MedCLIP: Contrastive Learning from Unpaired Medical Images and Text

Zifeng Wang[1], **Zhenbang Wu**[1], **Dinesh Agarwal**[1,2], **Jimeng Sun**[1]
[1]University of Illinois Urbana-Champaign
[2]Adobe
{zifengw2, zw12, jimeng}@illinois.edu, diagarwa@adobe.com

Abstract

Existing vision-text contrastive learning like CLIP (Radford et al., 2021) aims to match the paired image and caption embeddings while pushing others apart, which improves representation transferability and supports zero-shot prediction. However, medical image-text datasets are orders of magnitude below the general images and captions from the internet. Moreover, 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. In this paper, we decouple images and texts for multimodal contrastive learning thus scaling the usable training data in a combinatorial magnitude with low cost. We also propose to replace the InfoNCE loss with semantic matching loss based on medical knowledge to eliminate false negatives in contrastive learning. We prove that MedCLIP is a simple yet effective framework: it outperforms state-of-the-art methods on zero-shot prediction, supervised classification, and image-text retrieval. Surprisingly, we observe that with only 20K pre-training data, MedCLIP wins over the state-of-the-art method (using $\approx$ 200K data).

1 Introduction

Medical images such as X-rays, CTs, and MRIs are commonly used to diagnose, monitor, or treat medical conditions in clinical practice (FDA, 2022). With the rapid growth of medical images and the corresponding reports data, researchers have developed various deep learning models to support clinical decision making (Çallı et al., 2021).

Recently, large-scale image-text pre-training, e.g., CLIP (Radford et al., 2021), has achieved considerable successes in computer vision and natural language processing domains. CLIP is trained to predict the correct matching of a batch of images and text training examples. The joint-training of image and text representations on large-scale image-text pairs generates transferable representations and supports flexible downstream tasks. Inspired by success of CLIP, we believe the knowledge jointly learned from medical images and reports should be helpful for downstream clinical tasks.

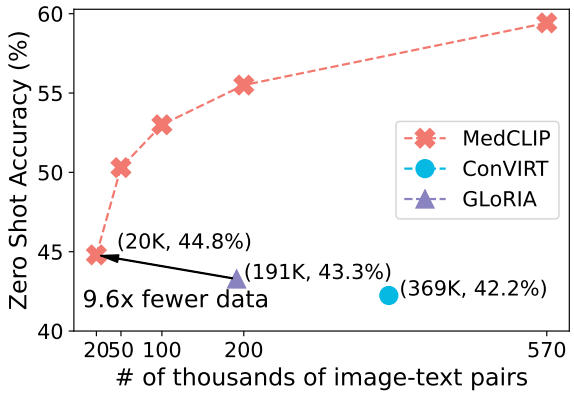


Figure 1: Zero-shot performance of MedCLIP, ConVIRT (Zhang et al., 2020), GLoRIA (Huang et al., 2021) when using different amounts of data for pre-training. ConVIRT and GLoRIA are trained on MIMIC-CXR (369K) and CheXpert (191K) dataset, respectively. Our method yields superior ACC than GLoRIA using near $1/10$ of pre-training data.

However, adopting vision-text pre-training on medical domain is a non-trivial task due to (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, while the total number of publicly available medical images and reports is orders of magnitude below; and (2) specificity of medical images and reports: compared to general domains (e.g., "cats" v.s. "dog"), the differences within medical domains are more subtle and fine-grained (e.g., "pneumonia" v.s. "consolidation"). In a nutshell, it is necessary to (1) address the data insufficiency issue; and (2) capture the subtle yet crucial medical meanings.

