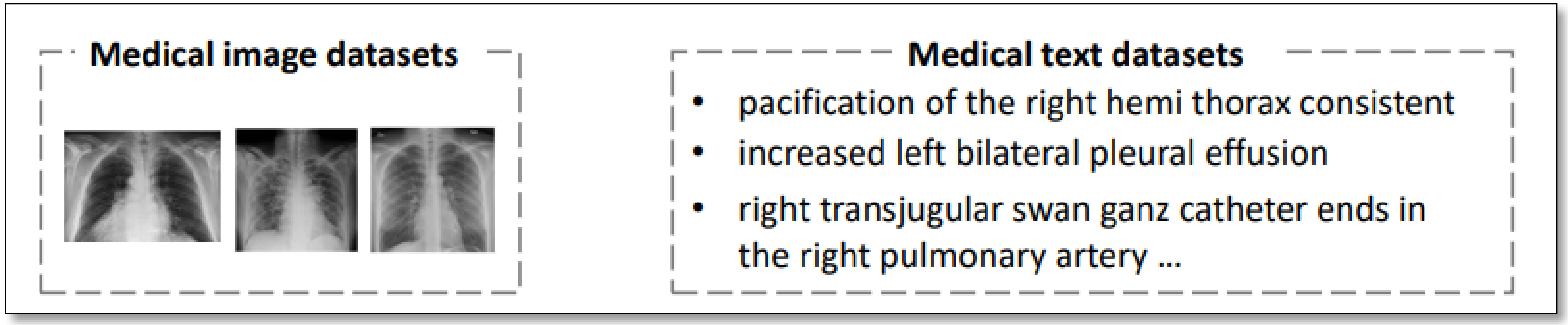
     the differences within medical domains are more **subtle and fine-grained**



Introduction

Challenges

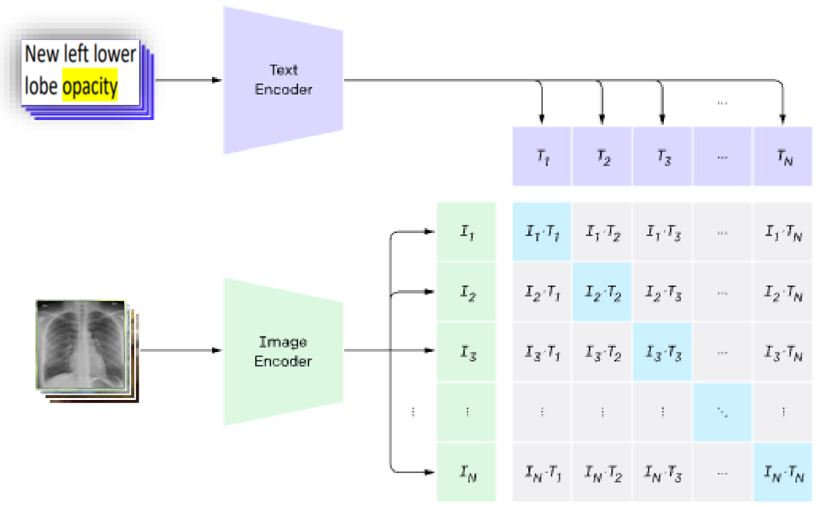
- **Limited usable data :**
  - ✓ Most medical image datasets **only provide the diagnostic labels** instead of the raw reports.
  - ✓ However, Previous methods need paired image and reports, leaving a vast number of medical **image-only** and **text-only datasets unused**.

