



ARTICLE

Adjusted Reasoning Module for Deep Visual Question Answering Using Vision Transformer

Christine Dewi^{1,3}, Hanna Prillysca Chernovita², Stephen Abednego Philemon¹,
Christian Adi Ananta¹ and Abbott Po Shun Chen^{4,*}

¹Department of Information Technology, Satya Wacana Christian University, Salatiga, 50711, Indonesia

²Department of Information Systems, Satya Wacana Christian University, Salatiga, 50711, Indonesia

³School of Information Technology, Deakin University, Burwood, VIC 3125, Australia

⁴Department of Marketing and Logistics Management, Chaoyang University of Technology, Taichung City, 413310, Taiwan

*Corresponding Author: Abbott Po Shun Chen. Email: chprosen@gm.cyut.edu.tw

Received: 18 August 2024 Accepted: 01 November 2024 Published: 19 December 2024

ABSTRACT

Visual Question Answering (VQA) is an interdisciplinary artificial intelligence (AI) activity that integrates computer vision and natural language processing. Its purpose is to empower machines to respond to questions by utilizing visual information. A VQA system typically takes an image and a natural language query as input and produces a textual answer as output. One major obstacle in VQA is identifying a successful method to extract and merge textual and visual data. We examine “Fusion” Models that use information from both the text encoder and picture encoder to efficiently perform the visual question-answering challenge. For the transformer model, we utilize BERT and RoBERTa, which analyze textual data. The image encoder designed for processing image data utilizes ViT (Vision Transformer), Deit (Data-efficient Image Transformer), and BeIT (Image Transformers). The reasoning module of VQA was updated and layer normalization was incorporated to enhance the performance outcome of our effort. In comparison to the results of previous research, our proposed method suggests a substantial enhancement in efficacy. Our experiment obtained a 60.4% accuracy with the PathVQA dataset and a 69.2% accuracy with the VizWiz dataset.

KEYWORDS

VQA; vision transformer; multimodal data; deep learning

1 Introduction

In the field of deep learning research, the Visual Question Answering (VQA) task is a multimodal challenge that involves computer vision [1,2] and natural language processing [3]. The primary goal of a general VQA model is to generate credible answers to questions related to visual content. This involves understanding and combining multimodal representations, specifically the features extracted from images and text-based questions. The model learns these representations by integrating visual features from images and linguistic features from questions.



VQA models indeed have a wide range of applications across different domains as follows: (1) Healthcare: VQA models are used to assist in medical diagnosis by analyzing medical images (like X-rays, and MRIs) and answering related questions [4]. This can help doctors make more accurate diagnoses and provide quick responses to patient inquiries. (2) Education: VQA models can be employed as educational tools to help students learn about various subjects through visual content. For example, students can ask questions about a historical image or a scientific diagram, and the model can provide informative answers [5]. (3) Accessibility: These models can help visually impaired individuals by describing images' content when they ask questions about visual scenes. This improves accessibility to visual information for those with vision impairments [6]. (4) Retail and E-commerce: VQA models can be used to enhance customer experience by answering questions related to product images, such as the color, size, or availability of an item in an online store [7]. (5) Security and Surveillance: In security applications, VQA models can be used to analyze surveillance footage and answer questions regarding the activities or objects detected in the video, helping in threat detection and incident analysis [8]. (6) Autonomous Vehicles: In the context of self-driving cars, VQA models can assist by analyzing images from the vehicle's cameras and answering questions related to road conditions, nearby objects, or potential hazards [9]. (7) Entertainment: In interactive media and gaming, VQA models can be used to create more immersive experiences by allowing users to ask questions about the visual content they are interacting with, such as scenes in a video game or movie [10].

The Visual Quality Assessment (VQA) is tough work since it incorporates a variety of vision-related activities, including object detection, scene detection, object counting, color detection, object segmentation, and a great deal more. This work has gotten much simpler because of the rapid advancements that have been made in deep learning models. Layer Normalization, a method for stabilizing and expediting training by the normalization of activations inside a layer, has been widely employed in natural language processing, particularly in Transformer-based models. Nonetheless, its application in VQA is still constrained. Layer Normalization may enhance learning efficiency and model robustness by stabilizing the training process and minimizing internal covariate shifts. Notwithstanding its demonstrated advantages in other areas, its direct implementation in VQA models remains insufficiently investigated. Despite the notable advancements of VQA systems, they continue to encounter substantial hurdles. Addressing obstacles such as biases, shallow learning, and challenges in reasoning and generalization necessitates additional research, particularly in investigating novel strategies like Layer Normalization to improve performance. Addressing these issues will advance the development of more resilient, generic, and reliable VQA systems [11,12].

Moreover, VQA is receiving a great amount of focus in both the academic and the practical environments. With an image and an input question, VQA can determine a textual response. These answers are associated with a particular object, color, and collection of characteristics that are present in that image. The extraction of the keywords of queries that are associated with the visual content of photographs is a simple process when done manually. To make an accurate prediction of a particular answer, this information of keywords and visual aspects is helpful. Due to unfortunate circumstances, the process that encompasses the extraction of features, the integration of relationships, and the prediction of answers ought to be accomplished automatically. As a result, the VQA is opposed to the idea of utilizing and combining both textual and visual characteristics in order to discover the answer [13].

In recent times, earlier methodologies have regarded the answer as a collection of possible words and sentences. The size of the vocabulary is a crucial factor in determining the success of the system's performance. As humans, we anticipate a system that can proficiently respond to all categories of inquiries. Unfortunately, no one can possess complete knowledge. Thus, it is impractical to construct a

system with an infinite vocabulary for providing answers. It is unrealistic to anticipate a system that can provide answers to every type of question. Given these rationales, we suggest refining this assignment by focusing on a certain sort of query. More precisely, we exclusively concentrate on Yes/No inquiries in a variety of our projects [14]. Typically, the VQA task datasets consist of questions that fall into three categories: Yes/No, Number, and Other. In an image retrieval system, a common query is to determine the presence of images and objects. It demonstrates that Yes/No questions are of interest in both research and application.

Vision Transformers represent a shift from traditional Convolutional Neural Networks (CNNs) to transformer-based architectures for image processing. Instead of using convolutions to extract features, ViTs break an image into patches, treat these patches as a sequence (like words in a sentence), and process them using the transformer architecture. Many research endeavors have effectively utilized Vision Transformers in the field of VQA [15–17], frequently integrating them with sophisticated language models and attention processes to develop cutting-edge VQA systems. These models have demonstrated encouraging outcomes, especially in situations that need comprehension of intricate settings and subtle inquiries.

This article’s contributions are as follows: (1) We investigate “Fusion” Models, which integrate information from the text encoder and picture encoder to carry out the visual question-answering task effectively. The text encoder can consist of a transformer model that operates on text, such as BERT, RoBERTa, and others. On the other hand, the image encoder can be a transformer model specifically designed for processing images, such as ViT, Deit, BeIT, and others. (2) We modified the reasoning module and added layer normalization to improve the performance result of VQA. (3) We conducted a thorough analysis and evaluation of the outcomes of our studies.

The remainder of this paper is arranged as follows. [Section 2](#) examines the related works. [Section 3](#) summarizes the methodology. [Section 4](#) discusses and describes the experiments and results obtained for the VQA. [Section 5](#) discusses the conclusions of the research paper and future research.

2 Related Works

2.1 Vision Transformer

Vision Transformers represent a shift from traditional Convolutional Neural Networks (CNNs) to transformer-based architectures for image processing. Instead of using convolutions to extract features, ViTs break an image into patches, treat these patches as a sequence (like words in a sentence), and process them using the transformer architecture.

Text-based transformer models, such as Bidirectional Encoder Representations from Transformers (BERT) [18–20], Robustly Optimized BERT Pretraining Approach (RoBERTa) [21], and others, have revolutionized the field of Natural Language Processing (NLP) by providing state-of-the-art performance on a wide range of language tasks. The transformer architecture forms the foundation of models like BERT and RoBERTa. It relies on the self-attention mechanism, which allows the model to weigh the importance of different words in a sentence when encoding their meaning. BERT was one of the first models to apply the transformer architecture in a bidirectional way, meaning it considers both the left and right context of a word simultaneously during training. RoBERTa is an optimized version of BERT that improves its training process and effectiveness [22]. Text-based transformer models like BERT and RoBERTa have set new benchmarks in NLP by providing deep contextual

understanding of language. Their ability to be pre-trained on massive amounts of data and then fine-tuned for specific tasks makes them versatile and powerful tools across various applications in natural language processing.

