TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

ADVANCED COLLEGE OF ENGINEERING AND MANAGEMENT

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

KALANKI, KATHMANDU

A Major Project Proposal Defense Report on

"Attention Based Automated Radiological Report Generation Using Multimodal Architecture"

[CT 707]

Submitted By:

Ankit Chhetri ACE077BCT012

Bhaskar Subedi ACE077BCT023

Biplov Belbase ACE077BCT031

A Major Project Proposal report submitted to the Department of Electronics and Computer Engineering in the partial fulfillment of the requirements for degree of Bachelor of Engineering in Computer Engineering

Kathmandu, Nepal.

Date: June 5, 2024

ACKNOWLEDGEMENT

We take this opportunity to express our deepest and sincere gratitude to our Academic Project Coordinator **Er. Laxmi Prasad Bhatt,** Department of Electronics and Computer Engineering for his insightful advice, motivating suggestions, invaluable guidance, help and support in this project selection and also for his/ constant encouragement and advice throughout our journey till the date.

We express our deep gratitude to **Er. Prem Chandra Roy**, Head of Department of Electronics and Computer Engineering, **Er. Dhiraj Pyakurel,** Deputy Head, Department of Electronics and Computer Engineering for their regular support, co-operation, and coordination.

The in-time facilities provided by the department are also equally acknowledgeable.

We would like to convey our thanks to the teaching and non-teaching staff of the Department of Electronics & Communication and Computer Engineering, acem for their invaluable help and support hitherto. We are also grateful to all our classmates for their help, encouragement and invaluable suggestions.

Finally, yet more importantly, we would like to express our deep appreciation to our grandparents, parents, siblings for their perpetual support and encouragement.

Ankit Chhetri	ACE077BCT012
Bhaskar Subedi	ACE077BCT023
Biplov Belbase	ACE077BCT031

Table of Contents

List of Table

List of Figures

Title **Page**

List of Abbreviations/Acronyms

CNN	Convolutional Neural Network
CVT	Convolutional Vision Transformer
CXR	Chest X-Ray
CIFAR	Canadian Institute of Advanced Research
LLM	Large Language Model
NLP	Natural Language Processing
RNN	Recurrent Neural Network
ViT	Vision Transformer
VTAB	Visual Task Adaptation Benchmark

CHAPTER 1
INTRODUCTION

1.1 Background

Good health is the first requisite of happiness and success in the life of people. People with sound physical and mental health have better productivity. Nepal's constitution has recognized right to health as the fundamental right of the citizens. It stipulates that the people have rights to free basic health services, emergency health services and access to information about health. Still, many Nepalese have no access to affordable and basic health facilities because state-run health centers lack sufficient infrastructure and human resources. This is a reason why the private hospitals operating in many parts of country cater to around 80 per cent health services, with the government hospitals providing only 20 per cent. This is really a matter of serious concern. The situation in the far-flung areas is far worse. Doctors often hesitate to work in the remote areas owing to low pay and geographical difficulties. The doctor-patient ratio in Nepal is 1:850 in the Kathmandu Valley and 1:150,000 in the rural areas. The World Health Organization recommends a doctor-patient ratio of 1:1,000. There are 32,218 registered doctors and 72,550 registered nurses in the country. Of them, only around 15,000 have been working in government health facilities [1]. Although public health service is state responsibility, its privatization has increased its cost beyond the capacity of majority of Nepalese. They have to spend a huge amount of their savings on the medical treatment and purchase of expensive medicines. The added financial burden is painful for the poor.

1.2 Motivation

The development of the intelligent automated system for the medical report generation with the help of the X- ray image can have far- reaching implements beyond just identifying the health problem. In Nepal, where healthcare access in rural areas is limited, this initiative is particularly vital. One of the major issues in Nepal is that it has extremely low ratio of doctors to patients, mainly in rural areas which apparently limits the access to better healthcare services. This project specifically targets the enhancement of early and accurate diagnosis of chest-related diseases which are prevalent and pose significant health challenges in the region. By deploying the portable chest X-ray units and conducting training programs for local healthcare workers in the rural areas we can bring advanced diagnostic capabilities directly to remote areas which significantly improves healthcare accessibility and reducing the dependency on medical specialists.

1.3 Statement of the Problem

Getting accurate chest X-ray results quickly is really important for finding and treating chest related problems. But there are big problems making this difficult, especially in rural areas. There aren't always enough experts to read the X-ray images. This causes delays in helping sick people, making them sicker or even putting their lives at risk. Considering Nepal's specific challenges like its geography and not having enough resources, fixing these problems is super important and crucial.

1.4 Project objective

- To enhance the process of chest X-ray report generation and diagnosis in Nepal.

1.5 Significance of the study

The significance of our project is multifaceted, with its primary impact being the enhancement of healthcare accessibility and quality in Nepal. By improving the process of chest X-ray report generation and diagnosis, particularly in rural and underserved areas, we aim to ensure that individuals have timely access to essential diagnostic tools, leading to earlier detection and treatment of chest related problems. It not only helps in the rural areas but also can be implemented at any places in the condition where due to some problem specialist presence is delayed. This not only improves health outcomes but also reduces the burden on healthcare facilities. Overall, our project holds the potential to significantly transform healthcare delivery, leading to improved patient outcomes, reduced healthcare costs, and a more resilient healthcare system in Nepal.

CHAPTER 2

LITERATURE REVIEW

Automated Radiological Report Generation is a derivative technique to describe clinical details of Chest X-ray images. It is a combination of computer vision and Natural Language Processing which have a strong societal impact. Description retrieval, template filling, and hand-crafted NLP techniques were some of the earlier methods of report writing. There were many advancements in automated medical report generation later, but the base arrangement of each method was to utilize an image encoder for converting CXR images into a latent space and then bring a decoder into play to generate medical reports. The problem was generically identified as an image-to-sequence problem. We have divided the review literature based on the encoder-decoder architectures used in automated radiological report generation [2].

Medical report generation process proposed a CNN-RNN architecture to generate captions for images [3]. These results were however too simple and lacked details. As more work was done in the field, attention was introduced with model's attention with RNN and CNN [4].

CNNs have proven to be extremely effective in image classification, object detection, and segmentation tasks, and have been utilized in a variety of applications including as self-driving cars, medical diagnostics, and facial recognition. CNNs were shown to be capable of classifying view orientations of chest radiographs with excellent accuracy [5].

