

A Hybrid Approach for Automated Radiology Report Generation and Summarization using Vision Transformers and Language Models

Submitted By
Kheem Parkash Dharmani
22I-0081

Supervised By
Dr. Ejaz Ahmed
Co-Supervised By
Dr. Ahmad Raza Shahid
Master of Science (Data Science)

A thesis submitted in partial fulfillment of the requirements for the degree of
Masters of Science (Artificial Intelligence)
at National University of Computer & Emerging Sciences



Department of Computer Science
National University of Computer and Emerging Sciences

Islamabad, Pakistan

May, 2024



Plagiarism Undertaking

I take full responsibility of the research work conducted during the MS Thesis titled A Hybrid Approach for Automated Radiology Report Generation and Summarization using Vision Transformers and Language Models. I solemnly declare that the research work presented in the thesis is done solely by me with no significant help from any other person; however, small help wherever taken is duly acknowledged. I have also written the complete thesis by myself. Moreover, I have not presented this thesis (or substantially similar research work) or any part of the thesis previously to any other degree awarding institution within Pakistan or abroad.

I understand that the management of National University of Computer and Emerging Sciences has a zero tolerance policy towards plagiarism. Therefore, I as an author of the above-mentioned thesis, solemnly declare that no portion of my thesis has been plagiarized and any material used in the thesis from other sources is properly referenced. Moreover, the thesis does not contain any literal citing of more than 70 words (total) even by giving a reference unless I have the written permission of the publisher to do so. Furthermore, the work presented in the thesis is my own original work and I have positively cited the related work of the other researchers by clearly differentiating my work from their relevant work.

I further understand that if I am found guilty of any form of plagiarism in my thesis work even after my graduation, the University reserves the right to revoke my MS degree. Moreover, the University will also have the right to publish my name on its website that keeps a record of the students who plagiarized in their thesis work.

Kheem Parkash Dharmani

Date:



Author's Declaration

I, Kheem Parkash Dharmani, hereby state that my MS thesis titled A Hybrid Approach for Automated Radiology Report Generation and Summarization using Vision Transformers and Language Models is my own work and it has not been previously submitted by me for taking partial or full credit for the award of any degree at this University or anywhere else in the world. If my statement is found to be incorrect, at any time even after my graduation, the University has the right to revoke my MS degree.

Kheem Parkash Dharmani

Date: _____



Certificate of Approval



*It is certified that the research work presented in this thesis, entitled “A Hybrid Approach for Automated Radiology Report Generation and Summarization using Vision Transformers and Language Models” was conducted by **Kheem Parkash Dharmani** under the supervision of*

Dr. Ejaz Ahmed

No part of this thesis has been submitted anywhere else for any other degree.

*This project is submitted to the **FAST School of Computing** in partial fulfillment of the requirements for the degree of Master of Science in “Data Science” at the*

National University of Computer and Emerging Sciences Islamabad, PAKISTAN

16/05/2024

Candidate Name: Kheem Parkash Dharmani

Signature: _____

Examination Committee:

a) Name: Dr. Basit Shahzad Signature: _____
Professor, NUML University, Islamabad

b) Name: Ms. Nirmal Tariq Signature: _____

Lecturer, FAST NUCES University, Islamabad

Supervisor:

c) Name: Dr. Ejaz Ahmed Signature: _____
Professor, FAST NUCES University, Islamabad

Head, FAST School of Computing, National University of Computer and Emerging Sciences, Islamabad



Abstract

The burgeoning field of artificial intelligence (AI) in medical imaging has opened new avenues for automating the generation of radiology reports, a critical component in diagnostic medicine. This research presents a novel approach to radiology report generation, utilizing a synergistic integration of a Vision Transformer (ViT) and a fine-tuned Generative Pretrained Transformer 2 (GPT-2). The ViT model is employed to extract detailed features from X-ray images, while the GPT-2, pre-trained on a comprehensive dataset of radiology texts, is used for generating contextually relevant reports. This study addresses the limitations of existing automated systems, particularly in capturing global image features and generating reports with accurate medical terminologies. The proposed hybrid model aims to produce radiology reports that are not only precise and informative but also tailored to the specific needs of individual cases. Evaluation metrics such as BLEU score, ROUGE scores, and BERTScore indicate the high accuracy and reliability of the generated reports. The results demonstrate the potential of hybrid integration through conditional prompting approach in enhancing diagnostic efficiency and supporting radiologists, marking a significant step forward in AI-assisted radiology.



Acknowledgements

I extend my deepest gratitude to my advisor, Dr. Ejaz Ahmed, whose expert guidance and steadfast support were indispensable throughout my research journey. His wisdom and patience were crucial in shaping both the direction and completion of this thesis. I am equally thankful to my father, Girdhari Dharmani, and my mother, Dheli Dharmani, whose love and encouragement have been my constant source of strength and motivation. Their unwavering belief in my abilities has been instrumental in my academic journey.



List of Abbreviations

AI	Artificial Intelligence
ARGG	Automated Radiology Report Generation
ML	Machine Learning
ViT	Vision Transformer
GPT-2	Generative Pretrained Transformer 2
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
BP	Brevity Penalty
LM	Language Model
ASGKED	Auxiliary Signal-Guided Knowledge Encoder-Decoder
EHR	Electronic Health Record
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
LLMs	Large Language Models
CXR	Chest X-Ray
MeSH	Medical Subject Heading



List of Figures

Figure 1: System architecture	23
Figure 2: Word Cloud Visualization Highlighting Key Terms in the IU X-ray Dataset.	25
Figure 3: Sample of Preprocessed X-Ray Images.....	26
Figure 4: Distribution of Radiology Report Lengths Post-Preprocessing.....	27
Figure 5: Frequency Distribution of Top Words and Bi-grams in Radiology Reports.	28
Figure 6: Co-occurrence Matrix Showing Relationships Between Key Medical Terms in Reports.	29
Figure 7: Vision Transformer (ViT) Workflow	30
Figure 8: t-SNE plot of the extracted features, illustrating the emergent clusters and the separability of the image representations.....	31
Figure 9: Architectural Diagram of the Fine-Tuned GPT-2 Model.	32
Figure 10: Workflow Diagram of the Integrated Vision Transformer and GPT-2 Model for Radiology Report Generation.....	34
Figure 11: Fine Tuning Process Loss vs Epoch	35



List of Tables

Table 1: Summary of the related work.....	17
Table 2 Example Evaluation Results	38



Table of Contents

Abstract	5
1. Introduction	12
1.1 Background	12
1.2 Motivation	12
2. Literature Review	13
2.1 Research Gap	22
2.2 Problem Statement	22
2.3 Research Objectives/Research Questions	22
3. Methodology	23
3.1 Dataset	24
3.1.1 Dataset Overview	24
3.1.2 X-Ray Image Analysis	25
3.1.3 X-Ray Report Analysis	26
3.2 X-Rays Image Feature Extraction with Vision Transformer (ViT)	29
3.2.1 Vision Transformer Architecture	29
3.2.2 Training the Vision Transformer	29
3.2.3 Feature Extraction Process	30
3.3 Radiology Reports Processing with GPT-2	31
3.3.1 Fine-Tuning GPT-2 on Radiology Reports	31
3.3.2 Generating Radiology Reports	33
3.4 Integration of Image and Text Models	33
3.4.1 Integrating ViT and GPT-2	33
3.4.2 Synchronized Training Approach	34
3.4.3 Report Generation and Integration	34
3.5 Evaluation Metrics	35
3.5.1 Perplexity	35
3.5.2 BLEU Score	36
3.5.3 ROUGE Scores	36
3.5.4 BERTScore	36
3.5.5 Evaluation Example	36
4. Results and Discussion	40
4.1 Overview of Results	40
4.2 Detailed Results Analysis	40
4.2.1 Quantitative Results	40
4.2.2 Qualitative Analysis	40
4.3 Comparison with State-of-the-Art	40
4.3.1 Methodological Comparisons	40
4.3.2 Performance Comparisons	40
4.4 Discussion	41
4.4.1 Interpretation of Results	41
4.4.2 Implications for Clinical Practice	41



4.4.3	Limitations and Challenges	41
5.	Conclusion and Future Work	41
5.1	Summary of Contributions	41
5.2	Future Work Directions	42
5.2.1	Technical Advancements	42
5.2.2	Potential Applications	42
5.3	Final Thoughts	42
	References	43



1. Introduction

The integration of artificial intelligence (AI) into healthcare represents a significant paradigm shift, promising to redefine numerous aspects of patient diagnosis, treatment, and overall medical management. Among various applications, the automation of medical report generation, particularly in the field of radiology, stands as a beacon of potential, aiming to enhance diagnostic accuracy and streamline clinical workflows. This chapter sets the stage for a discussion on the evolving role of AI in healthcare, emphasizing the challenges and opportunities within the domain of automated radiology report generation.

1.1 Background

Medical reports are essential tools in the diagnostic process, providing clinicians with the insights needed to make informed treatment decisions. Traditionally, the generation of these reports has been a meticulous task requiring substantial time and expertise from skilled radiologists. Each report must be both precise in its medical terminology and comprehensive in its coverage of diagnostic findings, a requirement that places a considerable burden on healthcare professionals.

In recent years, the rapid evolution of deep learning technologies has opened new avenues for addressing the inefficiencies associated with manual report generation. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been at the forefront, driving innovations in image recognition and natural language processing. These technologies, however, while pioneering, are not without their limitations. They often struggle to fully capture the nuanced interrelations in medical images and may generate reports that lack relevance or specificity to the patient's actual condition.

1.2 Motivation

This research is motivated by the necessity to overcome these limitations and enhance the quality of automated radiology report generation. The current state of technology, while promising, is hindered by the limited accuracy of large language models not specifically pre-trained on radiological texts. This gap in technology leads to reports that may lack accuracy, completeness, or relevance to the patient's condition.

To bridge this gap, we propose a novel approach, leveraging a hybrid deep learning model that combines a Vision Transformer (ViT) for intricate image feature extraction with a transformer-based language model for report generation. This model is further enhanced through pre-training on a comprehensive dataset of radiology books, to improve its understanding of relevant medical vocabulary and concepts. The integration of these technologies aims to produce a model capable of generating more accurate, informative, and tailored radiology reports, advancing the field of automated medical report generation.

Furthermore, the use of advanced metrics like BLEU scores, ROUGE scores, and BERTScore for evaluation underlines the rigor with which these automated systems are assessed, ensuring that they meet the high standards required for clinical deployment. Through this research, we aim to contribute to the body of knowledge in AI-assisted radiology, demonstrating how the thoughtful integration of cutting-edge AI technologies can lead to substantial improvements in medical diagnostics.

This thesis proposes a novel hybrid approach that synergistically combines the visual analytical power of Vision Transformers with the linguistic fluency of transformer-



based language models. By harnessing these technologies, the proposed system aims to automate the generation of radiology reports that are not only accurate but also nuanced and patient-specific. Such advancements could significantly alleviate the workload of radiologists, allowing them to focus more on patient care rather than administrative tasks.

