Original Research Article

Investigating Text-Guided Cross-Region Feature Alignment for Multimodal Disease Localization in Chest X-Ray Images

Sourya Potti

Dougherty Valley High School, 10550 Albion Rd, San Ramon, California, 94582, United States

ABSTRACT

Deep learning object detection techniques have been extensively applied to lung- and chest-related healthcare applications. Recent advances in text-guided object detection techniques have led to substantial performance improvements over image-based detection techniques. While such models employing traditional region–text similarity have been explored for detecting abnormalities in chest X-rays, the efficacy of models leveraging the concept of region-region similarity in this domain remains largely unexamined. Although such architectures have demonstrated effectiveness in natural scene contexts, their applicability to chest X-rays has been restricted due to the inherent challenges of the medical object detection task. This gap prompts the question of whether chest X-ray-based disease detection can be performed by training cross-region feature alignment architectures. In this study, this question is addressed by systematically investigating a text-guided region-region similarity based object detection architecture, dubbed CXR-CoDet. For this, this work investigates multiple training hyperparameter configurations (with varying learning rate, batch size, number of training iterations), number of support images needed for co-occurrence computation, different pretrained weights, different granularity of disease descriptions, and incorporation of medical information through the text encoder. This work also underscores the limitations of region-region similarity-based object detection architectures, particularly applied in medical imaging, and provides recommendations for improvements. Code is available at: https://github.com/souryatech/TGCRFA-CXR.git

Keywords: deep learning; object detection; chest X-Ray; medical imaging; text-guided object detection; disease localization

INTRODUCTION

Object detection has been a fundamental component of machine learning applications (1-4). Specifically, object detection has allowed for an in-depth understanding of visual scenes, with its ability to localize bounding boxes on classes on which object detection models have been trained (5). Such a model also allows for the comprehension of medical scenes, such as X-ray images, leading it to localize signs of specific diseases that it is trained on (1, 6). This is especially useful in diagnosing chest disease, since X-rays are crucial in that standpoint of diagnosis (1).

However, current research on Chest X-ray (CXR) detection has embodied object detection models that only work on the disease categories that they are trained on. Recently, research on object detection models has gone to the next level. Specifically, region-wise

Corresponding author: Sourya Potti, E-mail: sourya.tech.8@gmail.com.
Copyright: © 2025 Sourya Potti. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Accepted October 23, 2025
https://doi.org/10.70251/HYJR2348.362839

correspondence models have allowed for the diverse connection between the understanding of language concepts and their localization on visual scenes (2, 4). With such a model being prevalent, its implementation on Chest X-ray modalities opens a whole new door of possibilities for medical advancement, such as the model's detection of abnormalities using vast medical knowledge learned from text guidance (2, 4, 7, 8).

Currently, region-region similarity models are primarily trained on natural scenes and lack training on medical X-ray scenes, leading them to not have an in-depth understanding of medical language and localization (2). Therefore, the prospect of implementing these models on Chest X-ray modalities raises the question: Can *co-occurrence-guided object detection models achieve high performance in detecting abnormalities on chest X-ray images?* This paper investigates this question, specifically by training a co-occurrence-based model, utilizing region-region similarity, where support images are retrieved to train a current image for improving the correspondence between text embeddings and visual features, on a Chest X-ray detection dataset called VinbigdataCXR (9). For this, this work tunes specific hyperparameters of the model, which include optimizing the model's learning rate, number of iterations, replacing the regular CLIP-based text encoder (10) with a biomedical variant, and optimizing the number of support images utilized in region-region similarity.

In summary, this paper explores the performance gap between cross-region feature alignment models trained on chest X-ray (CXR) modalities and those trained on natural scene datasets. It further seeks to enhance the effectiveness of CXR-based models and examines whether their performance can be raised to meet or approach standard benchmark levels.

LITERATURE REVIEW

General Object Detection Advances

Co-Occurrence Guided Region-Word Alignment for Open-Vocabulary Object Detection (CoDet) represents a crucial shift in Open Vocabulary Object Detection (OVOD) as it introduces the concept of region-region similarity (2). Unlike traditional VLMs that generate basic region-text pairs and are pre-trained only on specific detection features, which decreases their nuance, CoDet juxtaposes region-text proposals with pre-existing image-text pairs through its co-occurrence alignment, which further decreases its ambiguity in aligning region-text pairs (2). This culminates in stronger weights in pairs that are more similar to their previous correspondences than pairs that contrast more. Ultimately, CoD*et al*lows for a more nuanced approach toward region-text alignment than traditional VLMs, which rely solely on pre-trained data; it demonstrates this through similarity between regions that strengthen its confidence in common region-text pairs, therefore solidifying its accuracy on repeatedly encountered correspondences. A new type of model, AggDET (4), additionally builds upon CoDet's introduction of region-region similarity with region-prototype similarity, which does not require much training data (4). This model is primarily based on the co-occurrence model trained on natural scenes from CoDet (2).

Frozen CLIP-based DETR with Language Model Instruction (LaMI-DETR) aims to improve region-word alignment models through enhancing image and text encoders while improving their understanding from text guidance. Specifically, LaMI-DETR proposes a frozen Contrastive Language-Image Pretraining (CLIP) image encoder while utilizing embeddings from the CLIP text encoder for weights (10, 11). This both leverages the capabilities of VLMs while preventing overfitting in base categories. LaMI-DETR also proposes the utilization of Large Language Models (LLMs) to better process text encodings. Specifically, it utilizes GPT-3.5 to generate visual descriptions with given text guidance, which then leads to the descriptions' grouping from a T5 model into categories with visual similarities (11). While CoDet (2) and AggDet (4) focus on enhancing vision-language alignment to get better results, LaMI-DETR instead focuses on improving its processing of text guidance, which significantly decreases the possibility of overfitting and improves inter-category relationships (11).