Existing works try to tackle the challenges above in different ways. ConVIRT (Zhang et al., 2020) jointly trains the vision and text encoders with the

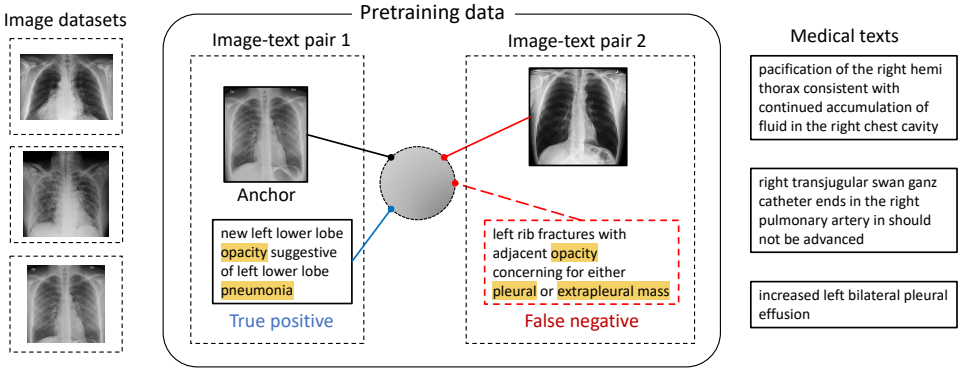


Figure 2: Demonstration of challenges in medical image-text contrastive learning. (1) Pre-training data only includes paired images and texts. However, many more image-only and text-only datasets are ignored. (2) False negatives appear. For an anchor image, previous methods treat paired texts (i.e., reports from the same patient's study) as positives and unpaired texts (i.e., reports from other patients' studies) as negatives. However, the negative texts can describe the same symptoms as the anchor texts.

paired medical images and reports via a bidirectional contrastive objective; GLoRIA (Huang et al., 2021) further models both the global and local interactions between medical images and reports to capture the pathology meanings from specific image regions. However, both works have significant limitations, as illustrated in Fig. 2.

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

- **False negatives in contrastive learning.** Both 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.

To handle the above challenges, we propose a simple yet effective approach, namely MedCLIP. It has the following contributions:

- **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 instead of expensively scaling up data.

- **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.

We make comprehensive evaluation on MedCLIP across four public datasets. Results show that MedCLIP reaches extremely high data efficiency, as shown in Fig. 1. Our method obtains better performances than the state-of-the-art GLoRIA (Huang et al., 2021) using only 10% pre-training data. Extensive experiments verify MedCLIP's transferability to various downstream tasks. It wins over baselines by a large margin: over 10% improvement of prediction ACC for zero-shot prediction and supervised image classification tasks on average; over 2% improvement of retrieval precision. Details are in §4.

2 Related Works

Vision-text representation learning was shown to learn good visual representations (Joulin et al., 2016; Li et al., 2017; Sariyildiz et al., 2020; Desai and Johnson, 2021; Kim et al., 2021; Wang et al., 2021a). But all of them work on paired image and captions from general domain, e.g., Flickr (Joulin et al., 2016) and COCO Captions (Desai and Johnson, 2021). Likewise, these methods do

not support cross-modal retrieval hence do not support zero-shot predictions either.

Many propose to learn visual-semantic embedding for vision-text retrieval (Liu et al., 2019; Wu et al., 2019; Lu et al., 2019; Huang et al., 2020; Chen et al., 2021) by attention or objection detection models; and by vision-text contrastive learning (Zhang et al., 2020; Jia et al., 2021; Yuan et al., 2021; Yu et al., 2022) or multiple vision and text supervision (Singh et al., 2021; Li et al., 2022). They all work on general domain where near infinite web images and captions are available, which dwarfs the scale of medical image-text data. This challenge hurdles the execution of self-supervised CL for large vision-text transformers. Though remedies like data augmentation (Li et al., 2021) and knowledge graph (Shen et al., 2022) were proposed, the magnitude of used data is still far larger than medical data.

Medical image-text representation learning was investigated based on contrastive learning as well (Zhang et al., 2020; Huang et al., 2021; Wang et al., 2021b). Nonetheless, they all work on paired medical images and texts so still encounter the lacking data challenge. Moreover, they all suffer from the false negative noises when adopting noise contrastive estimation (NCE) (Van den Oord et al., 2018) to perform instance discrimination (Wu et al., 2018), which undermines the representation quality (Arora et al., 2019; Zheng et al., 2021). Our work bridges the gap by making the full use of all available medical data to support medical image-text pre-training. And we harness medical knowledge tailored to eliminate false negatives in contrastive learning to improve the pre-training data efficiency.

3 Method

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

3.1 Vision and Text Encoder

`MedCLIP` consists of one visual encoder and one text encoder.

