FPT University

GRADUATION THESIS

Visual Question Answering for Medical Data Using a Visio-Linguistic Model

Tran Quang Duc

Tran Thi Kim Thanh

Le Viet Tien

Bachelor of Artificial Intelligence

Supervisor: Dr. Bui Van Hieu

Supervisor's signature

Major: Artificial Intelligence

University: FPT University

Hoa Lac campus, 12/2023

ACKNOWLEDGMENTS

We would like to thank our thesis supervisor, Dr. Bui Van Hieu, sincerely for his enduring support, long-lasting guidance and nurturing advisership, which has greatly illuminated and guided our most honored academic journey. His expertise and astute insights have undoubtedly an important role in influencing the direction of our intellectual conversation. We sincerely thank him for the many contributions he gave us with dedication.

Our heartfelt thanks go to FPT University for providing us with the tools, space and support we needed to complete our thesis. The University's commitment to education and research has truly inspired us to pursue our goals. As proud recipients of this distinguished academic institution, we are very proud to be part of its scientific heritage.

In addition, we would like to express our grateful thanks to our families and loyal friends, who have been the foundation of steadfast love, unrelenting assistance, and unfailing inspiration. Their unwavering belief in our ability and never-ending drive to succeed have irreversibly strengthened our will to achieve our goals. We are eternally grateful for their invaluable roles in our accomplishments and acknowledge that their contributions have been essential to our success.

In conclusion, everyone who has contributed to our academic journey feels a general sense of gratitude. Together, their contributions and roles have greatly enriched our academic journey and are invaluable. We sincerely thank everyone who has helped us along this journey and enabled us to reach this crucial point in our academic pursuits. I always appreciate everyone who has been a part of my academic career. Everything contributed has been an invaluable asset to our team and we will always be grateful for everything you have done. We appreciate all of your support at this crucial time in my life and for getting us to this point.

ABSTRACT

Recently, the research on Medical Visual Question Answering (Med-VQA) [1] is becoming significantly popular. Med-VQA intends to answer the question, given an image with vital clinic-relevant information, helps physicians in diagnosing diseases, giving patients better insights about illness. Med-VQA performs worse than general domain VQA due to a lack of accurate data such as the typical image as X-ray image. And another reason is proposed models are complicated in both image encoder and text encoder, which does not completely have outstanding performance. In order to deal with Med-VQA data limitation, recents studies primarily refine the fusion module which is responsible for synthesizing the question features and image features and provide models pre-trained by self-collection new dataset, overlooking the effect of question and image history.

In this thesis, we introduce a visio-linguistic model, the architecture employing an Associative Memory Module in the shape of separate storage of visual-linguistic individual experiences and their relationship to enhance context. Additionally, we introduce a Prototype Learning block to carry out stratified prototype learning on textual, visual embeddings utilizing morden Hopfield layers. Our model endeavors to acquire the most significant prototypes from the embeddings of texts and images with the augmentation of memory from associate memory modules. This is in contrast to directly acquiring concrete representations of joint features for different meanings in text and image. Then, by using these learned prototypes, more complex semantics can be represented for the answer. On VQA-RAD datasets, the proposed method accomplishes state-of-the-art performance with notable accuracy improvements of 0.45 %.

Keywords: Visio-linguistic, Computer Vision, Natural Language Processing, Prototype Learning, Visual Question Answering.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
OOD	Out-of-Distribution
RNNs	Recurrent Neural Networks
CA	Cross-Attention
SA	Self-Attention
SAM	Self-Attentive Memory
AMM	Associative Memory Module
CNN	Convolutional Neural Network
ViT	Vision Transformer
SOTA	state-of-the-art
GPT	Generative Pre-trained Transformer
KNN	K-nearest Neighbor
LVQ	Learning Vector Quantization
NMT	Neural Machine Translation
LLM	Large Language Model

1. Introduction

1.1. About Medical Visual Question Answering

Visual Question Answering (VQA) involves the integration of Computer Vision (CV) and Natural Language Processing (NLP) [1]. The purpose of the VQA system is to provide answers to questions related to a given image by analyzing its content. The recent investigation into medical Visual Question Answering (VQA) has exploded widely.

Medical Visual Question Answering (Med-VQA) is a task in machine learning wherein a medical image is presented alongside an associated query, and the objective is to furnish a precise response to said query. It is a harmonious fusion of the fascinating fields of CV, the art of visual data-understanding, the alluring world of NLP, the complex web of human language understanding, and the profound knowledge and expertise of the medical world. A Med-VQA integrated can potentially respond to doctor's requests, reduce the pressure on the healthcare system and increase the effectiveness of medical staff (Example Figure 1.1). Another application that parallels the benefits of Med-VQA is to be executed as a clinician or a researcher who examines body tissue and provides help to other healthcare systems to make a diagnosis [2]. The VQA medical system can act as a knowledgeable assistant alongside the responsibilities of medical professionals. The utilization of a VQA system as a secondary diagnostic tool presents the potential to mitigate the risk of misdiagnoses by offering an additional perspective that corroborates the physician's interpretation of medical images [3].

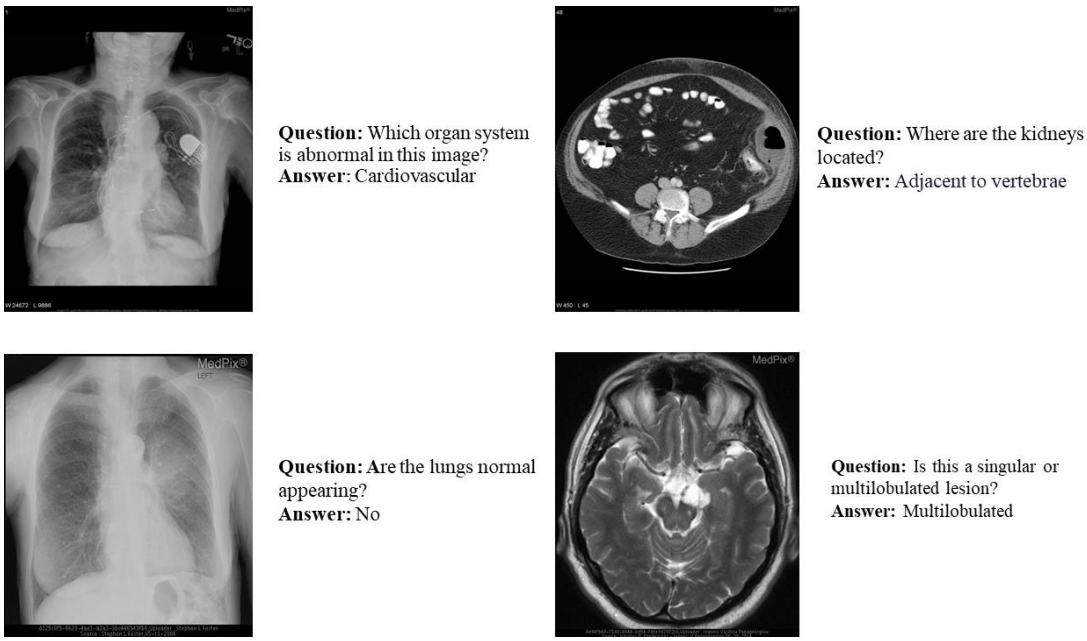

Figure 1.1: Med-VQA examples images, questions, and answers.

In the end, a sophisticated and comprehensive Med-VQA system can immediately evaluate patient images and provide a variety of answers. When a healthcare expert is not available, a Med-VQA system can grant comparable guidance in certain circumstances,

such as fully automated health checks. After visiting the hospital, patients search online for more information. Misleading information and misleading data can produce inappropriate results. To provide trustworthy answers anywhere, anytime, a Med-VQA can also be applied into an online examination platform.

Medical VQA has higher technical requirements than general area VQA for the following reasons. First, the high demand for expertise makes expert annotations expensive, and Question-Answer (QA) pairs cannot be artificially created from images, making it difficult to create a standard, large-scale Med-VQA dataset. In addition, answering queries around medical image's information requires a specific VQA model design. A lesion is microscopic, so the question also needs to be focused, comprehended on a fine-grained scale. Therefore, segmentation approaches may be asked to accurately identify the area of interest. Finally, a question could be extremely typical and require the model to be trained on a completely medical knowledge.

1.2. Background and challenges in the study

Recent advancements in Artificial Intelligence have unlocked novel avenues for clinical decision support. Notably, solutions focused on the automated interpretation of medical images have garnered significant attention because of capability applications in both image retrieval and aided analysis. Further, the Med-VQA system can comprehend medical images and answering questions related to their meaning hold promise for bolstering clinical training, decision-making, and patient-about learning. From a computational standpoint, this Med-VQA task presents a compelling challenge that necessitates the seamless integration of NLP and CV techniques. These approaches have demonstrated significant effectiveness in this particular specialty area [2]. Text-based queries and images that provide precise answers through the use of medical visual content have gained importance in recent developments. Models such as PMC-VQA [4] and MUMC [5] as shown in Figure 1.2 are prime examples of these models that exhibit remarkable performance and achieve innovative results rather than state-of-the-art results at the moment.

Medical Visual Question Answering (VQA) constitutes a distinct field within VQA, focusing on generating responses to inquiries posed in natural language regarding medical images. Although Med-VQA promises exciting advancements in healthcare, its application faces several challenges. The challenge of question diversity, interpretability, complementary medical data, large language models, generalizability, and integration into the medical pipeline represent six key obstacles that arise from the general medical requirements for developing robust and effective applications [2]. However, a prominent obstacle in medical VQA lies in the scarcity of large-scale labeled training data, which typically involves substantial costs for collection and construction. Thus, medical multimodal data diversity poses a challenge for VQA models to achieve Out-of-Distribution (OOD) generalization. Recent studies have indicated a risk of OOD

generalization in medical VQA, as shown in Figure 1.3, the models may answer based on questions and omit input images and due to correlations between the question and answer distribution.

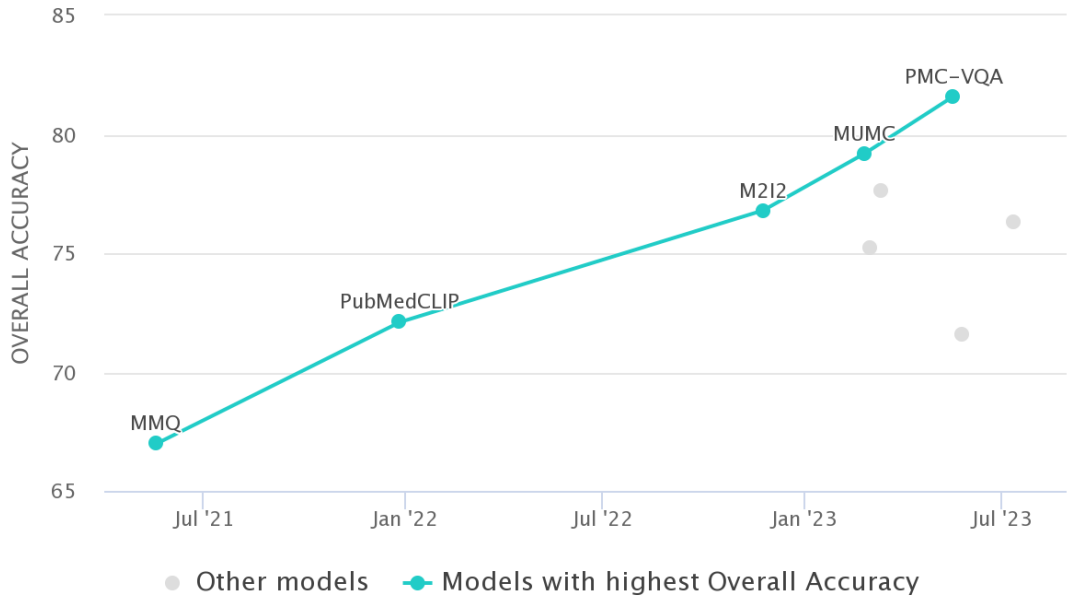

Figure 1.2: The graph of some approaches results on the VQA-RAD benchmark [6]

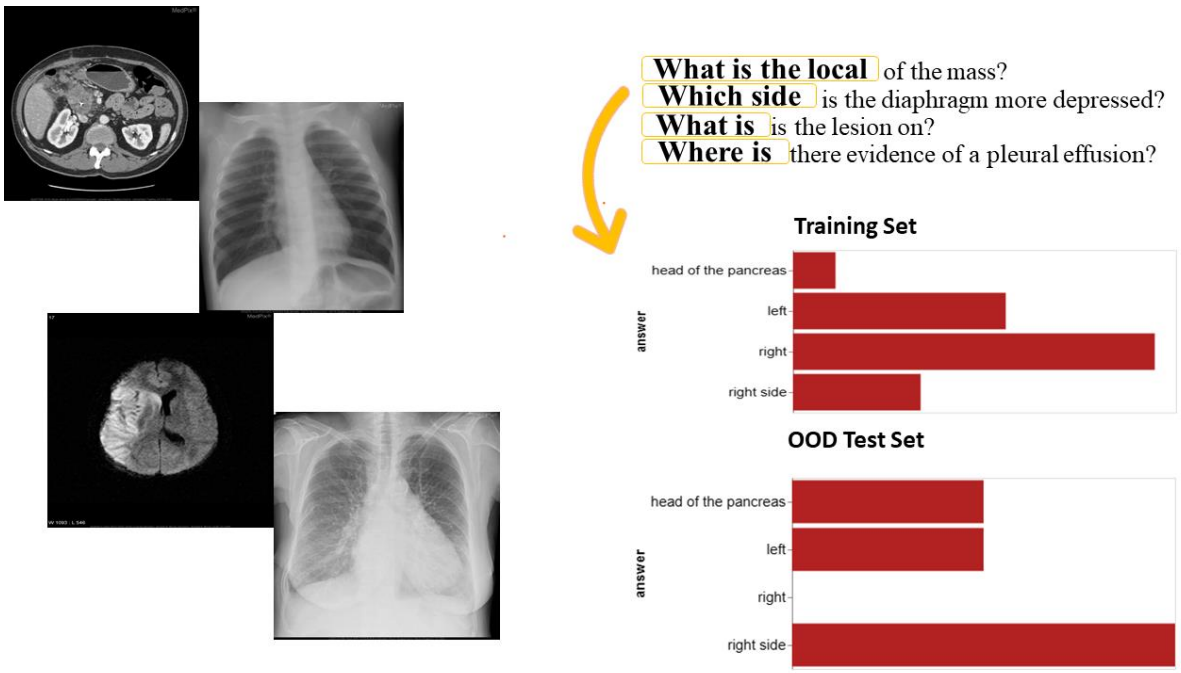

Figure 1.3: In the VQA-RAD dataset, the disparity in the allocation of responses in relation to a question prefix varies between the training and testing phases. This can be attributed to the flawed methodology in selecting models using the OOD test set.

