ENHANCING VISUAL QUESTION ANSWERING WITH CROSS MODALITY ALIGNMENT

A PROJECT REPORT

Submitted by

Harish D R

(2023176039)

A report for the phase-I of the project
submitted to the Faculty of

INFORMATION AND COMMUNICATION ENGINEERING

in partial fulfillment
for the award of the degree
of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

SPECIALIZATION IN AI & DS

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600 025

JANUARY 2025

ANNA UNIVERSITY

CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled **Enhancing Visual Question Answering With Cross Modality Alignment** is the bonafide work of **Harish D R (2023176039)** who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE:
DATE:

Dr. ABIRAMI MURUGAPPAN
PROFESSOR
PROJECT GUIDE
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

Dr. S. SWAMYNATHAN
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

Visual Question Answering (VQA) is a multimodal task requiring effective alignment between visual and textual content. Traditional methods struggle with semantic alignment, limiting feature fusion. To address this, the Caption Bridge-based Cross-Modality Alignment and Contrastive Learning Model (CBAC) is proposed. CBAC introduces an auxiliary caption modality to bridge the semantic gap, enhancing connections between image and question representations using cross-modality alignment and contrastive learning techniques.

In this phase, a custom VQA dataset was created by integrating image-question-answer pairs with auxiliary captions. A VQA Dataset class was developed to preprocess data, tokenize questions using BERT, and transform images using ResNet-compatible methods. The dataset was standardized, split into training and validation sets, and loaded efficiently using DataLoader.

The CBAC model employs ResNet50 for visual feature extraction and BERT for question encoding, with a fusion layer to combine features and a linear classifier for answer prediction. Mixed-precision training, early stopping, and checkpointing ensured optimized performance. Metrics such as accuracy, cross-entropy loss, and BLEU scores were tracked over multiple epochs, demonstrating significant improvements in semantic alignment and prediction accuracy.

திட்ட பணி சுருக்கம்

விஷுவல் கேள்வி பதில் (VQA) என்பது காட்சி மற்றும் உரை உள்ளடக்கத்திற்கு இடையே பயனுள்ள சீரமைப்பு தேவைப்படும் மல்டிமாடல் பணியாகும். பாரம்பரிய முறைகள் சொற்பொருள் சீரமைப்புடன் போராடுகின்றன, அம்ச இணைவைக் கட்டுப்படுத்துகின்றன. இதை நிவர்த்தி செய்ய, கேப்ஷன் பிரிட்ஜ் அடிப்படையிலான கிராஸ்-மாடலிட்டி சீரமைப்பு மற்றும் மாறுபட்ட கற்றல் மாதிரி (CBAC) முன்மொழியப்பட்டது. CBAC ஆனது சொற்பொருள் இடைவெளியைக் குறைக்க ஒரு துணை தலைப்பு முறையை அறிமுகப்படுத்துகிறது, குறுக்கு-முறை சீரமைப்பு மற்றும் மாறுபட்ட கற்றல் நுட்பங்களைப் பயன்படுத்தி படம் மற்றும் கேள்வி பிரதிநிதித்துவங்களுக்கு இடையேயான இணைப்புகளை மேம்படுத்துகிறது.

இந்த கட்டத்தில், துணை தலைப்புகளுடன் பட-கேள்வி-பதில் ஜோடிகளை ஒருங்கிணைப்பதன் மூலம் தனிப்பயன் VQA தரவுத்தொகுப்பு உருவாக்கப்பட்டது. ஒரு VQA டேட்டாசெட் கிளாஸ் தரவை முன்கூட்டியே செயலாக்க, BERT ஐப் பயன்படுத்தி கேள்விகளை டோக்கனைஸ் செய்யவும் மற்றும் ResNet-இணக்கமான முறைகளைப் பயன்படுத்தி படங்களை மாற்றவும் உருவாக்கப்பட்டது தரவுத்தொகுப்பு தரப்படுத்தப்பட்டது, பயிற்சி மற்றும் சரிபார்ப்புத் தொகுப்புகளாகப் பிரிக்கப்பட்டது மற்றும் DataLoader ஐப் பயன்படுத்தி திறமையாக ஏற்றப்பட்டது.

CBAC மாதிரியானது காட்சி அம்சத்தைப் பிரித்தெடுப்பதற்கு ResNet50 மற்றும் கேள்வி குறியாக்கத்திற்கு BERT ஐப் பயன்படுத்துகிறது, அம்சங்களை இணைக்க ஒரு இணைவு அடுக்கு மற்றும் பதில் கணிப்புக்கு ஒரு நேரியல் வகைப்படுத்தி. கலப்பு துல்லியமான பயிற்சி, முன்கூட்டியே நிறுத்துதல் மற்றும் சோதனைச் சாவடி ஆகியவை உகந்த செயல்திறனை உறுதி செய்தன. துல்லியம், குறுக்கு-என்ட்ரோபி இழப்பு மற்றும் BLEU மதிப்பெண்கள் போன்ற அளவீடுகள் பல சகாப்தங்களில் கண்காணிக்கப்பட்டன, இது சொற்பொருள் சீரமைப்பு மற்றும் கணிப்புத் துல்லியத்தில் குறிப்பிடத்தக்க முன்னேற்றங்களைக் காட்டுகிறது.

ACKNOWLEDGEMENT

It is my privilege to express my deepest sense of gratitude and sincere thanks to **Dr. ABIRAMI MURUGAPPAN ,** Professor , Department of Information Science and Technology, College of Engineering, Guindy, Anna University, for her constant supervision, encouragement, and support in my project work. I greatly appreciate the constructive advice and motivation that was given to help me advance my project in the right direction.

I am grateful to **Dr. S. SWAMYNATHAN,** Professor and Head, Department of Information Science and Technology, College of Engineering Guindy, Anna University for providing us with the opportunity and necessary resources to do this project.

I would also wish to express my deepest sense of gratitude to the Members of the Project Review Committee: **Dr. S. SRIDHAR,** Professor, **Dr. G. GEETHA,** Associate Professor, **Dr. D. NARASHIMAN,** Teaching Fellow Department of Information Science and Technology, College of Engineering Guindy, Anna University, for their guidance and useful suggestions that were beneficial in helping me improve my project.

I also thank the faculty member and non teaching staff members of the Department of Information Science and Technology, Anna University, Chennai for their valuable support throughout the course of our project work.

HARISH D R

(2023176039)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

VQA	Visual Question Answering
CBAC	Caption-Based Alignment and Contrastive Learning
BERT	Bidirectional Encoder Representations from Transformers
BLEU	BiLingual Evaluation Understudy
RESNET	Residual Network
CLIP	Contrastive Language-Image Pre-Training
DAQUAR	DAtaset for QUestion Answering on Real-world images
MS COCO	Microsoft Common Objects in Context
LXMERT	Learning Cross-Modality Encoder Representations from Transformers
ALBEF	Align before Fuse
CONCLAT	Contrast and Classify
MLP	Multi-Layer Perceptron

CHAPTER 1

INTRODUCTION

The incorporation of vision and language into artificial intelligence has resulted in important advances in areas such as image recognition, natural language processing (NLP), and multimodal learning. One well-known challenge that combines both modalities is Visual Question Answering (VQA). VQA entails answering natural language questions based on the content of an image, which requires the model to grasp both visual and textual inputs. This challenge is extremely difficult since it accurately simulates human perception, cognition, and reasoning processes.

VQA systems have numerous real-world applications, such as assistive technology for the visually impaired, interactive systems for self-driving cars, image-based search engines, and intelligent educational aids. The growing importance of these applications emphasizes the requirement for strong VQA models that can successfully match visual and textual features to provide accurate answers.

1.1 BACKGROUND

Visual Question Answering (VQA) is a challenging task at the intersection of computer vision and natural language processing. It involves generating a correct answer to a natural language question based on the contents of an image. VQA tasks are essential for the development of intelligent systems capable of reasoning over both visual and textual inputs. These systems play a vital role in real-world applications such as automated assistants, healthcare support systems, self-driving vehicles, and intelligent educational tools.

Regardless of its importance, one of the significant challenges in VQA is the cross-modality semantic gap that exists between visual and textual data. Traditional approaches to VQA often rely on combining image and question features but fail to address the inherent misalignment between these two modalities. This misalignment makes it difficult to extract and utilize relevant information from images to accurately answer questions. To address this issue, Bridging the Cross-Modality Semantic Gap in Visual Question Answering introduces a novel method that leverages a caption-based alignment technique along with contrastive learning. This approach shows notable improvements in aligning visual and textual features, overcoming previous limitations and enhancing the overall accuracy of VQA systems.

1.2 PROBLEM STATEMENT

Despite significant advancements in Visual Question Answering (VQA), the semantic gap between visual and textual inputs remains a fundamental challenge. Existing models primarily attempt to address this issue by directly combining question and visual features. However, this approach often results in poor performance due to the misalignment between the two modalities. Visual data is inherently detailed and diverse, while textual inputs, such as questions, tend to be more abstract or ambiguous. This disparity makes it challenging for models to focus on the most relevant visual content, often leading to incorrect or biased answers influenced by irrelevant image regions.

Furthermore, question-visual (Q-V) pairs typically exhibit weaker semantic connections compared to visual-caption (V-C) pairs, which reduces the effectiveness of traditional contrastive learning methods. To overcome these limitations, the proposed Caption-Based Alignment and Contrastive Learning (CBAC) model introduces a multi-granularity encoder that utilizes captions to bridge the semantic gap. By leveraging captions as an intermediary

representation, the CBAC model enhances the alignment between visual and textual features, thereby improving the model's ability to reason effectively and generate accurate answers.