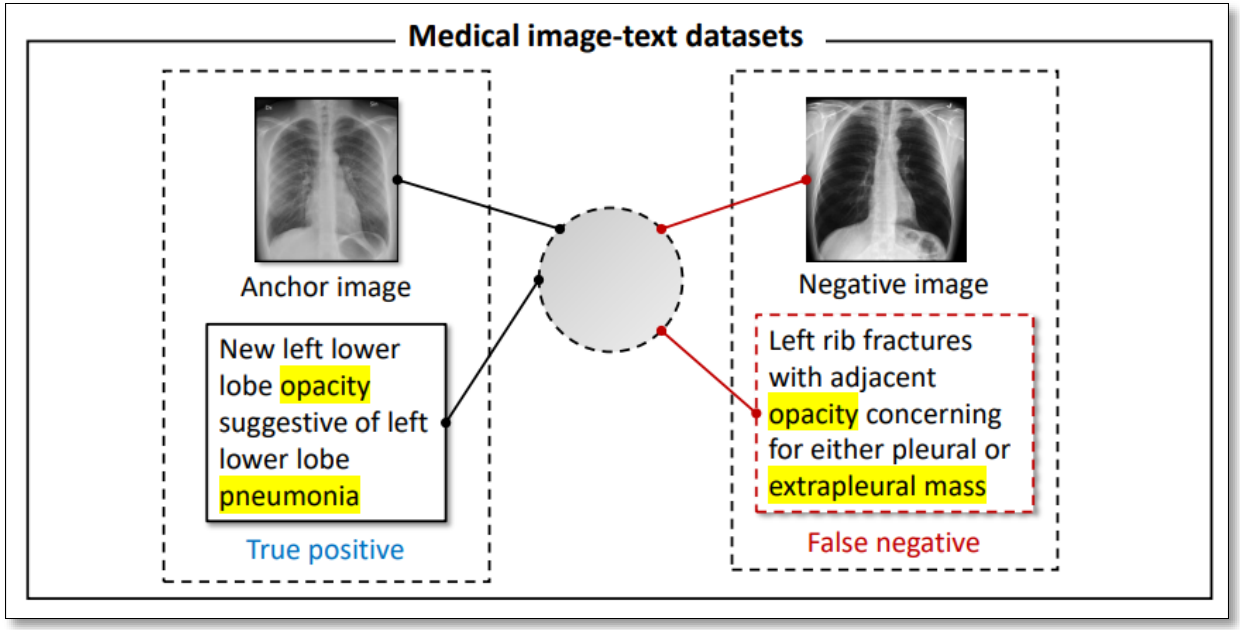


Introduction

Challenges

- **False negatives in contrastive learning :**
  - ✓ Previous methods try to push images and texts embeddings from different patients apart.
  - ✓ However, even though some reports do not belong to the target patient’s study, they can still describe the same symptoms and findings.
  - ✓ Simply treating the other reports as negative samples brings noise to the supervision and confuses the model.





Introduction

Contribution

- **Decoupling images and texts for contrastive learning :**
  - ✓ We extend the pre-training to cover **the massive unpaired images and texts datasets**, which scales the number of training data in a combinatorial manner.
  - ✓ It opens a new direction to expand multi-modal learning **based on medical knowledge.**
- **Eliminating false negatives via medical knowledge.**
  - ✓ We observe that images and reports from separate patients' studies may carry **the same semantics** but are falsely **treated as negatives** by previous methods.
  - ✓ Hence, we design a **soft semantic matching loss** that uses **the medical semantic similarity** between each image and report as the supervision signal.
  - ✓ This approach equips the model with the ability to **capture the subtle yet crucial medical meanings**.

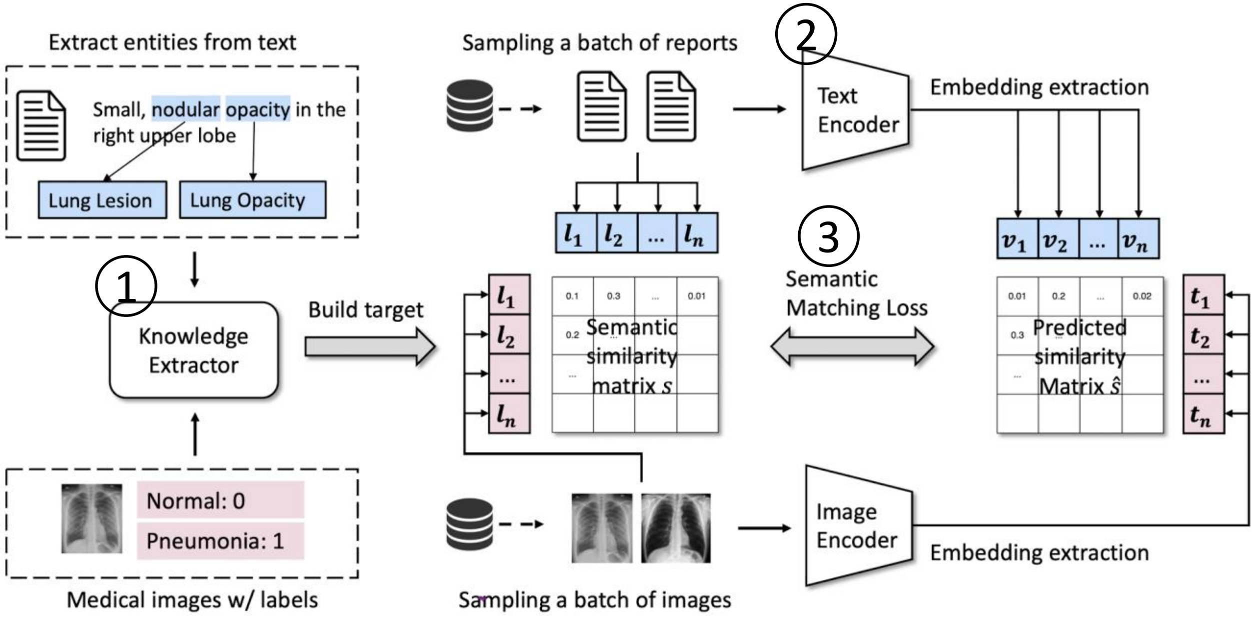
Introduction

Contribution

- **Decoupling images and texts for contrastive learning :**
  - ✓ We extend the pre-training to cover **the massive unpaired images and texts datasets**, which scales the number of training data in a combinatorial manner.
  - ✓ It opens a new direction to expand multi-modal learning **based on medical knowledge.**
- **Eliminating false negatives via medical knowledge.**
  - ✓ We observe that images and reports from separate patients' studies may carry **the same semantics** but are falsely **treated as negatives** by previous methods.
  - ✓ Hence, we design a **soft semantic matching loss** that uses **the medical semantic similarity** between each image and report as the supervision signal.
  - ✓ This approach equips the model with the ability to **capture the subtle yet crucial medical meanings**.

