Multi-Modal Large Language Model for Visual Question Answering on Medical Domain

A dissertation submitted in partial fulfilment of the requirements
for the award of the degree of

M.Tech.

in

Computer Science

By

Srimanta Singha (Roll No. CS2326)

under the supervision of

DR.Ujjwal Bhattacharya

CVPR Unit, ISI, Kolkata

INDIAN STATISTICAL INSTITUTE

BT ROAD, KOLKATA-700108, INDIA

CERTIFICATE

This is to certify that the dissertation entitled **"MultiModal Large Language Model for Visual Question Answering on Medical Domain"** submitted by Srimanta Singha, having roll number CS2326 to Indian Statistical Institute, Kolkata in partial fulfillment for the award of the degree of Master of Technology in Computer Science is a bonafide record of work carried out by him under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

11/04/2025

(Supervisor's Signature)

DR.Ujjwal Bhattacharya
Indian Statistical Institute
Kolkata

Acknowledgements

I extend my sincere appreciation to **Dr.Ujjwal Bhattacharya**, my advisor at the Computer Vision and Pattern Recognition Unit of the Indian Statistical Institute in Kolkata, for her guidance, continuous support, and inspiration. Her profound knowledge and creative suggestions have taught me a great deal in every subject and have shown me how to conduct solid research.

I am deeply grateful to all the teachers at the Indian Statistical Institute for their invaluable advice, insights, and instruction, which provided a crucial perspective to my research. Finally, I want to express my gratitude to my parents and extended family for their unwavering support. I also extend my sincere appreciation to all my friends for their continuous assistance and encouragement. I am thankful to everyone who has contributed to my growth and success, even if I have inadvertently missed mentioning them in the above list.

Date: 11/06/2025

Srimanta Singha.
Roll: CS2326
M.Tech CS, 2nd year
Indian Statistical Institute

Abstract

Artificial intelligence (AI) strategies such as Multimodal learning, which can integrate inputs of multiple modes, e.g., image and text, have shown significant promise in medical applications. In this dissertation, we present our related study of a Multimodal Large Language Model (MLLM) designed for Visual Question Answering (VQA) in the medical domain, based on both image and text input modalities to improve diagnostic reasoning and decision support. Our model processes medical images (e.g., chest X-rays, CT scans, and ultrasound images) along with clinical text to answer complex, domain-specific questions. We employ a cross-modal fusion mechanism to align visual features with textual embeddings, enabling the model to generate accurate and contextually relevant responses.

In this work, we have studied two different datasets, one is **ImageCLEF 2019** medical VQA dataset and the other is **MED-GRIT-270K** dataset.

First, we work on ImageCLEF 2019 medical VQA dataset and our approach demonstrates superior performance compared to existing multimodal baselines on same dataset, achieving state-of-the-art results in diagnostic precision and interpretability.

Furthermore, to address the limitations of existing datasets, we reformat ImageCLEF 2019 VQA into a descriptive answer-style dataset and fine-tune Vision-LLM on this enhanced dataset to improve its medical reasoning capabilities.

Second, to specialize the model for chest X-ray analysis, we extract a subset of radiology images and paired text from the MED-GRIT-270K dataset, then fine-tune the VLLM to create a robust chest X-ray AI system.

Contents

1. Introduction 1
- 1.1. Objectives . . . 1

2. Problem Definition 3

3. Existing Techniques and Related Work 5
- 3.1. Datasets . . . 5
- 3.2. **Existing Techniques in Medical VQA** . . . 6
  - 3.2.1. Image Encoders . . . 6
  - 3.2.2. Text Encoder (Question Embedding) . . . 6
  - 3.2.3. Multi-modal Factorized Bilinear (MFB) Fusion . . . 6
  - 3.2.4. LSTM as Decoder for Answer Generation . . . 7
- 3.3. **Related Works** . . . 7
- 3.4. Evaluation Metrics . . . 8

4. Proposed Method for Medical VQA 9
- 4.1. **Image Processing** . . . 9
- 4.2. **Proposed Architecture** . . . 10
  - 4.2.1. **Multimodal Factorized High-order (MFH) Block** 10
  - 4.2.2. **Image Attention Mechanism** . . . 10
  - 4.2.3. **Transformer-Based Decoder** . . . 12
- 4.3. **Architecture** . . . 12
- 4.4. **Model Performance & Results:** . . . 13

5. Adapting Vision-Language Large Models for Downstream Tasks 17

6. Conclusion 21

7. Future Work 22

A. References 23

1. Introduction

Artificial intelligence(AI) technology has a wide range of applications and has been used extensively to build parts of bigger systems since 2019s. The advancement of deep learning has led to an exponential development in AI applications. In essence, deep learning is a branch of AI. AI can be applied to a wide range of tasks, including object recognition, image captioning, machine translation, computer vision, and natural language understanding.

Medical Visual Question Answering (VQA) is an emerging field in artificial intelligence (AI) that combines computer vision and natural language processing to interpret and reason over medical images in response to text-based queries. It holds significant promise for clinical decision support, particularly in radiology and diagnostic medicine.

In this project, we explore multimodal learning approaches for medical VQA using two major datasets: ImageCLEF 2019 VQA-Med and MED-GRIT-270K. Initially, we focus on the ImageCLEF 2019 dataset and propose a transformer-based decoder architecture, which outperforms existing baseline models in terms of answer accuracy and interpretability. To further enhance model performance, we reformat the dataset to include descriptive-style answers, enabling the model to generate more informative and context-aware responses. This enhanced dataset is then used to fine-tune a Vision-Language Large Model (VLLM), resulting in a more robust and explainable medical VQA system.

1.1. Objectives

The primary objective of this project is to develop an advanced **Vision-Language Large Model (VLLM)** framework for **Medical Visual Question Answering (VQA)** that can effectively understand and reason over both medical images and clinical text. The project aims to enhance diagnostic support and interpretability through multimodal learning.

To achieve this, the project is guided by the following specific objectives:

- To design and implement a transformer-based decoder architecture tailored for the **ImageCLEF 2019 VQA-Med** dataset, demonstrating improved performance over existing multimodal baselines in terms of accuracy and interpretability.
- To enhance the ImageCLEF 2019 dataset by converting short-form answers into **descriptive answer formats**, allowing the model to produce more informative and clinically meaningful responses.
- To fine-tune a **Vision-Language Large Model (VLLM)** on the enhanced dataset, thereby building a robust and context-aware medical VQA system.
- To specialize the VLLM for **chest X-ray interpretation** by curating a subset of radiology image-text pairs from the **MED-GRIT-270K** dataset and fine-tuning the model for domain-specific question answering.

This project ultimately aims to bridge the gap between visual and textual medical data, showcasing how large language models can be effectively extended to multimodal tasks for real-world healthcare applications.

2. Problem Definition

In the field of medical diagnostics, interpreting complex visual data such as X-rays, CT scans, or pathology slides often requires domain-specific knowledge and significant clinical experience. With the increasing volume of medical imaging data, there is a pressing need for intelligent systems that can assist clinicians by automatically analyzing images and answering clinically relevant questions.

Medical Visual Question Answering (VQA) is a challenging multimodal task that involves providing accurate natural language answers to questions posed about medical images. This task requires models to jointly understand visual content and textual context, often involving nuanced medical terminology and reasoning.