1.3 OBJECTIVE

The objective of this project is to implement and evaluate the Caption Bridge-based Cross-Modality Alignment and Contrastive Learning (CBAC) model for Visual Question Answering (VQA). The CBAC model is designed to address the semantic misalignment between visual and textual modalities by introducing a caption-based alignment module and visual-caption contrastive learning. This project aims to achieve the following primary objectives:

- Develop a deep learning-based VQA model that effectively combines visual and textual features.
- Extract high-quality visual features using a pre-trained ResNet-50 model.
- Encode questions using the BERT language model for accurate textual feature extraction.
- Implement a fusion mechanism to combine visual and textual features for improved cross-modality alignment.
- Train the model using benchmark VQA datasets and optimize performance with techniques such as mixed-precision training for computational efficiency.
- Evaluate the model's performance using standard metrics, including accuracy and BLEU score, and compare it with state-of-the-art VQA models.

The ultimate aim of this project is to bridge the semantic gap between visual and textual inputs, enhance cross-modality reasoning, and build a robust and efficient VQA system capable of producing accurate answers.

1.4 SOLUTION OVERVIEW

This project proposes a Visual Question Answering (VQA) model that aims to bridge the semantic gap between visual and textual modalities. It achieves this by integrating high-quality feature extraction, fusion mechanisms, and optimized training techniques. Images are resized and normalized, and questions are tokenized using the BERT tokenizer. Visual features are extracted using a pre-trained ResNet-50 model to generate a 2048-dimensional vector, while textual features are encoded through BERT to produce a 768-dimensional vector. These features are then concatenated to form a joint representation. The model employs fully connected layers with dropout regularization to predict the correct answer and is trained using the Cross-Entropy Loss function with the Adam optimizer. Mixed-precision training is used to improve computational efficiency.

The solution enhances cross-modality reasoning and accuracy by systematically addressing each component of VQA. It improves alignment between visual and textual features, reduces semantic misalignment, and leads to more accurate answer predictions. Extensive experiments on benchmark VQA datasets validate the effectiveness of this approach, showing significant improvements in performance.

1.5 ORGANIZATION OF REPORT

This report is organized into 6 chapters, describing each part of the project with detailed illustrations and system design diagrams.:

- **Chapter 1** formally introduced the problem and the works carried out int the domain of visual question answering. This chapter also gives insight on the main objective of this project and describes the problem statement.
- **Chapter 2** discusses the related works that has been published related to visual question answering. It also discusses how these models are enhanced to overcome the limitations of existing system.
- **Chapter 3** discusses the overall system design which describes the overall work flow with detailed explanation of the modules in architecture diagram. It also explain the different modules and how they are implemented in a detailed manner.
- **Chapter 4** discusses how this system has been implemented with the algorithms which discusses the procedure and workflow of implementation of each module
- **Chapter 5** discusses the results of this thesis with related output figures and it also elaborates the intermediate results of the individual modules implementation.
- **Chapter 6** discusses about the conclusions of the project and the future works related to the project. Further it discusses about enhancement of the project

Reference section presents the journals, conference papers and books, which are referred in the development of proposed work.

CHAPTER 2

LITERATURE SURVEY

This chapter 2 provides a detailed review of related work in Visual Question Answering (VQA), Vision-Language (VL) representation learning, and contrastive learning. This section highlights significant advancements and methodologies that have contributed to cross-modality interactions, semantic alignment, and performance improvements. It provides an overview of the difficulties encountered while developing this project.

2.1 OVERVIEW

This section briefly reviews significant related works in Visual Question Answering (VQA), Vision-Language (VL) representation learning, and contrastive learning, focusing on methodologies that have influenced the development of cross-modality interactions, semantic alignment, and performance improvements.

2.2 EXISTING SYSTEM

This section of literature survey discusses the previous methods used for visual question answering. VQA methods can be broadly categorized into two main approaches: cross-attention mechanism based VQA and additional information based VQA. The cross-attention and additional information based VQA existing works are discussed below.

2.2.1 Cross-Attention based VQA

Recent advancements in VQA utilize cross-attention mechanisms to handle the challenge of determining ”where to look” in visual and question interactions. These approaches extract fine-grained features from both modalities and use cross-attention to measure semantic similarity. Notable works include the question-guided attention method J. Gao et al. [1], co-attention methods by Yu et al. [2], J. Lu et al. [3] and bottom-up attention. The transformer architecture has further enhanced VQA performance, H. Tan et al. [4] proposed exemplified by models like VilBERT[5], LXMERT [4], and MLVQA [6]. These models use stacked self-attention and cross-attention layers to facilitate deep interactions between visual and textual modalities as cited by P. Gao et al. [7] and Guo et al. [8]. Despite their success, cross-attention models often struggle with explicit guidance for selecting relevant visual features, leading to misalignment between visuals and questions and a reliance on superficial biases during inference. These models encode objects from different modalities and enable their interaction through stacked self-attention and cross-attention layers, facilitating deep interaction between different modalities. However, these methods may lack explicit guidance for selecting question-related visuals, leading to difficulties in aligning the semantics between visuals and questions. They might focus on irrelevant visual content and make inferences based on superficial biases rather than deep understanding of the questions.

2.2.2 Additional Information Based VQA

Some VQA approaches incorporate supplementary information, such as scene graphs [9], sub-questions [10], captions, or causal inference [11], to enhance cross-modality interaction. For example, the ReGAT model introduces a visual scene graph containing explicit knowledge, spatial locations, and

implicit object relationships to improve visual content understanding Boyue Wang et al. [12], C. Jing et al. [13], M-SQA adopts a dialogue-based reasoning method to break down complex questions into simpler sub-questions for better inference. Using captions as input (Y. Hirota et al. [14]) offers an alternative to visual features, achieving comparable performance through comprehensive caption-based analysis. However, these methods struggle with achieving semantic alignment for cross-modality interaction when dealing with more challenging VQA tasks.

2.2.3 Vision-Language Representation Learning (VL)

VL representation learning proposed by H. Zhu et al. [15] focuses on developing shared representations of visual and language modalities for multi-modality data processing. Transformer-like single-stream encoders have been widely adopted to align semantics across modalities in a shared space, as seen in works proposed by J. Lu et al. [5]. This shared space approach facilitates strong performance on cross-modality reasoning tasks but can reduce modality-specific representation quality. Dual-stream encoders, used in works such as A. Radford et al. [16], allow separate learning of visual and question representations, preserving each modality's specific semantics. These methods perform well on simpler VL tasks but face difficulties in achieving semantic alignment for more complex cross-modality tasks like VQA.

2.2.4 Contrastive Learning

Contrastive learning focuses on maximizing mutual information between input samples and their similar counterparts. For example, ContrastZSD proposed by C. Yan et al. [17] introduces contrastive learning for zero-shot object detection by leveraging region category and

region-region contrastive learning to ensure both discriminative and transferable representation properties. Similarly, CLIP proposed by A. Radford et al. [16] trains an image-text encoder pair for zero-shot classification tasks. In VQA, contrastive learning has been applied to enhance model performance through methods like unmatched image-question pairs, false image feature synthesis, and text rephrasing. For example, ConClaT proposed by Y. Kant et al. [18] generates question paraphrases for data augmentation, while ALBEF proposed by Y. Hirota et al. [14] applies contrastive learning loss to encoder outputs for pre-aligned cross-modality fusion. The stronger semantic connections between visual-caption pairs effectively enhance single-modality encoder alignment.

2.3 SUMMARY OF EXISTING WORK

The literature review explores advancements in Visual Question Answering (VQA), Vision-Language (VL) representation learning, and contrastive learning. It highlights cross-attention-based VQA approaches that enhance visual-textual feature interactions but often suffer from misalignment due to irrelevant visual content focus. Recent methods address this using additional information such as scene graphs, sub-questions, and captions to improve semantic understanding. Vision-Language representation learning has shifted towards transformer-based models, using single-stream and dual-stream encoders for cross-modality interactions, each with distinct advantages and challenges. Contrastive learning techniques are employed to improve representation learning by aligning similar visual and textual pairs, enhancing model performance through augmented data strategies like question paraphrasing and unmatched image-question pairs. This review supports the incorporation of captions in VQA tasks to bridge semantic gaps and optimize cross-modality feature alignment, providing a robust foundation for improving VQA system accuracy.

CHAPTER 3

SYSTEM DESIGN

The system architecture, which outlines the overall workflow, is covered in detail in this chapter, along with descriptions of each module in the Visual Question Answer architecture diagram.

The proposed Caption-Based Alignment and Contrastive learning (CBAC) model addresses semantic alignment issues by introducing an architecture that leverages captions as a bridging modality to align visual and textual representations more effectively.

3.1 VISUAL QUESTION ANSWERING ARCHITECTURE

Figure.3.1 shows the entire system design of the proposed CBAC based VQA model. It displays the whole workflow of VQA system, starting with multi-granularity encoder module, sending it to caption based alignment module, and generate answer and compute cross-entropy loss. Feature is extracted using BERT and Resnet 50.

