Evaluating CLIP and MedCLIP on Pediatric Chest X-Ray Images

Dante Miller
Department of Computer Science, Rice University
dm85@rice.edu

Abstract

Compared to standard image models, which apply a feature extractor with a linear classifier to learn the correct (image, label) relationships, CLIP is trained on both an image and text encoder to learn the correct (image, text) relationships. Given the general domain of the (image, text) relationships CLIP learns, it encounters limitations when dealing with subtlety and fine-grained detail such as medical images. To bridge this gap, MedCLIP was developed to fine-tune CLIP on pairings of (medical image, medical report), specifically for the goal of enhancing accuracy in predicting whether a patient has COVID/pnuemonia or not from adult lung chest X-ray images. In our our project we explore the use of zero-shot transfer and fine-tuning of both CLIP and MedCLIP on pediatric lung chest X-ray images, and compare their performance using standard metrics. We demonstrate that zero-shot learning, when comparing the two models, exhibits similar performance but reveals distinct trade-offs in terms of recall and precision. Additionally, we find that during fine-tuning, the performance of MedCLIP significantly surpasses that of CLIP.

1. Introduction

Chest radiography is a pivotal imaging modality in medicine, essential for screening and diagnosing a variety of disorders in patients, including those affecting the lung parenchyma, airways, heart, and mediastinum. It is also invaluable for assessing support devices and diagnosing life-threatening conditions such as pneumonia [5]. Crucially, the interpretation of chest radiographs significantly differs across age groups, as the radiographic appearance varies with age. Viral respiratory infections, particularly COVID-19, impose a significant clinical burden on the population, leading to increased sick visits, hospitalizations, and, unfortunately, higher mortality rates [4]. These concerns underscore the necessity of developing data-driven approaches and computational tools to aid physicians in diagnosing viral respiratory infections [6].

The application of computer vision models to medical imaging has been a significant challenge that many researchers have sought to address. The difficulty lies in accurately identifying and correlating specific areas within images with relevant medical diagnoses, such as disease types. Additionally, there has been a recent expansion in incorporating medical text into the medical image diagnosis challenge. This includes the development of MedCLIP [7], a variant of the CLIP [9] model fine-tuned on medical image datasets, such as CheXpert and MIMIC-CXR. Our work focuses on pediatric lung chest X-ray images obtained from a publicly available dataset [3] and a private dataset [8]. Examples of paired (medical image, text) samples can be seen in Figure 1.

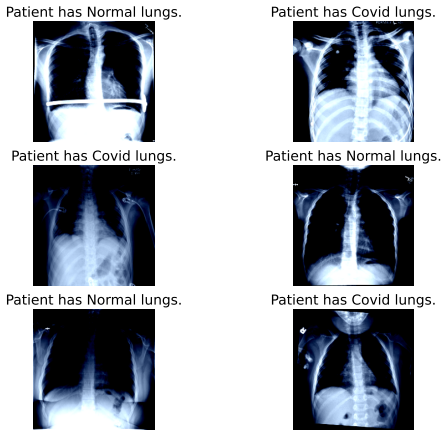

Figure 1. Samples of pediatric lung chest X-ray images and their respective labels are displayed.

2. Related Work

Early research in applying machine learning to medical diagnosis focused on convolution-based models for lung chest X-ray datasets [1][3][10]. TorchXRayVision [2], one of the state-of-the-art methods in medical diagnosis, trained a variety of convolution-based models on various chest X-ray image datasets. Similar efforts, akin to TorchXRayVision, have targeted convolution-based models and pediatric chest X-ray image datasets [1][3][10]. MedCLIP [9], a

