MedCLIP: Contrastive Learning from Unpaired Medical Images and Text

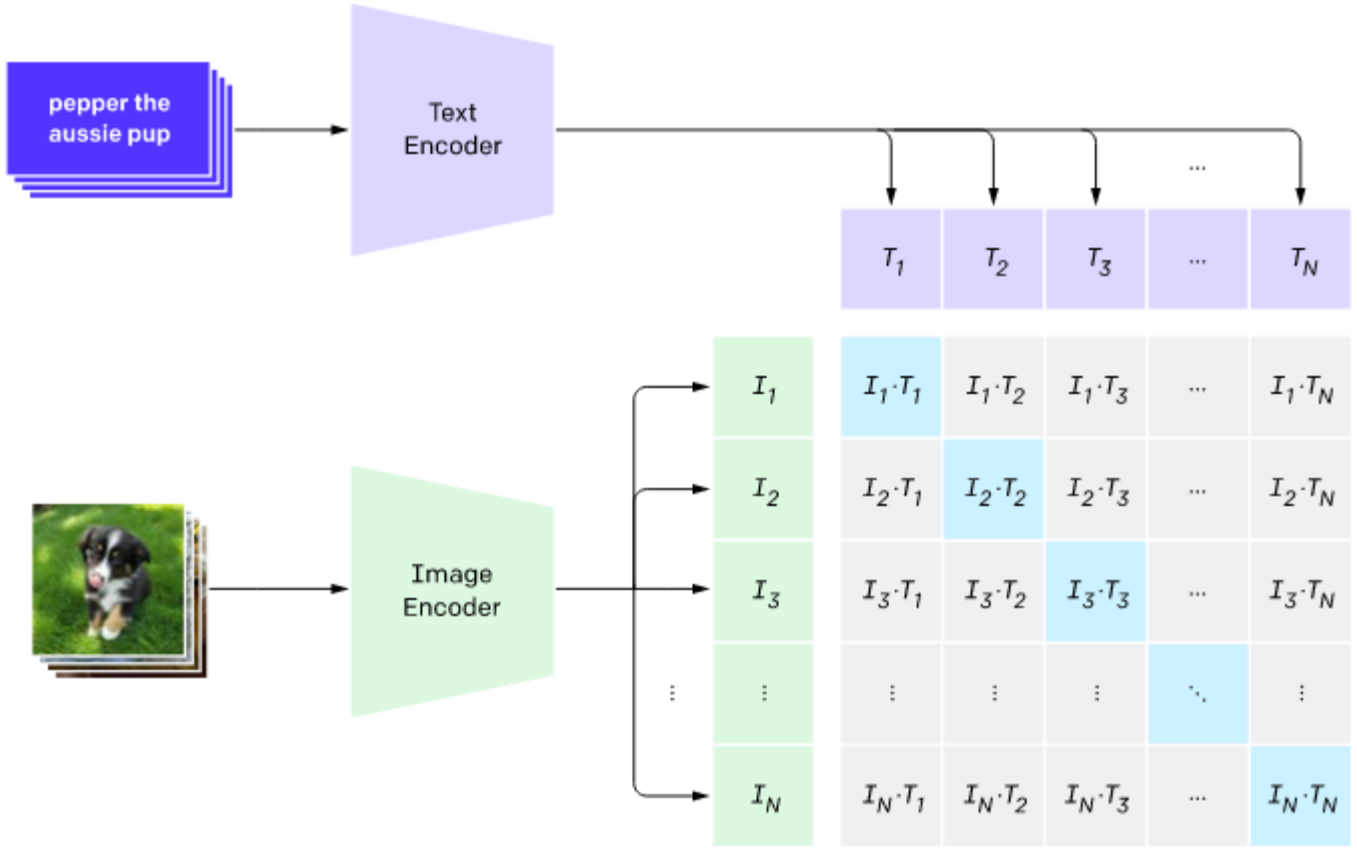
Presenter: Zihao Wei

2023/11/07

Wang, Zifeng, et al. "Medclip: Contrastive learning from unpaired medical images and text." *arXiv preprint arXiv:2210.10163* (2022).