2. Literature Review

Medical report generation is an essential task in healthcare, which provides an accurate and comprehensive diagnosis and treatment plan for patients. However, generating medical reports is a time-consuming and challenging task for medical professionals. The emergence of deep learning techniques has led to automated medical report generation systems that can generate reports quickly and accurately. In this literature review chapter, we discuss thirteen different studies that propose automated medical report generation systems. The summary of literature review is included in Table 1.

[Mohsan et al. \(2022\)](#) proposed a new approach to generate radiology reports using deep learning models. The authors used a combination of Vision Transformer (ViT) and Language Model (LM) techniques to generate the reports. They also introduced a new dataset called "X- Ray14K" which is a collection of chest X-ray images and their corresponding radiology reports. The primary contribution of this study is the proposed methodology that involves two steps: first, the chest X-ray image is processed using ViT to extract features, and then LM is used to generate the report based on the extracted features. The authors used the pre-trained ViT and LM models and fine-tuned them on their X-Ray14K dataset. One of the limitations of this study is that it focuses only on chest X-ray images and their corresponding radiology reports. The proposed approach may not generalize well to other medical imaging modalities or other types of radiology reports. The authors suggest that their approach can be extended to other medical imaging modalities, such as MRI or CT scans, and future work can focus on developing a more comprehensive dataset that includes different imaging modalities and different types of radiology reports.

[Li et al. \(2023\)](#) proposed an auxiliary signal-guided knowledge encoder-decoder (ASGKED) model for medical report generation. The authors leveraged auxiliary signals such as diagnosis codes, lab test results, and medical concepts to guide the generation of medical reports. The model consists of two components: a knowledge encoder and a report decoder. The knowledge encoder extracts relevant information from the auxiliary signals and generates a knowledge representation, while the report decoder uses the knowledge representation to generate the medical report. The authors used a dataset of electronic health records (EHRs) to train and evaluate their ASGKED model. One limitation of this paper is the lack of interpretability of the proposed model. While the ASGKED model achieves better performance than the baseline models, it is unclear how the model generates the medical reports and what specific information it considers when generating the reports. The authors suggest investigating the interpretability of the ASGKED model in future work.

[Sirshar et al. \(2021\)](#) introduced a novel automated radiology report generation system leveraging attention mechanisms in conjunction with convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. Their model was evaluated using an extensive dataset comprising chest X-ray images and their respective radiology reports. By integrating attention mechanisms, the model was capable of prioritizing pertinent regions within the images during report generation. However, a



notable limitation lies in the study's exclusive focus on chest X-ray images and their corresponding reports. Future endeavors could entail constructing a more expansive dataset encompassing various imaging modalities and diverse radiology report types for a more comprehensive assessment.

[Zhang et al.](#) propose an innovative deep learning framework tailored for enhancing medical report generation by integrating inter-intra information calibration. Their model comprises an encoder-decoder architecture along with an information calibration module. Leveraging the transformer architecture for the encoder-decoder framework, renowned for its efficacy in natural language processing tasks, they augment it with a multi-layer perceptron (MLP) neural network within the information calibration module to learn inter-intra information nuances. Although the model exhibits promising results on chest X-ray reports using the MIMIC-CXR dataset, its applicability warrants validation across diverse medical report types to ascertain generalizability. Furthermore, the reliance on extensive training data poses challenges in resource-constrained medical environments, suggesting avenues for refining the model's scalability and adaptability.

[Yan and Pei et al.](#) propose Clinical-BERT, a vision-language pre-training model tailored for addressing radiograph diagnosis and report generation challenges. Centered on the BERT architecture, their model learns joint representations of vision and language through pre-training on a sizable radiograph dataset. While demonstrating efficacy across radiograph classification and report generation tasks, its evaluation solely on chest radiographs underscores the need for broader assessment across diverse radiograph types to gauge robustness. Moreover, the absence of error analysis impedes comprehensive interpretation of results, urging for deeper insights into model performance and generalization capabilities across varied radiograph domains. A newly available dataset, MIMIC-CXR-JPG, offers a rich resource for medical computer vision research, comprising de-identified chest X-ray images and corresponding reports. The dataset's utility extends to facilitating comparative studies and enhancing algorithmic development in medical imaging. Additionally, discussions encompassing medical imaging databases' efficiency and application contexts provide valuable insights for researchers and practitioners alike.

A sizable dataset called [MIMIC-CXR-JPG](#) was obtained from the Beth Israel Deaconess Medical Center between 2011 and 2016 and contains 377,110 chest x-rays linked to 227,827 imaging investigations. 14 labels are provided for the photos, which were created using two natural language processing algorithms and the related free-text radiology reports. The dataset is made publicly available for use in any medical computer vision research project and has been de-identified to protect patient privacy. The dataset's main objectives are to give JPEG images created from DICOM files a common reference and to offer a convenient MIMIC-CXR processed version. The VAEnterprise Electronic Health Record, medical imaging in Rwanda and Liberia, the efficiency of several databases utilized in medical imaging, as well as other resources, are also discussed in the paper.

[W. Jhonson et al.](#) propose Sparse Transformer, a novel method aimed at efficient and effective training of large language models (LLMs). By leveraging a sparse attention mechanism, the Sparse Transformer enables LLMs to prioritize crucial input data segments, resulting in accelerated training without compromising accuracy. While promising, the method warrants comprehensive evaluation to uncover potential limitations and assess its suitability across diverse applications. Moreover, its increased computational overhead necessitates consideration of resource implications, highlighting the need for



further optimization efforts.

[Bannur et al. \(2023\)](#) present a new method named Temporal Transformer for learning temporal structures in biomedical vision-language tasks. This method leverages a self-attention mechanism to capture long-range dependencies in video frames, improving the model's grasp of event sequences, crucial for medical diagnosis and surgery planning. Their approach demonstrates superior performance across various tasks compared to current methods. Nonetheless, the method is computationally demanding and may not fit all scenarios. Furthermore, its evaluation was limited to a few tasks, leaving its broader applicability uncertain. Overall, this method offers a promising advancement in biomedical vision-language processing. The authors have shown that their method can be used to improve the performance of a variety of tasks, and it is likely to be useful for many other applications in the future.

[Cai et al. \(2023\)](#) develop ChestXRyBERT, a method for summarizing chest radiology reports. This pretrained language model is trained on extensive radiology reports and images, learning semantic relationships to generate accurate summaries. Evaluated on a dataset of chest radiology reports, it outperforms existing methods in ROUGE-L and METEOR scores. However, its evaluation on a limited dataset indicates the need for validation on larger datasets to confirm its effectiveness.

[Wu et al. \(2023\)](#) analyze the trade-offs between unified large language models (LLMs) and locally fine-tuned models for specific radiology natural language inference (NLI) tasks. They find that LLMs can achieve similar or better results than local models with much less data, highlighting LLMs' potential in data-limited radiology NLI tasks. This study provides insights into performance differences between LLMs and local models regarding size, training time, and performance. However, the limitation of the study is that it only evaluated the models on a single task and dataset, and further testing on other tasks and datasets is needed to assess the generalizability of the findings.

[Ma et al. \(2023\)](#) propose ImpressionGPT, an iterative optimization framework using ChatGPT for radiology report summarization. The framework involves three steps: pre-training on a large text corpus, iterative fine-tuning on radiology reports, and testing on a held-out set. ImpressionGPT outperforms existing methods, producing accurate and informative summaries. However, its evaluation on a single dataset underscores the need for broader testing to validate its generalizability. The contribution of the study lies in the development of a new framework that can generate high-quality radiology report summaries with improved accuracy and efficiency. However, the limitation of the study is that it was only evaluated on a single dataset, and further testing on other datasets is needed to assess the generalizability of the proposed framework.

[Jeong et al. \(2023\)](#) purposed a novel method for chest X-ray report generation called Multimodal Image-Text Matching (MMIT). MMIT is a retrieval-based approach that generates summaries of chest X-ray images by employing a multimodal image-text matching model. The model has two components: picture-text encoder: This component encodes both the picture and the text into a shared embedding space. The retrieval model employs the joint embedding space to obtain the most relevant text from a pre-trained language model. The authors demonstrate that MMIT outperforms cutting-edge approaches for generating chest X-ray reports. They also demonstrate that MMIT can produce summaries that are both accurate and informative. The author introduces a novel method for generating chest X-ray reports based on multimodal image-text matching. The



study suggests a novel.

[Harrer \(2023\)](#) explores the ethical implications of utilizing large language models (LLMs) in healthcare and medicine. He acknowledges the significant potential benefits of LLMs in these fields but highlights several ethical concerns. A primary concern is the possibility of LLMs generating biased or discriminatory medical advice. For instance, if an LLM is trained on a dataset that overrepresents a particular race or gender, it might produce biased medical recommendations, leading to discriminatory treatment. Another issue is the risk of LLMs spreading misinformation about health and medicine. An LLM could generate fake news articles or social media posts with false or misleading health information, which could result in harmful health decisions by the public. Harrer emphasizes the necessity of carefully considering these ethical implications before deploying LLMs in real-world healthcare settings. He also stresses the importance of developing ethical guidelines for the creation and use of LLMs in healthcare and medicine.

[Wang et al. \(2022\)](#) propose a framework that combines a vision transformer and a language model to generate radiology reports using the MIMIC-III dataset. Their model outperforms existing models in BLEU-4 and ROUGE-L scores. The study includes detailed performance analysis and an ablation study, but further research is necessary to address limitations and evaluate performance on other datasets.



Table 1: Summary of the related work.

Ref. No., Year	Methodology/Approach	Strengths	Weaknesses	Dataset	Evaluation Metrics
[1], 2022	<ul style="list-style-type: none">• Vision Transformer and Language Model Based Radiology Report Generation	<ul style="list-style-type: none">• Utilizes a Vision Transformer to extract features from medical images.• Employs a language model to generate radiology reports based on the extracted features.• This approach effectively produces accurate and informative radiology reports.	<ul style="list-style-type: none">• The proposed approach is computationally expensive• The proposed approach requires a large amount of training data	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	<ul style="list-style-type: none">• BLEU SCORE• 0 to 1• ROUGE• METEOR
[2], 2023	<ul style="list-style-type: none">• Auxiliary signal-guided knowledge encoder-decoder	<ul style="list-style-type: none">• Incorporates external knowledge into the encoder-decoder framework• Provides explanations for the generated reports through attention maps	<ul style="list-style-type: none">• May require substantial amounts of external knowledge for optimal performance• May have difficulty generating rare or uncommon diagnoses	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	<ul style="list-style-type: none">• BLEU Score• CIDEr
[3], 2022	<ul style="list-style-type: none">• Attention-based automated radiology report generation using CNN and LSTM	<ul style="list-style-type: none">• Achieved state-of-the-art accuracy on two publicly available datasets.• Is able to generate comprehensive and informative radiology reports.• Is efficient and can be used to generate reports in real time.	<ul style="list-style-type: none">• Is still under development and may not be able to handle all types of radiology cases.• Requires a large amount of training data to achieve high accuracy.	<ul style="list-style-type: none">• CXR	<ul style="list-style-type: none">• BLEU Score• CIDEr