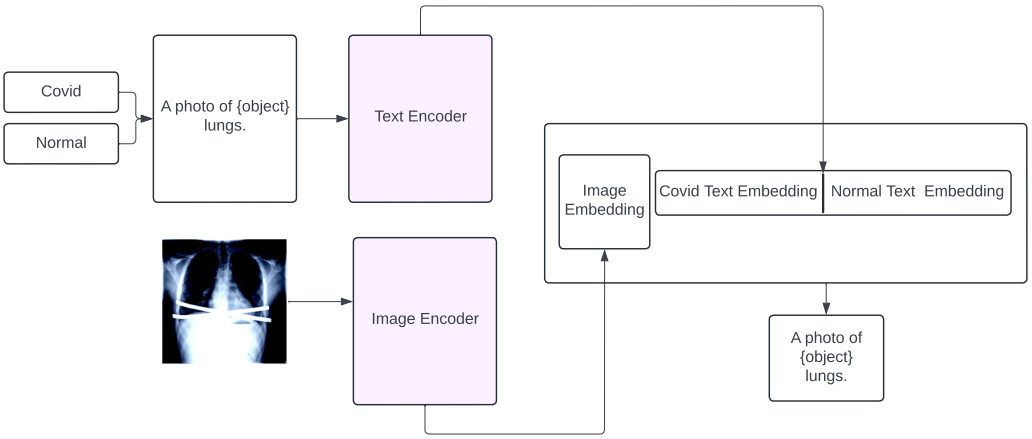

Figure 2. Diagram depicting the high-level conceptual framework of both CLIP and MedCLIP models.

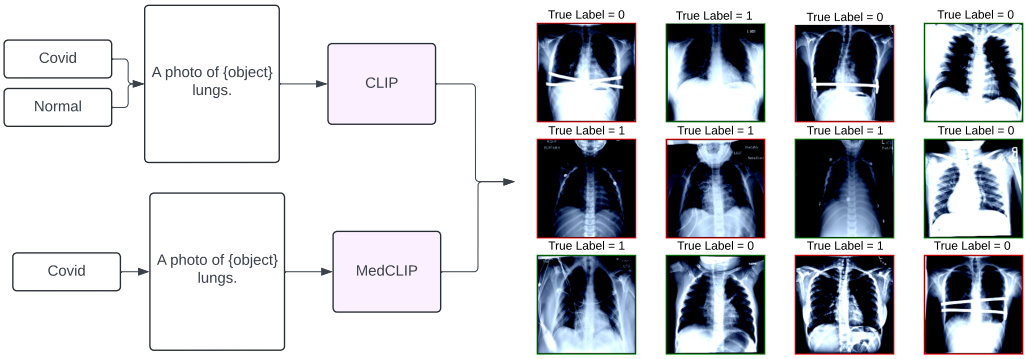

Figure 3. Diagram depicting an example of the expected inputs and outputs from both CLIP and MedCLIP. This is the result of performing zero-shot learning using CLIP on Ori. The green boundary indicates the classification is correct, whereas red indicates that it is incorrect.

variant of the CLIP model fine-tuned on medical image datasets such as CheXpert and MIMIC-CXR, focused on the direct applicability of CLIP to adult chest X-ray image datasets. Our work is centered on evaluating both CLIP and MedCLIP to determine whether directly applying a general-based CLIP model outperforms MedCLIP in terms of zero-shot learning and fine-tuning.

3. Method

3.1. CLIP

CLIP [7], developed by OpenAI, is a model designed to understand images within the context of textual descriptions. It consists of two main components: an image encoder and a text encoder. The image encoder converts images into high-dimensional vector spaces, while the text encoder performs a similar transformation for textual descriptions. It employs a contrastive learning approach to align the vector representations of images and texts within the same space and fine-tunes the parameters to minimize the distance between matching image-text pairs and maximize the distance between non-matching pairs.

3.2. MedCLIP

MedCLIP [9] builds on the foundational concepts of CLIP, integrating medical domain knowledge to enhance the processing and comprehension of medical images and reports. Although the framework includes a knowledge extractor for identifying specific medical terms within textual data, it is not utilized in our experiments, which focus solely on medical images and their corresponding labels. The developers introduced the concept of creating a semantic similarity matrix based on actual targets and report text. They generate a predicted similarity matrix using the predicted target vector representation from the image encoder and the representation from the text encoder. They then apply the symmetric similarity loss used in the original CLIP framework.

The ideas I mentioned for both CLIP and MedCLIP can be better understood by referring to Figure 2. The process begins with labels that are incorporated into a preset sentence description, which is then encoded using a text en-

Dataset	Pneumonia	Normal	Train	Validation	Test
P1 (UCSD)	1493	1583	1967	493	616
P2 (ORI)	404	419	526	132	165

Table 1. Distribution of pneumonia and normal cases across different datasets with train, validation, and test splits.