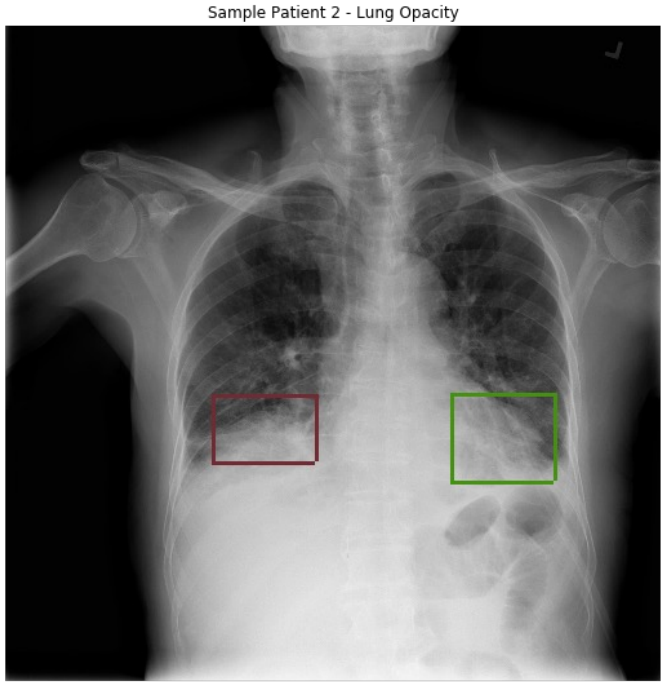
Quick Recap: CLIP

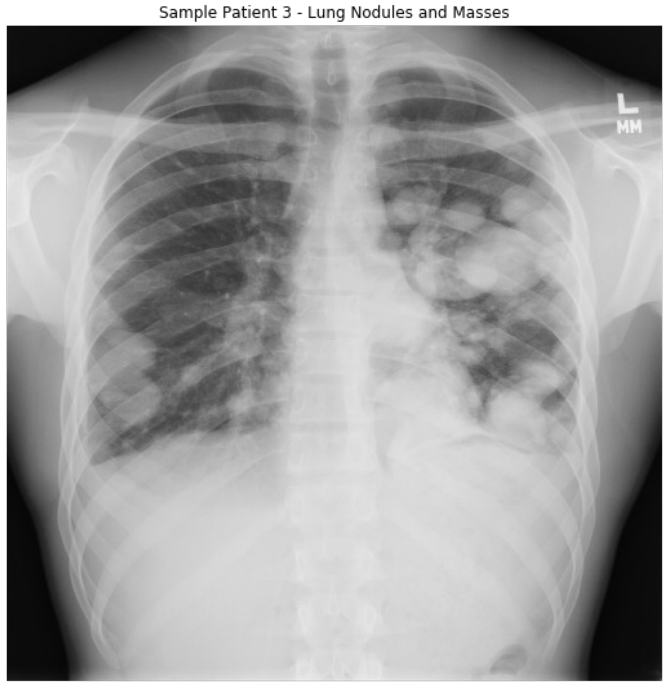
CLIP performs contrastive learning which aims to match the paired image and caption embeddings while pushing others apart by optimizing an InfoNCE loss.



Why naïve CLIP method doesn’t work in medical area?

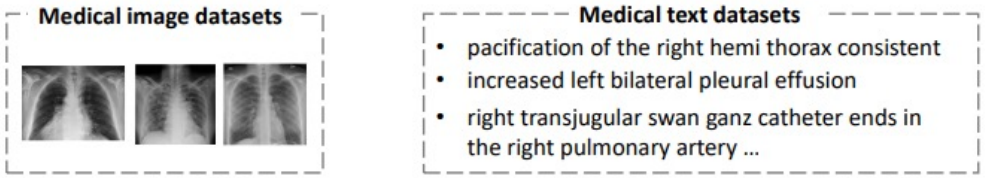
- Data insufficiency;
  - 400M vs 20K
- Medical domain is more subtle and fine-grained;
  - “Cat & dog” vs “pneumonia & consolidation ”



