**HC2**

FYP Progress Report

**Artificial Intelligence**

**Against**

**COVID-19**

by

Pre-training Model:  LU Weiqi, XU Mingshi

Explainable AI:  CHEN, Siyu, ZHOU Taichang

**HC2**

Supervised by

Prof. CHEN Hao (陳浩)

Advised by

Luo Luyang and Zhuang Jiaxin

Submitted in partial fulfillment of the requirements for COMP4981 and CPEG4901

in the

Department of Computer Science and Department of Computer Engineering

The Hong Kong University of Science and Technology

2022-2023

Date of submission: September 16, 2022

**Table of Contents**

[1 Pre-training Mode 4](#_Toc127224992)

[1.1 Introduction 4](#_Toc127224993)

[1.2 Literature Survey 4](#_Toc127224994)

[1.2.1 Contrastive Language Image Pre-training (CLIP) [1] 5](#_Toc127224995)

[1.2.2 CheXzero [3] 5](#_Toc127224996)

[1.2.3 MaskCLIP [2] 6](#_Toc127224997)

[1.2.4 Multi-Granularity Cross-modal Alignment (MGCA) [4] 7](#_Toc127224998)

[1.3 Methodology 8](#_Toc127224999)

[1.3.1 Pretrained CLIP Model 9](#_Toc127225000)

[1.3.2 Prompt Engineering 10](#_Toc127225001)

[1.4 Experiments 11](#_Toc127225002)

[1.4.1 Zero-Shot Classification 11](#_Toc127225003)

[1.4.2 Zero-Shot Segmentation 12](#_Toc127225004)

[1.5 Future Work 13](#_Toc127225005)

[2 Explainable Artificial Intelligence (XAI) 14](#_Toc127225006)

[2.1 Overview 14](#_Toc127225007)

[2.2 Literature Survey 15](#_Toc127225008)

[2.3 Methodology 19](#_Toc127225009)

[2.3.1 Post-hoc 19](#_Toc127225010)

[2.3.2 Ante-hoc 20](#_Toc127225011)

[2.4 Future Plan 22](#_Toc127225012)

[2.4.1 Post-hoc 22](#_Toc127225013)

[2.4.2 Ante-hoc 22](#_Toc127225014)

[3 References 23](#_Toc127225015)

[4 Appendix 25](#_Toc127225016)

[4.1 The 1st  Project Meeting 25](#_Toc127225017)

[4.2 The 2nd Project Meeting 25](#_Toc127225018)

[4.3 The 3rd Project Meeting 25](#_Toc127225019)

# Pre-training Mode

## Introduction

Much research on deep learning for medical images focuses on training an end-to-end model for infection diagnosis and lesion segmentation. However, pre-training models built based on large datasets are attracting increased attention due to their significant improvement on downstream tasks. The Contrastive Language Image Pre-training (CLIP) [1] model is one of the most important works in this area. Trained based on natural image text pairs, CLIP model is known for its zero-shot classification capability. MaskCLIP [2] shows that CLIP can also do zero-shot segmentation. Inspired by this work, we question whether we can fine-tune the CLIP model on the chest-Xray image text pair dataset and preserve its zero-shot capability on both classification and segmentation tasks.

In this project, we first investigated existing fine-tuning methods and evaluated some fine-tuned CLIP models based on medical data [3] on a zero-shot segmentation task. Due to our current knowledge, there do not exist pretraining models which can do zero-shot segmentation tasks on medical data. Our project aims to propose a pre-training method and pre-train an encoder with chest X-ray datasets to improve the performance of zero-shot lesion segmentation.

## Literature Survey

Contrastive Language Image Pre-training (CLIP) [1] is one of the most influential pre-training models in recent years and our model will also be constructed based on this work. Then we also survey a lot of fine-tuned pre-training methods and models. The representative works are also shown below. Although some limitations still exist in this work, we will try our best to distill the advantage of these works.

### Contrastive Language Image Pre-training (CLIP) [1]

CLIP is a pretraining model proposed by OpenAI in 2021. Compared with the previous standard vision models, CLIP shows a significant improvement in zero-shot segmentation tasks. The following are the training and zero-shot process of CLIP.

