Multimodal Attention Variants for Visual Question Answering

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

Doctor of Philosophy

Submitted by

Aakansha Mishra

Under the supervision of

Prof. Ashish Anand & Dr. Prithwijit Guha

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
Guwahati 781039, Assam, India

October 2023

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Aakansha Mishra
October 2023

Acknowledgements

I am profoundly thankful to my supervisors Prof. Ashish Anand and Dr. Prithwijit Guha for supervising my PhD thesis. I am thankful to them as they allowed me to work on the problem of my interest. I am grateful for the valuable feedback and input from my supervisors throughout my research, and recognize the role they played in shaping the direction of my work.

I am also thankful to my doctoral committee members – Prof. Shreemayee Borah, Dr. Vijaya Saradhi, and Dr. Amit Awekar for providing valuable feedback during the research. I hereby express my gratitude towards Prof. Pradeep K. Das, Prof. S.V. Rao and Dr. T. Venkatesh for all their moral support. I sincerely thank all the faculty members for their direct and indirect support.

I would also like to acknowledge the technical staff, system admins at IIT Guwahati and the ever-supportive administrative staff in the Dept. of CSE, Computer & Communication Centre, Academic Affairs, and Student Affairs, specially, Prabin Bharali, Gauri Khuttiya Deori, Rihu Mahato, Raktjeet Pathak, Bhriguraj Borah and Nanu Alan Kachari.

I am thankful to the Ministry of Human Resource & Development, Government of India, for providing me PhD fellowship. I am thankful to Dept. of Biotechnology (project no. BT/COE/34/SP28408/2018) and Dept. of CSE for providing necessary computational resources required for the thesis contributions.

I would like to express my gratitude to the many friends and colleagues who have supported me during this journey, including Pawan, Arijit, Akshay, Pradeep Bhale, Ujjwal Biswas and many others. A vote of thanks to my seniors and juniors, including Sunil Sahoo, Rakesh Pandey, M S Vasudevan, Prateek K., Abhishek, Mathew Francis, Mrinmoy Bhattacharjee, Shikha Baghel, Pankaj Chaudhary, NS Aswathy, Nidhi Ahlawat, Vanshali Sharma, and many others.

I would like to express my heartfelt gratitude to the most important people in my life. Firstly, I would like to thank my father, Mr. Ram P. Mishra, who has been a constant source of motivation and has always stood by my side through thick and thin. I am grateful for his support in all aspects of my life. I am also thankful for the

presence and blessings of my mother, Late Mrs. Krishna Mishra, whose eternal soul continues to inspire me every day. Additionally, I want to express my appreciation to my sister, Priyanka Mishra, for taking care of all responsibilities and providing unwavering support throughout my PhD journey. Furthermore, I would like to thanks Monika Mishra, D. K. Mishra, Anurag, Sonu for their care and support. Thanks to little stars Amogha, Anvika, Krati, who are source of positivity and reasons for smiles. Last but not least, I would like to extend my gratitude to the rest of my family members and relatives who have supported me unconditionally with their time, love, and blessings.

Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

Certificate

This is to certify that this thesis entitled "**Multimodal Attention Variants for Visual Question Answering**" submitted by **Aakansha Mishra**, in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy, to the Indian Institute of Technology Guwahati, Assam, India, is a record of the bonafide research work carried out by her under my guidance and supervision at the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam, India. To the best of my knowledge, no part of the work reported in this thesis has been presented for the award of any degree at any other institution.

Date: April 2023
Place: Guwahati

Prof. Ashish Anand	Dr. Prithwijit Guha
Professor	Associate Professor
Dept. of CSE	Dept. of EEE
IIT Guwahati	IIT Guwahati

Abstract

Visual Question Answering (VQA) is an exciting field of research that involves answering natural language questions asked about an image. This multimodal task requires models to understand the syntax and semantics of the question, interact with the relevant objects in the image, and infer the answer using both image and text semantics. Due to its complex behavior, VQA has gained considerable attention from both vision and natural language research community.

Most contributions to VQA focus on improving model performance by developing better mechanisms for attaining question and image representations that facilitate interaction between the two. However, despite the progress made, there is still room for improvement in terms of the accuracy of inferred answers. To address this, various methods have been introduced, such as attention mechanisms, that enable effective interaction between the two input modalities.

In this context, this work contributes to the ongoing efforts to improve VQA model performance. Specifically, novel VQA models are proposed that break down the problem into smaller components, making it easier to predict the answer. The focus is given on improving the attention mechanism for the two modalities, resulting in a richer and more accurate feature representation. This work demonstrates that improving VQA model performance can be achieved through multiple avenues, and by combining these approaches, we can achieve even better results. These findings have the potential to enhance the performance of VQA models and contribute to the development of more advanced AI systems that can accurately understand and respond to natural language questions about images.

The first model (ACA-VQA), Aggregated Co-attention based Visual Question Answering, aims to improve VQA performance by exploiting cross modality attention in multiple stages. The attention is aggregated at each stage to preserve the cues obtained from multiple stages. This proposal is benchmarked on the TDIUC and VQA2.0 dataset against state-of-the-art approaches. The experimental results demonstrated the efficacy of multistage co-attention mechanism.

The second model (CSCA-VQA) has an attention block containing both self-attention and co-attention on image and text. The self-attention modules provide

contextual information of objects (for an image) and words (for a question) crucial for inferring an answer. On the other hand, cross-modal attention aids the interaction of image and text. To obtain fine-grained information from the two modalities, dense attention blocks are cascaded multiple times. Benchmarking on the widely used VQA2.0 and TDIUC datasets demonstrates the efficacy of key components of the model and the stacking of attention modules.

The third contribution (DAQC-VQA) addresses two important issues in VQA: answer prediction in a large output answer space and obtaining enriched representation through cross-modality interactions. The DAQC-VQA system consists of three main network modules. The first module is a dual attention mechanism that helps in obtaining an enriched cross-domain representation of the two modalities. The second module is a question classifier subsystem that identifies input question category, that helps reduce the answer search space. The third module predicts the answer depending on the question category. All component networks of DAQC-VQA are trained in an end-to-end manner with a joint loss function.

Overall, this work contributes to the ongoing efforts to improve the accuracy of VQA models and enhance their ability to accurately understand and respond to natural language questions about images.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Terms	Abbreviations
VQA	Visual Question Answering
NLP	Natural Language Processing
CV	Computer Vision
AMT	Amazon Mechanical Turk
BERT	Bidirectional Encoder Representations from Transformers
CE	Cross Entropy
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
MLP	Multi Layer Perceptron
FFN	Feed Forward Network
FCNet	Fully Connected Network
KB	Knowledge Base
LSTM	Long Short-Term Memory Network
GRU	Gated Recurrent Unit
SOTA	State-of-the-Art
TDIUC	Tasks Directed Image Understanding Challenge
A-MPT	Arithmetic Mean Per Type
H-MPT	Harmonic Mean Per Type
AoI	Attention-on-Image
AoQ	Attention-on-Question
SA	Self Attention
CA	Co-Attention
ACA	Aggregated Co-attention

CSCA	Cascaded Self- and Co-attention
DAQC	Dual Attention & Question Categorization

List of Symbols

Terms	Symbol
I	Image
$\mathscr{I}$	Set of all images
q	Input natural language question
$\mathscr{Q}$	Set of all natural language questions
a	Answer corresponding to input (I, q)
$\mathscr{A}$	Set of all answers
$\hat{\mathbf{a}}$	Answer probability vector predicted by VQA system
â	Most probable answer predicted by VQA system
n_c	Total number of answers i.e. $\mid \mathscr{A} \mid$
n_{qc}	Total number of question categories
$\mathscr{L}_T$	Total loss for training VQA system
n_v	Number of salient image regions
n_w	Number of question words
$\mathbf{rI}$	Image Representation
$\mathbf{r_i}$	ResNet-101 embedding of i^{th} image region
$\mathbf{Eq}$	Question Representation
$\mathbf{eq_j}$	GloVe Word Embedding of j^{th} question word
$\theta_{\mathbf{v}}$	Aggregated visual attention
$\theta_{\mathbf{q}}$	Aggregated text attention
$\mathbf{F}_t$	Fused embedding at $t^{th}-$stage
$\mathscr{L}_t$	Loss for t^{th} stage
(K, Q, V)	Key, Query, Value matrices for multi-head attention
n_h	Number of heads for multi-head attention

$\alpha\mathbf{I}$	Attention scores for image regions
$\alpha\mathbf{Q}$	Attention scores for question word
ω_F	Fused embeddings
$\mathscr{L}_{QCS}$	Question Categorization System loss
$\mathscr{L}_{APS}$	Answer Prediction System loss

Chapter 1

Introduction

1.1 Overview

Humans rely on processing information from various modalities such as *vision*, *language*, and *audio* in their daily decision-making process. Hence, the ability to solve multimodal tasks such as Visual Question Answering (VQA), Image Captioning, Video Question Answering, Emotion Recognition, and Sentiment Analysis are considered some of the challenges for the next generation of artificial intelligence. Multimodal systems combine various modalities, thus expanding the boundaries of individual fields. In particular, multimodal systems that combine computer vision (CV) and natural language processing (NLP) have received considerable attention from the research community over the last decade [2–5].

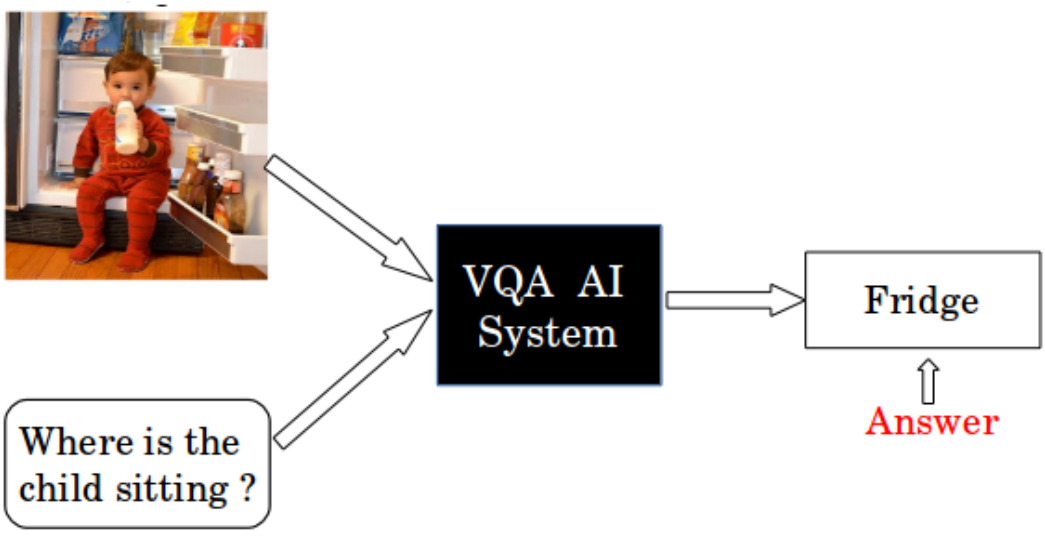


Fig. 1.1 Overview of a VQA System with a sample [6] illustration. An image is inpiut to the model and a related question, VQA system will predict the answer.

The combination of vision and language in a multimodal system equips machines with the intelligence to comprehend the content of an image and communicate in a way that mimics human behavior. This capability has given rise to a new and promising research area known as Visual Question Answering (VQA) [6, 2]. Humans are naturally adept at processing, reasoning, and analyzing visual, textual, and audio information, allowing them to answer questions related to real-world scenarios. VQA [2] seeks to replicate this ability in computational approaches, enabling machines to perform the same task. The primary objective of a VQA system [2, 6] is to generate a natural language response to questions asked about a given image. In Figure 1.1, a high-level overview of VQA task is shown. In this figure, a natural language question, "*Where is the child sitting?*", is asked about the given image, and a VQA system generates an answer "*Fridge*".

VQA is a challenging multimodal artificial intelligence task involving computer vision, natural language processing and commonsense reasoning. VQA has gained wide attention for several reasons. First, it has vast real-life applications for several human computer interaction tasks, assistive systems, automation systems etc. Some applications are as follows:

- *Assistive Navigation and Scene Interpretation* – Useful for assistance to visually impaired persons. They interact with an AI system to perceive their surroundings for deciding on further actions.
- *Surveillance Video Data Analysis* – Interactive natural language querying with surveillance systems for understanding object presence and activities in huge video surveillance datasets.
- *Teaching Purpose* – for kids through interactive sessions. Kids could interact and learn from VQA based AI system by asking questions.
- *Attribute based Annotation for huge data* – For data with defined set of attributes, annotation could be done by asking question if an object has some *'X'* attribute or not.
- *AI-based Personal Assistants* – Helpful in getting interactive assistance for routine daily tasks.
- *Object Recognition and Identification* – Can assist users in identifying as well as describing objects and reading text. A user with a visual impairment can inquire about the contents of a package, the labels on etc.
- *Accessibility in Smart Homes* – To control and interact with smart home devices and appliances through voice commands and questions. Users can be assisted about the status of lights, thermostats, security cameras etc.
- *Healthcare Support* – Can assist patients with various needs and in understanding their health conditions. For example, it can answer questions about medical instructions, explain diagrams on medical devices etc.

- *Employment and Workplace Assistance* - Could provide assistance to individuals with disabilities in the workplace by providing real-time assistance with tasks, interpreting complex visual data, and facilitating communication with colleagues.

Second, VQA, as a multimodal task, has the potential to significantly influence several downstream tasks including scene interpretation, intent detection etc. A brief discussion on VQA significance for such downstream tasks is presented in the Section 1.2.

Third, VQA poses a significant challenge for AI in comparison to conventional computer vision tasks such as image classification, object detection, and object recognition. The Visual Question Answering (VQA) task involves processing questions using commonsense reasoning and object detection, identifying relations among different objects, and understanding how they interact within an image. In some cases, it also requires accessing external knowledge bases for additional information. To infer the answer, VQA systems inherently solve multiple computer vision and natural language tasks as sub-tasks. This complexity makes VQA an AI-complete task, as it requires solving several challenging sub-problems to answer questions based on visual content. As a result, a wide range of techniques has been developed to tackle this issue, and much attention has been given to improving the performance in this area.

1.2 Significance of VQA

Visual Question Answering plays a crucial role in bridging the gap between computer vision and natural language understanding. By enabling the machines to comprehend and respond to questions asked about images, VQA surpasses the limitations of traditional unimodal tasks, promoting an improved understanding of visual content in context of text. This multimodal task finds application across numerous domains, from aiding visually impaired individuals in understanding their surroundings to enhancing image retrieval systems and robotics etc. VQA task is making machines as well as humans to understand and talk about pictures in a more natural way.

VQA has emerged as one of the multimodal task with the potential to significantly influence a spectrum of downstream tasks, including intent detection, entity extraction, scene interpretation etc. By extending the capabilities of AI models to process both images and text, it provides a richer context for understanding user queries. Downstream tasks in the context of VQA refer to tasks that can benefit from the capabilities and insights gained through VQA model. Some of these downstream tasks include:

- **Intent Detection:** Intent detection is one of the primary tasks to operate and do conversation with digital assistants to give relevant responses. Most of the existing works focus on identifying intents based on textual conversation. Intent detection could be approached as a VQA problem, where the query could be regarding the intention of the person in the image and answer shall be identified from a set of intentions. Multimodal intent detection could be more realistic as it takes into account visual modality along with text and will allow systems to identify intentions that might be ambiguous in text alone.

- **Entity Extraction:** The goal of entity extraction is to understand and categorize entities such as names of people, organizations, locations, numbers etc from a given text to gain relevant insights. Multimodal entity extraction aims to leverage relevant image information to improve the performance. To accomplish this, VQA is a possible way where entity extraction can be done by asking queries for image content for the relevant entities present.

- **Relation Extraction:** As VQA systems can be applied for entity extraction. It could further be extended to identify how the identified entities (objects, attributes) are related. The query would be regarding the spatial arrangements, contextual information in the objects while answer labels consists of possible relations and arrangements of the objects in the image.

- **Scene Interpretation:** VQA could contribute to a deeper understanding of scenes and images. This could be a valuable task in applications like autonomous driving, where a vehicle needs to comprehend its surroundings for safe navigation. Through querying VQA system autonomous vehicle can identify the objects, assess road conditions, and respond to dynamic scenarios, contributing to enhanced decision-making processes. Visually impaired can perceive their surrounding conditions by interacting through VQA system.

- **Annotating large datasets:** VQA could be used as a tool to automate the large-scale data annotation . Annotation could require the attributes, presence, location of objects present in the image. A VQA system could be trained once for the dataset that needs to be annotated. With this trained model, more data can be annotated.

1.3 Motivation

The attention mechanism was introduced in neural machine translation (NMT) [7] and has now become an inherent part of various machine learning algorithms. It helps to compute the features by putting emphasis on certain parts of the input followed by their re-weighting. This simple yet effective mechanism is found to be better than computing the global vector for the data. In VQA, the answer for textual question with reference to visual data needs an alignment of textual question words (or tokens)

with the image regions. This makes the attention mechanism a better choice for VQA compared with the global feature extraction. Is standard attention mechanism enough for VQA? This question is the primary motivational point behind this thesis. The direct alignment of the question features with the visual image is a better choice with respect to the global feature extraction. But the relation between the different regions of image is also important to be incorporated in the attention mechanism.

1.4 Potential Research Gap

In standard attention mechanism, the model attends to the visual features that are most relevant to the textual question to generate an output. However, this mechanism may not be enough to capture all the relevant information in the multimodal data. Further, multistage attention could give the VQA models the capability similar to the human behaviour of looking at the data multiple times to understand a complex task. By using multistage attention, the model can capture more fine-grained relationships between the textual and visual data. Hence, the model can improve its understanding of the visual and textual context to predict more accurate answers to the given questions. In many VQA applications, the answer space can be vast, making it difficult for the model to predict the correct answer from large search space. This motivates the development of question categorization based VQA. To address this, a question classification subsystem can be incorporated into the VQA system. The question classification subsystem can identify the category of the input question, which can then be used to reduce the answer search space. By limiting the answer search space to a subset of possible answers that are relevant to the question category, the model can improve its accuracy in predicting the correct answer.

1.5 Problem Definition

A visual question answering system $\mathscr{S}_{VQA}$ aims to estimate the probabilities of answers a ($a \in \mathscr{A}$) to an input (natural language) question q ($q \in \mathscr{Q}$) about an image I ($I \in \mathscr{I}$). Such a system is trained on the set of images $\mathscr{I}$, set of questions $\mathscr{Q}$ associated with images and set of all answers $\mathscr{A}$. This is generally achieved by using representative vector space embeddings of questions ($\mathbf{f}(q)$) and images ($\mathbf{g}(I)$) computed using deep neural networks. The answer probability vector $\hat{\mathbf{a}}$ and the most probable answer â are predicted by $\mathscr{S}_{VQA}$ as

$$\hat{\mathbf{a}} \quad = \quad \mathscr{S}_{VQA}(\,\mathbf{f}(\mathrm{q})\;,\;\mathbf{g}(\mathrm{I})\,)\hat{\mathbf{a}} = \underset{\mathbf{a}\in\mathbf{A}}{\arg\max}\, P(\,\mathbf{a}\mid \mathscr{S}_{VQA}(\,\mathbf{f}(\mathbf{q})\;,\;\mathbf{g}(I)\,)\,) \qquad (1.1)$$

The multimodal VQA task can be solved using a simple end-to-end trainable architecture, as shown in Figure 1.2. The architecture consists of three modules: feature extraction, feature fusion, and answer prediction. In this particular example, the image features are obtained from a pretrained neural network, such as CNN [8, 9], while the question is encoded using a recurrent neural network like LSTM [10]. The two feature representations are combined to produce a joint multimodal embedding. This joint embedding is then fed into a fully connected network for answer classification (within a predefined set of answers) or generation. Most of the existing literature has approached VQA as a classification task, and this thesis adopts the same approach.

VQA models are complex because they must comprehend the syntax and semantics of natural language questions, interact with relevant objects in the image based on the context of the question, and deduce the answer by combining information from both image and text semantics. Many research efforts in VQA have focused on enhancing performance by developing models that provide better mechanisms for obtaining question and image representations that facilitate stronger interactions between the two modalities. While this approach has yielded useful information, it is still essential to prioritize a correct inference of the answer.

To accomplish this, much of the research introduces various methods that facilitate strong interactions between the two input modalities. Notably, answering a question about an image requires focusing on specific parts or regions of the image. Conse-

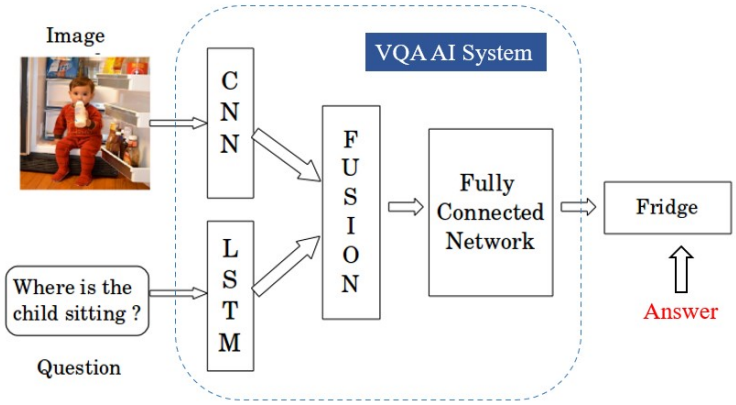


Fig. 1.2 Block diagram for basic VQA trainable architecture. Here, image is encoded through pretrained CNN and question encoding is obtained from LSTM. Both features are fused and fed to a fully connected network for answer classification

quently, there has been significant work on developing VQA methods based on the *attention* mechanism. The objective of this approach is to extract features from the attended modality, focusing more on the region that is most relevant to inferring the answer.

The main goal of this thesis is to develop VQA models that leverage:

- Multimodal information to improve the attention mechanism for each modality and obtain richer feature representations.
- Interaction between modalities by generating attention for both modalities in the context of each other using refined features.
- Prior information on question categories to increase the efficiency of answer classification by reducing the answer space.

In next section, each contribution towards thesis is summarized followed with the outline of thesis.

1.6 Contributions

There are multiple approaches to improving VQA model performance, and this thesis explores two different ways, either individually or in combination. As one of the approaches, the thesis explores enhancement of the attention mechanism for the two modalities to obtain an improved and comprehensive feature representations. The other approach is to break down the VQA model into smaller tasks, which reduces the search space for the final classification. The first and second contributions of this thesis focus on the former approach, while the third contribution combines both approaches. Figure 1.3 provides a schematic overview of the thesis. Following sections briefly describe each contribution.

1.6.1 Contribution 1 - Visual Question Answering with Aggregated Co-attention

Attention is one of the indispensable components of a VQA system [11–18]. The main objective of the first contribution is to further improve the existing dual attention mechanism by proposing the interaction between two modalities at multiple stages. Multistage attention attempts to mimic the human behavior to understand a complex scene (or image) or text by looking at the scene or reading the text multiple times. The proposed multistage attention based model first extracts the faster-RCNN-based visual features and LSTM-based encoding of question. These features are further

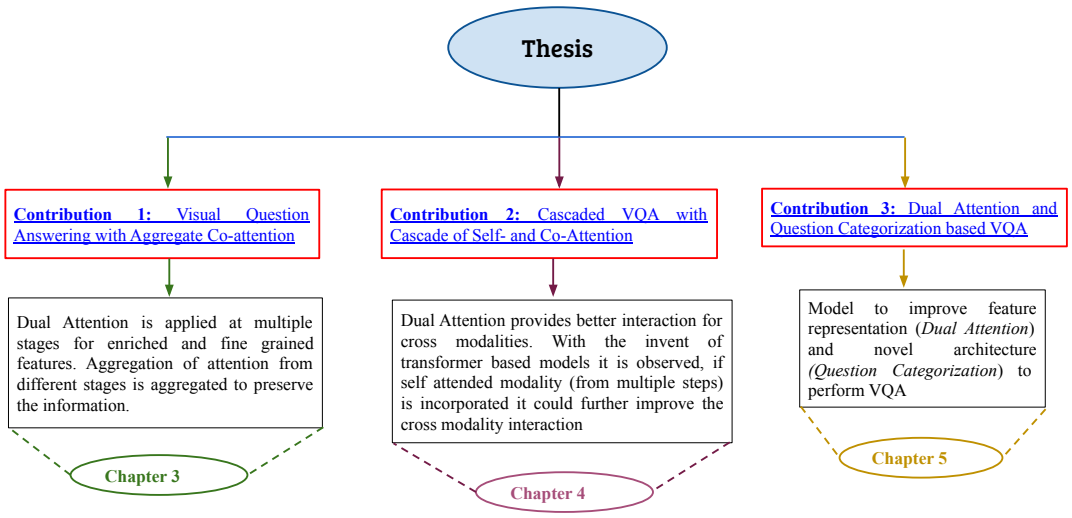


Fig. 1.3 Overview of thesis contributions

refined using cross-modality attentions in an iterative manner in multiple stages. Two types of cross-modality attention, namely, *Attention-on-Image* **(AoI)** and *Attention-on-Question* **(AoQ)** are used in the following manner. At a particular stage, say t, AoI is applied and thus obtained attended visual embedding is fused with LSTM encoding of question. And AoQ is employed on word embeddings of question in the context of the attended visual representation obtained from AoI. Thus, attended word embeddings are passed through LSTM to obtain attended question representation. To preserve attention scores from different stages for each modality, *aggregated attention* is incorporated. Aggregated attention implies the following."At any stage of attention (*visual* (AoI) or *textual* (AoQ)), the attention score will be taken as addition of attention scores from all previous stages of the corresponding modality". To perform the final answer classification by the proposed model, unified visual embedding obtained from each stage of *visual attention* are fused via point-wise multiplication. Similarly, question encoding from each stage of textual attention is combined to obtain the final textual features. These embeddings are fused to obtain the joint representation. This joint visual and text representation is used for the final model classification.

1.6.2 Contribution 2 - CSCA: VQA with Cascade of Self- and Co-Attention Blocks

The first contribution focuses on dual attention mechanism to reflect cross modality interactions or relations. It ignores emphasizing the relations of objects within image and relations among words within question. Hence, the second contribution includes *self-attention* on each modality along with dual-attention. Embeddings obtained after self-attention encodes contextual information within a single modality and that is used

to generate dual-attention based encoding. The process of *self-attention* and *cross-attention* comprises a block of dense attention mechanism. Such a dense attention block is employed in multiple stages to obtain enhanced representations.