Chest radiographs were used to construct a CNN-based model for the automated classification of pulmonary tuberculosis, obtaining high performance and indicating the promise for deep learning in the disease detection [6]. CNN was used to detect and classify abnormalities on chest radiographs, with good sensitivity and specificity [7].

Transformers was a revolutionary approach for sequence-to-sequence tasks such as Machine Translation [8]. It was recurrence and convolution free network. Transformers was first employed for text recognition. It leveraged the Transformer architecture for both image understanding and word piece-level text generation [9].

Transformers were first used in 2018 as Image generative model [10]. In 2020, Vision Transformer (ViT), also known as vanilla image transformer, was proposed to demonstrate Transformer in image classification which outperformed existing image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.) [11]. Vision Transformer (ViT) is the first implementation of a transformer in a deep neural network on large-scale image datasets.

This work has reviewed recent studies done in image processing to give more information about the performance of the two architectures and what distinguishes them. A common feature across all papers is that transformer-based architecture or the combination of ViTs with CNN allows for better accuracy compared to CNN networks. It has also been shown that this new architecture, even with hyper parameters fine-tuning, can be lighter than the CNN, consuming fewer computational resources and taking less training time as demonstrated in the works [12,13].

Introducing convolutions into the Vision Transformer architecture to merge the benefits of Transformers with the benefits of CNNs for image recognition tasks. Convolutional token embedding and convolutional projection, along with the multi-stage design of the network enabled by convolutions, make our CvT architecture achieve superior performance while maintaining computational efficiency. Furthermore, due to the built-in local context structure introduced by convolutions, CvT no longer requires a position embedding, giving it a potential advantage for adaption to a wide range of vision tasks requiring variable input resolution [14].

Memory-Driven Transformer was proposed for automated medical report generation. The decoder layer incorporates, Memory-Conditioned Layer Normalization, enhancing report generation capability [15]. Recently, Large Language Model (LLM) was employed using cyclic techniques for report generation [16].

Recently, large language models have demonstrated excellent capabilities to perform tasks with zero in-domain data, conduct logical reasoning, and apply commonsense knowledge in NLP tasks [17]. This leads us to ponder whether we can apply large language models to medical report generation tasks. As for long text generation, LLMs are equipped with an inherent understanding of grammar, syntax, and semantic coherence, making them well-suited for tasks requiring extended text generation, such as medical reporting.

Con transitive learning is a technique to improve representation learning. Dynamic graph combined with Contrastive Learning in Transformers [18]. This improved visual and text representation in medical report generation task. Furthermore, 3D shared subspace was also explored for representation improvement [19].

CHAPTER 3
REQUIREMENT ANALYSIS

3.1 Functional Requirements

Image Input: The system should accept chest X-ray images in standard formats (e.g., DICOM, JPEG).

Image Preprocessing: Preprocess input images to enhance quality and normalize for analysis.

Feature Extraction: Utilize convolutional vision transformers to extract relevant features from X-ray images.

Text Generation: Employ LLMs to generate descriptive medical reports based on extracted features.

Report Formatting: Ensure generated reports are formatted professionally and include necessary medical terminology.

Output Delivery: Provide the generated reports in a downloadable/printable format for user convenience.

3.2 Non-Functional Requirements

Performance: The system should be capable of processing X-ray images and generating reports within a reasonable timeframe.

Security: Implement robust security measures to protect patient data and ensure compliance with healthcare privacy regulations (e.g., HIPAA).

Usability: Design a user-friendly interface that is intuitive for both radiologists and healthcare professionals to interact with.

Training and Support: Offer comprehensive training and support resources for end-users to effectively utilize the system.

3.3 System Requirements

3.3.1 Hardware Requirements

1. **High-Performance Computing (HPC) Cluster**: Essential for handling large datasets and performing extensive computations.
2. **GPUs**: High-end GPUs such as NVIDIA Tesla V100 or A100 for training deep learning models. At least 2-4 GPUs recommended for parallel processing.
3. **Memory**: Minimum 128 GB RAM to manage large datasets and support GPU operations.
4. **Storage**: At least 1 TB of SSD storage to store datasets and model checkpoints efficiently.

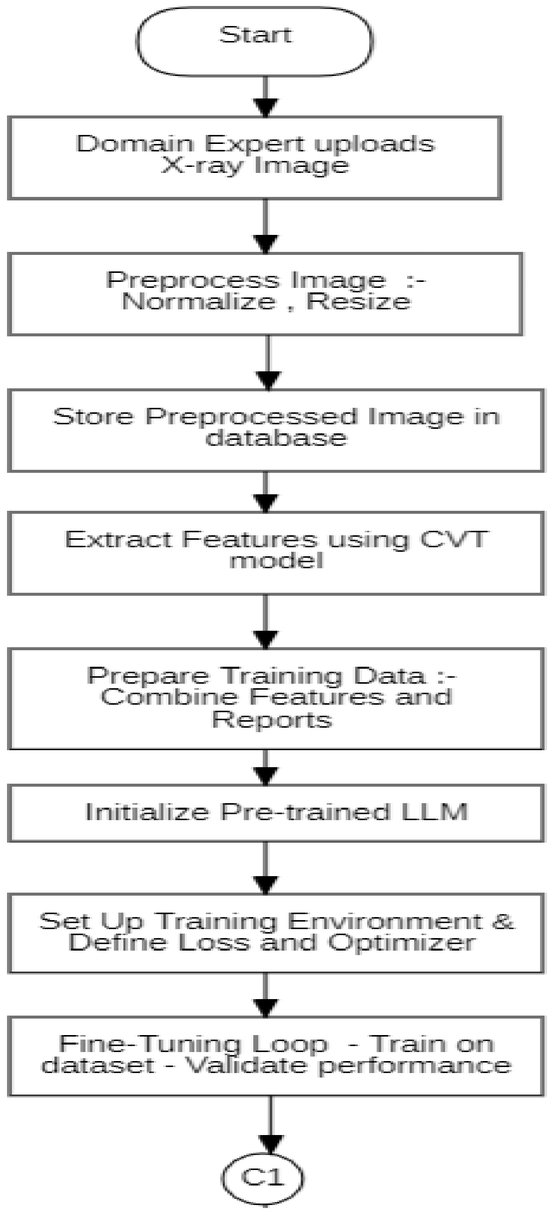
3.3.2 Software Requirements

1. **Operating System**: Windows
2. **Deep Learning Frameworks**:
    - **PyTorch**: Deep learning framework.
3. **Libraries and Dependencies**:
    - **NumPy**: For numerical operations.
    - **Pandas**: For data manipulation and analysis.
    - **scikit-learn**: For evaluation metrics and machine learning utilities.
    - **OpenCV**: For image preprocessing and augmentation.
    - **NLTK** or **spaCy**: For natural language processing tasks.
4. **Pre-trained Models and Tokenizers**:
    - **Transformers** library by Hugging Face: For access to pre-trained models like BERT, GPT-2, and Vision Transformer.
5. **CUDA**: For GPU acceleration.
6. **Jupyter Notebooks**: For interactive development and testing.

