REPRESENTATION LEARNING FOR INTERPRETATION OF CHEST X-RAYS

REPRESENTATION LEARNING FOR INTERPRETATION OF CHEST X-RAYS

BY

MEHRDAD ESHRAGHI DEHAGHANI, B.Sc.

A Thesis

submitted to the Computing & Software

and the School of Graduate Studies

of McMaster University

in partial fulfilment of the requirements

for the degree of

Masters of Science

Masters of Science (2025) McMaster University

(Computing & Software) Hamilton, Ontario, Canada

TITLE:	Representation Learning for Interpretation of Chest X-Rays
AUTHOR:	Mehrdad Eshraghi Dehaghani M.Sc (Computing & Software), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Mehdi Moradi
NUMBER OF PAGES:	xvii, 52

Lay Abstract

Chest X-rays are one of the most common medical imaging tools used by doctors to diagnose diseases, track how illnesses progress, and make treatment decisions. However, analyzing these images can be complex and time-consuming. This research explores how artificial intelligence (AI) can help improve chest X-ray analysis by automatically detecting important features and generating medical insights.

Using a large dataset containing detailed descriptions of chest X-rays, we trained an AI model to recognize anatomical structures and detect diseases. This AI system was then used for various tasks, including identifying specific diseases, tracking their progression, and even generating medical reports.

Our approach improves accuracy and efficiency, making it easier for doctors to interpret X-rays. This work highlights how AI can support medical professionals and enhance automated radiology analysis, ultimately leading to better patient care.

Abstract

Chest X-ray imaging is one of the most commonly performed diagnostic procedures in radiology, playing a critical role in detecting chest pathologies, monitoring disease progression, and guiding treatment decisions. This thesis investigates the application of representation learning as an upstream modality to enhance various downstream tasks in chest radiograph analysis, including localized disease classification, progression tracking and automated radiology report generation.

To achieve this, we utilize the Chest ImaGenome dataset, a subset of MIMIC-CXR, which comprises 242,072 scene graphs that describe individual chest X-rays. These scene graphs contain automatically extracted information from radiology reports, including patient demographics, anatomical bounding boxes, pathological findings, and progression of disease in each anatomical region. This structured information serves as supervisory labels for training models.

For the upstream representation learning task, we employ the DEtection TRansformer (DETR), a transformer-based object detection framework, to identify anatomical structures in chest X-rays and generate meaningful feature representations. These learned features are subsequently leveraged for multiple downstream tasks, including localized classifications via specialized classifiers and radiology report generation using a large language model.

Our approach achieves strong performance across these tasks, with an average ROC of 89.1% over nine disease categories in localized disease detection. Additionally, our method demonstrates effectiveness in tracking localized disease progression, achieving an average accuracy of approximately $\sim$67% and an average F1 score of $\sim$71%. Furthermore, it produces clinically relevant radiology reports.

The results highlight the effectiveness of a unified transformer-based architecture for chest X-ray interpretation, demonstrating its capability to achieve competitive performance across multiple tasks while minimizing reliance on handcrafted features or task-specific models. This work underscores the potential of representation learning to enhance automated chest radiograph analysis and improve clinical decision support.

To my family.

Acknowledgements

I would like to express sincere gratitude to my supervisor Dr. Mehdi Moradi, for his encouragement and patient guidance on my research path.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CV	Computer Vision
NLP	Natural language processing
CXR	Chest X-Ray
AUC	Area Under the Curve
RoI	Region of Interest
MLP	Multi-layer Perceptron
FFN	Feed-forward Network
DETR	DEtection TRansformer
ROC	Receiver-operating Characteristic Curve
IoU	Intersection over Union

mAP Mean Average Precision

PSA Pseudo Self-attention

Declaration of Academic Achievement

The following academic contributions have been made during the course of this research:

- ***Published:*** **Mehrdad Eshraghi Dehaghani**, Amirhossein Sabour, Amarachi B. Madu, Ismini Lourentzou, and Mehdi Moradi, "Representation Learning with a Transformer-Based Detection Model for Localized Chest X-Ray Disease and Progression Detection," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15001, Springer Nature Switzerland, Oct. 2024, pp. 578–587.

- ***Under Review:*** Amirhossein Sabour, **Mehrdad Eshraghi Dehaghani**, and Mehdi Moradi, "Labeling of large imaging datasets using LLMs: Efficient strategies for radiology report structuring," Medical Image Computing and Computer Assisted Intervention (MICCAI), 2025.

- ***Accepted:*** Kailin Chu, Amirhossein Sabour, **Mehrdad Eshraghi Dehaghani**, Ismini Lourentzou, and Mehdi Moradi, "Localized disease severity detection

in chest X-Ray images," IEEE Engineering in Medicine and Biology Society (EMBC), 2025.

- ***Accepted:*** Lyuyang Wang, Sommer Chou, **Mehrdad Eshraghi Dehaghani**, Gerry Wright, Lesley MacNeil, Mehdi Moradi, "Self-supervised learning for drug discovery using nematode images: method and dataset," IEEE Journal of Biomedical and Health Informatics, 2024.

Chapter 1

Introduction

Chest X-ray (CXR) imaging is one of the most widely used and essential diagnostic tools in clinical practice [20]. It provides radiologists with critical insights for identifying various thoracic pathologies and abnormalities, assessing disease progression over time, and conducting routine health screenings. The diagnostic value of CXR imaging plays a pivotal role in devising timely and effective treatment plans for patients. However, given its widespread use, the growing volume of imaging studies poses a significant burden on radiologists and healthcare systems—an issue exacerbated by the global shortage of trained radiology professionals [1, 14, 15]. Reducing this burden through automated image analysis tools can substantially assist clinicians and improve diagnostic workflows.

The automation of CXR interpretation has emerged as a major innovation in medical imaging, offering the potential to enhance both diagnostic accuracy and efficiency. With increasing demands for rapid and consistent image assessments—particularly in resource-limited settings—AI-powered tools provide a promising solution to support radiologists and improve patient outcomes. Recent advancements in artificial

intelligence (AI), particularly in the field of computer vision, have facilitated the development of highly effective models for medical image analysis, achieving remarkable levels of accuracy.

This advancement has been largely driven by the emergence of large-scale annotated CXR datasets [5, 6], which have empowered deep learning models to capture intricate visual patterns and subtle indicators of disease. AI-based systems have demonstrated strong performance across multiple tasks, including disease classification, abnormality localization, and automated report generation [10, 11, 18]. These tools are not only becoming more accurate but are also approaching the diagnostic performance of experienced radiologists [17, 19].

Traditional deep learning models, such as convolutional neural networks (CNNs) [9], initially drove much of this progress. More recently, transformer-based architectures—particularly vision transformers—have shown notable improvements in medical image analysis by capturing long-range dependencies and improving interpretability [8, 12]. Originally developed for general vision tasks, vision transformers are now being increasingly adopted in healthcare due to their ability to model global context, which is especially valuable for radiological image interpretation.

Despite this progress, most existing models focus on global disease classification—predicting which pathologies are present in a CXR without specifying their exact location. While global classification is valuable, spatial localization of abnormalities adds significant interpretability and clinical relevance, allowing radiologists to better understand and verify AI predictions. This thesis addresses key limitations of current models by proposing novel tasks and approaches for localized CXR interpretation.

Specifically, this work introduces new models capable of both identifying disease presence and localizing abnormalities within anatomically defined regions of interest (RoIs). In addition, we propose a new task—*localized disease progression monitoring*—which involves determining whether a pathology in each anatomical region has improved, worsened, or remained stable between two timepoints. Lastly, we present a method for generating free-text radiology reports directly from CXRs. A central contribution of this thesis is the use of *representation learning* to unify these tasks under a single model architecture. Rather than training separate models for each task, we develop a shared backbone that extracts informative image features, which are then utilized across various downstream tasks. This approach leads to more efficient training and inference, and reduces the overall computational cost.

To support these efforts, we utilize the Chest ImaGenome dataset [20], derived from MIMIC-CXR [7], which includes 242,072 frontal CXR images annotated with structured scene graphs. Each image is associated with bounding boxes for up to 29 anatomical regions and region-level attributes, such as disease presence, progression status, and corresponding report sentences. In Chapter 2, we address the tasks of localized disease classification and progression monitoring. In Chapter 3, we focus on the generation of free-text radiology reports.

1.1 Representation Learning for Localized Disease Detection and Progression Monitoring

To support localization in the model output and enable multiple downstream tasks using a single unified framework, the proposed representation learning module is

designed to perform two primary functions:

- **Detect anatomical regions of interest (RoIs)** that serve as the spatial basis for localization and analysis.
- **Extract feature representations** of these RoIs, which are then used by downstream models to perform classification and progression analysis.

To achieve these objectives, we employ *DEtection TRansformer (DETR)* [3], a transformer-based object detection model. DETR takes a CXR image as input and identifies relevant anatomical regions by predicting their classes and corresponding bounding box coordinates. We fine-tune DETR on the Chest ImaGenome dataset, selecting 12 of the 29 annotated regions that are more frequently referenced in radiology reports and more likely to exhibit pathological findings.