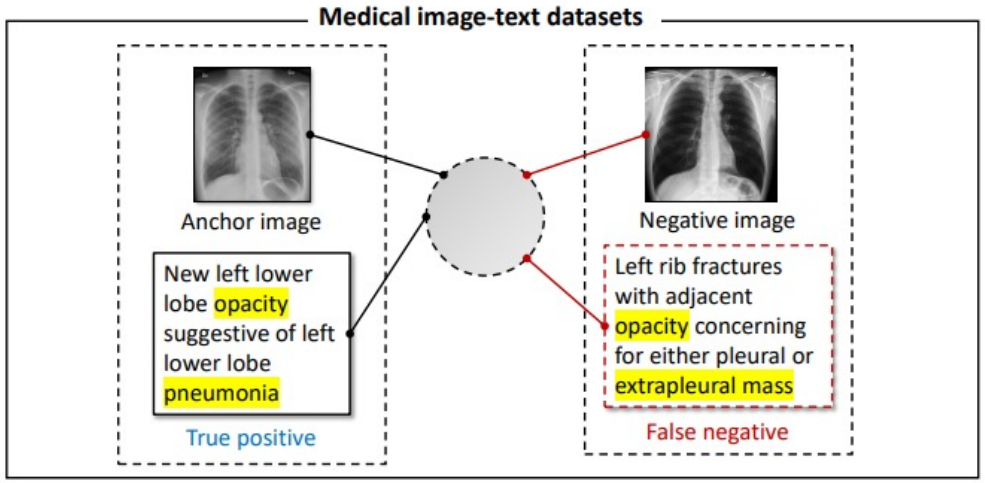


What are the problem that the previous works have?

- Data still not used sufficiently;



- False negatives in contrastive learning.