1.6.3 Contribution 3 - Dual Attention and Question Categorization based Visual Question Answering

The third contribution aims to develop a novel architecture of two levels for VQA along with dual modality attention. The first level acts as a question classifier. It classifies the given question into one of the pre-defined question categories. The second level consists of the multiple answer classifiers. Each answer classifier predicts answers amongst the answer subset belonging to a particular question category. This type of model reduces the answer search space as the answer to be predicted is only from the set of answers that are candidates for a question category and not from overall answer set. To extract better features, attention is applied on both the modalities in context of each other. To accomplish this, initial feature extraction is performed for both modalities (salient regions of image and LSTM encoding of question). Next, correlation is computed for each salient region of image in context of question to compute the attended visual representation. For each question, word attention score is calculated based on their correlation with attended and refined visual representation. A fused representation of the two modalities is obtained by combining the attended encodings. The fused embedding is then passed to a fully connected network to classify the question category. At the next level, one single classifier, corresponding to the question category predicted from previous level, is activated from a set answer classifiers.

For all three model proposals, extensive experiments are performed on two widely used publicly available VQA datasets, VQA2.0 [6] and TDIUC [1]. The comparative analysis with existing models have shown that the proposed models obtain competitive performance compared to relatively more complex models and outperforms several baseline models.

Outline of thesis

This thesis is organized as follows:

- **Chapter 1** presents the problem definition with motivation and a brief description of each contribution.
- **Chapter 2** discusses the existing literature for VQA. More detailed description is given for the methods in literature that are related to contributions of thesis.

- **Chapter 3** introduces the *'Multistage Aggregated Co-Attention based VQA'* model, which is based on iterative interactions between the two modalities. This model improves upon the attention mechanism by using a multistage aggregation approach that gradually incorporates information from both the question and image modalities.

- **Chapter 4** proposes a *'Dense Interaction Mechanism'* to enhance the interaction between the two modalities and obtain a more enriched representation. This mechanism uses a dense block architecture that facilitates multiple interactions between the two modalities to create a more comprehensive feature representation.

- **Chapter 5** presents the dual-attention and question categorizer-based VQA model (DAQC-VQA). This model uses a dual-attention mechanism that attends to both the question and image modalities and a question categorizer that reduces the answer space by categorizing questions based on their characteristics.

- **Chapter 6** concludes this thesis and provides a summary of the contributions and their impact on the VQA field. This chapter also identifies potential areas for future research that build upon the work presented in this thesis. Additionally, this chapter reflects on the significance of the contributions made to the field of VQA and discusses their implications for real-world applications. The findings and insights gained from this thesis can be useful in areas such as image and video search, autonomous driving, and robotics, among others. Furthermore, this chapter discusses the limitations of the proposed models and suggests possible directions for addressing these limitations in future research. The importance of developing more efficient VQA models that can be deployed in resource-constrained environments is also highlighted. Overall, this final chapter provides a comprehensive overview of the work presented in this thesis and highlights the potential impact and opportunities for future research in the field of VQA.

Chapter 2

Literature Survey

Chapter Highlights

- Summary of literature review of various VQA methods is presented.
- The various VQA datasets are briefly discussed.
- Proposed thesis contributions are summarized.

2.1 Introduction

This chapter presents relevant existing works for the VQA task. VQA systems can be broken into a modular structure with *feature extractor*, *multimodal fusion* and *classifier* being the fundamental modules. The objective of the *feature extractor* module is to obtain a representation or embeddings of the image and text modalities. Image embeddings are also referred to as visual representation or visual embeddings. Often such representations are obtained independently for each of the two modalities. Section 2.2 discusses most commonly used methods to obtain visual and text representations.

Multimodal fusion module combines representations of individual modalities to get a single or fused representation of the two modalities. The fused representation is fed to the *classifier* module to get a final answer. Section 2.3 discusses the different fusion methods adopted for the VQA task.

Final stage of answer prediction could be formulated as a sentence generation or a classification problem. In the sentence generation formulation, a decoder-based model can be designed. However, answer prediction as a classification task is the most commonly adopted approach and this thesis has used the same using the *classifier* module. The classifier module uses the fused representation as an input to a classifier.

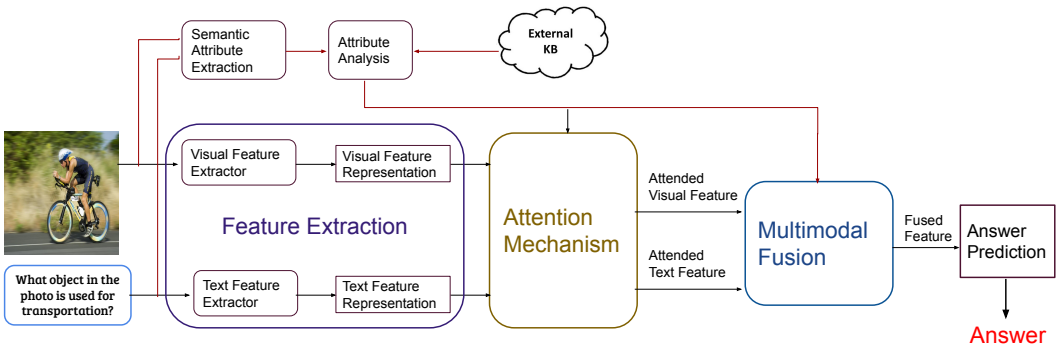


Fig. 2.1 General overview of VQA system with core modules. Feature extraction (visual and textual) is a common part for all approaches. For external knowledge based methods *semantic attribute extraction* is an additional module. Further these features are input to do the features improvement through attention mechanism. Attended feature representation are then fused to fed into answer classification network.

Figure 2.1 illustrates a view of such a modular structure of a VQA system. Apart from these fundamental modules, *attention* and *external KBs* modules help in obtaining enriched representation of the two modalities. For example, the attention mechanism module aims to interpret "where to focus" in one of the modalities. Some methods identify such regions in the image in context of the given question. While, some other methods identify such regions in both the modalities in context of the other modality. Recent methods also give emphasis on internal correlations among image regions and question words, and thus aim to encode internal feature dependency as well. This thesis discusses the different variants of attention mechanism in Section 2.4.

Subsequent sections discuss the existing works based on this modular structure.

2.2 Feature Extractor

This section discusses most commonly used feature extractors for the image and text modalities for the VQA task.

Visual Feature Extractor: The objective of this module is to obtain a representation for the given image. Initial methods utilized pretrained CNN-based networks for visual feature extraction. VGG [8] and ResNet [19] are the most commonly used such pretrained CNN networks. These models were trained on ImageNet [20] dataset for image classification task. Another set of methods [11, 21, 22] extracted features from customized CNN in an end-to-end framework.

Recent feature extractors use salient regions extracted through object detection by faster-RCNN [23]. The representation of these salient regions are then obtained from

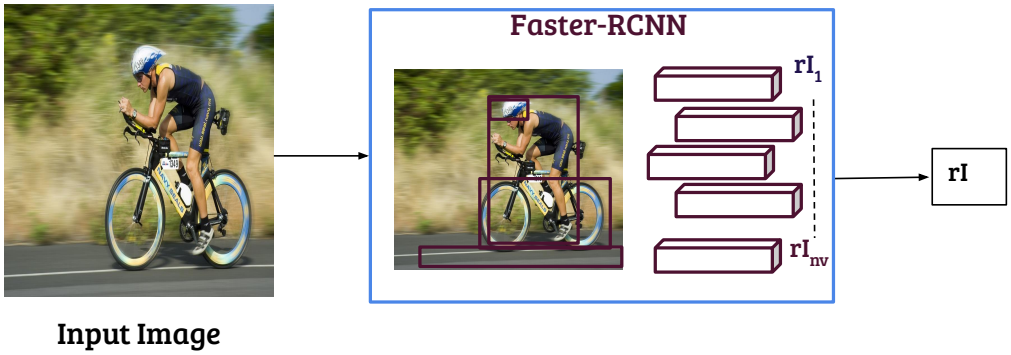


Fig. 2.2 Block diagram for visual feature extraction. Here, rI represents the ResNet-101 feature representation for n_v salient regions detected from faster-RCNN

pretrained ResNet [19]. ResNet-101 (ResNet with 101 layers) embeddings of these salient regions is used to represent the visual features. Figure 2.2 gives an overview of salient region based visual feature extraction.

To model complex relations for VQA task, visual and textual features are also represented through a graph [24–26]. Each node of graph correspond to image regions or words in the question. Edges of graph represents relation amongst different regions / words. The edges between nodes capture the relationships between them, such as spatial and semantic relationships between image regions and words in the question. By constructing these graphs, the model can reason over the relationships between visual and textual information, enabling more accurate and detailed answers to be generated. Graph-based features are able to capture the complex interactions between the visual and textual components of the VQA task, leading to improved accuracy and robustness of question answering systems.

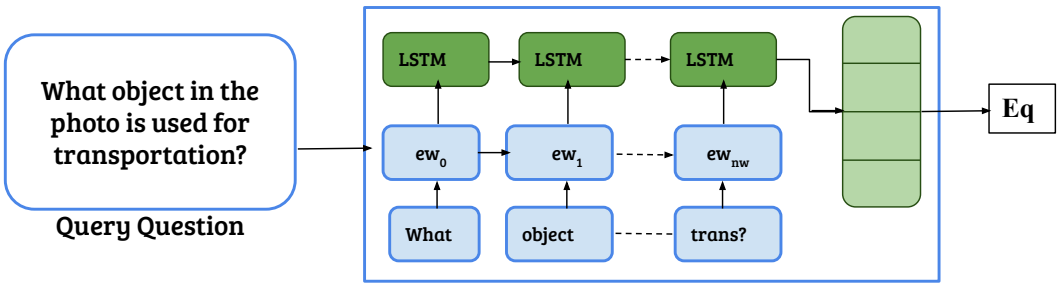


Fig. 2.3 Block diagram for textual feature extraction. Here, ew_j shows the GloVe word embedding of j^{th} word, n_w is the number of words in question. **Eq** is the question encoding obtained from the last hidden layer of LSTM.

Textual Feature Extractor: The objective of this module is to obtain a representation for the question. Initial approaches used one-hot encoding to represent words of the given question, which are fed to a LSTM network. Embedding or representation of the question is obtained from the last hidden layer of the LSTM network. Recent question feature extractors use pretrained word embeddings instead of one-hot encoding. Examples of most commonly used pretrained word embeddings include word2vec [16] and GloVe [12, 27]. Textual feature extraction from pretrained word embedding and LSTM is presented in Figure 2.3.

2.3 Multimodal Fusion

Feature fusion is an indispensable module for multimodal tasks such as the VQA task. Figure 2.4 illustrates the basic building blocks of fusion based VQA model.

Early methods for VQA task adapted the features extracted from pretrained deep networks. As discussed in the previous section, visual features are extracted primarily from last hidden layer of pretrained deep convolution network such as VGG [8] or ResNet-101 [19] trained on ImageNet [28] dataset for classification task. For textual features, question words are represented from GloVE [29] embedding and fed to LSTM [30]. Last hidden layer of LSTM is exploited as question encoding.

Antol et al. [2], in one of the first prominent works in VQA, used elementwise-summation to combine visual and text representations. They used pretrained VGG to get the image embeddings, while LSTM over one-hot encoding of question words were used to obtain the text representation. Jabri et al. [27] primarily targeted the multiple choice based questions. They concatenated image, question, answer triplet

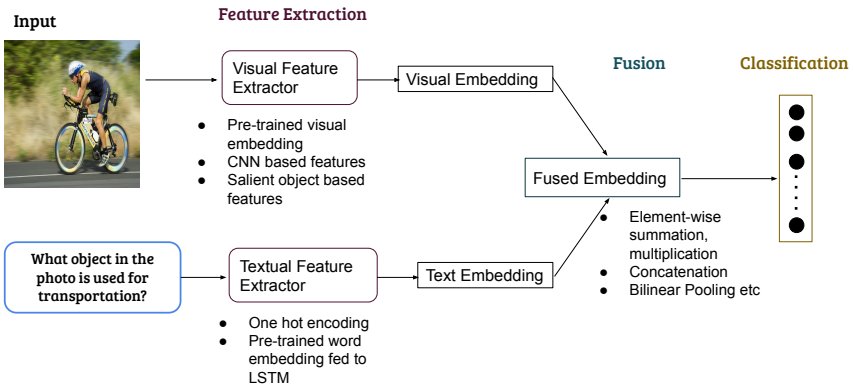


Fig. 2.4 An overview of fusion based VQA method. Features are extracted for visual and text modality through pre-trained network. After projecting the feature embeddings to same dimension, features are fused through elementwise summation, multiplication etc. for classification

encodings and fed to a logistic regressor. The regressor predicts whether the given triplet is correct. Visual features were obtained from pretrained ResNet-101 model. The question is encoded as an average of word2vec embeddings of each words.

Shrestha et al. [31] proposed unified model (RAMEN) for the VQA task that performs well on datasets comprised of images from different domains (synthetic and real-world images). It concatenates the features from two modalities at early stage. These fused bimodal embeddings are projected in a shared space for learning inter modality relationships. Further, a bi-directional GRU is used to aggregate the bi-modal embeddings and question encoding to capture the recurrent interaction. This aggregated representation is fed for answer classification.

Ren et al. [32] and Malinowski et al. [33] considered answer generation for the VQA task. Ren et al. in [32] has extracted the image features from last hidden layer of VGG net. This image embedding was treated as the first word of the question forwarded to a LSTM network. The answers were generated by taking the outputs from last hidden layer of the LSTM network. Authors in [33] fed CNN-based image features along with one hot encoding of question words to a LSTM network. Output of the LSTM network was used to generate answers.

Multimodal convolution is proposed by Ma et al. in [21] for the VQA task. After convolution, dual modality features were obtained by flattening the feature maps from the last layer. These features were then fused for answer classification. Gao et al. [34] proposed the fusion of multi-modality at early stages by a question-guided convolution kernel. It helped extract better spatial information, as kernels were generated based on the language features to convolve with the visual features.

Multimodal Compact Bilinear (MCB) pooling was introduced by Fukui et al. [35] to capture the complex interaction between the two modalities. It uses the outer product-based [36] interaction between visual and textual modalities and outputs a high-dimensional feature representation. The outer product is an expensive operation as each and every element of one modality interacts with that of others. To address the above problem, MCB leverages the approximation-based approach. This approach could be presented as their convolution instead of explicitly performing the outer product of two feature vectors. MCB outperforms the simple fusion mechanisms at the cost of computation and resource requirements. In order to deal with the complexity and computation issues in MCB, Kim et al. [37] have proposed another bilinear pooling-based solution termed Multimodal Low-rank Bilinear pooling (MLB). MLB is based on the Hadamard product of two modalities with two low-rank projection matrices. It could generate a low-dimension output vector and thus have fewer parameters. Though MLB's output is low-dimension, it is observed that it converges slowly. Multimodal Factorized Bilinear pooling (MFB) [38] was introduced to overcome the issues of

obtaining compact output features with robust, expressive representation like MCB and MLB. MFB is inspired by matrix factorization, where projection matrices are factorized as low-ranked matrices. As a natural extension of bilinear pooling, authors have also proposed generalized high order pooling, Multimodal Factorized High order pooling (MFH) [39] cascades multiple MFB blocks to learn the better and richer representations.

MUTAN [13] further reduces the parameter in the bilinear pooling-based approaches by decreasing the mono-modal embeddings' size and modeling their interaction as accurately as possible with a full bilinear fusion scheme. BLOCK [40] introduces the block-term decomposition for reducing the model parameters for bilinear fusion. Block Term Decomposition Pooling (BTDP) [41] is another bilinear interaction-based method that performs sparse bilinear interactions between modalities. It exploited the Block Term Decomposition theory [42–44] of tensors, resulting in a sparse and learnable block-diagonal core tensor for multimodal fusion. It is equivalent to conducting multiple tiny bilinear operations in different feature spaces.

DMRNet [45] has proposed multi-graph reasoning and fusion (MGRF) layer. It adopts pretrained semantic relation embeddings to reason complex spatial and semantic relations between visual objects. These relations are fused adaptively. Multiple layers of the MGRF module can be stacked to form Deep Multimodal Reasoning and Fusion Network (DMRFNet) for better reasoning and robust fused embedding. Lao et al. [46] has proposed a Multi-stage Hybrid Embedding Fusion (MHEF) mechanism, which comprises Dual Embedding Fusion (DEF), Latent Embedding Fusion(LEF), and Hybrid Embedding Fusion(HEF). DEF transforms one modality embedding into the reciprocal embedding space before fusion. Subsequently, DEF is incorporated with LEF to obtain novel HEF. HEF is applied in multiple stages to obtain better feature fusion.

The fusion of two modalities is an irreplaceable module for this task, but the representations that are to be fused always have a scope of improvement. To achieve the same, this thesis has contributed towards learning better cross modality interaction to obtain improved feature representations. The following section elaborates on the literature for attention mechanism based models for VQA.

2.4 Attention based Methods

In a VQA system, the attention mechanism is the way to weigh the features by correlating one modality in the context of another. Highly correlated features will get more attention compared to less correlated ones. Application of attention mechanism has evolved from *visual attention* [11, 47–50, 26, 51–53] to *co-attention* or *dual*

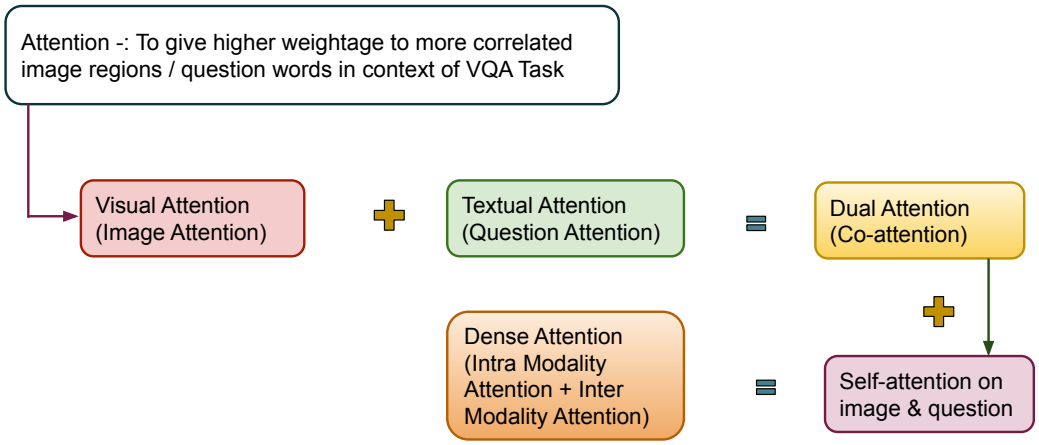


Fig. 2.5 Schematic representation for evolution of attention mechanism for VQA task.

attention or *cross-attention* [54] and is headed towards *dense attention* [55, 56]. A schematic representation of the evolution of attention mechanism is presented in Figure 2.5.

Early attention-based VQA models [11, 34, 47–50, 26, 51–55], focused on the image region(s) that is (are) most relevant to the given question. Models may capture irrelevant information while looking at the entire image, and that may adversely impact performance. The use of attention mechanism in VQA models has given significant performance improvement as the question mostly requires to focus on a small portion of the image. Thus, attention mechanism has become an integral part of every VQA model. In VQA, *visual attention models* aim to interpret “where to look" in the image for answering the question. In figure 2.6, the flow of *visual attention* based model is presented.

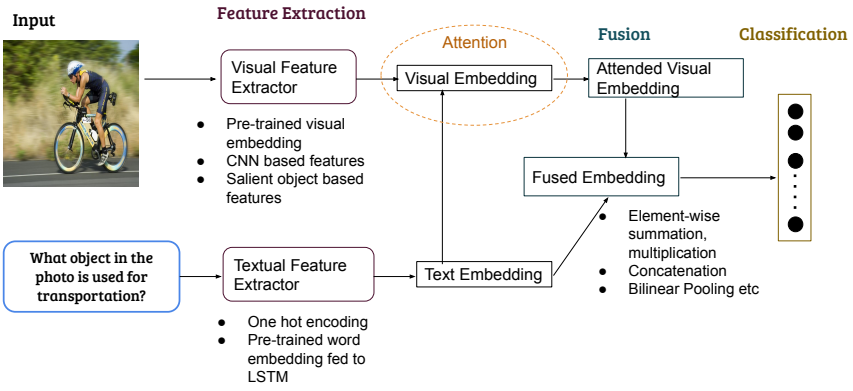


Fig. 2.6 An overview of *visual attention*. Attention is put only on visual modality in context of question. This attended visual embedding is then fused with question encoding for answer classification task.

Shi et al. [47] has proposed a model that predicts the answer by selecting an image region which is most relevant to question text. Yang et al. [11] proposed a multi-step attention-based method that allows reasoning over fine-grained information present in a question. They used question embeddings to generate attention distribution over image regions. The learned attention score is used to weight the image region embeddings. Weighted sum of image region embeddings is used as a visual feature for next step. The model proposed by Wu et al. [57] generates multi-step attention to reason over objects and progressively infer the answer. Sun et al. [58] introduced second order based visual attention module derived from multiple glimpses of visual attention. Farazi et al. [59] has proposed a question agnostic attention mechanism that first identifies object maps in the image. Further, attention is generated for visual features in context of the identified object maps.

An approach proposed by Anderson et al. [12] combines the top-down and bottom-up attention modules and shows significant improvement in the performance. In this work, the bottom-up model detects salient regions extracted using Faster-RCNN [23], while the top-down mechanism uses task-specific context to predict the attention score of the salient image regions. Shi et al. [16] proposed Question Type guided Attention (QTA), which used semantics of question category to generate attention on bottom-up, top-down image features extracted from ResNet and faster-RCNN respectively. Noh et al. [60] proposed a recurrent deep neural network with attention mechanism, where each node in the network can predict the answer. To optimize the network parameters, loss is aggregated from all units. Xi et al. [17] introduced a VQA model based on multi-objective visual relationship detection, where relevant image regions are extracted from question-guided attention. Further, an analysis of interrelationships between salient objects is given by word vector similarity. Here, the primary objective was to improve the detection of inter-relations among objects. Ding et al. [18] proposed two attention mechanisms, namely, *stimulus-driven* and *concept-driven*, which are inspired by human psychology for image caption generation tasks. Kim et al. have proposed Bilinear Attention Network (BAN) [61] that generates an attention map for two modalities from Hadamard product-based interaction. Further it uses a low rank bilinear pooling based fusion of two modalities for task of answer classification. Do et al. [62] have proposed the attention mechanism comprising trilinear interaction of image, question, and answer. As answers are unavailable during the test phase, knowledge distillation is used to transfer knowledge from the trilinear model to the bilinear model.

A few approaches exploit visual information in multiple ways from the image for a more informative visual representation. Lu et al. [53] proposed attention for image regions [19] and object proposals [23, 12]. The attended features (image regions and object proposals) are fused with question features via multiplication and projected to

a common space. Huang et al. [63] proposed object-level grounding for generating attention. They observed that, along with attention to image object regions in the context of question words, it is informative to generate the semantic similarity between question words and object labels.

Along with visual attention, attention to text also gives informative cues to infer the answer. All the question words are not equally important to answer the question. Only a few words would be more relevant. The attention on the words that leverage the visual space and vice versa is known as *co-attention*. In the VQA literature, *dual-attention* or *cross-modality attention* are also used to reflect the co-attention. Figure 2.7 shows the basic flow of *dual attention based* methods.

Lu et al. [54] have proposed that attention over the question, i.e., "what to see" is equally important as "where to look" to answer a question. Question embedding are extracted at the word level, phrase level, and question level. The attention is applied to the image and question at each level in parallel or alternatively. The parallel mechanism attends question and image simultaneously, while the alternative mechanism sequentially alternates between generating image and question attention. The usage of a stack of dense co-attention layers was proposed by Nguyen et al. [64]. Here, each word of a question interacts with each image region to generate image attention and image regions generates attention for each word of the question. Later attended embeddings are fused to feed into the classification network. The co-attention blocks are stacked in multiple layers to obtain refined representation. Zhang et al. [56] have proposed co-attention on each feature of image region guided by the question and each feature of the question guided by the image in multiple stages. To integrate the features, bilinear fusion is exploited with a residual module for multiple glimpses of images and questions.

The co-attention mechanism can be further strengthened using better intra-modality encoding. With the proposal of transformer [65] based dense attention mechanism,

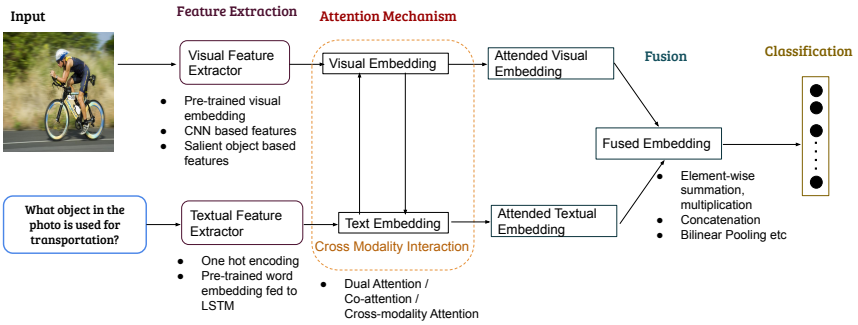


Fig. 2.7 An overview of *dual attention*. Here attention is given to both the modalities in context of each other to obtain the attended dual modality representation. These attended representations are then fused for answer classification.