Zero-shot object detection (ZSD) works by leveraging a language feature space to be able to detect unseen objects in a collection of region proposals. It does this by projecting the features of each region to the main text-embedding space, GloVe (12) in this case, and utilizes the word embeddings that pair with regions, as weights (13). ZSD is a more effective method due to its flexibility in recognizing unseen objects in a space. However, its limitations include the restriction of training samples that it receives to come from a limited number of seen classes, which leads to less effective and accurate training. While others try to solve this with the Generative Adversarial Network (GAN) (14), with some luck, there is still a lack of performance.

Chest X-Ray Detection Efforts

Zero Shot Detection for Chest X-Ray Modalities (ZSD-CXR) represents a key transformation in medical object detection techniques since it applies a richer understanding of the relationship between disease classes and image features to chest x-ray modalities (7). Specifically, this work leverages zero-shot detection on chest X-ray modalities to detect classes that were not specified during the training procedure, allowing for a much broader range of diseases that the model can detect. This also accounts for a significant limitation that has been pervasive in chest X-ray object detection, which is the limited amount of data available. This model utilizes a region-text framework for their architecture, creating a cosine similarity matrix between text encodings and region proposals. It was also trained on the MIMIC CXR (15) dataset and tested on the Vinbigdata Chest X-Ray Abnormalities dataset (9).

Weakly Supervised Object Detection in Chest X-Rays with Differentiable ROI Proposal Networks and Soft ROI Pooling (WSRPN) represents a significant breakthrough in chest X-Ray image analysis techniques due to the leveraging of weakly supervised object detection for a more specialized understanding of chest X-Ray disease localization. Specifically, this model utilizes WSD to not require the necessity of future supervision for new instances of chest X-ray datasets (3). This work addresses the issue that such a form of multiple instance learning is not applied to Chest X-ray modalities. This model works by generating bounding box proposals on the fly by utilizing a Region of Interest attention module. WSRPN (3) additionally trains on the CXR8 dataset (16).

Importance of This Work with respect to Literature

Region–region similarity–based techniques have recently gained traction in the object detection domain for natural scenes, as demonstrated by advances in general object detection frameworks (2, 4). However, a review of chest X-ray–based studies (3, 7) reveals that such approaches have not yet been adopted in the context of chest abnormality detection. This work bridges that gap by introducing and adapting the concept of region–region similarity to the medical imaging domain, specifically for chest abnormality detection.

METHODS AND MATERIALS

Dataset

This study utilized the VinBigdata Chest X-Ray Abnormalities dataset (VinDr CXR), constituting 15000 chest X Ray images with a 1024 × 1024 scale (9). Specifically, this dataset was also partitioned into a 94:6 split, where 14100 images were designated for training, and 900 images were designated for testing. This dataset is highly imbalanced by a ratio of 2868:215 (i.e., 13.34) as shown in the class-distribution plot in Figure 1. Besides, this dataset is particularly challenging as

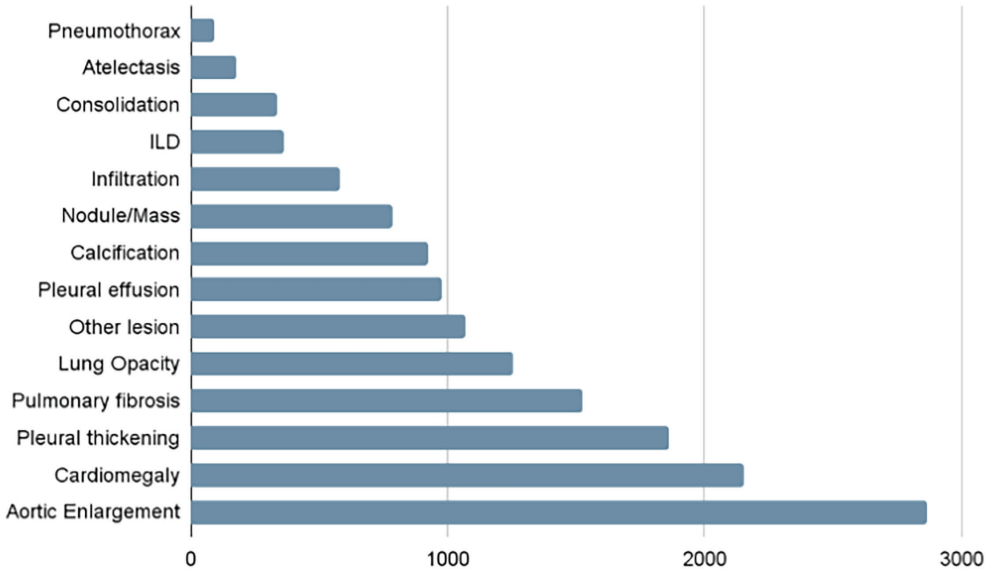

Figure 1. Represents the class distribution of the training split of the VinDr CXR Dataset (9).

it has a smaller number of images compared to larger datasets such as MIMIC-CXR (15), thereby posing a challenge to the model's performance on Chest X-ray modalities when finetuning it on this dataset.