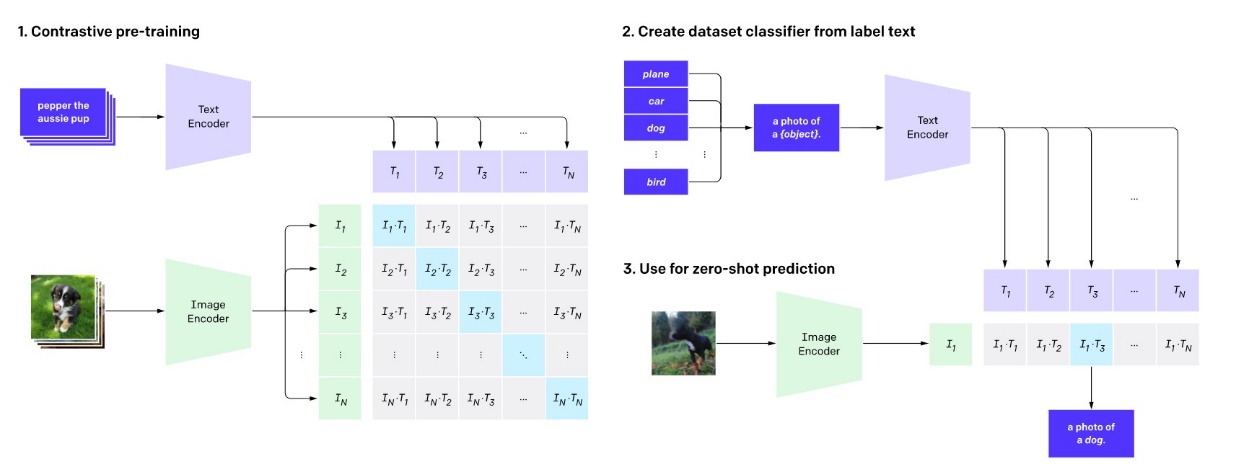


Figure 1: CLIP pre-trains an image encoder and a text encoder and minimizes the contrastive loss between corresponding image text pairs while maximizing this loss between uncorrelated pairs. Then they do zero-shot classification by using a very simple prompt text to transform the class label into a text, then check which text has the highest similarity with the image after encoding.

The CLIP can obtain good performance on zero-shot t classification and is a very powerful tool for the downstream task due to the following reason: Learning from publicly available text-image pairs from the internet, the CLIP dataset does not require expensive data labeling process, thus they trained CLIP on 400 million image-text pairs and achieved significant improvement.

### CheXzero [3]

The CheXzero is a self-supervised model which is fine-tuned on CLIP based on the medical image-text pair dataset. The main dataset used in their fine-tuning process is MIMIC-CXR, which contains 377,110 images corresponding to 227,835 radiographic reports. They preserved the zero-shot classification capability of CLIP and achieved relatively high accuracy on lesion classification tasks. The figure below shows how they finetune the CheXzero and use it to do the zero-shot classification task.

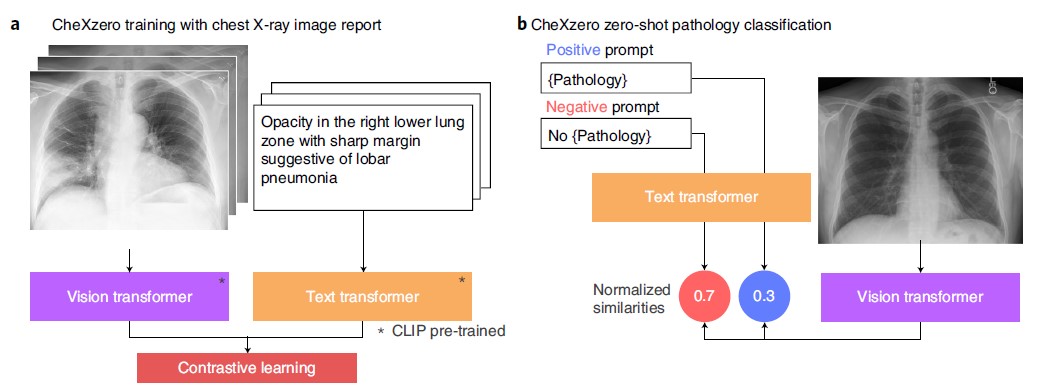


Figure 2: a. CheXzero is fine-tuned on the medical image dataset using the same way that OpenAI’s CLIP is pre-trained. Thus, the output pretraining model structure of CheXzero is the same as that of OpenAI’s CLIP. b. CheXzero uses a slightly different way when they build the prompt for the zero-shot classification. Instead of testing the similarities between the image encoding output and different classes of text embedding, they build a positive prompt and a negative prompt and do binary testing.

Although CheXzero was proven successful in zero-shot classification tasks on medical images, we found it lost the zero-shot segmentation capability found in the CLIP model of OpenAI. This point will be further illustrated in the experiment session.