CHAPTER 4

SYSTEM DESIGN AND ARCHITECTURE

4.1 Flowchart:



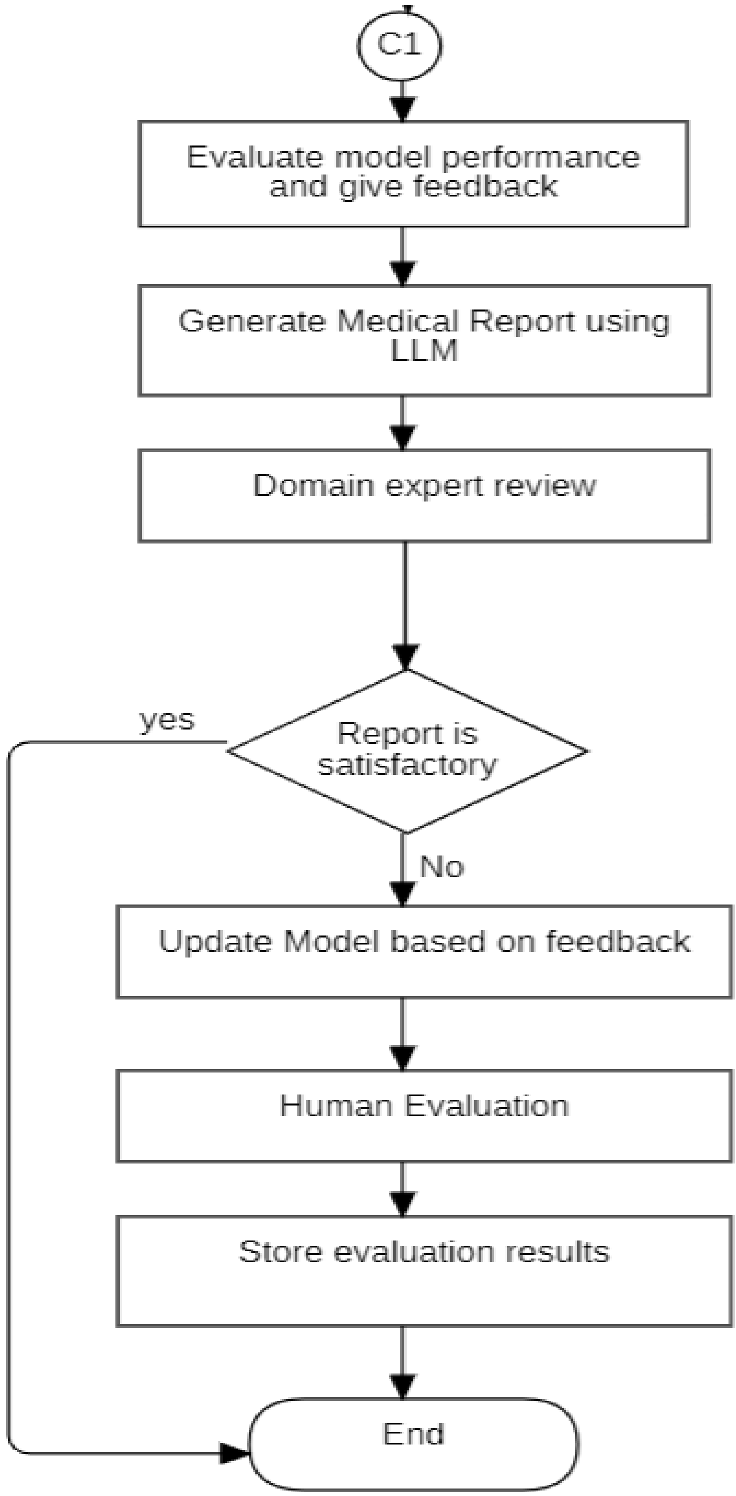


Figure 1: Flowchart

4.2 Use case Diagram:

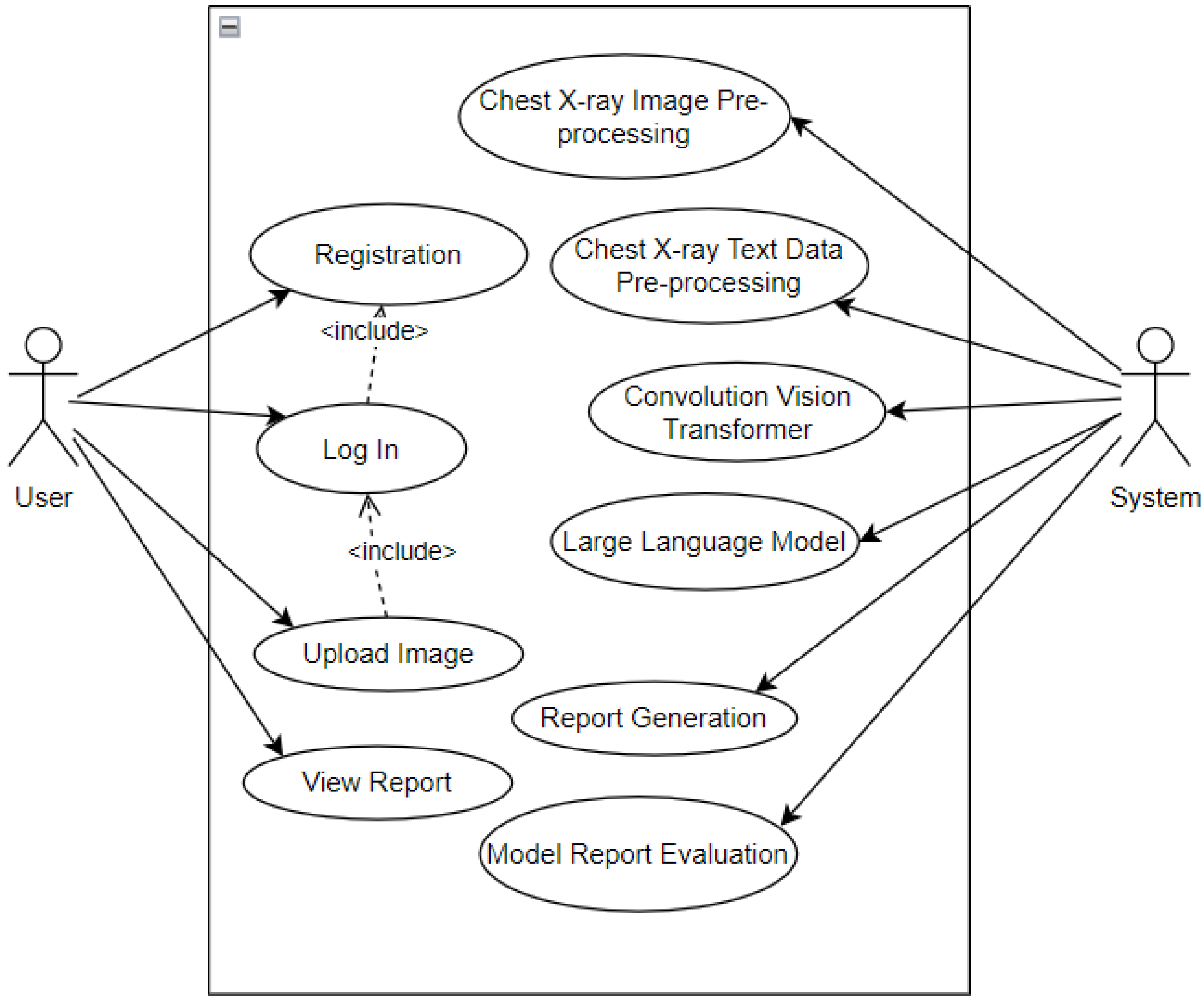


Figure 2: Use case diagram

4.3 DFD Level 0

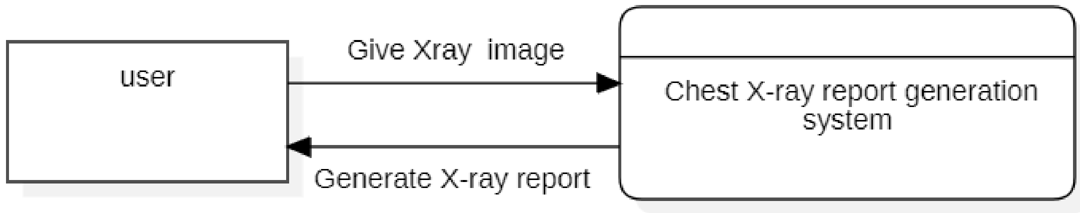


Figure 3: DFD level 0.