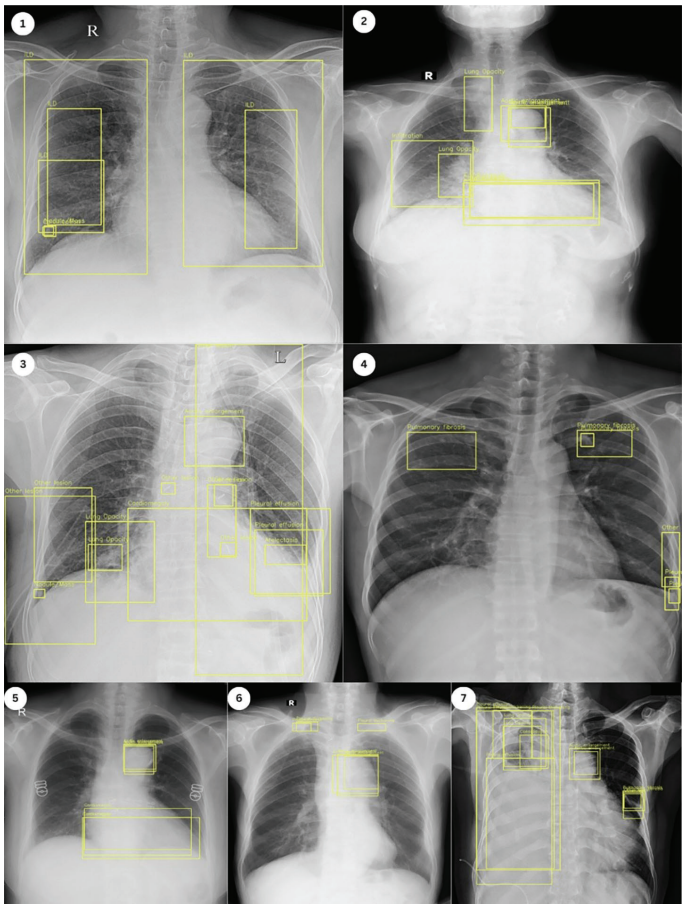

Figure 2. Demonstrates ground truth bounding box proposals of samples from the VinDr CXR dataset. The bounding boxes of all classes are included at least once in this figure. Image 1 represents a chest x-ray with bounding boxes for ILD, Calcification, and Nodule/Mass. Image 2 represents a chest x-ray with bounding boxes for Lung Opacity, Aortic Enlargement, Cardiomegaly, and Infiltration. Image 3 represents a chest x-ray with bounding boxes for Aortic Enlargement, Atelectasis, Pleural Effusion, Lung Opacity, Other Lesion, and Cardiomegaly. Image 4 represents a chest x-ray with bounding boxes for Pulmonary Fibrosis and Other Lesion. Image 5 represents a chest x-ray with bounding boxes for Cardiomegaly and Aortic Enlargement. Image 6 represents a chest x-ray with bounding boxes for Pleural Thickening, Pneumothorax, and Aortic Enlargement. Image 7 represents a chest x-ray with bounding boxes for Pleural Effusion, Consolidation, Pulmonary Fibrosis, Lung Opacity, Pleural Thickening, and Aortic Enlargement.

Co-Occurrence Guided Region Text Alignment in CXR-CoDet

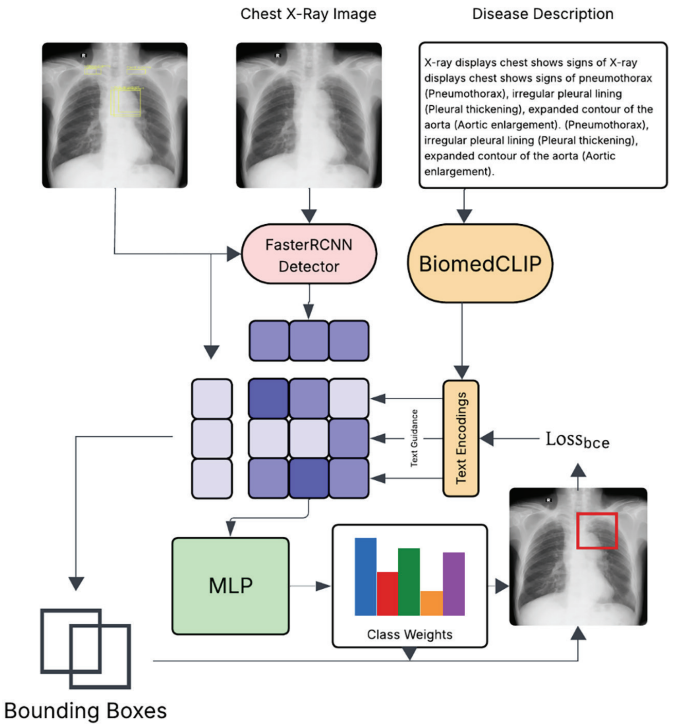

Figure 3. Shows a diagram of the CXR-CoDet architecture. Specifically, the Chest X-ray image is passed to a FasterRCNN (5) detector, and the disease descriptions are passed to BiomedCLIP (17). These encodings and bounding boxes are passed into a text-guided region-region similarity mechanism, which is then converted into class weights utilizing an MLP.

Specifically, this model includes a text encoder, a region proposal network, and an image encoder; it utilizes the mechanism of region-region similarity to create a similarity matrix that is then passed into an MLP for predicting final class weights. The different components of the model are elaborated in Figure 3 and below:

Text Encoder

This work's text encoder encodes the disease descriptions or findings corresponding to each input CXR image. This study initially utilized a pretrained CLIP text encoder (10). For this work, CLIP (10) has been kept frozen since the aim was to utilize an already trained text encoder. However, to integrate further medical knowledge in the text encoding, CLIP was replaced by a pretrained Biomedical CLIP (BiomedCLIP)

(17). The model was kept frozen to maintain a solidified understanding of chest-related terminology.

FasterRCNN Detector

This detector takes each input image and outputs a set of bounding boxes. FasterRCNN consists of a Region Proposal Network (RPN) that proposes regions of interest (ROIs) and a convolutional backbone that extracts features from these ROIs (5). The FasterRCNN (5) utilized a frozen ResNet50 convolutional backbone (18).

Similarity Estimator

In the model's co-occurrence region similarity estimator, a dot product is performed between the proposal of the query image and the average of the proposals of the support images. This product is then weighted by the encodings of the corresponding caption. This weighting allows for text guidance on the output of the model. With this effect of text on the output of the model, the model's performance can be changed through the optimization of what kinds of captions are fed in training. This estimator outputs a final similarity matrix with the location of higher values indicating regions of higher probability (2).

Multi-Layer Perceptron (MLP)

At the end, this model has a two-layer MLP that converts the aforementioned similarity matrix to output class probabilities.

RESULTS

Training & Implementation

This project utilizes a baseline model called CXR-CoDet, which is originally trained on natural scenes utilizing co-occurrence (2). However, since the model is being fine-tuned on a dataset (VinDr CXR (9)) that it was not made for, since the model was originally trained on natural scenes, it is not guaranteed that CoDet's training parameters, such as the number of iterations, the batch size, and the learning rate, will improve the model's performance.