Dataset	Method	Accuracy	Precision	Recall	AUC
ORI	CLIP	0.5747	0.5622	0.6040	0.5153
ORI	MedCLIP	0.5808	0.5925	0.4678	0.6015
UCSD	CLIP	0.6752	0.6963	0.5867	0.4645
UCSD	MedCLIP	0.7500	0.6790	0.9196	0.8588

Table 2. Zero-shot learning experimental results for CLIP and MedCLIP models on ORI and UCSD datasets.

coder. Similarly, an image of the lungs is encoded using an image encoder. These two encoders produce separate vector embeddings. Subsequently, we calculate a similarity matrix; in our case, it is actually a similarity vector, as we consider one image at a time. The vector embedding with the greatest similarity to the image embedding is then selected. While there are many more details, this explanation captures the high-level idea.

I also wanted to include an additional example of the input and output pairings for what is expected from both models, which can be observed in Figure 3. The CLIP model is based on finding the text for an associated class that matches the image. MedCLIP is quite similar to CLIP but emphasizes binary classification with sigmoid. What is interesting about MedCLIP is that it is trained on more specific information about the lungs. Examples include 'Patchy ground glass opacity in lower' and 'Confluent ground glass consolidation in mid', which is very intriguing as this deviates a bit from the structure that CLIP is normally trained with, and what we fine-tune the MedCLIP model with, which is 'a photo of Covid lungs'.

4. Code Availability

Detailed information about CLIP and MedCLIP can be found at their respective repositories: CLIP and MedCLIP. My code and experimentation details are available at MedCLIP-pediatric.

5. Experiments and Results

5.1. Data

For these experiments we considered two types of pediatric lung chest X-ray image datasets: P1 (ORI) [8] and P2 (UCSD)[3]. Further details about these datasets can be found in Figure 1.

5.2. Zero-Shot Transfer

Zero-shot transfer involves applying a model to a dataset it has not previously encountered. Neither the CLIP nor the MedCLIP models were exposed to the datasets or images we are using for evaluation. While CLIP was trained on a diverse, general corpus, MedCLIP received training on adult lung chest X-ray images for various tasks. However, pediatric chest X-ray images, the focus of our work, are considered a distinct domain. This distinction arises because the interpretation of chest radiographs varies significantly across different age groups.

5.3. Fine-tuning Contrastive Learning

We fine-tuned the models following the guidelines and code provided in their respective GitHub repositories, with particular emphasis on the calculation of loss during training. Unlike CLIP, where manual adjustment was necessary, MedCLIP automatically accommodates this aspect of the fine-tuning process.

5.4. Implementation Details

The majority of the code was implemented using PyTorch and consists of components, particularly some model parts, referenced in the CLIP and MedCLIP repositories, as detailed further in the code availability section above. Initially, the code extracts data from my Google Drive, followed by the application of preprocessing provided by both CLIP and MedCLIP. Thus, depending on the model used, the associated preprocessing is applied to the data. Subsequently, we visualize the images, with examples available in the results folder. After visualization, we perform zero-shot transfer on both datasets utilized in our experiments, with the results being saved accordingly. Following the zero-shot transfer, fine-tuning of both models using contrastive learning is performed, and these results are also saved. The code for the majority of the zero-shot and fine-tuning processes was written by myself, with references from both reposi-

Dataset	Method	Accuracy	Precision	Recall	AUC
ORI	CLIP	0.7879	0.8833	0.6543	0.6933
ORI	MedCLIP	0.7515	0.7500	0.7407	0.8257
UCSD	CLIP	0.9529	0.9787	0.9231	0.6986
UCSD	MedCLIP	0.9594	0.9507	0.9666	0.9900

Table 3. Fine-tuning experimental results for CLIP and MedCLIP models on ORI and UCSD datasets.

Dataset	Method	Accuracy	Precision	Recall	AUC
ORI	CLIP	0.6501	0.6272	0.7079	0.6515
ORI	MedCLIP	0.6063	0.5897	0.6510	0.6238
UCSD	CLIP	0.5813	0.5395	0.9370	0.7630
UCSD	MedCLIP	0.9629	0.9372	0.9900	0.9901

Table 4. Transfer-learning experimental results for CLIP and MedCLIP models on ORI and UCSD datasets. The models used were previously fine-tuned on the counterpart dataset.

tories to understand the inputs and outputs, as well as key aspects related to the loss, and the expected inputs and outputs.