Vision Encoder. We encode images into embeddings $\mathbf{v} \in \mathbb{R}^D$ using a vision encoder E_{img}. A projection head then maps raw embeddings to $\mathbf{v}_p \in \mathbb{R}^P$.

$$\mathbf{v} = E_{img}(\mathbf{x}_{img}) \tag{1a}$$

$$\mathbf{v}_p = f_v(\mathbf{v}) \tag{1b}$$

where f_v is the projection head of the vision encoder.

Text Encoder. We create clinically meaningful text embeddings $\mathbf{t} \in \mathbb{R}^M$ by a text encoder. We project them to $\mathbf{t}_p \in \mathbb{R}^P$ as

$$\mathbf{t} = E_{txt}(\mathbf{x}_{txt}) \tag{2a}$$

$$\mathbf{t}_p = f_t(\mathbf{t}) \tag{2b}$$

where f_t is the projection head and E_{txt} denotes the text encoder. This gives the same embedding dimension P as the vision encoder, suitable for contrastive learning.

3.2 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) due to the significant expense of supplying high-quality annotations by medical specialists as well as privacy and legal concerns. 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. The challenge is that for image-only, and text-only datasets, CLIP-like contrastive learning is infeasible. Also, we want to dig out all positive pairs to eliminate false negatives.

Suppose we have n paired image-text samples, m labeled images, and h medical sentences. Previous methods are only able to use n paired samples. By contrast, we decouple the n paired samples into n images and n sentences, respectively. Ultimately, we are able to obtain $(n+m) \times (n+h)$ image-text pairs by traversing all possible combinations, which results in $\frac{(n+m)\times(n+h)}{n}\times$ more supervision. For instance, in Fig. 2, previous method pretrains on 2 image-text pairs while `MedCLIP` is capable of exploiting $(2+3) \times (2+3) = 25$ samples in total.

To fulfill the additional supervision, we propose to leverage external medical knowledge to build the knowledge-driven *semantic similarity*. Unlike previous works that treat all positive samples equally (Khosla et al., 2020; Zheng et al., 2021; Wang and Sun, 2022), here we propose to differentiate samples via their semantic similarities.

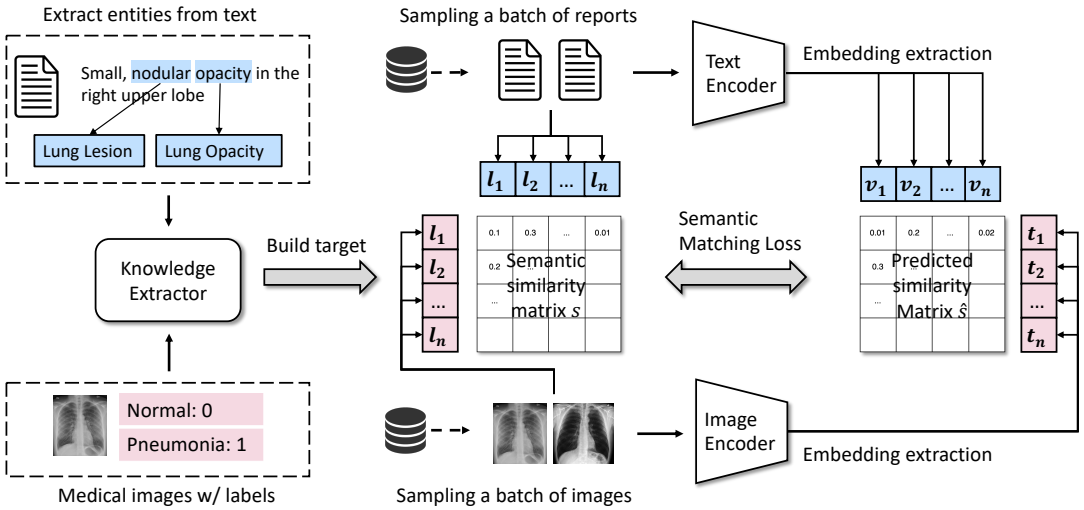


Figure 3: The workflow of MedCLIP. The knowledge extraction module extracts medical entities from raw medical reports. Then, a semantic similarity matrix is built by comparing medical entities (from text) and raw labels (from images), which enables pairing arbitrary two separately sampled images and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