A learning rate of 0.0005, a batch size of 32, and 90000 iterations have been initially utilized. The network utilizes a custom multiplier of 0.1 for the last two linear layers in the MLP. A learning rate scheduler that decays the current learning rate after both 60000 and 80000 iterations is also employed. Any separate oversampling techniques for the rare classes in the dataset are not applied. The annotations directly passed into the dataset are disease captions or findings and corresponding bounding box coordinates. 8 workers and a concept group size of 1 have been utilized for the data loading process. Additionally, additional co-occurrence feature augmentation was applied in region proposals, and also the number of top proposals for weakly-supervised feature alignment was kept to 32.

The model utilizes a ResNet50 backbone in its FastRCNN architecture, which utilizes batch normalization layers for its regularization. In the model architecture's multi-layer perceptron (MLP), two linear layers were utilized, with a ReLU activation following the first one and a flatten and softmax layer following the second one.

This model additionally has a data augmentation procedure based on randomly resizing the image as long as its shortest edge is above 800 pixels.

The loss function utilized for this model is Binary Cross-Entropy, a fundamental loss function utilized for similar object detection tasks entailing region-region similarity (2).

As observed in Figure 4, the loss appears to strictly converge at around 100000 iterations, with a marginal and sharp drop at around 80000 iterations. By convergence, the loss has dropped by around 225%.

Figure 4. Demonstrates the relation between image loss and the number of iterations based on a singular training procedure of the model on the VinDrCXR dataset. For this case, image loss has been shown, in which image loss is the discrepancy between the model's predicted output and the ground truth, since there is increased difficulty in understanding convergence with total loss, given the sharper changes in other losses. Additionally, image loss is the fundamental and most important loss since it relates the output to the ground truth.

Evaluation Metrics

The evaluation metrics that have been utilized for this work are AP50, APm, and APl. These metrics have been fundamentally utilized in traditional object detection tasks (19, 20, 21). These works are followed to select the aforementioned metrics, as elaborated below:

AP_{50}

Small abnormalities demand far greater localization precision than larger ones. Consequently, model predictions are more likely to diverge from the ground truth for smaller findings due to the higher spatial accuracy required. Employing an Intersection over Union (IoU) threshold of 0.5 provides a balanced and stable evaluation of bounding-box performance, reducing the disproportionate penalty that small abnormalities would otherwise incur.

AP_m and AP_l

For clinical applicability, the model must perform robustly across abnormalities of all sizes. Small and medium-sized findings are particularly challenging—and clinically critical—since they are easier to overlook, whereas larger abnormalities are generally easier to detect but often demand urgent intervention. Differentiating AP metrics by abnormality size enables a more granular assessment of model performance, highlighting which size ranges the model is most effective at detecting.

Performance on natural scenes

This work first performs continual fine-tuning of CoDet on the COCO dataset and presents the results of the experiment in Table 1.

Table 1. Performance of fine-tuned CoDet on the COCO dataset (in %)

AP50	APm	APl
46.787	32.128	41.053

Performance Analysis with CXR dataset

Next, different model parameters are optimized, including the learning rate, number of training iterations, batch size, variation in wording utilized in captions, and initial weights that the model is pretrained on. With the co-occurrence architecture of CXR-CoDet, the optimal number of support images based on a query image is also determined for a better AP50 score. Besides, this work hypothesizes that changing the weights in the similarity matrix product via

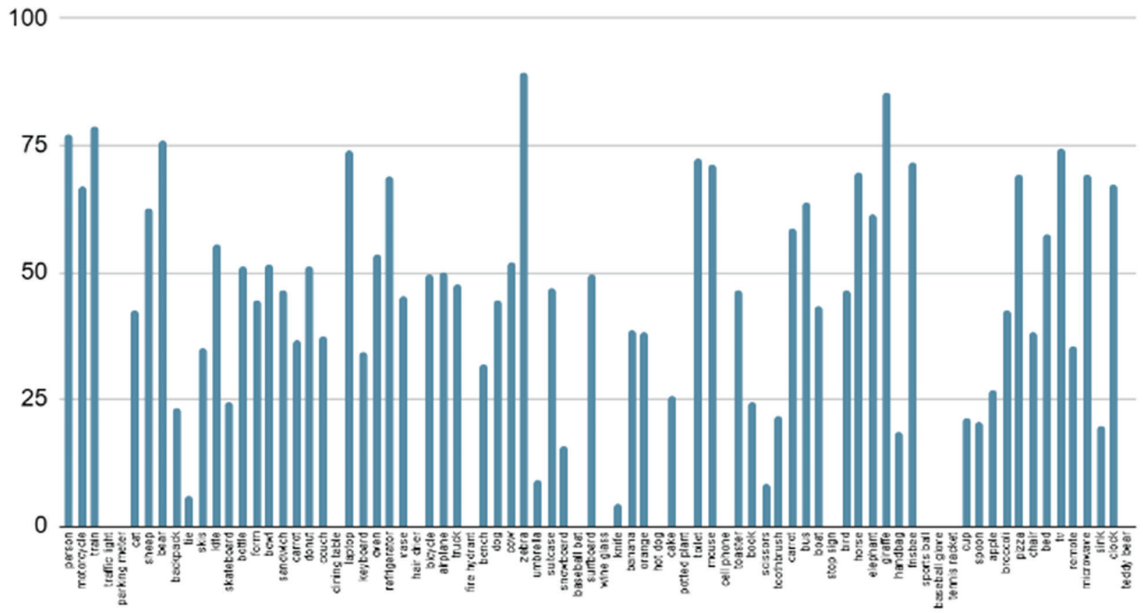

Figure 5. Shows CoDet's performance on each class of the COCO dataset in AP50 metrics. The CoDet model especially performs extremely well on the zebra, person, train, bear, and giraffe classes with an AP50 score over 75%. CoDet additionally achieves an AP50 score between 50% and 75% for classes such as sheep, toilet, mouse, and bed. CoD*et al*so achieves an extremely low AP50 score for the tie, umbrella, knife, cake, and scissors classes.

manipulating the text guidance (through the captions for each image passed during training) can potentially optimize the similarity matrix passed to the MLP. This raises the questions of whether varying the learning rate, making the image caption text more descriptive, changing the CLIP model to BiomedCLIP, or modifying the number of support images fed during training can improve the model performance. The subsections below address these questions by comparing the AP50, APm, and API scores of each configuration along with the individual performance of the models on each disease.