DETR utilizes a ResNet-50 backbone [4] to extract global image features. These features are tokenized and passed through a transformer encoder-decoder architecture, which ultimately produces predictions via a feed-forward network (FFN). The visual features used for downstream tasks are obtained from the decoder's final query embeddings—each representing a specific anatomical region. These feature vectors form the core of our representation learning approach.

For localized disease classification, each feature vector is passed through a simple FFN to predict the presence of 9 possible findings within its corresponding region. For progression monitoring, we process a pair of CXRs—acquired from the same patient at different timepoints—using the same representation learning module. The resulting region-specific feature vectors from both images are compared by computing their differences. We apply a self-attention mechanism to the resulting difference

vectors and feed them into another FFN to predict the progression status (improved, worsened, unchanged) for each anatomical region.

This chapter introduces a novel yet lightweight architecture that demonstrates the ability to perform both tasks—localized disease classification and progression monitoring—using a single shared model with high accuracy and low computational overhead.

1.2 Representation Learning for Clinical Report Generation

In this chapter, we explore the potential of vision-based representation learning for generating high-quality free-text radiology reports from chest X-ray images. Clinical reports are typically written by radiologists after carefully analyzing different anatomical regions in the CXR [16]. Inspired by this approach, we design a system that mimics this process by focusing on region-specific visual features to generate descriptive, coherent textual outputs.

As in the previous chapter, we leverage the DETR model as the backbone for representation learning. This time, DETR is fine-tuned to detect all 29 anatomical regions annotated in the Chest ImaGenome dataset. For each detected region, DETR produces a corresponding feature vector, which serves as a compact representation of that region's visual characteristics.

To generate the final report, we condition a Large Language Model (LLM) on these region-specific feature vectors. Given the remarkable progress in LLMs over recent years [13, 2], these models offer a powerful mechanism for generating fluent

and clinically meaningful text. We employ a lightweight LLM that takes as input the DETR-derived features and produces one or two descriptive sentences per selected anatomical region.

The conditioning process is achieved through *pseudo self-attention* [21], which involves embedding the visual feature vectors into the LLM's attention mechanism by modifying the query, key, and value representations accordingly. To determine which regions should contribute to the report, we employ a feed-forward network (FFN) that selects the most relevant regions based on their feature representations.

This chapter extends the application of vision-based representation learning previously used for classification and localization, to the more complex task of free-text report generation. Our results demonstrate that, despite its simplicity, the proposed model performs competitively with most of the prior state-of-the-art approaches while remaining lightweight and computationally efficient.

1.3 Contributions and Thesis Organization

This sandwich thesis follows McMaster University thesis requirements and terms of use, consisting of one published conference paper, and one in-processing conference paper. All of those two works are related to the topic of representation learning for interpretation of Chest X-rays. The references for these two papers are as follows:

- **Mehrdad Eshraghi Dehaghani**, Amirhossein Sabour, Amarachi B. Madu, Ismini Lourentzou, and Mehdi Moradi, "Representation Learning with a Transformer-Based Detection Model for Localized Chest X-Ray Disease and Progression Detection," in *Proceedings of the Medical Image Computing and*

Computer Assisted Intervention – MICCAI 2024, vol. LNCS 15001, Springer Nature Switzerland, Oct. 2024, pp. 578–587. Contributions:

  - We propose a new approach for chest X-ray interpretation using DETR-based anatomical features to jointly perform localized disease detection and progression monitoring.
  - We introduce localized disease progression monitoring and show that feature differences across regions enable strong performance, even with a simple multi-layer perceptron (MLP).
  - We provide ablation studies and qualitative examples highlighting the role of anatomical regions in monitoring disease progression.

- **Mehrdad Eshraghi Dehaghani**, and Mehdi Moradi, "Clinical Radiology Report Generation for Chest X-Ray Using Transformer-Based Representation Learning Model," *To be submitted to* the Machine Learning in Medical Imaging (MLMI) Workshop, MICCAI 2025. Contributions:

  - We introduce a representation learning transformer-based architecture designed to generate high-quality free-text radiology reports from chest X-ray images.
  - We assess the performance of our approach using both grammatical accuracy and clinical efficacy metrics.
  - We conduct a comprehensive comparison with prior works on this task and provide an analysis of the results.

The thesis structure is as follows:

- **Chapter 2:** The details of representation learning with a transformer based detection model for localized chest x-ray disease and progression detection.
- **Chapter 3:** The details of clinical radiology report generation for chest X-Ray using transformer-based representation learning model.
- **Chapter 4:** The thesis conclusion and discussion.

Bibliography

[1] Bastawrous, S., Carney, B.: Improving patient safety: avoiding unread imaging exams in the national va enterprise electronic health record. Journal of digital imaging **30**(3), 309–313 (2017)

[2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), `https://arxiv.org/abs/2005.14165`

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – European Conference on Computer Vision (ECCV) 2020. pp. 213–229. Springer International Publishing, Cham (2020)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385

[5] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)

[6] Johnson, A.E.W., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**(1), 1–9 (2016)

[7] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data pp. 1–8 (2019)

[8] Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Beyer, L., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

[9] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[10] Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 269–280 (2021)

[11] Liu, G., Yin, C., Wang, Y., Zhang, Q., Sun, J., Wang, X., Choy, J., Lu, L.: Clinically accurate chest x-ray report generation. In: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 4th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 106, pp. 249–269 (2019)

[12] Park, S., Kim, M.Y., Park, H., Lee, J.B., Shin, J.H., Lee, K.S., Ye, J.C.: Multi-task vision transformer using low-level chest x-ray feature corpus for covid-19 diagnosis and severity quantification. Medical Image Analysis **75**, 102299 (2022)

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8) (2019)

[14] Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ (2017)

[15] Rosenkrantz, A.B., Hughes, D.R., Duszak, R.: The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology **279**(1), 175–184 (2016)

[16] Stacy, K., Goergen, F.J., Pool, T.J., Turner, J.E., Grimm, M.N., Appleyard, C., Crock, M.C., Fahey, M.F., Fay, N.J., Ferris, S.M.: Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. JMIRO **57**(1), 1–7 (2013)

[17] Tang, Y., Tang, Y., Xiao, J., Summers, R.M.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digital Medicine **3**(1), 70 (2020)

[18] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9049–9058 (2018)

[19] Wu, J.T., Wong, K.C., Chung, K., Lai, V., Wu, W., Yip, S., Li, K.C., Leung, W.H., Nicholls, J.M., Tsang, W.C., et al.: Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. JAMA Network Open **3**(10) (2020)

[20] Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W.G., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest imagenome dataset for clinical reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[21] Ziegler, Z.M., Melas-Kyriazi, L., Gehrmann, S., Rush, A.M.: Encoder-agnostic adaptation for conditional language generation (2019)

Chapter 2

Representation Learning with a Transformer-Based Detection Model for Localized Chest X-Ray Disease and Progression Detection

2.1 Abstract

Medical image interpretation often encompasses diverse tasks, yet prevailing AI approaches predominantly favor end-to-end image-to-text models for automatic chest X-ray reading and analysis, often overlooking critical components of radiology reports. At the same time, employing separate models for related but distinct tasks leads to computational overhead and the inability to harness the benefits of shared

data abstractions. In this work, we introduce a framework for chest X-ray interpretation, utilizing a Transformer-based object detection model trained on abundant data for learning localized representations. Our model achieves a mean average precision of ∼94% in identifying semantically meaningful anatomical regions, facilitating downstream tasks, namely localized disease detection and localized progression monitoring. Our approach also yields competitive results in localized disease detection, with an average ROC 89.1% over 9 diseases. In addition, to the best of our knowledge, our work is the first to tackle localized disease progression monitoring, with the proposed model being able to track changes in specific regions of interest (RoIs) with an average accuracy ∼67% and average F1 score of ∼71%. Code is available at `https://github.com/McMasterAIHLab/CheXDetector`.

2.2 Introduction

Chest X-ray is the most common medical imaging modality. In recent years, with the introduction of multiple large-scale annotated chest X-ray datasets [4, 5], the field of automatic interpretation of chest X-ray images leveraging artificial intelligence has seen a great deal of activity. There are generally two lines of work in this area. An early wave of works has focused on accurate detection/classification for a limited number of diseases or findings [13]. The obvious flaw is the limited scope, as radiology reports are not merely lists of findings. Instead, they consist of localized descriptions, often with comparative and localized references to the progress of disease from a previous scan. In response to these shortcomings, subsequent works have expanded the scope of these early works by including a larger number of findings in their classification models than those labeled in the publicly available datasets [17]. With the

increased popularity of language models in medical imaging, the more recent wave of activity is focused on end-to-end training of image-to-sequence models that produce a complete radiology report given a chest X-ray image [7, 11]. This line of work addresses the problem of limited scope and application of disease classifiers. However, these models are often evaluated for their readability and similarity to radiologist reports, and not for the accuracy of the findings they list or their comprehensiveness [18]. As is common in generative models, these models can often produce factually incorrect language. Additionally, there has been a growing interest in disease detection with semantically meaningful localization [1, 10] as well as the monitoring of disease progression within image pairs, assessing whether a patient's condition has improved, deteriorated, or remained stable over time [6, 9]. Despite these advancements, to the best of our knowledge, the challenge of *localized* disease progression monitoring, predicting disease progression in specific anatomical regions, remains unexplored.