In particular, we split raw reports into sentences x_{txt}. MetaMap (Aronson and Lang, 2010) is used to extract entities defined in Unified Medical Language System (Bodenreider, 2004) from raw sentences. Follow the practice of Peng et al. (2018), we focus on 14 main entity types shown by Table 5. Likewise, for images with diagnosis labels, we leverage MetaMap to map the raw classes to UMLS conceptions thus being aligned with entities from texts, e.g., "Normal" maps to "No Findings". We build multi-hot vectors from the extracted entities for images and texts, as $\mathbf{l}_{img}$ and $\mathbf{l}_{txt}$, respectively. Therefore, we unify the semantics of images and texts. For any sampled x_{img} and x_{txt}, we can measure their semantic similarity by comparing the corresponding $\mathbf{l}_{img}$ and $\mathbf{l}_{txt}$.

3.3 Semantic Matching Loss

We bridge the images and texts through the built semantic labels $\mathbf{l}_{img}$ and $\mathbf{l}_{txt}$. During each iteration, we sample N_{batch} input images $\{\mathbf{x}_{image}\}$ and text $\{\mathbf{x}_{text}\}$ separately. Instead of defining positive pairing by searching equivalent labels, we propose to build soft targets s by

$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}. \tag{3}$$

s thus indicates the medical semantic similarity.

For an image i, we obtain a set of s_{ij} where $j = 1 \dots N_{batch}$ corresponds to the batch of texts. The soft target is computed by normalizing across j by softmax.

$$y_{ij}^{v \to t} = \frac{\exp s_{ij}}{\sum_{j=1}^{N_{batch}} \exp s_{ij}}. \tag{4}$$

Similarily, the reversed text-to-image soft targets are obtained by

$$y_{ji}^{t \to v} = \frac{\exp s_{ji}}{\sum_{i=1}^{N_{batch}} \exp s_{ji}}. \tag{5}$$

The logits are obtained by cosine similarities between image and text embeddings:

$$\hat{s}_{ij} = \tilde{\mathbf{v}}_i^{\top} \cdot \tilde{\mathbf{t}}_j, \tag{6}$$

where $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{t}}_j$ are normalized $\mathbf{v}_p$ and $\mathbf{t}_p$, respectively. The predicted similarity is also obtained by softmax function

$$\hat{y}_{ij} = \frac{\exp \hat{s}_{ij}/\tau}{\sum_{i=1}^{N_{batch}} \exp \hat{s}_{ij}/\tau}. \tag{7}$$

τ is the temperature initialized at 0.07. The *semantic matching loss* is hence the cross entropy between the logits and soft targets as

$$\mathcal{L}^{v \to l} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} y_{ij} \log \hat{y}_{ij}. \tag{8}$$

Likewise, we can compute $\mathcal{L}^{l \to v}$ and then reach to

$$\mathcal{L} = \frac{\mathcal{L}^{v \to l} + \mathcal{L}^{l \to v}}{2} \tag{9}$$

as the final training objective.

Table 1: Results of zero-shot image classification tasks on four datasets. We take an additional prompt ensemble version of each method (with subscript $_{ENS}$). We take the mean and standard deviation (STD) of accuracy (ACC) in five runs considering the randomness of prompt generation process. Best scores across a dataset are in bold.

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)
ConVIRT$_{ENS}$	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)
GLoRIA$_{ENS}$	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)
MedCLIP-ResNet$_{ENS}$	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)
MedCLIP-ViT$_{ENS}$	**0.5942(<0.01)**	0.5024(<0.01)	0.7943(<0.01)	**0.7682(<0.01)**

Table 2: Results of medical image classification tasks after fine-tuning. Best scores are in bold.

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**

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

4 Experiments

We conduct extensive experiments on four X-ray datasets to answer the following questions:

- **Q1.** Does the proposed pre-training method yield better zero-shot image recognition performances?
- **Q2.** Does the knowledge-driven supervision, i.e., semantic matching task, facilitate the contrastive image-text pre-training?
- **Q3.** Does MedCLIP bring better performance and label efficiency for downstream classification tasks with fine-tuning?
- **Q4.** Are the learned embeddings good at cross-modal retrieval tasks?
- **Q5.** How do the learned embeddings look like?