Impact of varying training parameters

In this subsection, this work analyzes the impact of several training parameters, namely: (a) batch size, (b) number of iterations, and (c) learning rate.

Table 2. Performance of CXR-CoDet with varying training hyperparameters (in %)

Hyperparametric values	AP50	APm	APl
Batch Size: 32	0.724	0.275	0.158
Batch Size: 64	0.594	0.054	0.151
Num. Iterations: 90000	0.724	0.275	0.158
Num. Iterations: 130000	0.742	0.310	0.186
Learning Rate: 0.0005	0.650	0.170	0.196
Learning Rate: 0.001	0.724	0.275	0.158

Variation of batch size. Table 2 shows the change in AP50, APm, and APl metrics with batch sizes: 32 and 64. This increase in batch size shows a lower AP50 score, a lower APm score, and a slightly lower APl score. This demonstrates that the model does not train at a level deep enough for the connection between visual features in the chest X-Ray and the encodings of the corresponding captions to improve. This would lead to an overall decrease in the model's performance on detecting lesions from a chest X-ray.

Variation of number of iterations. Table 2 shows the change in AP50, APm, and APl metrics with an increase in the number of iterations from 90000 to 130000. The AP50, APm, and APl metrics increase with a higher number of iterations. However, the AP50 and APl scores increase only by a small decimal amount from 0.03-0.04, which shows that the model has potentially converged, and increasing the number of iterations further might not be very effective in increasing the model performance. Therefore, this would only marginally improve the model's lesion-localization performance.

Variation of learning rate. Table 2 shows the change in AP50, APm, and APl metrics with a change in the learning rate from 0.0005 to 0.001. While the AP50 and APm scores increase, the APl score instead decreases marginally. This demonstrates that increasing the learning rate further will detrimentally affect the model performance, as it will become harder for the model to fully understand large features in the CXR dataset because of the decrease in APl score, eventually being more than marginal, leading to a lack of performance in detecting abnormalities that require much more urgent attention. While this decrease in performance is not noticeable in Figure 6 for such large abnormalities, with a much higher learning rate, the model will eventually have a decrease in performance for large lesions such as Cardiomegaly, given a larger change in the overall AP score for large abnormalities.

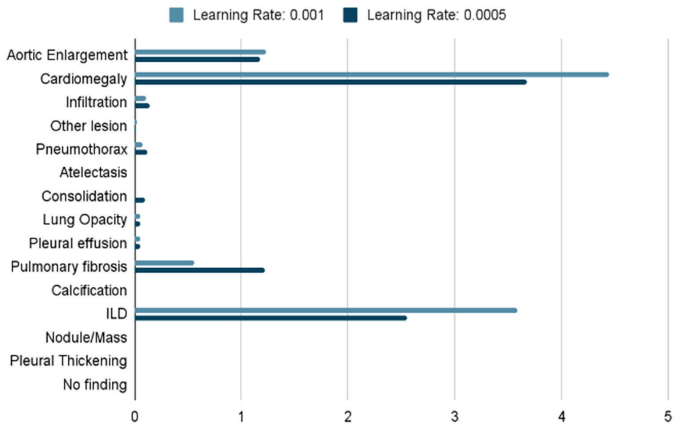

Figure 6. Shows an increase in performance for the majority classes with a learning rate of 0.001, such as Aortic Enlargement, Cardiomegaly, and ILD. However, there has been a decrease in performance in classes such as Pneumothorax and Pulmonary fibrosis, possibly due to a need for further learning of those classes, which can be achieved with a smaller learning rate.

Impact of initial weights utilized

Table 3 reports the changes in AP_{50}, AP_m, and AP_l metrics when the model's pretrained weights are switched from CoDet to Detic for the backbone, region proposal network (RPN), and detector, while

resetting the weights of the region–region similarity mechanism and the MLP. The increases in AP_{50} and AP_l indicate improved performance in detecting larger abnormalities, as reflected by the enhanced detection of large lesions such as Cardiomegaly (Figure 7). This suggests that the model becomes more effective at identifying clinically urgent, large-scale abnormalities.

However, this improvement comes at the cost of a decline in AP_m, suggesting reduced performance for smaller or normal-sized lesions, such as Pulmonary Fibrosis (Figure 7). These findings imply that while the model gains strength in detecting prominent abnormalities, it sacrifices precision for smaller findings. This reduction in precision likely arises from the inherently greater localization sensitivity required for small objects, where even minor bounding-box inaccuracies can significantly lower the Intersection over Union (IoU) with the ground truth—resulting in a lower AP_m score.

Table 3. Performance of CXR-CoDet with varying initial weights (in %)

Pretrained Weights	AP50	APm	APl
CoDet	0.724	0.275	0.158
Detic (Backbone & Detector) & Reset Weights (Region-Region Similarity & MLP)	0.949	0.032	0.274

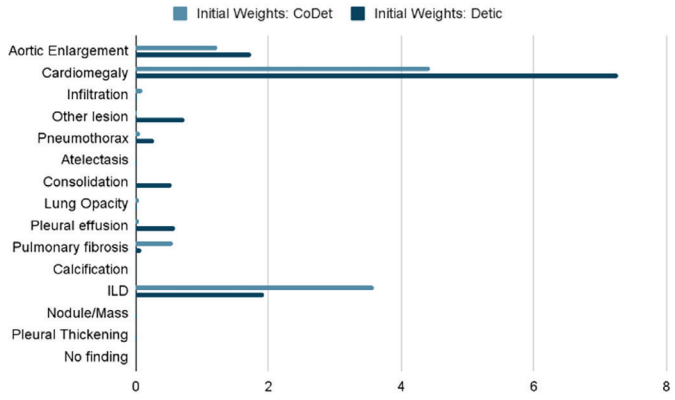

Figure 7. A closer look at Figure 6 shows a further increase in the AP50 score for major classes such as Cardiomegaly and Aortic Enlargement, and a small increase for most minor classes. However, there has also been a decrease in performance for classes such as ILD and Pulmonary fibrosis.