Traditional VQA systems struggle with medical data due to:

- Limited domain-specific datasets with high-quality annotations.
- Short, often vague answer formats that reduce reasoning richness.
- Lack of models fine-tuned on complex biomedical image-text pairs.

The goal of this project is to develop a **Vision-Language Large Model (VLLM)**-based system that can overcome these challenges by:

1. Processing diverse types of medical images.
2. Answering complex, clinically relevant questions in a descriptive and context-aware manner.
3. Supporting decision-making in diagnostic scenarios.

A few examples of questions associated with the medical image (X-Ray) of Fig. 2.1 are listed below.

Sample Questions:

- What abnormality is visible in the left lung?
- What organ system is visible in this image?

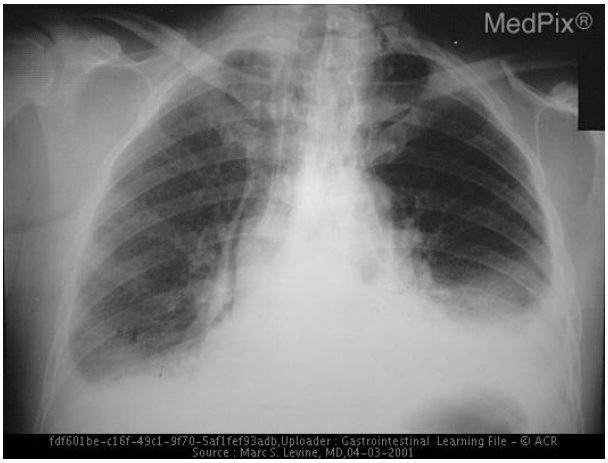


Figure 2.1.: Example of chest X-ray

- What is the most alarming finding in this image of Figure 2.1?

Most of the questions in the medical domain can be challenging for non-experts to answer accurately. While doctors are trained to interpret medical images, even for them, identifying the underlying issue is not always immediate and may require specialized analysis.

To address this, we develop an AI-based Visual Question Answering (VQA) system tailored for the medical domain. The system is trained on a diverse set of medical images and associated clinical questions, enabling it to generate accurate answers and provide meaningful interpretations of potential abnormalities.

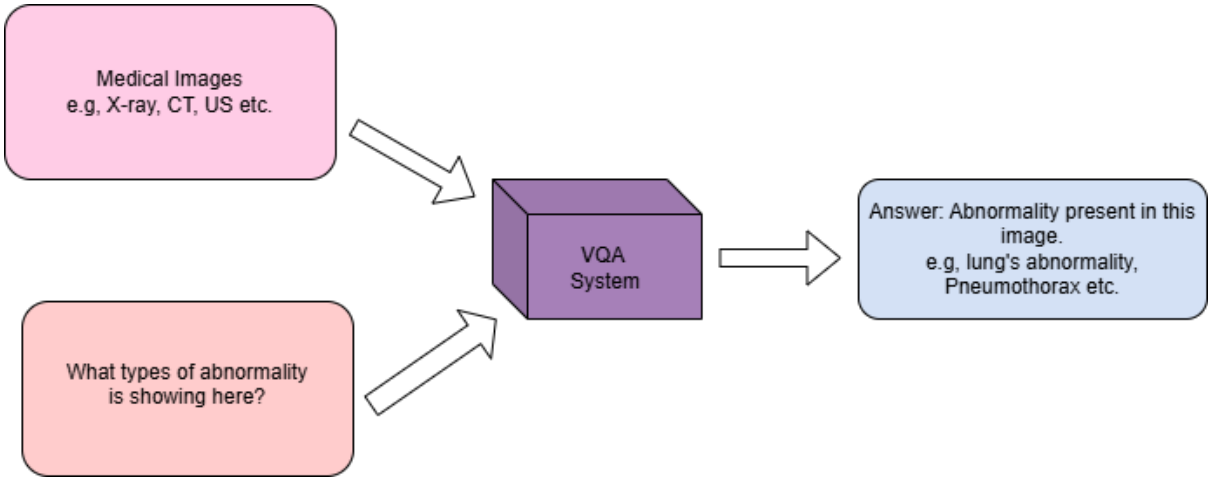


Figure 2.2.: VQA system workflow

3. Existing Techniques and Related Work

3.1. Datasets

Deep Learning (DL), a subset of Artificial Intelligence (AI), requires large and high-quality datasets to effectively train models. These models involve numerous hyperparameters that must be carefully tuned during the training process to achieve optimal performance. The more comprehensive and diverse the training data, the better the model can generalize to unseen examples. Therefore, the quality and quantity of data play a critical role in the success of any Machine Learning (ML) or Deep Learning (DL) approach.

Many existing techniques and related studies in the field of medical Visual Question Answering (VQA) are primarily based on the ImageCLEF 2019 VQA-Med dataset.

- **VQA-MED 2019:** The VQA-Med 2019 challenge focused on radiology images and included four primary categories of questions: Modality, Plane, Organ System, and Abnormality. These categories were structured to vary in difficulty and were designed to support both classification-based and generative answering approaches. In this second edition of the VQA-Med challenge, the questions were specifically crafted to target a single visual element per instance—for example: "What organ is primarily shown in this MRI?", "In what plane is this mammogram taken?", "Is this a T1-weighted, T2-weighted, or FLAIR image?", or "What is most alarming about this ultrasound?". These questions were intentionally designed to be answerable using only the visual content of the image, without requiring external medical knowledge or domain-specific reasoning.

3.2. Existing Techniques in Medical VQA

Medical VQA systems typically employ a **multimodal architecture** comprising:

3.2.1. Image Encoders

- **VGG16**: A classical CNN architecture with 16 weight layers, used for feature extraction from medical images. It applies sequential 3×3 convolutional filters and max-pooling to capture hierarchical visual patterns (e.g., lung opacities in X-rays). Despite its simplicity, its deep structure risks losing fine-grained details due to aggressive pooling.
- **ResNet152**: A residual network with 152 layers, leveraging skip connections to mitigate vanishing gradients. Its ability to preserve low-level features (e.g., microcalcifications in mammograms) through identity mapping makes it superior to VGG16 for high-resolution medical imaging.

3.2.2. Text Encoder (Question Embedding)

- **BERT**: A transformer-based model that generates context-aware embeddings for clinical questions. By pretraining on biomedical corpora (e.g., PubMed), it captures domain-specific semantics (e.g., distinguishing "consolidation" from "atelectasis"). In VQA, BERT is commonly used to encode the input question text, capturing rich semantic and syntactic information to produce meaningful question embeddings. These embeddings facilitate effective understanding and reasoning over the textual modality, crucial for accurate answer prediction.

3.2.3. Multi-modal Factorized Bilinear (MFB) Fusion

- **Multi-modal Factorized Bilinear Pooling (MFB)**: A fusion mechanism that combines visual ($\mathbf{V}$) and textual ($\mathbf{Q}$) features via:

$$\mathbf{z} = \text{SumPooling}(\sigma(\mathbf{U}^T\mathbf{V}) \odot \sigma(\mathbf{W}^T\mathbf{Q}), k) \tag{3.1}$$

  where $\mathbf{U}, \mathbf{W}$ are projection matrices, σ is ReLU, and k is the factor size. MFB outperforms simple concatenation or Hadamard product

by factorizing high-dimensional interactions, crucial for aligning radiology terms with image regions (e.g., linking "pleural effusion" to costophrenic angle blunting).