4.4 DFD Level 1

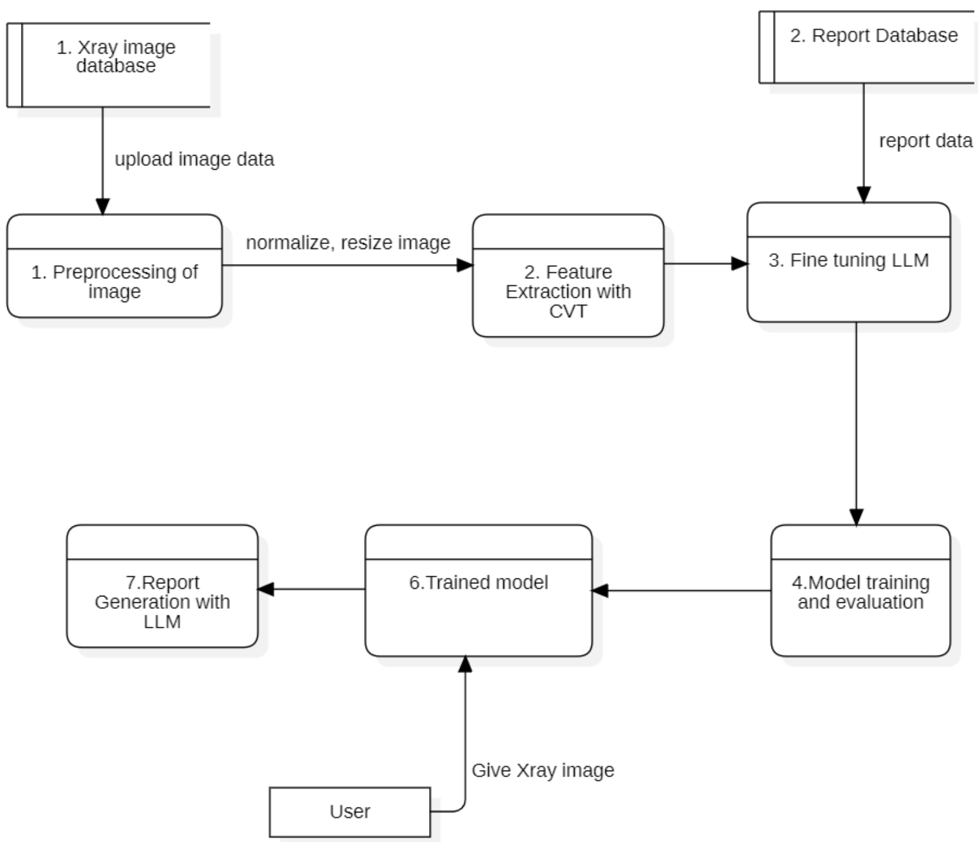


Figure 4: DFD level 1

4.5 Proposed Pipeline:

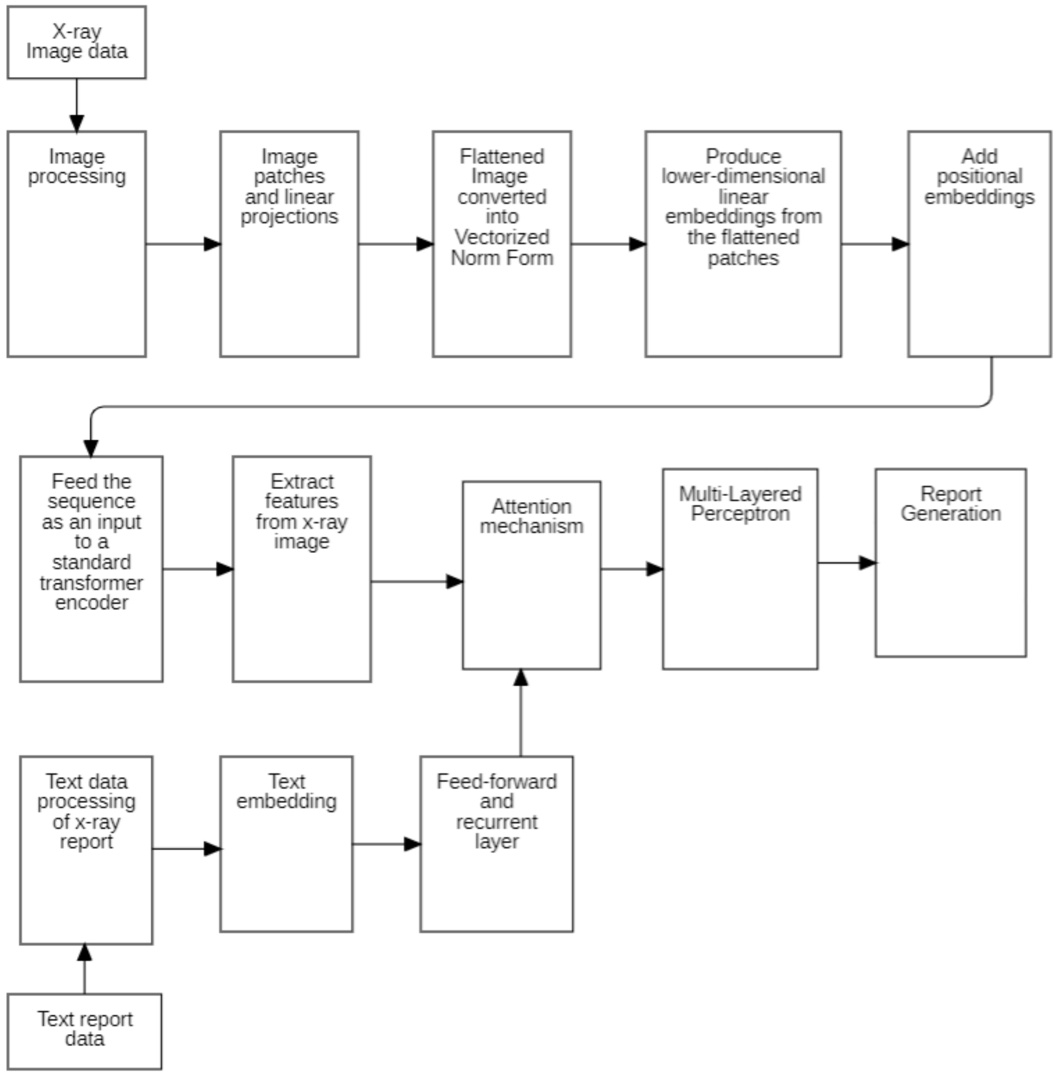


Figure 5: Pipeline for Attention-based Automated Radiology Report Generation

CHAPTER 5

METHODOLOGY

The methodology for our project consists of five main phases: data collection, preprocessing, feature extraction with a Convolutional Vision Transformer (CVT), report generation using a Language Model (LLM), and thorough evaluation. Here is a detailed breakdown of each step:

5.1 Data Collection

To develop a robust system for automated medical report generation, we will utilize several publicly available datasets, ensuring diversity and comprehensiveness:

- **MIMIC-CXR**: Comprising over 377,110 chest radiographs with corresponding radiology reports from the Beth Israel Deaconess Medical Center, this dataset is invaluable for training and validation (CAIMI).
- **CheXpert**: Contains 224,316 chest radiographs along with labeled reports from Stanford University, facilitating extensive model training (CAIMI).
- **PadChest**: Offers 160,868 chest X-rays with multi-label annotations, including radiological findings and diagnoses, from BIMCV (CAIMI).
- **OpenI Chest X-ray Dataset**: A collection of chest X-rays and corresponding radiology reports from the National Library of Medicine, useful for detailed model evaluation (CAIMI).

5.2 Data Preprocessing

5.2.1 Image Preprocessing

All images will be resized to a standard dimension (e.g., 224x224 pixels) and normalized to ensure consistent input for the model. This involves adjusting pixel values to have zero mean and unit variance.

5.2.2 Report Tokenization:

Radiology reports will be tokenized using a tokenizer compatible with the chosen language model, converting the text into a format suitable for model input (e.g., words or sub-words transformed into numerical tokens).

5.3 Feature Extraction with Vision Transformer

5.3.1 Model Selection:

We will employ a pre-trained Convolution Vision Transformer (CVT), leveraging its proven effectiveness in capturing complex spatial dependencies in images.