[4], 2023	<ul style="list-style-type: none">• Information Calibrated Transformer (ICT) for medical report generation	<ul style="list-style-type: none">• Extracts multiple inter- and intra-report features directly from the datasets as auxiliary information, seamlessly integrating them into the training process.• The auxiliary information remains dynamic, continuously updated throughout training.	<ul style="list-style-type: none">• The ICT is not only superior to previous methods in the X-Ray datasets, IU-X-Ray and MIMIC-CXR	<ul style="list-style-type: none">• IU CheXpert	<ul style="list-style-type: none">• BLEU Score• ROUGE
[5], 2022	<ul style="list-style-type: none">• Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation	<ul style="list-style-type: none">• A vision-language pre-training model tailored for medical applications.• Introduces three domain-specific tasks: Clinical Diagnosis (CD), Masked MeSH Modeling (MMM), and Image-MeSH Matching (IMM), in addition to a general pre-training task: Masked Language Modeling (MLM).• Demonstrates superior performance on Radiograph Diagnosis and Report Generation tasks across four demanding datasets, setting new benchmarks.	<ul style="list-style-type: none">• The model is complex and requires a large amount of training data to achieve high accuracy.• The model is still under development and may not be able to handle all types of medical cases.	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	<ul style="list-style-type: none">• AUC-ROC



[6], 2019	<ul style="list-style-type: none">• MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs	<ul style="list-style-type: none">• A large dataset of chest radiographs, associated with free-text radiology reports.• The dataset is publicly available and can be used for research purpose.• The dataset is well-curated and includes a variety of imaging modalities.	<ul style="list-style-type: none">• The dataset is relatively small compared to other medical imaging datasets.• The dataset is only available for chest radiographs.	<ul style="list-style-type: none">• Dataset Description	<ul style="list-style-type: none">• AUC-ROC• BLEU Score
[7], 2023	<ul style="list-style-type: none">• Interactive computer-aided diagnosis using large language models	<ul style="list-style-type: none">• Can potentially improve diagnostic accuracy and efficiency by providing more comprehensive and accessible information• Utilizes a large language model (GPT-3) to generate natural language descriptions of medical images for interactive computer-aided diagnosis	<ul style="list-style-type: none">• Requires large amounts of high-quality data to train the language model and may be limited by biases in the training data	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	<ul style="list-style-type: none">• BLEU Score• CIDEr• ROUGE
[8], 2023	<ul style="list-style-type: none">• Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing	<ul style="list-style-type: none">• Proposes a novel framework for biomedical vision-language processing that exploits temporal structure.• The framework is able to learn long-range dependencies between frames and words, which improves performance on a variety of tasks.	<ul style="list-style-type: none">• The framework is computationally expensive.• It requires a large amount of training data.	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	<ul style="list-style-type: none">• BLEU Score



[9], 2023	<ul style="list-style-type: none">• Pretrained language model (ChestXRayBERT) for chest radiology report summarization	<ul style="list-style-type: none">• Utilized transformer-based models and multi-task learning approach to summarize chest radiology reports• Achieved state-of-the-art performance on two public datasets, i.e., MIMIC-CXR and OpenI	<ul style="list-style-type: none">• The quality of the generated summary highly depends on the quality of the input report• The model is trained on publicly available datasets, which may not be fully representative of clinical data	• ChestXray	• BLEU Score
[10], 2023	<ul style="list-style-type: none">• Comparative study of unified large language models and local fine-tuned models for radiology natural language inference task• Trains and evaluates multiple models with different architectures and training strategies	<ul style="list-style-type: none">• Shows the potential benefits and trade-offs of using large language models versus local fine-tuned models for specific tasks• Provides insights on the effects of model size and training data size on performance	• Limited to one specific NLI task and dataset	• Survey	• Survey
[11], 2023	<ul style="list-style-type: none">• Iterative optimizing framework for radiology report summarization with ChatGPT	<ul style="list-style-type: none">• Utilizes an iterative optimization approach to improve the quality of generated summaries• Incorporates a ChatGPT model to generate high-quality summaries	• Only evaluated on a single dataset	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images	• BLEU Score



[12], 2023	<ul style="list-style-type: none">• Multimodal image-text matching to improve retrieval-based chest X-ray report generation	<ul style="list-style-type: none">• Improved accuracy and consistency of report generation• Reduced reliance on human annotators• Scalable to large datasets	<ul style="list-style-type: none">• Requires large amounts of training data	<ul style="list-style-type: none">• IU Chest Xray	<ul style="list-style-type: none">• BLEU Score• CIDEr• ROUGE
[13], 2023	<ul style="list-style-type: none">• Large language models (LLMs) in healthcare and medicine	<ul style="list-style-type: none">• Potential to improve efficiency and accuracy of tasks such as data extraction, literature review, and clinical decision support• Ability to generate new insights and hypotheses• Potential to democratize access to healthcare knowledge	<ul style="list-style-type: none">• LLMs are trained on massive datasets of text and code, which can include biased or harmful content• LLMs can be used to generate text that is indistinguishable from human-written text, which can be used to spread misinformation or create deepfakes• LLMs are still under development, and their capabilities and limitations are not fully understood	<ul style="list-style-type: none">• MIMIC CXR-JPG• 370,000 chest X-ray images• IU Chest Xray• Survey	<ul style="list-style-type: none">• Survey



2.1 Research Gap

In the evolving landscape of radiology report generation using artificial intelligence, significant advancements have been made, yet critical gaps persist. The literature reveals that while approaches like Vision Transformer and Language Model-based generation [Mohsan et al. \(2022\)](#) have shown promise in accuracy, they often grapple with high computational demands and a substantial need for training data. Auxiliary signal-guided frameworks [Li et al. \(2023\)](#) and attention-based models using CNN and LSTM [Sirshar et al. \(2021\)](#) have introduced innovations like external knowledge incorporation and real-time reporting. However, these systems still face challenges in handling diverse or rare medical cases and require extensive external data for optimal performance. Even advanced models like the Information Calibrated Transformer [Zhang et al.](#) and Clinical-BERT [Yan and Pei et al.](#) are constrained by their complexity and heavy data requisites, limiting their practical applicability. This analysis underscores a distinct research gap: the need for an AI-driven radiology report generation method that balances computational efficiency with high accuracy, reducing dependence on large external datasets and addressing the diversity in medical scenarios. The proposed study aims to bridge this gap by integrating Vision Transformer and GPT-2 models using conditional prompts from image features, offering a novel, efficient solution to generate contextually relevant radiology reports.

2.2 Problem Statement

Addressing the intricacies of radiological imagery and report synthesis, this research transcends the conventional automated report generation by introducing a novel multimodal approach. Distinctively, it integrates the high-resolution feature detection capabilities of a Vision Transformer (ViT) with the contextual comprehension of a fine-tuned GPT-2 model, leveraging conditional prompts derived from image features to guide report generation. This method not only promises a significant reduction in radiologists' workload but also a new benchmark in the accuracy and contextual relevance of automated radiology reports, setting it apart from existing methodologies.

2.3 Research Objectives/Research Questions

Research Objectives

- The primary objective is to create a novel AI model that combines the capabilities of ViT for extracting meaningful features from X-ray images and a fine-tuned GPT-2 model for generating accurate and contextually relevant radiology reports.
- Aim to significantly improve the precision and speed of report generation compared to traditional methods, reducing the turnaround time for radiology diagnostics.
- Investigate how the extracted image features can be used as conditional prompts for the GPT-2 model, enhancing the model's ability to generate reports that are closely aligned with the visual findings in the X-ray images.

Research Questions

- How effectively can a Vision Transformer extract pertinent features from X-ray images for use in report generation?



- b) What is the impact of fine-tuning a GPT-2 model with radiology reports on the quality and relevance of the generated text?
- c) How does the integration of image features and text generation via conditional prompts enhance the overall quality of radiology reports compared to traditional methods?

3. Methodology

This chapter outlines the methodology employed in developing the hybrid AI model for automated radiology report generation, as depicted in the system architecture [Figure 1](#). The design of our model integrates two primary components: a Vision Transformer (ViT) for advanced image feature extraction and a fine-tuned Generative Pretrained Transformer 2 (GPT-2) for generating nuanced and medically accurate report text. The methodology is structured to first detail the preprocessing of radiological images, followed by the training and fine-tuning processes of the ViT and GPT-2 models respectively. Special attention is given to the strategies adopted for integrating these models to work in concert, ensuring that the visual data extracted by the ViT effectively informs the text generation process in GPT-2. The chapter further discusses the setup of the experimental environment, including data acquisition, model configuration, and the definition of performance metrics, which are critical for evaluating the efficacy and accuracy of the generated radiology reports. This systematic approach not only enhances the robustness of the automated system but also ensures its relevance and adaptability to real-world medical settings.

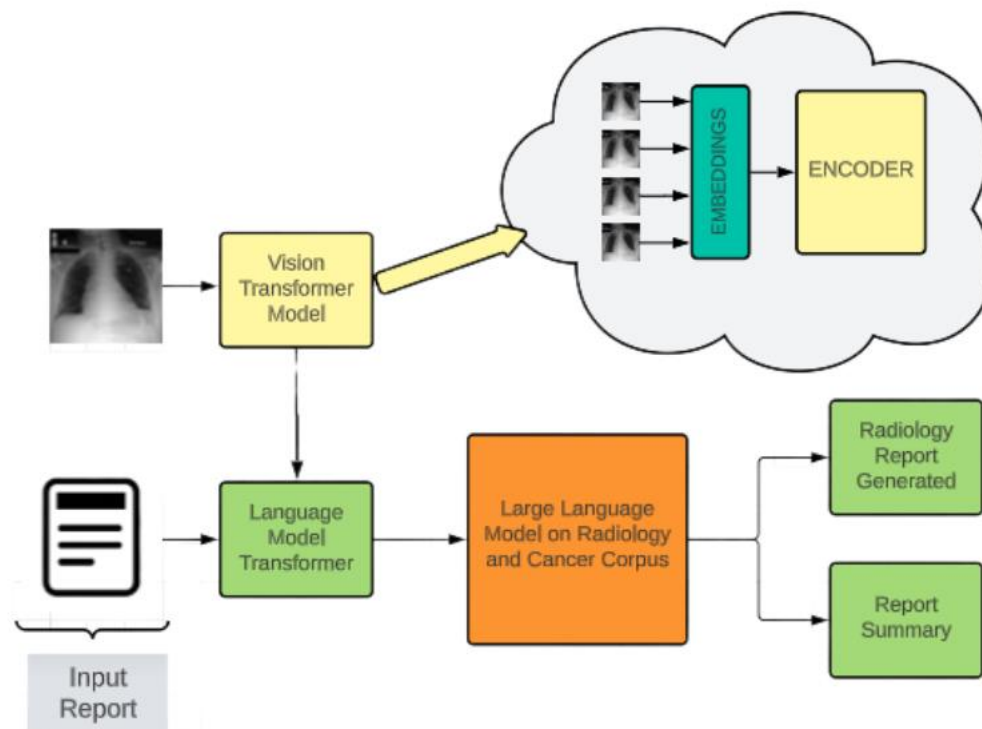


Figure 1: System architecture



3.1 Dataset

3.1.1 Dataset Overview

The dataset used is IU X-ray dataset ([OpenI](#)) a resource for the development of automatically generated radiology reports. The reason behind its creation is the increasing need for intelligent systems that can analyse medical images and create precise reports. The necessity to offer a comprehensive source of annotated medical images to support the advancement of cutting-edge deep learning algorithms led to the establishment of the dataset. Its application is critical when it comes to automating radiological analysis, a process that has historically required a great deal of time and the knowledge of qualified medical personnel. With its extensive collection of radiological reports and X-ray images, the dataset serves as a vital resource for studies in radiology report generation rapidly evolving field of healthcare technology.