### MaskCLIP [2]

In this paper, the author proposed a very interesting idea, as the CLIP model can do zero-shot classification on the whole image, it should be able to do zero-shot classification on every pixel, thus the classification task transferred into a segmentation task. Their work contains two main parts: 1. They implemented a zero-shot segmentation model called MaskCLIP, this part does not need any training and all the weight is the same as CLIP. 2. They build a dilated trainable structure and use the zero-shot segmentation result to guide the dilated model. The dilated model is called MaskCLIP+. In our project, we mainly focus on MaskCLIP. The following figure shows the MaskCLIP structure.

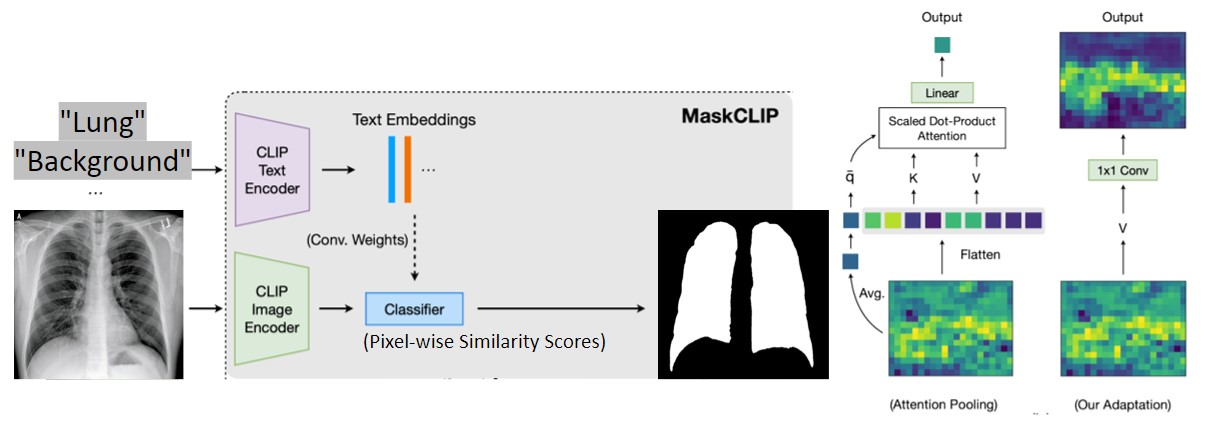


Figure 3: In the MaskCLIP model, it follows the same method as CLIP did on the classification task instead of it do zero-shot prediction on every pixel.

They extracted the pixel-wise feature map on the last attention layer of the CLIP model, then they use the same weight used on the output vector for the CLIP classification task to check every pixel vector in the image and use the label as the segmentation label.

### Multi-Granularity Cross-modal Alignment (MGCA) [4]

In the MGCA model, the images and the texts first pass through the image and text encoders respectively to obtain a series of token representations, and then realize the correspondence of three granularities through the following three modules:

Instance-wise Image-Text Alignment (ITA): Perform instance-level alignment, that is, the contrast loss of image text.

Cross-attention-based Token-wise Alignment (CTA): Token-level alignment based on the cross-attention mechanism. The starting point of this module corresponds to the previous pathological region level, and the CTA module is used to explicitly match and align local medical images and radiology reports. The idea is to perform token-level alignment, using cross-attention to compute a match between the generated visual and textual tokens.

Cross-modal Prototype Alignment (CPA): Both ITA and CTA regard samples from different instances as negative pairs, so it is possible to push away samples with similar semantics, such as pairs of the same disease, in the embedding space. Therefore, the CPA module is for disease-level alignment.

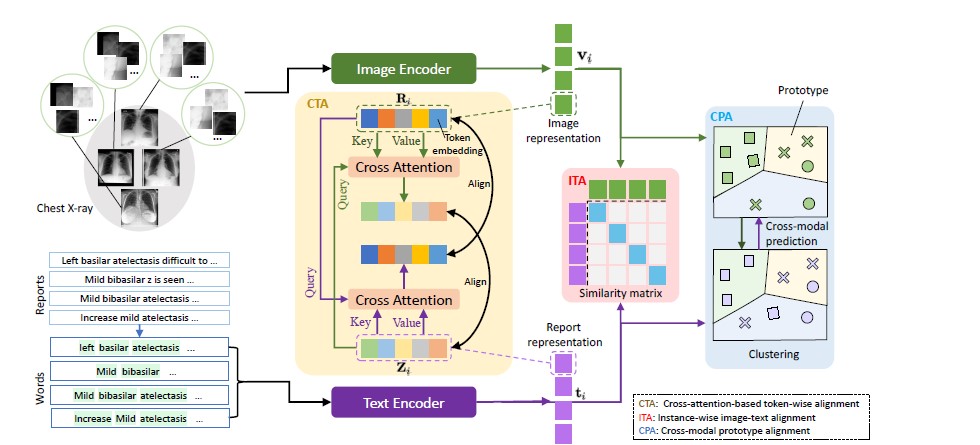


Figure 4: The model structure of MCGA.