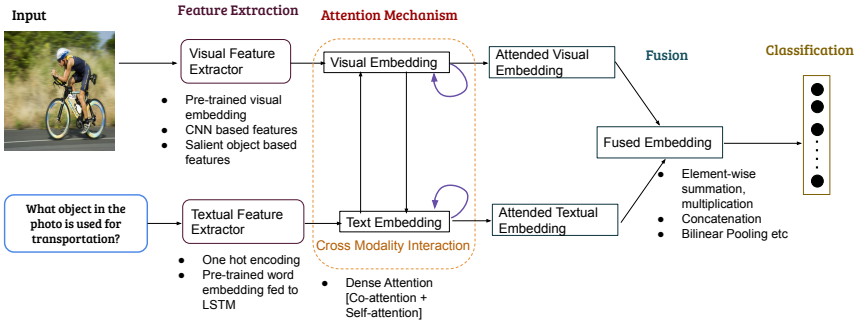


Fig. 2.8 An overview of *dense attention*. Dense attention mechanism incorporates attention on dual modality in context of itself and other modality.

a significant improvement was observed for text-based tasks. Recently, transformer-based dense attention has been introduced for multimodality tasks. Dense attention in multimodality tasks encodes intra-modality, and inter-modality features [14, 15, 66–69]. Figure 2.8 presents a general overview of *dense attention* based VQA models.

Gao et al. [14] proposed Dynamic Fusion with intra-and inter-modality Attention Flow (DFAF), a stacked network that uses inter-modality and intra-modality information for fusing features. It uses the average pooled features that can dynamically change intra-modality information flow. Yu et al. [66] proposed a deep co-attention network that follows encoder decoder-based architecture to generate dense self-attention and co-attention. It helps to obtain the fine-grained features for multimodal tasks. Multimodal Latent Interaction (MLIN) was proposed in [15] that leverages multimodal reasoning through the process of summarization, interaction, and aggregation. Lu et al. [69] proposed BERT architecture for multimodal (vision and language) learning. This model is pretrained on a large caption dataset for better transfer learning. Further, these pretrained models are fine-tuned for VQA tasks. Tan et al. [68] encoded the vision and language through a large-scale transformer model termed LXMERT. It is pretrained with a large amount of vision and language data on five multimodality tasks. These tasks helped in learning both intra-modality and cross-modality relationships.

Cross-modality attention is crucial for VQA systems and could be improved by exploiting multiple stages. And further, if information from multiple stages are preserved and flows in other subsequent stages, it could enhance the attention mechanism. At each stage, attention provide cues in a different way; hence a final decision should rely on all such cues. The cross-modality attention can be further enhanced by encoding fine-grained information of two modalities in an end-to-end manner.

Apart from the above-discussed VQA models, there are other models that utilises additional information. However, such extra information is not used by the proposed models of the thesis. Following section discusses other sets of VQA methods for the sake of completeness.

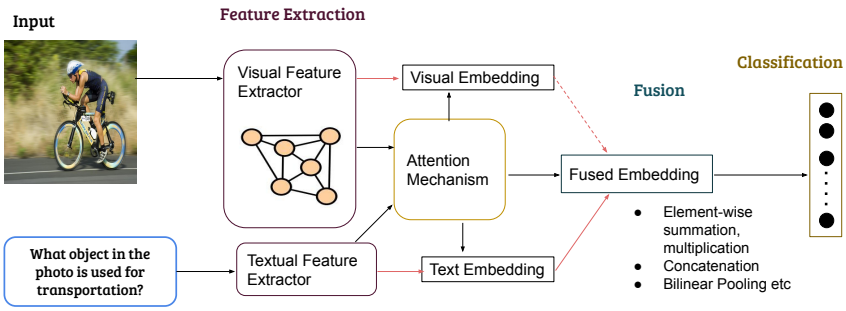


Fig. 2.9 An overview of *graph based* methods. Here the visual features are encoded as scene graph to model the relational reasoning capability to the models. These embeddings are treated as visual features to preform attention or directly fused with text modality.

2.5 Other Methods

Graph based Methods-: Attention based models have achieved a significant promising performance for VQA task. But these kind of models do not always have sufficient capability to deal with complex questions that require high level reasoning, like 'counting' related questions. At this end graph based models performs well. A general flow of *graph based* VQA model is presented in figure 2.9.

Initially, Teney et al. [24] have introduced graph based model for "abstract scenes" VQA. It uses a graph-based representation for image and question. The image is represented as a scene-graph while question as a parse tree. Zhang et al. [25] first utilizes graph model that specifically targets "counting" based questions. This method highly relied on the engineered relations between nodes of graph. Narasimhan et al. [70] first attempted to use graph convolution network (GCN) to answer factual questions based on the knowledge graph. This model mainly concentrated on the entity-relation graph extracted from the image and knowledge graph. However, this method heavily relies on external knowledge graph related to the domain. Question-Conditioned Graph (QCG) model was proposed by Norcliffe et al. [71]. Here, the objects proposed from faster-RCNN act as nodes and edges define the interaction between regions conditioned on question. For each node, a set of nodes is chosen from the neighborhood using strongest connection criterion. This leads to a question specific graph structure.

Graph based models in VQA mainly defines the similarity between objects as their semantic relationships. The difference between objects could also be more informative for establishing the relationship between nodes in the graph. To achieve this Wu et al. [26] has proposed an object difference based graph learner (ODA), that learns the semantic relations between the objects of image guided by question. By learning these relationships image was represented as an object graph encoded with structural

dependencies between objects. Xiong et al. [72] proposed to construct an entity-attribute graph from an image. A classifier is trained to infer the missing information that are crucial for answering the queries. And final answers are predicted with graph pattern matching. Cadene et al. [73] has proposed a multimodal relational network (MUREL) learned to reason over image regions based on interaction with question and models the relation between every pair of regions. To model the complex questions, authors in [74] have proposed a graph-based attention network *Relation aware Graph Attention Network (ReGAT)* to encode images in graphs. They have shown learning the inter-object relation in the question context through graph attention improves the model performance.

External Knowledge based Methods-: The question asked for an image can be very complex. All questions cannot be answered about an image using only the visual information present in it. Sometimes, to answer the question related to an image requires the external source of knowledge, e.g., *"Is the animal shown in picture is vegetarian?"*. It's difficult to answer the question without any external knowledge source. The knowledge-based VQA models try to leverage facts from an external knowledge base. External knowledge base (KB) helps as a guide for image question answering. Examples of such KBs include DBpedia [75], OpenIE [76–78], Freebase [79], NELL [80] , YAGO [81, 82], WebChild [83, 84], and ConceptNet [85]. KBs store information in computer readable structured format for efficient extraction. They consist of information like common sense knowledge, encyclopedic knowledge, and visual knowledge. This knowledge could be extracted through semantic attributes from image. Figure 2.10 presents the general work flow of *external knowledge based* VQA models.

Some popular KB-based VQA methods are: 'Ahab' proposed by Wang et al. [86], Ask Me Anything by Qi Wu [87] and FVQA by Wang et al. [88]. In 'Ahab', visual features are first extracted from an image using CNN and using these features, a query

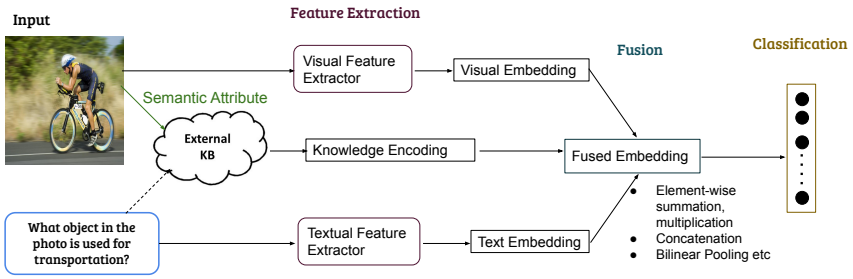


Fig. 2.10 An overview of *External Knowledge Base* method. Alongwith visual and text features, additional features are also extracted from an external knowledge base by querying through semantic attribute. Finally these knowledge based features are encoded to fuse with dual modality features for answer classification.

is made onto the KB. The answer is obtained by summarizing the results of the query. FVQA [88] was proposed to improve these methods by using LSTM to map visual features to queries. Ask me anything [87] uses a common embedding approach on top of the information from KB. Similar to the above two methods, visual semantic features are extracted using a CNN. A short description of the image is generated from KB using these features. This short description is then embedded using Doc2Vec into a fixed-sized vector. The embedded vector is finally fed into an LSTM to produce an answer.

Wu et al. [89] have observed that that CNNs and RNNs are not able to capture the high level concepts. To achieve this a model was proposed that is able to capture high level concepts through external knowledge for top-k attributes of image in text format. Zhu et al. [90] has proposed dynamic model that iteratively asks queries from the external knowledge source. The knowledge acquired from iterative queries is repeatedly stored in a memory bank after encoding. Another round of queries is made based on the knowledge acquired in memory bank from previous iterations. Narsimha et al. [91] proposed a knowledge base of facts associated with visual content. These facts are formatted in a way to identify the visual concepts in the image, an attribute or phrase associated with the visual entity, and relation between the entities. This knowledge has significantly improved the model performance. The model proposed by Song et al. [92] was based on commonsense reasoning and cross-modal BERT. To add commonsense along with image and questions encoding, relevant entities (bounding box) information was added from an external KB in form of a sentence. Authors of [93][94] figured that existing KB methods inject the information without selection. It resulted in noise for reasoning and hence led to several wrong answer predictions. To deal with this, they represented an image by a multimodal heterogeneous graph, which contained multiple layers of information corresponding to the visual, semantic and factual features. An intra-modal graph convolution network extracted relevant information from each modality and another cross-modal graph convolution aggregated information from cross modality. This process of selection was stacked multiple times to perform iterative reasoning predicted the optimal answer. Gui et al. [95] identified that using external knowledge just based on tags, or relevant concepts may not always be appropriate to add and could result in noisy information. They have proposed a novel way to extract knowledge (implicit and explicit). The implicit knowledge was added by using new prompts that extracted tentative answers and supporting evidence from a frozen GPT-3 model. To add the explicit knowledge, a contrastive learning-based knowledge retriever using the CLIP [96] model was added, where all the retrieved knowledge were centered around visually-aligned entities. These methods had high explainability in terms of the way they arrived at the results. Methods that used a single complicated CNN to map images and questions directly to answers, gave little insight into the computations performed to get the answer. On the contrary, KB

methods defined sequence and structured steps that give an advantage to these models when trying to understand the internal dynamics.

2.6 Dataset Description & Evaluation Metrics

TDIUC and *VQA2.0* are the two most commonly used datasets for the evaluation of VQA models. This section discusses these two datasets and the corresponding evaluation metrics. Other VQA related datasets are also briefly presented for completeness.

2.6.1 Task Directed Image Understanding Challenge Dataset (TDIUC)

Dataset Description: TDIUC[1] [1] is the largest available VQA dataset of real images. It consists of $1,654,167$ open-ended questions of 12 categories associated with $167,437$ images. The dataset provides categories of questions associated with images explicitly.

The questions in TDIUC are acquired from the following three sources: questions imported from existing datasets, questions generated from image annotations, and the questions generated through manual annotations. Figure 2.11 shows the category-wise sample distribution of questions. The distribution is highly biased due to the collection of the images and questions from natural sources. Few question classes and images are more frequent in nature; while, others are rare. The most significant number of questions (approximately 0.65 million) are in the 'Object Presence' (with Yes/No answers) category. On the other hand, the least number of questions (only 521) lies in the 'Utility Affordance' category. Studies [6, 97, 98] have shown that VQA models get affected by language prior bias. To avoid such issue, TDIUC introduced a special category 'Absurd'. This category contains questions having no semantic relation with the associated images. Such questions have a single answer, and that is 'Does-Not-Apply' [1]. Presence of the 'Absurd' category forces models to learn appropriate relations between the question(s) and the visual contents of the image(s) and prevents them to answer blindly with language prior.

Evaluation Metrics

Three evaluation metrics are defined by [1] for the VQA task. These are *Overall accuracy*, *Arithmetic-Mean Per Type (MPT)* and *Harmonic-Mean Per Type (MPT)*. The *Overall accuracy* is the ratio of the number of correctly answered questions to the total number of questions. VQA datasets are highly imbalanced as a few question

[1] https://kushalkafle.com/projects/tdiuc.html#download

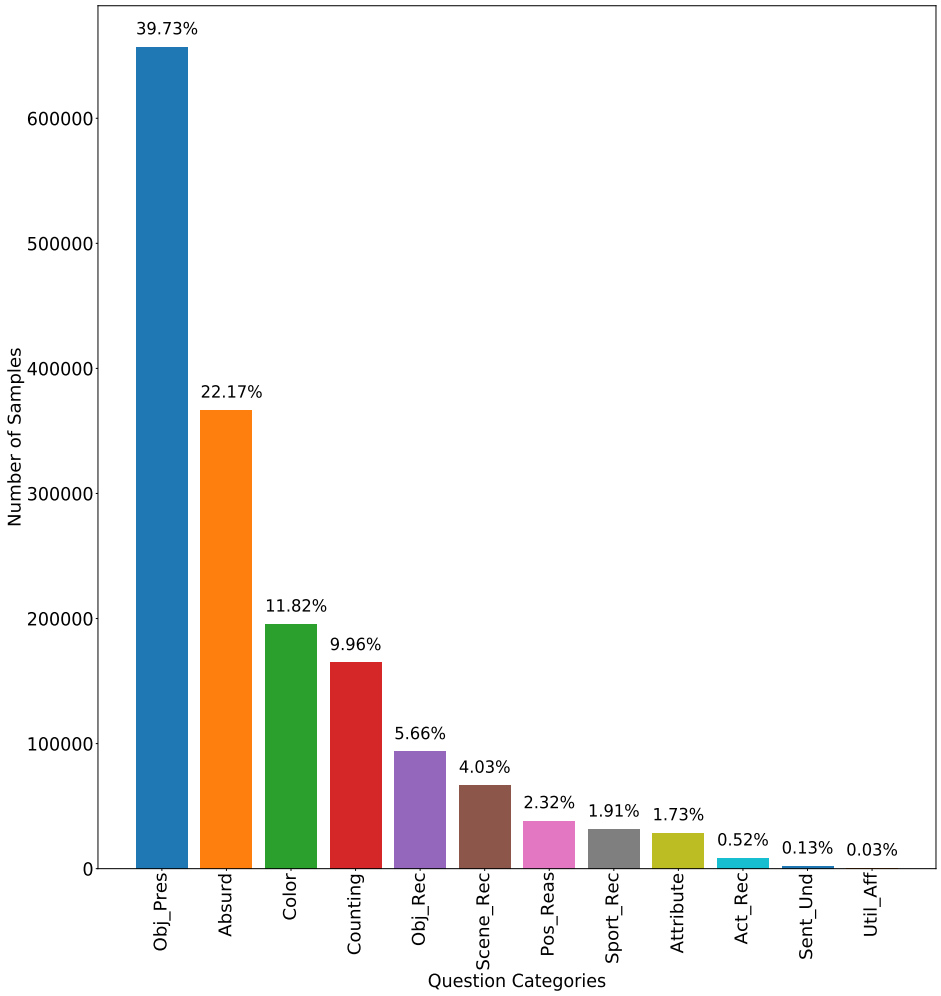


Fig. 2.11 Distribution of 12 Categories of TDIUC Questions [1].

categories are more frequent than others. *Overall accuracy* is not a good evaluation metric for such cases. The other two metrics *Arithmetic-Mean Per Type (MPT)* and *Harmonic-Mean Per Type (MPT)* [1] are generally used to achieve unbiased evaluation. *Arithmetic-MPT* computes the arithmetic mean of the individual accuracy of each question category. This evaluation metric assigns uniform weight to each question category. *Harmonic-MPT* reports the harmonic mean of individual question category accuracy. Unlike Arithmetic-MPT, the Harmonic-MPT measures the ability of a model to have a high score across all question categories.

2.6.2 VQA2.0

Dataset Description: VQA2.0[2] is one of the widely used VQA dataset of real images with a total of 0.7M question image pairs partitioned into train, validation and test set. For each question image pair, 10 human annotated answers are given. Figure 2.12 presents category-wise frequency of question in VQA2.0 dataset. VQA2.0 was introduced to reduce the language bias that existed in its preceding version. Agrawal et al. [6] came with an intuition for dealing with language prior that if a question is asked for two similar images with different answers, then to some extent most frequent answers that model infers blindly could be reduced (as same question has different answer). VQA2.0 [6] has been created to address language prior bias present in VQA1.0. For instance, for most of the questions related to sports in VQA1.0 dataset answer is “tennis” and for binary questions, answer is “yes”. As a result, models that answered “tennis" for every sports question and answered “yes" for every binary question give high performance to the overall metric. To overcome this, in VQA2.0, for each triplet (I,Q,A), where I is image, Q is question, and A is the answer respectively, another triplet (I',Q,A') is introduced. Here question Q makes sense for the complementary image I'. However, the answer A' to the question is something

[2]https://visualqa.org/download.html

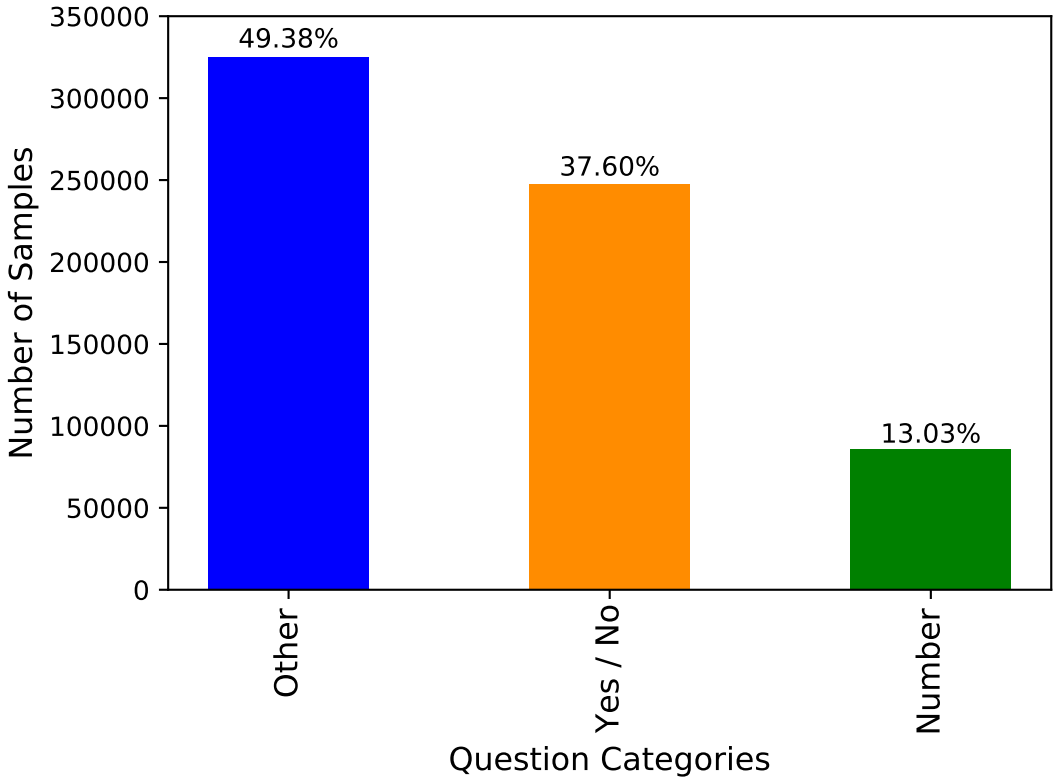


Fig. 2.12 Distribution of 3 Categories of VQA2.0 Questions [2]

different than A. By adding these new complementary images to the dataset, the impact of language priors on the models is reduced to a great extent.

Evaluation Metrics

VQA accuracy [2][6] is used for evaluating VQA1.0 and VQA2.0 (real and abstract) datasets. As in the VQA dataset for each question, answer is collected from 10 annotators. For evaluation of a model generated answer, the score is computed using the following formula:

$$\textbf{Accuracy(ans)} = min\left\{\frac{\textbf{\#humans that said ans}}{\textbf{3}}, \textbf{1}\right\} \tag{2.1}$$

It implies that an answer is correct if at least 3 annotators out of 10 respond that answer.

2.6.3 Other Datasets

There are other VQA datasets that are designed with varied motivations. Such motivations include the provision of visual reasoning capability, modeling in multilingual scenario, overcoming bias, use of external knowledge, etc. A statistical summary of all datasets is shown in 2.1 and brief summary is given below.

DAtaset for **QU**estion **A**nswering on **R**eal-world (DAQUAR) [99] was the first dataset for VQA task. It is based on real-world images from NYU-Depth v2 image dataset. The question-answer pairs are of two types based on the methods it is generated: synthetic and human. The synthetic question-answer pairs are based on a few templates while human question-answer pairs were collected using 5 human subjects. Images in DAQUAR are indoor scenes only; as a result questions are related to object and location from indoor scenes. Dataset exhibits bias due to human behavior of focusing mostly on specific objects compared to others. In DAQUAR human generated dataset, this bias exists for table and chair objects as the highly frequent answers with of more than 400 instances.

COCO-QA [32] is another dataset consisting of a large number of question-answer pairs for MS-COCO [100] images. COCO-QA dataset is generated by automatically converting the descriptions of MS-COCO images into question-answer pairs. As the questions are generated from the description of images, it is comparatively easier to answer them than the questions which were generated by human annotators. As a result, it is required only to get a high-level understanding of image instead of depth understanding and reasoning. The generated question-answer pairs are of four categories in COCO-QA, and these categories are object, number, color, and location.

FM-IQA [22] is the only available VQA dataset in a language other than English. It is built using images from COCO image dataset. FM-IQA consists of 316,193 Chinese question-answer pairs and their English translations. Unlike COCO-QA, type of questions are not limited to some set of categories. The questions in FM-IQA includes questions based on the fundamental understanding of image like the action of objects, questions related to the presence of object class in the image, questions related to object attributes, questions related to the positioning of objects and their relatedness with each other. Along with simple image understanding, datasets also contains a set of questions which requires a high level of reasoning and common sense. Human annotators from Amazon Mechanical Turk (AMT) crowd-sourcing platform have generated question-answers pairs for FM-IQA.

VQA1.0 [2] includes real and abstract scene images. VQA real image was the largest of all the existing dataset when it was introduced and generated using MS-COCO images. Human annotators from Amazon Mechanical Turk crowd-sourcing platform had generated question-answer pairs. MS-COCO consists of a wide variety of images with multiple objects and a different environment. As a result, VQA1.0 real consists of a diversified collection of question-answer pairs. In VQA1.0 questions are of open-ended as well as multiple choice types. To generate questions related to an image, the image is shown to three annotators, and to avoid repetition of questions, previously generated questions are shown to annotators. Detailed statistics of question-answer pairs are stated in Table 2.1.

Visual7W [101] is a VQA dataset which has dense annotations and objects' localization in the image. The visual7W dataset contains seven types of questions: **"what, where, when, who, why, how, and which"**. Compared to other VQA datasets, the questions in this dataset are more affluent, and the answers are longer on average. Question-answer pairs were generated through the AMT crowd-sourcing platform. Visual 7W consists of multiple choice type questions, with four associated answers for each question. In visual7W, object level grounding annotations are also provided by linking of objects present in a pair of question answers and drawing bounding boxes over those objects. There are 561,459 object groundings with an average of 12 bounding boxes associated with each image.

Similar to Visual7W, Visual Genome [102] question answering dataset consists of questions starting with **"what, where, when, who, why, how, and which"**. Human annotators generated this dataset on images from the visual genome image dataset. Question answer pairs in the visual genome are from free form Question Answer (QA) and Region-based Question Answer. For generating free form QA pairs, eight questions are asked for an image with at least three different W's from list of W's mentioned above. In region-based QA, pairs are generated based on a region with specifications like regions having more than 5K pixels and length of phrase of region

Table 2.1 Datasets for VQA [**Tr :** Training, **V :** Validation, **Ts :** Test, **OE :** Open Ended, **MCQs :** Multiple Choice Questions, **OW :** One Word, **MWs :** Multiple Words, **WUPS :** Wu-Palmer Similarity]

Year	Dataset	Image Source (Ques. Cat)	Q/I	No. of QA Pairs	Type	Ans. Len.	Evaluation Metrics
2015	**DAQUAR [99]**	NYUv2 (4)	8	6.7K 5.6K	OE	OW or MWs	WUPS
2015	**COCO-QA [32]**	MS-COCO (4)	1	78.7K 38.9K	OE	OW	WUPS
2015	**FM-IQA [22]**	MS-COCO	-	316K	OE	MW or	Human
2015	**VQAv1 [2]**	MS-COCO (20+)	5.6	248K 121K 244K	OE	OW(90%) or MW	VQA Accuracy
2015	**Abstract Scenes [2]**	Clipart (20+)	3	60K 30K 60K	OE	OW(90%) or MW	VQA Accuracy
2016	**Visual 7W [101]**	MS-COCO (7)	6.9	327K	MCQ's	OW or MW	Accuracy
2017	**Visual Genome [102]**	MS-COCO (7)	13.4	1.7M	OE	OW or MWs	Accuracy
2017	**VQAv2 [6]**	COCO (20+)	(-) (5.6)	443K 214K 447K	OE	OW(90%) or MWs	VQA Accuracy
2017	**Abstract Scenes [6]**	Clipart (20+)	3	60K 30K 60K	OE	OW(90%) or MWs	VQA Accuracy
2017	**CLEVR [103]**	MS-COCO		699K 149K 447K		OW	Accuracy
2017	**TDIUC [1]**	MS-COCO	12	1.6M	OE	OW	MPT Accuracy
2017	**OK-VQA [104]**	MS-COCO	1	14K	OE	OW	VQA Accuracy
2018	**GQA [105]**	MS-COCO & Flickr (10)	4	22M	OE		Accuracy

description more than four words. Free form QA pairs provide a diverse set of QA pairs while region based pairs add a set of factual QA pairs in the dataset.

Compositional **L**anguage & **E**lementary **V**isual **R**easoning (CLEVR) [103] dataset is generated to test various aspects of visual reasoning, which includes attribute identification, counting, comparison, spatial relationships, and logical operations. Images in CLEVR dataset are synthetic and are generated by random sampling of scene graph and render it using Blender. Nodes of a scene graph represent objects with attributes, and spatially related objects are connected through edges. The question in CLEVR is associated with a functional program that can be executed on a scene graph of the image to obtain the answer.

OK-VQA [104] is a knowledge based VQA dataset. It is proposed primarily for the set of questions for which only image content is not sufficient to answer. The answer mainly relies on an image or object's implicit property, which could be inferred from external knowledge bases.

2.7 Discussions

In recent years, there has been massive interest in the VQA task. Despite the vast literature, still there persists a gap between the existing VQA model and the way a human learns. Also, compared to human efficiency, the performance gap is significant. Table 2.2 presents the highlights of evolutionary progress made for the VQA task. It also provides pros and cons of different methods or paradigms. The methods are primarily separated by the main theme of the model or paradigm.