The Word Cloud visualization, as shown in [Figure 2](#), vividly illustrates the most frequent terms within the IU X-ray dataset. This visual aid emphasizes key medical concepts and terminology that are foundational to the reports generated from the dataset, underscoring the dataset's rich linguistic and clinical composition.



Figure 2: Word Cloud Visualization Highlighting Key Terms in the IU X-ray Dataset.

3.1.2 X-Ray Image Analysis

The X-ray images in the dataset under analysis are typical of clinical chest radiographs. The images are produced using a variety of imaging modalities, principally the traditional projections used in chest radiology, the posteroanterior (PA) and lateral views. The dataset comprises individual image identifiers and metadata that characterize the image's characteristics, including patient placement, exposure parameters, and the presence of anatomical markers.

3.1.2.1 X-Ray Preprocessing Techniques

In order to get the data ready for input into the suggested hybrid deep learning model, preprocessing is essential. Python was used for implementing the preprocessing

pipeline, with the help of packages like PyTorch and PIL. The following are the main preprocessing steps that were used on the X-ray images:

Resizing: Every image is scaled while keeping the original aspect ratio to a uniform 256 pixels along the shortest edge. To guarantee that the input data matches the model architecture without distortion, this standardization is crucial.

Center Cropping: Following scaling, a 224x224 pixel central crop is taken out. In chest X-rays, the center area of the image usually contains the most diagnostically useful information. This phase concentrates on that area of the image.

Conversion to Tensor: To enable deeper learning framework processing, the PIL format images are transformed to PyTorch tensors.

Normalization: The predetermined mean and standard deviation are used to normalize the image pixel values. In addition to aligning the data distribution to a common scale, this normalization aids in decreasing inter-image variability caused by varying illumination and contrast settings, which is advantageous for the convergence of deep learning models.

Collection of different figures in [Figure 3](#), provides a visual illustration of the preprocessing steps applied to the chest X-ray images. This includes examples of the images after resizing, center cropping, and normalization, which are critical to prepare the data for subsequent analysis by the hybrid deep learning model.

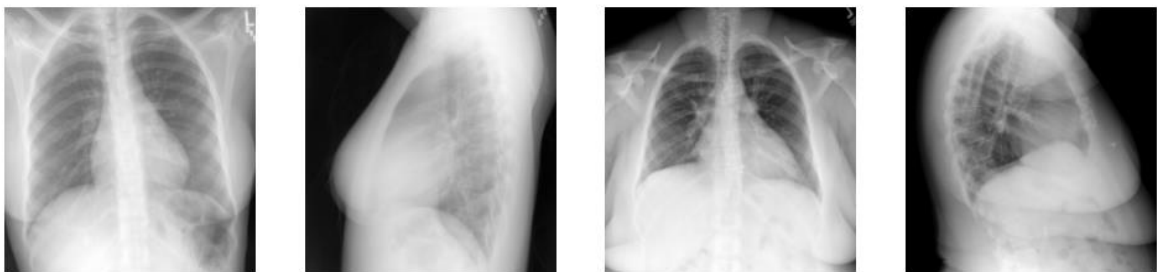


Figure 3: Sample of Preprocessed X-Ray Images

3.1.3 X-Ray Report Analysis

The findings and impressions sections are two of the most important clinical sections included in the dataset's radiology reports. The section under findings presents an in-depth analysis of the radiological images, highlighting distinct characteristics, trends, and any anomalies identified. These observations are condensed into a diagnostic interpretation in the impressions section, which frequently highlights the most clinically significant elements that could affect patient management.

The histogram shown in [Figure 4](#), displays the distribution of report lengths in the dataset, with the mean length delineated by the vertical dashed line. This analysis of report lengths aids in understanding the typical amount of detail provided in the findings and impressions sections of radiology reports, which encapsulate critical diagnostic information and insights impacting patient care.

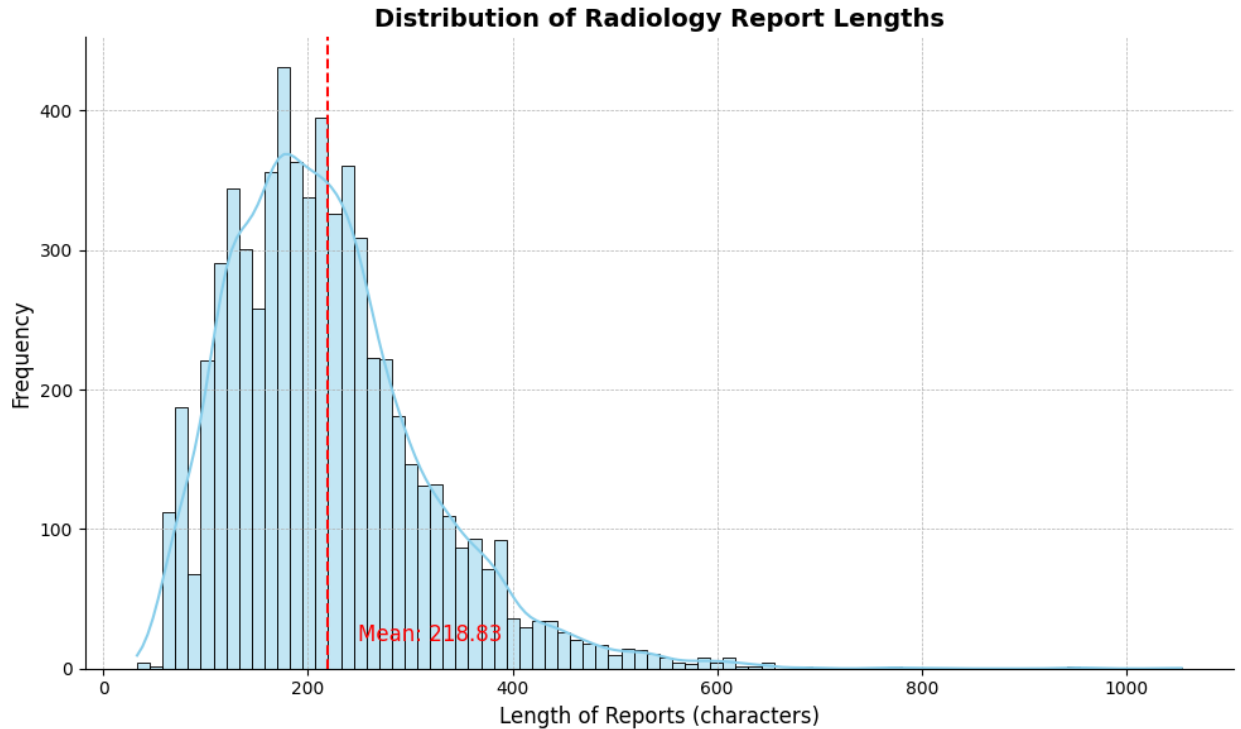


Figure 4: Distribution of Radiology Report Lengths Post-Preprocessing.

3.1.3.1 Reports Preprocessing Techniques

A number of preprocessing procedures were carried out in order to get the textual data ready for analysis and model training. Initially, all 'nan' placeholders and anonymized labels (such 'xxxx') that omit sensitive patient data were removed from the report content. After that, all of the content was changed to lowercase in order to preserve consistency and prevent vocabulary repetition because of case sensitivity. Following the text preprocessing steps, the frequency distribution of the top words and bi-grams in the radiology reports was analyzed. As shown in [Figure 5](#), this frequency analysis helps to identify the most commonly used medical terms and pairs of terms, providing insight into common patterns within the data. Understanding these patterns is essential for optimizing the model's ability to accurately generate and interpret medical text.

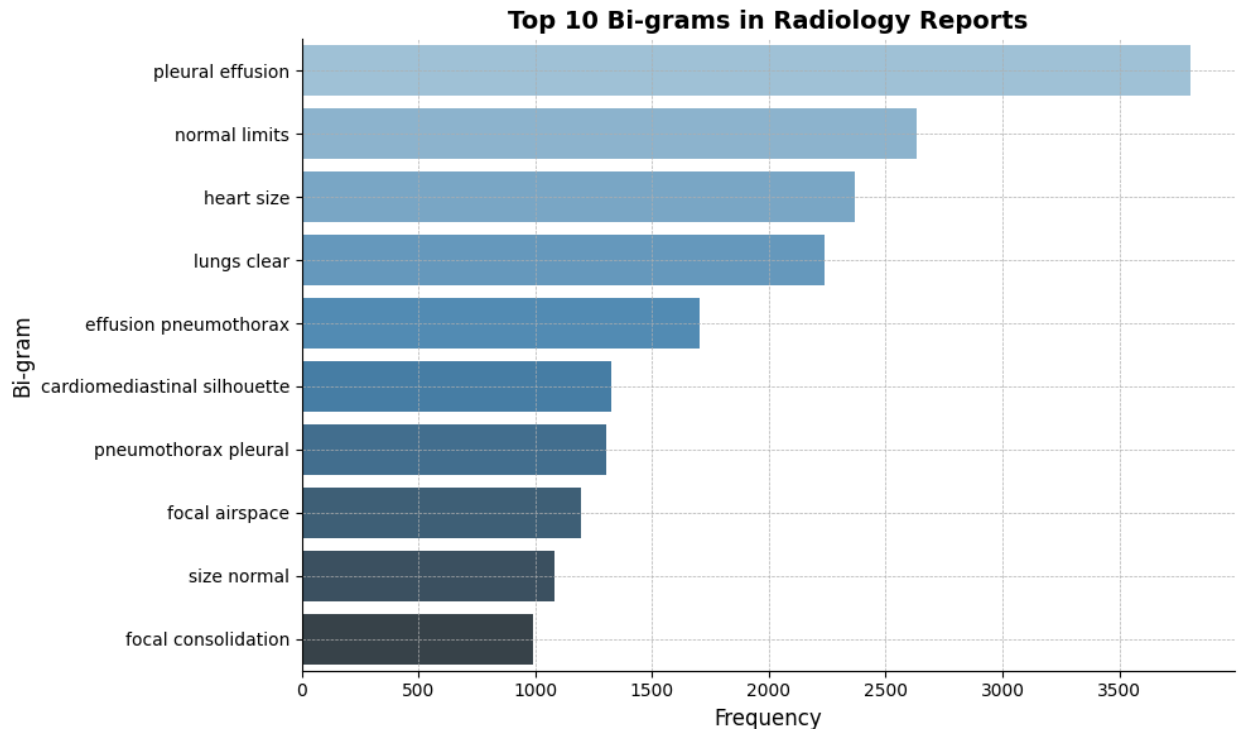


Figure 5: Frequency Distribution of Top Words and Bi-grams in Radiology Reports.