1.3. Our study objectives

The research object of the thesis “Visual Question Answering for Medical Data Using a Visio-Linguistic Model” is to explore algorithms to increase the efficiency of applying Med-VQA in reality and identify limitations of existing methods and solutions to improve approaches using multiple technical resolutions.

We conduct an in-depth study of object-based memory models using recurrent neural networks (RNNs) to achieve the main research goal (Hopfield, 1982) [7], independent of current methods. This research aims to recognize the inherent conditions of existing approaches and describe critical problems needing to be surpassed in the current state. Based on these experiments with previous methods, our proposed architect model will take advantage of various resolution and diverse scale methods to overcome the mentioned limitations, giving better performance-predicted output, while ensuring speed and resources in deployment.

The focus of the research goals is the development and creation of a novel architectural model. This architecture is intended to provide an advantage: the model can store the features and structures in its relational and element stores and use the rules it learned during training to reason across the two stores. Our architect aims to overcome the limitations of current methods by utilizing various techniques simultaneously, particularly in question open-up and answer in Med-VQA. Moreover, this architecture is optimized for computational efficiency, guaranteeing swift and resource-utilizing deployment in its operation.

With the model's conception, the study purpose is evaluating the ability of our model on other datasets.

1.4. Our study contributions

Our present three contributions from this study:

1. Presents the Associative Memory Module (AMM) approach to Med-VQA. The model was created as a two-storage model with relational storage and separate element storage. Because the two memories are separate, they must interact to enhance each other's representation. And use Hopfield layers to perform prototype learning from enriched visual-linguistic features to archives generalization.
2. As far as we can tell, our model represents the initial effort that uses Hopfield layers, further facilitating research on memorizing and reasoning in Med-VQA.
3. We achieved the new SOTA result for the VQA-RAD.

Our research introduces a promising model in VQA-RAD benchmark and other benchmarks in Med-VQA.

1.5. Thesis Structure

The thesis introduces an approach conducted experimentally that performs well compared to existing approaches.

Chapter 2, entitled "Theoretical Basis," is a section generalizing research in medical VQA tasks. This section shows the literature review of the research problem. This section also provides scientific evidence about what we are based on.

Chapter 3, entitled "Methodology," is a major section that represents most of our works. This chapter shows from the very beginning baseline model to the final model architecture after experiments. This chapter particularly describes the reason how the memory and prototype strategy can be applied to increase the accuracy of the Med-VQA task.

Chapter 4, entitled "Results and discussion," presents our competitive results and experiences we earned through the study. In this chapter, we will show our experiment for training, evaluation, and fine-tuning the model hyper-parameters. Besides, we compare each impact of the module to performance and resources of the model.

Chapter 5, entitled "Conclusion" is the final section summarizing our works, also characterizing our next steps with this study. Overall, as a journey, our study gets excellent performance despite lack of quantity in data, which leads to failure of generalizing dataset.

2. Theoretical Basis

In this section, we will show groundwork around several recent years to comprehend deeper the trend of research. Then, we will present the details of the foundation theories being applied in our approach.

2.1. Related work

2.1.1. History of Med-VQA methods

The comprehensive framework consists of three or four key components, contingent upon specific task requirements: A language encoder, a visual encoder, a fusion algorithm, and an answering unit. The visual encoder and language encoder have crucial roles in extracting visual features and contextual information from text, respectively. The fusion algorithm, in some instances, may be integrated into the final layers of language encoder [9, 10, 11], operating on the same principles. The answering component is realized as either a text classifier layer, employing a Softmax layer, or a text generator. The visual encoder encompasses versatile convolutional neural network (CNN) architectures, including but not limited to VGGNet [12], ResNet [13], InceptionNet, DenseNet, and EfficientNet [14], or Transformer-based backbones. Concurrently, the language encoder mirrors equivalent language models, such as Transformer [15] or any

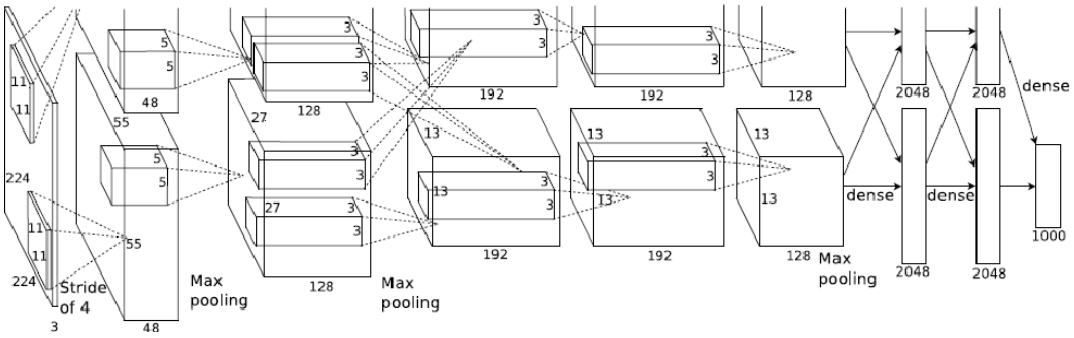

Figure 2.1: An illustration of the architecture of the CNN [12]

variation of BERT [16]. Pre-trained weights typically initialize these visual and language encoding models during the training stage of the Med-VQA task.

a. Visual encoder

Concerning the visual encoder in prevalent challenges, VGGNet [12], ResNet [13] in Figure 2.1, and the later Transformer-based Vision Transformer (ViT) are widely utilized. However, in mainstream applications, researchers often employ pre-trained models from ImageNet [17] on VGGNet and ResNet. While ImageNet pre-trained weights may not be directly applicable, they present a promising alternative in scenarios where medical datasets have limited image data.

The exploration of improved pre-trained models is a prominent focus for Med-VQA research and the broader medical AI community, rather than proposing entirely novel methodologies. Several emerging publicly available pre-trained weights on official medical datasets [9] have been introduced. Given the limited number of images in most medical VQA datasets (e.g., VQA-RAD [18] with 315 images, SLAKE-English [19] with 642 images, and OVQA [20] with 2001 images), strategies have been devised to address this constraint. These approaches involve utilizing pre-trained models on additional data, pre-training on more comprehensive datasets, and implementing contrastive learning. For instance, Xiaoman Zhang et al. employed a pre-trained model visual encoder from the PMC-VQA dataset [9]. Concurrently, researchers explore potential enhancements within the original dataset.

b. Language encoder

The language encoders mostly include LSTM [22], Bi-LSTM [23], GRU [24], BERT [16] and BioBERT [25]. Respectively, BERT methods gave higher results than others. Another research marks words to enhance the correlation between words in the same context [26]. However, NLP research has received less attention. Researchers develop more pre-training results on larger corpus and other typical textual documents like GPT-like LLMs or train a part of transformer BERT as Figure 2.2 for a variety of reasons.

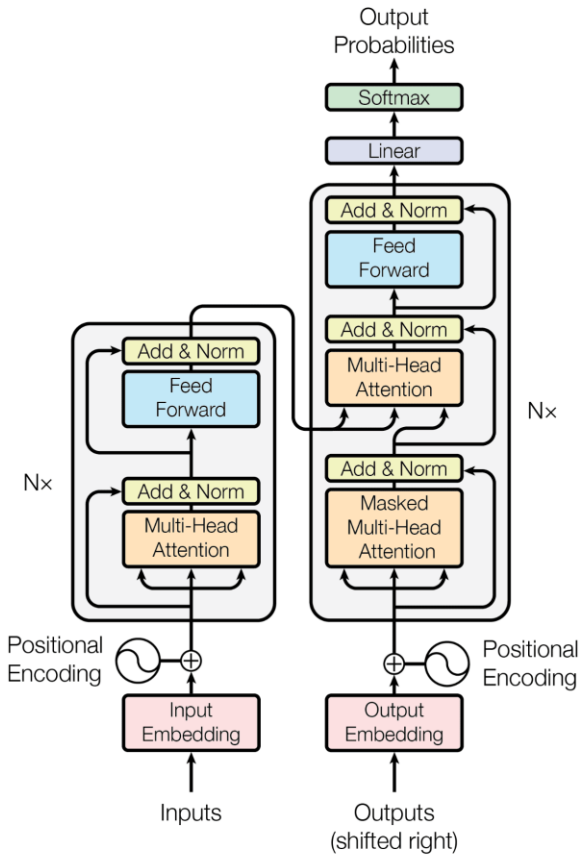

Figure 2.2: The Transformer - model architecture [15]

c. Fusion algorithm

The fusion algorithm merges the outputs from visual encoder and language encoder into hidden vectors representing both aspects. While it can be integrated into the language encoder, its effective implementation is essential for optimal results. Two primary fusion algorithms are employed: the attention mechanism and the pooling module.

The pooling module stands out as a significant technique for combining visual and language features. Common applications involve concatenation, summation, and element-wise product [27]. However, concatenation yields average performance [28], and the element-wise product may become computationally expensive when input vectors are large in any dimension [29].

d. Answer components

There are two choices for conducting output: classification mode and text generation mode. The classification mode has advantages on small, short answers. In contrast, text generation mode can give more accurate long answers. Anyway, It will be more difficult as the answers get longer and more complex in context, description.

2.1.2. Prior studies inspired our research

a. Prototype learning

Prototype learning is a method that is frequently applied in pattern recognition and computer vision applications with the goal of choosing or optimizing the most representative data points or anchors from the training set [30]. The K-Nearest Neighbors KNN technique is known as a popular Prototype learning approach [31], in which prototypes are selected by selecting the neighbors with the closest Euclidean distance from all training samples. The Learning Vector Quantization (LVQ) method is recommended [32] as manually picking prototypes from the entire training set and requires substantial memory. Prototype learning was improved in the classifier by considering the class boundaries of challenging categories with LVQ, allowing for classification of testing samples with a limited number of prototypes without examining whole training examples. As the era of deep learning approaches, research has shifted towards utilizing well-designed neural networks for autonomous prototype learning [33]. In [34], Yang et al. present an image recognition approach with convolutional prototype learning, simultaneously optimizing prototypes across various categories and a convolutional neural network (CNN). They find that incorporating prototype matching in decision-making enhances classification resilience. Dong et al. integrate the prototype learning technique semantic segmentation task by employing a sub-network to derive prototypes from the supporting set [35]. This technique shows success in applications such as action recognition [36] and face recognition [37].

In contrast to these methods, our study proposes prototype learning with morden Hopfield network [38] from text-image embeddings and enriched vision-language context. This approach simplifies the process by initially learning the most representative prototypes, aiding in the representation of more intricate semantics. Directly acquiring incredibly varied representations of combined text-image features for multiple semantics poses a notable challenge.

b. Memory Network

Memory networks have grown in popularity in language processing ever since Weston et al. [39] introduced a memory component to store simple event for the question answering problem. In general, input, scoring, attention, and response components make up memory networks. Unlike previous networks [39], Sukhbaatar et al. train memory networks in [40], eliminating the need to label supporting facts during the training phase. Building memory networks using neural sequence models with attention, Kumar et al. do so in [41]. A neuronal attention mechanism enables memory networks to focus on particular inputs when asked a question. This helps with many different computer vision and language processing issues, including machine translation [42, 43], picture classification [44], and image captioning [45]. The latest progress in neural architectures

that integrate memory or attention mechanisms includes developments such as neural stack-augmented RNNs [47], Turing machines [46] and hierarchical memory networks. [19]. In the term of Med-VQA, Pellegrini et al. [48] introduced a hierarchical arrangement of annotations through structured reports for X-Ray images. They subsequently incorporated context by referencing previously posed questions and their corresponding answers to predict responses. Recognizing the significant capabilities of memory networks in VQA [49, 50], we suggest employing an associative memory network to enhance the integration of vision-language context. For implementation, we use Modern Hopfield Network [38] to control learning and retrieval from memory. Our memory module thus notably differs from the existing approach.

2.2. Base theories

2.2.1. Attention mechanism

Self-attention, or intra-attention, serves as an attention mechanism that connects between different spots within a sequence to generate a representation of the entire sequence. It has demonstrated success across various tasks. For example: reading, summarizing, enriching [51, 52].