In this thesis, we try to overcome these gaps and propose solutions that try to mimic the way humans learn. To enhance the interaction of modality, we propose an aggregated multistage co-attention mechanism. Co-attention is core for VQA task. Multistage co-attention keeps on improving the information flowing from one stage to other. Further, aggregation of attention for dual modalities keeps on preserving the attention from different stages. This method is detailed in Chapter 3. We investigate the multistage co-attention mechanism and propose a dense attention-based mechanism to further improve bidirectional attention. Here an end-to-end model is proposed based on self and cross attention mechanism. Self-attention guides to encode the intra-modality contextual information; whereas, cross-attention is a crucial mechanism for dual-modality interaction. This model is elaborated in chapter 4. The existing question category information is another aspect that is rarely exploited in literature. By using question category whole answer search space could be reduced to respective question category based answers only. With this reduced answer search space, model could perform better with lesser confusion. This drives the proposal of the third

Table 2.2 Evolutionary Progress of VQA

Methods↓	Year	Pros	Cons
Fusion Based [2]	2015	- Simple - Low resource consumption	- Focus on the global features of the image [106] - Low performance due to limited interaction of modalities
Convolutional Attention Based [47]	2016	- Focus on salient image regions	- Issues with natural questions that contain reasoning and counting [107]
Graph Based [24]	2017	- Model complex relation - Good for reasoning	- Difficult to compared with attention based networks [107]
Object Based Attention [12]	2018	- Object level attention - More human like learning - Better Accuracy	- Issues with natural questions that contain reasoning and counting [107] - Trained on extracted region features
Transformer Based [108]	2019	- Generic - Powerful - High Performance	- Huge computational cost [109]
Large Vision Language Models [110]	2021	- Generic - Powerful - High Performance - End-to-End learning	- Required huge amount of training data [109] - Huge computational cost and resource requirements

contribution, i.e., *Dual Attention and Question Categorization based Visual Question Answering (DAQC-VQA)* discussed in Chapter 5. DAQC-VQA leverages the *Question Category* to answer the asked question. The detailed description, empirical results and analysis for reducing the answer search space alongwith its impact on different models *dual attention*, *aggregated co-attention* and *dense attention* is presented in chapter 5. We have validated all the proposed methods on two widely used datasets on real images, i.e., *VQA2.0* and *TDIUC*.

Chapter 3

Visual Question Answering with Aggregated Co-attention

Chapter Highlights

- Existing works from the literature demonstrate that attention on multi-modality in context of each other provides a better feature representation for image and question.
- This co-attention could be further improved if followed in multiple stages. A single stage of attention may not be able to extract sufficiently fine grained features suitable for the task. However, multiple glimpses through multistage co-attention might achieve that.
- A Multistage co-attention based model with corresponding aggregated attention of both modalities at each stage is proposed.
- Extensive experiments and analysis on *TDIUC* and *VQA2.0* show the efficacy of the proposed model in terms of *Overall Performance* and Question Category-wise Performance.
- The publications for this works are as follows:
  1. **Aakansha Mishra**, Ashish Anand, Prithwijit Guha, *Multistage Attention based Visual Question Answering*, IEEE International Conference on Pattern Recognition (ICPR), 2021, pp. 9407-9414
  2. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, *ACA-VQA: Aggregated Co-attention based Visual Question Answering*. [Accepted at Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), 2023]

3.1 Introduction

VQA models need to understand the syntax and semantics of the question, relate the question with the relevant object(s) of the image, and infer the answer using both image and text semantics. For example, in Figure 3.1, a VQA model needs to infer from the given question that it has to find the object *caboose* in the given image and identify its color. This task needs the understanding of syntax and semantics of the given question (textual domain). Further, the model has to understand that *caboose* is a part of a train and is not related to sky or any other objects in the given image (visual domain). Finally, the model has to identify the color of the specific part of the train. This example illustrates that the VQA task requires intricate understanding of both textual and visual domains.

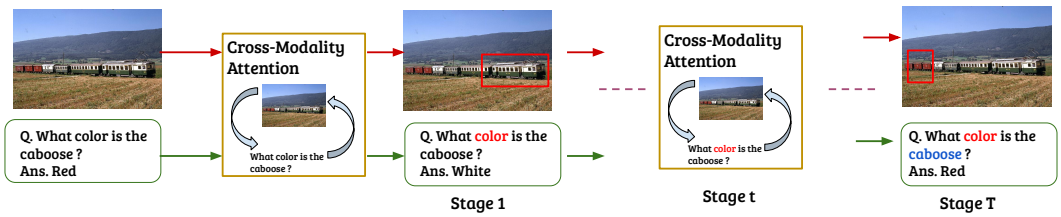


Fig. 3.1 An example to demonstrate the idea of proposed algorithm. The input question is asked for the color of the caboose in the given image. Initial stage (*Stage 0*) of attention on image provides high score to the front part of the train, and as a result the system answers the color as Green. In subsequent stages (*Stage 2* till *Stage T*) both modalities make each other stronger and the final stage predicts the correct answer *Red*.

A significant number of attention-based methods [12, 15, 47, 49, 73, 111] were proposed in the literature. These methods leverage upon attention from textual (question) to visual (image) domain to identify relevant region(s) of the image. Popular attention models leverage on the faster-RCNN [23] (method for object detection) features corresponding to the different regions of the image, and on LSTM framework [10] for question embedding. The learned embedding of the question provides the attention to the visual space and gives a weight to each region provided by the faster-RCNN. The region with higher weight is assumed to be highly correlated to the semantics of question. Most of the recent approaches use variants of the attention module and aim to obtain high-quality attention in the visual domain for a given question. Such unidirectional attention from textual to visual domain did help in improving the performance of VQA models. However, recent studies indicate that these models do overcome issues in identifying relevant region of the image, but still fall prey to language prior bias present in the VQA datasets.

It is always not possible to answer a question by reading at once or by looking at the image only one time. This corresponds to processing through shallow network. In

contrast, multiple glimpses of both image and question, aided by each other helps in answering. This correspond to VQA models [11, 58, 112] employing deeper networks with multiple stages of attention. Inspired from similar deeper VQA models, this proposal learns the co-attention (image attention and text attention) mechanism along with corresponding aggregation of dual modality attention from multiple stages.

To leverage the attention score from present stage in a multistage network may not be robust in a deeper model. Attention score obtained at each stage shows the importance of a specific image region (for image attention) or a question word (for text attention) for that stage. For salient features, attention score(s) obtained at each stage will be higher than other. Aggregating the attention scores from different stages could preserve the information in a better way. This could help in extracting better representation for both modalities, thereby resulting in significant improvement in the model's performance.

This work proposes a co-attention framework that considers textual to visual attention and visual to textual attention in an alternating fashion. The primary motivation to perform visual to textual attention is to improve question embedding in the context of visual features. Figure 3.1 shows an overview of the proposed method. Co-attention mechanism tries to improve the question embedding based on the previously learned attention on the image and further helps in obtaining a better representation of visual features. This co-attention mechanism is extended to multiple stages. It may help the model obtain a better understanding of the question, filter out the most relevant regions, and reason over the image objects to infer the answer. As shown in Figure 3.1, the first stage focuses on the front part of the train and not on the *'caboose'*. However, after a few stages, it does focus on the *'caboose'*. Similarly, for the question, two important terms *'color'* and *'caboose'* are recognized in the question by the model.

Extensive comparative experiments are conducted on the TDIUC [1] and VQA2.0 dataset [6] to evaluate the performance of the proposed model. Ablation analysis is also performed to show the importance of the multistage attention module, aggregation of attention and multistage loss in obtaining the performance gain. Key contributions of this work are as follows:

- Multistage co-attention based model with corresponding aggregated attention of dual modality at each stage.
- A multistage loss is proposed to overcome the gradient vanishing problem, since the deeper model is more likely to suffer from these problems.
- Experiments and ablation analysis on *TDIUC* and *VQA2.0* datasets show the efficacy of proposed method.

3.2 Proposed Method

Most existing works [12][73][15][111] treat VQA as a classification problem and train the system over all triplets $(\mathrm{I}, \mathrm{q}, \mathrm{a}) \in \mathbf{I} \times \mathbf{Q} \times \mathbf{A}$. Accordingly, the proposed framework treats VQA as a classification task.

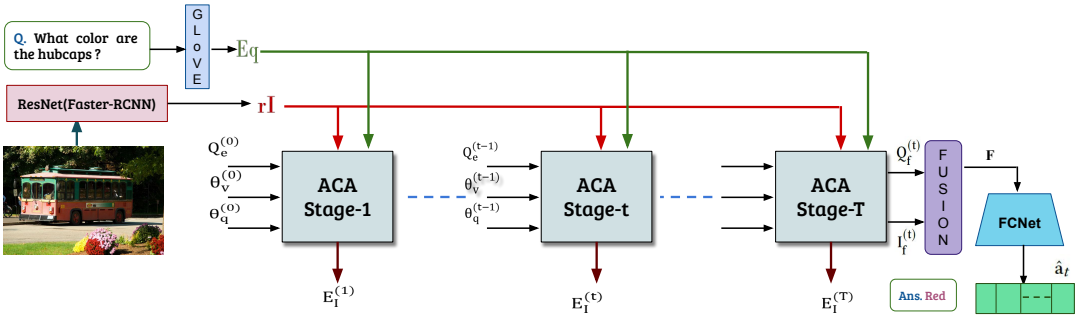


Fig. 3.2 Proposed ACA-VQA framework. Here **rI** and **Eq** are ResNet-101 features of n_v regions extracted from faster-RCNN and GloVe embeddings of words in questions. Each stage of **ACA** takes as input the LSTM encoding of question as well as region features and outputs the corresponding attended modality. After **T** stages of attention $I_f^{(t)}$ and $Q_f^{(t)}$ features are obtained which are then fused to obtain a unified representation, which is further passed to a fully connected network (**FCNet**) for answer classification.

Initially, the input image (I) features are extracted as object proposals by using pre-trained faster-RCNN network and the input question (q) is encoded using a LSTM network (Subsection 3.2.1). These features are used for cross-modal interaction through *Attention of Question on Image* (QoI) and *Attention of Image on Question* (IoQ) in each stage of *Co-Attention*. This model employs multiple such stages with *Aggregated Co-Attention* (Subsection 3.2.2). The model is learned in an end-to-end manner (Subsection 3.2.4) and the answers are predicted through an element-wise fusion of the attended features of both modalities (Subsection 3.2.3). The overall framework of *A*ggregated *C*o-*A*ttention based VQA (ACA-VQA) is depicted in Figure 3.2.

3.2.1 Feature Extraction

Visual Feature Extraction – Following existing literature [12], n_v object region proposals are obtained from the input image I by using the pretrained Faster-RCNN [23] network. The d_v dimensional ResNet-101 [19] embeddings of these region proposals are extracted further. Thus, salient region based image features $\mathbf{rI} \in \mathbb{R}^{d_v \times n_v}$ is represented as follows

$$\mathbf{rI} = [\mathbf{r}_1, \ldots \mathbf{r}_i, \ldots \mathbf{r}_{n_v}]; \quad \mathbf{r}_i \in \mathbb{R}^{d_v \times 1} \tag{3.1}$$

Textual Feature Extraction – The input question q is either trimmed or padded to a fixed length of n_w words. These n_w words are embedded in a d_w dimensional space with pretrained GloVe [29] embeddings ($\mathbf{eq} \in \mathbb{R}^{d_w \times 1}$). The question is thus represented as $\mathbf{Eq} \in \mathbb{R}^{d_w \times n_w}$. These GloVe embeddings are input to a LSTM network $\text{LSTM}_Q^{(0)}$ and its last hidden layer is used to represent the question encoding $\mathbf{Q_e}(0) \in \mathbb{R}^{d_q \times 1}$.

$$\mathbf{Eq} = [\mathbf{eq}_1, \ldots \mathbf{eq}_j, \ldots \mathbf{eq}_{n_w}]; \quad \mathbf{eq}_i \in \mathbb{R}^{d_w \times 1} \tag{3.2}$$

$$\mathbf{Q_e}(0) = \text{LSTM}_Q^{(0)}(\ \mathbf{Eq}\ ) \tag{3.3}$$

The input to the initial co-attention stage ($t = 1$) are $\mathbf{rI}$ and $\mathbf{Q_e}(0)$.

3.2.2 Cross-Modal Interaction through Aggregated Attention

The proposed framework ACA-VQA exploits an aggregated co-attention mechanism at multiple stages for interaction of two modalities. In the t^{th} stage ($t = 1, \ldots T$), the *visual attention scores* $\alpha_v^{(t)}$ are computed first. For this purpose, the visual features $\mathbf{rI}$ and the question embedding $\mathbf{Q_e}(t-1)$ from the previous stage are projected to spaces of common dimension d_{hv}.

$$\overline{\mathbf{rI}} = [\overline{\mathbf{r}}_1 \ldots \overline{\mathbf{r}}_{n_v}] = W_v^I \mathbf{rI} \tag{3.4}$$

$$\overline{\mathbf{Q_e}}(t-1) = W_q^I \mathbf{Q_e}(t-1) \tag{3.5}$$

Where, $W_v^I \in \mathbb{R}^{d_{hv} \times d_v}$ and $W_q^I \in \mathbb{R}^{d_{hv} \times d_q}$ are linear transformations. The visual attention scores $\alpha_v^{(t)} \in \mathbb{R}^{1 \times n_v}$ are estimated as follows.

$$u_v^{(t)}[i] = \beta_v^\intercal \left\{ \overline{\mathbf{r}}_i \odot \overline{\mathbf{Q_e}}(t-1) \right\} \tag{3.6}$$

$$\alpha_v^{(t)} = \text{SoftMax}\left(u_v^{(t)}[1] \ldots u_v^{(t)}[i], \ldots u_v^{(t)}[n_v] \right) \tag{3.7}$$

Where, $\beta_v \in \mathbb{R}^{d_{hv} \times 1}$ is a linear transformation. The attention score $\alpha_v^{(t)}[i]\,(i = 1 \ldots n_v)$ indicates the correlation between the i^{th} image region and the input question. These visual attention scores are aggregated across multiple stages $t = 1, \ldots T$.

The aggregated visual attention for the i^{th} image region ($i = 1, \ldots n_v$) at the t^{th} stage is computed as follows.

$$\theta_v^{(t)}[i] = (1 - \gamma_t)\theta_v^{(t-1)}[i] + \gamma_t \alpha_v^{(t)}[i] \tag{3.8}$$

Here, $\gamma_t = \dfrac{1}{t}$ $(t \geq 1)$ and $\theta_v^{(0)}[i]$ are initialized to zeros at $t = 0$.

$$\theta_v^{(0)}[i] = 0; \quad i = 1 \ldots n_v \tag{3.9}$$

The image region features $\mathbf{r}_i$ are weighed by the aggregated visual attention scores $\theta_v^{(t)}[i]$ to obtain the unified visual representation $\mathbf{E_I}(t)$ at the t^{th} stage.

$$\mathbf{E_I}(t) = \sum_{i=1}^{n_v} \theta_v^{(t)}[i]\mathbf{r}_i \tag{3.10}$$

At the t^{th} stage, the *question word attention scores* $\alpha_q^{(t)}[j]$ $(j = 1, \ldots n_w)$ are generated in context of the globally attended image representation $\mathbf{E_I}(t)$. The attention score $\alpha_q^{(t)}[j]$ is computed for each word representation $\mathbf{eq}_j$ $(j = 1, \ \ldots \ n_w)$. Initially, $\mathbf{E_I}(t)$ and the word embeddings $\mathbf{Eq} = [\mathbf{eq}_1 \ \ldots \ \mathbf{eq}_{n_w}]$ are projected to spaces of common dimension d_{hq}.

$$\overline{\mathbf{Eq}} = [\overline{\mathbf{eq}}_1, \ldots \overline{\mathbf{eq}}_j, \ldots \overline{\mathbf{eq}}_{n_w}] = W_q^Q \mathbf{Eq} \tag{3.11}$$

$$\overline{\mathbf{E_I}}^{(t)} = W_v^Q \mathbf{E_I}^{(t)} \tag{3.12}$$

Where, $W_q^Q \in \mathbb{R}^{d_{hq} \times d_w}$ and $W_v^Q \in \mathbb{R}^{d_{hq} \times d_v}$ are linear transformations. The attention scores $\alpha_q^{(t)}$ are computed in the following manner.

$$u_q^{(t)}[j] = \beta_q^\intercal \left(\overline{\mathbf{E_I}}(t) \odot \overline{\mathbf{eq}}_j\right) \tag{3.13}$$

$$\alpha_q^{(t)} = \text{SoftMax}\left(u_q^{(t)}[1] \ldots u_q^{(t)}[j] \ldots u_q^{(t)}[n_w]\right) \tag{3.14}$$

Here, $\beta_q \in \mathbb{R}^{d_{hq} \times 1}$. The attention score $\alpha_q^{(t)}[j]$ $(j = 1 \ldots n_w)$ indicates the correlation between the j^{th} question word and the input image. These word attention scores are aggregated across multiple stages $t = 1, \ldots T$.

The aggregated word attention for the j^{th} question word $(j = 1, \ldots n_w)$ at the t^{th} stage is computed as follows.

$$\theta_q^{(t)}[j] = (1 - \gamma_t)\theta_q^{(t-1)}[j] + \gamma_t \alpha_q^{(t)}[j] \tag{3.15}$$

Here, $\gamma_t = \dfrac{1}{t}$ $(t \geq 1)$ and $\theta_q^{(0)}[j]$ are initialized to zeros at $t = 0$.

$$\theta_q^{(0)}[j] = 0; \quad j = 1 \ldots n_w \tag{3.16}$$

The word embedding $\mathbf{eq}_j$ is weighed by the corresponding aggregated question attention score $\theta_q^{(t)}[j]$ $(j = 1, \ldots n_w)$. This leads to the attended question representation

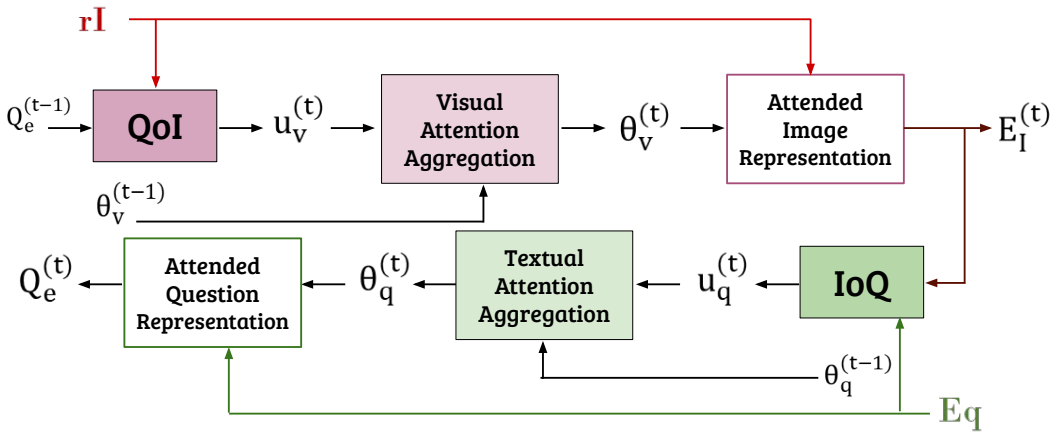


Fig. 3.3 Aggregation mechanism for visual and textual attention Here, θ_v^t and θ_q^t shows the aggregatd attention scores for visual and text features respectively.

$\widetilde{\mathbf{Eq}}(t)$ for the t^{th} stage. These weighted word embeddings are input to the LSTM network $\text{LSTM}_Q^{(t)}$ associated with the t^{th} stage and its last hidden layer is used to represent the question encoding $\mathbf{Q_e}(t) \in \mathbb{R}^{d_q \times 1}$.

$$\widetilde{\mathbf{Eq}}(t) = \left[\theta_q^{(t)}[1]\mathbf{eq}_1 \ldots \theta_q^{(t)}[j]\mathbf{eq}_j \ldots \theta_q^{(t)}[n_w]\mathbf{eq}_{n_w}\right] \tag{3.17}$$

$$\mathbf{Q_e}(t) = \text{LSTM}_Q^{(t)}\left(\widetilde{\mathbf{Eq}}(t)\right) \tag{3.18}$$

The linear transformations and the LSTMs associated with each stage are learned from data. The aggregation mechanism for textual and visual modality at t^{th} stage is demonstrated in Figure 3.3.

3.2.3 Answer Prediction

The aggregated co-attention based visual representation $\mathbf{E_I}(t)$ and question embedding $\mathbf{Q_e}(t)$ are obtained from the t^{th} stage. The representations obtained from the t^{th} stage are first projected to spaces of common dimension d_{hf} to obtain the embeddings $\mathbf{I_f^{(t)}}, \mathbf{Q_f^{(t)}} \in \mathbb{R}^{d_{hf} \times 1}$. These are fused through element-wise multiplication (denoted by $\odot$) to obtain the final multimodal representation $\mathbf{F_t} \in \mathbb{R}^{d_{hf} \times 1}$ for answer prediction at the t^{th} stage.

$$\mathbf{I_f^{(t)}} = W_f^I \mathbf{E_I}(t) \tag{3.19}$$

$$\mathbf{Q_f^{(t)}} = W_f^Q \mathbf{Q_e}(t) \tag{3.20}$$

$$\mathbf{F}_t = I_f^{(t)} \odot Q_f^{(t)} \tag{3.21}$$

Here, $W_f^I \in \mathbb{R}^{d_{hf} \times d_v}$ and $W_f^Q \in \mathbb{R}^{d_{hf} \times d_q}$ are the linear transformations for the respective image and question representations. The fused embedding $\mathbf{F}_t$ is input to a fully connected network $\text{FCNet}_{\text{ap}}^{(\text{t})}$ for the task of answer prediction. This network has a single hidden layer with $d_{hc}-$nodes. The output layer has $n_c = \mid \mathbf{A} \mid$ nodes. The hidden layer nodes host Sigmoid activation function, while the output layer nodes host the SoftMax activation function. The answer probability vector $\hat{\mathbf{a}}_t \in (0,1)^{n_c}$ is computed as follows.

$$\hat{\mathbf{a}}_t = \text{FCNet}_{\text{ap}}^{(\text{t})}(\mathbf{F}_t; d_{hc}; n_c) \tag{3.22}$$

3.2.4 Model Learning

The linear transformations W_v^I, W_q^I, β_v, W_q^Q, W_v^Q, β_q, W_f^I and W_f^Q are shared across the T stages. The $(T+1)$ LSTM networks (for different stages) $\text{LSTM}_Q^{(t)}$ $(t = 0, \ldots T)$ are used for computing the initial question embedding $\mathbf{Q_e}(0)$ and the ones for the next T stages. The fully connected networks $\text{FCNet}_{\text{ap}}^{(\text{t})}$ $(t = 1, \ldots T)$ are used for answer prediction. These linear transformations, the LSTMs and the fully connected networks are learned from the data.

Let, a be the ground-truth answer corresponding to the input image-question pair (I, q). An one-hot-encoded vector $\tilde{\mathbf{a}} \in \{0,1\}^{n_c}$ is constructed as a target prediction corresponding to the ground-truth answer a. A cross-entropy loss $\mathscr{L}_{CE}^{(t)}(\text{I}, \text{q}, \text{a})$ between the prediction $\hat{\mathbf{a}}_t$ and $\tilde{\mathbf{a}}$ is minimized over all image-question-answer triplets in for learning the VQA model parameters at t^{th} stage as follows.

$$\mathscr{L}_{CE}^{(t)}(I,q,a) = -\sum_{\forall (I,q,a) \in \mathbf{I} \times \mathbf{Q} \times \mathbf{A}} \sum_{r=1}^{n_c} \tilde{\mathbf{a}}[r] \log(\hat{\mathbf{a}}_t[r]) \tag{3.23}$$

The total loss ($\mathscr{L}_T$) is further defined as the summation of loss from each stage.

$$\mathscr{L}_T = \sum_{t=1}^{T} \mathscr{L}_{CE}^{(t)}(I,q,a) \tag{3.24}$$

This total loss $\mathscr{L}_T$ is minimized for all $(\text{I}, \text{q}, \text{a})$ triplets in $\mathbf{I} \times \mathbf{Q} \times \mathbf{A}$ and the error gradient is back-propagated to learn the model components in an end-to-end manner.

3.3 Results and Discussion

The proposed approach of ACA-VQA is benchmarked on the *TDIUC* and *VQA2.0* datasets. This section presents the quantitative analysis (Sub-section 3.3.1), ablation studies (Sub-section 3.3.2), and qualitative results of the experiments.

Table 3.1 Question category-wise model performance on the validation / test split of the TDIUC dataset. The proposed approach ACA-VQA is compared with several state-of-the-art methods.

Question Type	**RAU [1]**	**MFH [39]**	**QTA [16]**	**BAN [61]**	**CoR [57]**	**BLOCK [40]**	**ACA-VQA**
Scene Recognition	93.96	92.9	93.80	93.1	**94.68**	92.8	93.92
Sport Recognition	93.47	93.8	95.55	95.7	**95.94**	93.6	95.82
Color Attributes	66.86	67.0	60.16	67.5	73.59	68.7	**73.96**
Other Attributes	56.49	55.9	54.36	53.2	59.59	58.0	**60.46**
Activity Recognition	51.60	51.8	60.10	54.0	**60.29**	53.2	59.81
Positional Reasoning	35.26	34.7	34.71	27.9	39.34	36.1	**41.32**
Object Recognition	86.11	86.1	86.98	87.5	88.38	86.3	**88.41**
Absurd	96.08	93.3	**100.0**	94.47	95.17	90.7	96.19
Utility & Affordance	31.58	35.7	31.48	24.0	40.35	34.5	**40.94**
Object Presence	94.38	94.1	94.55	95.1	95.40	94.2	**95.54**
Counting	52.1	50.7	53.25	53.9	**57.72**	52.2	56.53
Sentiment Und.	60.09	63.3	64.38	58.7	**66.72**	66.1	65.93
Overall Accuracy	84.26	84.3	85.03	85.5	86.58	83.6	**86.82**
Harmonic Mean	59.00	68.3	60.08	54.9	65.65	61.1	**66.10**
Arithmetic Mean	67.81	60.3	69.11	67.4	72.25	68.9	**72.40**

3.3.1 Quantitative Results

Question category-wise comparison on TDIUC dataset – Table 3.1 demonstrates the results in terms of *Question category-wise performance* of the proposed model on the TDIUC dataset compared with other baseline models as per the availability. The last three rows present the *Overall Accuracy*, *Arithmetic-MPT* and *Harmonic-MPT* respectively.