MedCLIP - overview

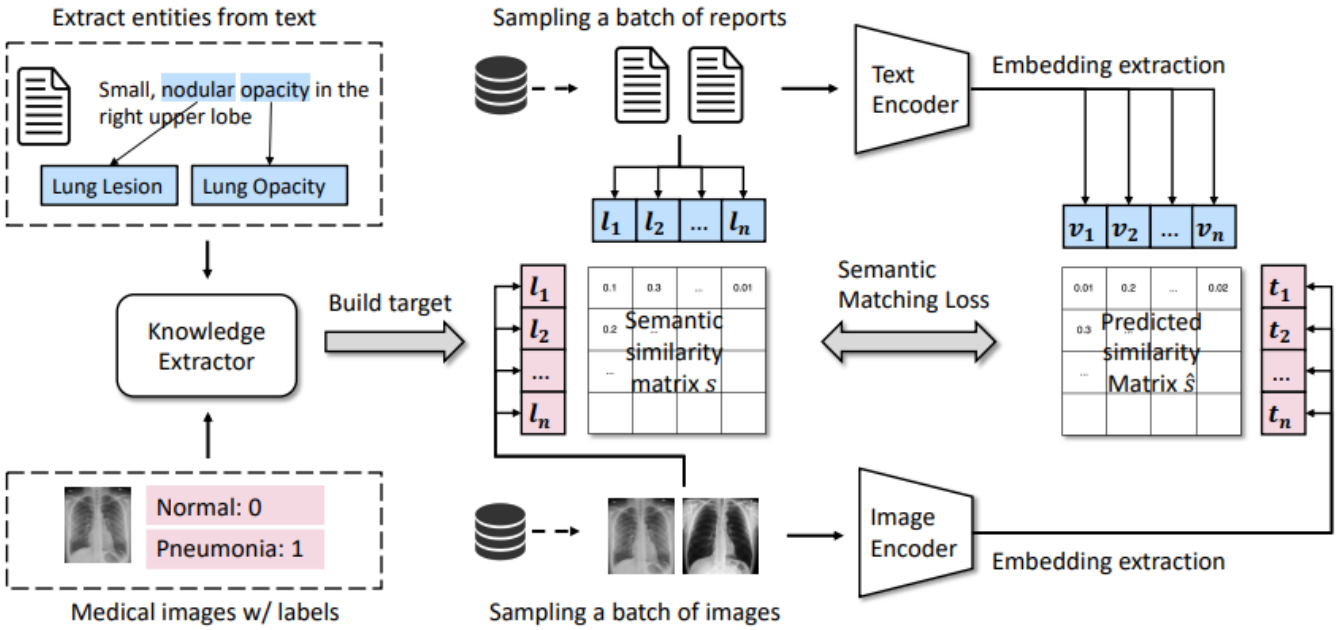


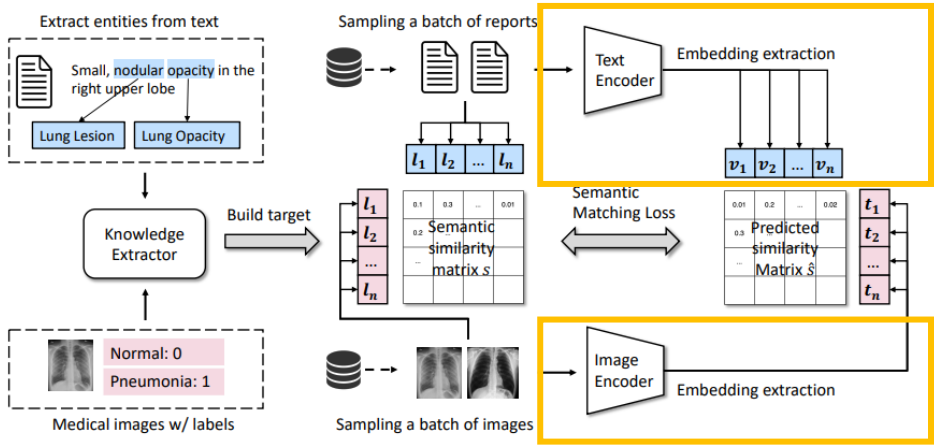
Figure 3: The workflow of MedCLIP. The knowledge extraction module extracts medical entities from raw medical reports. Then, a semantic similarity matrix is built by comparing medical entities (from text) and raw labels (from images), which enables pairing arbitrary two separately sampled images and texts. The extracted image and