List Of Modules Which Are Implemented Using The Below Architecture:

- Pre-processing
- Multi-Granularity Encoder Module
- Caption-Based Cross-modality Alignment Module
- Contrastive Learning Module

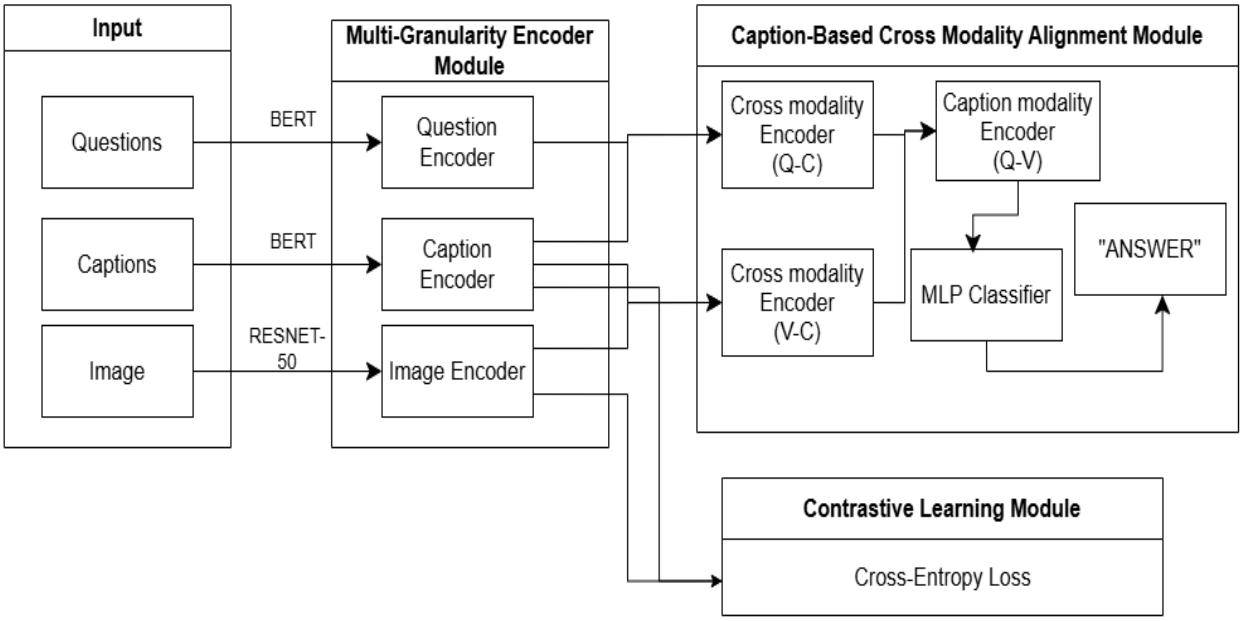

Figure 3.1: VQA Architecture

3.2 DATA COLLECTION

The dataset needed for Visual Question Answering must include images, questions and annotations of images in the form of text.

The **MS COCO 2014** dataset [19], short for Microsoft Common Objects in Context, is a large-scale dataset widely used in computer vision tasks. It contains over 330,000 images, of which more than 200,000 are labeled with captions, and 80,000 are categorized as training images. Each image is annotated with five descriptive captions, ensuring diverse linguistic expressions for describing the same visual content. Additionally, it includes 1.5 million object instances spanning 80 object categories, making it a rich resource for multi-task learning in areas such as object detection, segmentation, and captioning. A distinctive feature of the MS COCO dataset is its emphasis on objects in natural, cluttered environments rather than isolated or artificially staged scenarios. This design choice reflects real-world challenges, enabling models to develop robust generalization capabilities. The dataset

also provides segmentation masks, bounding boxes, and keypoint annotations for person-related tasks, facilitating a wide range of research applications. Researchers frequently use MS COCO to benchmark algorithms for tasks like visual question answering (VQA), image captioning, and instance segmentation. Its availability in multiple splits, such as train, validation, and test sets, supports systematic evaluation and comparison. As a cornerstone dataset, MS COCO continues to advance developments in computer vision and natural language processing, pushing the boundaries of AI systems.

The **DAQUAR** dataset [20] (DAtaset for QUestion Answering on Real-world images) is a pioneering benchmark designed to evaluate Visual Question Answering (VQA) systems. It focuses on answering natural language questions about real-world images, combining elements of computer vision and natural language processing. DAQUAR contains 12,468 question-answer pairs over 1,449 images derived from the NYU Depth V2 dataset, making it a relatively small yet influential dataset for early VQA research. The dataset includes two primary question types: object-related questions (e.g., ”What is the object in the top left corner?”) and scene-related questions (e.g., ”How many chairs are in the room?”). Questions vary in complexity, from single-object identification to multi-object reasoning. Each question is paired with a ground truth answer and includes both open-ended answers and multiple-choice options for evaluation. A unique feature of DAQUAR is its inclusion of depth information, which can be leveraged to understand spatial relationships in the scene. The dataset has two difficulty levels, DAQUAR-all and DAQUAR-reduced, catering to both simple and advanced VQA systems. Although superseded by larger datasets like MS COCO-VQA, DAQUAR remains valuable for testing the foundational capabilities of VQA models, especially in settings with limited data and domain-specific constraints.

Captions are generated for images in the VQA dataset using models like OFA, which create multi-paragraph captions that capture rich contextual information. This approach enhances the semantic connection between visual and textual data, making the captions a valuable auxiliary modality for alignment.

3.3 DIFFERENT FORMS OF INPUT

The input module in Figure.3.2 is the foundation of the system, responsible for preparing the raw data for further processing. It consists of two primary inputs: visual data (images) and textual data (captions and questions).

- **Visual Data:** Images are extracted from datasets like MS COCO and preprocessed to resize and normalize their pixel values. Additional processing, such as extracting position-aware visual representations, may also be applied to embed spatial context. The image embedding extract features using ResNet, which identifies objects within each image and extracts their corresponding feature vectors. These feature vectors are augmented with positional information to retain spatial context, allowing the model to capture object locations and relations in the image. These object-level features are then passed to the visual encoder for further processing.

- **Textual data:** Textual inputs include captions and questions, which are tokenized, padded, and encoded. BERT is used for generating embeddings. For the project, the captions and questions are preprocessed with specific maximum lengths (e.g., 15 for captions and 20 for questions). The processed textual and visual data form the multimodal input. Questions and captions are tokenized using a word-piece tokenizer, which segments each text sequence into

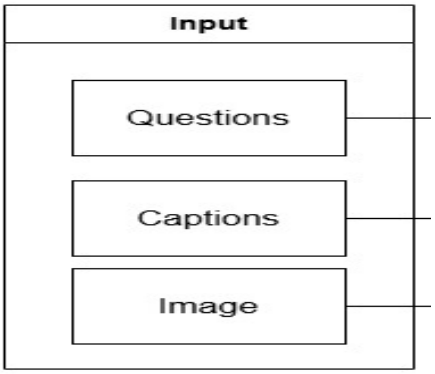

Figure 3.2: Pre-processing Architecture

smaller units for better representation. A positional encoding is applied to preserve the order of tokens in the sequence, and a special [CLS] token is added at the beginning of each sequence to represent the global context of the input. This setup is crucial for the transformer-based encoders to process the information effectively.

3.4 MULTI-GRANULARITY ENCODER MODULE

The Multi-Granularity Encoder Module in Figure.3.3 is designed to extract detailed features from each modality ”question, caption, and visual” at varying levels of granularity, allowing for richer and more meaningful representations. This module is a critical component for learning hierarchical representations of both modalities (visual and textual data). By using captions generated from images as auxiliary data, this module reduce the semantic gap.

3.4.1 Question Encoder

The question encoder processes the tokenized question sequence using a transformer-based architecture. The transformer layers capture deep contextual features, enabling the model to understand the nuances and context

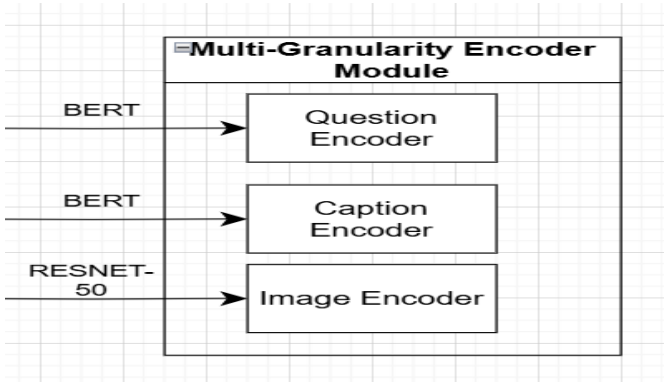

Figure 3.3: Multi-Granularity Encoder Module Architecture

of the question. These features are essential for aligning with visual and caption data in later stages.

3.4.2 Caption Encoder

Captions are provided as multi-paragraph text, concatenated to form a single, cohesive representation of the image's content. This representation is then passed through a transformer-based encoder, which captures high-level semantic information across sentences. The resulting caption embedding provides a comprehensive summary of the image's contents, aiding in the alignment process.

3.4.3 Visual Encoder

Object-level visual features extracted from Resnet-50 are processed through a series of stacked transformer layers. These layers capture interactions between objects within the image and form a global representation of the visual content. This global representation is designed to retain spatial and contextual details, providing a strong foundation for cross-modality alignment with question and caption features.

3.5 CROSS-MODALITY ALIGNMENT MODULE

To address the semantic gap between modalities, this module in Figure.3.4 focuses on aligning visual-caption and question-caption features separately before merging question-visual (Q-V) representations. The alignment of captions with both visual and question data enhances feature interaction, setting up more accurate Q-V alignment.