However, the Transformer built entirely on self-attention, released in 2017 [15]. The Transformer exhibited remarkable performance on machine translation tasks and English constituency parsing. Notably, this architecture stands out for its absence of recurrence or convolution.

$$Attention\ (Q, K, V)\ =\ softmax\ (\frac{Q\ K^T}{\sqrt{d_k}})\ V \quad (1)$$

2.2.2. EfficientNet architecture

EfficientNet constitutes a convolutional neural network as shown in Figure 2.3 founded on the principle of "compound scaling," aiming to address the enduring trade-off among model size (width, depth, resolution), accuracy, and computational efficiency. The core concept of compound scaling involves simultaneously scaling three crucial dimensions of a neural network: width, depth, and resolution.

Compound scaling

The procedure commences with a foundational model, acting as the initial reference point. Typically, this baseline model constitutes a neural network of moderate size that demonstrates proficiency in a given task but may lack optimization for computational efficiency.

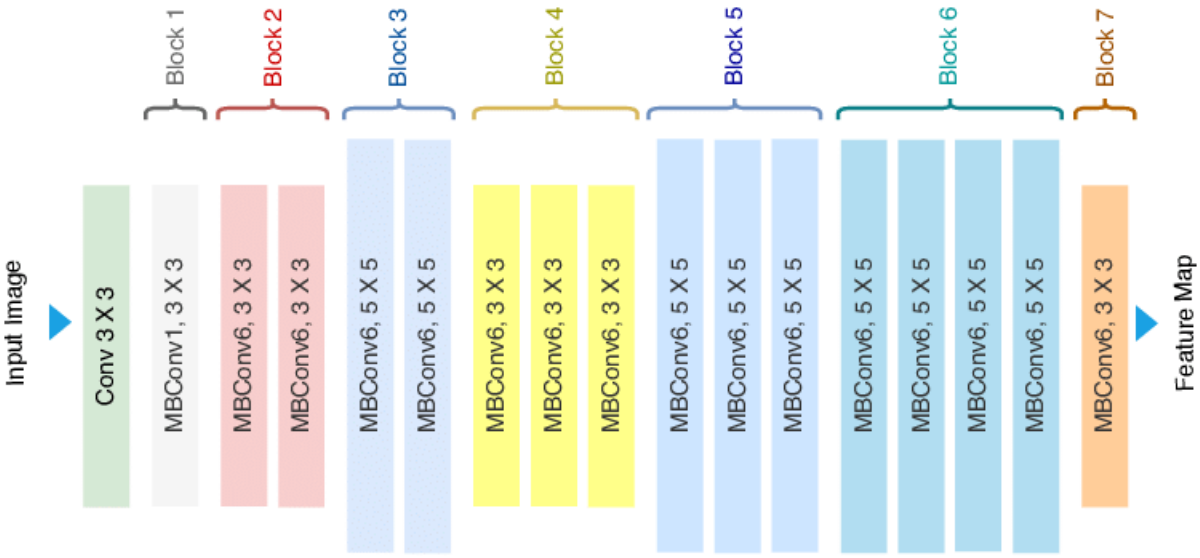

Figure 2.3: Architecture of EfficientNet-B0 with MBConv as Basic building blocks [13]

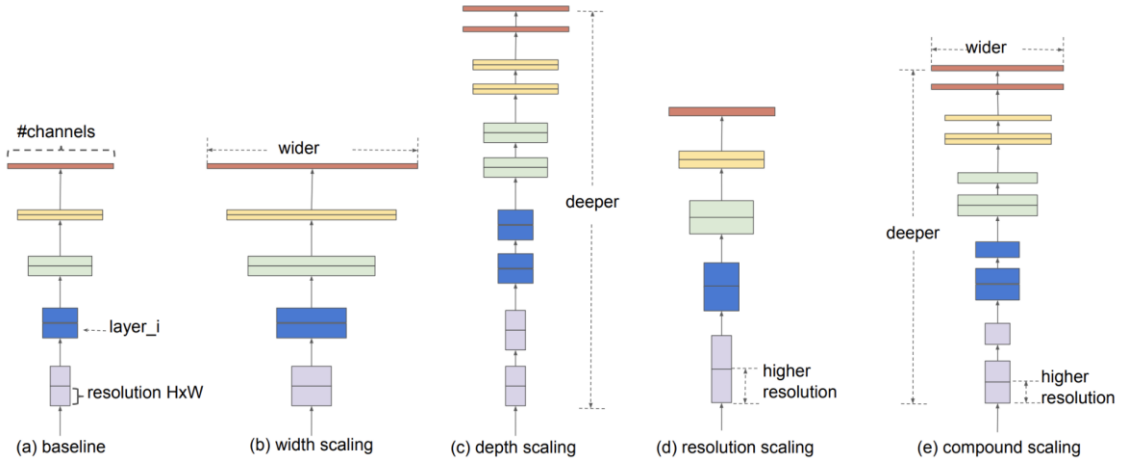

Figure 2.4: width, depth, resolution, compound scaling [14]

Subsequently, a compound coefficient is introduced as a user-defined parameter, determining the extent to which the neural network's dimensions should be scaled. This coefficient, represented by a single scalar value, uniformly adjusts the width, depth, and resolution of the model as Figure 2.4. By manipulating the φ value, one can regulate the overall complexity and resource demands of the model.

Through the adoption of the compound scaling methodology, EfficientNet adeptly navigates through an extensive array of model architectures, achieving an optimal equilibrium between accuracy and resource utilization. This notable capability for efficient scaling positions EfficientNet as a transformative force in the realm of deep learning. It facilitates SOTA performance across diverse CV tasks while maintaining adaptability to a spectrum of hardware constraints.

2.2.3. RoBERTa architecture

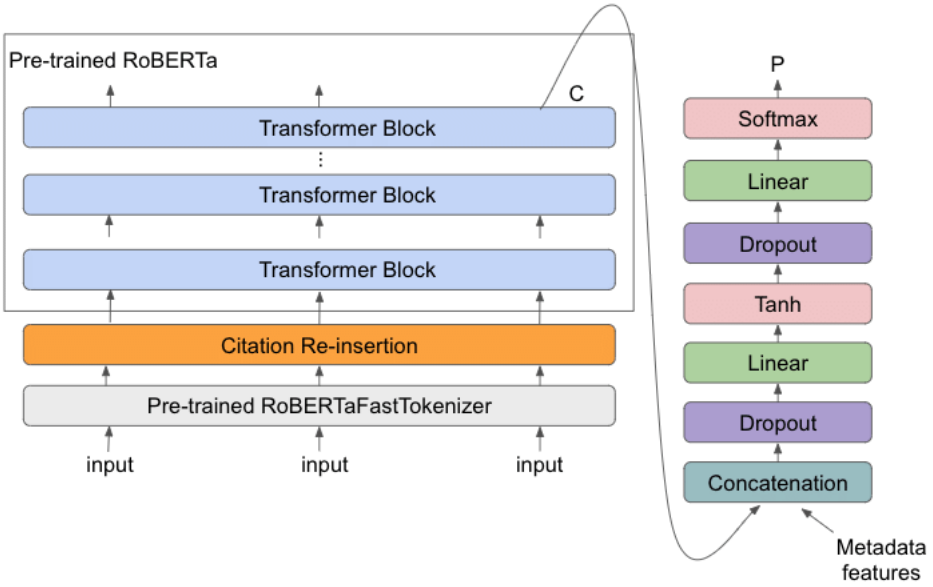

Figure 2.5: The RoBERTa model architecture [53]

RoBERTa, short for A Robustly Optimized BERT Pretraining Approach, introduces pivotal modifications to enhance performance and robustness, resulting in significant advancements across various natural language processing (NLP) tasks as shown in Figure 2.5. Key modifications encompass:

Increased training data: RoBERTa surpasses BERT by utilizing a significantly larger pre-training dataset, integrating 160GB of text data compared to BERT's 16GB. This expanded dataset enables RoBERTa to acquire a more comprehensive and nuanced comprehension of language.

Dynamic masking strategy: Departing from BERT's static masking strategy, RoBERTa adopts a dynamic masking approach that alters the masking of different tokens at each training step. This dynamic strategy encourages the model to prioritize learning contextual relationships between words, enhancing performance in downstream tasks.

Longer training schedule: RoBERTa undergoes an extensively prolonged pre-training period compared to BERT, involving 10 times the number of training steps. This extended duration allows the model to refine its internal representations, leading to heightened accuracy.

3. Methodology

3.1. Data augmentation and preprocessing

3.1.1. Data augmentation

Data augmentation is techniques to increase, enrich datasets in size, detail by creating new data from existing data. It is providing more data for training, therefore, results improve. Some current techniques produce negative examples by picking images

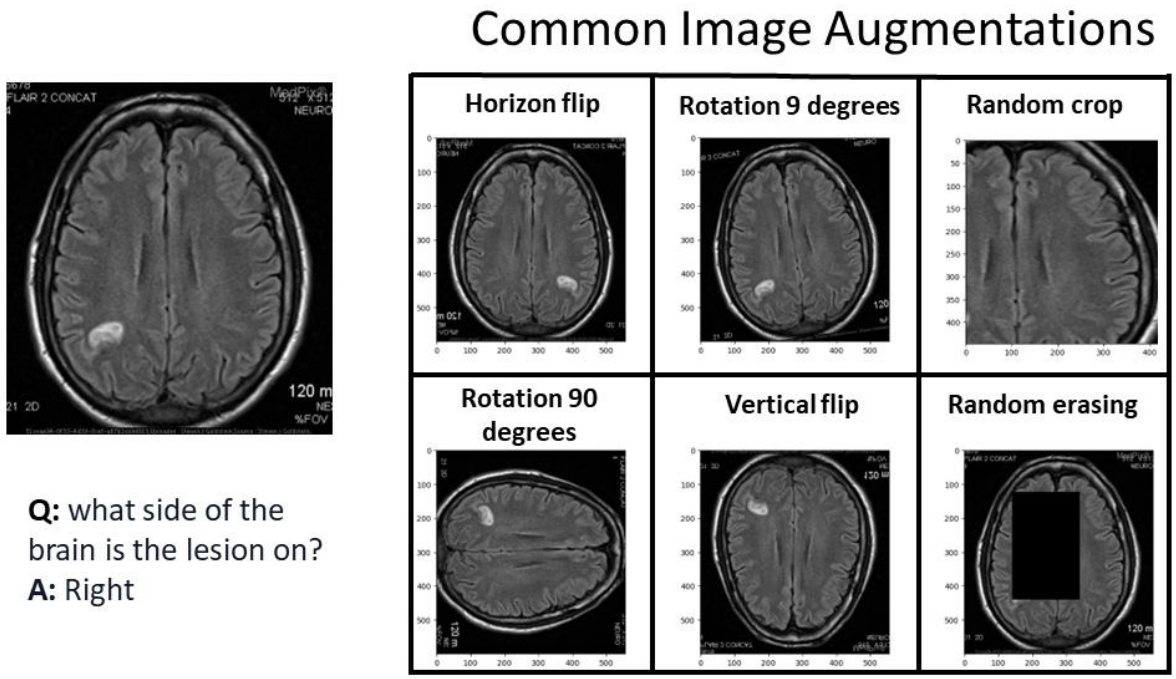

Figure 3.1: Common image augmentations apply on image of VQA-RAD

or questions at random [54, 55], or create new positive training samples by composing reasonable image-question (VQ) pairs. To create news VQ pairs, they rely on pre-defined rules to generate answers that are tailored for specific question types. However, these data augmentation methods almost always experience a significant drop in performance when tested on data from the same domain as the training data [56, 57] or the answer assignment techniques of these methods based on human annotations and lack generality [58]. Therefore, we use image transform in lieu of generating VQ pairs for data augmentation.

Image augmentation by spatial transformations may be useful in many computer vision applications. However, Elgendi el al (2021) [59] shows that it can negatively impact model performance in medical images. Especially on radiology images, for instance, consider the question and image in Figure 3.1. The bright area on the left part of the MRI (in the MRI process the left side of the image is corresponding to the right side of the brain) image is the brain lesion. As can be seen, the bright area is moved to the wrong place or disappears in horizon flip, Rotation 90 degrees, random crop and random erasing. Additionally, horizon and vertical flip would lead to non-physiology images. Thus, image augmentation without a clinical consideration may cause noise in the answering question.