3.2.4. LSTM as Decoder for Answer Generation

- **LSTM**: A recurrent network that generates descriptive answers autoregressively. It conditions on the fused MFB output ($\mathbf{z}$) and previous tokens, leveraging attention over image features to produce clinically coherent responses (e.g., "Right middle lobe infiltrate suggests pneumonia"). While prone to long-range dependency issues, its simplicity suits small-scale medical datasets.

3.3. Related Works

Visual Question Answering (VQA) has been extensively studied in the general domain, with numerous methods exploring diverse architectures for effective multi-modal reasoning. However, in the medical domain, the majority of research efforts have focused on classification-based VQA, where the model selects the most likely answer from a predefined set of choices. These approaches often rely on discriminative models that map image-question pairs to classification labels without explicitly generating answers.

For answer generation tasks in medical VQA, recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, are frequently employed as decoders. These models generate answers in a sequence-to-sequence manner, providing flexibility for open-ended responses, which is crucial in medical settings where answers are often descriptive.

To bridge the visual and textual modalities, several studies utilize Multimodal Factorized Bilinear (MFB) [1] pooling for cross-modal fusion. MFB effectively captures the interactions between image features and question embeddings, offering a more expressive joint representation compared to simple concatenation or element-wise fusion. In some cases, MFB is used as the sole fusion mechanism, without further sequential reasoning modules.

Despite these advances, medical VQA still faces challenges such as limited annotated data, domain-specific terminology, and the need for high reliability. As a result, integrating powerful feature encoders (e.g., CNNs

and transformers) with efficient fusion strategies and generative decoders remains an active area of research.

3.4. Evaluation Metrics

To evaluate the performance of the Visual Question Answering (VQA) model on medical images, we employed the **BLEU (Bilingual Evaluation Understudy)** score, a widely used metric for evaluating the quality of text generation tasks such as machine translation and image captioning. In the context of VQA, BLEU measures how closely the generated answer matches one or more ground-truth answers by computing n-gram overlaps. The BLEU score ranges from 0 to 1, where a score closer to 1 indicates higher similarity with the reference answer.

Mathematically, the BLEU-n score is defined as:

$$\text{BLEU-n} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where:

- p_n is the modified precision for n-grams,
- w_n is the weight for each n-gram length (typically uniform, e.g., $w_n = \frac{1}{N}$),
- BP is the **brevity penalty** to penalize overly short generated answers, defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases}$$

Here, c is the length of the candidate (generated) answer and r is the effective reference length.

Example:
If the reference answer is `"No signs of pneumonia"` and the model generates `"No pneumonia detected"`, the BLEU score evaluates the overlap of 1-grams, 2-grams, etc., and quantifies how many of the generated n-grams are present in the reference.

4. Proposed Method for Medical VQA

4.1. Image Processing

Medical images pose significant challenges for feature extraction due to their sparse nature, where large regions of the image are often uniformly black or contain irrelevant background information. To address this, we implement a dynamic cropping technique combined with image enhancement methods to isolate and emphasize the region of interest.

Initially, we enhance the input image using a series of preprocessing steps, including Gaussian blurring to reduce noise, followed by morphological operations such as erosion and dilation to refine object boundaries. After enhancement, we identify the four extreme points—top, bottom, left, and right—of the informative region by filtering out pixels with intensity values below a threshold of 25, determined empirically through trial and error. A bounding rectangle is then generated using these extreme points, and all areas outside this region are discarded.

The result is a cropped and enhanced image that retains only the clinically relevant content, thereby improving the accuracy and efficiency of visual feature extraction in the downstream VQA pipeline.

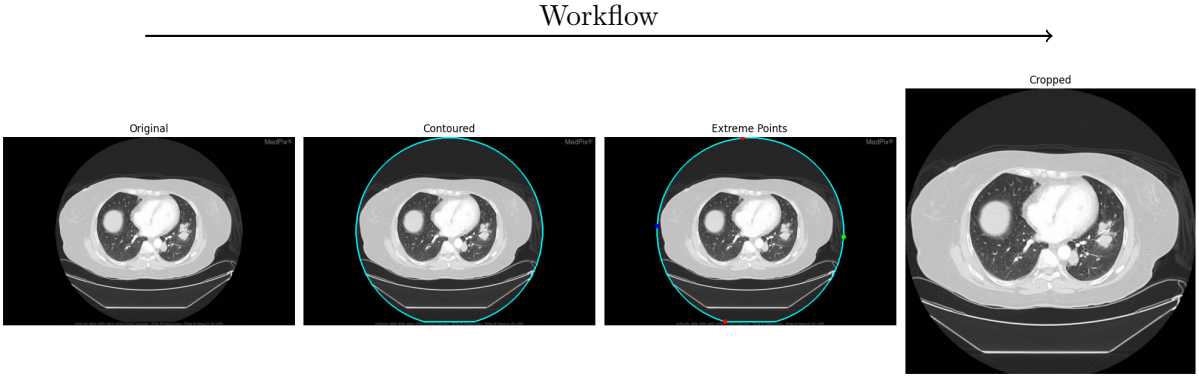


Figure 4.1.: Image Processing Flows

Note:
(a) The leftmost image shows the original input image.

(b) The second image (from left) highlights the contours detected after removing rare/unnecessary pixels (intensity < 25).
(c) The third image displays the extreme points identified on the largest contour.
(d) The rightmost image shows the cropped output, obtained by drawing a rectangle through these extreme points.

4.2. Proposed Architecture

In this section, we describe the key components of our proposed Visual Question Answering (VQA) architecture tailored for medical images. The overall model is designed to enhance multimodal interaction between medical images and clinical questions using advanced fusion, attention, and decoding strategies. The architecture is composed of the following main modules:

4.2.1. Multimodal Factorized High-order (MFH) Block

To capture more complex interactions between image and question modalities, we extend the traditional Multi-modal Factorized Bilinear (MFB) pooling approach by stacking multiple MFB layers, thereby forming a Multimodal Factorized High-order (MFH) block. This stacked structure allows the model to learn high-order correlations beyond simple bilinear fusion. In this block, visual features extracted from a convolutional neural network (e.g., ResNet-152 or VGG16) and textual features encoded via a pre-trained BERT model are passed through consecutive MFB layers. Each MFB layer captures pairwise interactions and produces intermediate joint representations. These intermediate outputs are concatenated to form a comprehensive high-order fused representation. The output of this MFH block serves as the input to the subsequent image attention mechanism.

4.2.2. Image Attention Mechanism

Following the MFH fusion, we apply a spatial image attention mechanism to emphasize the most relevant regions of the image with respect to the fused multimodal context. The attention module computes attention weights across all spatial locations and channels, producing a fine-grained relevance map.