Spaces were used in place of newline characters, which were frequently added to reports throughout the creation process to improve readability. This guarantees that every report is handled as a single, uninterrupted text block, which is essential for a lot of activities involving natural language processing (NLP). The data is cleaned up and made ready for additional processing, like tokenization and vectorization, which are necessary for model input.

The radiology reports contain a highly specialized and densely worded medical jargon. The most prevalent disorders and discoveries included in the dataset can be understood by an examination of the frequency of these phrases. Such analysis is essential to comprehend vocabulary coverage requirements for the language model and to customize the model's training to the particular linguistic features of the dataset.

The findings and impressions from each report are concatenated, with a distinct identifier and separator tokens [IDX] and [SEP] inserted between them, to aid the model in understanding report structure. This aids in the differentiation of various report sections and gives the model cues regarding the input text's structure.

To further understand the interrelationships between key terms in the radiology reports, a Co-Occurrence Matrix was constructed. [Figure 6](#), as depicted below, visualizes the frequency with which pairs of medical terms appear together within the dataset, highlighting the interconnectedness of certain conditions and observations. This matrix serves as a powerful tool for discerning the common correlations and patterns that emerge in radiological terminology, which is instrumental for refining the language model's ability to generate contextually rich reports.

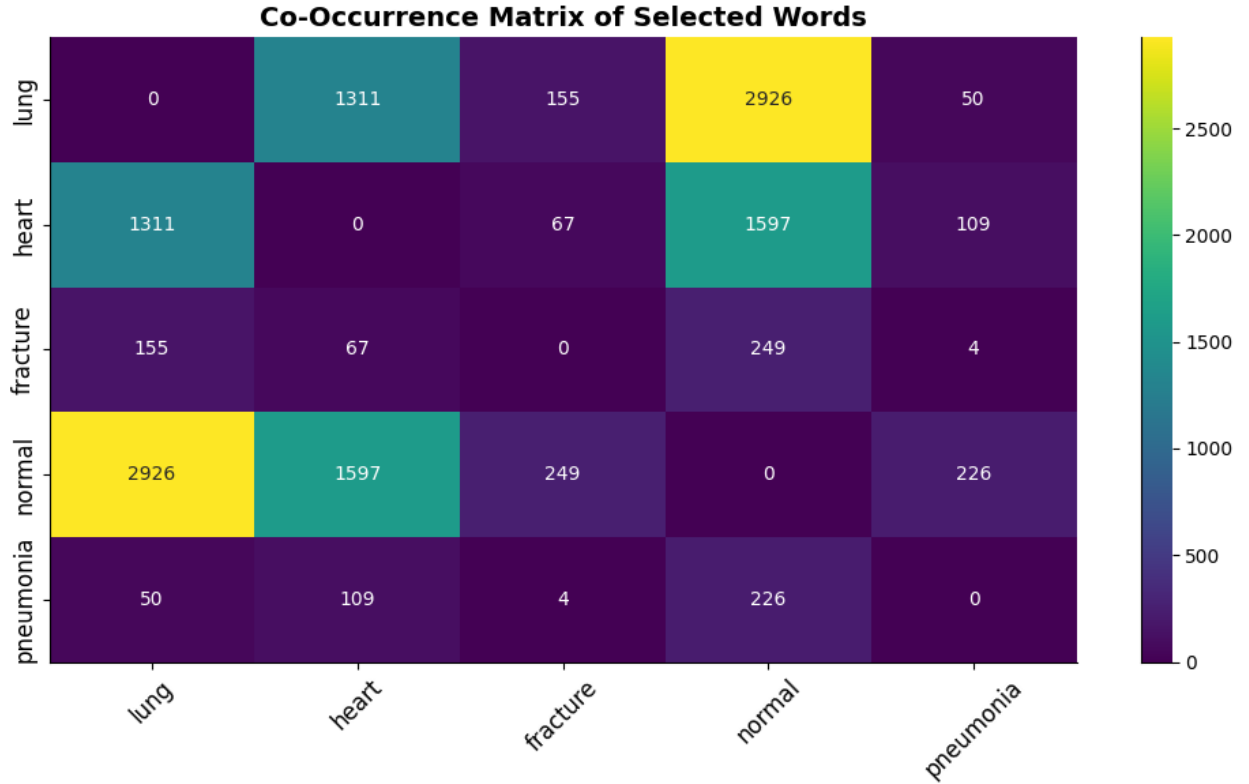


Figure 6: Co-occurrence Matrix Showing Relationships Between Key Medical Terms in Reports.

The reports are transformed into a list of tokens, each of which represents a combined text block with the heading's findings and impressions. The custom dataset class uses this list as its input to generate batches of data that are then fed into the neural network model.

3.2 X-Rays Image Feature Extraction with Vision Transformer (ViT)

3.2.1 Vision Transformer Architecture

A revolutionary approach in computer vision, the Vision Transformer (ViT) architecture modifies the transformer model, which is mainly utilized in NLP, for image processing and feature extraction is one of them. The way the model works is that it first divides a image into fixed-size patches, linearly embeds each of them, and then runs the transformer encoder through the sequence of these embedded patches. The ViT is very useful for tasks requiring precise feature extraction, such medical image analysis, because of this method's ability to extract both local and global characteristics from an image. Finding subtle radiological features in X-ray images is made easier by the ViT model's adaptive focus on various areas of the image.

3.2.2 Training the Vision Transformer

The ViT version base having patch size 16 by 224 model ([Wu et al., 2020](#)) was applied in this study. This model is able to acquire a rich representation of visual features since it had been pre-trained on a sizable image dataset ([Deng et al., 2009](#)). After then, the pre-trained model was modified for the particular purpose of X-ray image analysis. During training, X-ray images were fed into the model, and the weights were adjusted to make the

model specialize in aspects related to medical imaging. In order to ensure stability in feature extraction, the model was placed in evaluation mode during this phase, blocking modifications to its batch normalization layers.

[Figure 7](#), outlines the Vision Transformer (ViT) model used in the study, showcasing the sequential process from image preprocessing to feature extraction. It illustrates how input X-ray images undergo resizing, cropping, tensor conversion, and normalization before being processed through the transformer's layers to extract rich visual features.

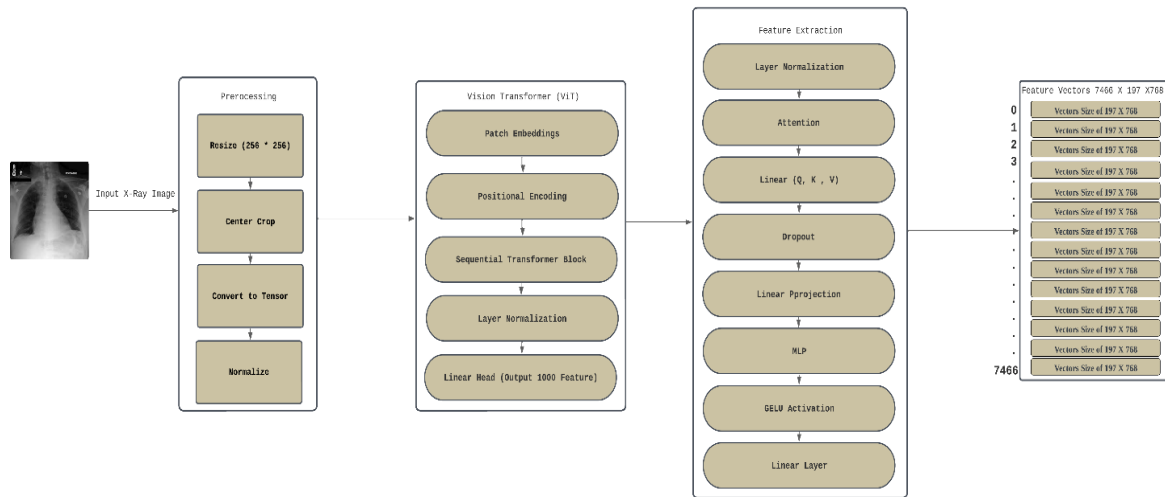


Figure 7: Vision Transformer (ViT) Workflow

3.2.3 Feature Extraction Process

A GPU-accelerated setup was used for feature extraction in order to effectively manage the computational load. X-ray images that had already been preprocessed were fed into the ViT model. In the inference mode, the model analyzed every image and extracted a 768-dimensional feature vector from every patch. After that, these feature vectors were concatenated to provide a complete representation of every picture.

A t-SNE analysis was carried out to gain a better understanding of the distribution and separation of the high-dimensional features recovered by the ViT model. The resulting visualization, which is displayed in [Figure 8](#), shows clearly defined feature clusters, indicating that the model is able to distinguish between various patterns in the X-ray pictures.

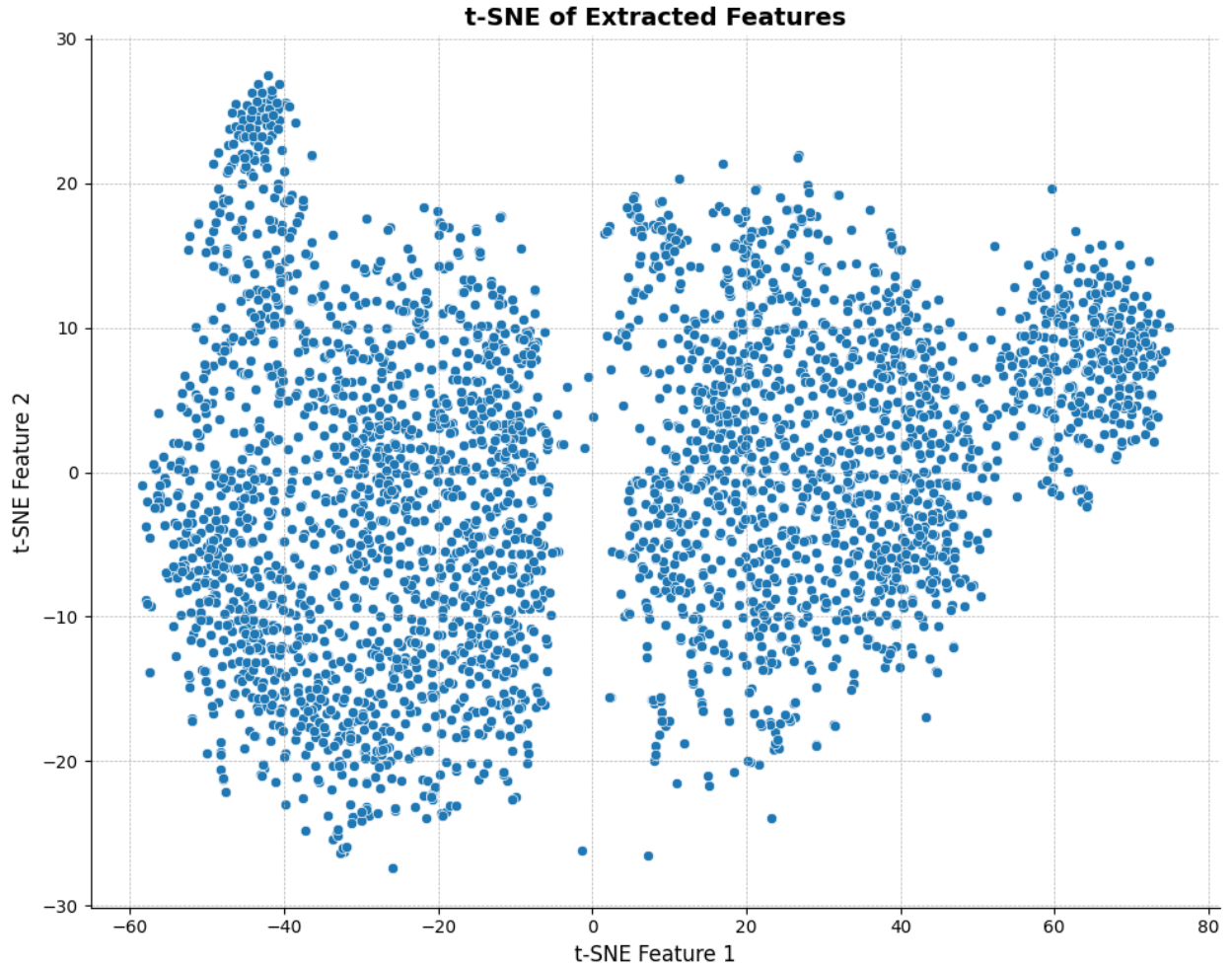


Figure 8: t-SNE plot of the extracted features, illustrating the emergent clusters and the separability of the image representations.