There are several ways to perform image augmentation by color modification. For instance, An RGB image is encoded as a 3-dimensional array, where each dimension represents a color channel (red or green or blue) and contains intensity values for that color. The colors of the image are like layers on top of its structure. We can remove or rearrange these layers without changing the basic shape of the image. Additionally, to augment an image, we can modify the hue, saturation, and value components of the image

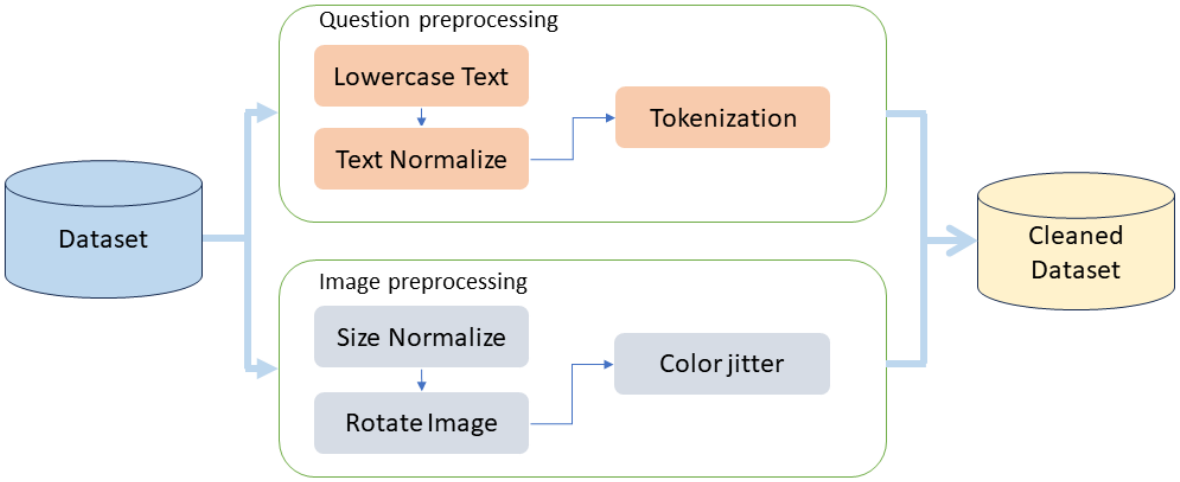

Figure 3.2: The overview of medical VQA data preprocessing process

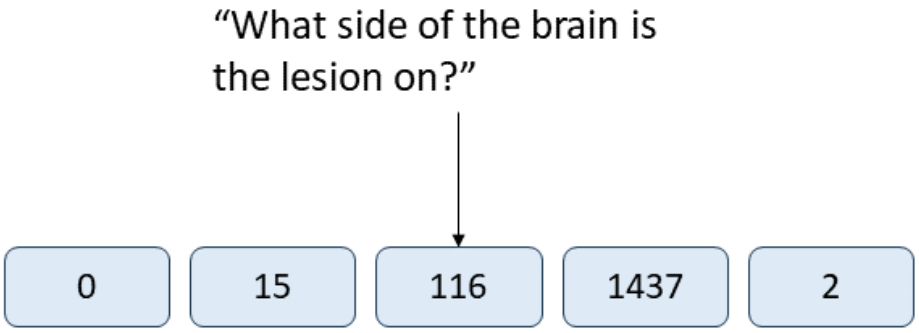

Figure 3.3: Illustration of tokenization

by isolating a specific color channel (such as blue, red, or green), and converting color spaces into one another. However, converting a color image should be performed carefully in medical VQA because one of the question categories is color.

In the context of this study, we use random rotation in range -9 to 9 degrees and color jitter which randomly adjust the brightness, contrast and hue of an image to augment the diversity of the Med-VQA dataset. Overall, the data augmentation improves generalizability and robustness of the model.

3.1.2. Data preprocessing for VQA-RAD dataset

For the purposes of training and evaluating our model, our study made use of the Visual Question Answering in Radiology Dataset, as introduced in Lau et al. [60]. However, It should be stressed that the questions in VQA-RAD can be answered with short answers. Consequently, the classification mode for output will have an advantage [2] with the dataset. In the preprocessing process, the answer will convert to a label.

Data preprocessing is crucial for most machine learning projects. By cleaning the data, we extract more information for training, leading to better results. Figure 3.2 shows how we prepare images and questions. In the question pre-processing process, the question will be apply the sequential techniques follow: (1) Converting texts into lowercase texts, (2) removing special characters and extra spaces, (3) Use Byte-Level Byte-Pair Encoding [61] pre-trained from [63] (See Fig 3.3 for more illustration of tokenization). The image pre-processing based on Section 3.1.1, we resize the image with size 488 x 488 x 3. Then, we apply data augmentation for the image. After data cleaning, the VQ pairs will be used to train and evaluate our medical VQA model.

3.2. Architecture overview

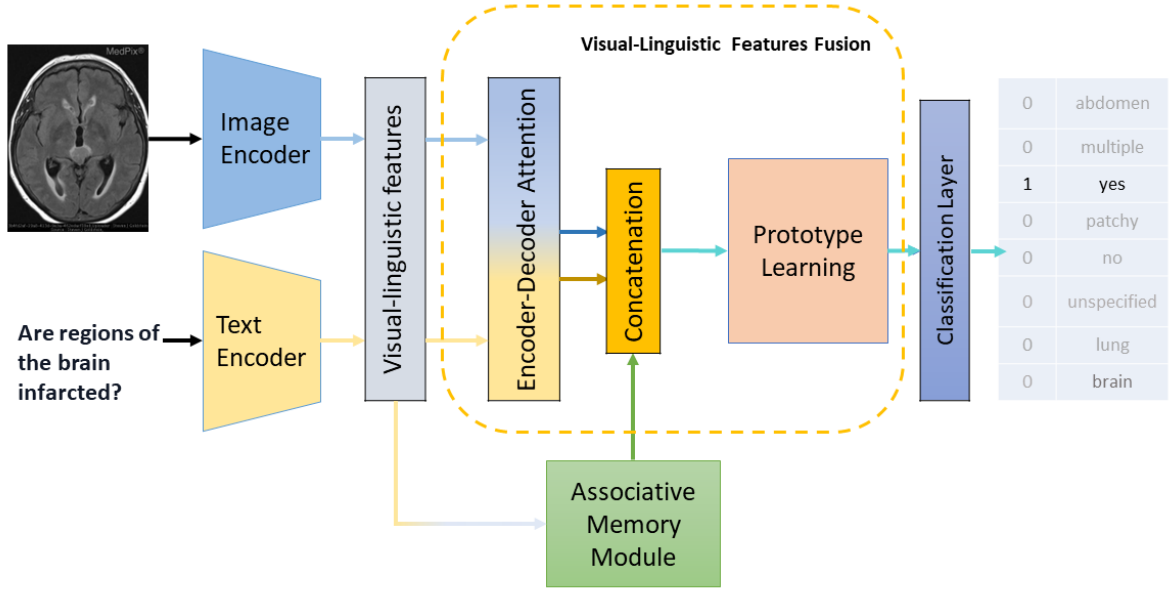

Figure 3.4: Overview of our model

In this section, we present the overview of our model and the pipeline to training the model. Illustrated in Figure 3.4, the model in a framework of four components: Input encoder, associative memory, visual-linguistic features fusion and the answering module. The training procedure encompasses the subsequent steps: First, our model encodes the input question and a related image. In particular, we use the spatial layout information-preserving outputs of the final pooling layer of EfficientNet [14] as visual features. To extract features, we feed it into RoBERTa [53]. The visual and linguistic features are concatenated then fed into an associative memory module, which consists of relational memory and item memory, to store and calculate relation between visual and linguistic features. Next, we use an encoder-decoder attention mechanism which can choose the greatest pertinent visual regions and linguistics from images and questions. Each linguistic feature and visual feature vector's weights are calculated by the encoder-decoder attention mechanism as visual-linguistic features. Output of AMM will concatenate to the result of

encoder-decoder attention to enrich visual-linguistic features. We use prototype blocks to conduct hierarchical prototype learning on visual-linguistic features with modern Hopfield layers. Finally, the output of Prototype Block will be used to classify answers.

The remainder of this chapter we will present each component of our model in detail.

3.3. Image features extraction

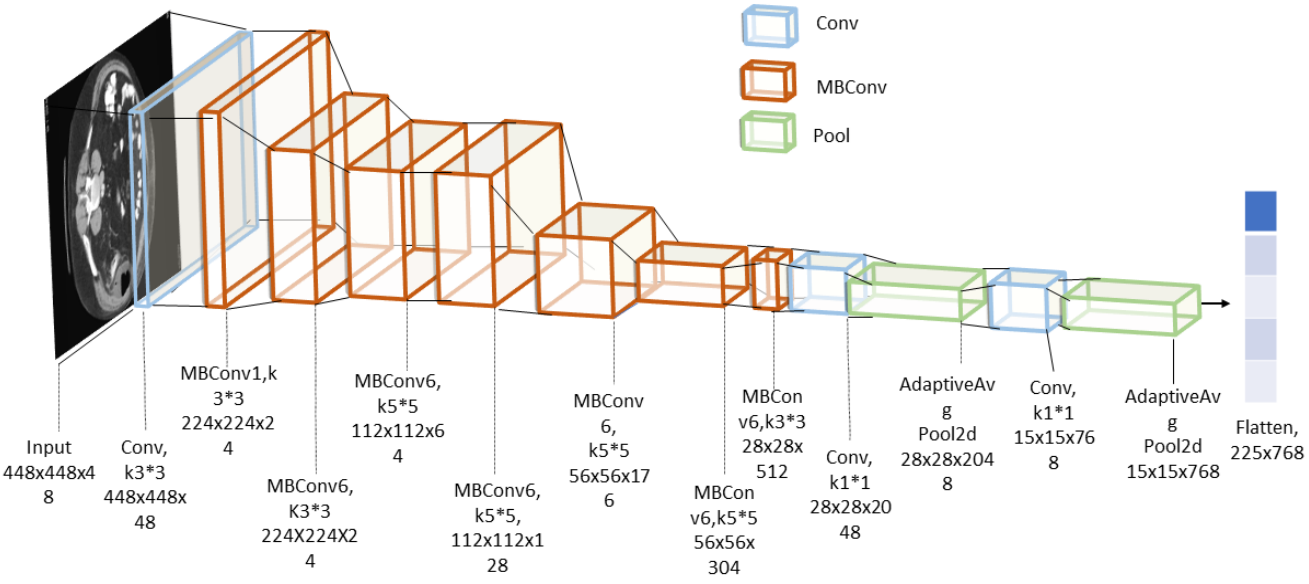

Figure 3.5: The architecture of the EfficientNet-b5 model.

To extract features, we make use of the pre-trained EfficientNet [14]. First, before feeding into CNNs we resize images to 448 × 448. Next, Outputs from EfficientNet's final convolution layer (The Conv, k1*1 in Figure 3.5) will be applied to a new convolution with Number of channels produced by the convolution embed size and pooling layers intended to rescale as features in the image. The primary reason is that it has an optimally balanced EfficientNet feature that allows it to be flexible with different hardware capacities and computational resources. The output of image embedding is $I \in \mathbb{R}^{s \times d}$, where s represents $H_{out} * W_{out}$ of scale pool layers with size is 2 and stride 1, d is embed size.

3.4. Question features extraction

In order to extract contextual information from the questions, we decided to initialize the question encoder with pre-trained RoBERTa (RadBERT-RoBERTa-4m) as shown in Figure 3.6. Initially, the Transformer model family has demonstrated its impressive performance although recurrent neural networks were the powerful solution in NLP for a long time. Transformer gives significant improvement in either accuracy or train costs [15]. Next and the most important, this pre-trained model was built from 4

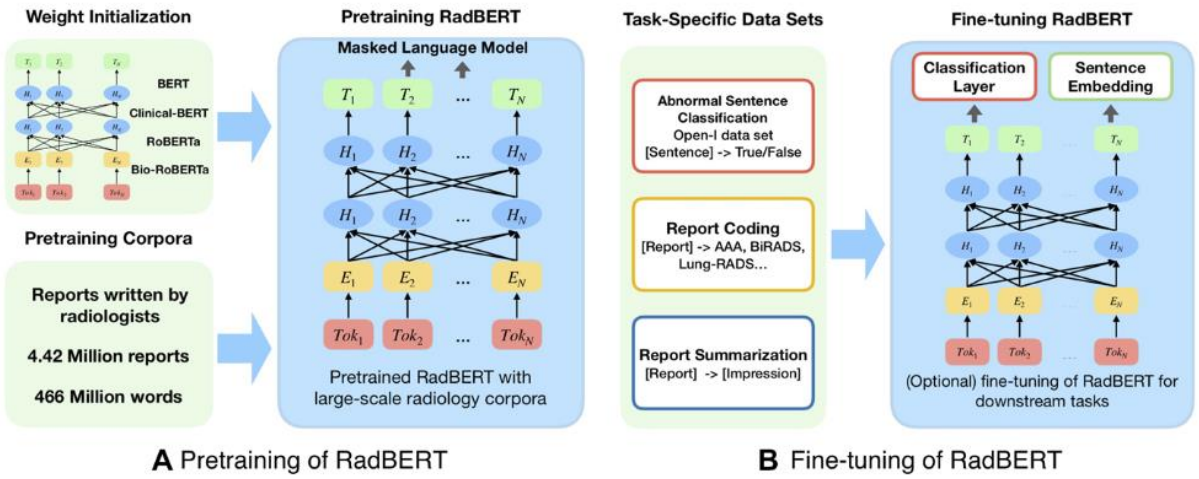

Figure 3.6: Pre-training and fine-tuning flowchart for BERT

million radiology reports, the same as the type of images in the VQA-RAD dataset, RAD stands for radiology. This pre-trained model outperforms all earlier medical domain models such as BioBERT, Clinical-BERT and BioMed-RoBERTa [64] in terms of medical language understanding.

3.5. Associative Memory Module (AMM)

SAM-Based Two-Memory (STM) is a memory model introduced by Hung el al. [65] to handle the memory limitations in Nth-farthest item and textual question answering. To achieve significant improvements in efficiency, the network's memory must be able to extract features and structural relationships in its item and relational memory units. Nonetheless, such only one memory module in a neural network has difficulty remembering relational representation. To deal with this challenge, the model was designed as a two memory model, which has relational memory separate item memory. Additionally, the two separate memories need to interact to enhance each other's representations (See Fig 3.7 for detail). Hung et al. [64] introduced a new operator called Self-attentive Associative Memory with outer product attention. In this thesis, we embrace this model with some modification in input to memorize the item and relation between image and question then enrich visual-linguistic features.