It is evident from Table 3.1 that ACA-VQA has demonstrated competitive performance with respect to most of the baseline methods, and outperforms in terms of AMPT and HMPT. These two measures provide the performance in a more uniform way compared to overall accuracy. In case of overall accuracy all categories are given equal weightage irrespective of the number of samples in each class. Specifically, for *'Utility Affordance'* ACA-VQA performs well by *5.03%*, for *'Other Attributes'* class, the model gains by *1.45%*. In terms of HMPT, ACA-VQA outperforms the

top-performing baseline CoR [57] by a margin of *0.68%*. On the other hand, in terms of AMPT metrics, the performance of the BAN2-CTI [62] model is surpassed by *0.20%* with the AMPT metrics of the ACA-VQA model.

Table 3.2 *Overall Accuracy* comparison with other state-of-the-art models on TDIUC dataset. The *Category* column indicates the type of methods to which the techniques in the *second column* belong.

Category	Methods	Overall Accuracy
FUSION	**MCB[36]**	81.86
	MLB[37]	83.10
	MUTAN[13]	82.70
	BLOCK[40]	85.96
VISUAL ATTENTION	**SAN[11]**	82.30
	BTUP[12]	82.91
	QCG[71]	82.05
	RAMEN[31]	86.86
	QAA[113]	84.60
DENSE ATTENTION	**DFAF[14]**	85.55
	MLIN⋆[15]	**87.60**
CO-ATTENTION	**Proposed**	86.82

Overall performance comparison on TDIUC dataset – Table 3.2 indicates the overall model performance compared with respect to the baseline models. This table presents the results for those baseline models whose category-wise accuracy are not available. ACA-VQA outperforms most of the baselines, except for DFAF and MLIN. These models are particularly complex and dense attention-based architectures.

Table 3.3 Performance of ACA-VQA evaluated on data that excluded samples from the 'Absurd' category during training.

Without Absurd					
Metrics	**MCB [36]**	**QTA [16]**	**BAN [61]**	**ACA-VQA** $\alpha_v + \alpha_q$	**ACA-VQA** $\theta_v + \theta_q$
Overall Accuracy	78.06	80.95	81.9	**84.06**	83.90
Arithmetic-MPT	66.07	66.88	64.6	69.91	**70.13**
Harmonic-MPT	55.43	58.82	52.8	63.08	**63.89**

The effect of language prior is considered a major issue in VQA literature [1][6]. Language bias often leads to blind prediction of a specific answer based on only the question. For example, consider a VQA dataset containing a large number of questions of similar type having binary answers as *'Yes'* or *'No'*. Additionally, the model will be biased towrds answering to any question with *'Yes'*, if it is the answer to majority questions. Thus, it is necessary to force the model to look into the visual content for answering. To this end, the *'Absurd'* question category is introduced that has only single answer *'DoesNotApply'*. Here, the input question is completely unrelated

to the given image. Training with the *'Absurd'* question category forces the VQA model to consider the visual component. The ACA-VQA model is trained without the *'Absurd'* question category to understand the effect of language bias. Table 3.3 reports the ACA-VQA performance with respect to three baseline models in terms of three metrics. ACA-VQA is found to outperform the baselines in terms of *arithmetic-MPT* and *harmonic-MPT*. However, the model provides higher *overall accuracy* without aggregation of attention.

Performance comparison on VQA2.0 dataset – ACA-VQA demonstrates impressive overall performance in comparison to baseline models, as well as state-of-the-art methods. Out of all the proposed baselines, proposed model surpasses the majority, with the exception of complex attention-based architectures: BAN, BAN2-CTI, DFAF, and MLIN.

Table 3.4 Comparison for VQA 2.0 validation split in terms of *Overall Accuracy* and *three categories* of questions

Category	Methods	Yes / No	Number	Other	Overall
FUSION	**MCB[35]**	77.37	36.66	51.23	59.14
	MLB[37]	81.89	42.97	53.89	62.98
	MUTAN[13]	81.09	41.87	54.69	62.71
	MFH[39]				61.60
VISUAL ATTENTION	**SAN[11]**	78.40	40.71	54.36	61.70
	RN[114]	80.51	41.92	54.75	62.74
	BTUP[12]	80.34	42.80	55.80	63.20
CO-ATTENTION	**BAN[61]**	–	–	–	66.0
	BAN2-CTI[62]	–	–	–	66.00
	DoG[115]	**82.16**	45.45	55.70	64.29
	CTDA[116]	81.26	43.24	55.67	63.65
	QAA[59]	–	–	–	60.5
DENSE ATTENTION	**DFAF[14]**	–	–	–	**66.21**
	MLIN⋆[15]	–	–	–	66.18
CO-ATTENTION	**ACA-VQA**	82.01	**46.45**	**56.87**	64.95

3.3.2 Ablation Analysis

ACA-VQA has two important components. First, the multistage attention modules, and second, the aggregation of attention in both modalities. Hence, the effect of the number of stages and that of aggregation need to be analyzed through ablation studies. Accordingly, a set of experiments are performed performed to identify the appropriate number of stages and to recognize the effective way of aggregating the attention of dual modalities. Another objective of this ablation study is to analyse the impact of stage-wise loss in comparison to a single loss applied at a final stage. The effect

of parameter sharing among different stages is also experimented. These different ablations are presented in Tables 3.5, 3.6, 3.7, and 3.8.

Table 3.5 Effect of attention aggregation on both modalities **without stage loss** on TDIUC dataset in terms of *Overall Accuracy*, *Arithmetic-MPT* & *Harmonic-MPT*. NA in first column indicates that aggregation of attention is *'Not Applicable'* with single stage model

Attention Mode	T =1	T = 2	T = 3	T = 4	T = 5
$\alpha_v + \alpha_q$	85.03 69.59 61.10	86.33 69.88 60.42	86.32 70.78 63.40	86.21 70.47 63.50	85.10 70.63 63.34
$\theta_v + \alpha_q$	NA	86.79 71.59 64.52	**86.81** 71.57 64.58	86.05 71.15 64.01	85.81 70.90 63.58
$\alpha_v + \theta_q$	NA	86.13 71.48 64.12	85.64 71.04 63.75	86.35 71.12 63.29	85.40 70.92 63.47
$\theta_v + \theta_q$	NA	86.74 **71.84** **65.27**	86.44 71.58 65.07	86.20 71.75 64.98	86.18 71.37 64.74

Tables 3.5, 3.6, 3.7, and 3.8 report the results of ablation analysis for different model variants by changing the number of co-attention stages from T $= 1 \ldots 5$. Table 3.5 demonstrates the performance in terms of TDIUC evaluation metrics. Here, the experiments are performed by sharing the linear transformation parameters (W_v^I, W_q^I, β_v, W_q^Q, W_v^Q, β_q, W_f^I, W_f^Q) for all stages with different attention aggregation strategies. The first row ($\alpha_v + \alpha_q$) demonstrates the model performances for no attention aggregation in either of the modalities. The second row ($\theta_v + \alpha_q$) shows the model performances for attention aggregation along the visual modality only. Similarly, the third row ($\alpha_v + \theta_q$) demonstrates the model performances for attention aggregation along the text modality only. The last row ($\theta_v + \theta_q$) shows the model performance for attention aggregation applied to both visual and text modalities. It is observed that the optimal results (second best overall accuracy of 86.74%, best AMPT of 71.84% and highest HMPT of 65.27%) are obtained for $T = 2$ co-attention stages with attention aggregation applied to both modalities.

Table 3.6 shows the model performances with stage-wise loss. Here, the effects of *no attention aggregation* (first row) and *attention aggregation in both modalities* (second row) are also studied. All linear transformation parameters and answer predictors are shared among the different stages. The model is found to perform best with $T = 4$ co-attention stages alongwith aggregation of dual modality attention. Additionally, it is observed from Tables 3.5 and 3.6, that training with stage-wise loss has increased the model's performances for all the cases ($T = 2$ onward). For example,

for $T = 4$, the overall accuracy, AMPT and HMPT have respectively increased by 0.62%, 0.35% and 0.65%.

Table 3.6 Effect of Aggregation on both modalities **with stage loss** on TDIUC dataset in terms of *Overall Accuracy*, *Arithmetic-MPT* & *Harmonic-MPT* .

Attention Mode	T =1	T = 2	T = 3	T = 4	T = 5
$\alpha_v + \alpha_q$	85.03 69.59 61.10	86.58 71.58 63.97	86.16 71.75 64.69	86.54 71.81 64.67	86.33 71.66 64.28
$\theta_v + \theta_q$	NA	85.94 71.84 65.27	86.83 71.78 65.36	86.82 **72.10** **65.63**	86.78 72.03 65.62

Table 3.7 shows the model performances where the training is performed without stage loss and all the linear transformation parameters are unshared among the different co-attention stages. The model is found to perform well till $T = 2$ stages only. It is observed that, further increment in the number of stages lead to performance degradation. This might be attributed to overfitting on account of stage-wise increment in model parameters.

Table 3.7 Effect of aggregation on both modalities **without stage loss and unshared parameters for classifier networks at each stage** on TDIUC dataset in terms of *Overall Accuracy*, *Arithmetic-MPT* and *Harmonic-MPT* .

Attention Mode	T =1	T = 2	T = 3	T = 4	T = 5
$\alpha_v + \alpha_q$	85.03 69.59 61.10	86.61 **71.77** 64.56	85.29 71.26 64.18	86.81 71.39 63.51	86.69 70.39 62.18
$\theta_v + \theta_q$	NA	85.35 71.42 **64.64**	**85.49** 70.78 63.30	85.63 69.74 60.39	85.12 68.02 55.77

Finally, Table 3.8 shows the results for model training with stage-wise loss. Here, the answer classifier network of each stage has unshared parameters while other set of linear transformation parameters are shared among different stages. This experimental setup is found to outperform all other configurations. The model keeps on improving the performance till $T = 3$ stages of aggregated co-attention. Further increment results in performance degradation.

It could be concluded from the ablation analysis that, the model demonstrates a performance gain till $T \leq 4$ for most experimental configurations. However, the performance degrades for $T > 4$. The impact of attention aggregation is found to be positive for all experimental configurations. The model is found to learn well with a comparatively lower number of parameters. However, unsharing of all parameters leads to overfitting thereby reducing model performance.

Table 3.8 Effect of Aggregation on both modalities **with stage-wise loss, shared linear transformation parameters and unshared answer predictors for each stage**. Results are reported on TDIUC dataset in terms of *Overall Accuracy*, *Arithmetic-MPT* & *Harmonic-MPT* .

Attention Mode	T =1	T = 2	T = 3	T = 4	T = 5
$\alpha_v + \alpha_q$	85.03 69.59 61.10	86.57 72.18 65.71	86.77 72.09 65.35	86.81 71.38 64.52	86.44 71.49 63.88
$\theta_v + \theta_q$	NA	85.60 71.64 64.12	**86.82** **72.40** **66.10**	86.61 71.75 64.36	86.73 71.13 62.91

In Figure 3.4, the parameter count for the ACA-VQA model is displayed. It is worth noting that the best performing model (which includes stage loss, only unshared answer predictor parameters, and shared linear transformations) has a relatively low parameter count. Figure 3.5 presents the parameter count and validation accuracy for the ACA-VQA model on the VQA2.0 dataset. The model's performance improves up to T = 2 stages, but beyond that, the performance decreases as the parameter count increases.

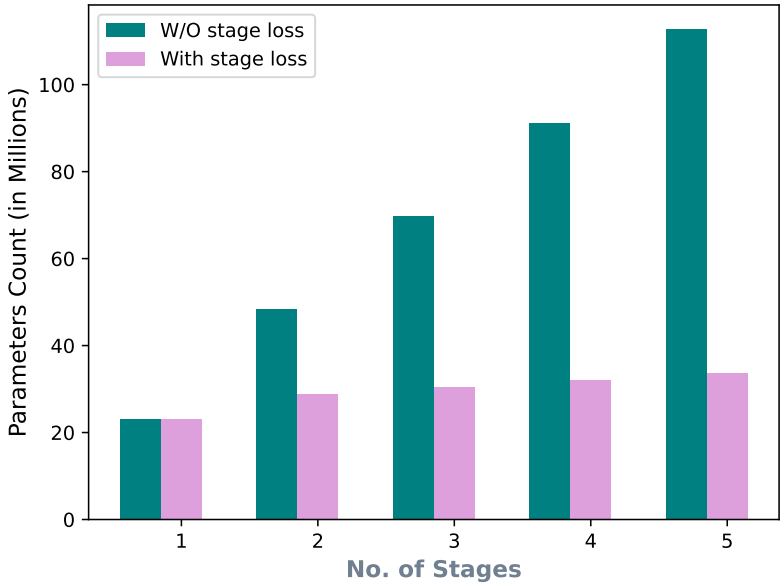


Fig. 3.4 Parameter Count (in Millions) for *TDIUC* dataset with respect to the number of stages incorporated in ACA-VQA model. W/O stage loss shows the count when all parameters are **unshared** and no stage loss is incorporated. With stage loss shows the parameters with **shared linear transformation** parameters and **unshared** answer predictor parameters.

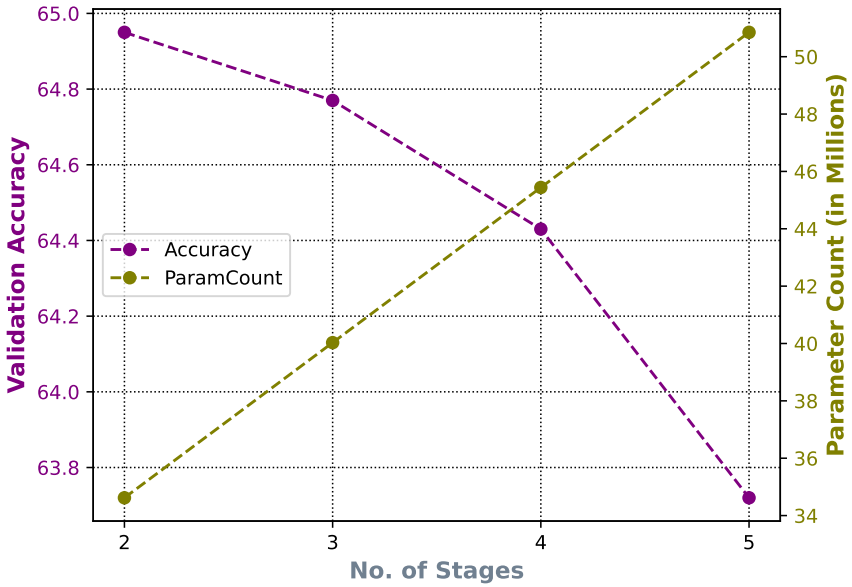


Fig. 3.5 Validation Accuracy and Parameter Count (in Millions) for *VQA2.0* dataset with respect to the number of stages incorporated in ACA-VQA model. Performance is reported with stage loss, with **shared linear transformation** parameters and **unshared** answer predictor parameters.

3.3.3 Qualitative Results

The qualitative results of ACA-VQA are demonstrated in figure 3.6. Each figure shows the top-2 salient regions obtained from the proposed model. Saliency of regions is defined by the visual attention obtained from the model. Results are demonstrated on different categories of TDIUC dataset like *'color'*, *'object presence'*, *'object recognition'* etc. It can be observed that the regions with the highest scores in the model are more likely to infer the most relevant concepts necessary to arrive at the correct answer. These regions serve as an indicator of the model's attention to specific areas of the input image that are crucial in making accurate answer predictions.

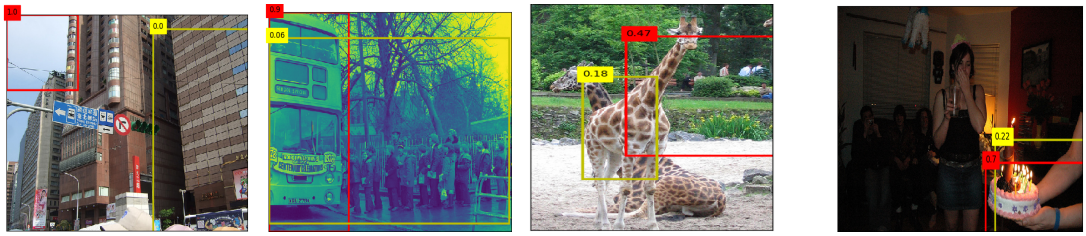
Consider the image shown in Figure 3.6a with associated question *"What color is the clock ?"*. Here, the most relevant object is *'clock'*. the corresponding attention score for the region belonging to clock is highest amongst all the proposed regions. Similarly, consider the *'object presence'* category of question *'Are there any knives in the picture ?'* associated to the image in Figure 3.6f. To answer this, the model has inspected the *'knife'* in the image. Similarly, consider the other images shown in Figures 3.6c, 3.6d, 3.6e, 3.6f, 3.6g, 3.6h, 3.6i, 3.6j, 3.6k, 3.6l, and their question-answer pairs. Here, the model has shown the ability to capture the most relevant regions for answer prediction.

(a) **Q.** What color is this clock ?
Ans: Gold ✔ (0.35, 0.31)
GT: Gold

(f) **Q.** Are there any knives in the picture ?
Ans: Yes ✔ (0.66, 0.33)
GT: Yes

(c) **Q.** What color is the woman's shirt ?
Ans: Blue ✔(0.57, 0.27)
GT: Blue

(d) **Q.** What vehicle is shown in the photo?
Ans: Airplane ✔ (0.49, 0.2)
GT: Airplane

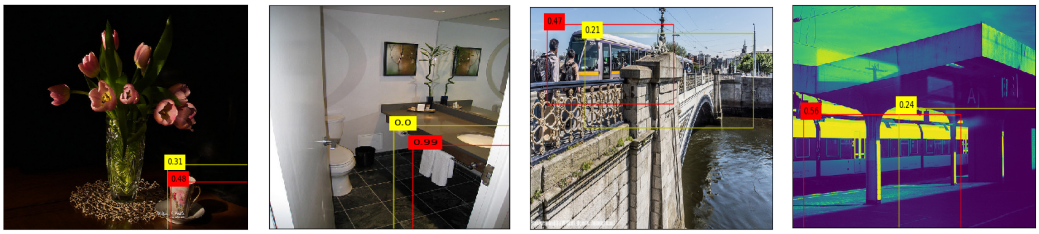


(e) **Q.** What color is the sky?
Ans: Blue ✔ (1.0, 0.0)
GT: Blue

(f) **Q.** Is there a vehicle in the photo ?
Ans: Yes ✔ (0.9, 0.06)
GT: Yes

(g) **Q.** What animal is in the picture ?
Ans: Giraffe ✔ (0.47, 0.18)
GT: Giraffe

(h) **Q.** What food is in the picture ?
Ans: Cake ✔ (0.7, 0.22)
GT: Cake



(i) **Q.** Are there any cup in the picture ?
Ans: Yes ✔ (1.0, 0.0)
GT: Yes

(j) **Q.** What colors are the towels ?
Ans: White ✔ (0.99, 0.01)
GT: White

(k) **Q.** Are there any vehicles in the picture ?
Ans: Yes ✔ (0.47, 0.21)
GT: Yes

(l) **Q.** What type of vehicle is this in picture ?
Ans: Train ✔ (0.56, 0.24)
GT: Train

Fig. 3.6 Qualitative results from proposed model on different category question for *TDIUC* dataset. Visualization of top-2 attention scores obtained for image regions. Top-2 salient regions identified with visual attention scores are represented as (top1, top2).

3.4 Error Case Analysis

In terms of the three defined evaluation metrics, proposed model demonstrates competitive performance compared to most of the models. However, it's important to highlight that it falls short in comparison to the other methods, which relies on relational reasoning techniques involving the encoding of relationships between objects through multiple sub-chains. We further analyse the failure cases of our model to gain better insight. Some of the analyzed failure cases are presented in the Figure-3.7. Our observation indicates the one of more of the following reasons could be most prominent bottleneck for the proposed approach.

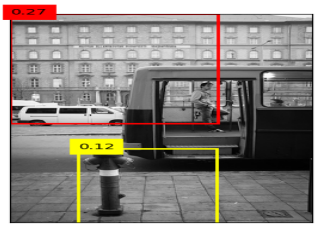
- **Questions about tiny visual content :** Some of the questions are highly fined-grained or are very specific in the following sense. The corresponding images contain very tiny visual information that can be very difficult even for the human to answer. In the Figure 3.7a, it can be observed that the identifying another person, who is mostly occluded and only a tiny fraction of head is visible, is very difficult to correctly answer the question on number of people present. In such scenarios model mostly gives incorrect answers.

- **Misidentification of Question Context:** Despite its proficiency in recognizing regions and keywords, the model occasionally falls in accurately identifying the specific context in question. This leads to incorrect predictions. As it could be seen in the Figure 3.7b, the question is about emotion but model is unable to understand the content and provides the object recognition result. However, the relation modeling could help in such cases to provide the context, *e.g.,* the relation between ball, grass, dog may help to draw the happiness conclusion.

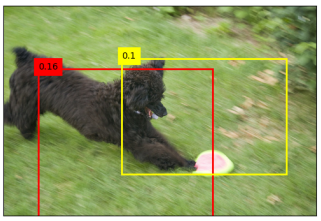
- **Misalignment with Ground Truth:** In some cases, the given ground truth may not be correct. For example, the given ground truth answer is three in Figure 3.7c, and the predicted answer is four. However, the predicted answer can be considered as correct since there are more laptops in the image compared to the ground truth.

- **Common-Sense Reasoning:** The approaches that models the common sense reasoning (through relation modeling, symbolic AI etc) outperforms the proposed model for some specific sets of questions. This provides the extra information about the image and helps to infer the better conclusion. However, one notable limitation of the proposed model is lack of common-sense reasoning. It appears to excel in identifying regions and keywords correctly but may produce answers that defy common-sense expectations. For instance, it might provide an activity-related response to a binary-type question, which highlights a gap in its reasoning abilities. The failure case presented in the Figures 3.7d, 3.7f, 3.7e occurs because of the weak relationship model among the objects.

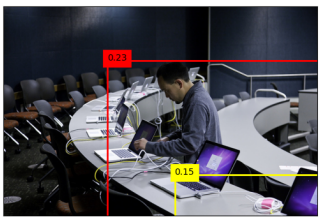
- **Lack of Relational Reasoning:** Notably, the model appears to lack the ability to perform relational reasoning effectively. It may struggle to comprehend and utilize the relationships between objects or elements within the image or question, resulting in sub-optimal answers. Developing this relational reasoning capacity



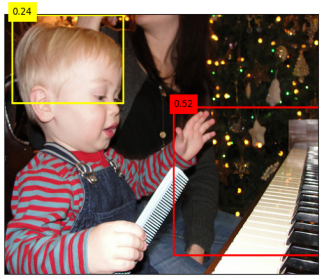
(a) **Q.** How many people are shown on the **bus**?
Ans. One✗ (0.27, 0.12)
GT: Two



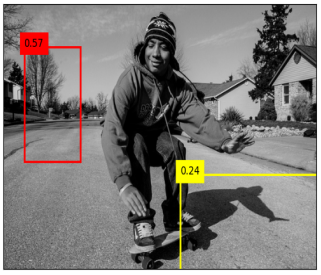
(b) **Q.** What is the **emotion** the dog is showing?
Ans. Yarn✗ (0.16, 0.10)
GT: Happiness



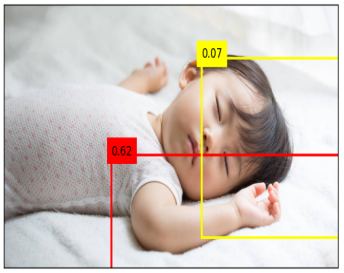
(c) **Q.** **How** many laptops are there?
Ans. Four✗ (0.23, 0.19)
GT: Three



(d) **Q.** What **season** is this?
Ans. Piano ✗ (0.52, 0.24)
GT: Christmas

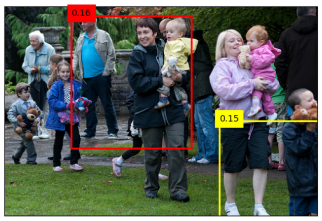


(e) **Q.** Is the person **happy**?
Ans. Street✗(0.57 0.24)
GT: Yes

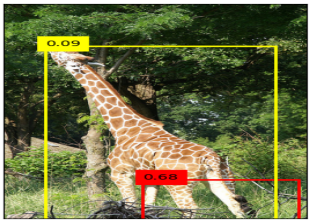


(f) **Q.** Is the boy **scared**?
A.Swimming✗ (0.62, 0.07)
GT: No

(g) **Q.** What are they **playing**?
Ans. Running✗(0.74, 0.04)
GT: Soccer



(h) **Q.** What is the girl **doing**?
Ans. Standing✗(0.16, 0.15)
GT: Walking



(i) **Q.** What is the giraffe **doing**?
Ans. Standing✗ (0.68, 0.09)
GT: Eating

Fig. 3.7 Wrong predictions from ACA-VQA

is essential for the model to provide more contextually accurate responses. For example, it can be observed in the Figure 3.7g that for the asked question about *game playing*, the proposed model responded *'running'* as the answer. This response is reasonable if other context and relation between the objects are ignored. However, if the relationship between the *ground*, *person*, and *ball* are present, the scenario is very different. We believe that incorporation of the ability to handle relations among objects in the proposed model may help significantly to overcome this drawback.

– **Region-Based Predictions:** In some instances, the model seems to rely heavily on region-based information for making predictions. However, it may struggle to pinpoint the most relevant regions, resulting in answers that do not align

with the question's. It can be observed in the Figures 3.7h, 3.7i, the results highly depend on the identified attention region. In both the cases, in the attended region the giraffe and one of the girls is *standing*. Hence, the predicted results can also be considered as valid. In fact, this aligns with the examiners' another comment regarding multilabel classification. Reformulating the task as multilabel classification with appropriate changes in the datasets will help resolve such errors.