Method

- In this section, we present the technical details of MedCLIP following the flow in Fig. 3.
- MedCLIP consists of 3 components
  1. vision and text encoders that extracts embeddings
  2. knowledge extraction that builds the semantic similarity matrix
  3. semantic matching loss that trains the whole model.



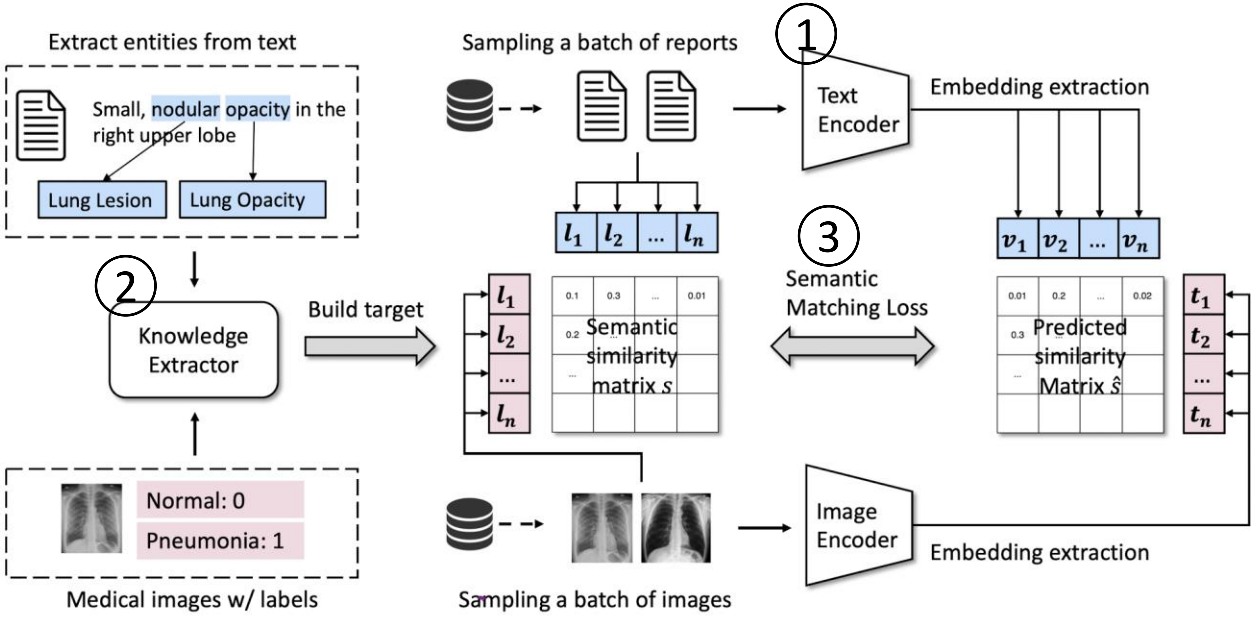
Method

Vision and Text Encoder

- MedCLIP consists of one visual encoder and one text encoder.
- **Vision Encoder.**
  a) We encode images into embeddings $\mathbf{v} \in R^{\mathbf{D}}$ using a vision encoder $\mathbf{E}_{img}$.
  b) A projection head then maps raw embeddings to $\mathbf{v}_p \in R^{\mathbf{P}}$.

$$\mathbf{v} = E_{img}(\mathbf{x}_{img})$$
$$\mathbf{v}_p = f_v(\mathbf{v})$$

- where $\mathbf{f}_{\mathbf{v}}$ is the projection head of the vision encoder.