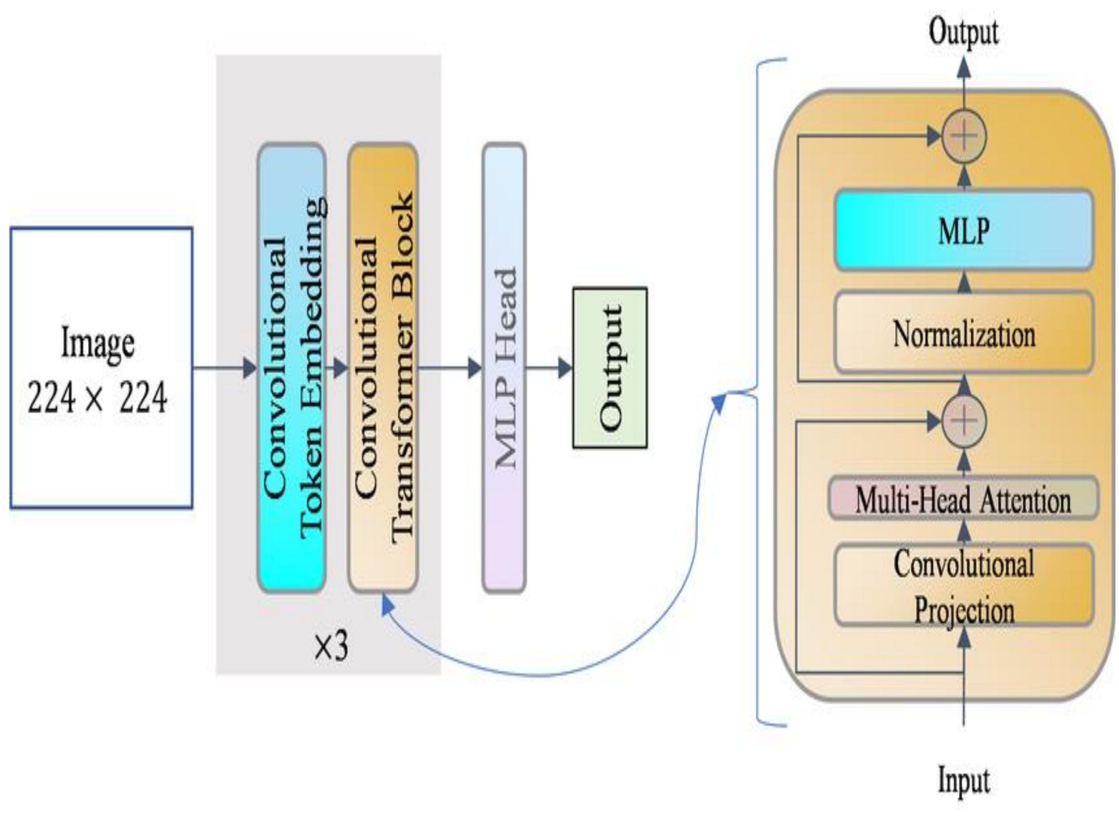


Figure 6: Convolution Vision Transformer
(Source:https://www.researchgate.net/publication/378130765/figure/fig5/AS:11431281223231110@1707683631709/The-architecture-of-CVT.ppm)

The Convolutional Vision Transformer (CVT), tailored for feature extraction in the context of chest X-ray report generation as it integrates convolutional layers and transformer encoders to encompass both fine-grained details and holistic context from images. The convolutional layers identify local features such as edges and textures within the X-ray images. These features are subsequently partitioned into consistent-sized patches, flattened, and converted into a sequential array of vectors through a linear embedding layer. To retain spatial information, positional encodings are incorporated into these patch embedding's. This series of embedded patches is then inputted into multiple transformer encoder layers. Here each transformer layer employs self-attention mechanisms that enable the model to evaluate the significance of distinct patches in correlation to one another that captures spatial relationships and comprehensive context. The output from the transformer encoders yields a comprehensive, high-dimensional feature representation of the entire image, which is then aggregated into a single feature vector through global pooling. This feature vector wrapped the useful and meaningful information from the X-ray image and can be used for subsequent tasks such as report generation. This hybrid approach accelerator the strengths of both convolutional and transformer architectures that provide an effective means of extracting the detailed and contextual features from the images.

5.3.2 Fine-tuning:

The CVT model will be fine-tuned on our chest X-ray datasets to tailor its feature extraction capabilities to the specific characteristics of medical images.

5.3.3 Training Process:

Dataset Splitting: The data will be split into training (70%), validation (15%), and test (15%) sets to ensure robust evaluation.

Augmentation and Optimization: Techniques such as data augmentation (e.g., flipping, rotation, scaling) and hyper parameter optimization (e.g., learning rate, batch size) will be used to enhance model performance and generalization.

5.4 Report Generation with Language Model

5.4.1 Model Selection

A powerful pre-trained language model (e.g., GPT-2) will be selected for generating the text reports.

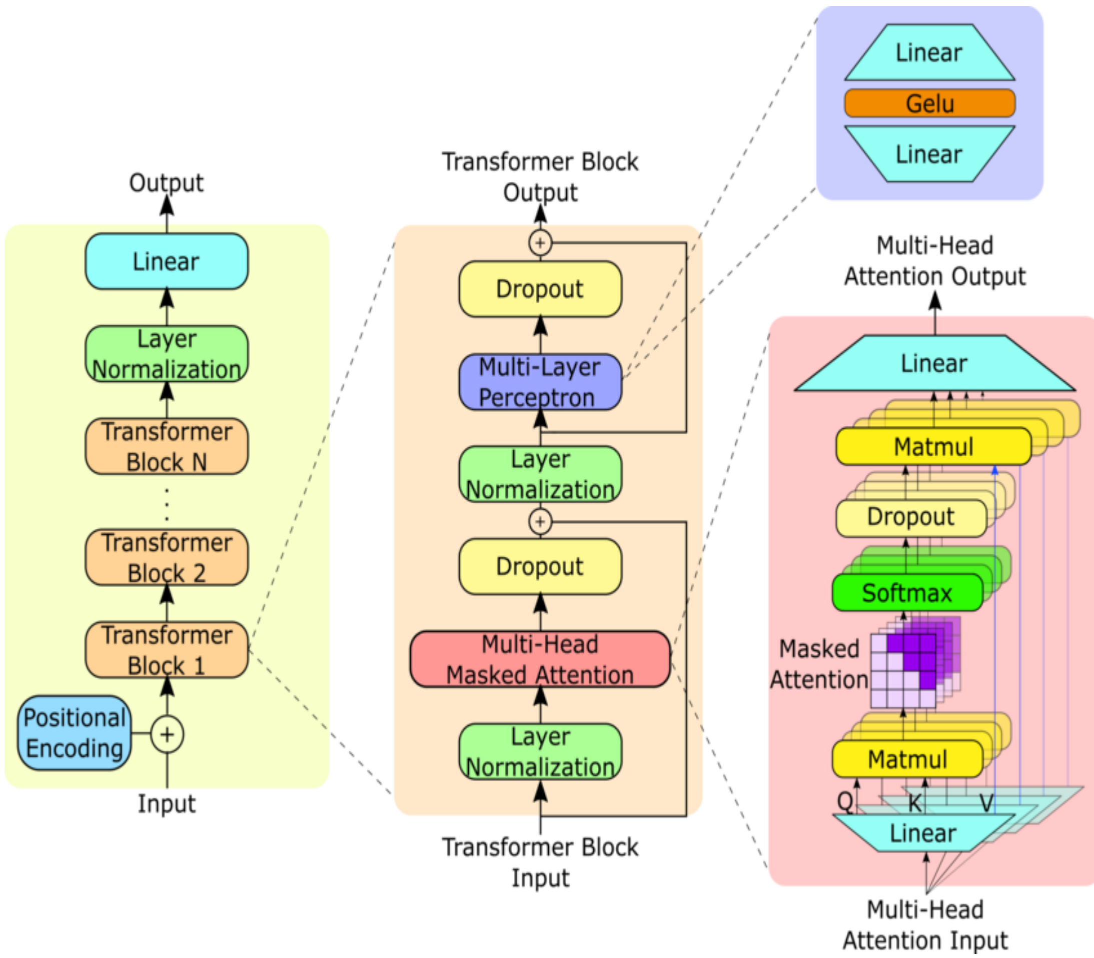


Figure 7: GPT 2

(https://www.researchgate.net/publication/373352176/figure/fig1/AS:11431281202501967@1698856108167/GPT-2-model-architecture-The-GPT-2-model-contains-N-Transformer-decoder-blocks-as-shown.ppm)