A transformer model that is specifically created to process images can serve as the image encoder in a multimodal model, such as the one that is utilized for Visual Question Answering (VQA). The Vision Transformer (ViT) [23], the Data-efficient Image Transformer (DeiT) [24], and the Bidirectional Encoder representation from Image Transformers (BEiT) are three of the most well-known image encoders that are generated using transformers [25].

2.2 Visual Questions Answering (VQA) and Layer Normalization

In recent years, there has been a growing interest in VQA in the general domain due to the success of deep learning. Significant advancements have been achieved, primarily through the utilization of deep Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) pipelines. Many works utilize various recurrent neural networks (RNNs), including LSTM [26], word2vec [27], and GloVe [28], to embed text. These RNNs are employed to record word relationships and extract textual characteristics. CNN architectures, including VGG16 [29], ResNet [30], and Faster R-CNN [31], were utilized to extract both low- and high-level visual signals. Subsequently, the feature representations were effectively merged to respond to the query question utilizing sophisticated attention techniques as stacking attention networks (SANs) [32].

Conventional VQA methods involve the integration of recurrent and convolutional neural networks. Typically, many methods employ a recurrent network, such as Long Short-term Memory (LSTM) [33] or Gated Recurrent Unit (GRU), to extract the textual properties. Furthermore, convolution neural networks are utilized to insert pictures [34]. Ultimately, it employs vector operations such as element-wise multiplication, concatenation, and other similar techniques to merge textual and visual characteristics to make predictions about answers [35].

The image encoder is a crucial component of VQA systems. The image encoder is tasked with extracting visual information from the input image. Historically, Convolutional Neural Networks (CNNs) such as ResNet or VGG were employed for this objective. However, newer methods utilize transformer-based models specifically developed for image processing, such as Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), or BEiT. The image encoder produces a collection of visual feature representations that capture various elements of the image, including objects, scenes, and spatial connections. Text Encoder: The question encoder analyzes the natural language question to extract semantic characteristics. Transformer-based models like as BERT, RoBERTa, or GPT are frequently employed for text encoding. The encoder converts the question into a vector representation that accurately captures the semantic and contextual information of the words in the query [36].

Multimodal Fusion refers to the process of combining information from multiple modes or sources into a unified representation [37]. After extracting the visual and textual components, it is necessary to integrate them coherently. This process is referred to as multimodal fusion. Various methodologies can be employed for fusion, including straightforward concatenation, attention processes, or more intricate approaches like bilinear pooling. The objective is to combine visual and textual information to create a unified representation that may be utilized to deduce the solution. Anticipated response: The combined representation is fed into either a classification layer or a decoder to provide the answer. The response may consist of a single word, a phrase, or even a complete sentence, contingent upon the intricacy of the assignment. The model is commonly trained in an end-to-end manner, where it optimizes for the correct answer by employing a loss function like cross-entropy.

Layer Normalization, often known as LN, is a method that is intended to stabilize and speed up the training of deep neural networks. It accomplishes this by normalizing the outputs of each layer across the features. Layer normalization, in contrast to batch normalization, normalizes across the batch dimension and the feature dimensions. This makes it particularly well-suited for models such as BERT and Vision Transformers (ViT), which have input shapes that can vary. The Layer Normalization technique is an essential component of BERT (Bidirectional Encoder Representations from Transformers) to ensure the stability of the training of the Transformer layers. These Transformer layers are made up of multi-head self-attention and feed-forward sub-layers. Both the self-attention mechanism and the feed-forward network are followed by applying LN. Moreover, LN standardizes the mean and variance of the aggregated inputs to the neurons within a single layer. In contrast to BatchNorm, which is contingent upon the dimensions of a mini-batch, LN exhibits fewer constraints. LN is adaptable to RNN and self-attention-based models. It has been implemented in advanced frameworks such as Transformer, BERT, and Transformer-X. LN enhances performance and is indispensable in these systems.

$$N(\mathbf{x}) = \frac{\mathbf{x} - \mu}{\sigma} \quad \mu = \frac{1}{H} \sum_{i=1}^H x_i \quad \sigma = \sqrt{\frac{1}{H} \sum_{i=1}^H (x_i - \mu)^2} \quad (1)$$

The Layer Normalization can be calculated in Eq. (1). N is the Normalization function, and $\mathbf{x} = (x_1, x_2, \dots, x_H)$ is the representation of input vector with dimension H . H denotes the number of hidden units in a layer. μ and σ are the mean and standard deviation of input [38,39].

VQA is a sophisticated computer vision task that entails supplying an algorithm with a natural language query associated with a picture, necessitating the generation of a natural language response for the specific question-image pair. Recently, various VQA models have been introduced to tackle this difficulty. The principal performance metric employed to assess these models is accuracy. The research community has started to recognize that accuracy alone is an inadequate criterion for evaluating model performance. Besides accuracy, models must also exhibit robustness, indicating that their output should be largely unaffected by slight perturbations or noise introduced to the input. This encompasses substituting words with synonymous terms, phrases, or sentences in input inquiries, or making minor adjustments to pixel values in the image [40]. Furthermore, an early VQA initiative, the DAQUAR dataset, innovates the use of soft evaluation for VQA. They employ a variant of the Wu-Palmer similarity on a lexical database to calculate a soft prediction score. Consequently, a forecast that is semantically akin to the target answer is no longer deemed false. This facilitates a more nuanced assessment of performance in VQA. Nevertheless, it is utilized solely for assessment, and this metric possesses inherent limitations, including its incapacity to differentiate colors [41].

A further instance of study that integrates both metrics based on the researcher [42]. During the evaluation process, they utilize the accuracy measure, defined as the ratio of properly anticipated responses to the total number of responses. Since responses are generated using open-ended processes utilizing LLMs and may exhibit differences, their research does not require a perfect correspondence between the prediction and the actual outcome. Their work assesses semantic similarity by cosine similarity in a vector space, with a threshold of 0.70. If two strings exhibit a high degree of semantic similarity, the prediction is deemed accurate; for instance, recognizing “couch” as valid for the label “sofa.”

2.3 Performance Evaluation (Wu & Palmer Similarity)

Performance metrics in Visual Question Answering (VQA) are crucial for evaluating how well a model understands and responds to questions about images. One such metric is **Wu & Palmer similarity**, which is used to assess the semantic similarity between words or phrases. It can be applied in VQA to evaluate how closely a model's generated answer aligns with the expected answer in terms of meaning, rather than just exact word matching. Wu & Palmer similarity is a semantic similarity measure based on the taxonomy structure of words in a lexical database like WordNet. It calculates the similarity between two words by considering the depths of the words in the taxonomy and the depth of their most specific common ancestor (also known as the Least Common Subsumer, LCS). Eq. (2) explains the Wu & Palmer similarity calculations [43,44].

$$\text{SimWP}_{XY} = 2 \times \text{depth}(N) / (\text{depth}(N1) + \text{depth}(N2)) \quad (2)$$

This means that $0 < \text{SimWP}_{XY} \leq 1$. The score can never be zero because the depth of the LCS is never zero (the depth of the root of a taxonomy is one). The score is one if the two input concepts are the same. Where SimWP is Wu & Palmer similarity, $N1$ and $N2$ are the number of arcs between the concepts X , Y and the ontology root R and N is the number of arcs between the LCS and the ontology root R as shown on Fig. 1 [43].

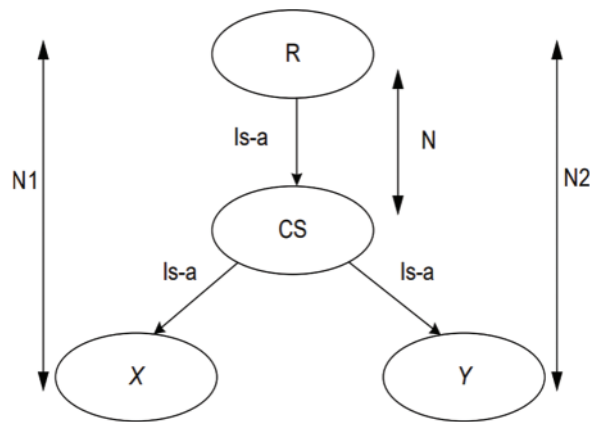


Figure 1: Wu & Palmer's ontology example

A quantitative metric that was developed expressly for the purpose of determining the degree of semantic similarity between two given words or sentences is known as the Wu & Palmer similarity. Because single-word answers are the primary focus of our assignment, the Wu & Palmer similarity metric is a good choice for evaluating potential solutions. On the other hand, due to the constraints introduced by its design, it might not be appropriate for phrases or sentences. A variant of the Wu & Palmer similarity score that is constructed using the WordNet taxonomy is made available by the Natural Language Toolkit (NLTK).