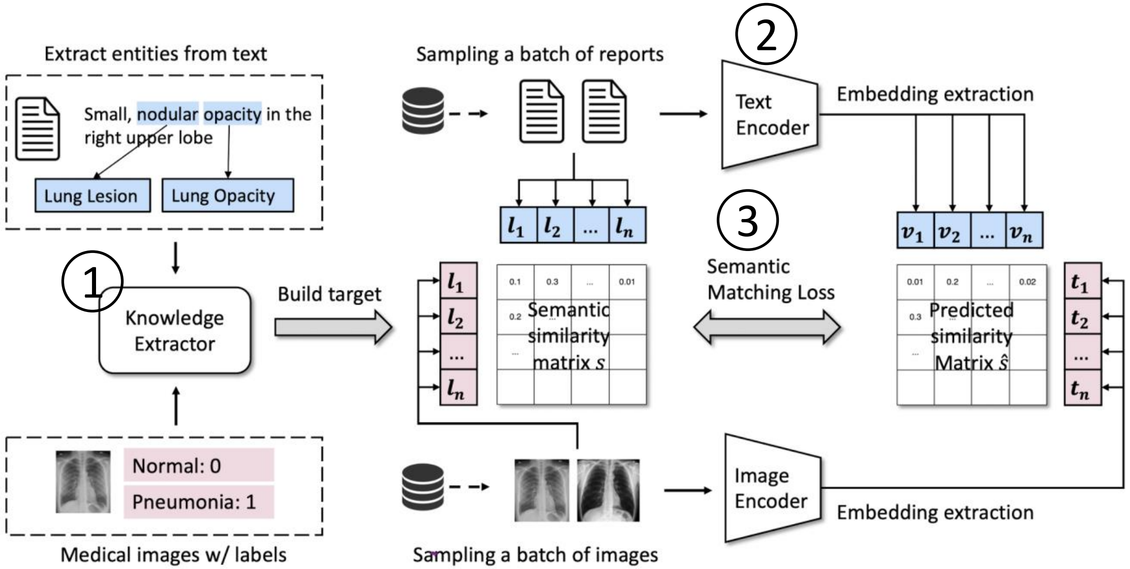


Method

Vision and Text Encoder

- MedCLIP consists of one visual encoder and one text encoder.
- **Text Encoder.**
  a) We create clinically meaningful text embeddings $\mathbf{t} \in R^{M}$ by a text encoder.
  b) We project them to $\mathbf{t}_p \in R^{P}$ as

$$\mathbf{t} = E_{txt}(\mathbf{x}_{txt})$$
$$\mathbf{t}_p = f_t(\mathbf{t})$$



- where $\mathbf{f_t}$ is the projection head and $\mathbf{E}_{txt}$ denotes the text encoder.

Method

Decouple Image-Text Pairs with Medical Knowledge Extractor

- Paired medical image text datasets are orders of magnitude less than the general paired image text (e.g., from the internet)
- To enhance medical multi-modal learning, we want to make the full use of all existing medical **image-text**, **image-only**, and **text-only** datasets.
- Suppose we have **n** paired image-text samples, **m** labeled images, and **h** medical sentences.
  Image sets : n+m
  text sets : n+h
- Knowledge extractor for **image-text** & **text-only** data : Use MetaMap with UMLS(Unified Medical Language System)
  https://gweissman.github.io/post/using-metamap-with-python-to-access-the-umls-metathesaurus-a-quick-start-guide/
- Knowledge extractor for **only-image data (w / labels)** : Match 14 keywords to label
- We build **multi-hot vectors** from the extracted entities for images and texts, as $\mathbf{l}_{img}$ and $\mathbf{l}_{txt}$, respectively.

Method

Table 5: 14 main finding types used in this paper.

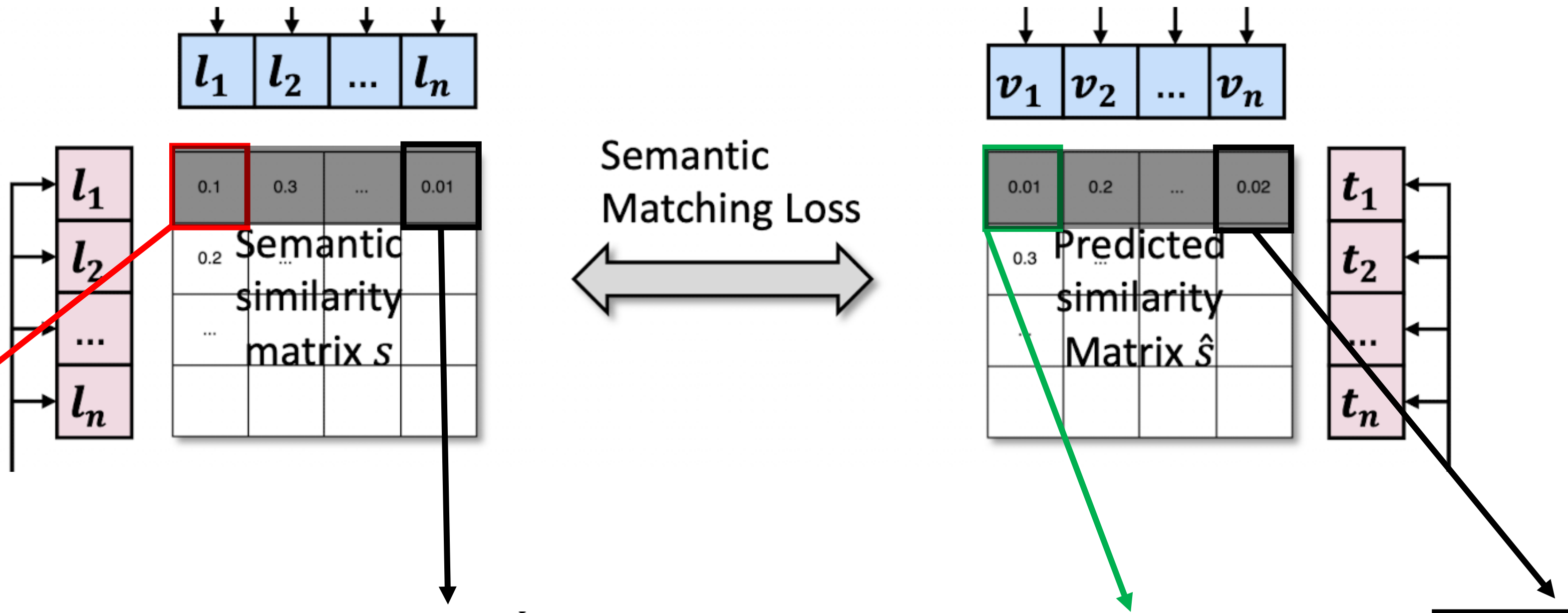
Finding types
No Finding
Enlarged Cardiomediastinum
Cardiomegaly
Lung Opacity
Lung Lesion
Edema
Consolidation
Pneumonia
Atelectasis
Pneumothorax
Pleural Effusion
Pleural Other
Fracture
Support Devices