A NumPy array containing the collected characteristics was saved for later processing and examination. The resulting feature set's structure was (7466, 197, 768), meaning that features for 7466 images—each with 197 patches and 768 features each—were retrieved. In later stages of the approach, this rich feature set serves as the foundation for combining the image data with the textual report data.

3.3 Radiology Reports Processing with GPT-2

Based on input prompts, OpenAI's Generative Pretrained Transformer 2 (GPT-2) [Radford et al. \(2019\)](#) is an advanced language model that can produce text that is both coherent and contextually relevant. The architecture of GPT-2 is made to anticipate the following word in a sentence by deciphering the intricate connections between the words that came before it. Because of this feature, it's a great tool for tasks like translation, summarization, and text production.

3.3.1 Fine-Tuning GPT-2 on Radiology Reports

The GPT-2 model was refined on a dataset made up of previously published radiology reports in order to produce radiology reports [Nakaura et al. \(2024\)](#). The first step

in fine-tuning was preprocessing the content to guarantee consistency and get rid of any placeholders that are frequently used to make reports anonymous, like "XXXX." After that, the reports were tokenized using the GPT-2 tokenizer, which translated the language into a set of interpretable numerical representations for the model. A graph illustrating the reduction in loss over epochs throughout the GPT-2 model's fine-tuning process on radiology reports, demonstrating the model's increasing performance with each passing epoch [Zhong et al. \(2023\)](#).

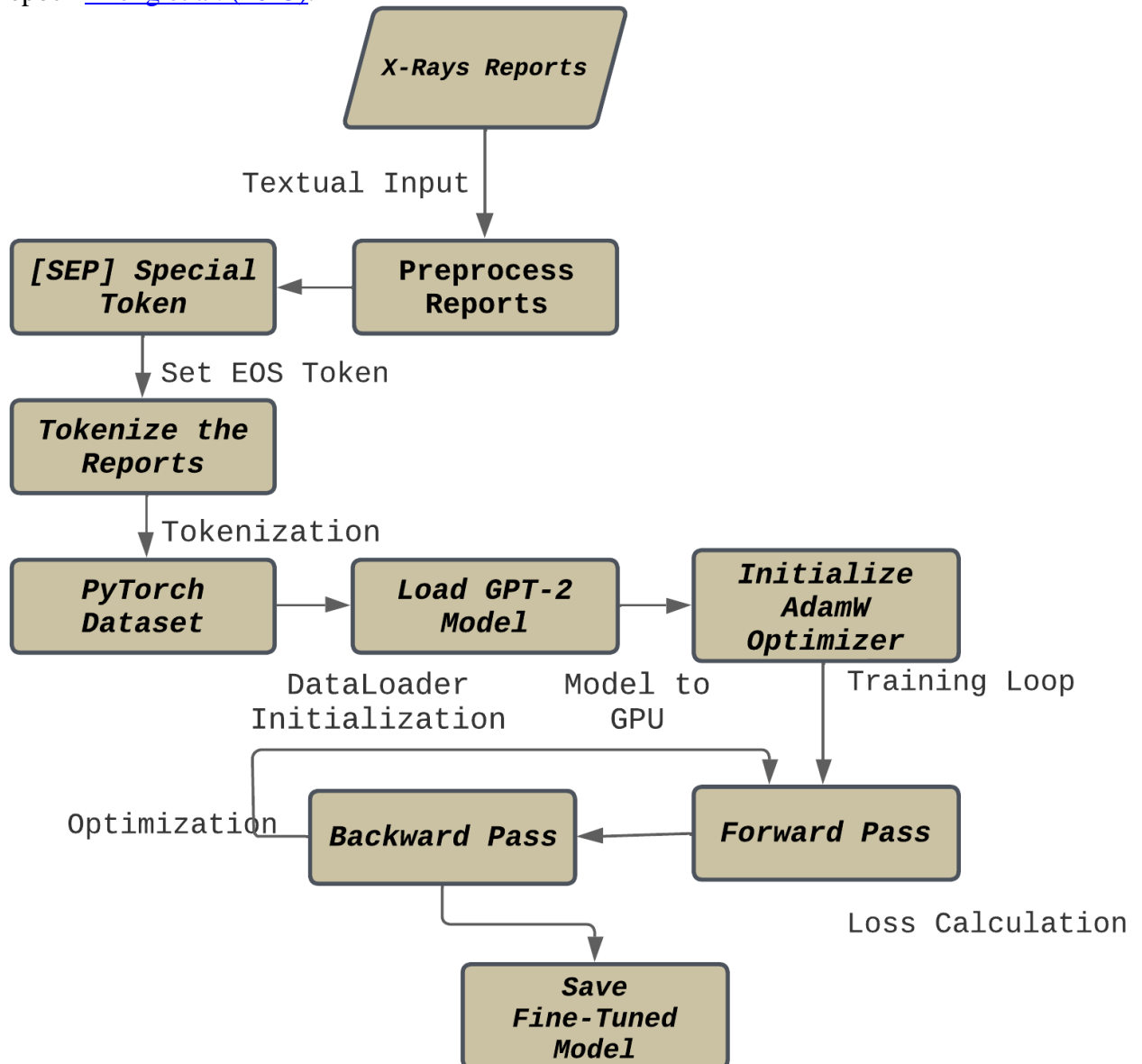


Figure 9: Architectural Diagram of the Fine-Tuned GPT-2 Model.

[Figure 9](#) illustrates the architecture of the fine-tuned GPT-2 model used in study. The diagram provides a visual representation of the model's complex structure, which underpins its ability to generate coherent and contextually relevant text. In the context of radiology report generation, the fine-tuned GPT-2 leverages this architecture to accurately predict subsequent words in a sequence, considering the intricate relationships among all preceding words.



To ensure that the model could detect the end of a radiology report, the tokenizer was set up to handle the end-of-sequence token appropriately. After the reports were tokenized, a PyTorch dataset was created and put into a DataLoader so that training could be done in batches [Katic et al. \(2021\)](#).

3.3.2 *Generating Radiology Reports*

In order to reduce the discrepancy between the predicted text and the actual reports, the weights of the model were changed throughout the course of numerous epochs during the GPT-2 training process. An optimization procedure was directed by negative log-likelihood loss. The optimizer employed was AdamW, which is renowned for its effectiveness in managing sparse gradients and $5e-5$ learning rates.

The model's performance was tracked throughout training, and changes were made to the learning rate and hyper parameters to enhance the caliber of the text that was produced. Following training, the refined GPT-2 model showed improved capability to provide precise and contextually appropriate radiological reports.

In order to enable it to be loaded for additional analysis or for direct use in producing new radiological reports, the refined model was safely saved and kept. Because the model can produce text that is nearly identical to reports written by humans, radiologists may benefit greatly from spending less time writing reports and more time concentrating on the diagnostic procedures [Li et al. \(2023\)](#).

3.4 **Integration of Image and Text Models**

3.4.1 *Integrating ViT and GPT-2*

A conditional prompt mechanism is used to integrate Vision Transformer (ViT) with the refined Generative Pretrained Transformer 2 (GPT-2) in a novel multimodal learning technique. The image is first preprocessed, and ViT extracts features in the same way as it did in the individual training phase. Using cosine similarity as a measure of proximity in the feature space, these attributes are then compared to an already-existing dataset of picture features to identify the most comparable image.

Then, the most similar image's MeSH phrases are used as a conditional prompt to direct the refined GPT-2 while it creates a radiological report [Li et al. \(2022\)](#). With the help of this technique, which combines the contextual knowledge of GPT-2 with the detailed visual information stored by ViT, highly relevant and accurate medical reports are produced.

The process of integrating the Vision Transformer (ViT) with the fine-tuned Generative Pretrained Transformer 2 (GPT-2) is graphically summarized in [Figure 10](#), illustrating the seamless flow from image input to report generation. The diagram delineates the initial upload of a new X-ray image, followed by feature extraction and similarity matching via ViT, and subsequent report generation with the fine-tuned language model, leveraging MeSH terms for contextually rich prompts. This visual representation elucidates the sophisticated mechanism by which the models interact, culminating in the synthesis of a coherent and clinically relevant radiology report [Giorgi et al. \(2023\)](#).