4.1 Datasets

CheXpert (Irvin et al., 2019) is a large dataset of chest X-rays with 14 observation labels collected from Stanford Hospital. Note that this dataset does not provide the corresponding medical reports to the public. We use the training split of this dataset for pre-training. For evaluation, we follow (Huang et al., 2021) and sample a multi-class classification dataset from the testing split, namely CheXpert-5x200. This multi-class classification dataset has 200 exclusively positive images for the five CheXpert competition tasks: Atelectasis, Cardiomegaly, Edema, Pleural, Effsion.

MIMIC-CXR (Johnson et al., 2019) is a large chest X-ray database with free-text radiology reports collected from the Beth Israel Deaconess Medical Center in Boston, MA. We use the training split of this dataset for pre-training. For evaluation, we also sample a MIMIC-5x200 dataset for the same five tasks above.

COVID (Rahman et al., 2021) is a publicly avail-

able x-ray dataset with COVID v.s. non-COVID labels. The positive and negative ratio is roughly 1:1. We use this dataset for evaluation.

RSNA Pneumonia (Shih et al., 2019) is a collection of pneumonia cases found in the database of chest x-rays made public by the National Institutes of Health. This is a binary classification dataset: pneumonia v.s. normal. We sample a balanced subset (i.e., 1:1 positive and negative ratio) and use it for evaluation.

4.2 Baselines

Random is a ResNet-50 (He et al., 2015) model with its default random initialization.

ImageNet is a ResNet-50 (He et al., 2015) model with weights pretrained on the standard ImageNet ILSVRC-2012 task (Deng et al., 2009).

CLIP (Radford et al., 2021) is a vision-text contrastive learning framework pre-trained on a dataset of 400M image-texts pairs collected from the internet.

ConVIRT works on vision-text contrastive learning in medicine. It employs a plain InfoNCE loss (Van den Oord et al., 2018) on paired X-rays and reports. We reproduce it based on their paper based on BioClinicalBERT text encoder and ResNet50 (He et al., 2016) vision encoder.

GLoRIA (Huang et al., 2021) entangles image subregions and words in inference by cross-attention which was argued to better capture key characteristics in images and reports. We implement it based on the official code and the provided pretrained weights [1].

4.3 Implementation Details

We use the BioClinicalBERT [2] as the backbone text encoder and Swin Transformer (Liu et al., 2021) with ImageNet (Deng et al., 2009) pre-trained weight as the backbone vision encoder. Both transformer-based models are drawn from the transformers library (Wolf et al., 2019). We also provide ablation study with ResNet-50 (He et al., 2015) as the vision encoder, which is in-line with previous works (Zhang et al., 2020; Huang et al., 2021).

MIMIC-CXR and CheXpert are used for pretraining where we held 5000 samples out out for evaluation. All images are padded to square then scaled to 224×224. For MIMIC-CXR, we combine the "Findings" and "Impression" sections of reports then split them into sentences. We remove sentences with less than 3 words. We take a linear projection head with output dimension 512, a learnable temperature τ initialized on 0.07. We utilize image augmentations to first scale to raw images to 256×256 then apply random crop with size 224×224; horizontal flipping with 0.5 probability; color jittering with brightness and contrast ratios from $[0.8, 1.2]$; random affine transformation with degree sampled from $[-10, 10]$, max translation rate 0.0625, and scale factor in $[0.8, 1.1]$. Other hyperparameters are: learning rate 5e-5, batch size 100, weight decay 1e-4, number of epochs 10, learning rate warmup ratio 0.1. We employ mixed-precision training such that the pretraining finishes in 8 hours on a single RTX-3090 GPU.

4.4 Q1. Zero-Shot Classification

We conduct zero-shot image classification evaluation on four datasets: CheXpert-5x200, MIMIC-5x200, COVID, and RSNA. The learned image-text encoders are used to support zero-shot prediction by matching the encoded image embeddings and the embeddings of created prompts for each disease class. We illustrate the results in Table 1.