Method

Semantic Matching Loss

$$\mathcal{L}^{v\to l} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} y_{ij} \log \hat{y}_{ij}.$$

$$\mathcal{L} = \frac{\mathcal{L}^{v\to l} + \mathcal{L}^{l\to v}}{2}$$



$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}.$$

$$y_{ij}^{v\to t} = \frac{\exp s_{ij}}{\sum_{j=1}^{N_{batch}} \exp s_{ij}}. \qquad y_{ji}^{t\to v} = \frac{\exp s_{ji}}{\sum_{i=1}^{N_{batch}} \exp s_{ji}}.$$

$$\hat{s}_{ij} = \tilde{\mathbf{v}}_i^{\top} \cdot \tilde{\mathbf{t}}_j,$$

$$\hat{y}_{ij} = \frac{\exp \hat{s}_{ij}/\tau}{\sum_{i=1}^{N_{batch}} \exp \hat{s}_{ij}/\tau}.$$

Dataset

Pretrain	# Images	# Reports	# Classes
MIMIC-CXR	377,111	201,063	-
CheXpert	223,415	-	14
Evaluation	**# Train (Pos.%)**	**# Test (Pos.%)**	**# Classes**
CheXpert-5x200	1,000 (-)	1,000 (-)	5
MIMIC-5x200	1,000 (-)	1,000 (-)	5
COVID	2,162 (19%)	3,000 (49%)	2
RSNA	8,486 (50%)	3,538 (50%)	2

Table 5: 14 main finding types used in this paper.

Finding types
No Finding
Enlarged Cardiomediastinum
Cardiomegaly
Lung Opacity
Lung Lesion
Edema
Consolidation
Pneumonia
Atelectasis
Pneumothorax
Pleural Effusion
Pleural Other
Fracture
Support Devices

Experiments

- **Implementation details**
  1. MedCLIP text encoder : BioClinicalBERT
  2. MedCLIP image encoder : Swin-transformer
  3. MedCLIP encoder output dimension : 512
  4. MIMIC CXR : combine the “Findings” and “Impression” sections of reports then split them into sentences.