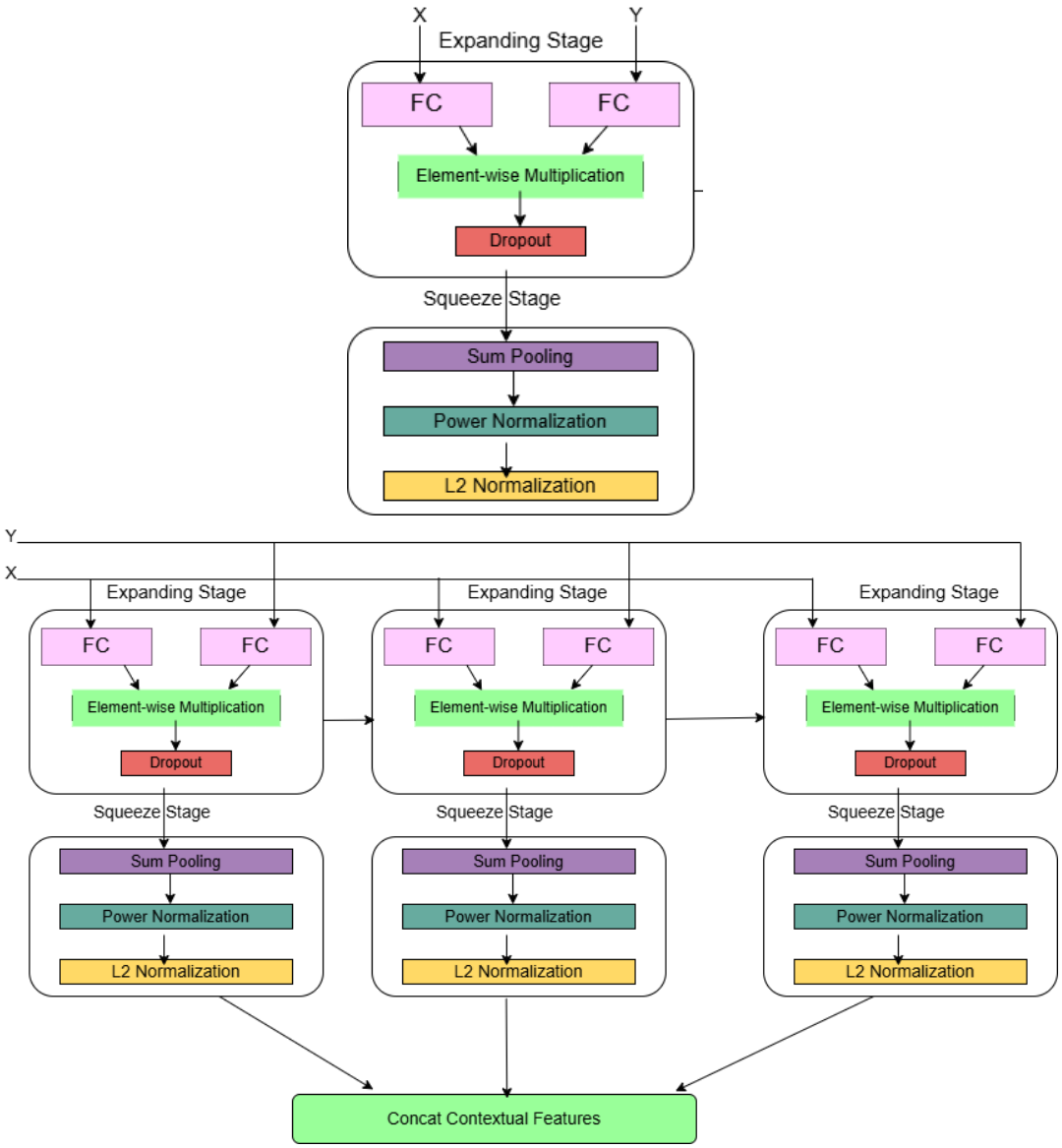


Figure 4.2.: MFB(top) & MFH(bottom) workflows

Instead of directly using the attended spatial feature map (obtained by element-wise multiplication of the attention weights with the image feature map), we perform a spatial compression. This is done by aggregating the attended feature map across spatial dimensions (e.g., using weigted sum pooling), resulting in a single compact feature vector that captures the most salient visual information.

This compressed attended image representation is then passed through an additional MFB layer, where it is again fused with the attended question representation. This second stage of cross-modal interaction further strengthens the alignment between visual and textual modalities.

4.2.3. Transformer-Based Decoder

To improve the answer generation process, we replace traditional recurrent decoder architectures such as LSTM with a Transformer decoder. The Transformer decoder is more adept at modeling long-range dependencies and parallel sequence generation, which is particularly beneficial in the medical domain where question-answer pairs may involve complex terminology and structure.

For word embeddings within the decoder, we employ pre-trained BERT embeddings instead of standard learnable embeddings. This choice ensures that the decoder benefits from rich contextual representations of medical terminology, improving the semantic relevance and grammatical coherence of generated answers.

The Transformer decoder receives the fused multimodal representation as context and generates the answer token-by-token in an autoregressive manner.

4.3. Architecture

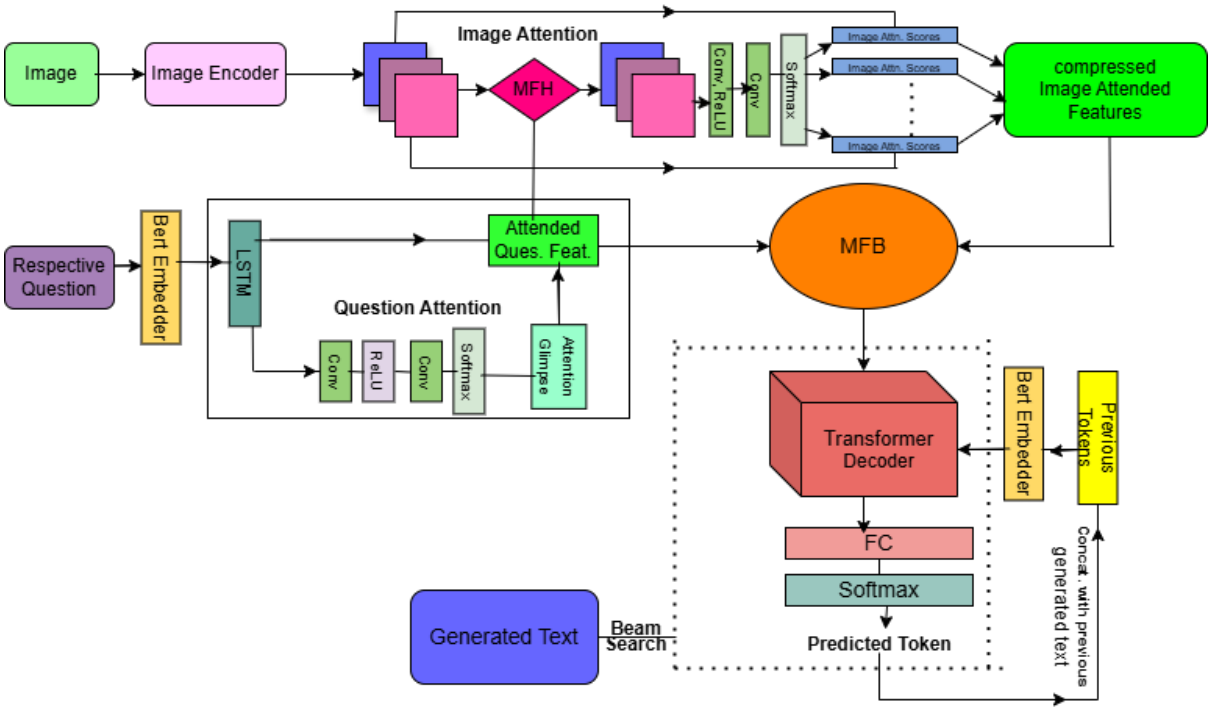


Figure 4.3.: Proposed Architecture of VQA with Transformer

4.4. Model Performance & Results:

In this section, we compare the results of our experiments with the baseline model on the original ImageCLEF-2019 Medical VQA dataset. The results presented below correspond to our proposed model, which incorporates a transformer-based decoder.