It can be found that our method outperforms all the other baselines by a great margin. `MedCLIP` is capable of benefiting from prompt ensemble to yield better performance. By contrast, the ensemble does not always lead to positive effect to the other two, especially that GLoRIA is usually harmed by ensemble. One reason might be that ConVIRT and GLoRIA cannot differentiate false negatives in contrastive pre-training, and those false negatives are incoporated with prompt ensemble which confuses the model. Besides, we observe that the original CLIP model yields bad predictions that are basically identical to random guess on all datasets. It demonstrates the discrepancy between the general internet image-text and the ones in medical domain.

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. To endow the model to detect COVID-19 infection,

[1] https://github.com/marshuang80/gloria

[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

we refer to the descriptions proposed in (Smith et al., 2020): *the presence of patchy or confluent, bandlike ground-glass opacity or consolidation in a peripheral and mid to lower lung zone distribution*, to build the prompts. This result demonstrates that contrastive pre-training of MedCLIP provides it with the transferability to out-of-domain classes.

4.5 Q2. Pre-training Data Efficiency

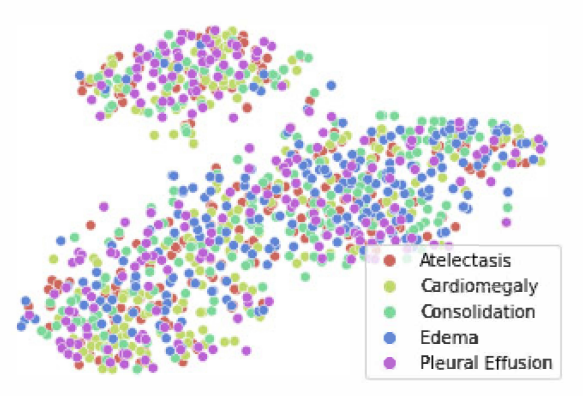
Data efficiency is a key challenge in CLIP based methods. As evident, CLIP uses 400M image-text pairs in the training phase, which is not just computationally expensive but also infeasible in medical domain due to limited data. To evaluate the data efficiency of MedCLIP, we subsample the pre-training data to 20K, 50K, and 200K, then pre-train MedCLIP and record the yielded model zero-shot prediction on CheXpert-5x200 data. Results show in Fig. 1.

We surprisingly find that with 20K data MedCLIP yields superior performance over GLoRIA that learns from the whole CheXpert dataset (around 200K image-text pairs). Likewise, MedCLIP beats ConVIRT that uses 369K data. When we include more training data, MedCLIP obtains a lift on its accuracy as well. We do not observe the saturation of zero-shot ACC at 570K samples (MIMIC-CXR plus CheXpert). It signifies the great capacity of MedCLIP on learning from multi-sourced data.

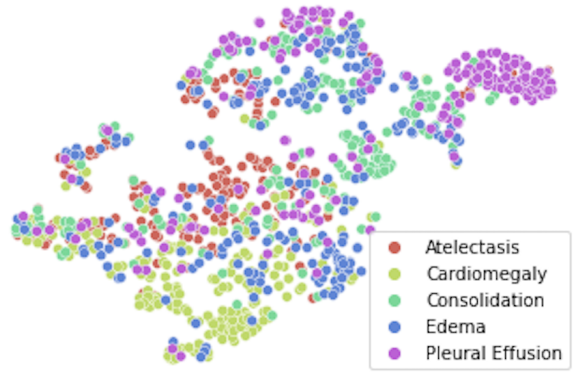
4.6 Q3. Fine-tune for Classification

We aim to evaluate the learned model transferability to downstream supervised tasks. We draw and froze the image encoder and fine-tune a randomly initialized linear classification head on the training data with cross-entropy loss. Results are in Table 2. We show that MedCLIP still achieves the best performances across all three methods. What is more, we surprisingly find that MedCLIP makes zero-shot prediction comparable with supervised learning models when contrasting Table 2 to Table 1.

On all datasets, the zero-shot MedCLIP performs better than the other supervised models. Specifically, we find on COVID, zero-shot MedCLIP performs better than its supervised counterpart, which demonstrates the supremacy of MedCLIP on low-resource scenarios. On the contrary, without pre-training on medical domain data, the ResNet baselines reach inferior performances.