After extracting features from the chest X-ray images using Computer Vision Transformer, here the GPT-2 (Generative Pre-trained Transformer 2) model is employed for generating text reports based on these extracted features. Here in this context, the GPT-2 architecture operates as a language generation model. To start with, the GPT-2 model used has a number of transformer decoders where input is image features that have been extracted. These layers are sequential in nature and produce text that accommodates both the accompanying context provided through the previous tokens generated and the extracted image features. Using its pre-training on a challenging text corpus to build up contextual understanding, GPT-2 generates reports about the findings observed in chest x-rays that are contextually meaningful. This integration brings together visual information and natural language generation abilities which facilitate automatic production of full-length detailed medical reports having high accuracy rates.

5.4.2 Integration:

The one-hot feature vectors of the image can then be fed to the language model that predicts next word. By combining CVT (Computer Vision Transformer) and GPT-2 (Generative Pre-trained Transformer 2), it takes advantage of computer vision as well as natural language processing. In its early stages, CVT deals with chest X-ray images and extracts high-level visual features that capture key patterns and structures in them. The above functions of these graphs serve as a meaningful input unit for generating quality text by models like GPT-2' that are effective at understanding and generating human-like text. A transformer decoders architecture pre-trained on large amounts of textual data provide a background for the GPT-2 system with an additional channel for feeding in those extracted visual features to generate good texts reports. This integrated system does not only interpret the visual information but also provides detailed and accurate medical findings present in the chest X-ray images through automated report generation, thus enabling quick production of medical reports.

5.4.3 Fine-tuning:

The language model will be fine-tuned on the tokenized radiology reports to enable it to generate coherent and medically accurate reports based on the image features.

5.4.4 Loss Function:

Cross-entropy loss will be used to measure the accuracy of the generated reports compared to the actual reports, guiding the optimization process.

5.5 Model Evaluation

5.5.1 Metrics

The performance of our integrated model will be given access using metrics such as BLEU score (measuring precision of generated text), ROUGE score (evaluating recall of generated text), and medical accuracy (specificity and sensitivity for clinical relevance).

5.5.2 Validation

Our model will be validate using a separate validation set to ensure that it generalizes well to unseen data. Cross-validation techniques will be employed to enhance reliability.

5.5.3 Benchmarking

Our model's performance will be comparing against the existing models to identify improvements and areas for the further enhancement of our model.