Given the weaknesses of direct image-to-text models and the variety of detection/classification tasks involved in chest X-ray interpretation, we propose to train and utilize a DEtection TRansformer (DETR) anatomical region detection model [2] to address multiple clinically relevant downstream tasks such as localized disease detection and localized disease progression monitoring. Specifically, previous works have provided large datasets of X-ray images with marked bounding boxes for anatomical regions ('lower lobe of right lung') [14, 15]. We define the detection of these bounding boxes as the task for training a DETR model. When trained, this model provides a rich feature vector for each anatomical region that can be used for both localized disease detection and localized disease progression monitoring. For each task, we train relatively compact models, using the features from the upstream model. We show

that the performance of our proposed framework is comparable to models specifically trained for these tasks. Our contributions can be summarized as follows:

(1) We introduce a novel approach for chest X-ray interpretation. By utilizing rich feature vectors generated by a DETR model trained for anatomical region detection, we address two clinically relevant downstream tasks simultaneously, localized disease detection and localized progression monitoring.

(2) We introduce the task of disease progression monitoring at a localized level. Our experimental results show that a simple model that extracts anatomical region feature differences can achieve competitive accuracy on this new task. We additionally demonstrate that a simple MLP architecture can jointly achieve competitive performance in localized disease detection.

(3) We further provide comprehensive ablation analysis with three model variations and qualitative examples to show the importance of anatomical regions in disease progression monitoring.

2.3 Methodology

2.3.1 Problem Definition

Let $\mathcal{C} = \{(\mathbf{X}, \mathbf{X}')_i\}_{i=1}^{N}$ be a set of CXR image pairs, where $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{H \times W \times C}$, and H, W and C are the height, width, and number of channels, respectively. Each image $\mathbf{X}$ is associated with a localized label set $\mathcal{Y}_i^1 = \{y_{i,k,m}\}_{k=1,m=1}^{K,M}$ where $y_{i,k,m}^1 \in \{0, 1\}$, indicating whether the label for the m-th finding appears in the k-th anatomical region of the image or not. In addition, each image pair $(\mathbf{X}, \mathbf{X}')_i$ is associated with a

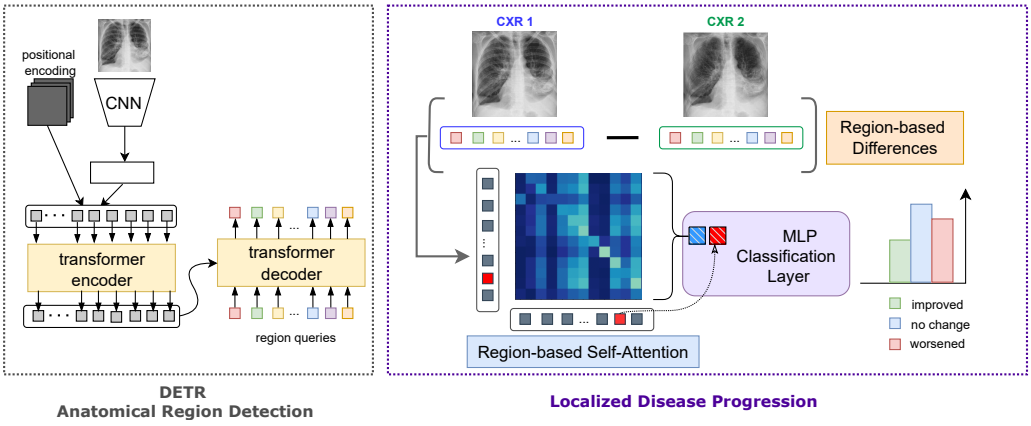

Figure 2.1: We first train a DETR anatomical region detection model on a large collection of CXR images. Given a pair of CXR images, the pretrained DETR decoder extracts visual features for anatomical regions of interest (RoI) for each image. These features are then used to compute region-based visual differences between the two CXRs. The information encoded in the difference vector is summarized through the self-attention mechanism that captures the importance of each RoI vector in relation to other RoIs and helps the model focus on relevant RoI changes. The resulting summary vector is concatenated with the region of interest (RoI) vector and fed into a multi-layer perceptron (MLP) classification layer for predicting whether the condition localized on the specific RoI has improved, worsened, or remained unchanged.

label set $\mathcal{Y}_i^2 = \{y_{i,k}\}_{k=1}^K$, where $y_{i,k}^2$ indicates whether the overall condition in the k-th anatomical region of the image pair $(\mathbf{X}, \mathbf{X}')_i$ has improved, worsened, or remained the same. The goal is to design a model that accurately predicts a set of labels indicative of the presence of pre-defined diseases at every anatomical Region of Interest (RoI) for an unseen image, and is also able to compare the two unseen images $(\mathbf{X}, \mathbf{X}')$ to predict localized progression labels as accurately as possible.

2.3.2 DETR region representation extraction backbone

To this end, we first utilize a DETR pre-trained region detection model for extracting feature representations of anatomical regions of interest (RoIs). The DETR region detection outputs a set of bounding boxes denoted as

$$\mathbf{B} = \{(b_1, c_1, s_1), (b_2, c_2, s_2), \ldots, (b_K, c_K, s_K)\}$$

where $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ represents the coordinates of the i-th bounding box, c_i denotes the class label associated with the box (chosen from a predefined set of K anatomical regions), and s_i represents the confidence score associated with each region query, indicating the likelihood of corresponding to a valid RoI. For each region query, the output of the last hidden state of the decoder can be represented as $\mathbf{F} = \{f_1, f_2, \ldots, f_i, \ldots, f_K\}$, where f_i represents the last hidden state of the decoder for the i-th region query. These hidden states serve as learned anatomical region feature representations, capturing the contextual information extracted by the decoder regarding the corresponding anatomical region query.

2.3.3 Localized Disease Detection

By utilizing the extracted feature representations $\mathbf{F} = \{f_1, f_2, \ldots, f_i, \ldots, f_K\}$ for the K anatomical regions of interest (RoIs), we train a compact feed-forward network to predict the presence or absence of particular disease in the respective RoIs. Let $y_{i,k,m}$ represent the ground truth label indicating the actual presence (1) or absence (0) of the m-th disease in the k-th RoI for the i-th sample. Similarly, let $\hat{y}_{i,k,m}$ represent the predicted probability for the presence or absence of the m-th disease in the k-th

RoI for the i-th sample. The binary cross-entropy loss $\mathcal{L}$ for a batch of N images can be defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}\left(\hat{y}_{i,k,m}\cdot\log(y_{i,k,m}) + (1-\hat{y}_{i,k,m})\cdot\log(1-y_{i,k,m})\right), \tag{2.3.1}$$

where N denotes the batch size and M the number of diseases.

2.3.4 Localized Disease Progression Monitoring

For localized disease progression monitoring, the goal is to predict, for each image pair $(\mathbf{X}, \mathbf{X}')_i$, whether the condition of a particular k-th anatomical region has improved, worsened, or remained unchanged. Let $\mathbf{F}_1 = \{f_{1,1}, f_{1,2}, \ldots, f_{1,k}, \ldots, f_{1,K}\}$ represent the feature vectors extracted from the regions of interest (RoIs) in the first image $\mathbf{X}_i$, and $\mathbf{F}_2 = \{f_{2,1}, f_{2,2}, \ldots, f_{2,k}, \ldots, f_{2,K}\}$ represent the feature vectors extracted from the corresponding RoIs in the second image $\mathbf{X}'_i$, where K is the number of RoIs. The model computes the differences of region vectors between the two images $\mathbf{F}_{\text{diff}} = [\mathbf{F}_2 - \mathbf{F}_1] = [f_{2,1} - f_{1,1}, \ldots, f_{2,k} - f_{1,k}, \ldots, f_{2,K} - f_{1,K}]$. To summarize the RoI information and capture the relationships between different RoIs, we employ a self-attention mechanism. The self-attention operation on $\mathbf{F}_{\text{diff}}$ can be denoted as:

$$\boldsymbol{\alpha} = \text{Softmax}\left(\frac{\mathbf{F}_{\text{diff}}\mathbf{F}_{\text{diff}}^{T}}{\sqrt{d_k}}\right), \tag{2.3.2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_k]$ represents the attention weights for each RoI, d_k is the dimension of the key vectors, and the softmax function is applied along the rows of the matrix. Next, the weighted sum of the difference feature vectors corresponding

to each RoI k is computed using the attention weights, $\boldsymbol{\alpha}\mathbf{F}_{\text{diff}}$, providing a summarized representation of the global RoI information considering the interdependencies between different RoIs. To perform the final localized disease progression prediction, we report results in both "Global" and "Region-focused" attention variants. In the "Global" approach, we average the rows of the self-attention output to obtain a single global vector to be used for all the RoIs of the chest x-ray, which is concatenated with $\mathbf{F}_k$ and passed through a classification layer. In the "Region-focused" approach, the k-th row of the self-attention output, corresponding to the k-th RoI, is concatenated with $\mathbf{F}_k$ and used directly for prediction, $\hat{y}_k = g\left([\mathbf{att}_k; \mathbf{F}_k]\right)$, where $\mathbf{att}_k$ is the self-attention output for the k-th RoI and $\mathbf{F}_k = [f_{2,k} - f_{1,k}]$ is the difference vector for the k-th RoI. The disease progression model loss for a batch of N images is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_{i,k}\cdot\log(\hat{y}_{i,k}) + (1-y_{i,k})\cdot\log(1-\hat{y}_{i,k})\right], \tag{2.3.3}$$

where $\mathcal{L}$ is the total loss over the entire batch, N is the batch size, K is the number of RoIs, and $y_{i,k}$, $\hat{y}_{i,k}$ are the ground truth label and model prediction for the k-th RoI of the i-th image.