3 Methodology

3.1 Research Workflow

Fig. 2a shows the VQA architecture default model that consists of several key components: (1) Image Encoder: Extract relevant features from the input image. Traditional image encoders use CNNs (e.g., ResNet, VGG) to process the image and extract feature maps. These features capture

information about objects, scenes, and spatial relationships. More recent approaches use vision transformers (e.g., ViT, DeiT, BEiT) to process images [45,46]. ViTs divide images into patches, embed these patches, and use self-attention mechanisms to capture global context and relationships between patches. (2) Text/Question Encoder used to process and encode the textual question into a vector representation. Text encoders typically use transformer-based models like BERT, RoBERTa, or GPT [47]. These models are pre-trained on large text corpora and fine-tuned for the VQA task, allowing them to capture the semantic meaning and context of the question [48]. (3) Reasoning Module: Combine the visual and textual features to create a unified representation that incorporates both modalities. One straightforward method is to concatenate the visual and textual features into a single vector. This approach may be followed by fully connected layers to learn a joint representation. More advanced methods use attention mechanisms to align and integrate visual and textual features. Cross-attention mechanisms allow the model to focus on relevant parts of the image when processing the question and vice versa. This method helps in effectively merging the information from both modalities. Some architectures use fusion layers that specifically combine visual and textual features. Techniques such as bilinear pooling or co-attention networks can be used to enhance the integration [49]. (4) Answer Decoder will generate the answer based on the combined representation of the visual and textual features [50,51].

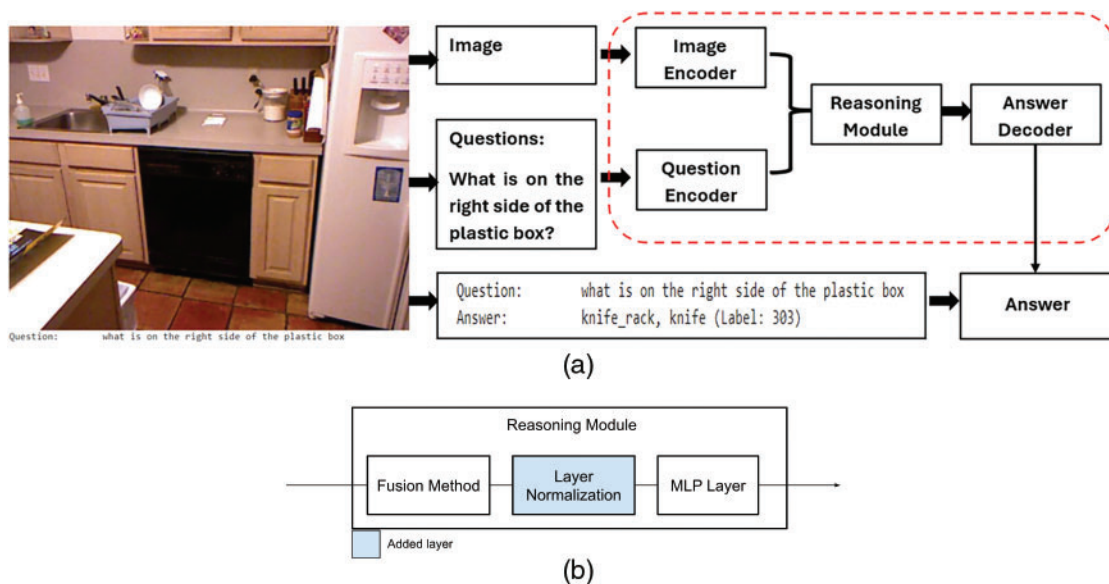


Figure 2: VQA architecture. (a) Default model, and (b) Modified reasoning module

Further, Fig. 2b illustrates our proposed method, which is a modified reasoning module. Additionally, our work includes the addition of layer normalizations to enhance the results of the experiment. Layer normalization is a technique used to stabilize and accelerate training by normalizing the inputs of each layer to have zero mean and unit variance. In the context of VQA layer normalization can be particularly beneficial when used in conjunction with Multilayer Perceptrons (MLPs) for fusion. Layer normalization helps to standardize the inputs to each MLP layer, which can prevent issues related to vanishing or exploding gradients. This is crucial in VQA where complex models with multiple layers and modalities are used. By normalizing the activations, layer normalization ensures that the distribution of activations remains stable throughout training, leading to more consistent and reliable

learning. Models with layer normalization often converge faster during training. This is beneficial for VQA models, which can be large and require significant training time.

The modified reasoning module consists of Layer Normalization added between the Fusion Method and MLP Layer. The Fusion Method combines visual and textual features into an input vector representation. Vector \mathbf{x}^{tv} represents the concatenated features, where $\mathbf{x}^{tv} = (x_1, x_2, \dots, x_{H_t+H_v}) \in \mathbb{R}^{H_t+H_v}$. H_t indicates the number of hidden units in the text encoder, such as BERT and RoBERTa, whereas H_v denotes the image encoder, such as ViT, DeiT, and BeiT. The additional layer normalizes the combined features using Layer Normalization, similar to Eqs. (3) and (4) [38,39].

$$N(\mathbf{x}^{tv}) = \frac{\mathbf{x}^{tv} - \mu^{tv}}{\sigma^{tv}} \quad \mu^{tv} = \frac{1}{H_t + H_v} \sum_{i=1}^{H_t+H_v} x_i \quad \sigma^{tv} = \sqrt{\frac{1}{H_t + H_v} \sum_{i=1}^{H_t+H_v} (x_i - \mu^{tv})^2} \quad (3)$$

$$\mathbf{h}^{tv} = \mathbf{g} \odot N(\mathbf{x}^{tv}) + \mathbf{b} \quad (4)$$

where \mathbf{h}^{tv} is the output of the Layer Normalization, with \mathbf{b} and \mathbf{g} are defined as the bias and gain parameters of the same dimension as $H_t + H_v$. \odot is a dot production operation for multiplication between two vectors.

3.2 Dataset

3.2.1 PathVQA

PathVQA is a collection of question-answer pairs specifically designed for pathology images. The purpose of this dataset is to be utilized for training and evaluating Visual Question Answering (VQA) systems specifically in the field of medical imaging. The dataset includes open-ended and binary “yes/no” inquiries. The dataset is derived from two publicly accessible pathology textbooks, namely “Textbook of Pathology” and “Basic Pathology”, as well as a publicly accessible digital library called “Pathology Education Informational Resource” (PEIR) [52,53]. The purpose of its creation was to assess AI models’ proficiency in responding to medical imagery inquiries, namely pathology slides. The collection comprises medical photos accompanied by associated inquiries and responses, encompassing many subjects such as disease diagnosis, identification of specific tissue structures, and others. The dataset at hand comprises a total of 5004 images and 32,795 question-answer pairings. Among the total of 5004 photographs, 4289 images are associated with a question-answer pair, while 715 images remain unused. Multiple instances of image-question-answer triplets appear more than once in the same split (training, validation, test, source: <https://huggingface.co/datasets/flaviagiammarino/path-vqa>) (accessed on 31 October 2024). After removing any duplicate image-question-answer combinations, the dataset consists of 32,632 question-answer pairings related to 4289 unique photos. Due to the complex nature of medical imaging and the need for specialist knowledge in the field to provide accurate responses, the PathVQA dataset poses a difficult task. This can be seen in Fig. 3, which illustrates the situation. Consequently, this attribute renders it a highly important asset for developing and experimenting with artificial intelligence models in healthcare and medical research.

3.2.2 VIZWIZ Dataset

The VizWiz dataset is a significant resource in the field of Visual Question Answering (VQA), particularly designed to address real-world challenges and support the development of assistive technologies for visually impaired users. The images and questions in the VizWiz dataset were collected

from real users who are blind or visually impaired. Users took photos using their smartphones and asked questions about the content of these images. This makes the dataset unique compared to other VQA datasets, which are typically curated and consist of high-quality, well-framed images. The dataset was introduced as part of the VizWiz project, which focuses on helping visually impaired people by allowing them to ask questions about images they capture in their daily lives (<https://vizwiz.org/tasks-and-datasets/vqa/>) (accessed on 31 October 2024). The dataset has 957 train-test images, with 195 of them labeled as “yes/no”. The VizWiz dataset consists of two classes: “yes” with 553 images and “no” with 599 images. For this study, we have selected the VizWiz-VQA 2023 edition and categorized it based on “yes/no” questions and responses [54,55].




	is gastrointestinal present?	yes
	is hyperplasia without atypia characterized by nests of...	yes
	is mucoepidermoid carcinoma present?	no

Figure 3: PathVQA dataset sample

A realistic and rigorous benchmark for the development of assistive technology is provided by the VizWiz dataset, which plays an important role in the research and development community of the VQA. Because of its emphasis on real-world applications and the inclusion of photos and queries from visually impaired users, it is an extremely significant resource for academics who are working toward the development of artificial intelligence systems that can have a discernible and beneficial effect on the lives of individuals. Fig. 4 shows the VizWiz dataset example.