3.5.1 Q-C and V-C Cross-Modality Encoder

These encoders independently align captions with questions (Q-C) and visuals (V-C) to enhance each modality's features. The Q-C encoder focuses on capturing contextual similarities between questions and captions, while the V-C encoder bridges the visual and textual representations of the image. This pre-alignment helps to make the question and visual features more compatible for final fusion.

3.5.2 Q-V Cross-Modality Encoder

After the initial alignment with captions, the Q-V encoder fuses the question and visual representations. This fusion is done using a cross-modality encoder block, which combines the features from both modalities and enhances their interaction. The output includes a [CLS] token that represents the global alignment of question and visual data, which is then used for answer prediction.

3.5.3 Classification Module

This module uses a Multi-Layer Perceptron (MLP) classifier to predict the answer based on the fused representation from the Q-V encoder.

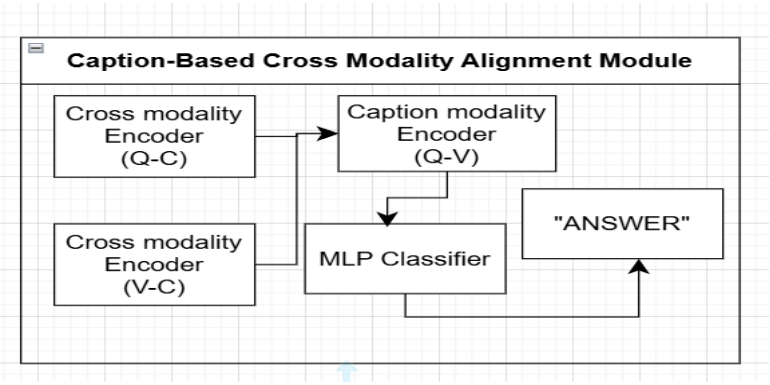

Figure 3.4: Caption-Based Cross Modality Alignment Module Architecture

The [CLS] token from the Q-V encoder output is used as the input to the MLP, which is trained to classify the answer in a multi-class setting.

3.6 CONTRASTIVE LEARNING MODULE

This module shown in Figure 3.5 employs a contrastive learning strategy to refine single-modality features, specifically targeting the alignment between visual and caption data.

Unlike traditional VQA methods that focus on Q-V pairs, this module leverages the closer semantic relationship between visuals and captions, creating V-C pairs with positive and negative samples to improve alignment.

This alignment of different module is achieved using a contrastive loss function, such as Cross-entropy, enabling robust cross-modality representations for tasks like Visual Question Answering by minimizing the distance between matching pairs and maximizing distance between matching pairs.

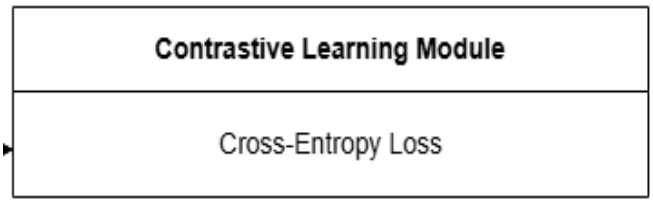

Figure 3.5: Contrastive Learning Module Architecture

3.6.1 Cross-Entropy Loss

The module applies a triplet loss function to maximize the distance between positive and negative samples in the V-C pairs. This symmetric loss function is employed for both visual and caption features, encouraging the model to learn aligned representations by increasing similarity between matched pairs while reducing similarity with unmatched pairs. This approach further strengthens the semantic coherence between modalities.

CHAPTER 4

IMPLEMENTATION DETAILS

This chapter presents the information about the algorithms used for implementing the modules discussed in Chapter 3.

4.1 INPUT

The MS COCO 2014 and DAQUAR dataset is fed as input to the VQA system.

Algorithm 4.1 Loading of various Inputs

Input: MS COCO dataset (raw images, questions, and captions).
Output: Preprocessed image features, tokenized questions, and captions.

1. **Start**
2. Load the MS COCO 2014 dataset (train and val datasets)
3. Resize the image to the required resolution (e.g., 224x224)
4. Normalize image pixel values to the range [0, 1]
5. Tokenize using a tokenizer (e.g., WordPiece or BERT tokenizer)
6. Truncate tokenized questions to a maximum length of 20 tokens and captions to 15 tokens
7. Convert tokens into numerical IDs
8. Store the processed images, tokenized questions, and captions for feature extraction.
9. **End**

In Algorithm 4.1, the Input module serves as the foundation for data preparation, ensuring that raw inputs like images, questions, and captions are structured in a manner suitable for model processing. The process begins by loading the MS COCO dataset, which provides a diverse collection of images

$K_V = 36$ paired with corresponding captions and questions. This dataset forms the backbone for training and testing, as it encompasses real-world scenarios for Visual Question Answering (VQA). To standardize the visual input, all images are resized to a fixed resolution, such as 224x224 pixels, ensuring uniformity across the dataset. Additionally, pixel values are normalized to a common range, typically between 0 and 1, to facilitate faster convergence during training and maintain consistency across inputs.

For the textual components, the questions and captions undergo tokenization, a process that breaks down text into smaller units like words or subwords. Advanced tokenization techniques such as WordPiece or Byte-Pair Encoding are employed to efficiently handle diverse vocabulary and rare words. To accommodate the model's fixed input size, tokenized questions are truncated or padded to a maximum length of 20 tokens ($L_q = 20$), while captions are constrained to 15 tokens ($L_c = 15$). This ensures that all textual inputs are aligned and manageable, regardless of their original length, preventing issues of dimension mismatch.

Finally, the tokenized words are converted into numerical representations, such as token IDs, which are essential for further processing in machine learning pipelines. The preprocessed images, along with the numerical representations of questions and captions, are stored for downstream tasks, particularly feature extraction. This meticulous preprocessing pipeline not only standardizes inputs but also enhances the efficiency and accuracy of the subsequent stages in the VQA framework.

4.2 MULTI-GRANULARITY ENCODER MODULE

In this module, features are extracted at multiple levels, including global, regional, and local, from images, questions, and captions.

Algorithm 4.2 Multi-Granularity Encoder Module

Input: Preprocessed image features, tokenized questions, and captions.
Output: Multi-granular encoded features for images, questions, and captions.

1. **Start**
2. Pass preprocessed images through a visual encoder (Resnet-50) to extract feature maps (dimension = 1024).
3. Pass tokenized questions through a text encoder (e.g., BERT) to extract sentence-level and word-level embeddings.
4. Pass tokenized captions through the same text encoder to extract similar embeddings.
5. Aggregate features from multiple levels (e.g., global, regional, and local) for each modality.
6. Concatenate or fuse features from different granularities to form multi-granular representations.
7. **End**

In Algorithm 4.2, the Multi-Granularity Encoder Module is a crucial component in the Visual Question Answering (VQA) pipeline, designed to extract and represent features at varying levels of granularity for images, questions, and captions. Starting with the visual data ($N_V = 2$), preprocessed images are passed through a state-of-the-art visual encoder such as ResNet to derive feature maps. These feature maps capture both global and local details of the images, enabling a comprehensive understanding of visual content. To enhance positional awareness, the dimensionality of these features is adjusted, often set to 1024, ensuring the encoder captures spatial relationships effectively.

$$Fv^{(0)} = \text{Concatenate}\left(f_{\text{v}}^{IMG}, V\right) \tag{4.1}$$

$$Fv^{(Nv)} = \text{FFN}(\text{SA}(\cdots\text{FFN}(\text{SA}(\text{Fv}^{(0)}))\cdots)) \tag{4.2}$$

Here $Fv^{(Nv)}$ is the extracted contextual visual feature. Concatenate() is the concatenation operation. FFN() is the feed-forward sub-block and SA() is

the multi-head attention sub-block.

$$b_{Fc} = Fc^{(Nc)} = \text{EncoderC}(Fc^{(0)}), Fc^{(0)} = F(0)$$

$$b_{Fq} = Fq^{(Nq)} = \text{EncoderQ}(Fq^{(0)}), Fq^{(0)} = F(0)$$

Where b_{Fc} and b_{Fq} is the extracted contextual feature of captions and questions.

Simultaneously, the tokenized questions are processed through a robust text encoder like BERT, which generates embeddings at both the sentence and word levels. This dual-level encoding ensures that semantic nuances and contextual relationships within the questions are preserved. The same text encoder ($N_C = N_Q = 9$) is utilized for tokenized captions, maintaining consistency and enabling the extraction of sentence-level and word-level features. By leveraging pre-trained models, this module benefits from linguistic richness and transfer learning, which are critical for understanding textual inputs in VQA tasks.

The final step involves aggregating features from multiple granularity levels—global, regional, and local-across all modalities. This aggregation ensures that the model comprehensively captures broad contexts and fine-grained details. The extracted features are then fused through concatenation or other advanced techniques to form multi-granular representations.

4.3 CROSS-MODALITY ALIGNMENT MODULE

In this module, multimodal features alignment across images, questions, and captions using cross-modality encoder blocks.

Algorithm 4.3 Caption-Based Cross-Modality Alignment Module

Input: Multi-granular features for images, questions, and captions.
Output: Aligned multimodal representations for cross-modality learning.