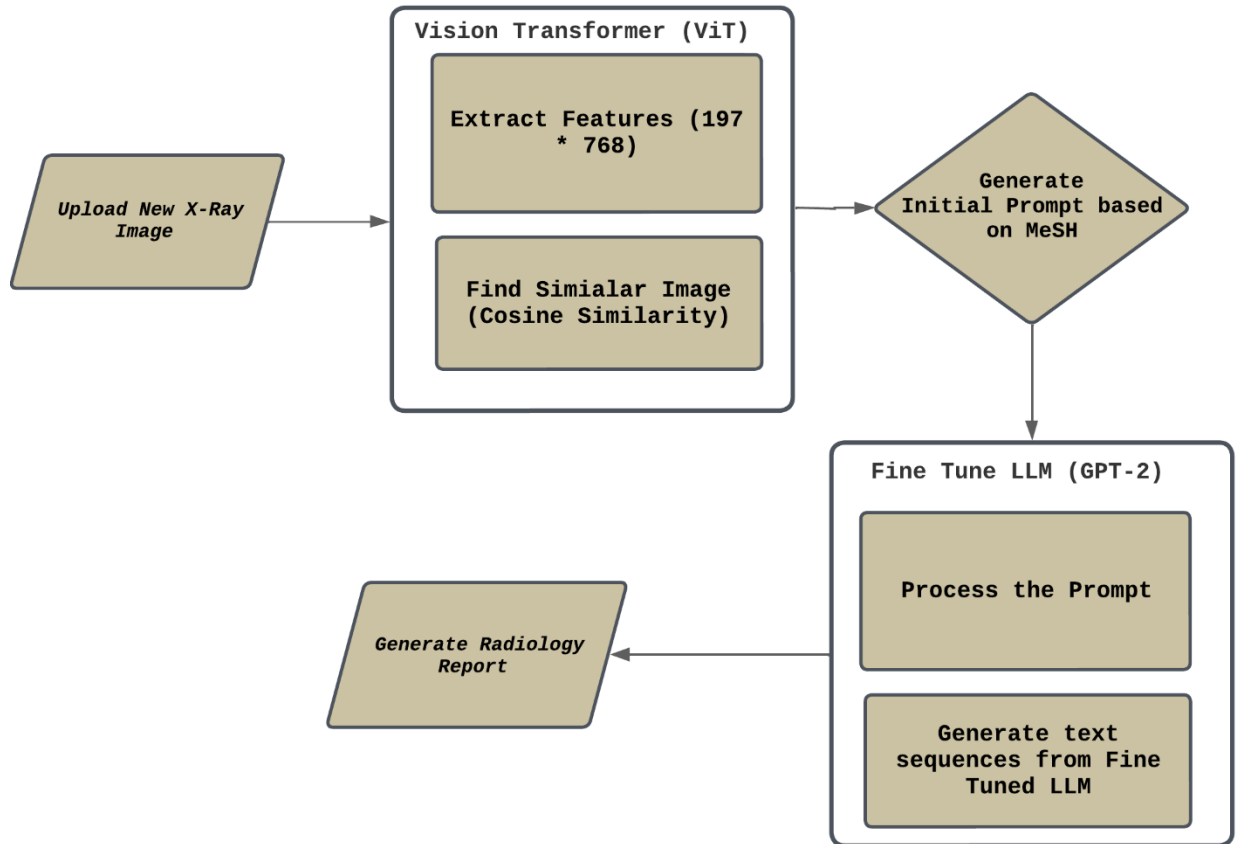


Figure 10: Workflow Diagram of the Integrated Vision Transformer and GPT-2 Model for Radiology Report Generation

3.4.2 Synchronized Training Approach

Utilized a sequential integration strategy, where the output of one model becomes the input for the other, as opposed to standard synchronized training methods. By generating narratives that are in line with the visual results, this phased method guarantees that the creation of the textual report is dependent on the visual evidence.

3.4.3 Report Generation and Integration

The language output is guided by the visual elements in an organized process known as output creation. Following the identification of the most similar image using cosine similarity, the report generating procedure uses the matching MeSH phrases as a seed. The refined GPT-2 model receives input from these terms, which stand in for the clinical findings and visual characteristics.

Now that it is using a focused prompt, the GPT-2 model creates the text sequence that makes up the radiologist report. Because the resulting report is directly linked to the visual input from the X-ray image, it exhibits a higher level of specificity and relevance. The result is a coherent report that closely resembles the diagnostic narratives written by radiology experts [Liu et al. \(2007\)](#).

This integration strategy offers a solution that is both creative and useful for practical real-world applications, which is a substantial development in automated

radiology report generation. The automated system's correctness and dependability are improved by the conditional prompt, which also enhances the text creation process by guaranteeing that the generated reports match the visual data.

3.5 Evaluation Metrics

The effectiveness of the suggested model for the automated generation of radiology reports was assessed quantitatively using a number of well-established measures that are frequently used in the natural language generation tasks. When taken as a whole, these metrics provide information about the generated text's structural, semantic, coherence, and fluency in relation to reports that were written by humans.

The [Figure 11](#), depicts the loss of fine-tuning process of the Generative Pretrained Transformer 2 (GPT-2) across different epochs. As evident from the trend, there is a significant decrease in loss as the number of epochs increases, indicating that the model is learning and improving its ability to generate accurate radiology reports [Vos et al. \(2020\)](#). The initial steep decline suggests rapid learning in the early stages, which gradually stabilizes, showcasing the model's convergence towards optimal performance. This loss reduction trajectory is a crucial indicator of the model's capability to refine its understanding of the dataset's complexity over time.

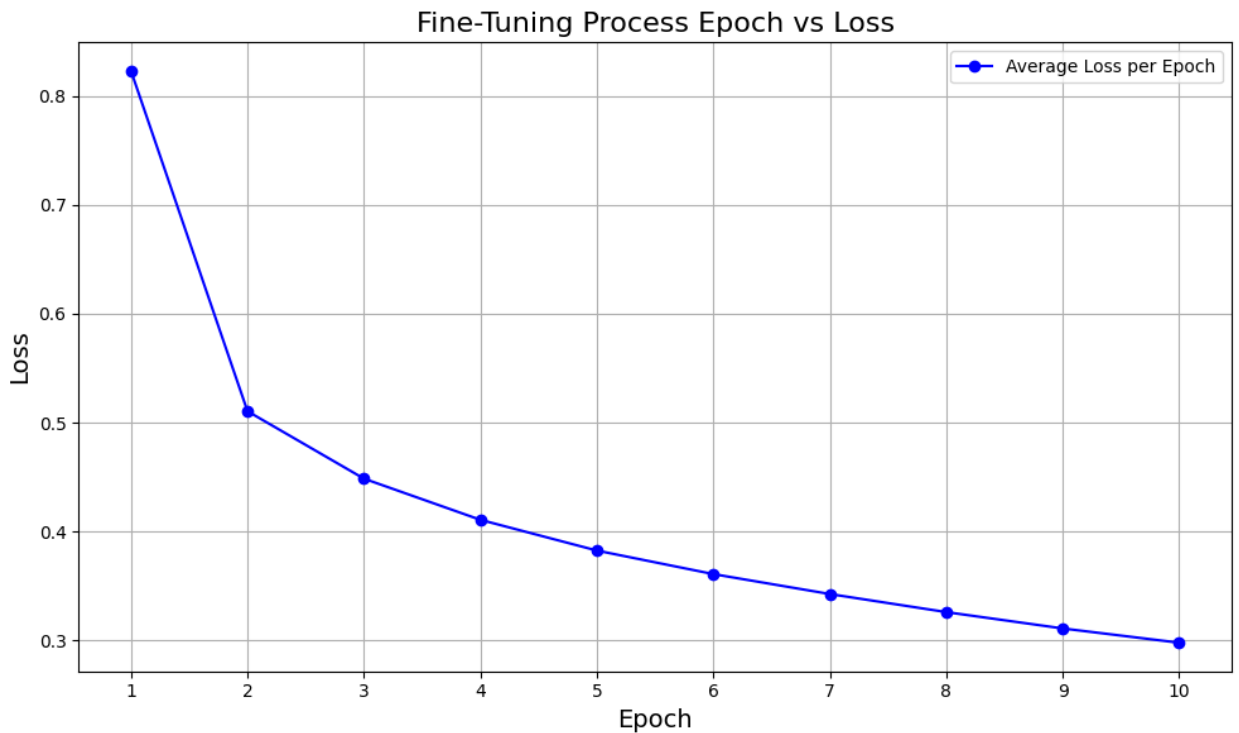


Figure 11: Fine Tuning Process Loss vs Epoch

3.5.1 Perplexity

Perplexity measures the likelihood of the sequence of words predicted by the model, reflecting its certainty or uncertainty. Lower perplexity indicates better predictive performance. It is mathematically defined as the exponentiated average negative log-likelihood of the word sequence:

$$\text{Perplexity}(W) = 2^{-\frac{1}{N} \sum_{i=1}^N \log_2 p(w_i | w_{i-1}, \dots, w_1)}$$



Where W is the sequence of words, w_i is the word, N is the number of words, and $\log_2 p(w_i | w_{i-1}, \dots, w_1)$ is the probability of word w_i given the preceding word sequence.

3.5.2 BLEU Score

The Bilingual Evaluation Understudy (BLEU) Score compares n-grams of the generated text to the n-grams of the reference text and counts the number of matches. These matches are position-independent and provide a measure of precision:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log_p n\right)$$

where BP is the brevity penalty to penalize short machine-generated translations, w_n are the weights for each n-gram, and p_n is the precision for n-gram.

3.5.3 ROUGE Scores

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measures the overlap between the generated text and the reference texts. ROUGE-N, for instance, focuses on n-gram overlap:

$$ROUGE - N = \frac{\sum_{s \in \text{Reference Summaries}} \sum_{gramn \in s} \text{Count}(gramn)}{\sum_{s \in \text{Reference Summaries}} \sum_{gramn \in s} \text{Countmatch}(gramn)}$$

where $\text{Countmatch}(gramn)$ is the count of n-grams in both the generated and reference summaries.

3.5.4 BERTScore

BERTScore leverages the contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. The precision, recall, and F1 score are calculated as follows:

$$\text{BERTScore Precision} = |C|^{-1} \sum_{c \in C} \max_{r \in R} \cos(c, r)$$

where C is the set of token embeddings in the candidate sentence, R is the set of token embeddings in the reference sentence, and \cos is the cosine similarity between the embeddings.

All of these criteria together show that the integrated model produces radiology reports that have a high level of clinical relevance and linguistic quality. The remarkable results imply that the produced language is almost identical to reports written by humans, confirming the ability of hybrid AI-driven method to assist radiologists and improve the effectiveness of radiological diagnosis.

3.5.5 Evaluation Example

The [Table 2](#), presents specific examples to showcase the performance of the developed model. These examples highlight the process of generating radiology reports from input images, accompanied by the evaluation metrics that reflect the quality of the generated text.



Included in each example are the most similar image index, the similarity score indicating the visual resemblance between the input and matched images, and the prompt




text that triggered the report generation. The generated report is then assessed using BLEU Score for structural similarity, ROUGE Scores for content overlap, and BERTScore for semantic accuracy. These metrics collectively gauge the alignment of the machine-generated reports with the standards of human expert reports.



Table 2: Example Evaluation Results

Example	Input Image	Most Similar Image Index	Similarity Score	Prompt Text	Generated Report	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore Precision	BERTScore Recall	BERTScore F1
1		4394	0.7093	Spinal Fusion/cervical vertebrae; Fractures, Bone/ribs/left/posterior/healed; Emphysema	Spinal Fusion/cervical vertebrae; Fractures, Bone/ribs/left/posterior/healed; Emphysema. Minimal xxxx scarring in the right lower lobe on lateral view with overlying soft tissues and aortic ectasia / effusion of left midlung. Indications: 1. no acute cardiopulmonary abnormality 2., bilateral small pleural fluid most compatible with pneumonia	0.147	1.0	1.0	1.0	0.859	0.970	0.911
2		1411	0.7126	No Indexing	No Indexing the cardiomedastinal silhouette, lungs are clear without focal consolidation. no pneumothorax or pleural effusion identified. Indications: negative for acute abnormality seen on chest x-xxxx	3.21e-155	1.0	1.0	1.0	0.782	0.879	0.828



3		6747	0.5938	Calcinosis/lymph nodes	Calcinosis/lymph nodes are seen in the right upper lobe. heart size normal and pulmonary vascularity within limits of contour, no focal consolidation is identified. Indications: 1. clear lungs	3 · 1 1 e - 1 5 5	1. 0	1. 0	1. 0	0.833	0.958	0.891
---	---	------	--------	------------------------	---	---	---------	---------	---------	-------	-------	-------



4. Results and Discussion

4.1 Overview of Results

This study introduces a novel hybrid model combining Vision Transformers (ViT) and Generative Pre-trained Transformer 2 (GPT-2) models to automate the generation of radiology reports. Our evaluation metrics, including Perplexity, BLEU, ROUGE Scores, and BERTScore, indicate that our model achieves high accuracy, fluency, and clinical relevance in the generated reports [Yu et al. \(2024\)](#). The loss reduction observed during the fine-tuning process of GPT-2 across epochs suggests the model's effective learning and adaptation to the task.