2.4 Experiments

Implementation Details. We employ DETR with ResNet-50 backbone as the initial model for fine-tuning and extracting feature representations of anatomical regions of interest. The model was trained using PyTorch-Lightning [3], with AdamW optimizer [8], weight decay of 10^{-4}, backbone learning-rate of 10^{-5} and a learning-rate of 10^{-3}. To avoid the gradient exploding problem, we use gradient clipping of 0.1.

The initial batch size for training is 13. For further increasing the batch size while considering the memory limitations, we use accumulated gradients of 5. The model was trained for 25 epochs. To filter out unwanted detections, we apply a threshold $\tau > 0.85$ to the scores associated with each region query. Only region queries with scores exceeding this threshold are considered valid detections. For the disease classification module, we use a feed-forward network (FFN) similar to DETR's FFN with an additional batch norm between layers. The input dimension for the FFN is 256, hidden dimension of 256, output dimension of 9 (equal to the number of studied finding classes). The model is trained for 100 epochs using Adam optimizer with a learning rate of 5×10^{-4} and a weight decay of 10^{-5}. For the localized disease progression detection module, a similar feed-forward network with an input dimension of 512, hidden dimension of 256 with two layers and output dimension of 3 is used. The model is trained for 100 epochs using Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-5}.

Dataset. We make use of the Chest ImaGenome dataset [15]. This dataset consists of two different sub-datasets: 1) Locally labeled data using a combination of rule-based natural language processing (NLP) and CXR atlas-based bounding box detection techniques [14, 16] to generate the annotations (silver dataset). This subset comprises $237,827$ frontal MIMIC-CXRs [5]. 2) Manually validated and corrected studies of 500 patients as ground truth. Chest ImaGenome is represented as an anatomy-centered scene graph with $1,256$ combinations of relation annotations between 29 CXR anatomical locations and their attributes. Each image is structured as one scene graph, resulting in approximately $670,000$ localized comparison relations between the anatomical locations across sequential exams. Rich representation features play a key

Table 2.1: Dataset characteristics for progression labels per anatomical regions at: right upper lung zone (RULZ), right mid lung zone (RMLZ), right lower lung zone (RLLZ), right costophrenic angle (RCA), right hilar structures (RHS), right apical zone (RAZ), left upper lung zone (LULZ), left mid lung zone (LMLZ), left lower lung zone (LLLZ), left costophrenic angle (LCA), left hilar structures (LHS), cardiac silhouette (CS)

Progression Label	RULZ	RMLZ	RLLZ	RCA	RHS	RAZ
Improved	957	2,338	5,681	5,249	7,799	744
Worsened	1,301	3,537	8,333	6,406	7,699	543
No Change	37,149	34,617	28,673	30,529	27,883	37,915
Total	39,407	40,492	42,687	42,184	43,381	39,202
Progression Label	LULZ	LMLZ	LLLZ	LCA	LHS	CS
Improved	678	2,382	6,305	5,225	7,749	1,722
Worsened	927	3,414	8,996	6,399	7,599	3,097
No Change	37,596	34,549	27,667	30,478	27,954	35,036
Total	39,201	40,345	42,960	42,102	43,302	39,855

role in performing the downstream task efficiently. To train the upstream model, we utilize the entire silver dataset of Chest ImaGenome with 70/10/20 split.

For localized disease progression, we consider the localized comparison relation data within Chest ImaGenome that pertains to cross-image relations for nine diseases of interest. Each comparison relation in the Chest ImaGenome dataset includes the DICOM identifiers of the two CXRs being compared, a set of comparison labels per some anatomical regions, and a set of disease names. In some cases, more than one progression label was assigned to one region. We excluded these samples from the dataset to focus on more accurate labels. The comparison is labeled as "no change", "improved" or "worsened", which indicates whether the patient's condition has remained stable, improved, or worsened, respectively. In contrast to [9] which reports global classification, we solve this problem at the local level, acquiring one progress label per anatomical location. We use $35,646$ CXR pairs in total that pertain to the

Table 2.2: Area under PR curves for the 12 anatomical locations with an mPA of 93.5%, calculated using [12]. Format for each cell: (Anatomical location: Area under PR curve, IoU threshold = 0.5)

RULZ	**RMLZ**	**RLLZ**	**RCA**	**RHS**	**RAZ**
0.984	0.961	0.962	0.818	0.963	0.968
LULZ	**LMLZ**	**LLLZ**	**LCA**	**LHS**	**CS**
0.983	0.964	0.955	0.743	0.960	0.959

nine diseases of interest. The distribution of the data is improved $(53,134)$, worsened $(58,251)$, and no change $(390,046)$. We use 70/10/20 train/validation/testing split across studies. Due to the large gap between the number of "no change" labels and two other classes, only for training, we consider a random subset of "no change" labels equal to the maximum number of labels in the other two classes. Table 2.1 provides further details. For localized disease classification, we utilize the silver dataset $(237,827$ CXRs) to train the model. We use 70/10/20 train/validation/testing split across studies. High-level statistics of the generated dataset based on findings and anatomical regions of interest are included in the supplementary material.

2.5 Results

Upstream detection network: Table 2.2 reports the area under precision-recall curves for the twelve target anatomical regions. The mean average precision (mAP) is 93.5%. For 10 of the 12 regions, the area under ROC curve is at or above 96%, with only right and left costophrenic angle being the exceptions.

Localized disease detection: Figure 2.2 shows the ROC curves obtained for the nine findings from the localized disease detection model. Despite a simple MLP architecture, the model provides an average AUC score of 89.1%. The closet benchmark

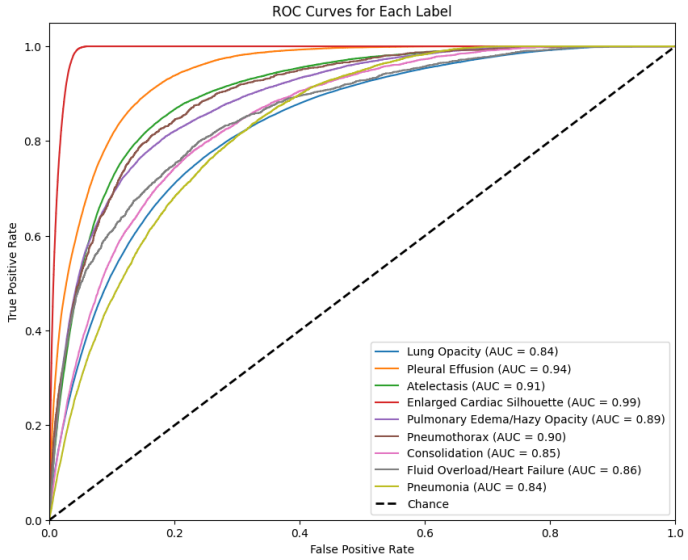

Figure 2.2: Localized Multi-label Disease Detection Results

Table 2.3: Localized Disease Progression Results (Accuracy/Weighted F1)

Anatomical location	Attention (Global)	Attention (Region-focused)	MLP
Right Upper Lung Zone	66.66/77.07	78.93/84.94	96.17/94.30
Right Mid Lung Zone	62.25/69.12	63.40/70.12	79.76/80.07
Right Lower Lung Zone	56.38/57.47	58.33/59.69	17.48/7.15
Right Costophrenic Angle	61.72/64.12	56.62/60.69	46.67/51.78
Right Hilar Structures	61.17/59.93	59.24/60.29	16.41/6.89
Right Apical Zone	66.69/78.11	83.81/89.53	97.44/96.17
Left Upper Lung Zone	66.56/77.75	81.01/86.85	97.15/95.75
Left Mid Lung Zone	63.80/70.09	69.23/74.30	83.34/81.82
Left Lower Lung Zone	55.56/55.91	54.55/55.76	19.98/8.73
Left Costophrenic Angle	61.86/64.43	61.50/64.38	38.19/42.64
Left Hilar Structures	67.88/63.42	61.87/63.20	16.59/9.29
Cardiac Silhouette	65.38/73.40	67.59/75.29	90.28/88.00
Weighted Average	**67.36/70.60**	**66.65/70.86**	**60.35/57.54**

to this work is [1] which reports localized disease detection on the same dataset with average AUC scores in the range of 89% to 93% for various models trained for the specific task of localized disease detection.