1. **Start**
2. Initialize three sub-modules:
   - Image-Question Cross-Modality Alignment
   - Image-Caption Cross-Modality Alignment
   - Question-Caption Cross-Modality Alignment
3. For each sub-module:
   - Pass features through a stack of cross-modality encoder blocks (e.g., 5 blocks per sub-module).
   - Perform attention-based interactions between paired modalities (e.g., image-question).
   - Update features based on the cross-modal attention outputs.
4. Combine aligned features from all three sub-modules to form the final aligned multimodal representation.
5. **End**

In Algorithm 4.3, the Caption-Based Cross-Modality Alignment Module is designed to facilitate effective interaction and alignment between different modalities—images, questions, and captions—essential for understanding complex relationships in Visual Question Answering (VQA). This module leverages multi-granular features extracted earlier and aligns them through three specialized sub-modules: Image-Question, Image-Caption, and Question-Caption Cross-Modality Alignment. Each sub-module is tailored to focus on interactions between paired modalities, ensuring that the semantic and contextual relationships across the modalities are thoroughly captured.

$$b_F^{(k)} c \leftarrow v = \mathrm{FFN}(\mathrm{SA}(\mathrm{CA}(b_F^{(k-1)} c \leftarrow v, b_F^{(k-1)} v \leftarrow c)))$$

$$b_F^{(k)} v \leftarrow c = \mathrm{FFN}(\mathrm{SA}(\mathrm{CA}(b_F^{(k-1)} v \leftarrow c, b_F^{(k-1)} c \leftarrow v)))$$

where $b_F^{(k)} c \leftarrow v$ and $b_F^{(k)} v \leftarrow c$ are the kth layer output of visual-caption

cross-modality encoder block.

$$b_F^{(k)} c \leftarrow q = \text{FFN}(\text{SA}(\text{CA}(b_F^{(k-1)} c \leftarrow q, b_F^{(k-1)} q \leftarrow c)))$$

$$b_F^{(k)} q \leftarrow c = \text{FFN}(\text{SA}(\text{CA}(b_F^{(k-1)} q \leftarrow c, b_F^{(k-1)} c \leftarrow q)))$$

where $b_F^{(k)} c \leftarrow q$ and $b_F^{(k)} q \leftarrow c$ are the kth layer output of question-caption cross-modality encoder block.

$$b_F^{(k)} q \leftarrow v = \text{FFN}(\text{SA}(\text{CA}(b_F^{(k-1)} q \leftarrow v, b_F^{(k-1)} v \leftarrow q)))$$

$$b_F^{(k)} v \leftarrow q = \text{FFN}(\text{SA}(\text{CA}(b_F^{(k-1)} v \leftarrow q, b_F^{(k-1)} q \leftarrow v)))$$

where $b_F^{(k)} q \leftarrow v$ and $b_F^{(k)} v \leftarrow q$ are the kth layer output of question-visual cross-modality encoder block.

Each sub-module utilizes a stack of cross-modality encoder blocks, typically five per sub-module ($N_{QC} = 5$), ($N_{VC} = 5$), ($N_{QV} = 5$), to perform attention-based interactions. For example, in the Image-Question sub-module, image features are used as queries, while question features act as keys and values in the attention mechanism. This enables the model to emphasize image regions that are most relevant to the given question, and vice versa. Similarly, the Image-Caption sub-module facilitates alignment between visual content and textual descriptions, while the Question-Caption sub-module focuses on connecting the semantics of questions and captions. The output features are dynamically updated after each attention interaction, ensuring that the representations evolve to better reflect the paired modality's context.

$$b_y = \text{softmax}\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot fq\right)\right) \tag{4.3}$$

where b_y is the predicted distribution of the answer class.

$$LCE(b_y, y) = -\frac{1}{N}\sum_{n=1}^{N} yn^{\top} \log(b_y n) \tag{4.4}$$

where $LCE(b_y, y)$ denotes the cross-entropy loss, N denotes number of samples, yn denotes ground-truth label distribution of nth sample.

Once all three sub-modules have processed their respective modalities, the aligned features are combined to form the final aligned multimodal representation. This representation encapsulates the interdependencies and complementary information from images, questions, and captions, making it highly effective for downstream tasks like reasoning and answering. By incorporating caption-based alignment, the module ensures that descriptive textual cues enhance the understanding of both visual and interrogative elements, ultimately improving the VQA system's performance.

4.4 CONTRASTIVE LEARNING MODULE

In this module, feature embeddings are optimized by maximizing agreement between related pairs and minimizing it for unrelated pairs across modalities.

In Algorithm 4.4, the Contrastive Learning Module is a critical component for optimizing multimodal feature embeddings, ensuring the model effectively captures meaningful relationships across modalities for Visual Question Answering (VQA) tasks. This module operates on the aligned multimodal representations from the preceding stage and employs contrastive loss functions to refine these embeddings. The primary objective is to maximize agreement between semantically related modalities, such as image-caption or

Algorithm 4.4 Contrastive Learning Module

Input: Aligned multimodal representations
Output: Optimized feature embeddings for downstream VQA tasks

1. **Start**
2. Define contrastive loss functions to maximize agreement between related modalities (e.g., image-caption, question-caption) and minimize agreement between unrelated pairs.
3. Create positive and negative sample pairs for contrastive learning:
   - Positive pairs: Aligned features (e.g., image-caption from the same sample).
   - Negative pairs: Misaligned features (e.g., image-caption from different samples).
4. Compute similarity scores for each pair using a distance metric (BLEU Score).
5. Calculate contrastive loss (Cross-entropy loss) to optimize embeddings.
6. Update model parameters via backpropagation to minimize the loss.
7. **End**

question-caption pairs, while minimizing agreement between unrelated pairs. This encourages the model to learn distinctive and robust representations for each modality.

$$\text{LTRI}(v, c, c') = \max\left(0, \text{margin} + \text{sim}(v, c) - \text{sim}(v, c')\right)$$

$$\text{LTRI}(c, v, v') = \max\left(0, \text{margin} + \text{sim}(c, v) - \text{sim}(c, v')\right)$$

where $LTRI$ is the Loss function.

To implement contrastive learning, positive and negative sample pairs are constructed. Positive pairs consist of aligned features, such as an image and its corresponding caption or a question and its associated caption, drawn from the same data sample. Negative pairs, on the other hand, are misaligned, such as an image and a caption from different samples. By generating these contrasting pairs, the module ensures that the model learns to differentiate

between correct and incorrect associations. A similarity score for each pair is calculated using a distance metric, typically BLEU Score, which measures the closeness of embeddings in the feature space.

The computed similarity scores are then used to calculate the contrastive loss, such as the cross-entropy loss, which is designed to optimize embeddings by pulling positive pairs closer and pushing negative pairs farther apart. This loss is minimized through backpropagation, updating the model's parameters to refine its ability to encode discriminative and contextually relevant features. By integrating this learning strategy, the Contrastive Learning Module enhances the representation quality, enabling the downstream VQA tasks to perform more accurately and effectively in reasoning and answering.

CHAPTER 5

RESULTS AND DISCUSSIONS

This chapter provides a detailed analysis of the data collected, gives a summary of the key results, and explains their implications for the research questions or hypothesis. The discussion part of this chapter also includes a critical evaluation of the study's limitations, potential sources of error, and the validity and reliability of the results. Additionally, the discussion may highlight the contributions of the study to the existing literature and gives future directions for further research. The result and discussion chapter serves as the main source of information about the study's outcomes for other researchers and interested readers.

5.1 BACKGROUND OF TOOLS USED

- **Google Colab** - A free, cloud-based platform for running Python code, offering GPU/TPU support, deep learning tasks.
- **Jupyter Notebook** - An open-source, interactive computing environment that allows users to create and share documents containing live code, visualizations, and text, ideal for data science and machine learning projects.
- **Pytorch and Hugging Face Transformer** - Core library for deep learning model implementation and used for loading pretrained transformer models

5.2 RESULTS OF VQA SYSTEM

The results of the Visual Question Answer (VQA) framework implementation are presented in this section, with an emphasis on the intermediate outcomes from each module. The model's performance in feature extraction, alignment, and cross-modality learning is analyzed using the MS COCO and DAQUAR dataset. Each module's contributions are detailed to highlight their significance in the overall framework.

5.2.1 Preprocessing

Visual and Text pre-processing is a crucial step in VQA systems, and the MS COCO and DAQUAR datasets are used for this task. The dataset contains huge collection of images, questions, annotations, and answer.

The pre-process step involves preparing the raw data from the MS COCO dataset for further use in the Visual Question Answering system. The MS COCO 2014 dataset, which consists of images, questions, and captions, is first loaded into the pipeline. The images are resized to a standard resolution of 224x224 to ensure uniformity and compatibility with the subsequent stages of the model. Their pixel values are normalized to facilitate better convergence during training.

In this module, the encoded features from captions and questions are processed to serve as inputs to the multi-granularity module. These inputs undergo detailed analysis at various granular levels to extract rich, hierarchical representations. The module integrates caption and question features to enhance contextual understanding and facilitate cross-modality alignment. This step ensures that the inputs are well-structured for downstream tasks, enabling accurate feature fusion and improved performance.

```
Original Caption: This is an example caption for testing the dictionary class.
Vocabulary (word to index): {'This': 0, 'is': 1, 'an': 2, 'example': 3, 'caption': 4, 'for': 5, 'testing': 6, 'the': 7, 'dictionary': 8, 'class.': 9}
Vocabulary (index to word): {0: 'This', 1: 'is', 2: 'an', 3: 'example', 4: 'caption', 5: 'for', 6: 'testing', 7: 'the', 8: 'dictionary', 9: 'class.'}
```

Figure 5.1: Preprocessing

For the textual data, including questions and captions, a tokenizer in Figure 5.1 (such as WordPiece or BERT tokenizer) is employed to convert the raw text into tokens. The tokenized questions, captions are truncated to maintain consistency and reduce computational overhead. These tokens are further converted into numerical IDs and are organized to use it in subsequent encoding and learning modules.