## Methodology

Inspired by MaskCLIP [2], we try leveraging the zero-shot capability of CLIP to do the zero-shot segmentation of lesions, e.g., pneumonia, pneumothorax, and pleural effusion, from Chest X-ray images. However, the default setting of MaskCLIP is for zero-shot segmentation of natural images, so it is necessary to modify it from two major perspectives: a pre-trained CLIP model, and prompt engineering. We will briefly explain these two modifications in the following sections.

### Pretrained CLIP Model

MaskCLIP bases its zero-shot segmentation result on the family of CLIP models pre-trained by OpenAI [1] on 400 million natural image-text pairs publicly available on the Internet. Even though the large-scale training dataset may contain a fraction of chest X-ray images with texts that can be directly downloaded from the Internet, the quality of these image-text pairs is usually not high. This is because chest X-ray images are often associated with the privacy of patients, and most chest X-ray datasets are private and require special approval to access. Therefore, the zero-shot capability of OpenAI’s CLIP on chest X-ray images is limited, and we have further verified this conjecture in our experiment part.

To improve the zero-shot capability of CLIP for chest X-ray images, we need to fine-tune OpenAI’s CLIP on chest X-ray image-text pairs. To the best of our knowledge, the largest chest X-ray dataset with image-text pairs is MIMIC-CXR [5]. This dataset contains 377,110 images with 227,835 corresponding radiology reports, where a radiology report may correspond to multiple images. Due to the high cost of fine-tuning OpenAI’s CLIP on the MIMIC-CXR dataset, we directly use the released checkpoints from the SOTA model named CheXzero, which is also fine-tuned on the MIMIC-CXR dataset to do the zero-shot learning. This greatly accelerates our research on the possibility of zero-shot segmentation of X-ray images from pre-training.

However, the training method of CheXzero is the same as OpenAI’s CLIP, and both models mainly focus on zero-shot classification. But the training data size of CheXzero is around 1,000 times smaller than the training data size of OpenAI’s CLIP. So, the zero-shot segmentation result of CXR may not be as satisfactory as that of natural images. So, to accelerate the pixel-wise matching with lesion types, i.e., lesion segmentation, a fine-grained pre-training method is needed to better align both local and global features of CXR images with class embeddings. We plan to follow the MGCA framework to fine-tune OpenAI’s CLIP to improve its zero-shot segmentation capability for Chest X-ray images.

### Prompt Engineering

Prompt engineering is also an important aspect that significantly affects the zero-shot segmentation performance of MaskCLIP. Under the setting of natural image segmentation, MaskCLIP uses 85 prompt templates in total to include most types of photo descriptions, for example, “a cropped photo of the {}.”, “a photo of many {}.”, and “a photo of the clean {}.”. For each class, MaskCLIP will fill all templates with the class name, encode them with the CLIP text encoder into text embeddings respectively, and compute the class embedding by averaging out the text embeddings. A series of experiments performed by MaskCLIP shows the possibility of zero-shot pixel-wise classification, i.e., segmentation, with such class embeddings.

However, under the context of medical imaging, the prompt templates for natural images may not be most suitable for zero-shot segmentation. Therefore, we write a series of prompt templates for CXR segmentation based on the descriptions of CXR images frequently occurred in MIMIC-CXR radiology reports. Some examples are “A limited lateral image with a {}.”, “Single frontal view of the chest with {}.”, and “Upright image of the chest with {}.”. We will show in the experiment part that such prompt templates can indeed increase the segmentation performance.

## Experiments

To explore the potential of CLIP in CXR segmentation, we perform the zero-shot classification and segmentation experiments on OpenAI’s and CheXzero’s CLIP using the framework of MaskCLIP. The details and findings of each experiment are illustrated in the following sections.

### Zero-Shot Classification

We choose the checkpoint named “best\_128\_5e-05\_original\_22000\_0.855.pt” from the GitHub repository of CheXzero (https://github.com/rajpurkarlab/CheXzero) as the CLIP model pre-trained on MIMIC-CXR dataset. The CheXzero model is pre-trained with batch size 128, learning rate 0.00005, and 22000 iterations. The architecture of CheXzero is ViT-B/32, which is the same as OpenAI’s CLIP. It is worth knowing that CheXzero is initialized with pre-trained weights of OpenAI’s CLIP and then fine-tuned on MIMI-CXR.