Regarding data characteristics, I used batch sizes of 256, and the resolutions are expected to be 224×224. For fine-tuning, the Adam optimizer is used; the CLIP model utilizes a learning rate of $lr = 1 \times 10^{-5}$, with weight decay set to 0.1 for the "ucsd" dataset and 1×10^{-5} for other datasets. Conversely, the MedCLIP model employs a learning rate of $lr = 1 \times 10^{-5}$, with weight decay adjustments of 1×10^{-2} for the "ucsd" dataset and 1×10^{-4} for other datasets. Both models are trained across 50 epochs, incorporating early stopping with a patience of 20. For future work, I want to implement parameter tuning but due to the sake of the time I had in this project I found the parameters through experimentation.

5.5. Results

The experimental results presented in the tables demonstrate the performance of the CLIP and MedCLIP models across pediatric datasets ORI and UCSD, considering different scenarios such as zero-shot learning, fine-tuning, and transfer learning. When examining Table 2 for the zero-shot learning experiment, both models generally perform better on the UCSD dataset than on the ORI dataset, with MedCLIP showing a significant advantage in recall and AUC on UCSD. Focusing on the fine-tuning experiment, the results from Table 3 indicate superior performance after fine-tuning, particularly for MedCLIP, which exhibits remarkable improvements on the UCSD dataset where both models achieve high precision and AUC values. For the transfer learning experiment, according to Table 4, while MedCLIP excels in accuracy and AUC on UCSD, CLIP struggles, particularly on ORI in terms of precision and recall. From these experiments, we noted that fine-tuning provided better overall results, especially enhancing MedCLIP's performance. Evaluating the transfer learning experiment is challenging, as UCSD tends to be simpler in terms of identifying the (image, label) relationships compared to ORI. However, from this analysis, we can conclude that CLIP performs better with more detailed data, as it generalizes better from UCSD to ORI compared to MedCLIP, which performs worse.

It should also be mentioned that MedCLIP is trained on adult chest X-ray images, which differ in terms of radiographical features from those used when diagnosing COVID-19, so the knowledge gained might not transfer as effectively; thus, the performance discrepancies make sense. What surprised me was how easily MedCLIP could be fine-tuned, although the results when generalizing to other datasets were intriguing. I think its failure to generalize from UCSD to ORI is understandable, as it's easier to learn from UCSD, but what is learned might not be accurate. I am curious about the outcomes when MedCLIP is fine-tuned from ORI to UCSD, as we observed more improvements, which makes me wonder if what is learned is actually relevant, or if I am overlooking it because it's applied to UCSD. Nonetheless, it must be relevant, as the model trained on CLIP did not perform as well. Overall, this area warrants further investigation, particularly because MedCLIP is trained on adult chest X-ray images. Understanding more about what features it is actually learning could be crucial, as perhaps the model is functioning correctly, but the transfer learning ability between these two distinct domains might pose challenges. Considering potential avenues for this application, exploring MedCLIP's performance on another adult chest X-ray dataset might be interesting to determine whether it is the model or the features that impact its generalization ability.

References

[1] M. E. Chowdhury et al. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.

[2] J. P. Cohen et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022.

[3] D. S. Kermany et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[4] G. Nino et al. Pediatric lung imaging features of covid-19: A systematic review and meta-analysis. *Pediatric Pulmonology*, 56(1):252–263, 2021.

[5] S. Padash et al. Pediatric chest radiograph interpretation: how far has artificial intelligence come? a systematic literature review. *Pediatric Radiology*, 52(8):1568–1580, 2022.

[6] H. H. Pham et al. Pedicxr: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *Scientific Data*, 10(1):240, 2023.

[7] A. Radford et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, page PMLR, 2021.

[8] Texas Children's Hospital. Data provided by texas children's hospital, 2023. Private communication.

[9] Z. Wang et al. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[10] R. M. Wehbe et al. Deepcovid-xr: An artificial intelligence algorithm to detect covid-19 on chest radiographs trained and tested on a large u.s. clinical data set. *Radiology*, 299(1):E167–E176, 2021.