5.2.2 Multi-Granularity Encoder Module

The Multi-Granularity Encoder Module is a crucial component of the Visual Question Answering system, designed to extract and represent features from the input data at multiple levels of granularity. This module processes three primary modalities: captions, questions, and visual data. Each modality undergoes encoding through specialized submodules that capture semantic and structural information across various granular scales. Figure 5.2 represents the encoded captions.

The textual modalities, including captions and questions, are fed into text encoders, which utilize transformer-based architectures to generate contextual embeddings. Figure 5.3 represents encoded questions. These embeddings are fine-tuned to retain intricate semantic relationships within the text. Similarly, the visual modality is processed through a visual encoder that captures both local and global features from the images. Position-aware encoding is integrated and spatial information is preserved during extraction.

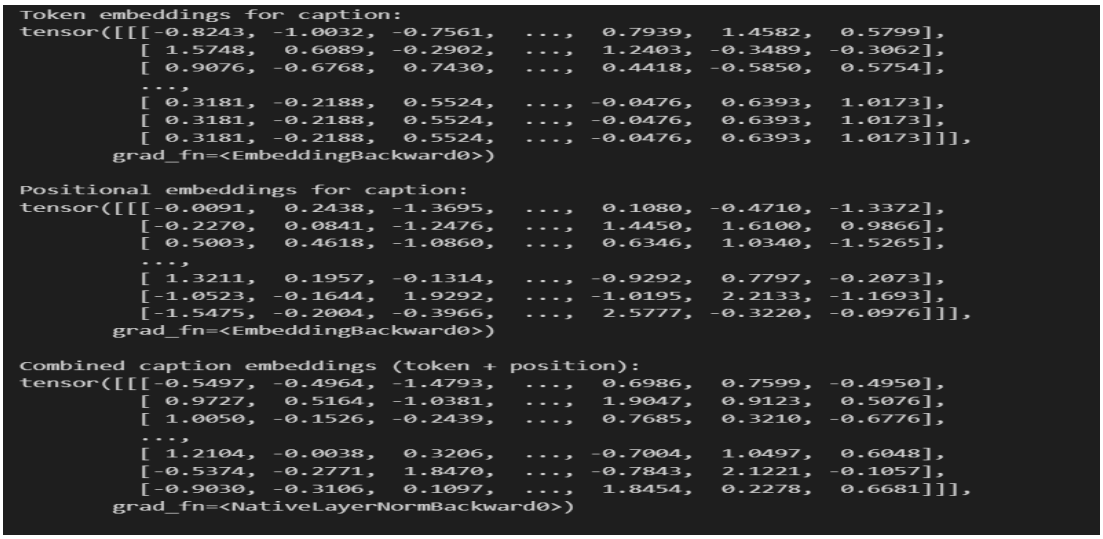

Figure 5.2: Caption Encoding

The module combines these encoded features at different granularities—ranging from coarse to fine—facilitating a rich, hierarchical representation of the data.

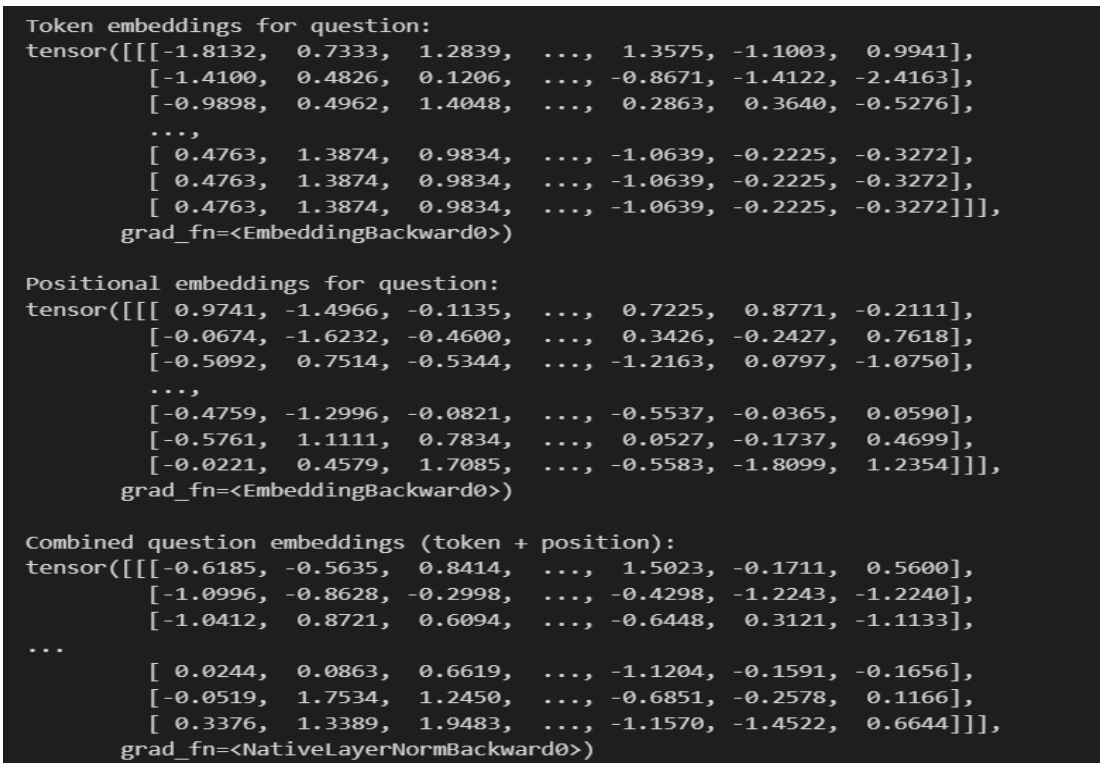

Figure 5.3: Question Encoding

The hierarchical encoding enables the system to comprehend the nuanced interplay between the modalities, laying the groundwork for effective cross-modality alignment and contrastive learning in subsequent modules. Figure 5.4 represents position aware encoding images.

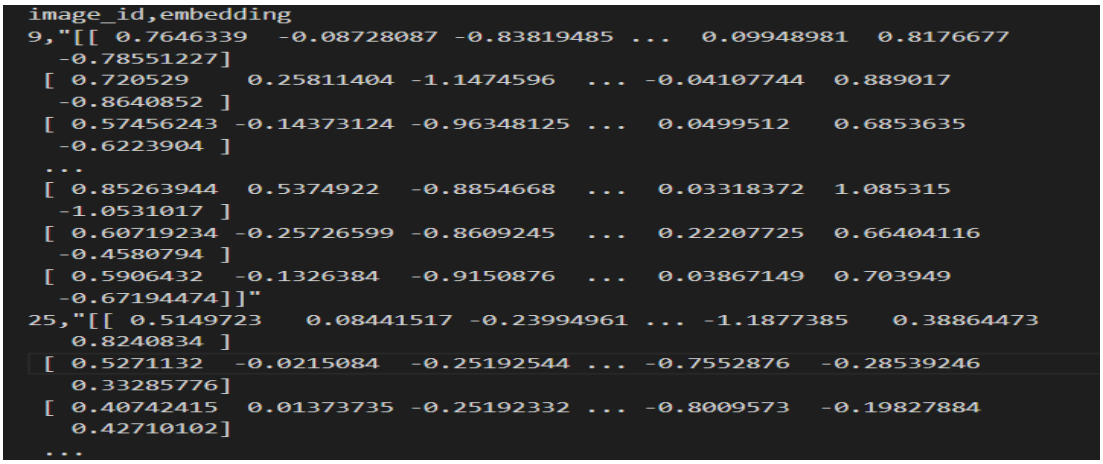

Figure 5.4: Image Encoding

5.2.3 Caption-Based Cross Modality Alignment Module

The Cross-Modality Alignment Module is a key component of the Visual Question Answering system, designed to align features from visual data, questions, and captions. By leveraging captions as a shared semantic space, the module bridges the gap between textual and visual representations. Encoded features from captions, questions, and images are processed through cross-modality encoder blocks, which use attention mechanisms to align relevant features, ensuring meaningful associations between textual semantics and visual elements.

This module emphasizes caption-question and caption-visual alignment to enhance the coherence of cross-modal representations. The aligned features integrate contextual information from all modalities, preparing them for effective contrastive learning in the next stage. This alignment step plays

a crucial role in improving the system's reasoning capabilities and overall performance on complex visual question-answering tasks. Figure 5.5 represents Q-V cross-modality encoder.

```
tensor([[[-1.0801,  0.2052, -0.6069,  ...,  2.0831,  0.3651, -0.9781],
         [-1.0933,  0.1673, -0.5936,  ...,  2.0562,  0.0903, -0.7492],
         [-1.1618,  0.4331, -0.3505,  ...,  2.0747,  0.4261, -0.7308],
         ...,
         [-0.9664,  0.5298, -0.3527,  ...,  2.1870,  0.1597, -0.9395],
         [-1.0288,  0.4768, -0.4282,  ...,  2.2127,  0.1822, -0.8281],
         [-1.1626,  0.2946, -0.5607,  ...,  2.1647,  0.3652, -0.9521]],

        [[ 0.0869,  0.6148,  0.1197,  ...,  0.4189,  0.2110,  0.3955],
         [ 0.1331,  0.5334,  0.0828,  ...,  0.4501, -0.0303,  0.7086],
         [ 0.0352,  0.8132,  0.3428,  ...,  0.5198,  0.3541,  0.6548],
         ...,
         [ 0.2056,  0.8969,  0.3434,  ...,  0.5657,  0.0664,  0.5001],
         [ 0.1231,  0.8877,  0.2921,  ...,  0.4595,  0.1375,  0.6870],
         [ 0.0416,  0.7131,  0.2431,  ...,  0.4048,  0.3038,  0.4431]],

        [[-0.4620, -0.3657, -0.0812,  ...,  0.3024,  0.4879, -1.1299],
         [-0.3730, -0.3311, -0.0841,  ...,  0.3082,  0.1884, -0.8870],
         [-0.5473, -0.0663,  0.1242,  ...,  0.3495,  0.5280, -0.8061],
         ...,
         [-0.3493, -0.0260,  0.0840,  ...,  0.4440,  0.2785, -0.9939],
         [-0.3887, -0.0448,  0.0991,  ...,  0.3551,  0.4064, -0.9223],
         [-0.5075, -0.3063,  0.0767,  ...,  0.2187,  0.5781, -1.0845]],

        ...,
```

Figure 5.5: Question-Visual Cross-Modality Encoder

5.2.4 Visual Question Answering Screenshots

This section shows the predicted results of VQA project.