Experiments

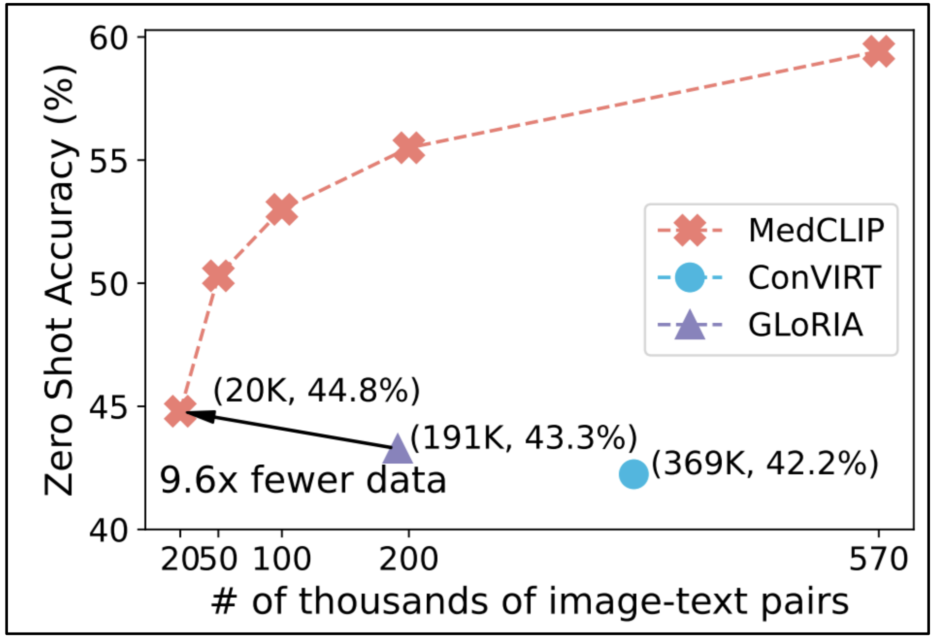
- **Zero-shot classification**

ACC(STD)	CheXpert-5x200	MIMIC-5x200	COVID	RSNA
CLIP	0.2016(0.01)	0.1918(0.01)	0.5069(0.03)	0.4989(0.01)
CLIP_{ENS}	0.2036(0.01)	0.2254(0.01)	0.5090(<0.01)	0.5055(0.01)
ConVIRT	0.4188(0.01)	0.4018(0.01)	0.5184(0.01)	0.4731(0.05)
$\text{ConVIRT}_{\text{ENS}}$	0.4224(0.02)	0.4010(0.02)	0.6647(0.05)	0.4647(0.08)
GLoRIA	0.4328(0.01)	0.3306(0.01)	0.7090(0.04)	0.5808(0.08)
$\text{GLoRIA}_{\text{ENS}}$	0.4210(0.03)	0.3382(0.01)	0.5702(0.06)	0.4752(0.06)
MedCLIP-ResNet	0.5476(0.01)	0.5022(0.02)	**0.8472(<0.01)**	0.7418(<0.01)
$\text{MedCLIP-ResNet}_{\text{ENS}}$	0.5712(<0.01)	**0.5430(<0.01)**	0.8369(<0.01)	0.7584(<0.01)
MedCLIP-ViT	0.5942(<0.01)	0.5006(<0.01)	0.8013(<0.01)	0.7447(0.01)
$\text{MedCLIP-ViT}_{\text{ENS}}$	**0.5942(<0.01)**	0.5024(<0.01)	0.7943(<0.01)	**0.7682(<0.01)**

1. Baseline models use traditional contrastive learning. So, they generate false negatives, which aggravate ensemble model
2. Interestingly, MedCLIP yields over 0.8 ACC on COVID data while **there is no COVID-19 positive image** available during the course of **pre-training**.
3. This result demonstrates that contrastive pre-training of MedCLIP provides it with the **transferability to out-of-domain classes**.

Experiments

- **Pre-training Data Efficiency**



Experiments

- **Fine-tune for Classification**

ACC	CheXpert -5x200	MIMIC -5x200	COVID	RSNA
Random	0.2500	0.2220	0.5056	0.6421
ImageNet	0.3200	0.2830	0.6020	0.7560
CLIP	0.3020	0.2780	0.5866	0.7303
ConVIRT	0.4770	0.4040	0.6983	0.7846
GLoRIA	0.5370	0.3590	0.7623	0.7981
MedCLIP	**0.5960**	**0.5650**	**0.7890**	**0.8075**

< Supervised manner >

ACC(STD)	CheXpert-5x200	MIMIC-5x200	COVID	RSNA
CLIP	0.2016(0.01)	0.1918(0.01)	0.5069(0.03)	0.4989(0.01)
CLIP_{ENS}	0.2036(0.01)	0.2254(0.01)	0.5090(<0.01)	0.5055(0.01)
ConVIRT	0.4188(0.01)	0.4018(0.01)	0.5184(0.01)	0.4731(0.05)
$\text{ConVIRT}_{\text{ENS}}$	0.4224(0.02)	0.4010(0.02)	0.6647(0.05)	0.4647(0.08)
GLoRIA	0.4328(0.01)	0.3306(0.01)	0.7090(0.04)	0.5808(0.08)
$\text{GLoRIA}_{\text{ENS}}$	0.4210(0.03)	0.3382(0.01)	0.5702(0.06)	0.4752(0.06)
MedCLIP-ResNet	0.5476(0.01)	0.5022(0.02)	**0.8472(<0.01)**	0.7418(<0.01)
$\text{MedCLIP-ResNet}_{\text{ENS}}$	0.5712(<0.01)	**0.5430(<0.01)**	0.8369(<0.01)	0.7584(<0.01)
MedCLIP-ViT	0.5942(<0.01)	0.5006(<0.01)	0.8013(<0.01)	0.7447(0.01)
$\text{MedCLIP-ViT}_{\text{ENS}}$	**0.5942(<0.01)**	0.5024(<0.01)	0.7943(<0.01)	**0.7682(<0.01)**

< Unsupervised manner >

1. we surprisingly find that MedCLIP makes **zero-shot prediction comparable** with **supervised learning** models when contrasting Table 2 to Table 1.

Experiments

- **Image-Text Retrieval :**

  We choose **CheXpert-5x200 to evaluate the semantic richness** of the learned representations.