We run OpenAI’s CLIP and CheXzero on pneumonia, pneumothorax, and lung datasets to evaluate their zero-shot classification capabilities. Pneumonia, pneumothorax, and lungs datasets contain 390, 130, and 200 positive samples respectively. The negative samples of pneumonia and pneumothorax dataset are normal lungs, while the negative samples of the lung dataset are X-ray images of hand bones. We can see from the table that OpenAI’s CLIP performs well in lung classification while poorly in lesion classification. In contrast to OpenAI’s result, CheXzero performs well in pneumonia classification, fairly well in pneumothorax classification, but poorly in lung classification.

The reason behind the classification performance of OpenAI’s CLIP is that the pre-training dataset includes a considerable number of Chest X-rays. So, there is no doubt that it can do well in lung classification. However, due to the lack of radiology reports publicly available on the Internet, OpenAI’s CLIP cannot perform well in the fine-grained classification of Chest X-ray images by lesion types.

The opposite result given by CheXzero is due to its fine-tuning on the MIMIC-CXR dataset. The rich medical diagnoses from MIMIC-CXR enable CheXzero to do fine-grained classification. However, the MIMIC-CXR only contains Chest X-rays, so the capability of zero-shot classification of lungs is lost due to the lack of negative samples.

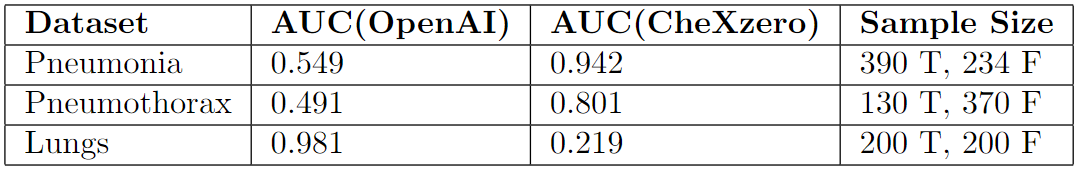


Figure 5: Performance of OpenAI’s CLIP and CheXzero on pneumonia, pneumothorax, and lung classification tasks.

### Zero-Shot Segmentation

To evaluate their zero-shot segmentation capabilities, we run OpenAI’s CLIP and CheXzero on human, lung, and pneumothorax segmentation datasets. Human, lung, and pneumothorax test datasets contain 91, 1000, and 1378 samples respectively. The natural prompt type corresponds to the default prompt templates of MaskCLIP for natural images, while the medical prompt type corresponds to the medical prompt templates created by us. Human/Lung IoU is the IoU for a certain class. mIoU means the average of the IoU of the target class and the IoU of the background.

From the figure below, we can see that OpenAI’s CLIP performs well in lung segmentation while CheXzero performs poorly in this task. This is reasonable since OpenAI’s CLIP can do well in lung classification, but CheXzero lost its capability of lung classification. OpenAI’s CLIP can even perform slightly better when we change to prompt type from natural to medical.

However, it surprises us that CheXzero cannot perform zero-shot segmentation of pneumothorax even though it performs well in pneumothorax classification. It is worth noting that we use the same dataset for pneumothorax classification and segmentation tasks. The result does not conform to the behavior of OpenAI’s CLIP in that the capabilities of zero-shot classification and segmentation for a certain class coexist. We conjecture that the abnormal behavior of CheXzero is that the zero-shot classification is not good enough to induce the zero-shot segmentation capability, since the AUC of CheXzero on the pneumothorax classification task is only 0.801, which is still much smaller than that of OpenAI’s CLIP on lung classification task (0.981).

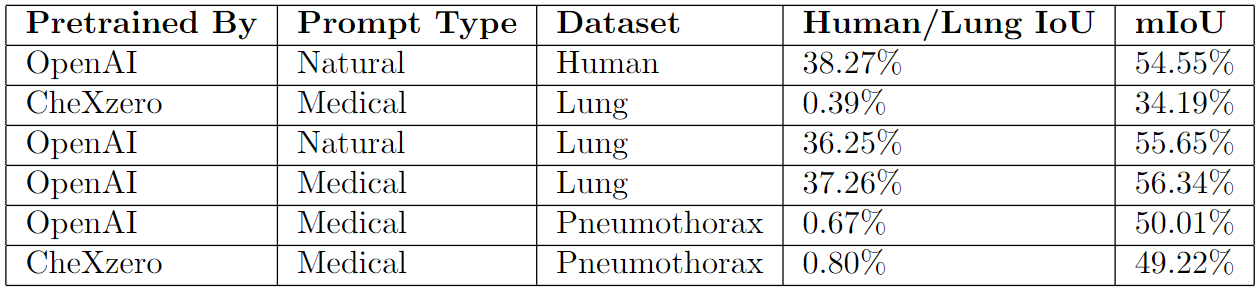


Figure 6: Performance of OpenAI’s CLIP and CheXzero on human, lung, and pneumothorax segmentation tasks.