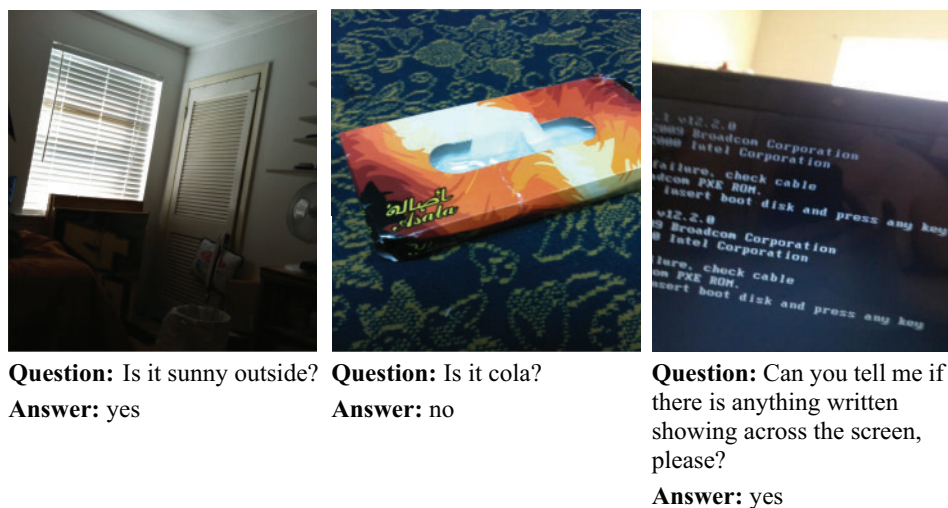


Figure 4: VizWiz dataset example

4 Experiment Results and Discussions

4.1 Experiment Results

Our experiment setting parameter can be seen in Table 1. We set the specified initial value for the random number generator to ensure consistent results when reproducing an experiment. Guarantees

consistent outcomes each time the code is executed, seed=12345. Our experiment will evaluate, log, and save every 100 steps. We restrict the maximum number of saved checkpoints, allowing only 3 to be stored. We retain only the most recent checkpoints and remove any older ones. In the context of model training, using the Wu and Palmer (WUP) Score as a metric involves selecting the best model based on how well it captures semantic similarity between word pairs according to this measure. When the evaluation method WUP parameter is set, the training process will evaluate the model's performance based on the Wu and Palmer Score. The model that achieves the highest or lowest (depending on the specific implementation and objective) WUP Score on the validation set during training is considered the best. In many cases, a higher Wu and Palmer Score indicates better performance, meaning the model has successfully learned to predict word pairs that are semantically similar according to their positions in the taxonomy.

Table 1: Experiment setting parameter

Parameters	Explanations
seed=12345	Sets the random seed for reproducibility. Ensures that the results are the same every time the code is run
evaluation_strategy="epoch"	Evaluation strategy: "steps" or "epoch"
eval_steps=100	Evaluate every 100 steps
logging_strategy="epoch"	Logging strategy: "steps" or "epoch"
logging_steps=100	Log every 100 steps
save_strategy="epoch"	Saving strategy: "steps" or "epoch"
save_steps=100	Save every 100 steps
save_total_limit=3	Limits the total number of saved checkpoints to 3, keeping only the most recent ones and deleting older checkpoints.
metric_for_best_model="wups"	Metric used for determining the best model
per_device_train_batch_size=32	Batch size per GPU for training
per_device_eval_batch_size=32	Batch size per GPU for evaluation
remove_unused_columns=False	Whether to remove unused columns in the dataset
num_train_epochs=50	Number of training epochs
fp16=True	Enable mixed precision training (float16)
dataloader_num_workers=8	Number of workers for data loading
load_best_model_at_end=True	Whether to load the best model at the end of training

Fig. 5 exhibits the training process of Bert Vit 50 Epoch VizWiz dataset. The model is assessed after each epoch using the assessment approach of "epoch" to monitor its performance. Performance measures like loss and accuracy are recorded at consistent intervals, such as every epoch, to track the success of training. The model is stored after every epoch using the "epoch" save approach. Storage efficiency is ensured by retaining only the most recent three checkpoints. The model with the greatest WUP score (or the selected measure) on the validation set is picked as the final model after 50 epochs. The purpose of this training method is to ensure that the BERT-ViT model acquires the ability to seamlessly incorporate visual and textual information, resulting in precise and contextually suitable responses within the VizWiz dataset.

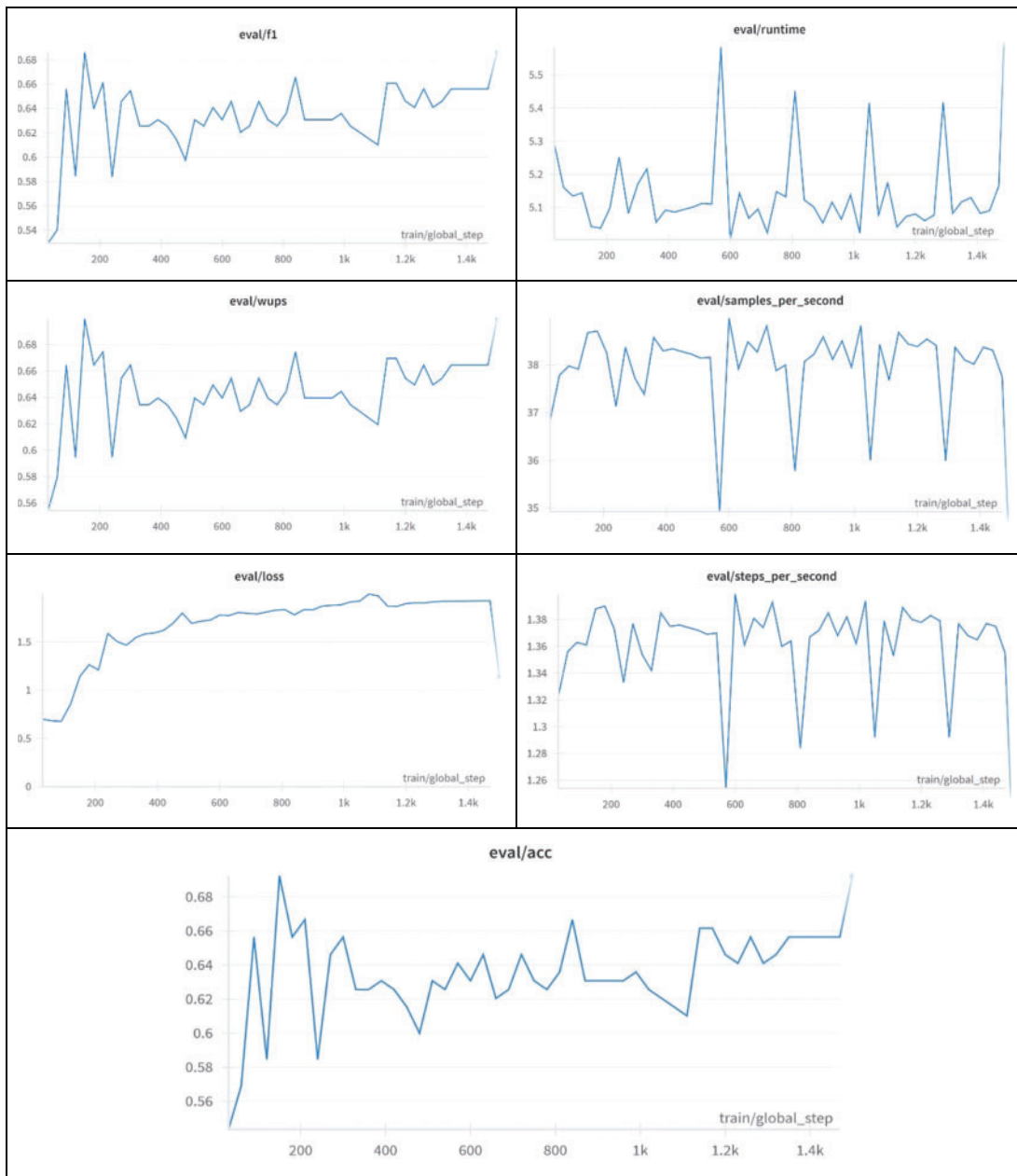


Figure 5: Training bert ViT 50 epoch VizWiz dataset

Table 2 displays the experimental findings using the PathVQA and VizWiz datasets, specifically with the default model. We utilize the BERT, and RoBERTa, for the text transformation and ViT, DeiT, and BeiT models for image transformation. Each training session in the experiment consists of 50 epochs. Based on the experiment results for the PathVQA dataset, BERT and ViT achieved the highest Wu & Palmer Score of 0.601, Accuracy of 0.595, and 198 M number of trainable parameters. Similarly, BERT and ViT, after training for 50 epochs, attain a maximum Wu & Palmer Score of 0.695, an Accuracy of 0.687, and have a total of 196 million trainable parameters for the VizWiz dataset.