CHAPTER 6

EXPECTED OUTPUT

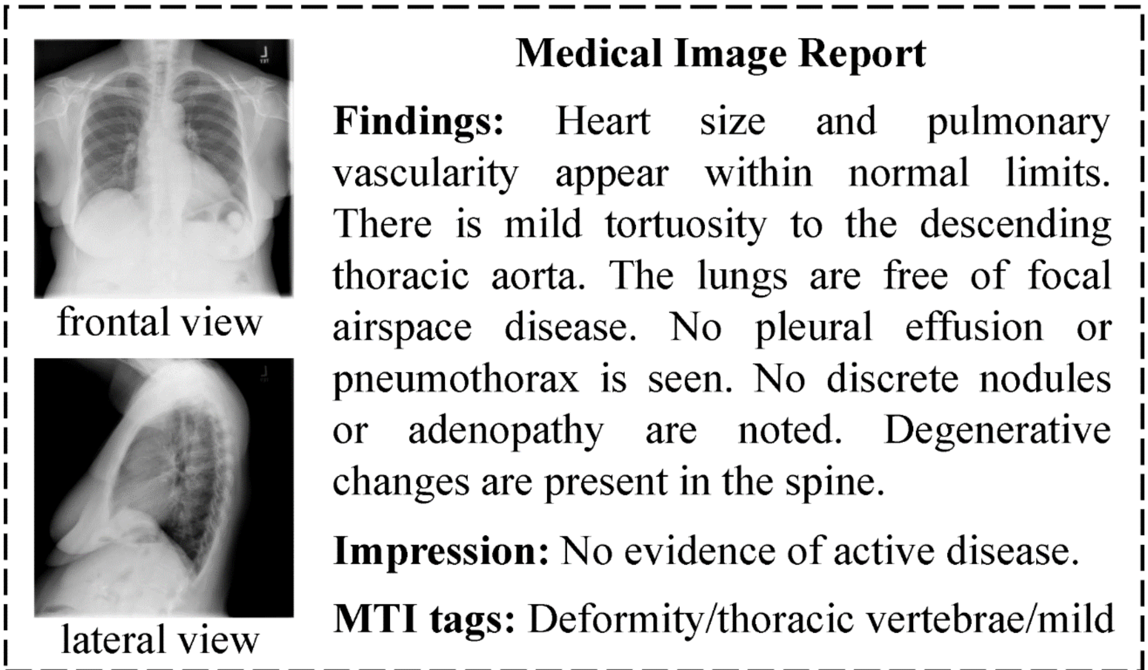


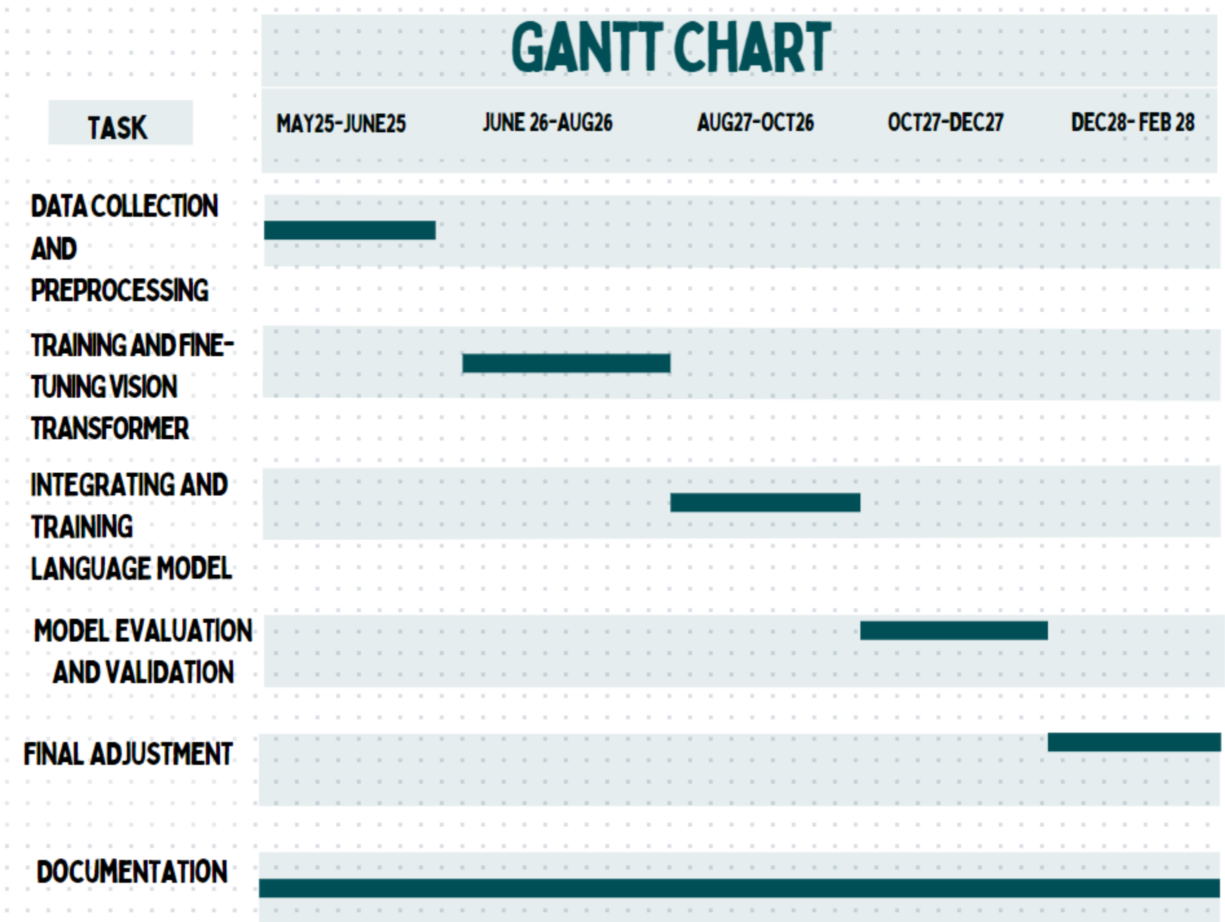
Figure 8: Expected Output

(Source: https://www.mdpi.com/bioengineering/bioengineering-10-00966/article_deploy/html/images/bioengineering-10-00966-g001.png)

CHAPTER 7

TIME SCHEDULE

Table 1 GANTT CHART



REFERENCES

[1] “Hurdles To Hospitals,” The Rising Nepal, Mar. 13, 2024. https://risingnepaldaily.com/news/23880#:~:text=The%20doctor-patient%20ratio%20in,patient%20ratio%20of%201%3A1%2C000 (accessed May 23, 2024).

[2] F. F. Alqahtani, M. M. MOHSAN, K. ALSHAMRANI, J. ZEB, S. ALHAMAMI, and D. ALQARNI, “CNX-B2: A Novel CNN-Transformer Approach For Chest X-ray Medical Report Generation,” Feb. 19, 2024. https://www.semanticscholar.org/paper/CNX-B2%3A-A-Novel-CNN-Transformer-Approach-For-Chest-Alqahtani-Mohsan/abb3949208e490c4ea5fc70ad092f414ab52d30a (accessed May 23, 2024).

[3] I. Allaouzi, M. B. Ahmed, B. Benamrou, and M. Ouardouz, ‘‘Automatic caption generation for medical images,’’ in Proc. 3rd Int. Conf. Smart City Appl. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 1–6.

[4] J. Yuan, H. Liao, R. Luo, and J. Luo, ‘‘Automatic radiology report generation based on multi-view image fusion and medical concept enrichment,’’ in Medical Image Computing and Computer Assisted Intervention. MICCAI 2019 (Lecture Notes in Computer Science), vol. 11769. Midtown Manhattan, NY USA: Springer, 2019, pp. 721–729.

[5] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum, and J. Mongan, ”Highthroughput classification of radiographs using deep convolutional neural networks,” J. Digital Imaging, vol. 30, no. 1, pp. 95-101, Feb. 2017.

[6] P. Lakhani and B. Sundaram, ”Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks,” Radiology, vol. 284, no. 2, pp. 574-582, Aug. 2017.

[7] M. Cicero, A. Bilbily, E. Colak, D. Kontos, and J. Mermelstein, ”Training and validating a deep convolutional neural network for computeraided detection and classification of abnormalities on frontal chest radiographs,” Investigative Radiology, vol. 52, no. 5, pp. 281-287, May 2017.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[9] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, “Trocr: Transformer-based optical character recognition with pre-trained models,” in AAAI 2023, February 2023. [Online].