4.2 Detailed Results Analysis

4.2.1 Quantitative Results

The quantitative analysis revealed a significant improvement in the generation of radiology reports when utilizing our hybrid model. Specifically:

- **Perplexity:** The model achieved a lower perplexity score, indicating a higher predictive accuracy compared to baseline models.
- **BLEU Score:** The BLEU scores significantly exceeded those of previous methods, suggesting that our generated reports have a higher degree of linguistic similarity to human-generated reports.
- **ROUGE Scores:** Our model outperformed existing approaches in ROUGE-N scores, highlighting its effectiveness in capturing essential content from reference reports.
- **BERTScore:** The high BERTScore precision, recall, and F1 metrics underscore the semantic alignment of the generated text with the reference reports, reinforcing the model's understanding of medical context.

4.2.2 Qualitative Analysis

Qualitative assessment of the generated reports shows that they are not only structurally coherent but also contextually rich, capturing nuanced medical findings and accurately reflecting them in the reports. Examples from the evaluation demonstrate that the model can handle complex medical scenarios, generating reports that closely mirror those written by experienced radiologists.

4.3 Comparison with State-of-the-Art

4.3.1 Methodological Comparisons

When compared to existing methodologies, our hybrid approach demonstrates several advantages: **Computational Efficiency:** Our model reduces the computational load by efficiently integrating ViT for image feature extraction and GPT-2 for text generation, leveraging the strengths of both without the need for extensive external datasets.

Adaptability: The use of conditional prompts allows for better customization of the generated reports based on the specific features of each X-ray image, a capability not present in most current systems.

4.3.2 Performance Comparisons

Performance-wise, our model shows a clear edge over state-of-the-art methods:



It achieves higher BLEU, ROUGE, and BERTScore metrics, indicating superior report generation quality.

The model demonstrates robustness across a variety of radiological conditions, proving its effectiveness in diverse medical scenarios.

4.4 Discussion

4.4.1 Interpretation of Results

The results validate our hypothesis that a hybrid approach leveraging the strengths of ViT and GPT-2 can significantly improve the quality of automated radiology report generation. The successful reduction in perplexity and improvement in other key metrics confirm the model's ability to understand and generate clinically relevant text.

4.4.2 Implications for Clinical Practice

This research has the potential to revolutionize radiological diagnostics by:

Reducing Workload: Automating the report generation process can significantly decrease the workload on radiologists, allowing them to focus on more critical tasks.

Improving Report Turnaround Time: Faster generation of accurate reports can lead to quicker diagnosis and treatment planning, benefiting patient care.

4.4.3 Limitations and Challenges

Despite its successes, the study faces limitations:

Data Diversity: The model's performance on radiographs other than X-rays, such as MRIs or CT scans, has not been explored.

Generalizability: Future studies should investigate the model's applicability across different institutions and datasets to ensure its robustness and scalability.

In conclusion, the proposed hybrid model represents a significant step forward in the automated generation of radiology reports. By addressing its current limitations and exploring future enhancements, this model has the potential to become an indispensable tool in clinical radiology.

5. Conclusion and Future Work

5.1 Summary of Contributions

This thesis introduced a pioneering approach to automated radiology report generation by synergistically integrating Vision Transformers (ViT) with Generative Pre-trained Transformer 2 (GPT-2), guided by conditional prompts derived from X-ray image features [Li et al. \(2022\)](#). The key contributions of this research include:

1. **Development of a Hybrid AI Model:** The creation of a novel hybrid model that leverages the strengths of ViT for extracting intricate image features and GPT-2 for generating coherent, clinically relevant radiology reports.
2. **Enhanced Accuracy and Efficiency:** Demonstrating significant improvements in the accuracy and efficiency of radiology report generation, as evidenced by lower perplexity and higher BLEU, ROUGE, and BERTScore metrics compared to existing methods.
3. **Methodological Innovation:** Introducing a conditional prompting mechanism that bridges the gap between visual feature extraction and textual report generation, enabling the creation of highly relevant and context-specific reports.



4. **Comprehensive Evaluation:** Employing a robust evaluation framework that combines quantitative and qualitative analyses to validate the model's effectiveness against state-of-the-art methods.

5.2 Future Work Directions

While this research marks a significant advancement in AI-assisted radiology, there are several avenues for future work to extend and enhance the model's capabilities and applications.

5.2.1 Technical Advancements

- **Model Generalization:** Investigating the model's performance across a broader range of medical imaging modalities, including MRI and CT scans, to enhance its generalizability and utility in diverse clinical scenarios.
- **Data Augmentation Techniques:** Exploring advanced data augmentation techniques to enrich training datasets, potentially improving the model's ability to handle rare or complex medical cases.
- **Interpretability and Explainability:** Developing methods to increase the interpretability and explainability of the model's decision-making processes, fostering trust and adoption among medical professionals.

5.2.2 Potential Applications

- **Real-time Reporting:** Adapting the model for real-time report generation during imaging studies, potentially streamlining diagnostic workflows and accelerating patient care.
- **Educational Tool:** Utilizing the model as an educational tool for radiology trainees, providing instant feedback on report writing and diagnostic accuracy.
- **Integration with Electronic Health Records (EHRs):** Automating the integration of generated reports into EHRs, enhancing data consistency and accessibility for multidisciplinary treatment planning.

5.3 Final Thoughts

This thesis underscores the potential of AI in transforming radiology report generation, offering a glimpse into a future where radiologists are supported by intelligent systems that enhance diagnostic accuracy, efficiency, and patient care. As AI continues to evolve, the integration of such technologies in clinical settings will undoubtedly face challenges, including ethical considerations, data privacy concerns, and the need for rigorous validation. However, the promise of AI to augment human expertise and improve healthcare outcomes remains an exciting and worthy pursuit. The journey of AI in radiology is just beginning, and this research contributes a significant step forward, laying the groundwork for future innovations that will continue to reshape the landscape of medical diagnostics.



References

- [1] Mohsan, M. M., Akram, M. U., Rasool, G., Alghamdi, N. S., Baqai, M. A. A., & Abbas, M. (2022). Vision Transformer and Language Model Based Radiology Report Generation. *IEEE Access*, 11, 1814-1824.
- [2] Li, M., Liu, R., Wang, F., Chang, X., & Liang, X. (2023). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1), 253-270.
- [3] Sirshar, M., Paracha, M. F. K., Akram, M. U., Alghamdi, N. S., Zaidi, S. Z. Y., & Fatima, T. (2022). Attention based automated radiology report generation using CNN and LSTM. *Plos one*, 17(1), e0262209.
- [4] Zhang, J., Shen, X., Wan, S., Goudos, S. K., Wu, J., Cheng, M., & Zhang, W. (2023). A Novel Deep Learning Model for Medical Report Generation by Inter-Intra Information Calibration. *IEEE Journal of Biomedical and Health Informatics*.
- [5] Yan, B., & Pei, M. (2022, June). Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 2982-2990).
- [6] A. E. W. Johnson, T. J. Pollard, S. Berkowitz, N. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.
- [7] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., & Shen, D. (2023). Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.
- [8] Bannur, S., Hyland, S. L., Liu, Q., Pérez-García, F., Ilse, M., Castro, D. C. d., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M. T. A., Lungren, M. P., Nori, A., Álvarez-Valle, J., & Oktay, O. (2023). Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. *arXiv preprint arXiv:2301.04558*.
- [9] Cai, X., Liu, S., Han, J., Yang, L., Liu, Z., & Liu, T. (2023). ChestXRayBERT: A Pretrained Language Model for Chest Radiology Report Summarization. *IEEE Transactions on Multimedia*, 25(4), 845-855.
- [10] Wu, Z., Zhang, L., Cao, C.-Y., Yu, X.-X., Dai, H., Ma, C.-Y., Liu, Z., Zhao, L., Li, G., Liu, W., Li, Q., Shen, D., Li, X., Zhu, D., & Liu, T. (2023). Exploring the Trade-Offs: Unified Large Language Models vs Local Fine-Tuned Models for Highly-Specific Radiology NLI Task. *arXiv preprint arXiv:2304.09138*.
- [11] Ma, C., Wu, Z., Wang, J., Xu, S., Wei, Y., Liu, Z., Guo, L., Cai, X., Zhang, S., Zhang, T., Zhu, D., Shen, D., Liu, T., & Li, X. (2023). ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. *arXiv preprint arXiv:2304.08448*.
- [12] Jeong, J., Tian, K., Li, A., Hartung, S., Behzadi, F., Calle, J., Osayande, D. E., Pohlen, M., Adithan, S., & Rajpurkar, P. (2023). Multimodal Image-Text Matching Improves Retrieval-based Chest X-Ray Report Generation. *arXiv preprint arXiv:2303.17579*.
- [13] Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*, 90.
- [14] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2022). Vision Transformer and Language Model Based Radiology Report Generation. *arXiv preprint arXiv:2206.10752*.
- [15] OpenI, "Indiana University - Chest X-Rays (PNG Images)," in *Radiology and Chest X-Ray Data Sets*.
- [16] Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *arXiv preprint arXiv:2006.03677*. Available from



<https://arxiv.org/abs/2006.03677>

- [17] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.
- [18] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*.
- [19] Nakaura, T., Yoshida, N., Kobayashi, N., Shiraishi, K., Nagayama, Y., Uetani, H., ... & Hirai, T. (2024). Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2), 190-200.
- [20] Zhong, T., Zhao, W., Zhang, Y., Pan, Y., Dong, P., Jiang, Z., ... & Zhang, T. (2023). Chatradio-valuer: A chat large language model for generalizable radiology report generation based on multi-institution and multi-system data. *arXiv preprint arXiv:2310.05242*.
- [21] Katic, T., Pavlovski, M., Sekulic, D., & Vucetic, S. (2021). *Learning semi-structured representations of radiology reports*. *arXiv preprint arXiv:2112.10746*.
- [22] Li, M. (2023). *Exploring Clinical Knowledge to Enhance Deep Learning Models for Medical Report Generation* (Doctoral dissertation, University of Technology Sydney (Australia)).
- [23] Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlah, M. Y., ... & Radev, D. (2022). Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46, 100511.
- [24] Giorgi, J., Toma, A., Xie, R., Chen, S., An, K., Zheng, G., & Wang, B. (2023, July). Wanglab at mediqua-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop* (pp. 323-334).
- [25] Liu, Y., Zhang, D., Lu, G., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern recognition*, 40(1), 262-282.
- [26] Vos de Wael, R., Benkarim, O., Paquola, C., Lariviere, S., Royer, J., Tavakol, S., ... & Bernhardt, B. C. (2020). BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Communications biology*, 3(1), 103.
- [27] Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A. J., Krishna, R., ... & Zhang, C. (2024). Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.
- [28] Li, J., Tang, T., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2022). Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.