Upon determining the optimal combination of experiment 1 with the default model, we observe that BERT and ViT exhibit the highest level of performance for both datasets.

Table 2: Experiment results with PathVQA and VizWiz dataset (Default model)

Dataset	Text transformer	Image transformer	Epoch	Wu & Palmer Score	Accuracy	F1	No. of trainable parameters
PathVQA	BERT	ViT	50	0.601	0.595	0.042	198M
PathVQA	BERT	DeiT	50	0.591	0.587	0.039	198M
PathVQA	BERT	BeiT	50	0.584	0.578	0.039	198M
PathVQA	RoBERTa	ViT	50	0.601	0.595	0.041	213M
PathVQA	RoBERTa	DeiT	50	0.589	0.583	0.037	213M
PathVQA	RoBERTa	BEiT	50	0.596	0.588	0.041	213M
VizWiz	BERT	ViT	50	0.695	0.687	0.68	196M
VizWiz	BERT	DeiT	50	0.664	0.656	0.636	196M
VizWiz	BERT	BeiT	50	0.654	0.646	0.644	196M
VizWiz	RoBERTa	ViT	50	0.664	0.656	0.623	211M
VizWiz	RoBERTa	DeiT	50	0.634	0.626	0.611	211M
VizWiz	RoBERTa	BEiT	50	0.604	0.595	0.593	211M

We are currently implementing our proposed model by modifying the reasoning module to include the normalization layers mentioned in Table 3. The model we propose can enhance the performance of both datasets in the experiments. The Wu & Palmer Score for the PathVQA dataset increased from 0.601 to 0.609, while accuracy improved from 0.595 to 0.604. In addition, the VizWiz dataset experienced an enhancement in Wu & Palmer Score, increasing from 0.695 to 0.70, as well as an improvement in accuracy, rising from 0.687 to 0.692.

Table 3: Experiment results with PathVQA and VizWiz dataset (Modified reasoning module)

Dataset	Text transformer	Image transformer	Epoch	Wu & Palmer Score	Accuracy	F1	No. of trainable parameters
PathVQA	BERT	ViT	50	0.609	0.604	0.048	198M
PathVQA	BERT	DeiT	50	0.603	0.596	0.046	198M
PathVQA	BERT	BeiT	50	0.589	0.583	0.040	198M
PathVQA	RoBERTa	ViT	50	0.620	0.614	0.051	213M
PathVQA	RoBERTa	DeiT	50	0.613	0.606	0.045	213M
PathVQA	RoBERTa	BEiT	50	0.597	0.590	0.045	213M
VizWiz	BERT	ViT	50	0.700	0.692	0.686	196M
VizWiz	BERT	DeiT	50	0.659	0.651	0.646	196M
VizWiz	BERT	BeiT	50	0.654	0.646	0.644	196M

(Continued)

Table 3 (continued)


Dataset	Text transformer	Image transformer	Epoch	Wu & Palmer Score	Accuracy	F1	No. of trainable parameters
VizWiz	RoBERTa	ViT	50	0.659	0.651	0.646	211M
VizWiz	RoBERTa	DeiT	50	0.644	0.636	0.606	211M
VizWiz	RoBERTa	BEiT	50	0.644	0.636	0.623	211M

4.2 Discussions

Layer normalization is an essential approach in the training of deep learning models, especially for tasks such as Visual Question Answering (VQA). The following are the primary advantages of implementing layer normalization in VQA training: (1) Layer normalization aids in stabilizing the training process by mitigating the internal covariate shift, which refers to the alteration of input distribution to each layer during training. This stabilization results in a higher degree of consistency in learning, enabling the model to converge more rapidly and consistently. (2) By normalizing the inputs of each layer, it helps in maintaining gradients within a reasonable range, preventing issues like exploding or vanishing gradients that can otherwise hinder learning. (3) Layer normalization can act as a form of regularization, reducing the risk of overfitting. This is particularly beneficial in VQA, where models must generalize well to unseen image-question pairs. (4) Layer normalization is particularly beneficial in deep architectures like transformers used in VQA, where the model depth can lead to unstable training. It ensures that the activations in deeper layers remain well-behaved, leading to better overall performance.

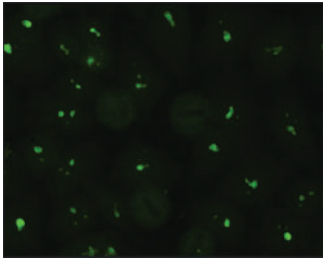

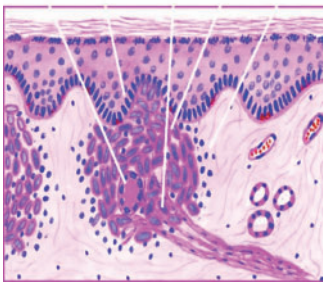
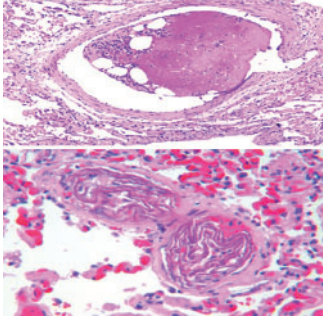
Table 4 shows the PathVQA dataset testing result, as we can see our proposed model BERT+ViT (Modified Reasoning Module) can improve the predicted answer from the default models and improve the similarity score. In Table 4, image number 1, the default model predicted response number “no” with a similarity score of 0.2105. However, our proposed model accurately predicted answer number “yes” with a similarity value of 1.0. Image number 4 was associated with the anticipated response “endocrine” with a similarity score of 0.285. On the other hand, our proposed model predicted the answer “joints” with a similarity score of 1.1.

Table 4: PathVQA dataset testing result

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
1		<p>Question: Is the heart heavier? Answer: yes (Label: 4089) Predicted answer: no Similarity: 0.21052631578947367</p>	<p>Question: Is the heart heavier? Answer: yes (Label: 4089) Predicted answer: yes Similarity: 1.0</p>

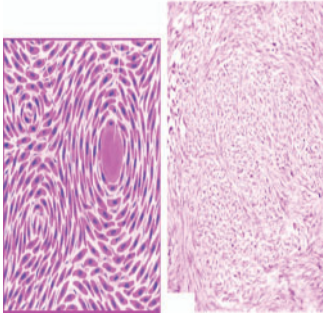
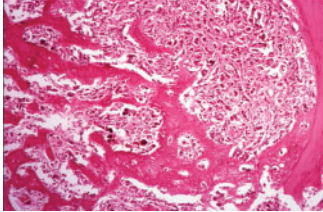
(Continued)

Table 4 (continued)

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
2		<p>Question: Is a nucleolar pattern typical of antibodies against nucleolar proteins? Answer: yes (Label: 4089) Predicted answer: no Similarity: 0.21052631578947367</p>	<p>Question: Is a nucleolar pattern typical of antibodies against nucleolar proteins? Answer: yes (Label: 4089) Predicted answer: yes Similarity: 1.0</p>
3		<p>Question: What is present? Answer: lung (Label: 2096) Predicted answer: cardiovascular Similarity: 0.0</p>	<p>Question: What is present? Answer: lung (Label: 2096) Predicted answer: joints Similarity: 0.667</p>
4		<p>Question: What is composed of epithelioid cells with sparse Langhans' giant cells and lymphocytes? Answer: granuloma (Label: 1589) Predicted answer: epithelial Similarity: 0.0</p>	<p>Question: What is composed of epithelioid cells with sparse Langhans' giant cells and lymphocytes? Answer: granuloma (Label: 1589) Predicted answer: strictures Similarity: 0.728</p>
5		<p>Question: How many small pulmonary arterioles are packed with laminated swirls of fetal squamous cells? Answer: two (Label: 3647) Predicted answer: testicular teratomas Similarity: 0</p>	<p>Question: How many small pulmonary arterioles are packed with laminated swirls of fetal squamous cells? Answer: two (Label: 3647) Predicted answer: two Similarity: 1.0</p>