In summary, the proposed model exhibits strengths in object and keyword identification but faces challenges in accurately recognizing relations among specific objects and making predictions that align with both the question and ground truth. Additionally, enhancing the model's common-sense reasoning capabilities is crucial for improving answer quality. In light of these insights, there arises a further necessity to delve into the explainability and interpretation of the attention models that underlie the proposed approach.

3.5 Discussions

In this chapter, a co-attention mechanism is presented that extracts attended features of both modalities by operating across multiple stages. To preserve the information from different stages, the attention scores obtained in both modalities are aggregated across stages. This ensures that the information from all stages are preserved and utilized in the model. A multistage loss is introduced to prevent the issues of gradient vanishing. Also, most parameters are shared across co-attention stages to avoid performance degradation due to overfitting. The proposed method (ACA-VQA) is benchmarked on the TDIUC and VQA2.0 datasets. It was observed that the proposed multistage aggregated co-attention model demonstrates competitive performance compared to state-of-art methods. The model's efficacy was further confirmed through ablation analysis, which demonstrated the effectiveness of utilizing multistage attention, aggregation, and stage-wise loss in the model.

Co-attention is a vital component of VQA models that facilitates the extraction of cross-modality information for inferring the answer. In this context, intra-modality interaction can enhance the model's capability by focusing on the within-modality details that should have higher weightage. To this end, the next chapter proposes an approach that utilizes intra-modality interaction followed by cross-modality attention. This approach aims to extract richer features from the input data and improves the model's overall performance on the VQA task.

Chapter 4

CSCA: VQA with Cascade of Self- and Co-Attention Blocks

Chapter Highlights

- The proposal on ACA-VQA (Chapter 3) focused on a dual attention mechanism to reflect cross-modality interactions. However, it did not emphasize the inter-object (within image) and inter-word (within question) interactions.
- The present chapter includes self-attention on each modality along with cross-attention. Embeddings obtained after self-attention encodes contextual information within a single modality and that is used further to generate cross-attention based encoding.
- The process of self-attention and cross-attention comprises a block of dense attention mechanism. Such dense attention blocks are further cascaded to obtain enhanced representations.
- Detailed performance analysis conveys that the proposed model is comparable with existing state-of-art VQA models on two benchmark datasets: *TDIUC* and *VQA2.0*.
- The publication for this works is:
  - **Aakansha Mishra**, Ashish Anand, Prithwijit Guha, *"CSCA: VQA with Cascade of Self- and Co-Attention Blocks"* [Manuscript Under Review]

4.1 Introduction

The previous chapter proposed a cross-modal multistage aggregated attention-based VQA model. This chapter proposes another multistage VQA model by incorporating dense attention on dual modality features. This dense attention mechanism is a

combination of cross-modal and intra-modal interactions. This approach aims to extract richer features from the input data by focusing on both the inter- and intra-modality interactions. The proposed model's dense attention mechanism allows it to capture both the cross-modality and intra-modality information effectively. By incorporating dense attention, we can capture more detailed information and improve the model's performance on the VQA task. Overall, our proposed approach aims to improve the performance of VQA models by leveraging the strengths of both cross-modal and intra-modal interactions. By doing so, we can extract more informative features from the input data and improve the model's ability to answer questions accurately.

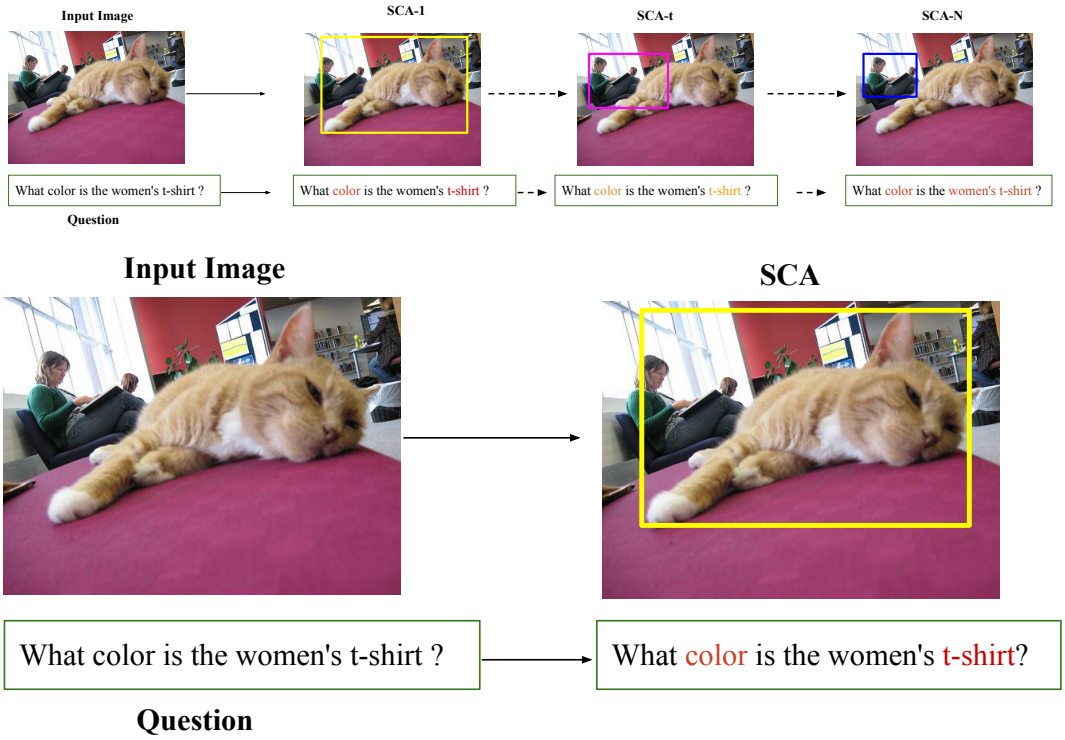


Fig. 4.1 An example to illustrate the attention relevance for dual modality through cascaded SCA module [top figure]. Without cascaded module the attention is not getting refined and hence not able to give better attention [bottom figure].

Recent attention-based models have taken inspiration from transformer-based models [65] to include self-attention (SA) as well. The SA helps in incorporation of internal correlation within a modality. For text modality, SA encodes internal correlation among words to obtain informative representation of the given sentence. Similarly, for image modality, SA helps in encoding correlation among the salient regions of image. Figure 4.1 shows an example for illustration. The given question is "*What color is the women's shirt?*. Salient regions within image include 'woman'. It is likely to be informative if the region consisting of 'women' keeps the contextual information such as "dress she is wearing", "hair color" as well as correlation with

other salient objects. Here, women's shirt could be one of the more correlated region with respect to some other salient objects. SA helps in encoding such information.

Based on the advantages of each of the following modules: self-attention (SA), co-attention (CA) and cascade of attention mechanisms, this work proposes combining them together in a systematic manner. Towards this objective, the proposed model builds one self- and co-attention based attention block (SCA), that combines both SA and CA in a specific way. For each of the text and image modalities, a specific SA module obtains a feature representation for the respective modality. Then the co-attention module uses self-attended representation of one modality and attends (takes attention) on the self-attended representation of the other modality to obtain cross-modality contextual representation for the second modality. Thus, there are two SA modules (one for each text and image modalities) and two co-attention modules within a single SCA block (Figure 4.2). In one complex attention block of SCA, both modalities guide itself to capture internal correlation and each other to learn the robust representation of each of the visual and textual domains.

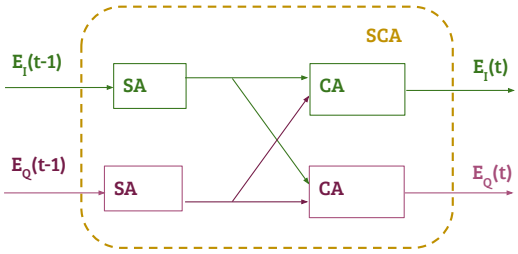


Fig. 4.2 Overview of the proposed model. An attention block, referred to as SCA, comprises of *self-attention (SA)* and *co-attention (CA)* modules. Multiple such attention blocks are cascaded, where output of some $(t-1)^{th}$ block is presented as input to the t^{th} block.

The proposed model exploits the niche attributes of the different attention mechanisms and further combines them together in a dense attention module (SCA block). A Cascade of multiple SCA blocks (CSCA) is used to extract fine-grained information. Figure 4.2 gives the overview of the t^{th} SCA block which takes representations of question and image of the $(t-1)^{\text{th}}$ block as input, and provides the improved representation of question and image.

CSCA model comprises of self- and cross-attention blocks cascaded in an alternate manner. Initially, self attention mechanism is applied for each of vision and text feature inputs to obtain representations encoding the intra-modality contextual information. These representations are then used as inputs to cross attention mechanism. Initial attention blocks may not be able to capture the relevant semantics for intra and inter modalities. While after learning through cascades of attention blocks, model learns the better multimodal feature representations.

The figure 4.1 [top row] serves as an illustrative example to highlight the impact of the cascaded Self and Cross-Attention (SCA) module within the model. In the initial SCA block, the model's attention is directed towards a broad range of image regions, encompassing various objects such as 'cat,' 'women,' 'window,' and others. Additionally, it exhibits a degree of focus on specific words like 'color' and 't-shirt,' as indicated by their attention scores. With the inclusion of multiple SCA blocks, notably after the t^{th} block, the model's attention gradually refines, shifting towards more concentrated image regions. This transition is accompanied by changes in word attention scores. Ultimately, in the final SCA block, the model's attention is concentrated on the most salient image region within the context of the given question. Simultaneously, the attention mechanism for the question becomes finely tuned to the most pertinent words that enable accurate responses.

When contrasting the results with and without the SCA module 4.1 [bottom row], it becomes evident that a single round of attention may not suffice to capture all the relevant image regions and question words effectively. The cascaded SCA module contributes to a progressive and refined attention process, enhancing the model's ability to grasp contextual information and produce more accurate answers.

To analyse and evaluate the model performance, extensive experiments are performed on two widely used VQA datasets: *VQA2.0* [6] and *TDIUC* [1]. Ablation analysis experiments are also performed to understand the impact of the important components of the proposed model. Primary contributions of this work are as follows:

- A dense attention based VQA model comprising of cascaded attention blocks.
- The core of each attention block consists of self-attention and co-attention so that the two modalities guide each other to obtain an enriched representation.
- Extensive performance evaluation along with ablation analysis of the proposed model on the two benchmark datasets – *TDIUC* and *VQA2.0*.

4.2 Proposed Approach

The proposed framework treats VQA as an answer classification task following existing works like [12][14][2][6][15]. The input image I ($I \in \mathscr{I}$) and the associated natural language question q ($q \in \mathscr{Q}$) are first subjected to feature extraction (Subsection 3.2.1). Pretrained deep networks are used to extract features from a few salient image regions. The network embeddings are used to represent the input image. Similarly, a pretrained network is used to obtain the word embeddings of the associated input question. These word embeddings collectively represent the input question. The feature embeddings of both image and text modalities are subjected to self-attention mechanism (Subsection 4.2.2) for capturing the relationships among different regions of I and words of

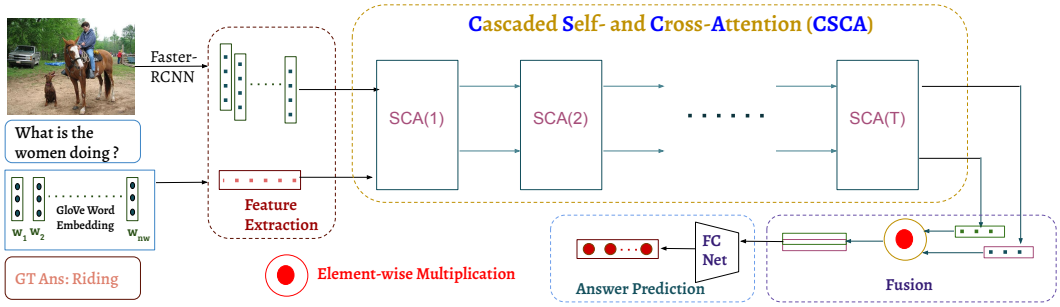


Fig. 4.3 Functional block diagram of the proposed approach. Initial feature extraction stage is followed by a cascade of self-attention and co-attention mechanisms. Final attended features are fused through element-wise multiplication and are fed to a fully connected network for answer classification.

q. The self-attended representations of these two modalities are further processed by co-attention modules (Subsection 4.2.3). This single stage of **S**elf and **C**o-**A**ttention mechanism cascade forms a single SCA block (Figure 4.2). Multiple SCA blocks are cascaded to obtain further fine grained representations of both modalities. The embeddings obtained from the final SCA block are fused (Subsection 4.2.4) and fed to the answer classification network (Subsection 4.2.5) to predict the answer probability vector $\hat{\mathbf{a}}$. The framework of proposed model is presented in Figure 4.3

4.2.1 Feature Extraction

Visual features ($\mathbf{rI} \in \mathbb{R}^{d_v \times d_{nv}}$) and *Textual features* ($\mathbf{Eq} \in \mathbb{R}^{d_w \times n_w}$) are extracted by the process detailed in section 3.2.1. All feature embeddings in $\mathbf{rI}$ and $\mathbf{Eq}$ are projected to spaces of common dimension (d, say) to obtain the respective initial feature embedding matrices as $\mathbf{rI}(0)$ and $\mathbf{E_q}(0)$.

$$\mathbf{rI(0)} = W_c^I \mathbf{rI} \tag{4.1}$$

$$\mathbf{E_q(0)} = W_c^Q \mathbf{Eq} \tag{4.2}$$

Here, $W_c^I \in \mathbb{R}^{d \times d_v}$ and $W_c^Q \in \mathbb{R}^{d \times d_w}$ are linear transformations. These representations are provided as input to the self- and co-attention modules.

4.2.2 Self-Attention

The self-attention (SA) mechanism is one of the key components of the proposed model. It is incorporated for both textual (question as collection of words) and visual (image as top-n_v salient regions) modalities. At the t^{th} ($t = 1, \ldots T$) block, the input to SA are $\mathbf{rI}(t-1)$ and $\mathbf{E_q}(t-1)$. Following [65], the SA uses *keys* and *queries*,

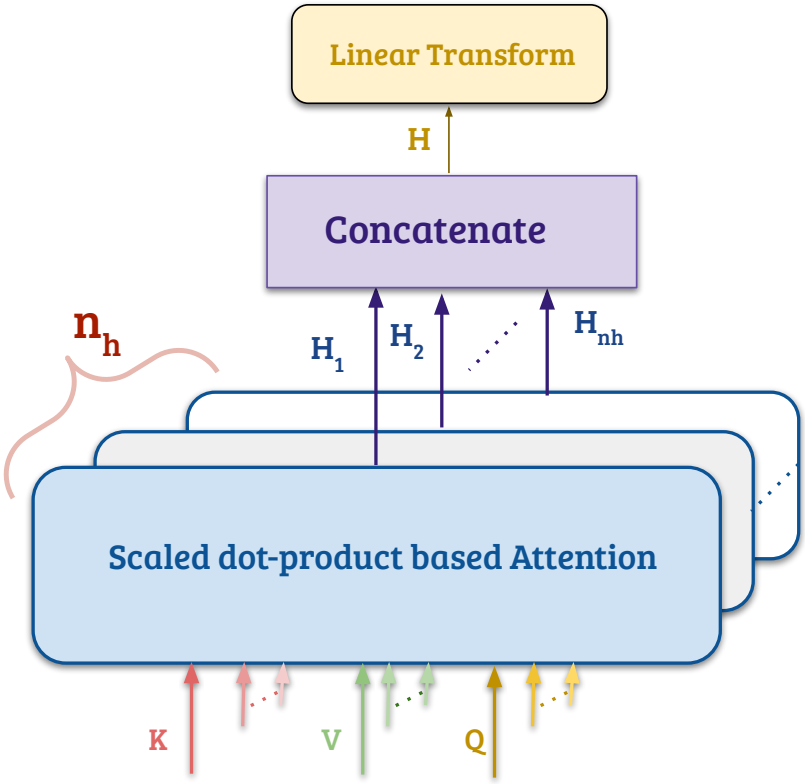


Fig. 4.4 Multihead Attention Mechanism. Here, n_h heads simultaneously process input data K, V and Q in parallel. The output of all heads are concatenated to present the attended representation.

both of dimension d_{KQ} and values of dimension d_{VS} respectively. The *Multi-Head Attention* [65] is incorporated to capture the attention from different aspects. For this, n_h parallel heads are added, where each head is considered to learn the relationships from different view (for image) and context (for question).

Let $\mathbf{E_M} = \{\mathbf{em}_1 \ldots \mathbf{em}_l\}$ be a matrix of l feature embeddings, where $\mathbf{em} \in \mathbb{R}^{d_m \times 1}$ and $\mathbf{E_M} \in \mathbb{R}^{d_m \times l}$. For visual features, $\mathbf{E_M} = \mathbf{rI}(t-1)$, $l = n_v$ and $d_m = d$. Similarly, for question features, $\mathbf{E_M} = \mathbf{E_q}(t-1)$, $l = n_w$ and $d_m = d$.

The query ($Q_S^{(h)}$), key ($K_S^{(h)}$) and value ($V_S^{(h)}$) matrices for the h^{th} head can be respectively expressed as follows

$$Q_S^{(h)} = \left(W_h^{QS}\right)^{\mathsf{T}} \mathbf{E_M} \tag{4.3}$$

$$K_S^{(h)} = \left(W_h^{KS}\right)^{\mathsf{T}} \mathbf{E_M} \tag{4.4}$$

$$V_S^{(h)} = \left(W_h^{VS}\right)^{\mathsf{T}} \mathbf{E_M} \tag{4.5}$$

where, $W_h^{QS} \in \mathbb{R}^{d_m \times d_{KQ}}$, $W_h^{KS} \in \mathbb{R}^{d_m \times d_{KQ}}$ and $W_h^{VS} \in \mathbb{R}^{d_m \times d_{VS}}$ are linear transformations. Using $\{Q_S^{(h)}, K_S^{(h)}, V_S^{(h)}\}$, the inner product of query is performed with all the keys and is divided by $\sqrt{d_{KQ}}$ for more stable gradients [65]. The *SoftMax* function is

applied on the inner product to obtain the attention weights for question words and salient image regions. A scaled inner product based attention is computed for all the heads in the following manner.

$$\mathbf{H_h} = \left(V_S^{(h)}\right) \text{SoftMax}\left(\frac{{Q_S^{(h)}}^\top K_S^{(h)}}{\sqrt{d_{KQ}}}\right) \tag{4.6}$$

$$\mathbf{MH}(\mathbf{E_M}) = W_{mh}\mathbf{H}; \quad \mathbf{H} = [\mathbf{H_1} \ldots \mathbf{H_h} \ldots \mathbf{H_{nh}}] \tag{4.7}$$

Here, $W_{mh} \in \mathbb{R}^{d_m \times (n_h \times d_{VS})}$ is the linear transformation. The output ($\text{MH}(\mathbf{E_M})$) of multi-head attention module is passed through fully connected feed forward layers with ReLU activation and dropout to prevent overfitting. Further, residual connections [19] followed by layer normalization are applied on top of fully connected layers for faster and more accurate training. The layer normalization is applied over the embedding dimension only. Finally, the self-attended embeddings of the input feature $\mathbf{E_M}$ are obtained as $\mathbf{SE_M} = \{\mathbf{sem}_1 \ldots \mathbf{sem}_l\}$ where $\mathbf{sem} \in \mathbb{R}^{d_m \times 1}$ and $\mathbf{SE_M} \in \mathbb{R}^{d_m \times l}$. For visual modality $\mathbf{SE_M} = \widetilde{\mathbf{rI}}(t-1)$ and for text $\mathbf{SE_M} = \widetilde{\mathbf{E_q}}(t-1)$.

4.2.3 Co-Attention

For cross-modal interactions, the co-attention module intakes the representations of two modalities and generates attention in context of each other. To facilitate this, the self-attended embeddings $\widetilde{\mathbf{E_q}}(t-1)$ and $\widetilde{\mathbf{rI}}(t-1)$ are taken as input. For generating image attention in context of question words, keys and values are generated from self-attended intermediate question representation while the query is obtained from the image itself (following Equation 4.6). Thus, the query ($Q_C^{(h)}$), key ($K_C^{(h)}$) and value ($V_C^{(h)}$) are respectively computed as follows.

$$Q_C^{(h)} = \left(W_h^{QC}\right)^\intercal \widetilde{\mathbf{E_q}}(t-1) \tag{4.8}$$

$$K_C^{(h)} = \left(W_h^{KC}\right)^\intercal \widetilde{\mathbf{rI}}(t-1) \tag{4.9}$$

$$V_C^{(h)} = \left(W_h^{VC}\right)^\intercal \widetilde{\mathbf{E_q}}(t-1) \tag{4.10}$$

Here, $W_h^{QC} \in \mathbb{R}^{d_m \times d_{KQ}}$, $W_h^{KC} \in \mathbb{R}^{d_m \times d_{KQ}}$ and $W_h^{VC} \in \mathbb{R}^{d_m \times d_{KV}}$ are linear transformations. Similarly, for cross-modal question attention, the query is obtained from self-attended question embeddings. While the keys and values are obtained from self-attended image embeddings. These queries, keys and values are similarly processed following Equations 4.6 and 4.7 to obtain the multi-head attention. This is fed to fully connected layers with ReLU, dropout, skip connections and layer normalization. The

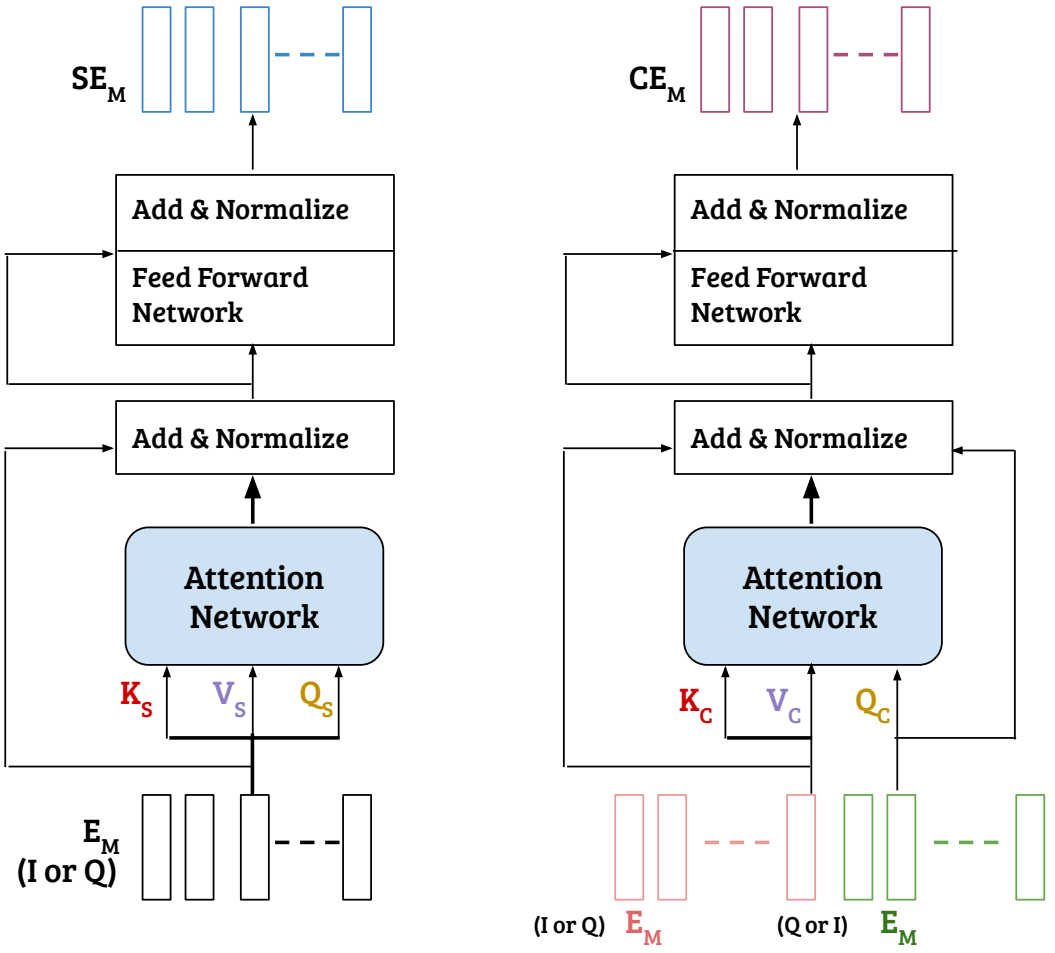


Fig. 4.5 *Self-attention* and *Co-attention* mechanism overview. Here, E_M denotes the input modality. For self-attention K_S, V_S, Q_S are obtained from the same input modality, while for cross modal attention Q would be from another modality.

output of this network provides the final output of the co-attention module. Figure 4.5 demonstrates the overview of the *self-attention* and *co-attention* mechanism.

4.2.4 Cascading & Fusion

A single SCA block comprising of *self-attention* (intra-modality interaction) and *co-attention* (inter-modality interaction) generates an enriched representation $(\mathbf{rI}(t), \mathbf{E_q}(t))$ of its input visual and textual features.

Existing works [11][112] suggest the stacking of multiple such blocks to obtain further fine grained representations. This is accomplished by cascading multiple SCA block to T stages. Let $\mathbf{rI}(T) \in \mathbb{R}^{d \times n_v}$ and $\mathbf{E_q}(T) \in \mathbb{R}^{d \times n_w}$ be the respective visual and question representations obtained from the final (T^{th}) SCA block.

The feature representations are obtained by averaging the attended embeddings of two modalities. So, the final visual embedding $\mathbf{I}_f$ is obtained as follows.

$$\mathbf{I}_f = \frac{1}{n_v} \sum_{i=1}^{n_v} \mathbf{rI}(T)[:, i] \tag{4.11}$$

Similarly, the question encoding $\mathbf{Q}_f$ is evaluated in the following manner.

$$\mathbf{Q}_f = \frac{1}{n_w} \sum_{j=1}^{n_w} \mathbf{E_q}(T)[:, j] \tag{4.12}$$

The unified multimodal representation $\mathbf{F} \in \mathbb{R}^{d \times 1}$ is obtained by fusing $\mathbf{I}_f$ and $\mathbf{Q}_f$ through element-wise multiplication.

$$\mathbf{F} = \mathbf{I}_f \odot \mathbf{Q}_f \tag{4.13}$$

The fused embedding $\mathbf{F}$ is fed to a fully connected network for answer prediction.