(a) CLIP



(b) MedCLIP

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

4.7 Q4. Image-Text Retrieval

We choose CheXpert-5x200 to evaluate the semantic richness of the learned representations by all models through the image-text retrieval task. Since 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 CheXpert-5x200 dataset. This gives rise to 1,000 images and 1,000 sentences as the retrieval dataset. We use Precision@K to calculate the precision in the top K retrieved reports/sentences by checking if the report belongs to the same category as the query image.

We display the results by Table 4. It can be seen that MedCLIP achieves the best performances across all methods. This indicates that our method efficiently provide the required semantic information to retrieve texts. We find that there is an increase precision for MedCLIP with the higher K. Analysis of this phenomenon is present in the Appendix A.

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**

4.8 Q5. Embedding Visualization

We also demonstrate the effectiveness of our representation learning framework by plotting t-SNE (Van der Maaten and Hinton, 2008) of image embeddings produced for CheXpert-5x200 images. We compare its embeddings with CLIP model embeddings. As visible in Fig. 4, our model produces better clustered representation. Whereas, CLIP model t-SNE plot is homogeneous. It is because most medical X-Rays share pretty high overlapping while only small lesion regions are different. Nonetheless, MedCLIP still detects clusters by the lesion types.

5 Conclusion

In this work, we propose a decoupled medical image-text contrastive learning framework named MedCLIP. It significantly expands the training data size with a combinatorial magnitude. Meanwhile, the introduction of medical knowledge sheds light on alleviating false negatives. As a result, MedCLIP yields an excellent pretraining data efficiency: it wins over the state-of-the-art baseline by 1% ACC with around $10\times$ fewer data. Moreover, we verify the prominence of MedCLIP on zero-shot prediction, supervised classification, and image-text retrieval tasks. It is expected to support a foundational model for the medical domain and handle medical diagnosis when facing diverse diseases with low resource requirements.

Limitations

This work leverages medical domain knowledge to decouple contrastive learning on medical images and texts. Hence, it significantly expands the available training data for pretraining. Meanwhile, the proposed knowledge-guided semantic matching loss debugs the false negatives appearing in naive contrastive learning. It still encounters failure cases where incorrect semantic tags are detected or missing detection of negation or uncertainty phrases. A remedy can be introducing learning from noisy data techniques (Wang et al., 2020, 2022) to alleviate the noises in the extracted semantic similarity matrix.

Another concern is that though we prove MedCLIP is able to reach comparable zero-shot prediction accuracy to the finetuned counterpart, it is still not amenable to practical use. We suppose the reasons include (1) the prompt-based inference relies on the prompt quality and (2) more pretraining data is desired to further enhance the pretraining. Specifically, for the (1) point, it is promising to leverage prompt-learning methods (Zhou et al., 2022) to automate the model application to downstream tasks instead of executing manual prompt engineering.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning*, pages 9904–9923. International Machine Learning Society (IMLS).

O Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–70.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. 2021. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173.

FDA. 2022. Medical imaging.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2017. Learning visual n-grams from web data. In *IEEE International Conference on Computer Vision*, pages 4183–4192.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. 2021. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. In *European Conference on Computer Vision*, pages 153–170. Springer.

Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. 2022. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*.

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. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. FLAVA: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*.

David L Smith, John-Paul Grenier, Catherine Batte, and Bradley Spieler. 2020. A characteristic chest radiographic pattern in the setting of the covid-19 pandemic. *Radiology: Cardiothoracic Imaging*, 2(5):e200280.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. 2022. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021a. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. 2021b. Self-supervised image-text pre-training with mixed data in chest x-rays. *arXiv preprint arXiv:2103.16022*.

Zifeng Wang and Jimeng Sun. 2022. Transtab: Learning transferable tabular transformers across tables. *arXiv preprint arXiv:2205.09328*.