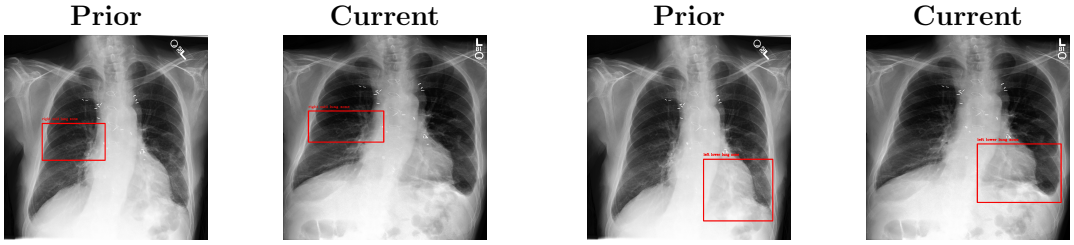

Figure 2.3: Examples of model predictions obtained by our model compared against the ground-truth labels and CheXRelFormer model [9]. Lung Opacity pathology (left) and Pneumonia pathology (right).

Localized disease progression: The results for localized progression labeling are presented in Table 2.3 for the three model variations. It is clear that the introduction of the attention layer in this classifier has improved the results compared to the baseline of MLP. Both the global and region-focused attention architectures outperform the MLP model, with a slight edge for global attention which provides an average accuracy $\sim$ 67% and an average F1 score of $\sim$ 71%. For this application, we do not have a current direct comparison from previous work. We trained a CNN Siamese network equivalent to the one in [15] with 3 classes, on 9 diseases and 12 anatomical regions, maintaining our original train/test splits. This simple model delivered a weighted average accuracy of only $\sim$ 34% and F1 score of $\sim$ 32%. Authors in CheXRelFormer [9] report global disease progression with an average accuracy of 49% across the diseases. Figure 2.3 highlights the advantage of localized over global progress classification. The progression labels in different anatomical locations can be inconsistent. As the figure shows, while our local model correctly classifies the progression label for both regions of interest, the global label by definition provides a single label that cannot be correct for both regions.

2.6 Conclusions

In this study, we presented a novel approach for interpreting chest X-rays, leveraging rich feature vectors derived from a DETR model trained specifically for anatomical region detection. By harnessing these feature vectors, we concurrently addressed two clinically significant downstream tasks: localized disease detection and localized progression monitoring. Furthermore, we introduce the novel task of disease progression monitoring at a localized level, demonstrating that extracting anatomical region feature differences can achieve competitive accuracy in this domain. Our experiments showcase the effectiveness of representation learning combined with simple architectures in achieving competitive performance for localized disease detection and progression.

Bibliography

[1] Agu, N.N., Wu, J.T., Chao, H., Lourentzou, I., Sharma, A., Moradi, M., Yan, P., Hendler, J.: Anaxnet: Anatomy aware multi-label finding classification in chest x-ray. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 804–813. Springer International Publishing, Cham (2021)

[2] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – European Conference on

Computer Vision (ECCV) 2020. pp. 213–229. Springer International Publishing, Cham (2020)

[3] Falcon, W., The PyTorch Lightning team: PyTorch Lightning (Mar 2019). https://doi.org/10.5281/zenodo.3828935, `https://github.com/Lightning-AI/lightning`

[4] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019)

[5] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data pp. 1–8 (2019)

[6] Karwande, G., Mbakwe, A.B., Wu, J.T., Celi, L.A., Moradi, M., Lourentzou, I.: Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. pp. 581–591. Springer (2022)

[7] Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3334–3343 (2023)

[8] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2018)

[9] Mbakwe, A.B., Wang, L., Moradi, M., Lourentzou, I.: Hierarchical vision transformers for disease progression detection in chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 685–695. Springer (2023)

[10] Müller, P., Meissen, F., Brandt, J., Kaissis, G., Rueckert, D.: Anatomy-driven pathology detection on chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 57–66. Springer (2023)

[11] Nguyen, H., Nie, D., Badamdorj, T., Liu, Y., Zhu, Y., Truong, J., Cheng, L.: Automated generation of accurate & fluent medical X-ray reports. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3552–3569. Association for Computational Linguistics (Nov 2021)

[12] Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., da Silva, E.A.B.: A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics **10**(3) (2021). https://doi.org/10.3390/electronics10030279, `https://www.mdpi.com/2079-9292/10/3/279`

[13] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning (2017)

[14] Wu, J., Gur, Y., Karargyris, A., Syed, A.B., Boyko, O., Moradi, M., Syeda-Mahmood, T.: Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In: Proceedings of the 17th International Symposium on Biomedical Imaging (ISBI). pp. 799–803. Institute of Electrical and Electronics Engineers (IEEE) (2020)

[15] Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W.G., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest imagenome dataset for clinical reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[16] Wu, J.T., Syed, A., Ahmad, H., et al.: Ai accelerated human-in-the-loop structuring of radiology reports. In: Proceedings of the Americal Medical Informatics Association (AMIA) Annual Symposium (2020)

[17] Wu, J.T., Wong, K.C.L., Gur, Y., Ansari, N., Karargyris, A., Sharma, A., Morris, M., Saboury, B., Ahmad, H., Boyko, O., Syed, A., Jadhav, A., Wang, H., Pillai, A., Kashyap, S., Moradi, M., Syeda-Mahmood, T.: Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. Journal of the American Medical Association (JAMA) Network Open **3**(10), e2022779–e2022779 (10 2020)

[18] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)

2.7 Supplementary material

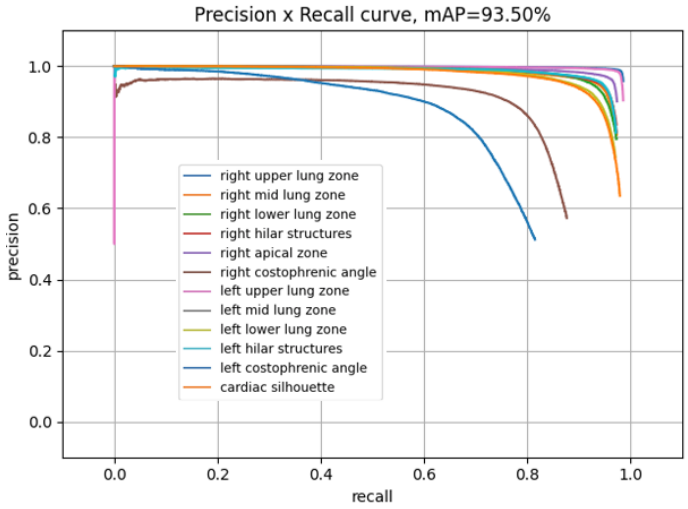

Figure 2.4: Precision-Recall curves for backbone representation learning module. Plot created using [12]

Table 2.5: Anatomical regions characteristics for representation learning model using DETR

Progression Label	Train	Val	Test	Total
Right Upper Lung Zone	166,259	23,918	47,341	237,518
Right Mid Lung Zone	162,889	23,400	46,422	232,711
Right Lower Lung Zone	162,983	23,414	46,446	232,843
Right Costophrenic Angle	166,137	23,900	47,297	237,334
Right Hilar Structures	162,883	23,399	46,419	232,701
Right Apical Zone	162,836	23,380	46,414	232,630
Left Upper Lung Zone	166,226	23,910	47,313	237,449
Left Mid Lung Zone	162,146	23,294	46,240	231,680
Left Lower Lung Zone	162,582	23,372	46,377	232,331
Left Costophrenic Angle	165,875	23,873	47,233	236,981
Left Hilar Structures	162,170	23,299	46,233	231,702
Cardiac Silhouette	166,490	23,949	47,387	237,826
Total	1,969,476	283,108	561,122	2,813,706

Table 2.6: Dataset characteristics for disease classification labels per anatomical regions at: right upper lung zone (RULZ), right mid lung zone (RMLZ), right lower lung zone (RLLZ), right costophrenic angle (RCA), right hilar structures (RHS), right apical zone (RAZ), left upper lung zone (LULZ), left mid lung zone (LMLZ), left lower lung zone (LLLZ), left costophrenic angle (LCA), left hilar structures (LHS), cardiac silhouette (CS)