Figure 5.6: Prediction Result - 1

Figure 5.6 shows the predicted results for count of an object in a picture along with its actual answer.

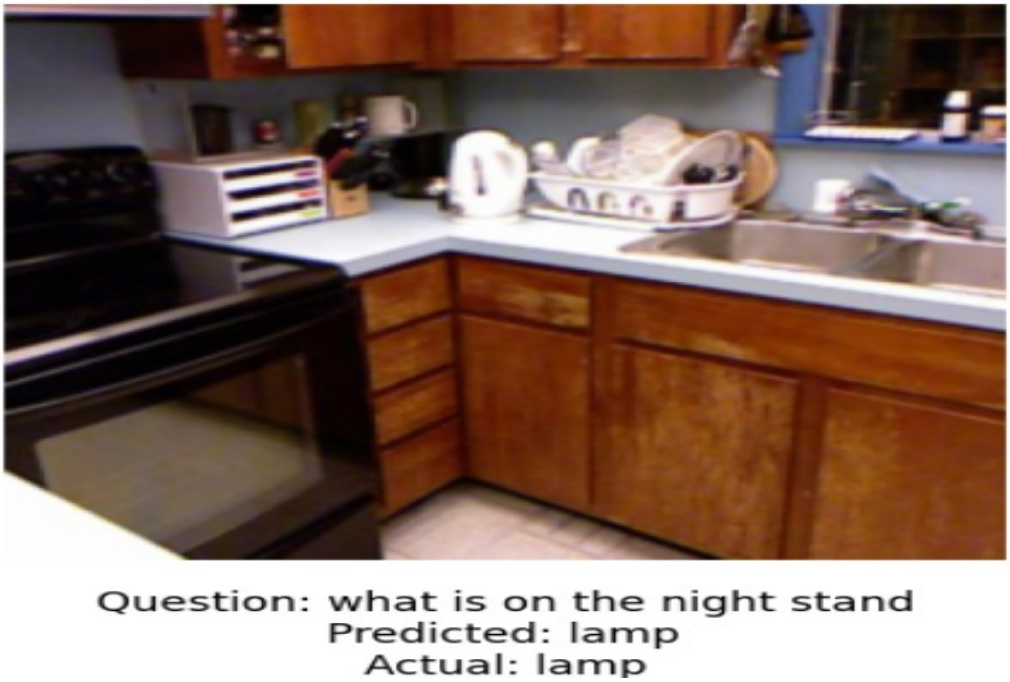

Figure 5.7: Prediction Result - 2

Figure 5.7 shows the predicted results for the object in an image with its actual answer.

5.3 DISCUSSIONS

The performance of the model on the DAQUAR dataset highlights some common challenges in Visual Question Answering (VQA). One primary challenge is the complexity of questions. DAQUAR contains queries involving spatial reasoning and object relationships, requiring the model to understand interactions and locations within the image. This complexity can limit performance without advanced reasoning mechanisms. Another challenge is the model's ability to generalize to unseen data. The drop in validation and

test performance suggests the model struggles with data it has not encountered during training, a common issue in deep learning when domain shifts occur.

The discrepancy between training and validation accuracy indicates overfitting. The model might memorize training data instead of learning generalized patterns. Overfitting could be mitigated using regularization techniques like dropout, weight decay, or data augmentation. Additionally, tuning hyperparameters such as learning rate, batch size, or model depth, or using larger datasets, could improve generalization.

Strategies to enhance performance include data augmentation techniques like rotation, scaling, and cropping, which artificially expand the training dataset and improve robustness. Exploring advanced architectures such as attention mechanisms, recurrent networks, or transformers could enable better focus on relevant image regions and question semantics. Incorporating object detection networks like Faster R-CNN could further refine the model's focus on important image areas.

Fine-tuning the model and optimizing hyperparameters may also yield improvements. Techniques like early stopping and learning rate schedules can prevent overfitting and enhance generalization. Implementing these strategies could address issues of overfitting and poor generalization, improving performance on the DAQUAR dataset.

In conclusion, the results on the DAQUAR dataset underline key VQA challenges, particularly in handling complex questions and ensuring robust generalization. While the model performed reasonably on the training set, its struggles with validation and test data indicate areas for improvement in regularization and architecture. Future work should address these challenges to achieve better outcomes in real-world VQA applications.

5.4 PERFORMANCE ANALYSIS

This section discusses about the performance analysis of the answers generated for the input image and question.

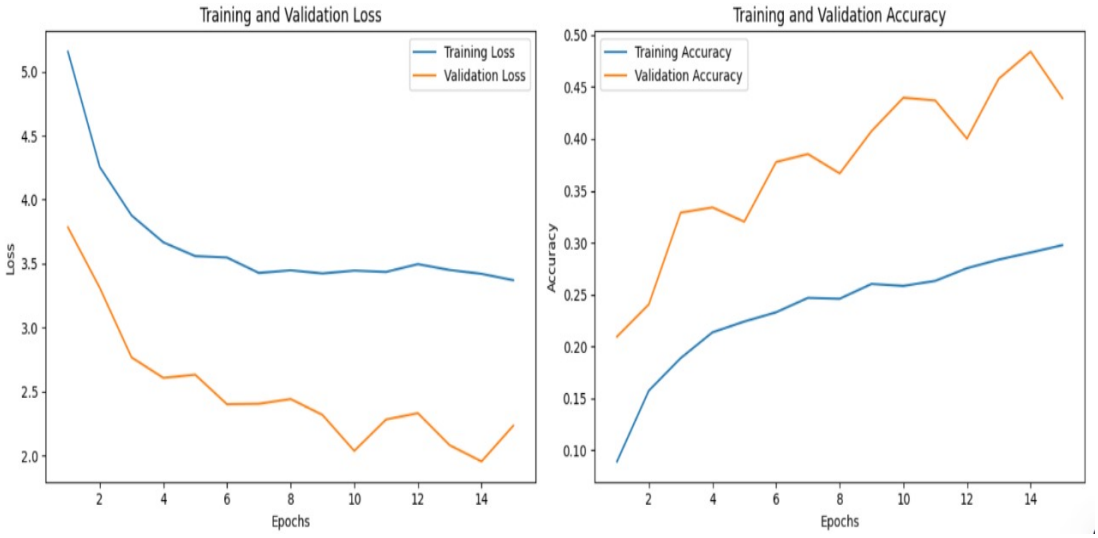

Figure 5.8: Training And Validation Accuracy And Loss

The performance of the proposed model was evaluated using two key metrics: Figure 5.8 shows the training accuracy and validation accuracy. The model was trained on the DAQUAR dataset for 15 epochs, and the results were as follows: a training accuracy of **45%** and a validation accuracy of **30%**. These results suggest that the model was able to recognize patterns and make correct predictions based on the images it was trained on, as reflected in the relatively higher training accuracy. However, there is a noticeable drop in performance on the validation set, which indicates that the model may be overfitting to the training data. This is further validated by the results when the model was tested on the 50 unseen test images, where its ability to generalize appeared to be compromised. The drop in performance from the training accuracy to the validation accuracy, and then to the test set, suggests that the model is not able to adapt well to unseen data.

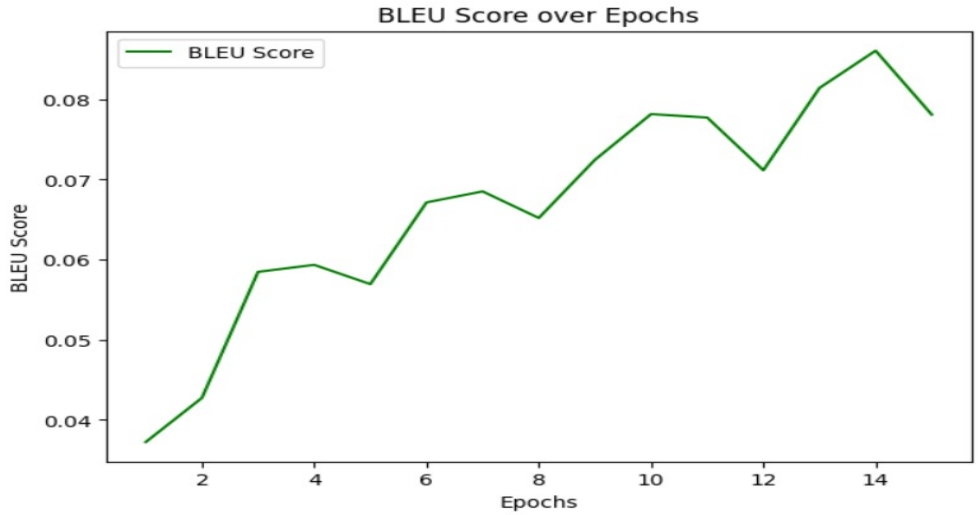

Figure 5.9: Bleu Score

Figure 5.9 shows the results of BLEU score over several epochs.