(Continued)

Table 4 (continued)

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
6		<p>Question: What shows whorls of smooth muscle cells which are spindle-shaped, having abundant cytoplasm and oval nuclei?</p> <p>Answer: microscopy (Label: 2248)</p> <p>Predicted answer: stroma</p> <p>Similarity: 0.111</p>	<p>Question: What shows whorls of smooth muscle cells which are spindle-shaped, having abundant cytoplasm and oval nuclei?</p> <p>Answer: microscopy (Label: 2248)</p> <p>Predicted answer: microscopy</p> <p>Similarity: 1.0</p>
7		<p>Question: What is present?</p> <p>Answer: joints (Label: 1922)</p> <p>Predicted answer: endocrine</p> <p>Similarity: 0.285</p>	<p>Question: What is present?</p> <p>Answer: joints (Label: 1922)</p> <p>Predicted answer: joints</p> <p>Similarity: 1.0</p>

The testing results of the VizWiz dataset are displayed in [Table 5](#). In image number 1 and 2, both the default model and our suggested model achieve perfect accuracy, correctly predicting all the answers with a similarity score of 1.0. For the third image, the default models predicted a negative answer with a similarity score of 0.2105. However, our proposed method accurately detected a positive answer with a similarity score of 1.0.

Table 5: VizWiz dataset testing result

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
1		<p>Question: Is it sunny outside?</p> <p>Answer: yes (Label:1)</p> <p>Predicted answer: yes</p> <p>Similarity: 1.0</p>	<p>Question: Is it sunny outside?</p> <p>Answer: yes (Label:1)</p> <p>Predicted answer: yes</p> <p>Similarity: 1.0</p>




(Continued)

Table 5 (continued)

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
2		<p>Question: It's it cola? Answer: no (Label: 0) Predicted answer: no Similarity: 1.0</p>	<p>Question: It's it cola? Answer: no (Label: 0) Predicted answer: no Similarity: 1.0</p>
3		<p>Question: Can you tell me if there is anything written showing across the screen please? Answer: yes (Label:1) Predicted answer: no Similarity: 0.21052631578947367</p>	<p>Question: Can you tell me if there is anything written showing across the screen please? Answer: yes (Label:1) Predicted answer: yes Similarity: 1.0</p>
4		<p>Question: this is a picture of a home i'm planning to buy, does it look like its in a safe neighborhood? Answer: yes (Label:1) Predicted answer: no Similarity: 0.21052631578947367</p>	<p>Question: this is a picture of a home i'm planning to buy, does it look like its in a safe neighborhood? Answer: yes (Label:1) Predicted answer: yes Similarity: 1.0</p>

(Continued)

Table 5 (continued)

No.	Images	BERT+ViT (Default model)	BERT+ViT (Modified reasoning module)
5		<p>Question: Can you see anything about these articles of clothing that does not match?</p> <p>Answer: no (Label: 0)</p> <p>Predicted answer: yes</p> <p>Similarity: 0.210</p>	<p>Question: Can you see anything about these articles of clothing that does not match?</p> <p>Answer: no (Label: 0)</p> <p>Predicted answer: no</p> <p>Similarity: 1.0</p>
6		<p>Question: Does Buddy's face look cute or not?</p> <p>Answer: yes (Label: 1)</p> <p>Predicted answer: no</p> <p>Similarity: 0.210</p>	<p>Question: Does Buddy's face look cute or not?</p> <p>Answer: yes (Label: 1)</p> <p>Predicted answer: yes</p> <p>Similarity: 1.0</p>
7		<p>Question: Does this need to be refrigerated?</p> <p>Answer: no (Label: 0)</p> <p>Predicted answer: yes</p> <p>Similarity: 0.210</p>	<p>Question: Does this need to be refrigerated?</p> <p>Answer: no (Label: 0)</p> <p>Predicted answer: no</p> <p>Similarity: 1.0</p>

Tables 6 and 7 present a performance comparison between the VQA and past study findings. Our proposed method demonstrates a significant enhancement in performance compared to earlier research results, achieving a 60.4% improvement with the PathVQA Dataset and a 69.2% improvement with the VizWiz datasets. BioMedLM [56] is a language model based on GPT2 that has been specifically trained on collections of biological texts. They demonstrate exceptional performance compared to their general counterparts in specialized biological language tasks, such as question answering or relation extraction. Their experiment attains a mere 57.2% accuracy when applied to PathVQA datasets. Furthermore, the Up Down [57] the technique achieved just 59.6% accuracy when applied to the VizWiz dataset.

Table 6: Performance comparison with previous research results with path VQA dataset

Method	Dataset	Accuracy (%)
MEVF [58]	PathVQA	44.8
MMQ [59]	PathVQA	47.1
VQAMix [60]	PathVQA	48.6
AMAM [61]	PathVQA	50.4
BioGPT LoRa [56]	PathVQA	47.9
BioMedLM LoRa [56]	PathVQA	57.2
Proposed method	PathVQA	60.4

Table 7: Performance comparison with previous research results with VizWiz dataset

Method	Dataset	Accuracy (%)
CNN+LSTM [62]	VizWiz	54
FT+VQA [63]	VizWiz	68.1
Up Down [57]	VizWiz	59.6
Proposed method	VizWiz	69.2

5 Conclusions

This paper examines “Fusion” Models, which use information from both the text encoder and picture encoder to efficiently perform the visual question-answering task. A text encoder typically comprises a transformer model, such as BERT, RoBERTa, or similar architectures, which processes textual data. Alternatively, the image encoder can be a transformer model that is explicitly tailored for image processing, such as ViT, DeiT, BeiT, and other similar models. To enhance the performance result of VQA, we made modifications to the reasoning module and implemented layer normalization. To thoroughly analyze and evaluate the results of our studies, we carried out a comprehensive study. The findings of our experiments indicate that a modified reasoning module has the potential to improve the performance outcomes associated with both the PathVQA and VizWiz datasets.

By normalizing the inputs inside each layer, layer normalization makes it possible for the model to converge more quickly, which in turn reduces the amount of time required for training overall. During the process of training huge models on enormous datasets, such as PathVQA and VizWiz, this is especially useful. The stabilization of learning, the enhancement of generalization, and the facilitation of the integration of multimodal data are all areas in which layer normalization plays a crucial role in the transformation of the training process for VQA models. These advantages result in models that are more robust and efficient, which, in the end, leads to an improvement in the performance of VQA systems on tasks that are similar to those found in both datasets. Our proposed method indicates a significant performance improvement when compared to the findings of past research. Specifically, we achieved a 60.4% improvement with the PathVQA Dataset and a 69.2% improvement with the VizWiz datasets.

A fresh opportunity to explore the practical side of research to create and improve VQA models that are more suited and compact in the future has been opened up as a result of this success. We will

explore other VQA datasets, combine them with Explainable Artificial Intelligence in our future study, and develop more effective techniques to create joint embedding spaces that seamlessly integrate visual and textual information.

Acknowledgement: This research is supported by the Vice-Rector of Research, Innovation, and Entrepreneurship at Satya Wacana Christian University.