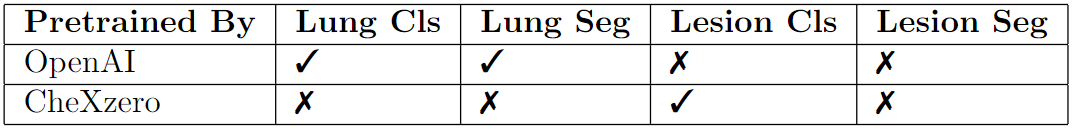


Figure 7: The capabilities of OpenAI’s CLIP and CheXzero on lung classification, lung segmentation, lesion classification, and lesion segmentation tasks.

## Future Work

We will do more testing of CheXzero on lesion segmentation tasks, e.g., pneumonia segmentation, under the framework of MaskCLIP. After that, we will develop our pre-train method to obtain a better feature map by contrastive learning on MIMIC-CXR. We hope our pretraining can increase the capabilities of zero-shot classification and segmentation compared to CheXzero.

# Explainable Artificial Intelligence (XAI)

## Overview

In recent years, machine learning has achieved great success in many fields such as computer vision, natural language processing, and speech recognition. Machine learning models have also been widely applied to some important real-world tasks, such as face recognition, automatic driving, malicious Software detection, intelligent medical analysis, etc.

However, due to the lack of interpretability of machine learning, it suffers from many constraints in terms of application. For example, the DNN model is like a black box, given input and obtained a decision result, but we cannot know exactly the decision basis behind it and whether the decision it makes is reliable. The lack of interpretability will likely pose a serious threat to many DNN-based applications in practical tasks, especially in safety-sensitive tasks. For example, an automatic medical diagnosis model that lacks explainability may bring wrong treatment options to patients, and even seriously threaten the lives of patients.

In data mining and machine learning contexts, interpretability is defined as the ability to explain or present intelligible terms to humans. The interpretability of machine learning models can be generally divided into 2 categories:

* Ante-hoc interpretability

It refers to making the model itself have interpretability by training a model with a simple structure and good interpretability or a self-explanatory model that combines interpretability into a specific model structure.

* Post-hoc interpretability

Refers to explaining trained machine learning models by developing interpretability techniques. According to different interpretation goals and objects, post-hoc interpretability can be divided into global and local interpretability. Global interpretability aims to help people understand the overall logic behind the complex model and the internal working mechanism, and local interpretability aims to help people understand the decision-making process and basis for each input sample of the machine learning model.

In this work, we review some related work in the Ante-hoc and Post-hoc categories and improve some models or algorithms in these papers. Finally, we achieve the main interpretability of the model by capturing some unique medical conditions.

## Literature Survey

The concept of XAI has already been around for a few years, and a lot of related works have been published. After reviewing over 30 papers, we discovered several papers related to our work the most. They are discussed below.

* **Network dissection [6], [7]**

Network dissection is trying to match units in the layer with concepts given the segmentation mask of those concepts. It will first calculate an activation heatmap about the concerns for the model and upsample it to the input size. Then it calculates the overlapping ratio between the activation map and the segmentation map for certain concepts. The concept that exceeds the threshold will be considered as matched. The problem with this method is that it requires the segmentation of concepts in the dataset, which causes an additional workload for data preparation. Additionally, the method cannot match abstract concepts that cannot be visualized, with the units, since it’s hard to offer such a segmentation.

图表

中度可信度描述已自动生成

Figure 8: Sample outputs of network dissection.

* **Causal Analysis [8]**

The Causal analysis divides the network into two parts, the part for figuring out concepts and the part for building the causal relationship between the concepts and the output. The authors have built a logistic regression to match the concept's appearance. The parameters learned from the regression can also be used as a mask of the concept in the feature map. To match abstract concepts with the output, direct effect, and indirect effect are calculated by the modified output from a conditional GAN. Finally, an interpretable model contains regressions and tree functions to surrogate the model.

The problem with this method is that it doesn’t look into the inner part of the convolutional layers. Instead, it uses regressions to simulate them, which may not be capable of imitating the model precisely.

图示

描述已自动生成

Figure 9: Workflow for causal analysis.