Impact of varying disease descriptions

Table 4 shows the difference between more elaborate disease descriptions and less diverse & less elaborative disease descriptions. Unlike unelaborate descriptions, elaborative disease descriptions have more diversified vocabulary and sentence structure (e.g. "opacity seen in lung areas" having different vocabulary and structure than "hazy regions in lungs noted" as shown in Table 4) along with more in-depth description on the visual features associated with each abnormality in the X-Ray, (e.g. "spread out haziness in the lungs (Infiltration)" instead of "Infiltration", demonstrated in Table 4).

Table 4. Examples of elaborate and unelaborate disease descriptions

Description	Example
Elaborative Description 1	"Chest X-ray shows spread-out haziness in the lungs (Infiltration), abnormal lump found on X-ray (Nodule/Mass), hazy regions in lungs noted (Lung Opacity), consolidated patches in lungs (Consolidation)."
Elaborative Description 2	"Scan indicates opacity seen in lung areas (Lung Opacity), evidence of lung consolidation (Consolidation), signs of lung infiltration (Infiltration), nodule or mass seen in lung (Nodule/Mass)."
Elaborative Description 3	"Scan indicates possible tumor-like shape noted (Nodule/Mass), spread-out haziness in the lungs (Infiltration), cloudy spots visible in lungs (Lung Opacity), consolidated patches in lungs (Consolidation)."
Unelaborate Description 1	"Radiograph reveals Consolidation, Infiltration, Lung Opacity, Nodule/Mass."
Unelaborate Description 2	"Chest X-ray showing Consolidation, Infiltration, Lung Opacity, Nodule/Mass."
Unelaborate Description 3	"Abnormalities: Consolidation, Infiltration, Lung Opacity, Nodule/Mass."

Table 5 presents the changes in AP_{50}, AP_m, and AP_l metrics corresponding to different types of disease descriptions provided to the network during training. Incorporating more diverse and detailed textual descriptions led to higher AP_{50} and AP_m scores, indicating an improved understanding of the relationship between medical knowledge and the detection of normal or smaller abnormalities. This effect is reflected in the model's enhanced performance

for classes such as Pulmonary Fibrosis and a modest improvement for Aortic Enlargement (Figure 8).

In contrast, the AP_l score decreased slightly, which may be attributed to the model achieving a more balanced representation of chest abnormalities. As the model's focus broadened from larger lesions to a wider range of findings, performance for large abnormalities—such as Cardiomegaly (Figure 8)—declined. Overall, this trade-off suggests that incorporating richer disease descriptions promotes a more evenly balanced model, capable of precise detection of smaller, less urgent but clinically significant lesions, while maintaining reasonable performance for large, high-priority abnormalities.

Table 5. Performance of CXR-CoDet with varying caption descriptions (in %)

Captions	AP50	APm	APl
Elaborative	0.724	0.275	0.158
Unelaborate	0.715	0.044	0.213

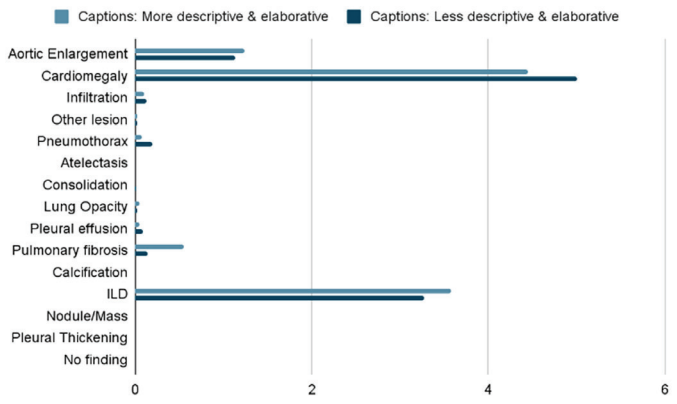

Figure 8. Shows an increase in AP50 for classes such as Aortic Enlargement, Pulmonary Fibrosis, and ILD with more elaborate disease captions. However, this has led to a decrease in AP50 for major classes such as Cardiomegaly, possibly due to the class imbalance in the dataset and the model's biased training towards majority classes with less descriptive captions.

Impact of varying number of support images

Table 6 summarizes the variation in AP_{50}, AP_m, and AP_l metrics as the number of support images in the region–region similarity mechanism increases from 1 to 3. While AP_{50} remains largely unchanged, AP_m decreases and AP_l increases. This pattern indicates that the model's performance improves for large abnormalities—as evidenced by the slight gains in detecting large lesions such as ILD and Cardiomegaly (Figure 9)—but at the expense of smaller and medium-sized findings. The selective increase in AP_l, without corresponding improvements in the other metrics, suggests that additional support images shift the model's focus toward larger lesions. Consequently, this configuration produces a model better suited for detecting high-risk, large-scale abnormalities that demand immediate clinical attention.

Table 6. Performance of CXR-CoDet with varying number of support images (in %)

Num. Support Imgs	AP50	APm	APl
1	0.724	0.275	0.158
3	0.724	0.060	0.250

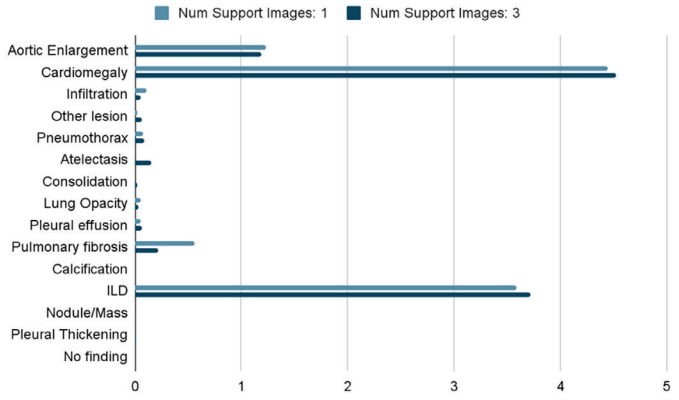

Figure 9. Shows a steady increase in performance for major classes such as Cardiomegaly and a slight decrease in Aortic Enlargement with an increasing number of support images. This has also led to an increase in performance for ILD and a decrease in performance for Pulmonary fibrosis.