text embeddings are paired to match the semantic similarity matrix.

MedCLIP – Vision and Text Encoder

- The encoder part of the MedCLIP is like CLIP;
  - Swin & RN50 as Vision encoder
  - BioClinicalBERT as Text encoder

- Note a projection head is mapping vision feature and text feature into same dimension

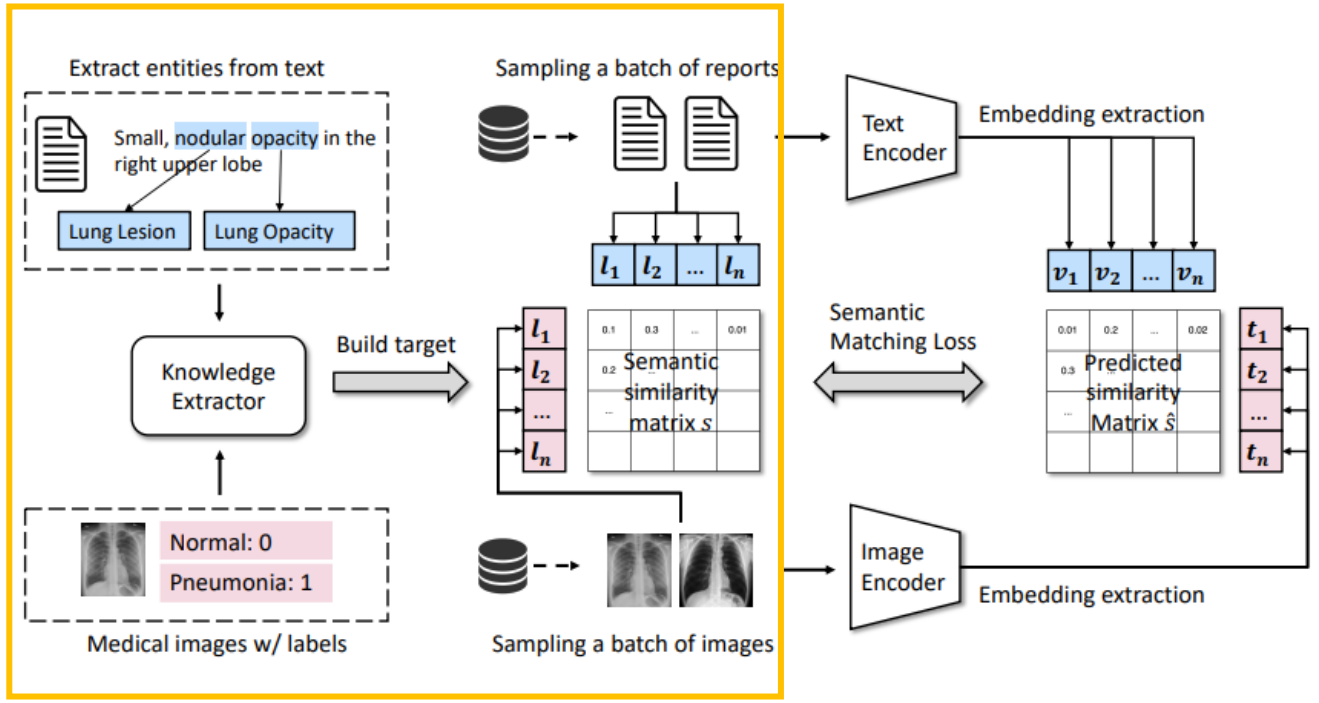
$$\mathbf{v} = E_{img}(\mathbf{x}_{img})$$
$$\mathbf{v}_p = f_v(\mathbf{v})$$