Finding Label	RULZ	RMLZ	RLLZ	RCA	RHS	RAZ
Lung Opacity	10,922	25,132	70,428	57,163	59,432	6,196
Pleural Effusion	663	3,697	15,788	56,107	0	0
Atelectasis	1,739	9,751	43,299	2,162	2,472	0
Enlarged Cardiac Silhouette	0	0	0	0	0	0
Hazy Opacity/Pulmonary Edema	1,479	2,404	6,807	1,685	35,350	0
Pneumothorax	115	197	1105	834	0	4603
Consolidation	1420	4192	7954	499	1301	0
Heart Failure/Fluid Overload	588	189	631	202	1249	0
Pneumonia	1940	6493	13647	0	2960	219
Finding Label	LULZ	LMLZ	LLLZ	LCA	LHS	CS
Lung Opacity	8,177	26,130	80,840	59,380	58,360	0
Pleural Effusion	526	5,066	18,933	18,166	0	0
Atelectasis	1,155	13,434	55,872	2,616	2,043	0
Enlarged Cardiac Silhouette	0	0	0	0	0	58,908
Hazy Opacity/Pulmonary Edema	1,479	2,391	6,805	1,673	35,196	0
Pneumothorax	77	125	729	594	0	0
Consolidation	900	4,382	9,045	518	1,014	0
Heart Failure/Fluid Overload	587	187	633	201	1,241	6,636
Pneumonia	1,074	5,862	13,907	0	2,270	0

Table 2.7: # Images (number of images) and # Bboxes (number of bounding boxes) utilized in order to generate the dataset for localized disease classification task

Finding Label	# Images	# Bboxes
Lung Opacity	102,675	323,175
Pleural Effusion	51,328	111,267
Atelectasis	49,507	94,280
Enlarged Cardiac Silhouette	41,195	41,195
Hazy Opacity/Pulmonary Edema	24,958	66,558
Pneumothorax	4,448	5,904
Consolidation	10,470	21,887
Heart Failure/Fluid Overload	4,694	8,723
Pneumonia	17,142	33,844

Chapter 3

Clinical Radiology Report Generation for Chest X-Ray Using Transformer-Based Representation Learning Model

3.1 Abstract

High-quality clinical reports in radiology play a critical role in improving diagnostic accuracy and guiding more effective and personalized treatment plans. In this paper, we explore the application of transformer-based representation learning for using anatomical region detection to automate radiology report generation in chest X-ray imaging. We propose a simple framework that leverages an upstream representation learning model to detect anatomical regions and extract fine-grained feature

representations from chest radiographs. These feature embeddings are then used to identify regions with a higher likelihood of pathological findings, prioritizing them for detailed analysis and report generation. Finally, the region-specific feature encodings are passed to a large language model (LLM) to generate interpretable, and contextually relevant radiology reports. The obtained results confirm the efficacy of the proposed method, highlighting the potential of representation learning in enhancing various radiology tasks, including anatomical localization, disease detection, and automated report synthesis.

3.2 Introduction

Chest X-ray imaging is widely used in clinical practice for both routine health assessments and the diagnosis of various pathologies, aiding in the development of efficient and effective treatment plans. Radiologists interpret these images and document their findings in free-text reports, which serve as critical references for patient care. However, the increasing demand for chest X-ray imaging, coupled with a shortage of trained radiologists, has led to significant workloads and potential delays in diagnosis [2, 27, 28]. To address this challenge, automatic clinical report generation has emerged as an active area of research in artificial intelligence (AI), aiming to reduce the burden on radiologists while maintaining high diagnostic accuracy.

Interpreting a chest X-ray involves analyzing different anatomical regions and describing relevant findings for each one. Given that chest X-ray reports follow a free-text structure—where each sentence typically corresponds to a specific anatomical region—generating high-quality reports presents several challenges. Many existing AI-based methods fail to capture all necessary information, often producing

incomplete reports or generating inaccurate findings [19]. A major limitation of these approaches is their reliance on raw visual features of the image without leveraging structured anatomical representations. This results in a lack of interpretability, making it difficult to justify the use of such models in clinical decision-making [9, 10, 19].

Recent progress in representation learning has demonstrated significant potential across a range of medical imaging application, including localized disease classification, progression monitoring [8], and severity assessment [12]. This technique involves training an upstream model to extract generalizable features, which can then be applied to downstream tasks. Transformer-based models [32], in particular, have demonstrated significant potential in addressing complex challenges in various fields such as computer vision [13]. In the context of report generation, a recent study [31] attempted to improve accuracy and explainability by adopting a representation learning approach that mimics a radiologist's workflow [30]—detecting anatomical regions and generating structured, region-specific descriptions.

Following that work and previous success stories, we explore the use of transformer-based object detection models as an upstream representation learning module for free-text clinical report generation. Specifically, we leverage the DEtection TRansformer (DETR) [4] model to identify anatomical regions of interest and extract corresponding feature vectors. These feature representations help determine which regions are more likely to exhibit pathological findings, guiding the generation of detailed and explainable clinical reports. By integrating these features into a large language model (LLM), we aim to produce structured reports where each sentence corresponds to a specific anatomical region, thereby enhancing interpretability and clinical applicability.

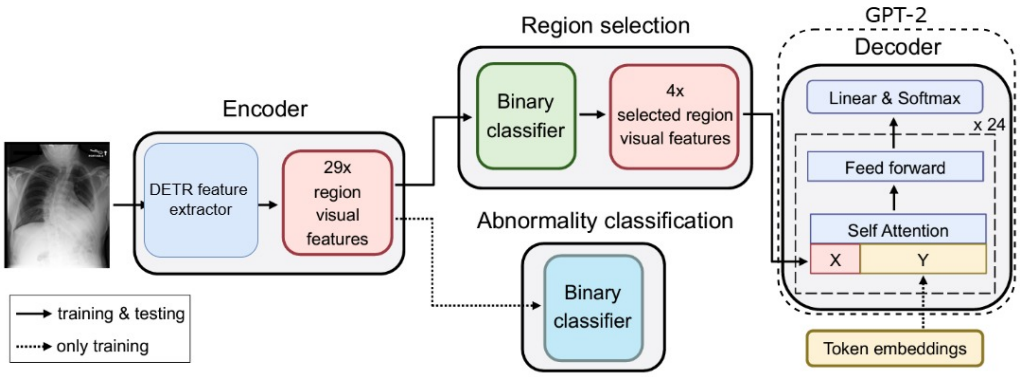

Figure 3.1: Overview of the model architecture. The model consists of three main stages. First, DETR is employed for anatomical region detection. Given an input chest X-ray (CXR), DETR identifies and extracts visual features for 29 possible regions in the image. In the second stage, a region selection module, implemented as a multi-layer perceptron (MLP), determines which regions should be used for generating textual descriptions. Additionally, an abnormality classification module, structurally similar to the region selection module, is used during training to enhance DETR's focus on clinically significant regions by identifying abnormalities. The final stage incorporates a GPT-2 Medium [25] decoder, which integrates the visual features of the selected regions and generates one or two descriptive sentences per region. These sentences are subsequently processed and merged to form the model's final output.

3.3 Methodology

3.3.1 Overview

Our model is consisted of three major components. The first component being the upstream region representation module, makes use of DETR to detect and extract the features of at most 29 anatomical regions of interest (RoIs). The next module being the region selection module, with the use of the features extracted in the first stage as input and a simple feed-forward neural network, makes a decision on whether or not to select each detected region to produce a report for. During training, there is an abnormality detection model which is an structurally identical feed-forward network

as the region selection model and determines if a region is abnormal, meaning that it contains a disease. This will help DETR to come up with richer features which subsequently helps region selection model to perform better. The third stage is a large language model pre-trained on PubMed [22] and conditioned on the region features which produces 1-2 sentences per selected RoI.

3.3.2 Problem Definition

Let $\mathcal{C} = \{\mathbf{X}_i\}_{i=1}^{N}$ be a set of chest X-ray (CXR) images, where each image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ has height H, width W, and C channels. Each image $\mathbf{X}$ is associated with a radiology report $\mathcal{Y} = \{s_1, s_2, ..., s_T\}$, where s_t represents the t-th sentence describing the observed findings in a particular anatomical region. The objective is to design a model that generates a structured and clinically accurate report $\hat{\mathcal{Y}}$ that closely matches the ground truth $\mathcal{Y}$. The proposed model consists of three main stages: anatomical region detection, region selection with abnormality classification, and report generation.

3.3.3 Anatomical Region Detection

In the first stage of our model, we utilize a pre-trained DEtection TRansformer (DETR) model to detect anatomical regions of interest (RoIs) within a given chest X-ray (CXR) image $\mathbf{X}$. The DETR model outputs a set of bounding boxes, each corresponding to a distinct anatomical region, along with its associated class label and confidence score. Formally, the output of DETR can be represented as:

$$\mathbf{B} = \{(b_1, c_1, s_1), (b_2, c_2, s_2), \ldots, (b_K, c_K, s_K)\}, \tag{3.3.1}$$

where $b_i = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ represents the coordinates of the i-th bounding box, c_i is the class label assigned to the detected region (selected from a predefined set of K anatomical regions), s_i denotes the confidence score for the detection, reflecting the likelihood of the region being a valid anatomical structure. For each region query, the final hidden state of the DETR decoder provides a feature representation that encapsulates contextual information about the detected anatomical region. We denote this set of learned representations as:

$$\mathbf{F} = \{f_1, f_2, \ldots, f_K\}, \tag{3.3.2}$$

where f_i corresponds to the last hidden state of the decoder for the i-th detected region. These feature representations serve as high-level embeddings that encode both spatial and semantic characteristics of anatomical regions. The extracted region features $\mathbf{F}$ are then passed to subsequent model components, including the region selection module and the abnormality classification module, to determine which regions should contribute to the final report. By leveraging DETR's self-attention mechanisms, the model efficiently localizes key anatomical structures, providing a strong foundation for clinically meaningful text generation in later stages.