Impact of embedding medical knowledge

Table 7 shows the change in AP50, APm, and APl metrics with a change in the pretrained text encoder utilized in the model architecture. The utilization of BiomedCLIP instead of regular CLIP has improved the model's performance w.r.t. AP50, APm, and APl, demonstrating the further understanding that the model has of relevant medical knowledge and its connection to visual features utilizing BiomedCLIP. This is additionally demonstrated in an increase in

performance for most of the classes in Figure 10, leading to the change in the text encoder from CLIP to BiomedCLIP, improving the overall performance of the model to detect most lesions, smaller ones that require precision in addition to larger ones that require immediate attention.

Table 7. Performance of CXR-CoDet with varying text encoder (in %)

BiomedCLIP v. CLIP	AP50	APm	APl
BiomedCLIP	0.724	0.275	0.158
CLIP	0.132	0.115	0.139

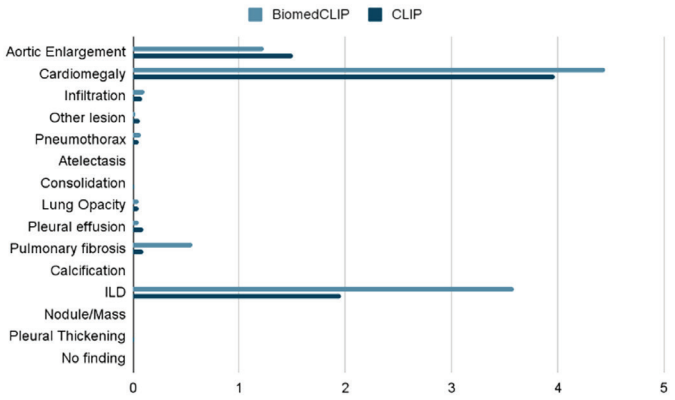

Figure 10. Shows the AP50 metrics per category with regular CLIP. This shows a full increase in AP50 metrics for most classes with BiomedCLIP utilized as the text encoder.

Summary of experiments

From this work's results, according to Table 8, it is observed that increasing the number of iterations has slightly increased the model's performance, as expected by standard ML conventions. Increasing the learning rate has led to an increase in overall model performance, along with an increase in performance on major abnormalities in the dataset, but a decrease in performance on abnormalities that require more close visual detail, as expected from standard ML conventions. Increasing the number of support images has led to a varied change in performance, with some larger sized abnormalities achieving higher performance and others achieving lower performance. Additionally, making captions more descriptive has improved the alignment between language and visual features, but at the cost of decreasing the model's performance in larger abnormalities. Training the model's detector from pretrained Detic weights instead of CoDet weights through the CXR-CoDet architecture has led to a higher performance of the model on larger features and a lower performance on a few smaller abnormalities. Most importantly, utilizing BiomedCLIP has significantly boosted the model's overall performance, which can be explained by BiomedCLIP's stronger understanding of medical language embeddings, making it the most effective parameter change.

Table 8. Change in performance with varying parameters

Change in settings/parameters	ΔAP50	ΔAPm	ΔAPl
Num. Support Imgs: 1 to 3	0.000	-0.215	0.092
Text embedding: CLIP to BiomedCLIP	**0.592**	**0.160**	**0.019**
Captions: Unelaborative to Elaborative	0.009	0.231	-0.055
Pretrained Weights: CoDet to Detic	0.225	-0.244	0.116
Batch Size: 32 to 64	-0.13	-0.221	-0.007
Num. Iterations: 90000 to 130000	0.018	0.035	0.028
Learning Rate: 0.0005 to 0.01	0.074	0.105	-0.038

DISCUSSION & CONCLUSION

This work examined the feasibility of applying cross-region feature alignment based object detection architectures to chest X-ray abnormality localization tasks. Different configurations of the hyperparameters of the model, such as the learning rate, batch size, number of iterations, number of support images, the descriptiveness of the captions, the initial pretrained weights, and the text encoder, were experimented with. These experiments were conducted to demonstrate the impact of these factors in improving the performance of the model.

The findings of this study highlight the inherent difficulty of applying region–region similarity in isolation within the chest abnormality detection domain. This is reflected in the comparatively low performance observed across the region–region similarity bottlenecks. The limitation likely stems from the imperfect nature of the region–region similarity

based image retrieval process, which can effectively capture visual differences between distinct objects but struggles to model pathological variations, given their subtle and heterogeneous manifestations. Furthermore, chest X-ray images often contain overlapping signs of multiple conditions, further complicating the differentiation of abnormalities. The absence of textual or semantic guidance in region–region similarity methods exacerbates this challenge, making it difficult to align medical terminology and clinically meaningful cues with corresponding visual patterns.

Overall, this work serves as an initial exploration of how cross-region feature alignment architectures can be employed as a bottleneck in real-world healthcare applications. The goal is not to demonstrate superior performance, but rather to conduct a deeper analysis of the potential of region–region similarity within the chest X-ray abnormality detection domain. For this reason, the study does not include direct performance comparisons with standard medical object detection methods. Nevertheless, this work acknowledges that incorporating comparisons with established approaches—such as PPAD (22), YOLO (23), and Mask R-CNN (21)—would broaden the scope and strengthen the conclusions of this work, and this remains an important future direction. To further address the limitations of region–region similarity in chest X-ray analysis, one promising strategy is to blur regions outside abnormalities, encouraging the model to focus on clinically relevant areas. Although radiologist performance was not included here due to limited access, the technical nature of this study, and the absence of radiologist annotations in the utilized dataset, future work will aim to foster collaboration between radiologists and AI systems for more clinically meaningful evaluation. Finally, this work believes that region–region similarity based architectures hold considerable promise in chest X-ray analysis, as indicated by the performance improvements observed with specific hyperparameter configurations. Future extensions may involve replacing the Faster R-CNN–based natural object detector with a chest X-ray–specific abnormality detector and integrating knowledge graphs into the similarity mechanisms to enhance model interpretability and performance.