3.5.1. Overall AMM's architecture

AMM dynamically updates its item memory $\mathcal{M}_t^i$ using gating mechanisms at each timestep based on input x^t (Eq. 9). The output of both the relational memory and the item memory is add and passed as input to the Self-Attentive Memory (SAM), the process leads to the generation of a novel relational representation, which is subsequently employed to update the state of the relational memory $\mathcal{M}_t^r$ (Eq. 12-13). The relational memory transfers its wisdom to the item memory (Eq. 14) and contributes information to the final output value (Eq. 15).

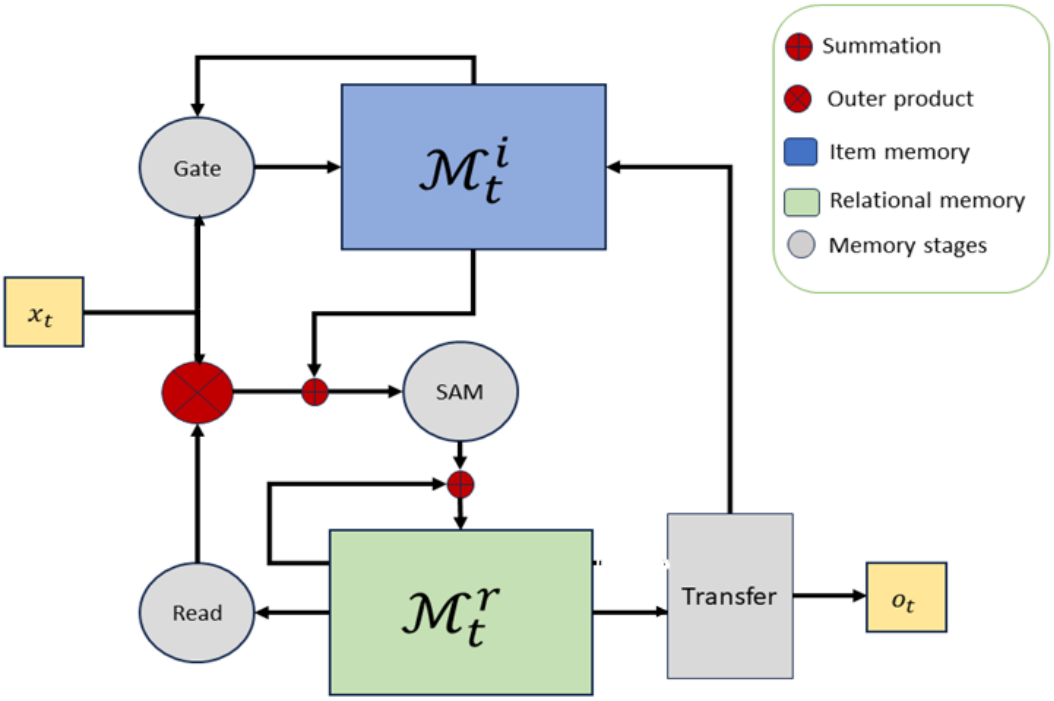

Figure 3.7: Illustration of Associative Memory Module

3.5.2. Self-Attentive Memory

a. Outer product attention (OPA)

OPA is a more advanced version of dot-product attention. Dot product attention (DPA) for be expressed as follows,

$$A^{\circ}(q, K, V) = \sum_{i=1}^{n_{kv}} Softmax(q.k_i)v_i \quad (2)$$

Where $A^{\circ} \in \mathbb{R}^{d_v}$; $q, k \in \mathbb{R}^{n_{kv}}$; "." is a dot product, with single query q and n_{kv} pairs of key-value.

OPA can be form as,

$$A^{\otimes}(q, K, V) = \sum_{i=1}^{n_{kv}} F(q \odot k_i) \otimes v_i \quad (3)$$

Where $A^{\otimes} \in \mathbb{R}^{d_{qk} \times d_v}$; $q, k_i \in \mathbb{R}^{d_{qk}}, v \in \mathbb{R}^{d_v}$, $\otimes$ is outer product, $\odot$ is element-wise multiplication and F is chosen as the tanh function.

The key distinction between OPA and DPA lies in their functionalities. DPA prioritizes the retrieval of a single, relevant item based on the query and key-value pairs. In contrast, OPA concentrates on forming a new relational representation, capturing the intricate connections and relationships between multiple items. By forming a relational representation, OPA manages to capture every individual bit-level association that exists between the value and the corresponding key-scaled query. This offers two benefits: (i) a DPA cannot provide higher-order representational capacity and (ii) retrieve stored items with a contraction operation by a type of memory that can be later recalled.

b. Self-attentive Associative Memory (SAM)

Given a memory $M \in \mathbb{R}^{n \times d}$and parameter weights $W_q \in \mathbb{R}^{n_q \times n}$,$W_k \in \mathbb{R}^{n_{kv} \times n}$,$W_v \in \mathbb{R}^{n_{kv} \times n}$, the operator retrieves n_q queries, n_{kv} keys and values from M as M_q, $M_k$and M_v, respectively,

$$M_q = LN(W_q M) \quad (4)$$

$$M_k = LN(W_k M) \quad (5)$$

$$M_v = LN(W_v M) \quad (6)$$

LN is denoted layer normalization operation [66]. Then, SAM generates a relational representation $SAM_\theta(M) \in \mathbb{R}^{n_q \times n \times n}$, the l-th element of the first dimension is defined as,

$$SAM_\theta\ (M)\ [l]\ = A^{\otimes}\ (M_q[l], M_k, M_v\ )$$
$$= \sum_{i=1}^{n_{kv}} F(M_q[l]\ \ M_k[i])\ \otimes\ M_v\ [i]) \quad (7)$$

where s = 1, ..., n_q. W_q[j], $W_k[i]$ and $W_v[i]$ is the j-th row of matrix W_q, the i-th row vector of matrix W_k and W_v, in the order given.

From M, model receive items to form $SAM_\theta\ (M)$ a new set of hetero-associative memories using Eq. 7. Every representation we mention is the relationship between a specific query and its associated collection of value. And preserving the greatest possible retrieval for the item memory is the role of the keys.

3.5.3. Associative Memory based on two Memory Model

The model consists of two memory unit $\mathcal{M}_t^i \in \mathbb{R}^{n \times n}$ for items and $\mathcal{M}_t^r \in \mathbb{R}^{n_q \times n \times n}$for relationships. At each timestep, we use the the previous state of memories $\mathcal{M}_{t-1}^i$,$\mathcal{M}_{t-1}^r$ and current input data x_t. To produce and new state of memories $\mathcal{M}_t^i$,$\mathcal{M}_t^r$ and output o_t . The following are descriptions of the memory stages,

$\mathcal{M}_t^i$-**Write** The data from the input is distributed throughout the rows of the item memory as associative memory. When an input x_t is received, the item memory is updated according,

$$X_t\ =\ f_1(x_t)\ \otimes\ f_2(x_t) \quad (8)$$

$$\mathcal{M}_t^i\ =\ \mathcal{M}_{t-1}^i\ + X_t \quad (9)$$

where $f1$ and $f2$ are fully connected neural networks that output a d-dimensional vector. The gating mechanisms of LSTM are utilized to enhance the performance of Eq. 9 as,

$$\mathcal{M}_t^i\ =\ F_t(\mathcal{M}_{t-1}^i, x_t)\ \odot\ \mathcal{M}_{t-1}^i\ + I_t(\mathcal{M}_{t-1}^i, x_t)\ \odot\ X_t \quad (10)$$

where I_t and F_t are input and forget gate with detail as,

$$F_t(\mathcal{M}_{t-1}^i, x_t) = W_F x_t + U_F tanh(\mathcal{M}_{t-1}^i) + b_F \quad (11)$$

$$F_I(\mathcal{M}_{t-1}^i, x_t) = W_I x_t + U_I tanh(\mathcal{M}_{t-1}^i) + b_I \quad (12)$$

$\mathcal{M}_t^r$-**Read** as relationships stored in relation memory, for rebuilding the previously seen items can be read to the relational memory.

$$v_t^r = softmax(f_3(x_t)^\top)\mathcal{M}_{t-1}^r f_2(x_t) \quad (13)$$

where f_3 is a fully connected neural network with n_q dimensional outputs vector. The read information from the previous state of $\mathcal{M}^r$ provides an additional input coming to the constructing relational process.

$\mathcal{M}_t^r$-**Write** $\mathcal{M}_t^i$-**Read** SAM will be used to and construct a candidate relational memory and read from $\mathcal{M}_t^i$ as follow,

$$\mathcal{M}_t^r = \mathcal{M}_{t-1}^r + \alpha_1 SAM_\theta(\mathcal{M}_t^i + \alpha_2 v_t^r \otimes f_2(x_t)) \quad (14)$$

where α_1 and α_2 are scaling hyper-parameters. The combination of the present item memory $\mathcal{M}_t^i$ and the current input data x_t and the relationship between the extracted item as an input for SAM from the previous relational memory v_t^r. With information from the far-off past, v_t^r enhances the relational memory.

$\mathcal{M}^r$- **Transfer** In this stage, by using high dimensional transformation, the $\mathcal{M}_t^r$ is transferred to the item memory is relational knowledge,

$$\mathcal{M}_t^i = \mathcal{M}_t^i + \alpha_3 G_1 \circ V_f \circ \mathcal{M}_t^r \quad (15)$$

where V_f is a function use to the input tensor be flattens the first two dimensions, G_1 is a Multilayer perceptron neural network that maps $\mathbb{R}^{(n_{kv} \times d) \times d} \rightarrow \mathbb{R}^{d \times d}$ and α_3 is a combining hyper-parameter. With trivial G_1, the transfer acts as though long-term memory results of the relational memory are added to the item memory. Hence, the useful in enhancing long-term memory is transfer $\mathcal{M}^r$-Transfer. Furthermore, we produce the relational memory to an output $o_t \in \mathbb{R}^{n_o}$ at each timestep. We alternatively apply high dimensional transformations and flatten like this,

$$o_t = G_2 \circ V_l \circ G_3 \circ V_l \circ \mathcal{M}_t^r \quad (16)$$

where n_r is a hyper-parameter, V_l is a function that the input tensor flattens the last two dimensions. G_2 and G_3 are Fully Connected neural networks that map $\mathbb{R}^{n_q \times (dd)} \rightarrow \mathbb{R}^{n_q \times n_r}$ and $\mathbb{R}^{n_q \times n_r} \rightarrow \mathbb{R}^{r_o}$. As opposed to the contraction (Eq. 11), it does not easily reconstruct the stored memory in the distillation process. It is able to capture the bi-linear representations that are kept in the relational memory through high-dimensional transformations. Because of this, the output is helpful for relational and item learning, even if it is in vector form because it contains rich representation.

3.6. Fusion module

Before delivering the Encoder-Decoder, we first present its fundamental component Encoder-Decoder attention and Prototype learning.

3.6.1. Encoder-Decoder attention

The Encoder-Decoder attention is a composite module that combines two basic attention units: Cross-Attention and Self-Attention, as illustrated in Figure 3.8. In [29], to improve the attending features' ability to be represented, the concept of multi-head attention is introduced. This involves parallelizing 'heads,' where each head represents an individual scaled dot-product attention function.

$$MA(Q,K,V) = [head_1, head_2, \ldots, head_h]W^o$$
$$head_i = A(QW_i^Q, KW_i^K, VW_i^V) \quad (17)$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_h}$ are the projection matrices for the i-th head, and $W^o \in \mathbb{R}^{(h*d_h) \times d}$. d_h is the dimensionality of the output features from each head.

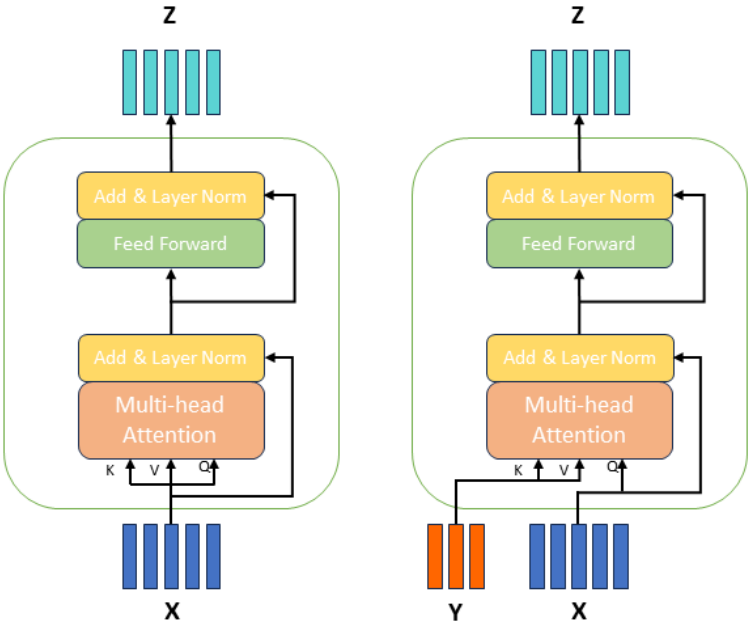

Figure 3.8: Self-Attention (left) and Cross-Attention (Right).

The encoder-decoder model (See in Figure 3.9) is motivated by the proposed Transformer model in [15]. Given image features X and Question Features Y as input. The encoder-decoder attention strategy might be viewed as an encoder that learns the characteristics of the attended question $Y^{(L)}$ with L stacked SA units and a decoder to use $Y^{(L)}$ to become familiar with the attended image features $X^{(L)}$ with stacked Self-Cross-Attention units.