* **Prototypical Part Network (ProtoPNet) [9]**

This paper aims to define an interpretable form of image processing (this looks like that), in the same way, that humans think about image classification tasks. This network dissects the image by finding prototypical parts and combines evidence from the prototypes to make a final classification. Besides, it will update the prototypes as the comparison goes on.

许多不同颜色的鸟

描述已自动生成

Figure 10: Prototypical parts and prototype.

* **Neural Prototype Trees (ProtoTree) [10]**

Based on ProtoPNet, this paper explores more on how human thinks in order. In addition to checking if this image looks like that, this model follows the structure of the decision tree and can get the decision made by tracing one path from root to leaf. Each node of the tree can be considered as a binary decision maker and finally this model output the possibility of each classification. This model applies pruning methods without sacrificing accuracy, resulting in a small tree with only 8 prototypes(layers) along a path to classify a bird from 200 species.

图示

中度可信度描述已自动生成

Figure 11: The decision tree of classification of birds.

In conclusion, except for the second one, the rest don’t directly choose the medical images dataset as input. But what we believe is that all of these four technologies or algorithms are applicable to medical images, specifically, MRI images.

## Methodology

Based on the main articles we selected, we divided the whole XAI project into two directions to try: Post-hoc and Ante-hoc. Post-hoc directions mainly combine the models and algorithms of network dissection and causal analysis. Ante-hoc mainly applies the idea of ProtoTree.

### Post-hoc

#### Train models and Use hooks to get feature map.

We build a 3D residual network based on ResNet-34, in which the residual block and normal convolution block appear alternatively. This model is based on a 3D Magnetic resonance imaging (MRI) MRI dataset from Duke University. Before training the model, the images in this dataset are normalized and introduced from the .dcm file into the tensors by the pydicom library. Besides, since the dataset is relatively small and only contains 922 patients’ MRI image data, we did the data augmentation method by rotation, mirror, crop, and translation. Finally, this ResNet model can reach an accuracy of 0.85~0.88.

Feature map, or activation map, is the output of a certain layer of the model every time after the forward() has been implemented. We can use the hook() function provided by PyTorch to memorize this output but don’t modify the input or output of the layers while we train the model. This feature map to some extent mains the learning outcome of this AI model and is vital for us to investigate how this model learns MRI images.

#### Using regression methods (Linear, Logistic, Lasso) to match feature map and clinical concepts.

Based on the 3D feature map obtained in 3.3.1.1 and the logistic regression method in Casual Analysis [8], we matched some of the binary professional clinic labels from the dataset with some layers of the model. Besides, we also consider other regression methods such as Linear SGD, and Lasso. After the regression, we got a 3D weight matrix, and we choose the largest 2 percent as the threshold. The results show that Lasso regression cannot successfully distinguish different labels since all thresholds are close to zero and linear SGD is super sensitive to the learning method and always needs manual adjustment. After comparison, the linear regression method’s outcome demonstrates great intuitive results and fits reality the most, as shown in Figure 12. Normally, we believe that the ER and HER2 labels are highly related, so we can notice that in Figure 12, the high-weight area of ER and HER2 are highly overlapped.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated Chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 12: The weight matrix of 5 chosen labels by Linear regression.

### Ante-hoc

The ProtoTree is a prototype-based self-interpretable method, as the method is going to use prototypes and a soft decision tree, where prototypes uncover what has been learned from the CNN networks, and the decision tree is that people can easily sum up the probability from the leaves to the root. Use the output of convolutional layers as features that can be matched with learnable prototypes representing attributes of the training set. Each prototype is correlated with one node in the decision tree, where the distances between features and prototypes are computed. The distance determines the weight of the leaf and right subtrees. The leaf nodes in the decision tree will build a small classification, and the final decision is based on the weighted average of leaves.

图示

描述已自动生成

Figure 13: A sample workflow of ProtoTree.

Additionally, to solve the overfitting problem during the training, we have introduced the Fidelity Similarity [11] to features, so that the features tend to be distinct and represent different clinical concepts. The Fidelity Similarity is chosen as features may not be vectors but matrices, whereas other functions like cosine similarity cannot be used.