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

MedCLIP – Decoupling

- Target at using all available data.
  - Image-text: n
  - Image-only: m
  - Text-only: h
  - n × n -> (n+m) ×(n+h)
- Construct semantic similarity matrix utilizing multiple-hot vectors.
  - Knowledge driven by medical entity recognition and labels.



MedCLIP – Semantic Matching Loss

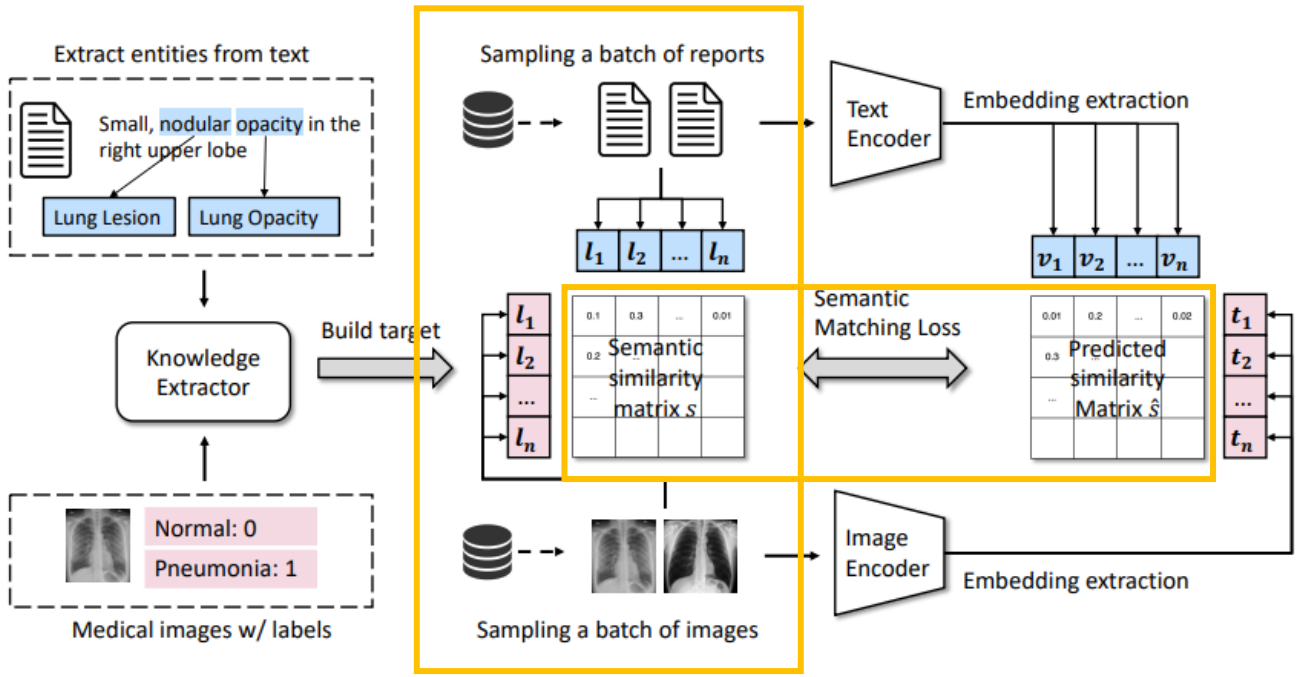
- Sample a batch of images and text {x_images} and {x_text} with size n.
- Calculate vectors {I_img} and {I_txt} using previous step.
- Get predicted embeddings {v} and {t}
- Build Ground truth matrix:

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

- Predicted matrix:

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

- Target: Cross entropy with the normalized soft targets



$$y_{ij}^{v \to t} = \frac{\exp s_{ij}}{\sum_{j=1}^{N_{batch}} \exp s_{ij}}. \qquad \hat{y}_{ij} = \frac{\exp \hat{s}_{ij}/\tau}{\sum_{i=1}^{N_{batch}} \exp \hat{s}_{ij}/\tau}. \qquad \mathcal{L}^{v \to l} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} y_{ij} \log \hat{y}_{ij}. \qquad \mathcal{L} = \frac{\mathcal{L}^{v \to l} + \mathcal{L}^{l \to v}}{2}$$

Experiment – Setup

- Datasets
  - CheXpert
  - MIMIC-CXR
  - COVID
  - RSNA Pneumonia
- Baseline
  - Random RN50
  - Imagenet RN50
  - CLIP
  - ConVIRT
  - GLoRIA

Table 3: The statistics of used datasets. Pos.%: positive sample ratio.

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

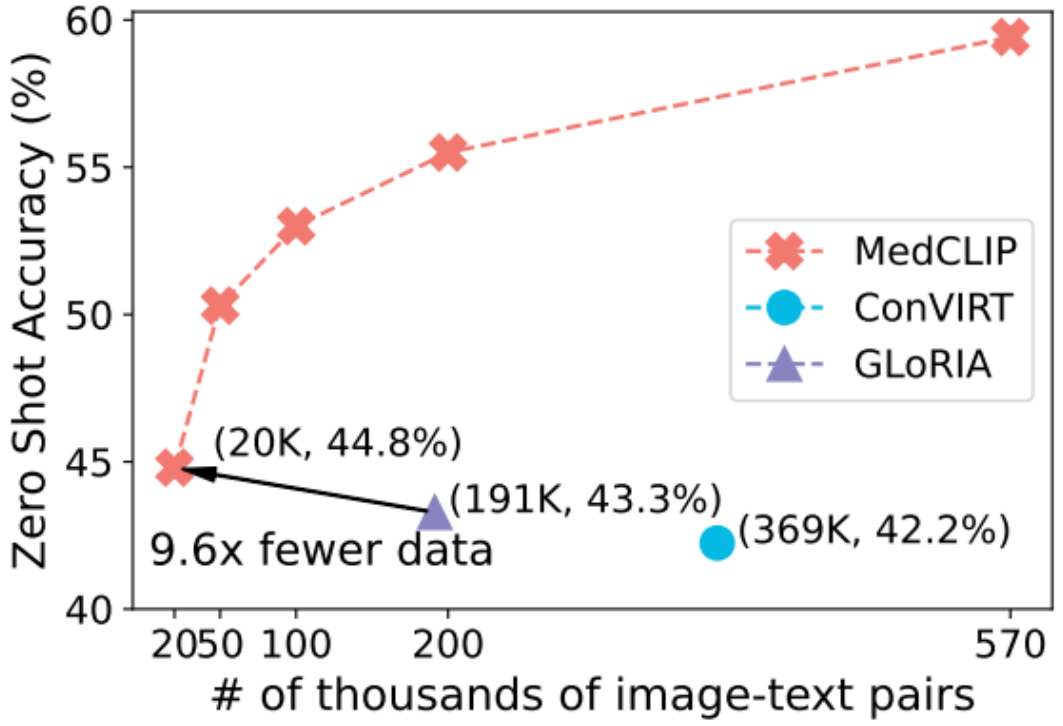
Experiment – Zeroshot classfication

- MedCLIP generally have better overall performance.
- MedCLIP can benefit from prompt ensemble to yield better performance.
- MedCLIP could transfer out-of-domain classes.

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)**

Experiment – Data efficiency

- Efficient training with 20K data.
- Not saturation with given more data.



Experiment – Image Classifications

- Mistake in paper: Linear probing instead of finetuning.

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**

Experiment – Image Text Retrieval

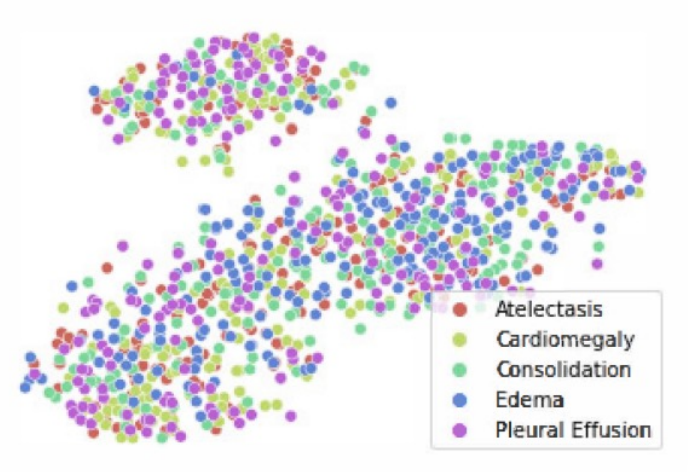
- Higher K, given better Gain
- Efficiently provide semantic information to retrieve texts.

Table 4: Results of Image-Text retrieval tasks on CheXpert5x200 dataset. We take the Precision@{1,2,5,10} to measure the performance of various models in this task. Best within the data are in bold.

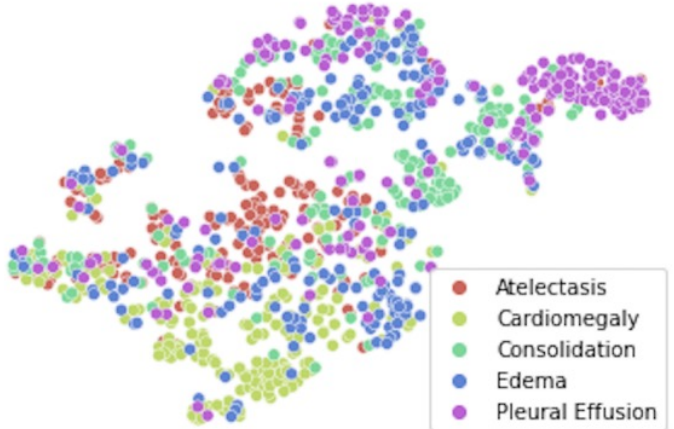
Model	P@1	P@2	P@5	P@10
CLIP	0.21	0.20	0.20	0.19
ConVIRT	0.20	0.20	0.20	0.21
GLoRIA	**0.47**	0.47	0.46	0.46
MedCLIP	0.45	**0.49**	**0.48**	**0.50**

Experiment – Visualization

- MedCLIP give better clustered representation.
- MedCLIP detect clusters by the lesion types.



(a) CLIP



(b) MedCLIP

Figure 4: Embeddings visualization of CheXpert5x200 images by CLIP and MedCLIP. Dimension reduced by t-SNE.

Discussion & Limitations

- Hand craft multiple hot representation has the difficulties in generalization.
- Unfair comparison with CLIP, which should be retrained with data from the same domain.

Quiz

- What are the main problems that MedCLIP solved?
- How does MedCLIP construct the semantic similarity matrix?