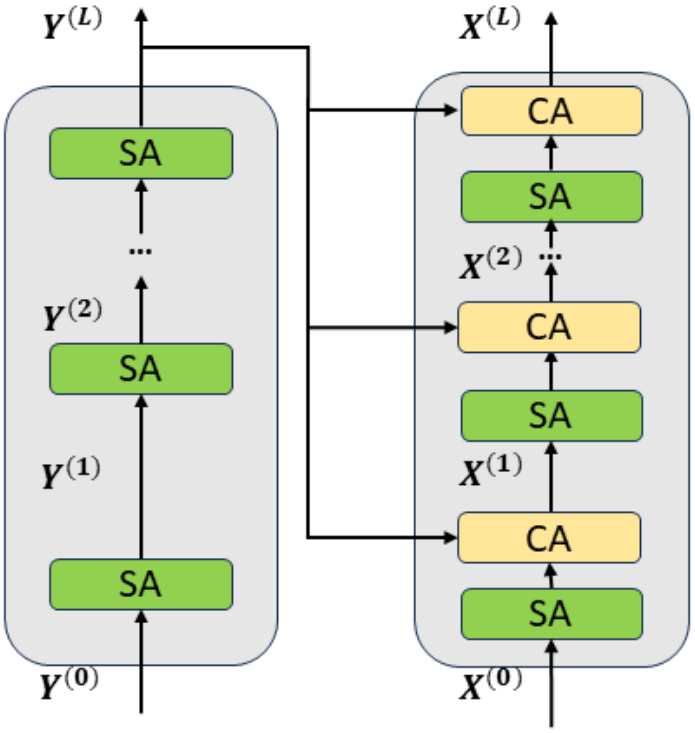

Figure 3.9: Encoder-Decoder attention.

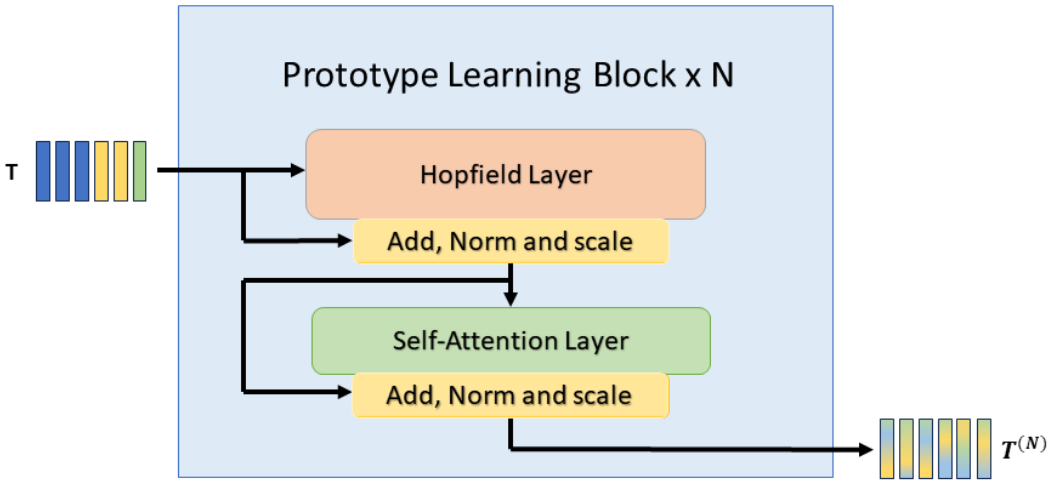

Figure 3.10: Detail of Prototype Leaning Block.

3.6.2. Prototype Learning Block

It is extremely difficult to directly learn genuine but highly different semantic representations because of the enormous modality gap that exists between inquiries and medical images. We apply using the visual-linguistic feature space for prototype learning (Figure 3.10). More specifically, we apply the Hopfield layer and the self-attention layer that make up the prototype learning block. Each sub-layer uses a residual connection, and

the Normalization and Scale layers come next.

Hopfield Layer. The initial proposal of the Hopfield layer, primarily designed for recognizing stored patterns or static queries, is found in the current Hopfield network [38]. Our objective is to enhance the network's capability to autonomously leverage clusters of high-level semantic concepts inherent in the data samples. This enhancement allows the network to systematically discover the most representative prototypes using the visual-linguistic features that are input. These acquired prototypes can be considered as code words representing more intricate semantic concepts. The Hopfield layer employed in this study is depicted in Figure 3.11. Formally, let $R \in \mathbb{R}^{(s+l)\times d}$ denote the input visual-linguistic feature R. The Prototype content and matrix prototype lookup matrix are then defined as $W_{store}, W_{lookup} \in \mathbb{R}^{n_{prot}\times d}$, where n_{prot} is the number of prototypes. The output $Z \in \mathbb{R}^{s\times d}$ can be be formulated as:

$$Z = softmax(\beta R W_{lookup}^{T}) W_{store}, \tag{18}$$

Where $\beta = \frac{1}{\sqrt{d}}$ is a scaling scalar.

With Eq. 18 the prototype learning can achieved by Hopfield layer the from two perspective : 1) Saving the stored patterns and learning the most representative prototypes in W_{store}, each row in W_{store} the equate to a stored visual-linguistic prototype; and 2) Using the prototype learning, the learnt prototypes from lookup matrix W_{lookup} with learning to represent the input visual-linguistic features X. Thus, the $softmax(.)$ in (*) can be comprehended as the probability of each class (prototype) in the mapping.

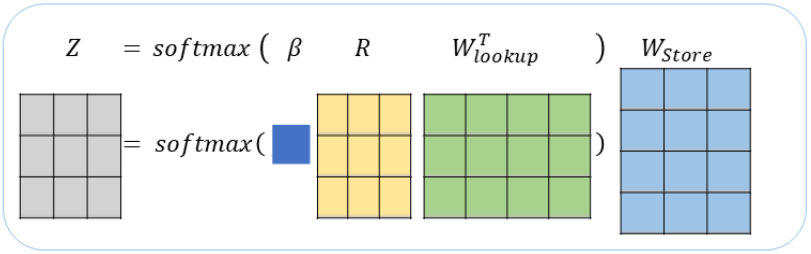

Figure 3.11: The Illustration of the Hopfield layer.

3.6.3. Features fusion

After the Encoder-Decoder attention stage, the output image features $X(L) \in \mathbb{R}^{s\times d_{embed}}$ and question features $Y(L) \in \mathbb{R}^{l\times d_{embed}}$ already contain rich information about the attention weights over the image regions and question words. Moreover, with the memory features $M \in \mathbb{R}^{1\times d_{embed}}$ representing relationships between arbitrary stored items from AMM. We concatenate all of them into $M \in \mathbb{R}^{(s+l+1)\times d_{embed}}$, then feed to the N prototype learning block to learn the prototype of each class $Z \in \mathbb{R}^{(s+l+1)\times d_{embed}}$.

3.7. Answer components and loss function

The answer will be classified with output of visual-linguistic features fusion Z. First, the fused features Z is projected to vector $z \in \mathbb{R}^{1\times d_{embed}}$ by an average pooling operator. Then, we feed vector z into a sequential fully connected layers ($Dropout$ - FC_1- $ReLU$ - FC_2), where $FC_1 \in \mathbb{R}^{d_{embed}\times n_{hidden}}$, n_{hidden} is number of hidden nodes, $FC_2 \in \mathbb{R}^{n_{hidden}\times n_{class}}$, n_{class} is the number of classify classes.

In reality the class in the medical VQA dataset usually imbalance [2]. Thus, this means that the model can be biased toward a class. Nevertheless, we found it beneficial of loss function with Focal Loss [67], which affords us to address class imbalance during training. We define the model losses mathematically below:

$$L_{Focal}(p_t) = -(1-p_t)^{\gamma} log(p_t) \qquad (19)$$

where γ is hyper-parameter and p_t probability of class t. As mentioned previously, the output of FC_2 is $y = [y_1; y_2; y_3; \ldots; y_{n_{class}}] \in \mathbb{R}^{n_{class}}$. Thus, we use $softmax(y)$ to calculate the probability of each class $p = [p_1; p_2; p_3; \ldots; p_{n_{class}}] \in \mathbb{R}^{n_{class}}$, $p_i \in [0, 1]$. With t is ground-true class, the model loss is $L_{Focal}(p_t)$.

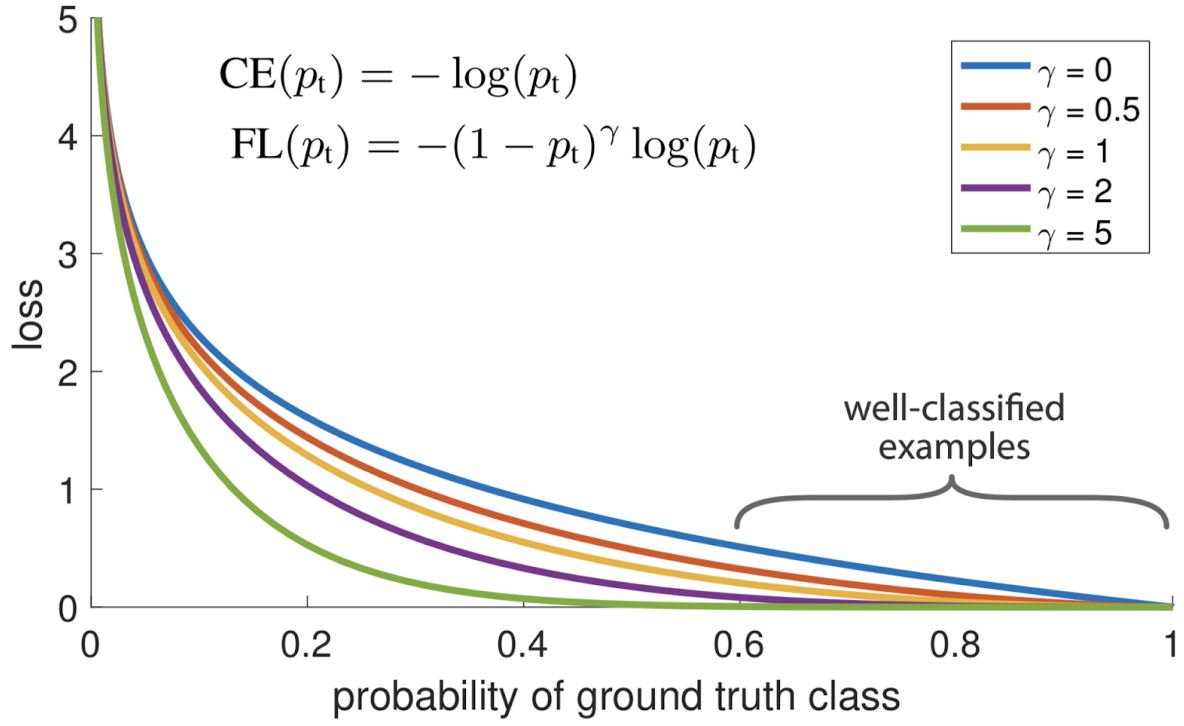

Figure 3.12: The Focal Loss down weights with factor of $(1 - p_t)$ easily [67]. CE is Cross Entropy, FL is Focal Loss.

4. Results and discussion

4.1. Dataset

4.1.1. Visual question answering in Radiology (VQA-RAD dataset)

VQA-RAD [48] is a medical dataset published in 2018 for only radiology. The images are equally distributed into the abdomen organs, chest and head. The dataset VQA-Rad [48] 315 radiology images distributed evenly into head, chest, and abdomen and contains 3515 Question-Answer pairs tested by clinicians. Each image can have one or many questions, but we take each individual question as the input. The questions are classified into 11 classes, like color, abnormality, organ system, modality, attribute, counting, object/condition presence, positional reasoning, plane, size, and others. Most of the answers are closed-ended which can be answered with short word type (58%), while remaining ones are open-ended with unrestricted length of the answers.

4.2. Evaluation metric

As mentioned in Section 2 above, Answer Components can be either classification type or text-generating (language) type. Respectively, in classification tasks, metrics are accuracy, F1 score in general. It treats output as a class and calculates exactly the distance of predicted answer to ground truth answer. In language tasks, metrics, which are used to evaluate sequence tasks (translation, summarization, caption, et al.) can be BLEU score, which estimates the similarity of the two sentences. BLEU score is the most popular choice in language tasks, but in med-VQA datasets, sentences are shorter than in other datasets, thus it becomes inefficient. However, there are many other metrics currently being used in evaluating med-VQA datasets based on Answer components, but in this study, we choose accuracy as the main metric because we defined this as a classification task and to compare with other approaches.

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

4.3. Specific Implementation

We have taken great care to ensure that this work is thorough and complete in implementing our model. In this section we explain the implementation process in detail.

Code: We reused a pre-existing codebase from the SAM model [65] 1 to build our own model. [65]. In addition, we created the data processing, data loading, training-test-inference procedures and visualization as well as all the source code ourselves.

Frameworks and libraries: Python programming and the PyTorch framework are used to code our approach. We can easily apply the model strategy during training and inference with PyTorch as it makes reading, preprocessing, and feeding the training data effortless. A tensor board tool is another feature of PyTorch that makes it easier for users to monitor training progress and view evaluation results, test losses, and training losses. In addition to PyTorch, several other built-in Python libraries are also used, including Numpy for matrix calculation, Pandas and Matplotlib for result analysis, and additional auxiliary libraries such as Einops, Pydash, torchinfo, Shapely, Timm, Torch-Lightning and Transformers.