4.2.5 Answer Prediction

The fused embedding $\mathbf{F}$ is fed to fully connected network with single hidden layer of dimension d_{hp}. The number of labels at the output layer is n_c ($n_c = |\mathscr{A}|$). The output answer probability vector $\hat{\mathbf{a}}$ is predicted as follows.

$$\hat{\mathbf{a}} = \text{FCNet}\left(\mathbf{F}; d_{hp}; n_c\right) \tag{4.14}$$

4.2.6 Model Learning

Let, a be the ground-truth answer corresponding to the input image-question pair (I, q). An one-hot-encoded vector $\tilde{\mathbf{a}} \in \{0,1\}^{n_c}$ is constructed as a target prediction corresponding to the ground-truth answer a. This model uses cross-entropy loss for answer prediction and is defined as

$$\mathscr{L}_c(I, q, a) = -\sum_{j=1}^{n_c} \tilde{\mathbf{a}}[j] log(\hat{\mathbf{a}}[j]) \tag{4.15}$$

The combined set of parameters for proposed model includes the ones for feature extraction, blocks of dense attention and fusion mechanism.

Table 4.1 Category-wise comparison of CSCA with previous state-of-the-art methods on the TDIUC dataset

Question Type	**SAN [11]**	**RAU [1]**	**MCB [36]**	**QTA [16]**	**BAN [61]**	**CSCA**
Scene Recognition	92.3	93.96	93.06	93.80	93.1	**94.48**
Sport Recognition	95.5	93.47	92.77	95.55	95.7	**95.85**
Color Attributes	60.9	66.86	68.54	60.16	67.5	**75.51**
Other Attributes	46.2	56.49	56.72	54.36	53.2	**60.89**
Activity Recognition	51.40	51.60	52.35	60.10	54.0	**61.00**
Positional Reasoning	27.9	35.26	35.40	34.71	27.9	**42.14**
Object Recognition	87.50	86.11	85.54	86.98	87.5	**89.11**
Absurd	93.4	96.08	84.82	**100.0**	94.47	97.28
Utility & Affordance	26.3	31.58	35.09	31.48	24.0	**40.35**
Object Presence	92.4	94.38	93.64	94.55	95.1	**96.34**
Counting	52.1	48.43	51.01	53.25	53.9	**60.70**
Sentiment Und.	53.6	60.09	66.25	64.38	58.7	**67.19**
Overall Accuracy	82.0	84.26	81.86	85.03	85.5	**88.12**
Harmonic Mean	53.7	59.00	60.47	60.08	54.9	**67.05**
Arithmetic Mean	65.0	67.81	67.90	69.11	67.4	**73.34**

Table 4.2 Comparing *Overall Accuracy* of CSCA for TDIUC dataset

Model	**Overall Accuracy**	**Arithmetic Mean**
BTUP[12]	82.91	68.82
QCG[71]	82.05	65.67
BAN2-CTI[62]	87.00	72.5
DFAF[14]	85.55	**NA**
RAMEN[31]	86.86	72.52
MLIN[15]	87.60	NA
CSCA	**88.12**	**73.34**

Table 4.3 Performance of CSCA on TDIUC data (except Absurd category samples) trained without 'Absurd' Category samples

Metrics	**MCB [36]**	**QTA [16]**	**BAN [61]**	**BAN2-CTI [62]**	**CSCA**
Overall Accuracy	78.06	80.95	81.9	85.0	**85.30**
Arithmetic-MPT	66.07	66.88	64.6	70.6	**71.21**
Harmonic-MPT	55.43	58.82	52.8	63.8	**65.40**

4.3 Results and Discussion

The proposed approach of CSCA-VQA is benchmarked on the *VQA2.0* and *TDIUC* datasets. This section presents the experimental setup details (Sub-section 4.3.1), quantitative analysis (Sub-section 4.3.2), basic analysis (Sub-section 4.3.3), ablation studies (Sub-section 4.3.4), and qualitative results (Sub-section 4.3.5) of the experiments.

4.3.1 Implementation Details

For visual feature representation, $n_v = 36$ (for TDIUC) and $n_v = 100$ (for VQA2.0) image regions are extracted. Dimensions of each image region feature is taken as $d_v = 2048$. The question length set to $n_w = 14$ words by trimming or padding (as necessary). The GloVe word embeddings of $d_w = 300$ dimensions are considered. The hidden space dimensions are kept as: $d = 512$, $d_{KQ} = 64$. For multi-head attention, $n_h = 8$ number of heads are used. Model is trained for 15 epochs with batch size of 64 samples for experiments and analysis. The number of hidden layer nodes of the answer prediction sub-system is set to $d_{hp} = 1024$. The Adamax optimizer [117] is used with a decaying step learning rate. The initial learning rate is set to 0.002 and it decays by 0.1 after every 5 epochs. The proposed model is built on PyTorch framework and is trained on NVIDIA-GTX 1080 GPU.

4.3.2 Quantitative Results

Overall Performance & Category-wise Performance Comparison on TDIUC Dataset – Tables 4.1 and 4.2 present the respective class-wise and overall performance for the TDIUC dataset. In terms of the overall accuracy, Arithmetic-MPT (AMPT) and Harmonic-MPT (HMPT) measures, the proposed model CSCA exhibits better performance compared to most of the baseline methods. Also, in terms of class-wise accuracy, CSCA leads in all except one class. A significant relative gain of 12.6% is observed compared to the next best performing model for the *'Counting'* category of questions. Table 4.3 presents the results for different models trained *'Without Absurd'* category of questions. It is observed that CSCA performs better than the existing ones for all three defined evaluation metrics.

Overall Performance & Category-wise Performance Comparison on VQA2.0 Dataset – Table 4.4 demonstrates the results on test-dev and test-std splits of the VQA2.0 dataset. Performance of the proposed model CSCA is comparable with that of the best among the existing methods. The models LXMERT [68], ViLBERT [69] are pretrained for multiple vision and language based tasks and are fine-tuned for VQA.

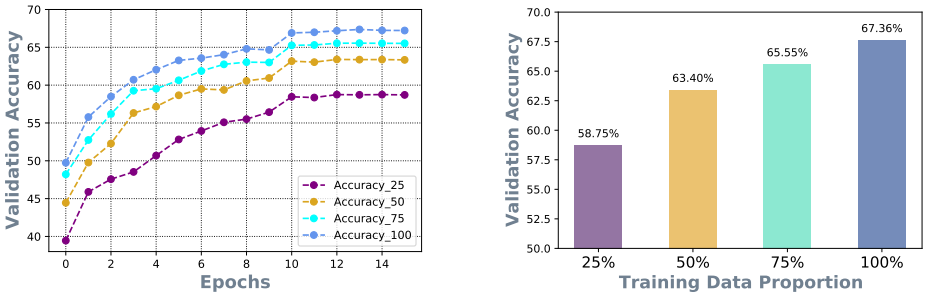
Here, CSCA has obtained 67.36% accuracy on the validation set. This is around 1% improvement over the best performance among the existing methods.

Table 4.4 Model performance on VQA 2.0 dataset: Validation, Test-Dev & Test-Std splits. CSCA is compared with several state-of-the-art methods including *Fusion based*, *Visual Attention* and *Dense Attention* based methods (separated by horizontal lines).

Methods	**Val**	**Test-Dev**				**Test-Std**
	Overall	**Yes / No**	**Number**	**Other**	**Overall**	**Overall**
MCB [35]	59.14	78.46	38.28	57.80	62.27	53.36
MLB [37]	62.98	83.58	44.92	56.34	66.27	66.62
MUTAN [13]	62.71	82.88	44.54	56.50	66.01	66.38
MFH [39]	62.98	84.27	49.56	59.89	68.76	–
BLOCK [40]	64.91	83.14	51.62	58.97	68.09	68.41
SAN [11]	61.70	78.40	40.71	54.36	61.70	–
BTUP [12]	63.20	81.82	44.21	56.05	65.32	65.67
BAN [61]	65.81	82.16	45.45	55.70	64.30	–
v-VRANet [118]	–	83.31	45.51	58.41	67.20	67.34
ALMA [119]	–	84.62	47.08	58.24	68.12	66.62
ODA [120]	64.23	83.73	47.02	56.57	66.67	66.87
BAN2-CTI [62]	66.00	–	–	–	–	67.4
CRANet [121]	–	83.31	45.51	58.41	67.20	67.34
CoR [57]	65.14	84.98	47.19	58.64	68.19	68.59
MUREL [73]	65.14	84.77	49.84	57.85	68.03	68.41
DFAF [14]	66.66	86.09	53.32	60.49	70.22	70.34
MLIN [15]	66.53	85.96	52.93	60.40	70.18	70.28
LXMERT [68]	–	–	–	–	–	**72.5**
ViLBERT [69]	–				70.55	70.92
CSCA	**67.36**	86.57	**53.58**	**61.06**	**70.72**	71.04

4.3.3 Basic Analysis

Effect of Training Data Size on Performance – An analysis is performed to observe the effect of the variation of training dataset size on model performance. The primary objective of this experiment was to ascertain whether a model trained on a smaller dataset can provide similar performance as the one learned from the complete set. To explore this, the model is trained with four different datasets obtained from the original VQA2.0 dataset. The first three datasets are obtained by random shuffling of all samples of the VQA2.0 dataset followed by the extraction of 25%, 50% and 75% samples. The fourth one is the complete VQA2.0 dataset (i.e. 100%). Other experimental setups like hidden dimension, number of answer classes are kept similar to the original setup for all variants of the dataset. The Epoch-wise performances for the four different datasets are shown in Figure 4.6a. As expected, the model

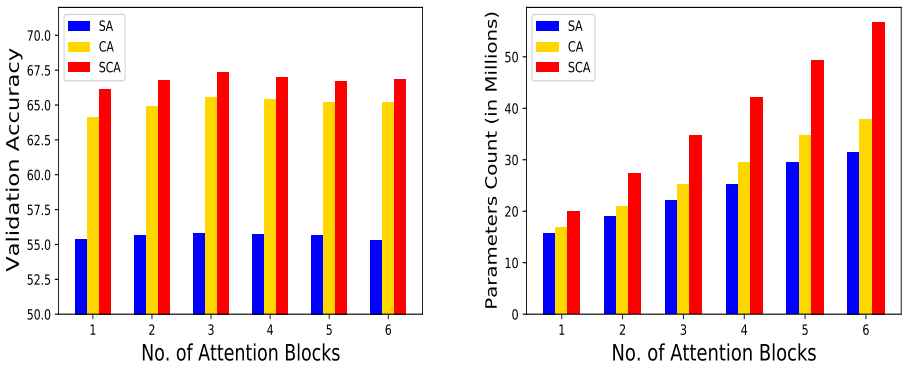


(a) (b)

Fig. 4.6 Illustrating the learning curves on training datasets formed with different amounts of instances from VQA2.0. (a) Performance on validation set of VQA2.0 with respect to the number of epochs. (b) Overall accuracy for VQA2.0 dataset with different proportion of the training data.

performance improved with an increase in training dataset size. It can be observed that in all four settings, the model performance evolves over a different number of epochs in a similar fashion. However, Figure 4.6b indicates that the relative gain achieved by increasing the training dataset size from 25% to 50% is significant compared to that by increasing from 50% to 75% or 75% to 100%. This observation may be attributed to the fact that in a collection of randomly shuffled datasets, not many novel instances were encountered during the subsequent increase of the training data.

Effect of Number of SCA Blocks – In one pass, it is difficult for a model to grasp all relevant information through a representation. Thus, attention blocks in cascade extract the fine-grained information and pass it on to the next one for further refinement. A set of experiments are performed to identify the optimal number of blocks in the cascade. Additionally, the effect of different independent attention mechanisms (SA only, CA only, SCA) for answer prediction is also analyzed. In Figure 4.7a, overall performance for validation split of VQA2.0 dataset is given with respect to varying number of blocks. Figure 4.7b shows the parameter counts with respect to the number of blocks. As per expectation, it is observed that the models perform poorly with single attention blocks (SA only, CA only, SCA). However, the performance is observed to rise only up to four number of blocks. Increasing the number of blocks beyond four does not lead to any further performance improvement. However, adding more blocks also lead to an increase in the number of model parameters (Figure 4.7b). Furthermore, one can observe that only CA module can perform better than using only the SA module. This is as per the expectation. Similarly, Figure 4.8 shows that the model performance keeps improving until the fourth SCA block for the TDIUC dataset. The model performance starts deteriorating with a further increase in the number of blocks.



(a) (b)

Fig. 4.7 Number of attention blocks incorporated. (a) Validation accuracy for VQA2.0 *'val'* split with respect to attention blocks. (b) Parameter counts with respect to attention blocks

4.3.4 Ablation Analysis

The proposed model performs self-attention on the two modalities to obtain intra-modality correlated features. Then the co-attention module uses respective representations of the two modalities to obtain cross-modality correlated features by performing attention for one modality in the context of another. In this ablation analysis, we examine the impact of individual attention module in various combinations to understand their importance. We also analyze the set of correct predictions obtained in these settings.

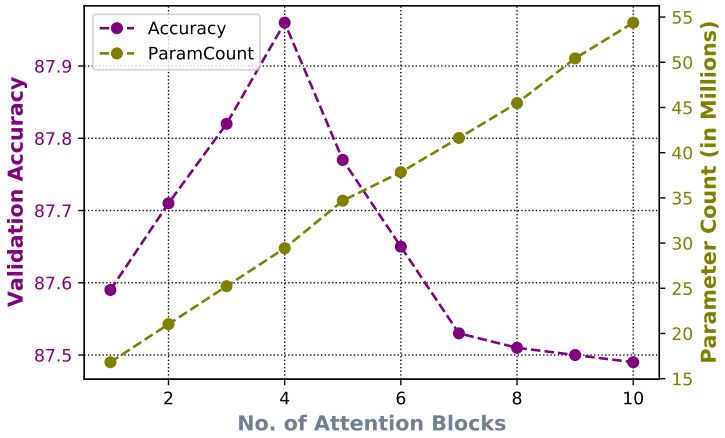


Fig. 4.8 *Validation Accuracy* and Parameter Count (in Millions) for *TDIUC* dataset with respect to the number of SCA blocks incorporated in the VQA model.

Table 4.5 and 4.6 present the results of ablation analysis experiments in terms of performance and complexity. The complexity is expressed in terms of the number of model parameters. The first row of the table shows the model performance when neither of the attention is incorporated. The features for both modalities are fused

Table 4.5 Evaluating model performance on VQA2.0 dataset to investigate the effect of *different basic attention modules* of the proposed model

SA	CA	Yes / No	Number	Other	Overall Accuracy	Parameter (in Millions)
✗	✗	69.95	36.42	50.19	55.80	22
✓	✗	79.08	40.75	49.96	59.69	15
✗	✓	81.17	44.63	56.34	64.13	25
✓	✓	**84.92**	**49.51**	**58.71**	**67.36**	**42**

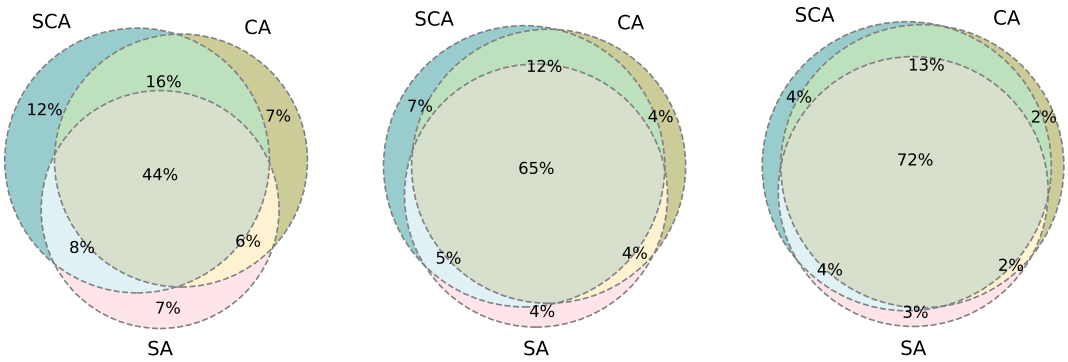
Table 4.6 Evaluating model performance on TDIUC dataset to investigate the effect of *number of attention blocks* and self-attention & cross attention.

SA	CA	Overall Accuracy	Parameter (in Millions)
✗	✗	69.18	7
✗	✓	70,46	21
✓	✗	87.42	25
✓	✓	**88.12**	**36**

directly via element-wise multiplication without applying self- or co-attention. Second row shows the performance when *only self-attention* (SA only) is incorporated on both modalities and answer prediction is based on the fused embedding of the self-attended representations of the individual modalities. Here, the fused representation is obtained via element-wise multiplication. Third row shows the results when *only co-attention* (CA only) is incorporated on image and question in the context of the other. The last row shows the results from the proposed model that comprises of both *self-attention* and *co-attention* in cascade (SCA).

As per expectation, the model without any attention mechanism provides the lowest performance (first row). The "SA only" model provides lower performance as it lacks the interaction of two modalities and learns a comparatively poor representation (second row). Co-attention is the crucial component for multi-modality that is found to perform better than *self-attention*. In terms of computational complexity, a simple fusion-based model uses the least number of parameters, while the proposed model (SCA) requires the highest number of parameters. However, the performance improvement, especially for VQA2.0 dataset, overcomes the complexity issue. We observe that the change in model performance is similar for both datasets in this analysis.

Figure 4.9 shows the model's performance over various attention mechanisms for the different types of questions category on VQA2.0 dataset. The following are observed from the results for the *'Number'* category of questions. While using the SA only and CA only blocks, the respective models show the overall performances of 65% and 73%. Models using SA and CA attention individually predicts 7% of samples correctly that are not correctly classified by any of the other models. Similarly,



(a) 'Number' Questions (b) 'Other' Questions (c) 'Yes/No' Questions

Fig. 4.9 SCA: Self-Attention & Co-attention, SA: Only self-attention is applied on text and visual features, CA: Cross-Modality Attention on text as well as on visual features guided by each other.

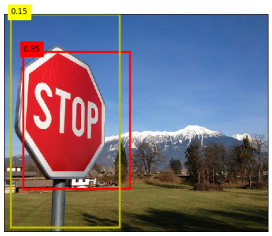
the model using SCA block classifies 12% of samples correctly that are not correctly classified either by the models using SA or CA only. Thus, the models using SCA blocks achieved the best performance. The same pattern was observed over the other question types i.e., *'Yes/No'* and *'Other'*. The detailed result for all the question types are shown in figure 4.9.

4.3.5 Qualitative Results

The qualitative results are presented in Figure 4.10 to demonstrate the efficacy of the proposed model. For this, two salient regions of a given image with the highest attention scores are highlighted. These are the attention scores obtained after cascading $T = 4$ SCA blocks. The question words that obtain the highest attention scores are also highlighted. As evident from Figure 4.10a, the proposed model CSCA is able to focus on relevant image regions and question words. The top-2 salient regions corresponding to the binary question *"Are there any cows in the picture?"* are the ones that capture the *cows* and hence, the model responds by the answer *'Yes'*. Similarly, Figures 4.10b, 4.10c, 4.10d, 4.10e, 4.10f, 4.10g, 4.10h, 4.10i, 4.10j, 4.10k, 4.10l show that the model is trying to identify the salient image regions and relevant question words to predict the appropriate answer.

However, the model made errors as well. One of the reasons was incorrect attention to image regions. As shown in Figure 4.11a, the model's focus is primarily on the position from where it seems like this room is a kitchen. If the attention is given to other regions, the answer will likely change to '*living room*'. In 4.11b for question '*What color is the wall in back of the desk ?*', the model focuses on the other side of the

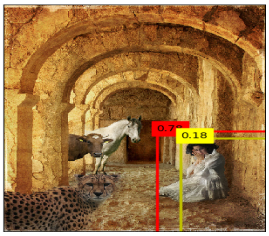
(a) **Q.** Are there any cows in the picture ?
Ans: Yes ✔ (0.59, 0.21)
GT: Yes



(b) **Q.** What shape is the stop sign ?
Ans: Octagon ✔(0.85, 0.15)
GT: Octagon

(c) **Q.** How many potted plants ?
Ans: 2 ✔ (0.80, 0.17)
GT: 2

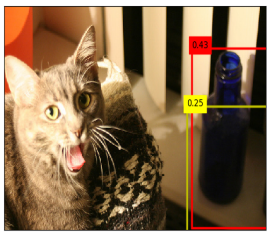
(d) **Q.** What is the color of the ground ?
Ans: White ✔ (0.22, 0.22)
GT: White



(e) **Q.** What is the lady doing ?
Ans: Sitting ✔ (0.78, 0.18)
GT: Sitting

(f) **Q.** What is the weather like ?
Ans: Cloudy ✔ (0.37, 0.28)
GT: Cloudy

(g) **Q.** What colors are the chair ?
Ans: Brown ✔ (0.78, 0.18)
GT: Brown

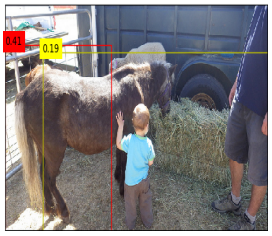


(h) **Q.** Are there any bottles in the photo ?
Ans: Yes ✔ (0.43, 0.25)
GT: Yes

(i) **Q.** How many horses are there?
Ans:1 ✔ (0.31, 0.25)
GT: 1

(j) **Q.** Is there any food in the photo?
Ans: Yes ✔ (0.53, 0.23)
GT: Yes

(k) **Q.** Are there any umbrellas in the photo?
Ans: Yes ✔ (0.32, 0.15)
GT: Yes



(l) **Q.** Is there a horse in the photo ?
Ans: Yes ✔ (0.41, 0.19)
GT: Yes

Fig. 4.10 Qualitative results for our proposed method CSCA. Attention for image obtained with cascade of $T = 4$ SCA blocks is presented. (top1, top2) attention score values correspond to the top two attention weight obtained for top-2 salient regions, that are relevant to infer the answer. The question words shown in blue are the ones that get the highest attention score.

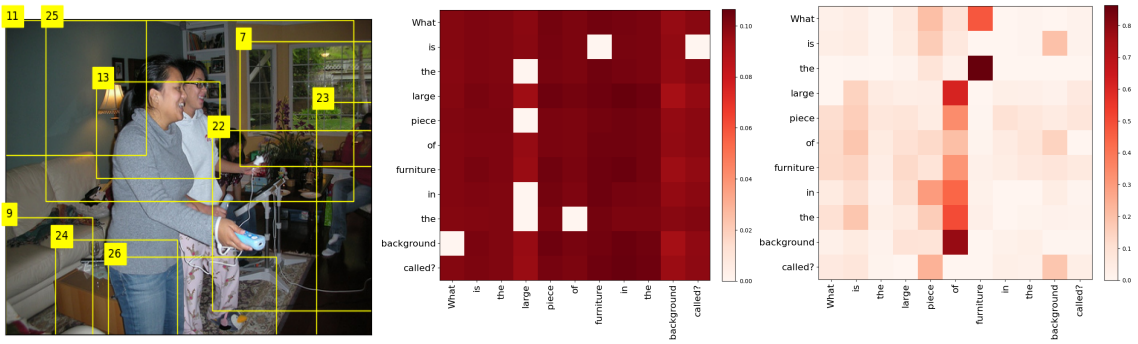
desk instead of the back. The predicted answer is ‘green’, the color on the side-wall of the desk.

(a) **Q.** What room is shown in the picture ?
Ans: Kitchen ✗ (0.48, 0.13)
GT: Living Room

(b) **Q.** What color is the wall in back of the desk ?
Ans: Green ✗ (0.79, 0.15)
GT: Gray

Fig. 4.11 Failure cases where wrong attention leads to incorrect answer prediction.

Attention Map Visualization – Consider the image-question pair (I, q) shown in Figure 5.7(a). The attention maps obtained from the self- and cross-attention modules of the first (SCA(1)) and final SCA block (SCA(4)) are visualized in Figures 4.12 – 4.13. Figure 4.12a shows salient regions in context of text semantics. Figures 4.12b, 4.12c show the Question-on-Question (QoQ) self-attention map for the question obtained from the SA modules associated with q in SCA(1) and SCA(4) respectively. It can be observed that after SCA(1), the self-attention map consider most words in the question. However, after SCA(4), the model is seen to focus on more relevant words like ‘*what*’, ‘*large*’, ‘*background*’ and ‘*furniture*’, while the attention on remaining words reduces. Figures 4.13a, 4.13b show the Question-on-Image (QoI) cross-attention map obtained



(a) (b) (c)

Fig. 4.12 Attention of Question-on-Question (QoQ) as output of SA module (associated with question) of (a) Bounding box visualization (b) first SCA block $SCA(1)$, and (c) final SCA block $SCA(4)$. Note the differential attention values and corresponding changes in correlation between relevant word pairs like (*what*, *furniture*), (*the*, *furniture*), (*background*, *of*) etc. Here, *Dark* color signifies higher attention value.