Funding Statement: This paper is supported by the National Science and Technology Council, Taiwan (Grant number: NSTC 111-2637-H-324-001-).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Christine Dewi; Abbott Po Shun Chen; Hanna Prillysca Chernovita; data collection: Stephen Abednego Philemon; Christian Adi Ananta; analysis and interpretation of results: Christian Adi Ananta; Christine Dewi; Hanna Prillysca Chernovita; Abbott Po Shun Chen; draft manuscript preparation: Abbott Po Shun Chen; Christine Dewi; Hanna Prillysca Chernovita; Stephen Abednego Philemon. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

- [1] T. Kunlamai, T. Yamane, M. Suganuma, P. J. Chun, and T. Okatani, "Improving visual question answering for bridge inspection by pre-training with external data of image-text pairs," *Comput. Civ. Infrastruct. Eng.*, vol. 39, no. 3, pp. 345–361, 2024. doi: [10.1111/mice.13086](https://doi.org/10.1111/mice.13086).
- [2] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu and J. Gao, "Vision-language pre-training: Basics, recent advances, and future trends," *Found Trends Comput. Graph. Vis.*, vol. 14, no. 3–4, pp. 163–352, 2022. doi: [10.1561/0600000105](https://doi.org/10.1561/0600000105).
- [3] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick and A. Van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vis. Image Underst.*, vol. 163, pp. 21–40, 2017. doi: [10.1016/j.cviu.2017.05.001](https://doi.org/10.1016/j.cviu.2017.05.001).
- [4] J. Y. I. Melvin, S. Gawade, and M. Shrimali, "Novel approach to integrate various feature extraction techniques for the visual question answering system with skeletal images in the healthcare sector," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 9, pp. 4138–4145, 2023. doi: [10.17762/ijritcc.v11i9.9781](https://doi.org/10.17762/ijritcc.v11i9.9781).
- [5] Y. Cheng, "Application of a neural network-based visual question answering system in preschool language education," *IEIE Trans. Smart Process. Comput.*, vol. 12, no. 5, pp. 419–427, 2023. doi: [10.5573/IEIESPC.2023.12.5.419](https://doi.org/10.5573/IEIESPC.2023.12.5.419).
- [6] K. Ramnath, L. Sari, M. Hasegawa-Johnson, and C. Yoo, "Worldly Wise (WoW)-cross-lingual knowledge fusion for fact-based visual spoken-question answering," in *NAACL-HLT 2021–2021 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol.*, 2021, pp. 1908–1919. doi: [10.18653/v1/2021.naacl-main.153](https://doi.org/10.18653/v1/2021.naacl-main.153).
- [7] R. Lin *et al.*, "PAM: Understanding product images in cross product category attribute extraction," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2021, pp. 3262–3270. doi: [10.1145/3447548.3467164](https://doi.org/10.1145/3447548.3467164).
- [8] S. Lu, Y. Ding, M. Liu, Z. Yin, L. Yin and W. Zheng, "Multiscale feature extraction and fusion of image and text in VQA," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, 2023, Art. no. 54. doi: [10.1007/s44196-023-00233-6](https://doi.org/10.1007/s44196-023-00233-6).

- [9] S. Atakishiyev, M. Salameh, H. Babiker, and R. Goebel, "Explaining autonomous driving actions with visual question answering," in *IEEE Conf. Intell. Transp. Syst.*, 2023, pp. 1207–1214. doi: [10.1109/ITSC57777.2023.10421901](https://doi.org/10.1109/ITSC57777.2023.10421901).
- [10] H. Li *et al.*, "TG-VQA: Ternary game of video question answering," in *IJCAI Int. Joint Conf. Artif. Intell.*, vol. 2023, pp. 1044–1052, 2023. doi: [10.24963/ijcai.2023](https://doi.org/10.24963/ijcai.2023).
- [11] L. Gao, H. Zhang, N. Sheng, L. Shi, and H. Xu, "Learning neighbor-enhanced region representations and question-guided visual representations for visual question answering," *Expert. Syst. Appl.*, vol. 238, 2024, Art. no. 122239. doi: [10.1016/j.eswa.2023.122239](https://doi.org/10.1016/j.eswa.2023.122239).
- [12] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis and A. Smola, "Multimodal chain-of-thought reasoning in language models," Feb. 2023. doi: [10.48550/arXiv.2302.00923](https://doi.org/10.48550/arXiv.2302.00923).
- [13] A. U. Khan, H. Kuehne, C. Gan, N. D. V. Lobo, and M. Shah, "Weakly supervised grounding for VQA in vision-language transformers," in *Lecture Notes in Computer Science*, vol. 13695, pp. 652–670, 2022. doi: [10.1007/978-3-031-19833-5_38](https://doi.org/10.1007/978-3-031-19833-5_38).
- [14] T. Le, H. T. Nguyen, and M. Le Nguyen, "Multi visual and textual embedding on visual question answering for blind people," *Neurocomputing*, vol. 465, pp. 451–464, 2021. doi: [10.1016/j.neucom.2021.08.117](https://doi.org/10.1016/j.neucom.2021.08.117).
- [15] S. Al-Hadhrami, M. E. B. Menai, S. Al-Ahmadi, and A. Alnafessah, "An effective Med-VQA method using a transformer with weights fusion of multiple fine-tuned models," *Appl. Sci.*, vol. 13, no. 17, 2023, Art. no. 9735. doi: [10.3390/app13179735](https://doi.org/10.3390/app13179735).
- [16] H. Xia, R. Lan, H. Li, and S. Song, "ST-VQA: Shrinkage transformer with accurate alignment for visual question answering," *Appl. Intell.*, vol. 53, no. 18, pp. 20967–20978, 2023. doi: [10.1007/s10489-023-04564-x](https://doi.org/10.1007/s10489-023-04564-x).
- [17] D. Koshti, A. Gupta, M. Kalla, and A. Sharma, "TRANS-VQA: Fully transformer-based image question-answering model using question-guided vision attention," *Intel. Artif.*, vol. 27, no. 73, pp. 111–128, 2024. doi: [10.4114/intartif.vol27iss73pp111-128](https://doi.org/10.4114/intartif.vol27iss73pp111-128).
- [18] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," *Front. Artif. Intell.*, vol. 6, 2023, Art. no. 411. doi: [10.3389/frai.2023.1023281](https://doi.org/10.3389/frai.2023.1023281).
- [19] A. Bello, S. C. Ng, and M. F. Leung, "A BERT framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, 2023, Art. no. 506. doi: [10.3390/s23010506](https://doi.org/10.3390/s23010506).
- [20] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *J. Big Data*, vol. 9, no. 1, 2022, Art. no. 9. doi: [10.1186/s40537-022-00564-9](https://doi.org/10.1186/s40537-022-00564-9).
- [21] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis," *Appl. Sci.*, vol. 13, no. 6, pp. 1–25, 2023. doi: [10.3390/app13063915](https://doi.org/10.3390/app13063915).
- [22] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," Jul. 2019. doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692).
- [23] H. M. Abd Alaziz *et al.*, "Enhancing fashion classification with Vision Transformer (ViT) and developing recommendation fashion systems using DINOVA2," *Electronics*, vol. 12, no. 20, 2023, Art. no. 4263. doi: [10.3390/electronics12204263](https://doi.org/10.3390/electronics12204263).
- [24] T. Jumphoo, K. Phapatanaburi, W. Pathonsuwan, P. Anchuen, M. Uthansakul and P. Uthansakul, "Exploiting data-efficient image transformer-based transfer learning for valvular heart diseases detection," *IEEE Access*, vol. 12, pp. 15845–15855, 2024. doi: [10.1109/ACCESS.2024.3357946](https://doi.org/10.1109/ACCESS.2024.3357946).
- [25] H. Bao, L. Dong, S. Piao, and F. Wei, "BEIT: BERT pre-training of image transformers," 2022. doi: [10.48550/arXiv.2106.08254](https://doi.org/10.48550/arXiv.2106.08254).
- [26] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, vol. 2016, pp. 21–29. doi: [10.1109/CVPR.2016.10](https://doi.org/10.1109/CVPR.2016.10).
- [27] K. Yi *et al.*, "Clevrer: Collision events for video representation and reasoning," in *8th Int. Conf. Learn. Rep., ICLR*, 2020, pp. 1–20.
- [28] K. Yi, A. Torralba, J. Wu, P. Kohli, C. Gan and J. B. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," *Adv. Neural Inform. Process. Syst.*, vol. 2018, pp. 1031–1042, 2018.