- **VGG16:** Compared to baseline [1] BLEU scores of 0.605 for BLEU-1,

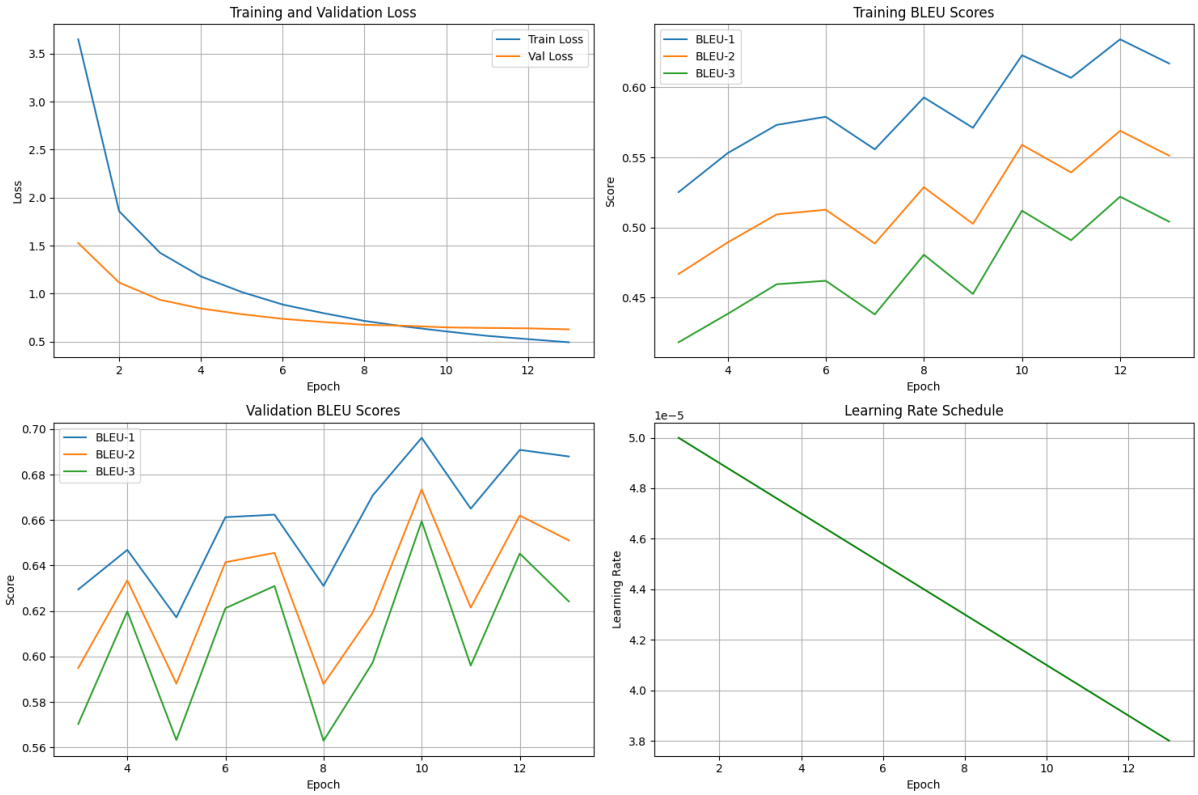


Figure 4.4.: Experimental results using VGG16

0.303 for BLEU-2, and 0.073 for BLEU-3, our model achieved significantly higher scores of **0.6908, 0.6619, & 0.6452** at epoch **12** respectively using VGG16, demonstrating substantial improvement in multi-level n-gram matching.

Description: The experimental results of VGG16 are shown in Fig. 4.4, which illustrates that the model continuously learns during training. Although minor fluctuations are observed, the BLEU score starts to decline after the 12^{th} epoch, triggering early stopping. The initial learning rate was set to 0.00005, using a linear scheduler.

- **ResNet-152:** Compared to baseline [1] BLEU scores of 0.605 for BLEU-1, 0.303 for BLEU-2, and 0.073 for BLEU-3, our model achieved

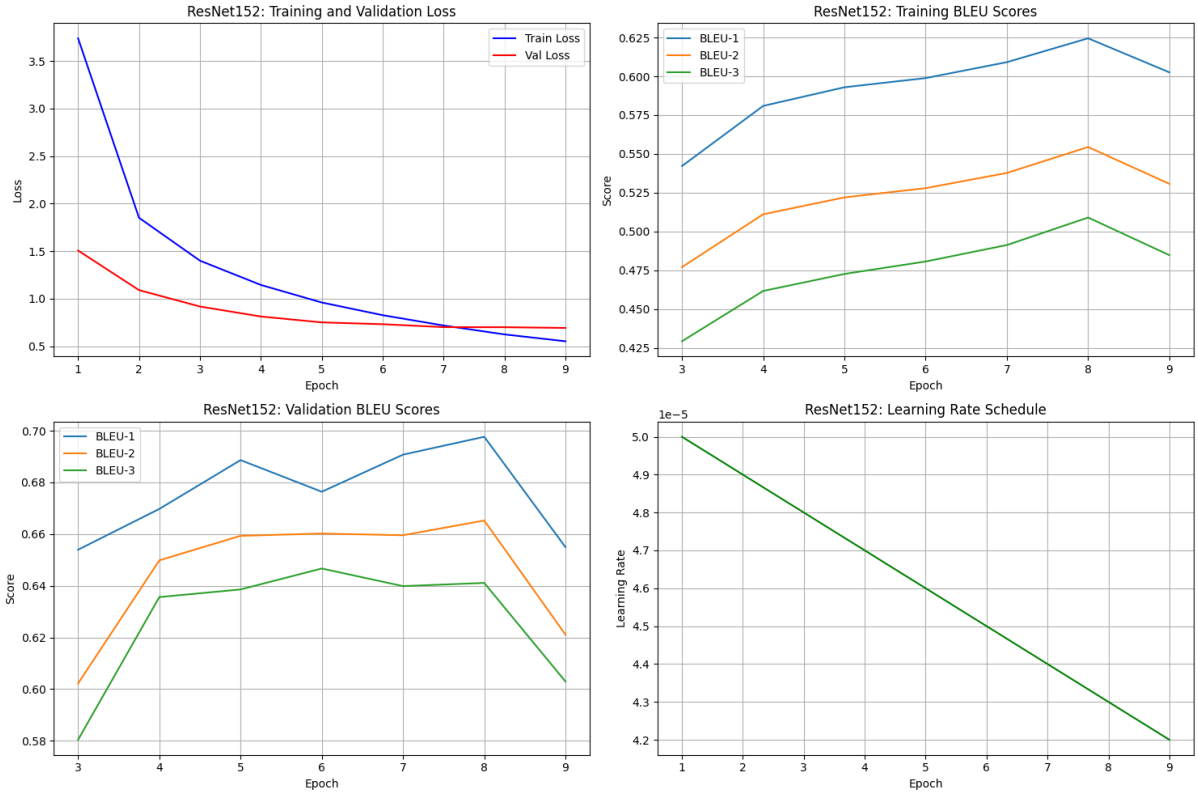


Figure 4.5.: Experimental results using ResNet-152

significantly higher scores of **0.6976, 0.6652, & 0.6411** at epoch **8** respectively using ResNet-152.