ACKNOWLEDGEMENTS

I express my gratitude to my mentor, Pramit Saha, for his invaluable guidance and detailed feedback throughout the process of this research project. His dedication to research has additionally inspired me to persevere in this research project.

FUNDING SOURCES

There are no funding sources to declare.

PREPRINT INFORMATION

A preprint of this manuscript titled *"Investigating Text-Guided Cross-Region Feature Alignment for Multimodal Disease Localization in Chest X-ray Images"* has been posted on the Authorea preprint server. It is available at https://www.authorea.com/users/958477/articles/1327348-investigating-text-guided-cross-region-feature-alignment-for-multimodal-disease-localization-in-chest-x-ray-images. All authors have approved the posting of this preprint, and it is not under formal peer review elsewhere.

CONFLICT OF INTERESTS

The author declares that there is no conflict of interest related to this work

REFERENCES

1. Cheng Y-C, Hung Y-C, Huang G-H, *et al*. Deep learning-based object detection strategies for disease detection and localization in chest X-ray images. *Diagnostics*. 2024; 14 (23): 2636. https://doi.org/10.3390/diagnostics14232636
2. Zhu Y, Wang X, Hu H, *et al*. Co-Occurrence guided region-word alignment for open-vocabulary object detection (CoDet). In Advances in Neural Information Processing Systems (NeurIPS), 2023. New Orleans, LA, USA. p. 71078-71094.
3. Müller P, Meissen F, Kaissis G, and Rückert D. Weakly supervised object detection in chest X-rays with differentiable ROI proposal networks and soft ROI pooling. arXiv preprint, 2024. arXiv:2402.11985. Available from: https://arxiv.org/abs/2402.11985 (accessed on 2025-08-30).
4. Zheng Y and Liu K. Training-free boost for open-vocabulary object detection with confidence aggregation. arXiv preprint, 2024. arXiv:2404.08603. Available from: https://arxiv.org/abs/2404.08603 (accessed on 2025-08-30).
5. Ren S, He K, Girshick R and Sun J. Faster R-CNN: Towards real-time object detection with region

proposal networks. In Advances in Neural Information Processing Systems (NeurIPS), 2015. Montreal, QC, Canada. p. 91-99.
6. Albuquerque C, Henriques R and Castelli M. Deep learning-based object detection algorithms in medical imaging: Systematic review. *Heliyon*. 2025; 11 (18): e41137. https://doi.org/10.1016/j.heliyon.2024.e41137
7. Talius E and Sayana R. Zero-shot object detection for chest X-rays. Stanford University CS231n Course Report, 2022. Stanford, CA, USA.
8. Gao Y, Li R, Croxford E, Caskey J, *et al*. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*. 2025; 4: e58670. https://doi.org/10.2196/58670
9. Le HD, Le HT, Nguyen NH, *et al*. VinDr-CXR: An open dataset of chest X-rays with radiologists' annotations. *Scientific Data*. 2022; 9: 429. https://doi.org/10.1038/s41597-022-01498-w
10. Radford A, Kim J, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML), 2021. Vienna, Austria. p. 8748-8763.
11. Du J, Fang H, Wang L, *et al*. LaMI-DETR: Frozen CLIP-based DETR with language model instruction. In Proceedings of the European Conference on Computer Vision (ECCV), 2024. Malmö, Sweden. p. 312-328. https://doi.org/10.1007/978-3-031-73337-6_18
12. Pennington J, Socher R and Manning C. GloVe: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. Doha, Qatar. p. 1532-1543. https://doi.org/10.3115/v1/D14-1162
13. Bansal A, Sikka K, Sharma G, Chellappa R and Divakaran A. Zero-shot object detection. In Proceedings of the European Conference on Computer Vision (ECCV), 2018. Munich, Germany. p. 384-400. https://doi.org/10.1007/978-3-030-01246-5_24
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, *et al*., Generative adversarial nets. In Advances in Neural Information Processing Systems (NeurIPS), 2014. Montreal, QC, Canada. p. 2672-2680.
15. Johnson AEW, Pollard TJ, Berkowitz SJ, *et al*. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*. 2019; 6: 317. https://doi.org/10.1038/s41597-019-0322-0
16. Wang X, Peng Y, Lu L, Lu Z, *et al*. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. Honolulu, HI, USA. p. 2097-2106. https://doi.org/10.1109/CVPR.2017.369
17. Zhang S, Xu Y, Usuyama N, *et al*. BiomedCLIP: A multimodal biomedical foundation model. Microsoft Research, 2022. Available from: https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224 (accessed on 2025-08-30).
18. He K, Zhang X, Ren S and Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. Las Vegas, NV, USA. p. 770-778. https://doi.org/10.1109/CVPR.2016.90
19. Tian J, Jin Q, Wang Y, *et al*. Performance analysis of deep learning-based object detection algorithms on COCO benchmark: a comparative study. *J Eng Appl Sci*. 2024; 71: 76. doi:10.1186/s44147-024-00411-z.
20. Lin TY, Dollar P, Girshick R, He K, *et al*. Feature Pyramid Networks for Object Detection. arXiv (preprint). 2016 Dec 8 Available from: https://arxiv.org/pdf/1612.03144 (accessed on 2025-10-11)
21. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. arXiv (preprint). 2017 Mar 17 Available from: https://arxiv.org/pdf/1703.06870 (accessed on 2025-10-11)
22. Sun Z, Gu Y, Liu Y, Zhang Z, *et al*. Position-Guided Prompt Learning for Anomaly Detection in Chest X-Rays. arXiv (preprint). 2024 May 20. Available from: https://arxiv.org/html/2405.11976v1
23. Mustafa Z, Nsour H. Using Computer Vision Techniques to Automatically Detect Abnormalities in Chest X-rays. *Diagnostics (Basel)*. 2023 Sep 18; 13 (18): 2979. doi:10.3390/diagnostics13182979. PMID:37761345; PMCID:PMC10530162.