# Zero-Shot Object Detection for Chest X-Rays

*Ellie Talius, Ruhi Sayana*

*Department of Computer Science, Stanford University*

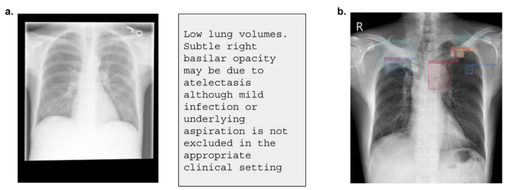
## Introduction

- Using deep learning for object detection in medical imaging is challenging
  - Requires **large amounts of labelled data**
  - Expensive, time-consuming to annotate
- OpenAI's CLIP model uses contrastive learning to build associations between images and text → can be used to understand medical image and radiology report pairs
  - Promising for zero-shot object detection
- **No previous research on performing zero-shot object detection on chest x-ray images**

## Problem Statement

- Input and Output: **Chest x-ray image → labelled bounding boxes** containing pathologies present in chest x-ray image
- We evaluate using mean Average Precision (mAP)

## Datasets

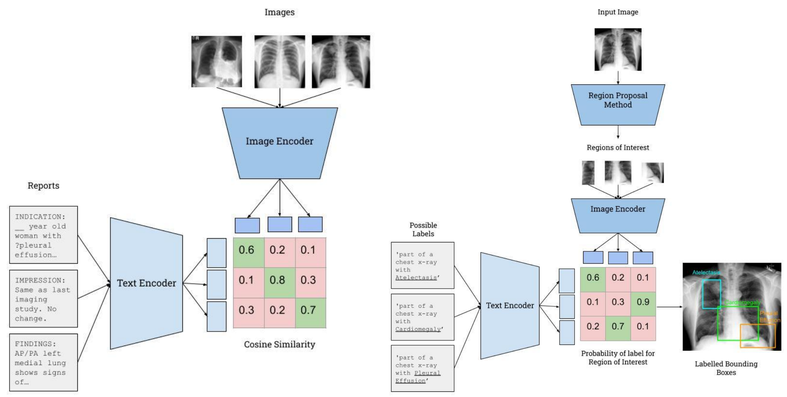



- **Training:** MIMIC-CXR (300,000 images and reports)
- **Test:** Kaggle VinBigData Chest X-ray Abnormalities detection (3,000 images)
- Preprocessing: Resizing to 224 x 224, and 3 data augmentation set ups (random crop, random horizontal flip, full transformation pipeline)

## References

**[1]** Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

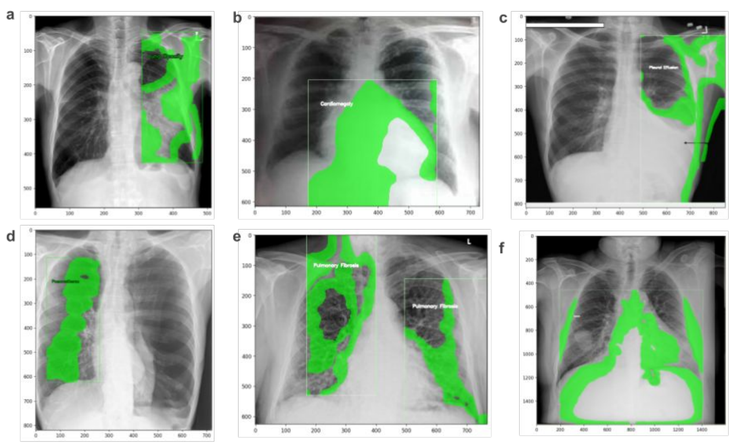
## Methods




- **Training:** Train CLIP architecture on a mini-batch of image - text pairs
  - Model learns to **maximize cosine similarity between correct pairs** and minimize cosine similarity of incorrect pairs
- **Evaluation**: Use learned image and text encoders to predict labelled bounding boxes
  - Use superpixel segmentation or selective search to generate RoI's
  - Encode image crops with image encoder and labels with text encoder
  - Return boxes with similarity greater than threshold (t=0.5)

## Experiments

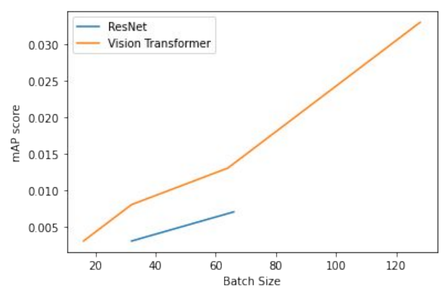
### Pathology Detection




### Zero-Shot vs. Fully Supervised

| Supervision | Model | mAP Score |
|---|---|---|
| Fully Supervised | Detectron 2 | **0.235** |
| Zero-Shot | CLIP-Single-No Train | 0.022 |
| | CLIP-Multiple-No Train | 0.017 |
| | Best CLIP | **0.045** |

### Batch Size




### Text Query

| Prompts | mAP Score |
|---|---|
| 1) a chest x-ray with {} | **0.036** |
| 2) part of a chest x-ray with {} | 0.025 |
| 3) {} | **0.035** |
| 4) {} present in a chest x-ray | **0.031** |
| 5) crop of a chest x-ray showing {} | 0.027 |

### Data Augmentation

| Data Augmentation | mAP Score |
|---|---|
| Random Crop | 0.035 |
| Random Flip | 0.025 |
| All Transforms | 0.045 |

### Region Proposal Selection

| Region Proposal Method | mAP Score |
|---|---|
| Superpixel Segmentation | 0.033 |
| Selective Search | 0.042 |

## Analysis

- Fully supervised models outperform zero-shot learning for pathology detection in chest x-rays
- Architecture and training improvements:
  - Using a **vision transformer backbone** on CLIP leads to better performance compared to ResNet, which is used in SOTA object detection models.
  - **Increasing batch size** improves the performance of the model with either backbone.
- Using a **series of data augmentations** to increase the size of the training set leads to better model performance.
  - Applying single transforms does improve model generalizability
- **General text queries** paired with test images improve model performance
- Using **selective search** to select regions for pathology detection improves performance over superpixel segmentation
  - Selective search optimizes for more features in the image, leading to better expressivity
  - Significant time cost

## Conclusions + Future Work

- We are able to perform **zero-shot object detection** for pathologies in chest x-ray images
- Key takeaway: zero-shot object detection can greatly benefit the medical field by helping automate and verify chest x-ray diagnosis **without the need for expensive labelled data** for training
- Our results are currently limited by the amount of compute (restricting batch size to a maximum of 128) and the use of a slow region proposal and object detection pipeline
- **Future steps:**
  - Developing a region proposal method that operates on the model i**mage embeddings instead of raw pixels**
  - Switching to a **queue-based method for training** to avoid the need for a large batch size