[10] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, ‘‘Image transformer,’’ in Proc. 35th Int. Conf. Mach. Learn., vol. 80, J. Dy and A. Krause, Eds., Jul. 2018, pp. 4055–4064. [Online]. Available: https://proceedings.mlr.press/v80/parmar18a.html

[11] Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner,and X. Zhai, ‘‘An image is worth 16 × 16 words: Transformers for image recognition at scale,’’ in Proc. 9th Int. Conf. Learn. Represent. (ICLR), 2021.

[12] Wang, H. Traffic Sign Recognition with Vision Transformers. In Proceedings of the 6th International Conference on Information System and Data Mining, Silicon Valley, CA, USA, 27–29 May 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 55–61.

[13] Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A. Single-Layer Vision Transformers for More Accurate Early Exits with Less Overhead. Neural Netw. 2022, 153, 461–473.

[14] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, andL. Zhang, “Cvt: Introducing convolutions to vision transformers,” in Proceedings of the IEEE/CVF international conferenceon computer vision, pp. 22–31, 2021.

[15] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1439–1449.

[16] W. Chen, X. Li, L. Shen, and Y. Yuan, “Fine-grained image-text alignment in medical imaging enables cyclic image-report generation,” arXiv preprint arXiv:2312.08078, 2023.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” 2023.

[18] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, “Dynamic graph enhanced contrastive learning for chest x-ray report generation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3334–3343.

[19] J. You, D. Li, M. Okumura, and K. Suzuki, “Jpg-jointly learn to align: Automated disease prediction and radiology report generation,” in Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 5989–6001.