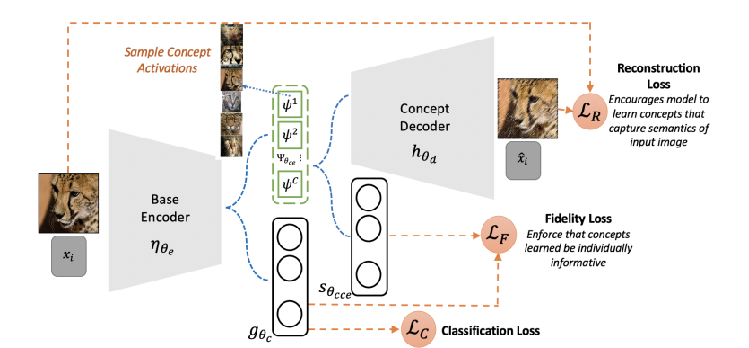


Figure 14: An overview of the proposed framework (concept activations denote images that maximally active each concept)

## Future Plan

### Post-hoc

Instead of computing certain labels in 2.3.1.2, in the future, we need to compute the weight matrix of all the clinic labels contained in the Duke dataset. Then, use the weight of regression we obtained in 2.3.1.2 to represent the importance of layers to concepts and filter the most activated concept for layers in the model.

### Ante-hoc

Combine the features figured from the Post-hoc methods to the ProtoTree so that the model can learn those important concepts by itself, which means we can understand the order of decisions made by the model from the root to lead along the decision tree.

# References

[1] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *Arxiv*, 2021, doi: 10.48550/arxiv.2103.00020.

[2] C. Zhou, C. C. Loy, and B. Dai, “Extract Free Dense Labels from CLIP,” *Arxiv*, 2021, doi: 10.48550/arxiv.2112.01071.

[3] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning,” *Nat Biomed Eng*, vol. 6, no. 12, pp. 1399–1406, 2022, doi: 10.1038/s41551-022-00936-9.

[4] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, “Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning,” *Arxiv*, 2022, doi: 10.48550/arxiv.2210.06044.

[5] A. E. W. Johnson *et al.*, “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs,” *Arxiv*, 2019, doi: 10.48550/arxiv.1901.07042.

[6] B. Zhou, D. Bau, A. Oliva, and A. Torralba, “Interpreting Deep Visual Representations via Network Dissection,” *Ieee T Pattern Anal*, vol. 41, no. 9, pp. 2131–2145, 2017, doi: 10.1109/tpami.2018.2858759.

[7] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, “Understanding the role of individual units in a deep neural network,” *Proc National Acad Sci*, vol. 117, no. 48, pp. 30071–30078, 2020, doi: 10.1073/pnas.1907375117.

[8] S. Singla, S. Wallace, S. Triantafillou, and K. Batmanghelich, “Using Causal Analysis for Conceptual Deep Learning Explanation,” *Arxiv*, 2021, doi: 10.48550/arxiv.2107.06098.

[9] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This Looks Like That: Deep Learning for Interpretable Image Recognition,” *Arxiv*, 2018, doi: 10.48550/arxiv.1806.10574.

[10] M. Nauta, R. van Bree, and C. Seifert, “Neural Prototype Trees for Interpretable Fine-grained Image Recognition,” *2021 Ieee Cvf Conf Comput Vis Pattern Recognit Cvpr*, vol. 00, pp. 14928–14938, 2021, doi: 10.1109/cvpr46437.2021.01469.

[11] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, “A Framework for Learning Ante-hoc Explainable Models via Concepts,” *2022 Ieee Cvf Conf Comput Vis Pattern Recognit Cvpr*, vol. 00, pp. 10276–10285, 2022, doi: 10.1109/cvpr52688.2022.01004.

# Appendix

## The 1st  Project Meeting

**Date**: Jan 17, 2023

**Time**: 4:30 pm

**Place**: ZOOM

**Present**: LU Weiqi; XU Mingshi; Luo Luyang.

1. Reviewed the progress of MaskCLIP+ experiments on CXR segmentation tasks.
2. Planned to investigate the zero-shot segmentation capability of CheXzero under the framework of MaskCLIP.

## The 2nd Project Meeting

**Date**: Feb 6, 2023

**Time**: 6:30 pm

**Place**: F2F

**Present**: LU Weiqi; XU Mingshi; Luo Luyang.

1. Discussed the reasons why CheXzero could not perform lung segmentation under the framework of MaskCLIP.
2. Planned to study the correlation between the zero-shot classification and segmentation capabilities of CheXzero.

## The 3rd Project Meeting

**Date**: Feb 11, 2023

**Time**: 8:30 pm

**Place**: Zoom

**Present**: LU Weiqi; XU Mingshi; CHEN, Siyu; ZHOU Taichang; Prof. CHEN Hao; Luo Luyang.

1. Reported the results of experiments on CheXzero and OpenAI’s CLIP for lesion classification and segmentation tasks.
2. Decided to fine-tune OpenAI’s CLIP on the MIMIC-CXR dataset to obtain the zero-shot lesion segmentation capability.