Environment: For implementation, debugging and data analysis we use Google Colab virtual machines with the following configuration: 12.7 GB of RAM and Intel Xeon CPU, along with GPU T4 with 15GB of VRAM. For training, we use cloud computing platform Vast.ai to rent virtual machines equipped RTX 3090 Ti with 24 GB of VRAM and AMD Ryzen 9 5900X.

Hyperparameters: Table 1 provides a detailed breakdown of the hyperparameters used during our training process. There are a few changes that we made from the original work:

Parameter	Value
Learning rate	0.0001
Epochs	100
Batch size	16
Optimizer	Adam
lr scheduler	Step lr
Focal loss's gamma	2
N-bit precision	16 bit mixed
Random seed	42

Table 1: Overall training hyper parameters

We set the input size of the image as 488 x 488 and the max padding size of the tokenizer for the question is 30. The Adam optimizer algorithm with a learning rate of 1e-5 was our choice. We also incorporated the StepLR Scheduler during training to make the model learn better. as shown in Figure 4.1. Then train the model batch size 16, and max num epochs 100.

Parameter	Notation	Value
General Hidden size	d, n_o, d_{embed}	512/768/1024
Question max length padding	l	20/30/40
Number of prototypes	n_{prot}	500/1000/1500
Number of Prototype learning block	N	8/10/12
Number of memory slot	n_q	1/6/12
Dropout	Dropout	0.4

Table 2: Overall model hyper parameters

We must choose the hyper-parameter appropriate for each module because of the model trade-off between accuracy and resource. As shown in Table 2, a model is considered positive if it achieves highest performance with lowest resource so we use some combination of model hyper-parameters to evaluate before training.

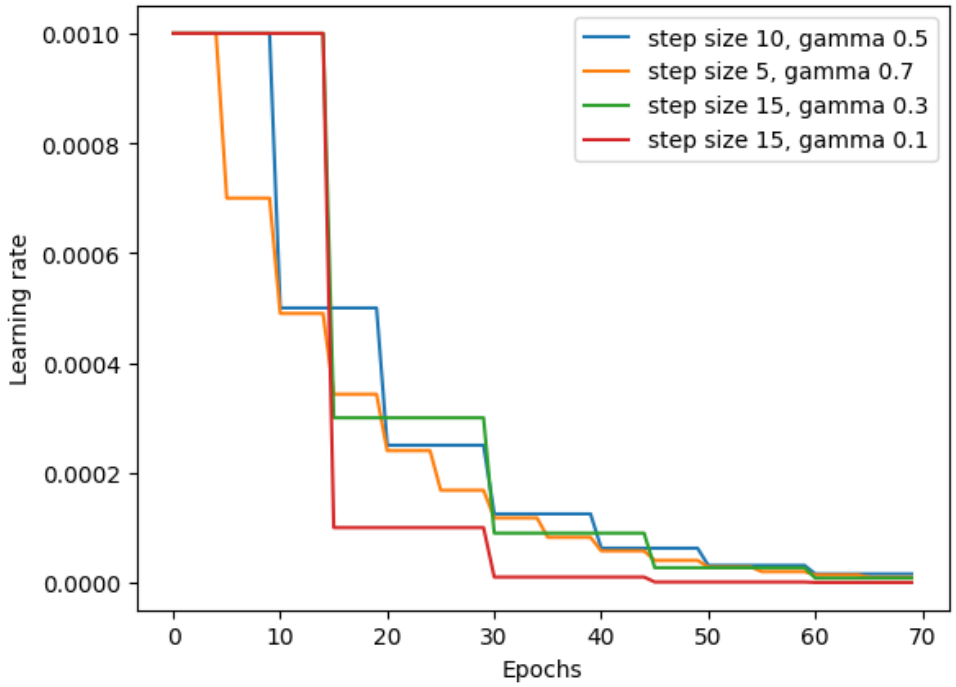

Figure 4.1: The learning rate changes following epochs of StepLR. We use step size 15 and gamma = 0.3 for our model.

4.4. Results and analysis

Our method demonstrated superior performance compared to the current state-of-the-art approach, as indicated in Table 3. In the evaluation on VQA-Rad datasets, our approach surpassed other methods in both open-ended and overall performance. Table 3 demonstrates how our proposed model performs better than the current state-of-the-art method Q2ATransformer by an absolute margin of 0.45% overall, with improvements of 0.3% and 0.72% in open and closed-ended scenarios, respectively. But compared to the best methods on closed-ended questions, our model 11.19% absolute improvement on open-ended questions and 4.43% overall. Based on these results, we can see our model shows significant benefits when answering open-ended questions, which supports our idea by using associative memory to enrich the memory of the network and prototype learning to learn representative prototypes from visual-linguistic features.

On the other hand, we also train and compare our model with fine tuning the model hyper-parameters follow table 4. We compare performance and resources in each module to discover fine parameters for our model with RTX 3090 Ti with 24GB of RAM and AMD Ryzen 9 5900X.

Table 3: Comparisons our method with the state-of-the-art methods on the VQA-RAD test set.

Methods	Closed	Open	Overall
BAN-VQAMix [20]	74.0	53.8	65.9
CMSA-MTPT [32]	77.3	56.1	68.8
MMQ-BAN [5]	75.8	53.7	67.0
FITS [8]	**82.0**	68.2	76.5
hi-VQA [15]	-	-	76.3
Q2ATransformer [40]	81.2	<u>79.19</u>	<u>80.48</u>
Ours	<u>81.98</u>	**79.39**	**80.93**

To find the main generalization of the model. First, we try general hidden size with the rest of parameters as the same as the original work. As shown in Table 4, the general hidden size set to 1024 is better than the others. However, the model size increased 43M compared to 768 but the accuracy just increased 0.4%. Therefore, we trade-off the accuracy for resources and set the general hidden size to 768 for other hyper-parameters.

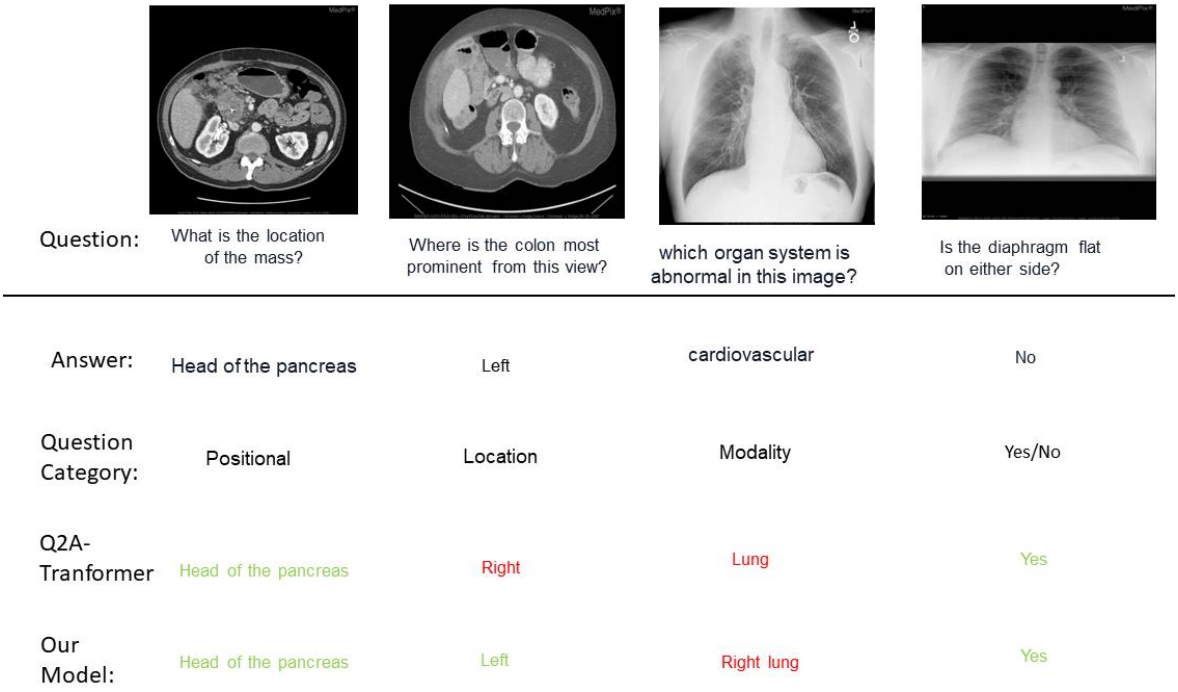

Figure 4.2: Examples from VQA-RAD dataset.

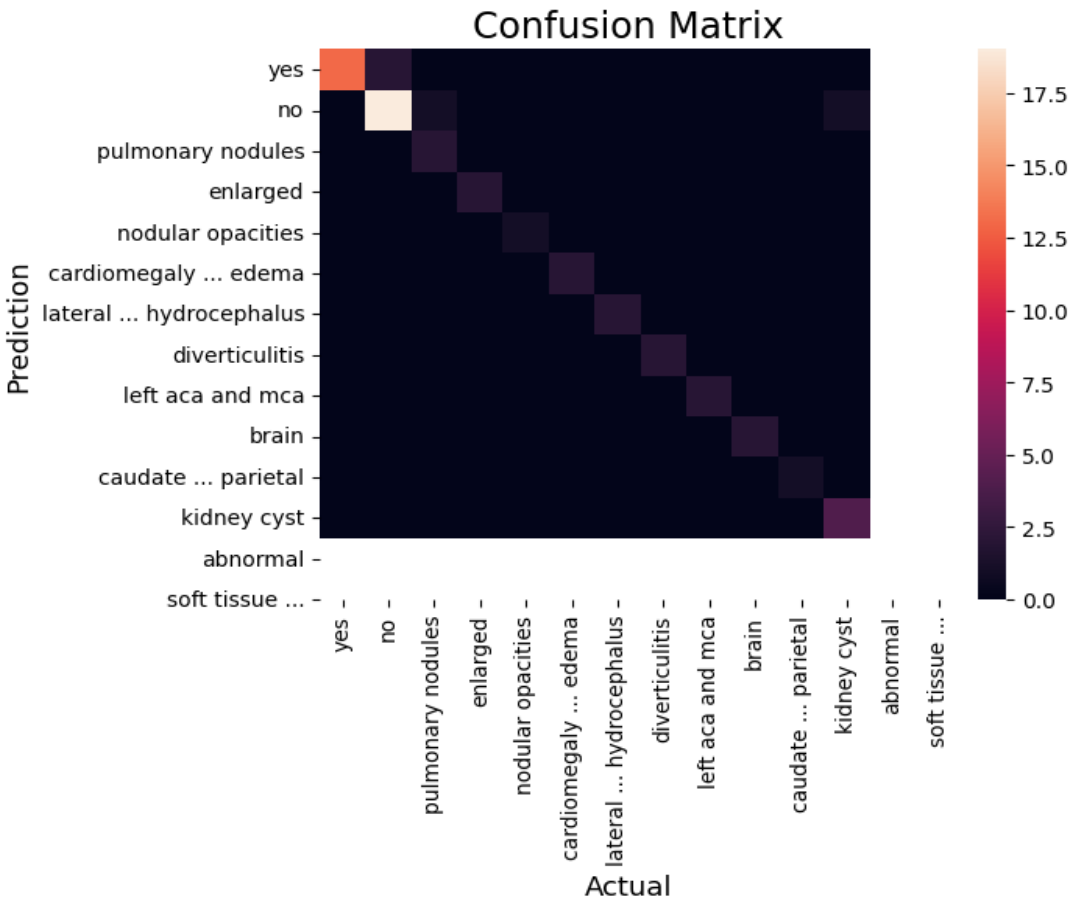

Figure 4.3: Confusion Matrix of abnormality questions.

Table 4: Comparisons with modified size of hidden on the VQA-RAD test set.

General hidden size	Model size (parameter)	Accuracy (%)
512	187.5M	65.3
768	213M	73.8
1024	**255M**	**74.2**

Table 5: Comparisons with modified max padding length.

Max padding length	Accuracy (%)
10	62.4
30	**73.8**
50	68.1

We fine-tune the parameters following the sequence of architecture starting with max padding length of question. Our tests showed in table 5 that the model's learning capacity can be limited by using small or excessively large values for these hyperparameters. Additionally, we note that the input question tokens have a maximum length of 32. Therefore, 30 is the maximum padding length suitable for encoding the query.

In conclusion, our model overcomes the OOD problem as shown in Figure 4.2 and generalizes. Figure 4.3 shows a confusion matrix of question abnormality questions, our model predicts almost respond well to non-binary questions.

4.4.1. The impact of the Associative Memory Module

We adopt STM[27] as the AMM for our model, with the two memory modules serving as the relational and item memory. The AMM of our model uses a memory network strategy, directly capturing the bi-linear representations that are kept in the relational memory through high-dimensional transformations and enriching the visual-linguistic features. Our experiments on AMM are present on table 6.