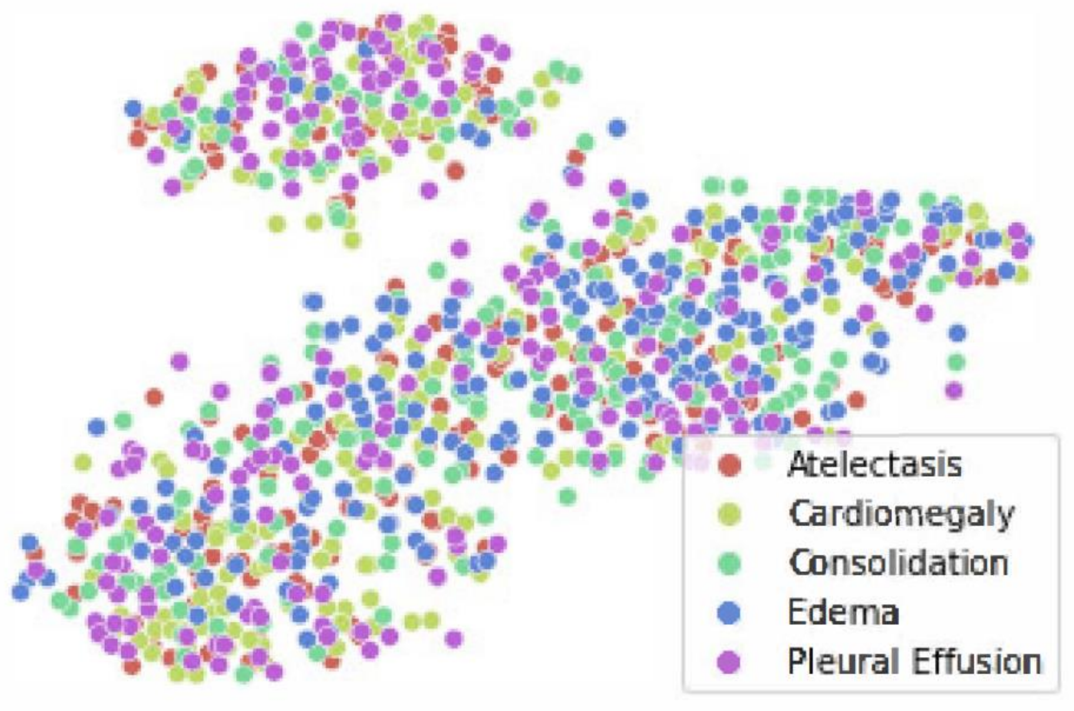
  CheXpert-5x200 do not have **report data** publicly available, we used **MIMIC-CXR dataset** to come up with reports/sentences.

  We sampled **200 sentences** for **each of the 5 classes** as present in CheXpert5x200 dataset.

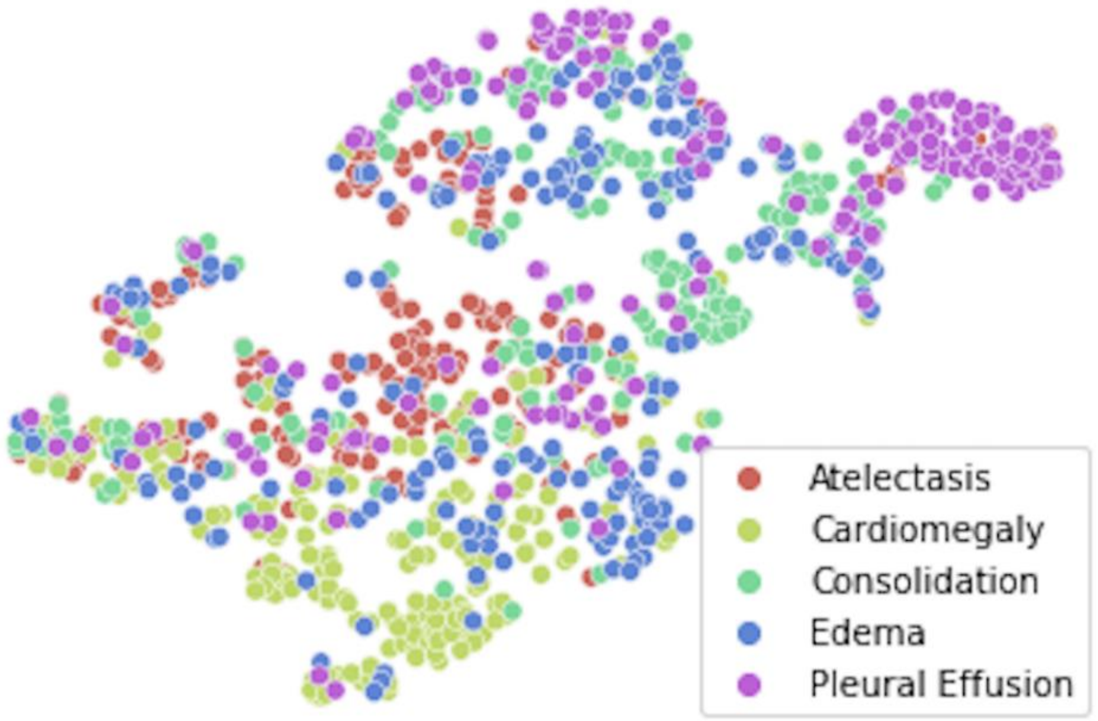
Model	P@1	P@2	P@5	P@10
CLIP	0.21	0.20	0.20	0.19
ConVIRT	0.20	0.20	0.20	0.21
GLoRIA	**0.47**	0.47	0.46	0.46
MedCLIP	0.45	**0.49**	**0.48**	**0.50**