Description: The experimental results of ResNet-152 are shown in Fig. 4.5, which illustrates that the model continuously learns during training. Here, less fluctuations are observed than VGG16, and the BLEU score starts to decline after the 8^{th} epoch, triggering early stopping. The initial learning rate was set to 0.00005, using a linear scheduler. Also, we got slighlty better result compared to VGG16.

- **GoogLe-Net:** Compared to baseline [1] BLEU scores of 0.605 for BLEU-1, 0.303 for BLEU-2, and 0.073 for BLEU-3, our model achieved significantly higher scores of **0.6954, 0.6565, & 0.6318** at epoch **7** respectively using GoogLe-Net .

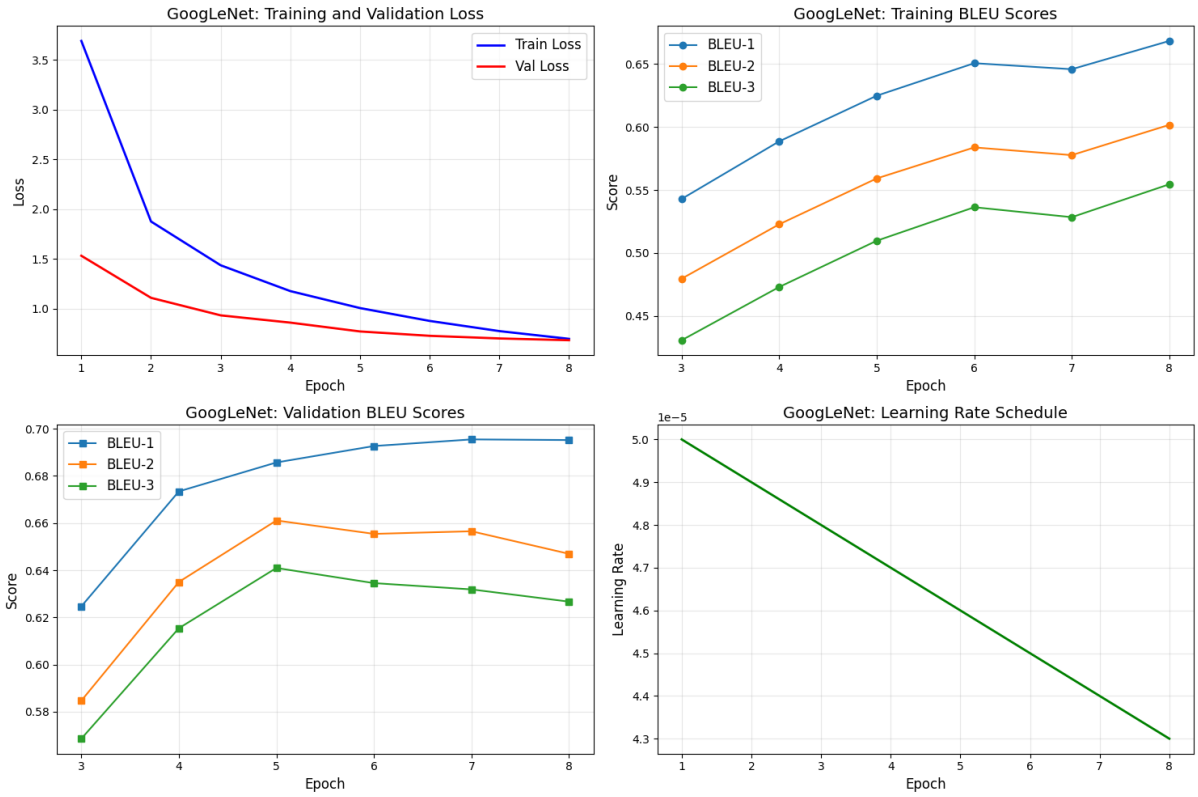


Figure 4.6.: Experimental results using GoogLe-Net

Description: The experimental results of GoogLe-Net are shown in Fig. 4.6, which illustrates that the model continuously learns during training. Here, very less fluctuations are observed than VGG16 & ResNet-152, and the BLEU score starts to decline after the 7^{th} epoch, triggering early stopping. The initial learning rate was set to 0.00005, using a linear scheduler. Also, we got better result compared to both VGG16 & ResNet-152.

- **MobileNetV2:** Compared to baseline [1] BLEU scores of 0.605 for BLEU-1, 0.303 for BLEU-2, and 0.073 for BLEU-3, our model achieved significantly higher scores of **0.6794, 0.6458, & 0.6282** at epoch **7** respectively using MobileNetV2.

Description: The experimental results of MobileNetV2 are shown in Fig. 4.7, which illustrates that the model continuously learns during training but the model has very poor performance on validation set than previous three models, and the BLEU score starts to decline after the 7^{th} epoch, triggering early stopping. The initial learning rate was set to 0.00005, using a linear scheduler.