3.3.4 Region Selection and Abnormality Classification

A multi-layer perceptron (MLP) is used as a region selection module to determine the importance of each detected region for report generation. The selection probability for each region is given by:

$$p_k = \sigma(\text{MLP}_{\text{select}}(\mathbf{f}_k)), \quad k = 1, \ldots, K, \tag{3.3.3}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. Regions with p_k exceeding a predefined threshold are retained for report generation. Additionally, an abnormality classification module, structurally similar to the region selection MLP, is incorporated only during training to refine DETR's focus on clinically relevant regions. The abnormality score for each region is computed as:

$$a_k = \sigma(\text{MLP}_{\text{abnormal}}(\mathbf{f}_k)), \quad k = 1, \ldots, K. \tag{3.3.4}$$

This module is will also improve the quality of representation features extracted by DETR.

3.3.5 Report Generation

The final stage of our model utilizes a GPT-2 Medium decoder to generate textual descriptions of the detected anatomical regions. Given the set of selected regions $\mathcal{R}' = \{r_j\}_{j=1}^{M}$ and their corresponding visual features $X = \{\mathbf{f}_j\}_{j=1}^{M}$, the decoder incorporates both textual and visual context to generate the final report. GPT-2 is an autoregressive neural model that leverages self-attention mechanisms to generate each token based on the context provided by all preceding tokens. Mathematically, the standard self-attention mechanism in GPT-2 is given by:

$$\text{SA}(Y) = \text{softmax}((YW_q)(YW_k)^\top)(YW_v), \tag{3.3.5}$$

where Y represents the token embeddings, and W_q, W_k, W_v denote the query, key, and value projection parameters. To ensure that report generation is informed by the most clinically relevant regions, we make use of pseudo self-attention (PSA) [36] mechanism,

which integrates region-level visual features into the self-attention computation:

$$\mathrm{PSA}(X, Y) = \mathrm{softmax}\left((YW_q)\begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^{\top}\right)\begin{bmatrix} XU_v \\ YW_v \end{bmatrix}, \tag{3.3.6}$$

where X consists of the selected region features $\mathbf{f}_j$, and U_k, U_v are newly initialized key and value projection matrices for the visual features. By integrating pseudo self-attention, the model conditions text generation not only on previously generated tokens but also on visual information extracted from diagnostically relevant regions. This ensures that the final report is both contextually and clinically aligned with the detected abnormalities.

3.3.6 Training

The training procedure is organized into three consecutive stages. Initially, the object detection module is trained independently. In the second phase, the binary classification modules are incorporated, and both the detector and classifiers are jointly optimized. Finally, in the third phase, the entire model undergoes end-to-end training, allowing all parameters to be updated. To optimize the language model, we utilize only those region visual features that correspond to reference sentences in the training data. This ensures that the region selection module is trained to accurately identify regions that require textual descriptions. Note that we freeze the language model and only parameters associated with injecting the feature vectors into the language model are trained. For regions associated with multiple sentences, the target sentences are concatenated to enable the model to learn to generate multiple-sentence outputs where applicable. The total loss function used for training is defined as:

$$\mathcal{L} = \lambda_{\text{obj}} \cdot \mathcal{L}_{\text{obj}} + \lambda_{\text{select}} \cdot \mathcal{L}_{\text{select}} + \lambda_{\text{abnormal}} \cdot \mathcal{L}_{\text{abnormal}} + \lambda_{\text{language}} \cdot \mathcal{L}_{\text{language}}, \quad (3.3.7)$$

where $\mathcal{L}_{\text{obj}}$ represents the loss function for the object detection module, which is based on DETR. The terms $\mathcal{L}_{\text{select}}$ and $\mathcal{L}_{\text{abnormal}}$ denote the weighted binary cross-entropy losses for the region selection and abnormality classification modules, respectively, while $\mathcal{L}_{\text{language}}$ is the cross-entropy loss for the language model.

The loss weights are determined based on performance observed on the validation set and are set as follows: $\lambda_{\text{obj}} = 1.0$, $\lambda_{\text{select}} = 5.0$, $\lambda_{\text{abnormal}} = 5.0$, and $\lambda_{\text{language}} = 2.0$.

3.4 Experiments

Implementation Details. We employ DETR with a ResNet-50 backbone as the initial model for fine-tuning and extracting feature representations of anatomical RoIs. As mentioned, the model is trained in three stages. In the first step, only the DETR is trained, using a learning rate of 10^{-4} for DETR and 10^{-5} for the backbone. In the second step, both DETR and the classifiers are trained. The learning rate for DETR remains the same as in the first step, while the classifiers are trained with a learning rate of 5×10^{-5}. In the third and final step, all parts of the model are trained: DETR is trained as in the first step, and both the classifiers and the language model are trained with a learning rate of 5×10^{-5}. We use the AdamW optimizer [18], with a weight decay of 10^{-4} for DETR and 10^{-2} for the rest of the model. The batch size is 128 for the first two steps and 4 for the third step, which is accumulated to 64. We used one NVIDIA H100 GPU with 80 GB of RAM for training. The first two steps

were trained for 60 epochs, and the last step for 5 epochs.

To filter out duplicate detections of any RoIs, we post-process the DETR output. Based on the confidence score of the model per anatomical region, we remove the duplicate region with the lower confidence score. Both the region selection and abnormality detection modules share the same architecture, consisting of a single feed-forward network (FFN). The input dimension of the FFN is 256, matching the dimension of each feature vector extracted by DETR. The hidden dimension is 64, and the output dimension is 1, indicating whether a region should be selected for report generation or identified as abnormal.

All images are single-channel grayscale, resized to 512×512 while maintaining their original aspect ratio through padding as needed, and normalized to have zero mean and unit variance. Data augmentation is applied to each image with a probability of 50%, including one or more of the following: color jitter with 20% variation in brightness and contrast, Gaussian noise with zero mean and variance sampled from the range $[10, 50]$, or an affine transformation consisting of translation up to $\pm 2\%$ of the image dimensions and rotation up to $\pm 2^{\circ}$.

Dataset. For training and evaluation, we utilize the Chest ImaGenome v1.0.0 dataset [34], which is derived from the MIMIC-CXR dataset [11]. MIMIC-CXR consists of a large collection of chest X-ray images paired with corresponding free-text radiology reports. The Chest ImaGenome dataset extends this by providing automatically generated scene graphs for each frontal chest X-ray in the dataset. Each scene graph encodes information about 29 distinct anatomical regions within the chest, represented as bounding box coordinates. Additionally, it includes textual descriptions

for each region whenever such descriptions are present in the associated radiology report. We follow the official 70/10/20 dataset split, which consists of 166, 512 images for training, 23, 952 for validation, and 47, 389 for testing. We adopt the "findings" section of the radiology reports from the MIMIC-CXR dataset as our reference text, following prior studies. The "findings" section provides a summary of observations recorded by radiologists. To ensure meaningful reference reports, we exclude cases where the "findings" section is empty. After this filtering step, the test set contains 32,711 images with corresponding reference reports.

3.5 Results

To measure the quality of the generated reports by the model, we evaluate at both the report and sentence levels using widely adopted natural language processing (NLP) metrics. At the report level, we compute BLEU [23], METEOR [1], ROUGE-L [14], and CIDEr-D [26], which quantify the similarity between the generated and target reports by measuring n-gram overlap. At the sentence level, we use METEOR, as it is well-suited for both sentence and report-level evaluation, unlike metrics such as BLEU. Table 3.2 shows the results for NLP metrics.

Conventional NLP metrics primarily focus on lexical overlap and may not accurately reflect the clinical correctness of generated reports [3, 17, 24]. To address this limitation, we follow previous works in reporting clinical efficacy (CE) metrics [6, 17, 20]. These metrics evaluate the agreement between generated and reference reports based on the presence status of key clinical observations, providing insight into the diagnostic accuracy of the generated text.

Clinical efficacy (CE) metrics evaluate the context alignment between generated

Region	RL	LL	SP	MED	CS	AB	**Average**
IoU	0.931	0.921	0.875	0.871	0.832	0.923	**0.876**

Table 3.1: Micro-averaged object detection results across six prominent regions: *right lung* (RL), *left lung* (LL), *spine* (SP), *mediastinum* (MED), *cardiac silhouette* (CS), and *abdomen* (AB). Nearly all of the 29 anatomical regions are typically detected in each CXR.

radiology reports and their reference counterparts concerning a set of clinically observations. We follow the previous work which aggregates example-based scores across 14 observations [20]. Table 3.3 shows the results for CE metrics.

To implement these assessments, we first utilize CheXbert [29], a BERT-based [7] information extraction system, to classify each observation into four categories: positive, negative, uncertain, or no mention. The observations include 12 disease types along with "Support Devices" and "No Finding." We convert these multi-class labels into binary classes with grouping positive and uncertain labels as the positive class, while negative and no mention are considered negative. For the final evaluation, example-based precision, recall, and F1 scores are computed over all 14 observations.