Zifeng Wang, Hong Zhu, Zhenhua Dong, Xiuqiang He, and Shao-Lun Huang. 2020. Less is better: Unweighted data subsampling via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6340–6347.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

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

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. 2021. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

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

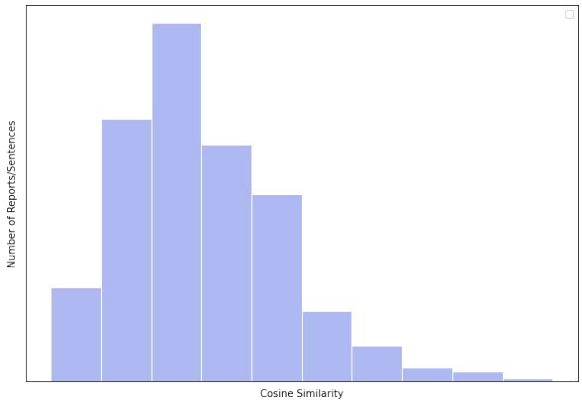
Finding types
No Finding
Enlarged Cardiomediastinum
Cardiomegaly
Lung Opacity
Lung Lesion
Edema
Consolidation
Pneumonia
Atelectasis
Pneumothorax
Pleural Effusion
Pleural Other
Fracture
Support Devices

A Analysis of Image-text retrieval results

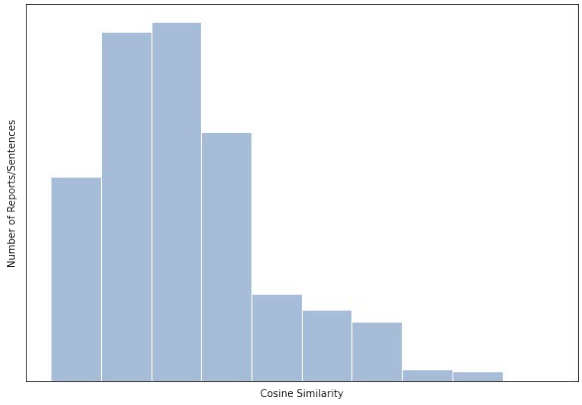
We observed that there is certain increase in `MedCLIP` Precison@K metric in Table 4. This phenomenon is sort of counterintuitive. As mentioned in Section 4.7, we used CheXpert-5x200 dataset for images and MIMIC-CXR for text in our image-text retrieval task. Each of them, i.e. images and text reports/sentences, has 1000 rows. Particularly, 200 images and sentences/reports are there for each class in CheXpert-5x200.

After running the main experiment we decided to plot some graphs to gain insight. Settings for the same is described as following: We pick up a class (e.g. Atelectasis) and for each image in that class we pick up sentences/reports, from top 10 (based on cosine distance) out of all the texts retrieved, that belong to the same class as that of the image in consideration. We also pick up their cosine similarity score. Finally, we plot a histogram (for each class) where x-axis is the cosine distance and the height of the bins of histogram represent the number of texts(retrieved in previous step) that have cosine similarity score of the bin. We have plotted few such plots and can be seen in Figure 5.

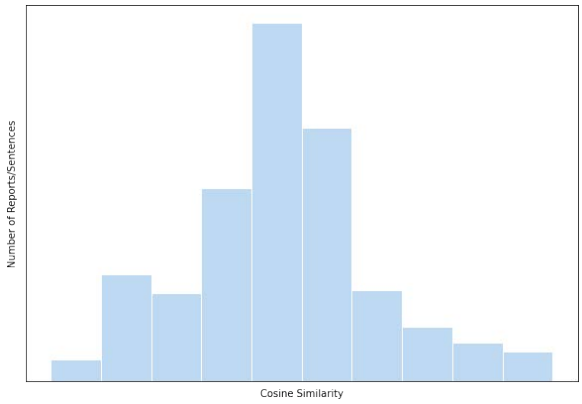
Intuitively, we are plotting number of texts that would appear in Precision@K of a particular class given a cosine similarity threshold. Now as the histogram tells, there are more text centered around relatively smaller cosine similarity score. Further, as the K increases in the Precision@K, intuitively cut-off cosine similarity(or threshold) score would decrease and hence more text from the same class start showing up in the result. This explains why Precision@K increased. In this sense, it is beneficial to investigate to address the non-smooth anisotropic distribution of image and text embeddings.



(a) Histogram Plot for Atelectasis class



(b) Histogram Plot for Cardiomegaly class



(c) Histogram Plot for Consolidation class

Figure 5: Visualization of the similarity distributions computed based on `MedCLIP` embeddings.