- [29] S. Liu, X. Zhang, X. Zhou, and J. Yang, "BPI-MVQA: A bi-branch model for medical visual question answering," *BMC Med. Imaging*, vol. 22, no. 1, 2022, Art. no. 6799. doi: [10.1186/s12880-022-00800-x](https://doi.org/10.1186/s12880-022-00800-x).
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020. doi: [10.1007/s11263-019-01228-7](https://doi.org/10.1007/s11263-019-01228-7).
- [31] Z. Ma, W. Zheng, X. Chen, and L. Yin, "Joint embedding VQA model based on dynamic word vector," *PeerJ Comput. Sci.*, vol. 7, 2021, Art. no. e353. doi: [10.7717/peerj-cs.353](https://doi.org/10.7717/peerj-cs.353).
- [32] Y. Jiang, Y. Yang, Y. Xu, and E. Wang, "Spatial-temporal interval aware individual future trajectory prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 10, pp. 5374–5387, 2023. doi: [10.1109/TKDE.2023.3332929](https://doi.org/10.1109/TKDE.2023.3332929).
- [33] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NIPS'15: Proc. 28th Int. Conf. Neural Inform. Process. Syst.*, May 2015, pp. 2953–2961. doi: [10.5555/2969442.2969570](https://doi.org/10.5555/2969442.2969570).
- [34] Q. Tran *et al.*, "Explain by evidence: An explainable memory-based neural network for question answering," in *COLING 2020-28th Int. Conf. Comput. Linguist.*, 2020, pp. 5205–5210. doi: [10.18653/v1/2020.coling-main.456](https://doi.org/10.18653/v1/2020.coling-main.456).
- [35] C. Kolling, J. Wehrmann, and R. C. Barros, "Component analysis for visual question answering architectures," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–12. doi: [10.1109/IJCNN48605.2020.9206679](https://doi.org/10.1109/IJCNN48605.2020.9206679).
- [36] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu and W. Zheng, "The multi-modal fusion in visual question answering: A review of attention mechanisms," *PeerJ Comput. Sci.*, vol. 9, no. 1, pp. 1–29, 2023. doi: [10.7717/peerj-cs.1400](https://doi.org/10.7717/peerj-cs.1400).
- [37] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, 2023. doi: [10.1016/j.inffus.2022.09.025](https://doi.org/10.1016/j.inffus.2022.09.025).
- [38] L. B. Jimmy, R. K. Jamie, and G. E. Hinton, "Layer normalization," 2016. doi: [10.48550/arXiv.1607.06450](https://doi.org/10.48550/arXiv.1607.06450).
- [39] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," in *Proc. 33rd Int. Conf. Neural Inform. Process. Syst.*, vol. 32, pp. 4381–4391, 2019. doi: [10.5555/3454287.3454681](https://doi.org/10.5555/3454287.3454681).
- [40] M. Ma, T. Tohti, Y. Liang, Z. Zuo, and A. Hamdulla, "A focus fusion attention mechanism integrated with image captions for knowledge graph-based visual question answering," *Signal Image Video Process.*, vol. 18, no. 4, pp. 3471–3482, 2024. doi: [10.1007/s11760-024-03013-7](https://doi.org/10.1007/s11760-024-03013-7).
- [41] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Estimating semantic structure for the VQA answer space," 2020, *arXiv2006.05726*.
- [42] Ö. Özdemir and E. Akagündüz, "Enhancing visual question answering through question-driven image captions as prompts," Apr. 2024. doi: [10.1109/CVPRW63382.2024.00163](https://doi.org/10.1109/CVPRW63382.2024.00163).
- [43] D. Guessoum, M. Miraoui, and C. Tadj, "A modification of Wu and Palmer semantic similarity measure," *Tenth Int. Conf. Mobile Ubiquitous Comput., Syst., Serv. Technol., UBICOMM 2016 Tenth Int. Conf. Mob. Ubiquitous Comput. Syst. Serv. Technol.*, Venice, Italy, vol. 1, no. 1, pp. 42–46, 2016.
- [44] M. A. Pratama and R. Mandala, "Improving query expansion performances with pseudo relevance feedback and wu-palmer similarity on cross language information retrieval," 2022. doi: [10.1109/ICAICTA56449.2022.9932984](https://doi.org/10.1109/ICAICTA56449.2022.9932984).
- [45] Y. Bazi, M. M. Al Rahhal, L. Bashmal, and M. Zuair, "Vision-language model for visual question answering in medical imagery," *Bioengineering*, vol. 10, no. 3, Mar. 2023, Art. no. 380. doi: [10.3390/bio-engineering10030380](https://doi.org/10.3390/bio-engineering10030380).
- [46] L. Bashmal, Y. Bazi, F. Melgani, R. Ricci, M. M. Al Rahhal and M. Zuair, "Visual question generation from remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 3279–3293, 2023. doi: [10.1109/JSTARS.2023.3261361](https://doi.org/10.1109/JSTARS.2023.3261361).
- [47] C. Dewi, R. -C. Chen, H. Yu, and X. Jiang, "XAI for image captioning using SHAP," *J. Inf. Sci. Eng.*, vol. 39, no. 4, pp. 711–724, 2023. doi: [10.6688/JISE.202307_39\(4\).0001](https://doi.org/10.6688/JISE.202307_39(4).0001).
- [48] Y. Ding, M. Liu, and X. Luo, "Safety compliance checking of construction behaviors using visual question answering," *Autom. Constr.*, vol. 144, 2022, Art. no. 104580. doi: [10.1016/j.autcon.2022.104580](https://doi.org/10.1016/j.autcon.2022.104580).

- [49] Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, M. A. Al Zuair, and F. Melgani, "Bi-modal transformer-based approach for visual question answering in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022. doi: [10.1109/TGRS.2022.3192460](https://doi.org/10.1109/TGRS.2022.3192460).
- [50] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *AAAI 2020-34th AAAI Conf. Artif. Intell.*, 2020, pp. 13041–13049. doi: [10.1609/aaai.v34i07.7005](https://doi.org/10.1609/aaai.v34i07.7005).
- [51] C. Dewi, R. C. Chen, and H. Yu, "Weight analysis for various prohibitory sign detection and recognition using deep learning," *Multimed. Tools Appl.*, vol. 79, no. 43–44, pp. 32897–32915, 2020. doi: [10.1007/s11042-020-09509-x](https://doi.org/10.1007/s11042-020-09509-x).
- [52] U. Naseem, M. Khushi, and J. Kim, "Vision-language transformer for interpretable pathology visual question answering," *IEEE J. Biomed. Heal. Inform.*, vol. 27, no. 4, pp. 1681–1690, 2023. doi: [10.1109/JBHI.2022.3163751](https://doi.org/10.1109/JBHI.2022.3163751).
- [53] U. Naseem, M. Khushi, A. G. Dunn, and J. Kim, "K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering," *IEEE J. Biomed. Heal. Inform.*, vol. 28, no. 4, pp. 1886–1895, 2024. doi: [10.1109/JBHI.2023.3294249](https://doi.org/10.1109/JBHI.2023.3294249).
- [54] P. Dognin *et al.*, "Image captioning as an assistive technology: Lessons learned from VizWiz 2020 challenge," *J. Artif. Intell. Res.*, vol. 73, pp. 437–459, 2022. doi: [10.1613/jair.1.13113](https://doi.org/10.1613/jair.1.13113).
- [55] D. Gurari *et al.*, "VizWiz grand challenge: Answering visual questions from blind people," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3608–3617. doi: [10.1109/CVPR.2018.00380](https://doi.org/10.1109/CVPR.2018.00380).
- [56] T. van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. M. Snoek, and M. Worring, "Open-ended medical visual question answering through prefix tuning of language models," in *Lecture Notes in Computer Science*, Switzerland: Springer, 2023, pp. 726–736. doi: [10.1007/978-3-031-43904-9_70](https://doi.org/10.1007/978-3-031-43904-9_70).
- [57] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086. doi: [10.1109/CVPR.2018.00636](https://doi.org/10.1109/CVPR.2018.00636).
- [58] B. D. Nguyen, T. -T. Do, B. X. Nguyen, T. Do, E. Tjiputra and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *Lecture Notes in Computer Science*, Switzerland: Springer, 2019, pp. 522–530. doi: [10.1007/978-3-030-32251-9_57](https://doi.org/10.1007/978-3-030-32251-9_57).
- [59] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *Lecture Notes in Computer Science*, Switzerland: Springer, 2021, pp. 64–74. doi: [10.1007/978-3-030-87240-3_7](https://doi.org/10.1007/978-3-030-87240-3_7).
- [60] H. Gong, G. Chen, M. Mao, Z. Li, and G. Li, "VQAMix: Conditional triplet mixup for medical visual question answering," *IEEE Trans. Med. Imaging*, vol. 41, no. 11, pp. 3332–3343, 2022. doi: [10.1109/TMI.2022.3185008](https://doi.org/10.1109/TMI.2022.3185008).
- [61] H. Pan, S. He, K. Zhang, B. Qu, C. Chen and K. Shi, "AMAM: An attention-based multimodal alignment model for medical visual question answering," *Knowl.-Based Syst.*, vol. 255, Nov. 2022, Art. no. 109763. doi: [10.1016/j.knosys.2022.109763](https://doi.org/10.1016/j.knosys.2022.109763).
- [62] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 398–414, Apr. 2019. doi: [10.1007/s11263-018-1116-0](https://doi.org/10.1007/s11263-018-1116-0).
- [63] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," Apr. 2017. doi: [10.48550/arXiv.1704.03162](https://doi.org/10.48550/arXiv.1704.03162).