3.6 Conclusions

In this study, we investigated the integration of DETR, a transformer-based model, for automated radiology report generation from chest X-ray images. Our approach utilized DETR as a representation learning model to extract comprehensive feature representations for various anatomical regions, facilitating the extraction of meaningful visual information. These extracted features were then leveraged for generating free-text radiology reports.

Our results demonstrate that our method is capable of producing grammatically

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
R2Gen [6]	0.353	0.218	0.145	0.103	0.142	0.277	0.406†
CMN [5]	0.321	0.218	0.148	0.106	0.146	0.278	-
PPKED [15]	0.360	0.224	0.149	0.106	0.149	0.284	0.237
$\mathcal{M}^2$ TR. PROGRESSIVE [21]	0.378	0.232	0.154	0.107	0.145	0.272	-
Contrastive Attention [16]	0.350	0.212	0.152	0.109	0.151	0.271	-
AlignTransformer [35]	0.378	0.235	0.156	0.112	0.158	0.278	-
$\mathcal{M}^2$ Trans w/ NLL [19]	-	-	-	-	-	-	0.445
$\mathcal{M}^2$ Trans w/ NLL+BS+f_{CE} [19]	-	-	-	-	-	-	0.492
$\mathcal{M}^2$ Trans w/ NLL+BS+f_{CEN} [19]	-	-	-	0.114	0.124	-	0.509
ITA [33]	0.395	0.253	0.170	0.121	0.147	0.284	-
CvT-212DistilGPT2 [20]	0.392	0.245	0.169	0.124	0.153	0.285	0.361
RGRG [31]	0.373	0.249	0.175	0.126	0.168	0.264	0.495
Ours	0.369	0.241	0.166	0.118	0.158	0.254	0.491

Table 3.2: Evaluation metrics for the full report generation task using natural language processing (NLP) techniques. Reported metrics include BLEU-1 to BLEU-4, METEOR, ROUGE-L, and CIDEr. A dash (–) indicates unavailable values. The CIDEr score marked with † is referenced from [19].

Method	P	R	$\mathbf{F}_1$
R2Gen	0.331	0.224	0.228
$\mathcal{M}^2$ Trans w/ NLL	-	-	-
$\mathcal{M}^2$ Trans w/ NLL+BS+f_{CE}	-	-	-
$\mathcal{M}^2$ Trans w/ NLL+BS+f_{CEN}	-	-	-
CMN	0.334	0.275	0.278
Contrastive Attention	0.352	0.298	0.303
$\mathcal{M}^2$ TR. PROGRESSIVE	0.240	0.428	0.308
CvT-212DistilGPT2	0.359	0.412	0.384
RGRG [31]	0.461	0.475	0.447
Ours	0.354	0.290	0.301

Table 3.3: Clinical efficacy (CE) metrics example-based averaged over 14 observations. Dashed lines indicate missing values. All results except RGRG are cited from [20].

coherent reports with a quality comparable to state-of-the-art models. While the proposed architecture may exhibit certain limitations in capturing clinical context as effectively as existing SOTA models, the strong potential of transformer-based vision

models suggests promising avenues for future research. Exploring alternative architectures or complementary methodologies could further enhance the clinical accuracy and effectiveness of such models in radiology report generation.

Bibliography

[1] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72 (2005)

[2] Bastawrous, S., Carney, B.: Improving patient safety: avoiding unread imaging exams in the national va enterprise electronic health record. Journal of digital imaging **30**(3), 309–313 (2017)

[3] Boag, W., Hsu, T.M.H., Mcdermott, M., Berner, G., Alesentzer, E., Szolovits, P.: Baselines for chest x-ray report generation. In: ML4H Workshop, pp. 126–140 (2020)

[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – European Conference on Computer Vision (ECCV) 2020. pp. 213–229. Springer International Publishing, Cham (2020)

[5] Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: ACL, pp. 5904–5914 (2021)

[6] Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP, pp. 1439–1449 (2020)

[7] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018), `http://arxiv.org/abs/1810.04805`

[8] Eshraghi Dehaghani, M., Sabour, A., Madu, A.B., Lourentzou, I., Moradi, M.: Representation Learning with a Transformer-Based Detection Model for Localized Chest X-Ray Disease and Progression Detection . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15001. Springer Nature Switzerland (October 2024)

[9] Geis, R., Brady, A.P., Wu, C.C., Spencer, J., Ranschaert, E., Jaremko, J.L., Langer, S.G., Kitts, A.B., Birch, J., Shields, W.F.: Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. Radiology **293**(2), 436–440 (2019)

[10] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5), 1–42 (2018)

[11] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data pp. 1–8 (2019)

[12] Kailin Chu, Amirhossein Sabour, M.E.D.I.L., Moradi, M.: Localized disease severity detection in chest x-ray images. In: IEEE Engineering in Medicine and Biology Society (EMBC) (2025)

[13] Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., Zhai, X.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)

[14] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74–81 (2004)

[15] Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: CVPR, pp. 13753–13762 (2021)

[16] Liu, F., Yin, C., Wu, X., Ge, S.: Ping zhang, andXuSun. contrastiveattention for automatic chest x-ray report generation. In: ACL-IJCNLP, pp. 269–280 (2021)

[17] Liu, G., Hsu, T.M.H., Mcdermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: MLHC, pp. 249–269 (2019)

[18] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2018)

[19] Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: NAACL, pp. 5288–5304 (2021)

[20] Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warmstarting (2022)

[21] Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. In: EMNLP, pp. 2824–2832 (2021)

[22] Papanikolaou, Y., Pierleoni, A.: Dare: Data augmented relation extraction with gpt-2 (2020)

[23] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318 (2002)

[24] Pino, P., Parra, D., Messina, P., Besa, C., Uribe, S.: Inspecting state of the art performance and nlp metrics in image-based medical report generation (2020)

[25] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8) (2019)

[26] Ramakrishna Vedantam, C.L., Zitnick, D.: Cider: Consensus-based image description evaluation. In: CVPR, pp. 4566–4575 (2015)

[27] Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ (2017)

[28] Rosenkrantz, A.B., Hughes, D.R., Duszak, R.: The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology **279**(1), 175–184 (2016)

[29] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In: EMNLP, pp. 1500–1519 (2020)

[30] Stacy, K., Goergen, F.J., Pool, T.J., Turner, J.E., Grimm, M.N., Appleyard, C., Crock, M.C., Fahey, M.F., Fay, N.J., Ferris, S.M.: Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. JMIRO **57**(1), 1–7 (2013)

[31] Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 7433–7442. IEEE (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.00718, `http://dx.doi.org/10.1109/CVPR52729.2023.00718`

[32] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), `http://arxiv.org/abs/1706.03762`

[33] Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., Chen, J.: An inclusive task-aware framework for radiology report generation. In: MICCAI, pp. 568–577 (2022)

[34] Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W.G., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest imagenome dataset for clinical reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)

[35] You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: MICCAI, pp. 72–82. Springer (2021)

[36] Ziegler, Z.M., Melas-Kyriazi, L., Gehrmann, S., Rush, A.M.: Encoder-agnostic adaptation for conditional language generation (2019)

Chapter 4

Conclusion

In this thesis, we explored the use of upstream transformer-based vision models for representation learning to interpret and extract clinically meaningful information from chest X-rays (CXRs). We proposed a novel architecture that leverages these learned representations to perform various downstream tasks, including localized disease classification and progression monitoring. Additionally, we extended this approach to investigate its potential for generating free-text clinical radiology reports from chest X-rays.

In Chapter 2, we discussed the limitations of previous models in accurately identifying localized pathologies within CXRs and emphasized the clinical importance of tracking disease progression. We introduced the novel task of localized disease progression monitoring and proposed a simple yet effective architecture capable of jointly performing both classification and progression monitoring. Utilizing DETR as a backbone for representation learning, our model achieved an average ROC of 89.1% across nine pathological findings for localized disease detection, which is competitive

with existing state-of-the-art methods. Furthermore, our model attained an average accuracy of approximately 67% and an F1 score of around 71% in the localized progression monitoring task.

In Chapter 3, we investigated the ability of our approach to generate high-quality radiology reports from CXRs. We utilized the same representation learning architecture used in Chapter 2 and designed a dual-network setup consisting of two identical neural networks: one responsible for selecting relevant image regions to describe, and another, used only during training, to refine the model's focus on abnormal regions. These components were coupled with a large language model to form the downstream model architecture and be used to generate coherent and clinically relevant report text. Our experiments demonstrated that the proposed method outperformed or was competitive with most of the existing models in terms of grammatical fluency and in terms of clinical efficacy metrics.

Despite the promising results, our method has limitations, particularly in report generation. In certain clinical efficacy metrics, our approach underperforms compared to methods such as RGRG. Future work could explore alternative architectures or enhancements to our current framework to address these shortcomings. Additionally, there is potential to extend this work by generating more structured and interpretable clinical reports, which could aid radiologists in more robust and systematic analysis of chest X-rays.