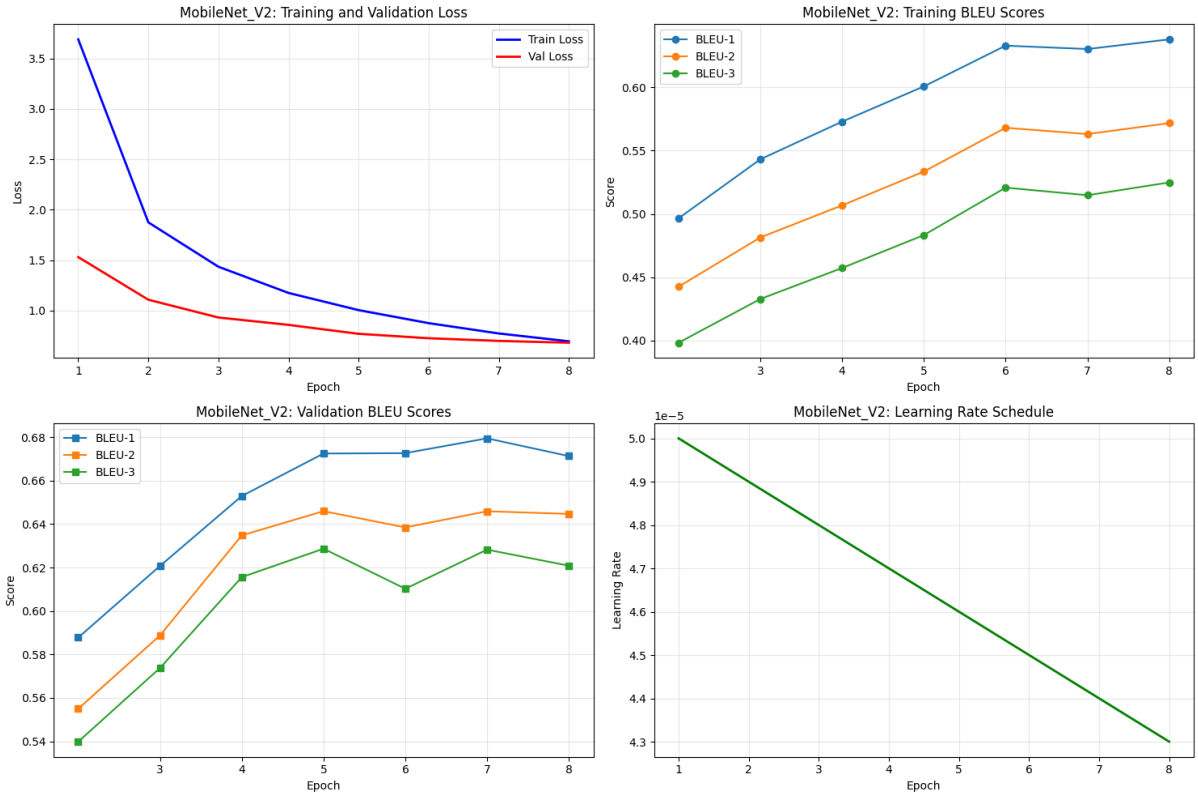


Figure 4.7.: Experimental results using MobileNetV2

Table 4.1.: Model Performance Comparison with Baseline

Model	**Epoch**	**BLEU-1**	**BLEU-2**	**BLEU-3**
Baseline [1]	-	0.605	0.303	0.073
VGG16	12	0.6908	0.6619	0.6452
ResNet-152	8	0.6976	0.6652	0.6411
GoogLeNet	7	0.6954	0.6565	0.6318
MobileNetV2	7	0.6794	0.6458	0.6282

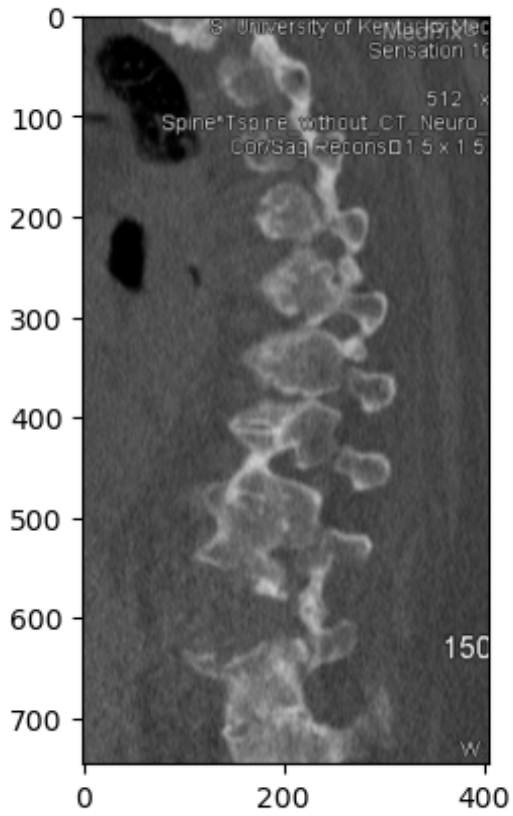
5. Adapting Vision-Language Large Models for Downstream Tasks

To adapt vision-language large models (VLLM) for the medical Visual Question Answering (VQA) task, we extended the original ImageCLEF dataset by transforming short length answers into descriptive sentence-level responses. For example, an original answer such as "non-contrast" was rephrased to a full descriptive sentence like "Yes, the displayed CT scan is a non-contrast image." This transformation aims to align the answer format with the generation capabilities of language models, enabling them to produce more natural and context-aware responses. We fine-tuned the Florence-2 base model, which consists of 0.23B parameters, on this enhanced dataset to enable the model to understand and generate descriptive medical answers effectively.

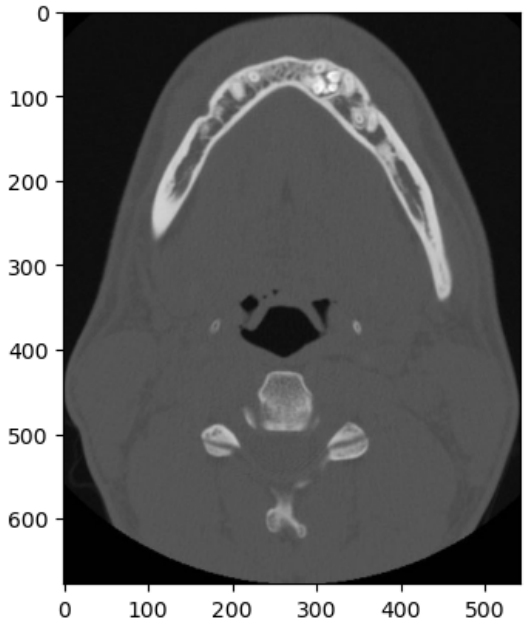
- **Florence-2:** is a general-purpose Vision-Language model developed by Microsoft, designed for broad cross-domain understanding and image-text tasks. Although the base Florence-2 model (0.23B) is not specifically trained on medical imaging data, it possesses strong foundational vision-language alignment capabilities. To better adapt it for medical VQA, we introduced a task-specific token, `<SpatialVQA>`, at the beginning of each question during fine-tuning. This specialized tag helps the model focus on the spatial reasoning and domain-specific nature of the medical queries. The model was then fine-tuned on our modified ImageCLEF dataset with descriptive answers, allowing it to learn both the structure and semantics required for generating medically relevant responses.

- **Results on ImageCLEF-2019(Descriptive Answers Format)**: the Florence-2 VLLM model was fine-tuned on ImageCLEF-2019(Descriptive Answers Format) dataset. The result of some sample data is showing in the Figure: 5.1

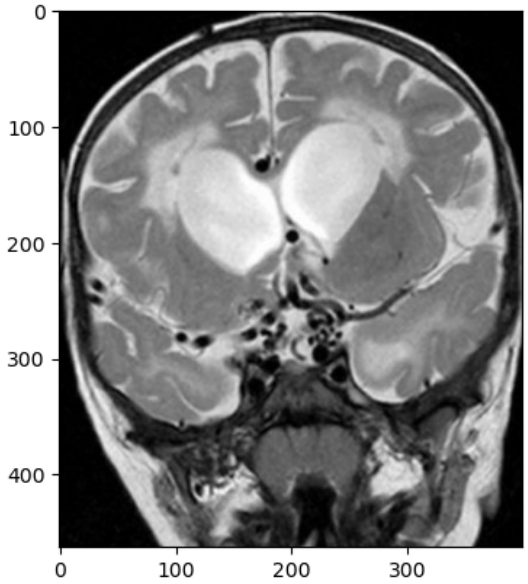
- **Results on Chest X-rays from MED-GRIT-270K**: Again, the Florence-2 VLLM model was fine-tuned on chest X-ray question-answer



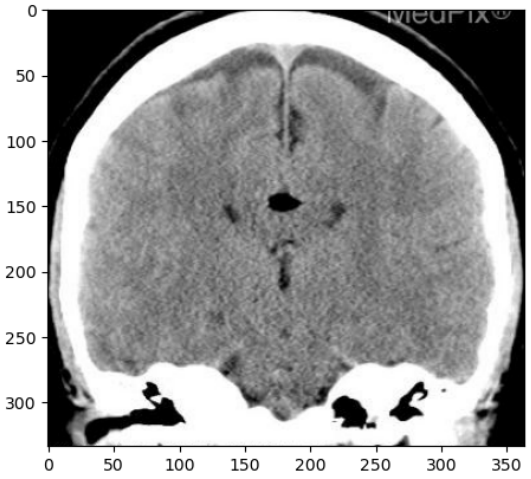
(a) **Question:** `<s><SpatialVQA> what part of the body is being imaged here?</s>`
Original: the part of the body being imaged here is the spine and its contents.
Generate: the part of the body being imaged here is the spine and its contents.