Experiments

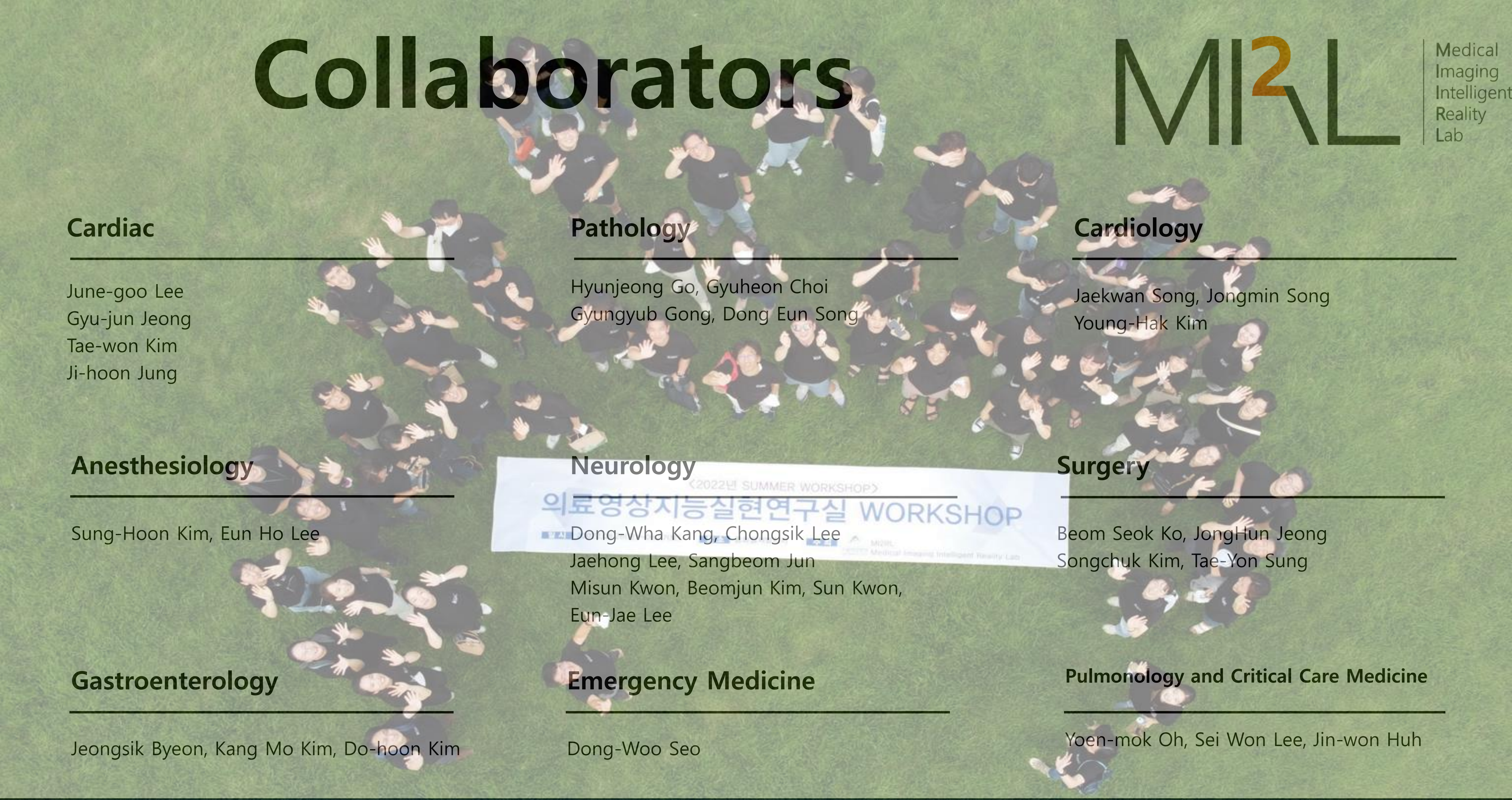
- **Embedding Visualization**



(a) CLIP



(b) MedCLIP



Collaborators

Medical
Imaging
Intelligent
Reality
Lab

Cardiac

June-goo Lee
Gyu-jun Jeong
Tae-won Kim
Ji-hoon Jung

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