As shown on Table 6, our model achieves groundbreaking accuracy with accuracy, but the training time is much higher without AMM. Furthermore, AMM helps models converge faster and learn better (See in Figure 4.2). Therefore, the memory module

Table 6: Comparison of models with different hyper parameters of AMM

Model	Accuracy (%)	Average training time (s/epoch)
w/o AMM	62.4	61
$n_q = 1$	68.8	65
$n_q = 6$	75.2	96
$n_q = 12$	**79.7**	**119**

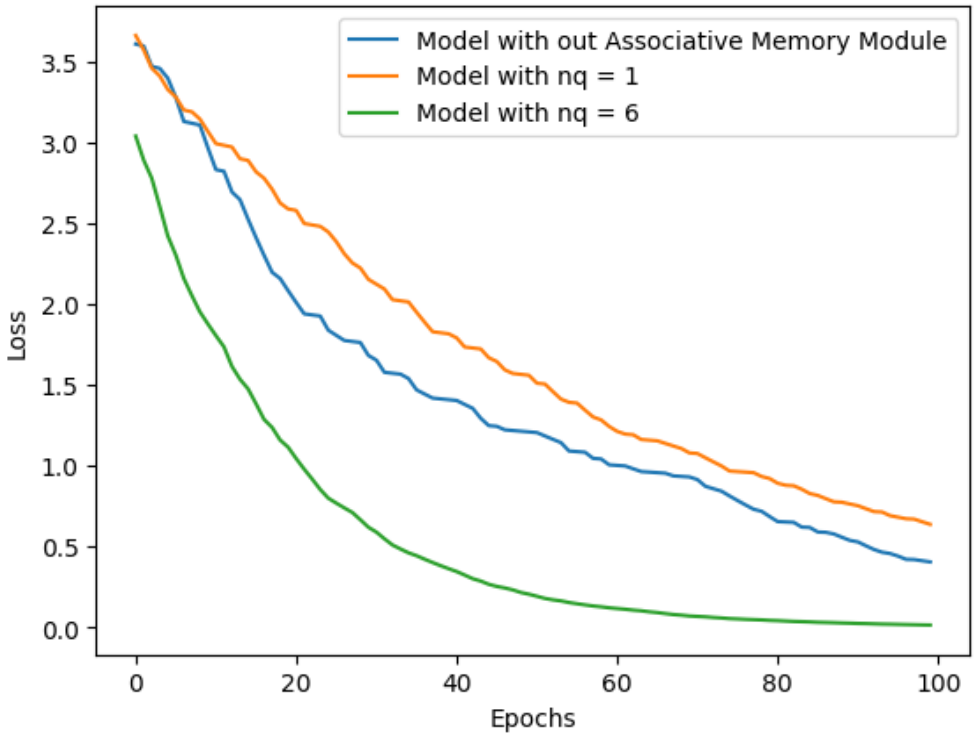

Figure 4.4: Training process of model with AMM hyper-parameter modification and the others hyper-parameter same as above.

facilitates for the features fusion module learning a set of representative examples for each class. However, we also found that AMM consumes a lot of resources (See Figure 4.3). After all, AMM with $n_q = 12$ We decided to trade-off resources for precision with one of the reasons presented in the next part.

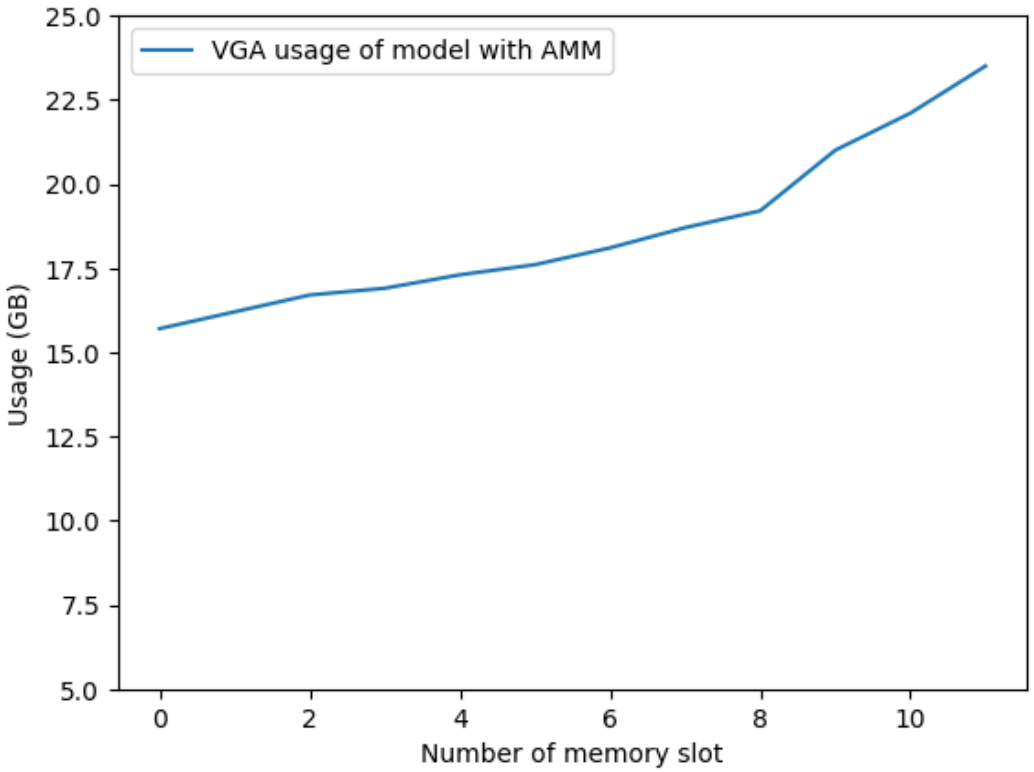

Figure 4.5: GPU consumption of model on VQA-RAD. The usage is calculated on the entire model process with batch size 16 and similar to the above hyper-parameter.

4.4.2. The impact of Prototype Learning

Generalization performance of the model: Prototype learning is a technique that involves learning a set of prototypes or representative examples for each class in a classification problem. These prototypes act as exemplars of each class, embodying the key features that differentiate one class from another. The Hopfield layer facilitates hierarchical learning, where coarse-to-fine the visual-linguistic features are extracted from the text and image embeddings. This allows the model to grab both local and global semantic information effectively.

Table 7: The model accuracy (%) of each set number prototype and number of block prototype learning.

No of prototype/block	5	10	15
500	80.1	80.47	79.96
1000	80.24	**80.93**	80.24
1500	80.18	80.51	80.04

5. Conclusion

In this thesis, we have introduced a new architecture in medical VQA, that is based on the concept of Associative Memory a to enrich the visual features and Prototype Learning to represent classes. In addition, our model is designed to overcome the data limitations and improve model generalization. In a benchmarking experiment the architecture outperformed all other methods and is demonstrated to memorize and visual-linguistic reasoning in answering visual medical questions.

However, there are some limitations in our proposed method that need to be resolved. First, the Associative Memory Module works fine but consumes quite a bit of VRAM. This can be improved with optimizing the AMM or change to a new core operator in AMM. Second, the Prototype Learning is not performed as well as we expected, and the effective architecture needs to be created to improve its performance.

In conclusion, this thesis has proposed a new approach for the Med-VQA problem, exhibiting considerable increases in accuracy compared to state-of-the-art models. The future work highlighted in this thesis further improves the proposed method and contributes to the model generalization. The results of this thesis will contribute to the VQA and multimodal.

REFERENCES

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C., Parikh, D., 2015. VQA: Visual question answering, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 2425–2433.

[2] Zhihong Lin, Donghao Zhang, Qingyi Tac, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. arXiv preprint arXiv:2111.10056, 2022.He, X., Zhang, Y., Mou, L., Xing, E., Xie, P., 2020. PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 .

[3] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al., 2020. Human-computer collaboration for skin cancer recognition. Nature Medicine 26, 1229–1234.

[4] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang1 , Weidi Xie 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. arXiv preprint arXiv:2305.10415v

[5] Li, P., Liu, G., He, J., Zhao, Z., & Zhong, S. (2023). Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. Lecture Notes in Computer Science, 374-383.

[6] Papers with Code - VQA-RAD Benchmark (Medical Visual Question Answering). (n.d.). Papers With Code. Retrieved December 12, 2023, from https://paperswithcode.com/sota/medical-visual-question-answering-on-vqa-rad?metric=Overall%20Accuracy

[7] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

[8] a

[9] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415.

[10] Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., & Xie, W. (2023). Pmc-clip: Contrastive language-image pre-training using biomedical documents. arXiv preprint arXiv:2303.07240.

[11] Li, P., Liu, G., Tan, L., Liao, J., & Zhong, S. (2023, April). Self-Supervised Vision-Language Pretraining for Medial Visual Question Answering. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[12] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image

recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[14] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.
[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.
[18] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5, 1–10.
[19] Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021, April). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1650-1654). IEEE.
[20] Huang, Y., Wang, X., Liu, F., & Huang, G. (2022, July). OVQA: a clinically generated visual question answering dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2924-2938).
[21] Liu, B., Zhan, L.M., Wu, X.M., 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 210–220.
[22] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[23] Zhang, S., Zheng, D., Hu, X., & Yang, M. (2015, October). Bidirectional long short-term memory networks for relation classification. In Proceedings of the 29th Pacific Asia conference on language, information and computation (pp. 73-78).
[24] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
[25] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical

text mining. Bioinformatics, 36(4), 1234-1240.
[26] Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C., 2021. Mmbert: Multimodal bert pretraining for improved medical vqa, in: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1033–1036. doi:10.1109/ISBI48211. 2021.9434063.
[27] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas. pp. 457– 468.
[28] Yu, Z., Yu, J., Fan, J., Tao, D., 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 1839–1848.
[29] Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Transactions on Neural Networks and Learning Systems 29, 5947–5959. doi:10.1109/TNNLS.2018.2817340.
[30] Zeithamova D, Maddox WT, Schnyer DM. Dissociable prototype learning systems: evidence from brain imaging and behavior. J Neurosci. 2008 Dec 3;28(49):13194-201. doi: 10.1523/JNEUROSCI.2915-08.2008. PMID: 19052210; PMCID: PMC2605650.
[31] Cunningham, Padraig & Delany, Sarah. (2007). k-Nearest neighbour classifiers. Mult Classif Syst. 54. 10.1145/3459665.
[32] A. Sato and K. Yamada, “Generalized Learning Vector Quantization,” in Advances in Neural Information Processing Systems 8 (NeurIPS), 1996, pp. 423–429.
[33] Leng, Y., Yu, L., & Xiong, J. (2019, October). Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review. In 2019 international conference on multimodal interaction (pp. 395-403).
[34] Yang, Hong-Ming & Zhang, Xu-Yao & Yin, Fei & Liu, Cheng-Lin. (2018). Robust Classification with Convolutional Prototype Learning. 3474-3482. 10.1109/CVPR.2018.00366.
[35] Dong, Nanqing & Xing, Eric. (2018). Few-Shot Semantic Segmentation with Prototype Learning.
[36] Zhu, Yi & Li, Xinyu & Liu, Chunhui & Zolfaghari, Mohammadreza & Xiong, Yuanjun & Wu, Chongruo & Zhang, Zhi & Tighe, Joseph & Manmatha, R. & Li, Mu. (2020). A Comprehensive Study of Deep Video Action Recognition.
[37] L. Li, X. Mu, S. Li and H. Peng, "A Review of Face Recognition Technology," in IEEE Access, vol. 8, pp. 139110-139120, 2020, doi:

10.1109/ACCESS.2020.3011028.
[38] Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2020). Hopfield Networks is All You Need. ArXiv. /abs/2008.02217.
[39] Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. arXiv preprint arXiv:1410.3916.
[40] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. Weakly supervised memory networks. arXiv:1503.08895, 2015.
[41] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In Proc. Int. Conf. Mach. Learn., 2016.
[42] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Proc. Conf. Empirical Methods in Natural Language Processing, 2014.
[43] Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014.
[44] M. F. Stollenga, J. Masci, F. J. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In Proc. Advances in Neural Inf. Process. Syst., 2014.
[45] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proc. Int. Conf. Mach. Learn., 2015.
[46] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. arXiv:1410.5401, 2014
[47] A. Joulin and T. Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. In Proc. Advances in Neural Inf. Process. Syst., 2015.
[48] Pellegrini, C., Keicher, M., Özsoy, E., & Navab, N. (2023, October). Rad-ReStruct: A Novel VQA Benchmark and Method for Structured Radiology Reporting. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 409-419). Cham: Springer Nature Switzerland.
[49] S. Chandar, S. Ahn, H. Larochelle, P. Vincent, G. Tesauro, and Y. Bengio. Hierarchical memory networks. arXiv:1605.07427, 2016.
[50] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap. Meta-learning with memory-augmented neural networks. In Proc. Int. Conf. Mach. Learn., 2016
[51] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733, 2016.

[52] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017.
[53] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
[54] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: ICCV. pp. 19–27 (2015)
[55] Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., Hengel, A.v.d.: On the value of out-of-distribution testing: An example of goodhart's law. In: NeurIPS (2020)
[56] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., Zhuang, Y.: Counterfactual samples synthesizing for robust visual question answering. In: CVPR. pp. 10800–10809 (2020)
[57] Chen, L., Zheng, Y., Niu, Y., Zhang, H., Xiao, J.: Counterfactual samples synthesizing and training for robust visual question answering. arXiv (2021)
[58] Kil, J., Zhang, C., Xuan, D., Chao, W.L.: Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In: EMNLP (2021)
[59] Elgendi, M., Nasir, M. U., Tang, Q., Smith, D., Grenier, J. P., Batte, C., ... & Nicolaou, S. (2021). The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective. Frontiers in Medicine, 8, 629134.
[60] Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5, 1–10
[61] Wang, C., Cho, K., & Gu, J. (2019, September 7). Neural Machine Translation with Byte-Level Subwords. arXiv.Org. https://arxiv.org/abs/1909.03341.
[62] Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A., &Hsu, C.-N. (2022). RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence, 4(4). https://doi.org/10.1148/ryai.210258
[63] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.
[64] Le, H., Tran, T., & Venkatesh, S. (2020, November). Self-attentive associative memory. In International Conference on Machine Learning (pp. 5682-5691). PMLR.
[65] Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

[66] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988)