(b) **Question:** `<s><SpatialVQA> what part of the body is being imaged here?</s>`
Original:the part of the body being imaged here is the face, sinuses, and neck.
Generate: the part of the body being imaged here is the face, sinuses, and neck.

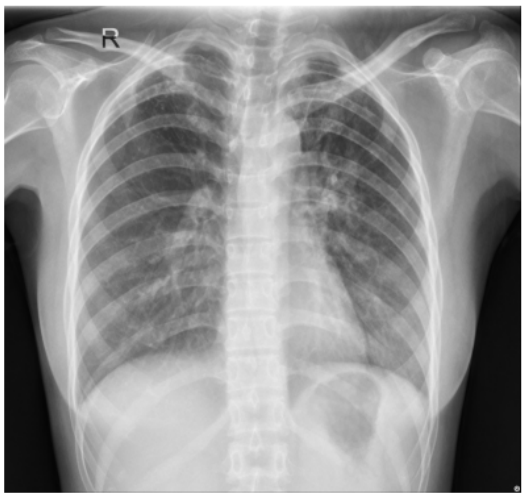


(c) **Question:** `<s><SpatialVQA> the mri shows what organ system?</s>`
Original:the mri shows the skull and its contents.
Generate: the mri shows the skull and its contents.

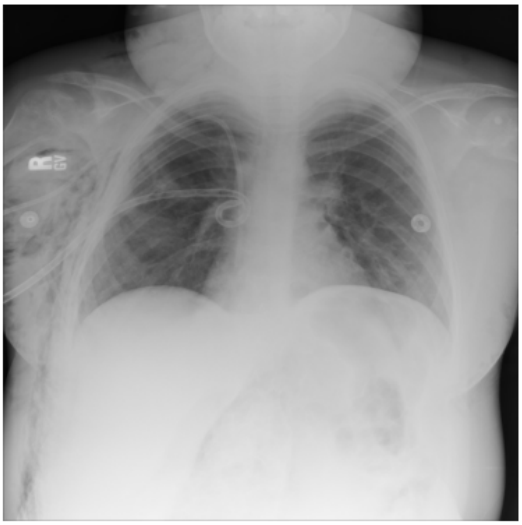


(d) **Question:** `<s><SpatialVQA> what is the primary abnormality in this image?`
Original:the primary abnormality is a corpus callosum lipoma.
Generate: the primary abnormality is glioblastoma multiforme.

Figure 5.1.: Examples of descriptive answer generation by the fine-tuned VLLM model on ImageCLEF-2019 dataset.



(a) **Question:** `<s><MVQA> what can be inferred from the presence of pulmonary tuberculosis in the lungs on the x-ray image?</s>`
Original: the presence of pulmonary tuberculosis in the lungs on the x-ray image indicates a bacterial infection that primarily affects the lungs. symptoms such as coughing, chest pain, and coughing up blood may be present. treatment typically involves a combination of antibiotics over an extended period to effectively manage the infection, with regular monitoring and follow-up with healthcare providers being essential for successful treatment outcomes.
Generate: the presence of pulmonary tuberculosis in the lungs on the x-ray image indicates a bacterial infection affecting the lungs. symptoms such as coughing, chest pain, and coughing up blood may be present. treatment typically involves a combination of antibiotics over an extended period to manage the infection effectively.



(b) **Question:** `<s><MVQA> what is in the region at <loc_301> <loc_164> <loc_492> <loc_248>?</s>`
Original: pneumothorax.
Generate: pneumothorax.

Figure 5.2.: Examples of descriptive answer generation by the fine-tuned VLLM model on chest X-ray from MED-GRIT-270K.

pairs from the MED-GRIT-270K dataset to create a specilized VLLM model on Chest X-ray Images. The result of some sample data is showing in the Figure 5.2

6. Conclusion

In this work, we addressed the complex task of medical Visual Question Answering (VQA) using a combination of classical and large-scale vision-language approaches. Initially, we constructed a baseline architecture employing VGG16 as the image encoder, BERT for question encoding, and MFB-based cross-modal fusion. This configuration demonstrated strong performance with BLEU-1, BLEU-2, and BLEU-3 scores of 0.6976, 0.6652, and 0.6411(using ResNet-152), respectively—significantly outperforming the baseline scores of 0.605, 0.303, and 0.073.

To further enhance the answer quality and better align with real-world clinical use-cases, we extended the ImageCLEF dataset by converting short length answers into descriptive, sentence-level responses. Leveraging this enhanced dataset, we fine-tuned a Florence-2 base Vision-Language Large Model (VLLM) of size 0.23B. Since Florence-2 is not pre-trained on medical data, we introduced a task-specific token `<SpatialVQA>` to better guide the model during fine-tuning. This allowed the model to adapt to the medical domain and generate context-aware, descriptive answers.

Again, we extracted the chest X-ray data from MED-GRIT-270K and fine-tuned Florence-2 to create a VLLM model specialized for the chest region of the human body. Overall, our results demonstrate the effectiveness of both classical deep learning methods and modern VLLMs for medical VQA tasks. The fine-tuning of a general-purpose VLLM on a domain-adapted dataset proved especially promising, offering a scalable pathway for more interpretable and clinically relevant VQA systems in the medical field.

7. Future Work

While the current work demonstrates promising results in adapting both classical and large-scale Vision-Language Models (VLLMs) for medical Visual Question Answering (VQA), several directions remain open for future exploration. First, the Florence-2 model, although fine-tuned on enhanced descriptive datasets and domain-specific tags like `<SpatialVQA>`, still lacks native medical domain knowledge. Future efforts can focus on pretraining or further domain-adaptive pretraining on large-scale medical image-text pairs to boost model specialization. Additionally, the current fine-tuning was focused solely on chest X-rays using MED-GRIT-270K; expanding this work to include other modalities such as MRIs or pathology slides could improve the model's generalizability. Exploring multi-turn question answering and reasoning capabilities within the same framework could also bring VQA closer to practical clinical decision support. Lastly, incorporating human-in-the-loop feedback or explainability techniques can further improve the reliability and trustworthiness of such models in real-world medical settings.

A. References

[1] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," IEEE International Conference on Computer Vision (ICCV), pp. 1839–1848, 2017.

[2] X. Huang, H. Huang, L. Shen, Y. Yang, F. Shang, J. Liu, and J. Liu, "A Refer-and-Ground Multimodal Large Language Model for Biomedicine," (MICCAI-2024), Computer Vision and Pattern Recognition.

[3] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. M¨uller, and M. Lungren, "Overview of ImageCLEF 2018 medical domain visual question answering task," in CLEF (Working Notes), 2018.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in IEEE ICCV, 2015.
[5] Z. Lin, D. Zhang, Q. Tac, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical Visual Question Answering: A Survey," *arXiv preprint*, arXiv:2111.10056, 2021.

[6] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," in *CLEF (Working Notes)*, 2019.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-training," 2018. [Online]. Available: `https://openai.com/research/language-unsupervised`

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] B. Boecking, "Biomedical Word Embeddings with Subword Information and MeSH," *Scientific Data*, vol. 6, no. 1, pp. 1–9, 2019.

[11] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Z. Liao, Q. Wu, C. Shen, A. Van Den Hengel, and J. Verjans, "AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-Based Sentence Mapping Approach for Medical Domain Visual Question Answering," in *CLEF 2020 Working Notes*, 2020.