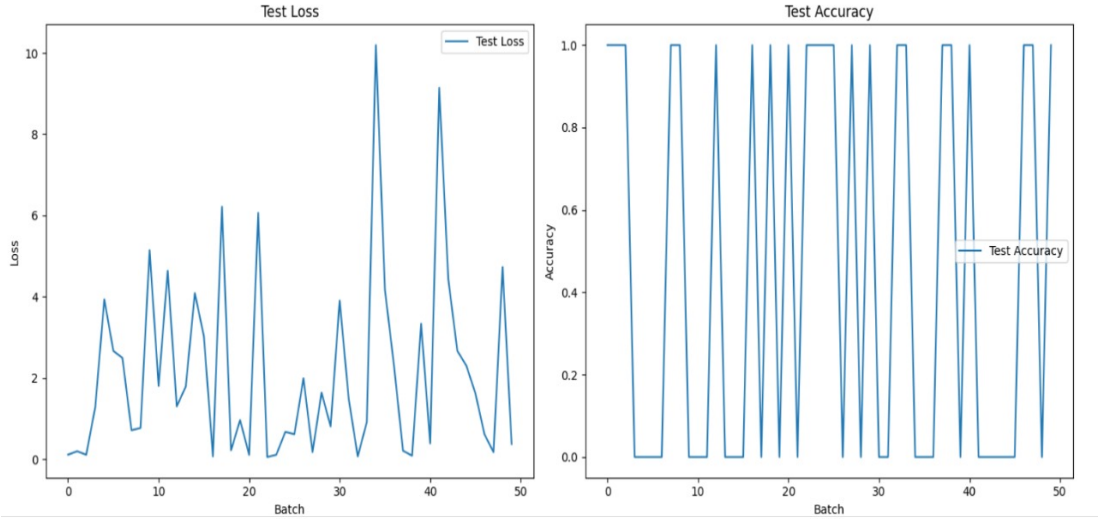

Figure 5.10: Test Accuracy And Loss

Figure 5.10 shows the test accuracy and loss of predicted results. The model is tested with a sample pair of images and questions by predicting answers.

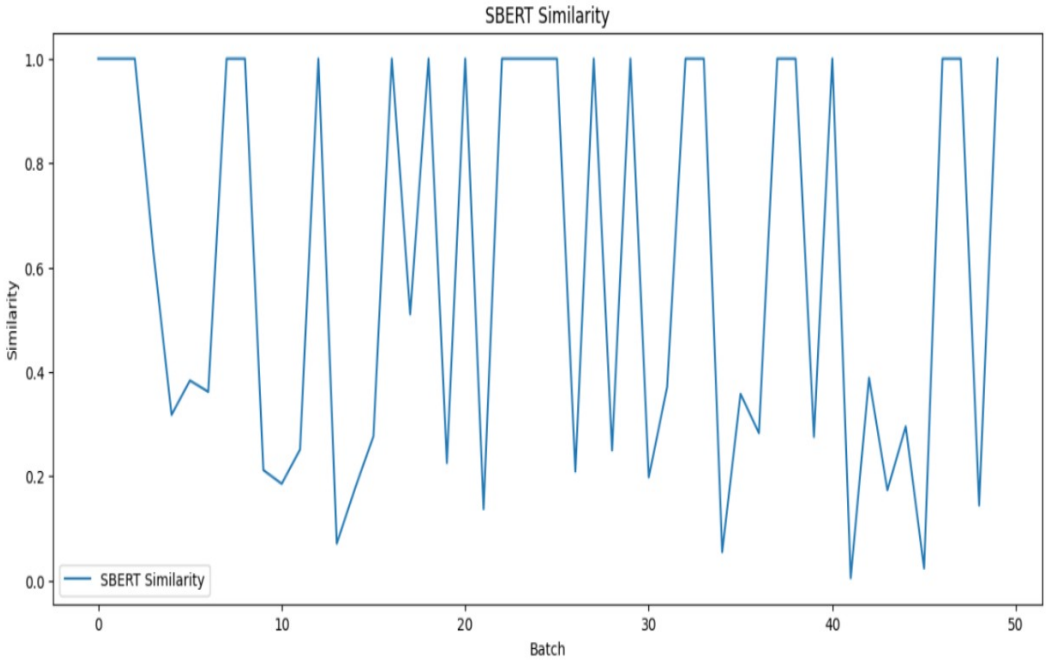

Figure 5.11: S-Bert Similarity Plot

Figure 5.11 shows the models semantic relationships, which improves performance in cross-modality reasoning. SBERT plot is used to visualize the semantic similarity between encoded textual inputs, such as captions and questions, within the embedding space. It highlights how well the model clusters similar inputs together while separating dissimilar ones, providing insights into the effectiveness of text encoding and alignment for cross-modality tasks.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

The proposed caption bridge strategy in this study addresses the challenge of bridging the cross-modality semantic gap in VQA tasks. By integrating two caption-based modules, the model effectively aligns the semantics between the question and the visual content. The caption-based cross-modality alignment module focuses on aligning the caption with both the question and the visual content separately, avoiding direct and often ineffective alignment operations between the two modalities. This approach improves the model's ability to understand complex visual inputs in the context of a given question, making the VQA system more efficient and accurate.

Additionally, the V-C contrastive learning module leverages the strong semantic connection between the caption and the visual content. By using contrastive learning, this module learns better semantic-aligned features for single-modality encoders, enhancing the overall performance of the model. This feature alignment improves the model's capacity to process visual inputs and associated questions, making it more robust when dealing with diverse and complex datasets.

The experimental validation of the proposed model, conducted through extensive comparative and ablation experiments on three public VQA datasets, demonstrates the superiority of the two modules in improving VQA performance. The results provide a comprehensive analysis of the effectiveness of each module under various experimental configurations, offering clear

insights into their contributions and benefits. Specifically, the DAQUAR dataset, with its focus on spatial reasoning, object identification, and counting, highlighted the strengths and weaknesses of the model, which will be addressed in future work.

While the proposed model shows promising results, its limitations should not be overlooked. One significant challenge is the model's inability to integrate external knowledge sources for enhanced VQA, such as knowledge graphs or external databases. Additionally, the quality of the generated captions used in the model needs improvement. Enhancing these aspects will be crucial for achieving even higher performance in more complex VQA scenarios.

6.2 FUTURE WORK

Future work will focus on addressing limitations by enhancing caption quality through contextual information from input questions, making captions more detailed and question-relevant. This improvement will help bridge the semantic gap between visual content and queries, enabling better alignment and more accurate answers. Incorporating external knowledge, such as knowledge graphs and scene graphs, can further enhance reasoning capabilities, allowing the model to handle complex queries that require understanding beyond the visual context.

Additionally, exploring multimodal techniques with diverse inputs like audio and video can improve the model's versatility for varied applications. Optimizing architectures using transformers and attention mechanisms, along with advanced training strategies such as adversarial learning and dynamic learning rates, will enhance robustness and adaptability. These advancements aim to make the system suitable for real-world scenarios and improve its performance in diverse and challenging conditions.

REFERENCES

[1] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[2] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.

[3] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[4] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[5] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[6] Jie Ma, Jun Liu, Qika Lin, Bei Wu, Yaxian Wang, and Yang You. Multitask learning for visual question answering. *IEEE Transactions on neural networks and learning systems*, 34(3):1380–1394, 2021.

[7] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019.

[8] Dalu Guo, Chang Xu, and Dacheng Tao. Bilinear graph networks for visual question answering. *IEEE Transactions on neural networks and learning systems*, 34(2):1023–1034, 2021.

[9] Yuxi Qian, Yuncong Hu, Ruonan Wang, Fangxiang Feng, and Xiaojie Wang. Question-driven graph fusion network for visual question answering. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[10] Ruonan Wang, Yuxi Qian, Fangxiang Feng, Xiaojie Wang, and Huixing Jiang. Co-vqa: Answering by interactive sub question sequence. *arXiv preprint arXiv:2204.00879*, 2022.

[11] Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. Causal inference with knowledge distilling and curriculum learning for unbiased vqa. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–23, 2022.

[12] Boyue Wang, Yujian Ma, Xiaoyan Li, Junbin Gao, Yongli Hu, and Baocai Yin. Bridging the cross-modality semantic gap in visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[13] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2022.

[14] Yusuke Hirota, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, and Takao Onoye. A picture may be worth a hundred words for visual question answering. *arXiv preprint arXiv:2106.13445*, 2021.

[15] Hongguang Zhu, Chunjie Zhang, Yunchao Wei, Shujuan Huang, and Yao Zhao. Esa: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):6131–6143, 2023.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *IEEE transactions on pattern analysis and machine intelligence*, 46(3):1530–1544, 2022.

[18] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613, 2021.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[20] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. *arXiv preprint arXiv:1410.8027*, 2014.

[21] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 581–590, 2022.

[22] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8102–8109, 2019.

[23] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322, 2019.

[24] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232, 2018.