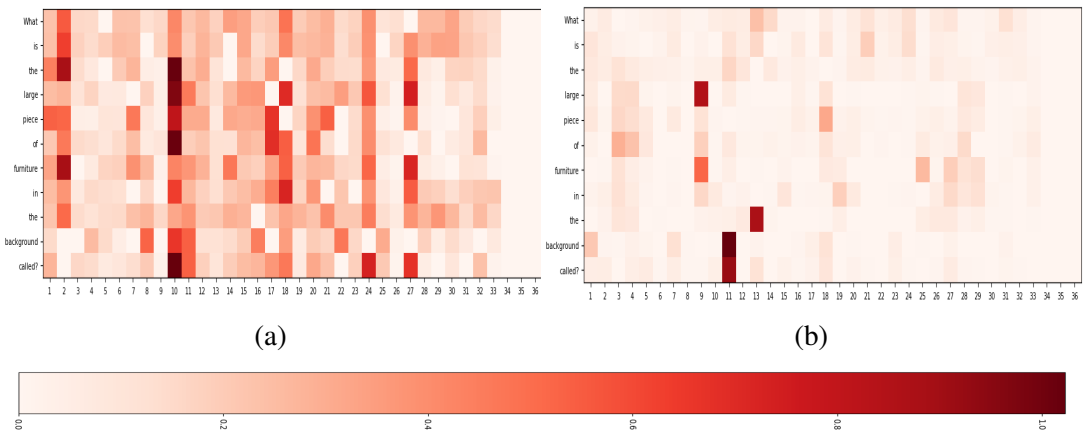


Fig. 4.13 Attention Map Visualization for (I, q) pair of Figure 5.7(a): Attention of Question-on-Image (QoI) as output of CA module (associated with image) of (a) first SCA block $SCA(1)$, and (b) final SCA block $SCA(4)$. Here, *Dark* color signifies higher attention.

from the CA modules associated with I in SCA(1) and SCA(4) respectively. The figure-4.13a and 4.13b shows the cross-attention using the question word to the image region. The attention map obtained after SCA(1) shows a poor attention distribution (Figure 4.13a). However, this significantly improves at the output of the CA module of last block SCA(4) (Figure 4.13b). The correlations between image-regions and question-words are represented by this attention map. It is observed that specific words like *large*, *furniture*, *background* have developed higher correlations with few most semantically relevant regions e.g., $9, 11$. Also, the remaining image-regions and question-words have very low correlation and they mostly do not contribute to the final joint embedding.

4.4 Discussions

This chapter presented a VQA model that incorporated a dense attention mechanism to extract informative features from the input data. The dense attention mechanism was achieved by exploiting both self-attention and co-attention. This enabled the model to capture both the intra-modality and cross-modality interactions of the input features. The self-attention mechanism allowed the model to obtain improved representations within a single modality. For instance, in the case of an image, a salient region interacts with every other region, and the final representation inherits the contextual information for all regions. Similarly, for an input question, self-attention provides a representation of every single word, thereby capturing the contextual information for other words as well. The proposed model also utilized cross-modal interaction between two modalities, which was further strengthened by self-attention of the two modalities. Attention blocks were cascaded multiple times to facilitate refined cues of

visual and textual features, enhancing the model's ability to extract more informative features from the input data. The model's effectiveness was confirmed through detailed experiments and analysis performed on two benchmark VQA datasets. The results demonstrated that the proposed dense attention-based VQA model outperformed several state-of-the-art VQA models, thereby achieving significant improvements in the accuracy.

In the next chapter, the VQA problem is approached by leveraging the prior question category information associated with every image-question pair. This information is utilized to reduce the search space for answer classification. Instead of classifying the answer from a larger space, it is predicted from a smaller set partitioned based on the question category information. This approach aims to improve the efficiency of the VQA model and reduce the computational overhead of the classification process.

Chapter 5

Dual Attention and Question Categorization based Visual Question Answering

Chapter Highlights

- The available VQA datasets have question category information associated with each image, question, answer triplet. However, existing works in the literature rarely used this information.
- Inspired from human behavior of question answering, this chapter proposes a novel architecture for VQA that exploits the question category information.
- Additionaly, an improved feature representation is obtained by *dual attention*.
- *Extensive Experiments* are performed on two benchmark VQA datasets, viz., *TDIUC* and *VQA2.0* to demonstrate the efficacy of the proposal in terms of *Overall Performance* and Question Category-wise Performance.
- Publications related to this chapter are as follows:
  1. **Aakansha Mishra**, Ashish Anand, Prithwijit Guha, *CQ-VQA: Visual Question Answering on Categorized Questions*. In Proceedings of International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8.
  2. **Aakansha Mishra**, Ashish Anand, Prithwijit Guha, *Dual Attention and Question Categorization based Visual Question Answering*, IEEE Transaction on Artificial Intelligence (TAI), vol. 4(1), pp. 81–91, 2022

5.1 Introduction

Attention mechanism has played a significant role in improving the VQA model performance. Initial VQA models adopting attention mechanism [6, 11–18, 97] focused

on finding relevant regions in image pertaining to the given question. Attention on image regions using textual features has become a default component of VQA models. However, recent studies [54][49][122][112] have indicated that image conditioned attention on question further helps models obtain improved question representation. Thus, to obtain an enriched representation with cross-modality interactions, dual attention mechanism is incorporated with the text and image modalities.

Furthermore, some studies [16][123] indicate that reducing the answer search space with the help of *Question Categorizer* helps in performance improvement. Such an approach is motivated by the general human behaviour for answering a question. For example, consider the input question *"What is the color of grass?"*. Realizing the fact that the question is about *color*, helps in simplifying the task in choosing the answer as a color name. Similarly, a VQA model may first identify the question category (*color*, say). Thus, instead of exploring the entire answer space, the question category information helps the VQA model to focus on a smaller search space (e.g. answers specific to *color* category only).

Motivated by the above observations, this chapter proposes the **D**ual **A**ttention and **Q**uestion **C**ategorization based Visual Question Answering system (DAQC-VQA). The DAQC-VQA combines subsystems for *Dual Attention*, *Question Categorization* and *Answer Prediction*. Figure 5.1 illustrates the overview of DAQC-VQA. The proposed model uses a dual attention mechanism to obtain enriched cross-domain textual and image features at the first stage. The question classifier subsystem uses the fused features of the two modalities to obtain the question category. Question classification is followed by an activated answer prediction network corresponding to the predicted question category.

The key contributions of this work can be summarized as follows.

- *Dual Attention* – Attention on Image (AoI) and Attention on Question (AoQ) to obtain an enriched representation of both modalities.
- *Question Categorization* – Question type identification for answer space reduction leading to performance improvement in answer prediction.
- *Extensive Experiments* on two benchmark VQA datasets, viz., *TDIUC* and *VQA2.0* to demonstrate the efficacy of DAQC-VQA in terms of *Overall Performance* and Question Category-wise Performance.

5.2 Proposed Method

The *Visual Question Answering* (VQA) system predicts an answer probability vector $\hat{\mathbf{a}}$ in response to an input image $\mathrm{I} \in \mathscr{I}$ and an associated natural language question

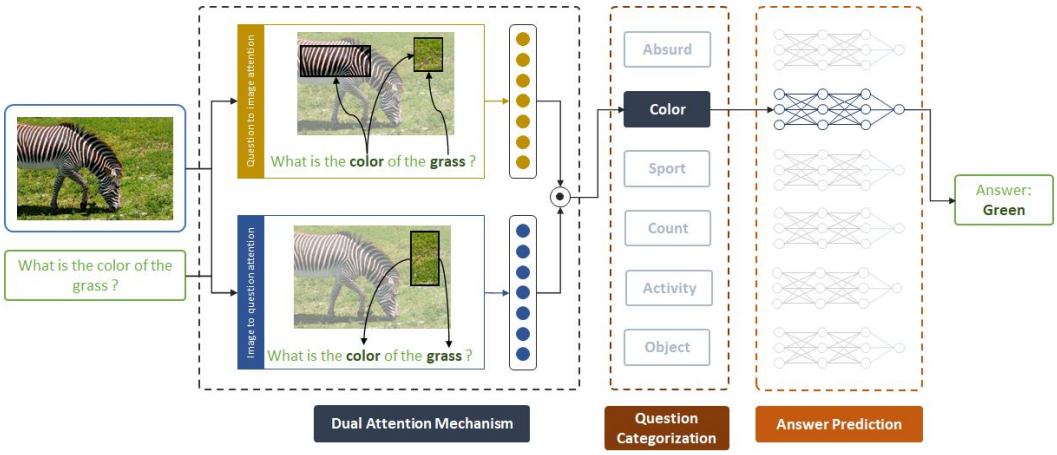


Fig. 5.1 An overview of the proposed DAQC-VQA system. Image and question will be input to the model. Attention is generated for salient regions of image in context of question and for each word of question in context of attended image. On the basis of attended fused representation question category will be predicted. On the basis of predicted question category, answer will be classified from the set of answers corresponding to that category.

$q \in \mathscr{Q}$. Following recent works [12][2][97][98], this proposal formulates the VQA task as a classification problem. Here, $\hat{\mathbf{a}}$ is predicted using features of the inputs (I, q).

A pretrained object proposal network (Faster-RCNN [23]) is used to capture the most prominent regions of the input image I. These region proposals are further processed by the pretrained ResNet-101 network [19] for visual feature extraction. Similarly, the GloVe embeddings [29] of the words in q are processed by a LSTM network for computation of question encoding. Detailed descriptions of visual and textual feature extraction are provided in the Subsection 3.2.1.

These visual and textual features (or embeddings) are *attended* to further focus on the prominent image regions and words in q. This attention mechanism is elaborated in Subsection 5.2.2. These attended visual and textual features are then fused by elementwise multiplication to obtain a (multimodal) joint embedding (Subsection 5.2.3). This joint embedding is used further for answer prediction.

Classification of large number of categories (here, the number of answers $n_c = |\mathscr{A}|$) is often considered a hard problem. However, the answer set $\mathscr{A}$ can be decomposed into its subsets by using an additional (and often available) information on question categories. Instances of such question categories are *Yes/No*, *Color specific*, *Object specific*, *Action specific* etc. For example, the *Yes/No* category questions will have a two element answer set $\{yes, no\}$. Similarly, other question categories will have their corresponding answer sets of lesser size.

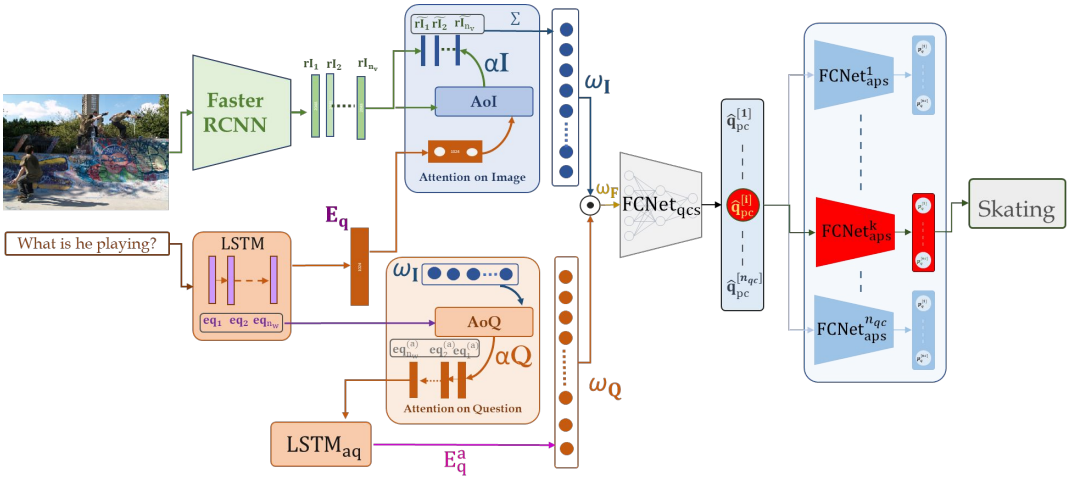


Fig. 5.2 The functional block diagram of DAQC-VQA. Features are extracted from two input modalities. These feature representations are exploited for generating attention scores in context of the other modality. AoI represents the Attention-on-Image module based on LSTM based question encoding. AoQ shows the Attention-on-Question in context of attended visual representation. $\text{FCNet}_{\text{qcs}}$ is the question category classifier. And, $\text{FCNet}^{(l)}_{\text{aps}}$ represents the l^{th} answer prediction sub-module, which performs the final answer prediction.

Let $\text{q}_\text{c} \in \mathscr{Q}_\mathscr{C}$ ($\mid \mathscr{Q}_\mathscr{C} \mid = n_{qc}$) be the category label of input question q. This proposal first classifies an input question q to one of these n_{qc} categories. Each question category corresponds to an answer set $\mathscr{A}_m$ ($\cup_{m=1}^{n_{qc}} \mathscr{A}_m = \mathscr{A}$). The question categorization stage leads to the selection of one from n_{qc} independent answer prediction subsystems. The final predicted answer is provided by the selected subsystem.

The components of the VQA system are trained by using the attended features of the joint visual and textual modalities to minimize loss functions over question categories and answers for all input pairs $(\text{I}, \text{q}) \in \mathscr{I} \times \mathscr{Q}$ and corresponding ground-truth pairs $(\text{q}_\text{c}, \text{a}) \in \mathscr{Q}_\mathscr{C} \times \mathscr{A}$. Figure 5.2 illustrates the functional block diagram of DAQC-VQA. This consists of the following subsystems – (a) visual and (b) textual feature extraction; attention mechanism for (c) image and (d) text embeddings; (e) feature fusion; (f) question categorization, and (g) answer prediction. These subsystems are detailed in the following subsections.

5.2.1 Feature Extraction

Visual features (**rI**) are extracted by the process detailed in section 3.2.1 and are given as follows

$$\mathbf{rI} = [\mathbf{r}_1, \ldots \mathbf{r}_i, \ldots \mathbf{r}_{n_v}] \quad , \quad \forall \mathbf{r}_i \in \mathbb{R}^{d_v \times 1} \tag{5.1}$$

where $\mathbf{r}_i$ is the embedding of the i^{th} ($i = 1, \ldots n_v$) image region.

Textual features (**Eq**) are obtained as defined in 3.2.1 and are given as follows

$$\mathbf{Eq} = [\mathbf{eq}_0, \ldots \mathbf{eq}_j, \ldots \mathbf{eq}_{n_w}] \;\; \& \;\; \forall \mathbf{eq}_j \in \mathbb{R}^{d_w \times 1} \tag{5.2}$$

A question encoding $\mathbf{q_e} \in \mathbb{R}^{d_q \times 1}$ is further obtained by processing **Eq** by a LSTM network $LSTM_Q$.

5.2.2 Attention Mechanism

DAQC-VQA exploits dual attention on the following two modalities. These are *Attention on Image* (AoI) and *Attention on Question* (AoQ), and are discussed as follows.

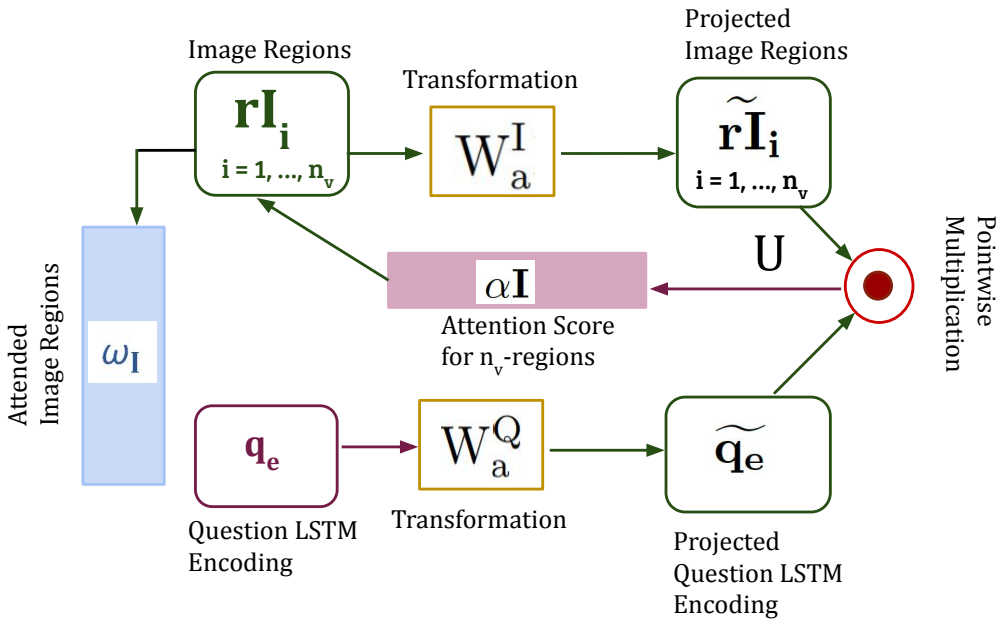


Fig. 5.3 Block diagram for demonstrating the Attention on Image guided by question.

Attention on Image (AoI) signifies the focus on important image regions with respect to the words in the question. The task of this subsystem is to compute an unified and attended visual embedding $\omega_\mathbf{I}$ using the attention scores $\alpha_I^{(i)} \in (0,1)$ corresponding to each $\mathbf{r}_i$ $(i = 1, \ldots n_v)$. Here, a higher value of $\alpha_I^{(i)}$ indicates a greater correlation between the i^{th} image region and q. The embedding $\omega_\mathbf{I}$ is obtained as the attention score weighted sum of the image region embeddings. The unattended visual features $\mathbf{r}_i$ and the question encoding $\mathbf{q_e}$ are first transformed to a d_{hi} dimensional

space followed by a non-linear transformation (Equations 5.3 and 5.4).

$$\tilde{\mathbf{r}}_i = \sigma\left(W_a^I \, \mathbf{r}_i\right) \tag{5.3}$$

$$\tilde{\mathbf{q}}_\mathbf{e} = \sigma\left(W_a^Q \, \mathbf{q}_\mathbf{e}\right) \tag{5.4}$$

Here, $\tilde{\mathbf{r}}_i \in \mathbb{R}^{d_{hi} \times 1}$ $(i = 1, \ldots n_v)$ are the transformed visual features, $\tilde{\mathbf{q}}_\mathbf{e} \in \mathbb{R}^{d_{hi} \times 1}$ is the transformed question encoding, and $W_a^I \in \mathbb{R}^{d_{hi} \times d_v}$ and $W_a^Q \in \mathbb{R}^{d_{hi} \times d_q}$ are the transformation matrices.

The attended visual features $\widetilde{\mathbf{rI}}_\mathbf{i}$ $(i = 1, \ldots n_v)$ are element-wise multiplied (designated by $\odot$) with $\tilde{\mathbf{q}}_\mathbf{e}$ to produce intermediate embeddings, say $\mathbf{U} = [\mathbf{u}_1, \ldots \mathbf{u}_i, \ldots \mathbf{u}_{n_v}]$. The attention scores $\alpha\mathbf{I} = [\alpha_I^{(1)}, \ldots \alpha_I^{(i)}, \ldots \alpha_I^{(n_v)}]$ are obtained by linearly transforming $\mathbf{U}$ by using the parameter vector $W_a^A \in \mathbb{R}^{1 \times d_{hi}}$ (Equations 5.5 and 5.6).

$$\mathbf{u}_i = \widetilde{\mathbf{rI}}_i \odot \tilde{\mathbf{q}}_\mathbf{e} \tag{5.5}$$

$$\alpha\mathbf{I} = \text{SoftMax}\left(W_a^A \, \mathbf{U}\right) \tag{5.6}$$

The parameters of W_a^I, W_a^Q and W_a^A are learned during overall model training. The final unified and attended visual feature representation $\omega_\mathbf{I}$ is obtained as the attention score weighted sum of the n_v attended visual embeddings (Equation 5.7). The functional block diagram for obtaining attended visual representation is presented in Figure 5.3.

$$\omega_\mathbf{I} = \sum_{i=1}^{n_v} \alpha_I^{(i)} \times \mathbf{r}_i \tag{5.7}$$

The **Attention on Question** (AoQ) subsystem aims at generating attended question embedding by processing the attended visual representation $\omega_\mathbf{I}$. Each word embedding $\mathbf{eq}_j$ $(j = 1, \ldots n_w)$ is assigned a weight based on $\omega_\mathbf{I}$. A higher attention score is attributed to a word that is more relevant to the image semantics. The primarily attended visual embedding $\omega_\mathbf{I}$ and the word embeddings $\mathbf{eq}_j$; $(j = 1 \ldots n_w$ are non-linearly transformed to a d_{hq}-dimensional space (Equations 5.8 and 5.9).

$$\overline{\mathbf{eq}}_j = \sigma\left(U_a^Q \, \mathbf{eq}_j\right) \tag{5.8}$$

$$\overline{\omega} = \sigma\left(U_a^I \, \omega_\mathbf{I}\right) \tag{5.9}$$

Here, $\overline{\mathbf{eq}}_j \in \mathbb{R}^{d_{hq} \times 1}$ $(j = 1, \ldots n_w)$ are the transformed word embeddings, $\overline{\omega} \in \mathbb{R}^{d_{hq} \times 1}$ is the transformed visual representation, and $U_a^Q \in \mathbb{R}^{d_{hq} \times n_w}$ and $U_a^I \in \mathbb{R}^{d_{hq} \times d_v}$ are the transformation matrices.

The transformed embeddings $\overline{\mathbf{eq}}_j$ $(j = 1, \ldots n_w)$ and $\overline{\omega}$ are element-wise multiplied to obtain the intermediate embeddings $\mathbf{V} = \left[\mathbf{v}_1, \ldots \mathbf{v}_j, \ldots \mathbf{v}_{n_w}\right]$. These are further

transformed to obtain the word-wise attention scores $\alpha\mathbf{Q} = \left[\alpha_Q^{(1)}, \ldots \alpha_Q^{(j)}, \ldots \alpha_Q^{(n_w)}\right]$ using the parameter vector $U_a^A \in \mathbb{R}^{1 \times n_w}$ (Equations 5.10 and 5.11).

$$\mathbf{v}_j = \overline{\mathbf{eq}}_j \odot \overline{\omega} \tag{5.10}$$

$$\alpha\mathbf{Q} = \text{SoftMax}\left(U_a^A \, \mathbf{V}\right) \tag{5.11}$$

The word embeddings $\mathbf{eq}_j$ are multiplied with their corresponding attention scores $\alpha_Q^{(j)}$ to obtain the attended embeddings $\mathbf{eq_j^{(a)}} \in \mathbb{R}^{d_w \times 1}$ $(j = 1, \ldots n_w)$. These attended embeddings $\mathbf{E_q^a} \in \mathbb{R}^{d_w \times n_w}$ are input to the LSTM network $LSTM_{aq}$ to obtain the final attended question encoding $\omega_\mathbf{Q} \in \mathbb{R}^{d_q \times 1}$ (Equations 5.12, 5.13 and 5.14).

$$\mathbf{eq_j^{(a)}} = \alpha_Q^{(j)} \times \mathbf{eq}_j \tag{5.12}$$

$$\mathbf{E_q^a} = \left[\mathbf{eq_1^{(a)}}, \ldots \mathbf{eq_j^{(a)}}, \ldots \mathbf{eq_{n_w}^{(a)}}\right] \tag{5.13}$$

$$\omega_\mathbf{Q} = LSTM_{aq}\left(\mathbf{E_q^a}\right) \tag{5.14}$$

The parameters of U_a^Q, U_a^I, U_a^A and $LSTM_{aq}$ are learned during overall model training.

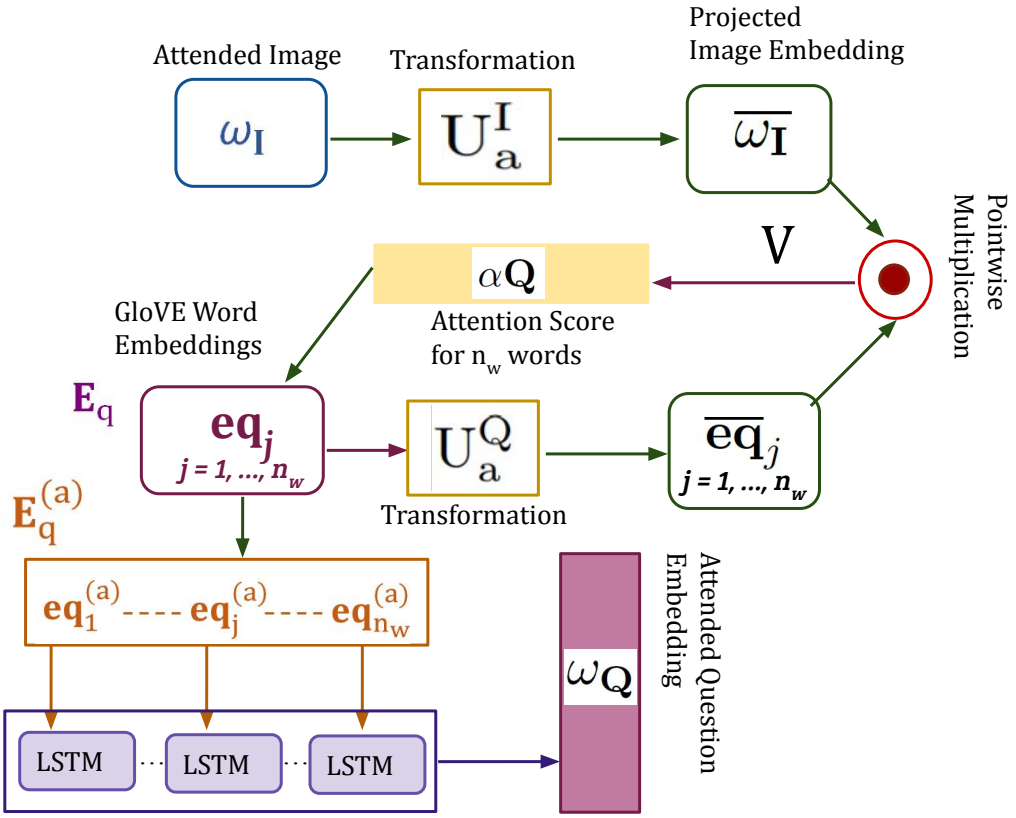


Fig. 5.4 Block diagram for demonstrating the Attention on Question guided by attended visual representation.

5.2.3 Fusion

The attended visual ($\omega_{\mathbf{I}}$) and textual ($\omega_{\mathbf{Q}}$) embeddings are linearly transformed to $\overline{\omega_{\mathbf{I}}} \in \mathbb{R}^{d_f}$ and $\overline{\omega_{\mathbf{Q}}} \in \mathbb{R}^{d_f}$ respectively. The transformed embeddings are fused by element-wise multiplication to obtain the joint multimodal embedding $\omega_{\mathbf{F}}$ (Equations 5.17).

$$\overline{\omega_{\mathbf{I}}} = V_f^I \omega_{\mathbf{I}} \tag{5.15}$$

$$\overline{\omega_{\mathbf{Q}}} = V_f^Q \omega_{\mathbf{Q}} \tag{5.16}$$

$$\omega_{\mathbf{F}} = \overline{\omega_{\mathbf{I}}} \odot \overline{\omega_{\mathbf{Q}}} \tag{5.17}$$

The parameters of the transformation matrices $V_f^I \in \mathbb{R}^{d_f \times d_v}$ and $V_f^Q \in \mathbb{R}^{d_f \times d_q}$ are learned during overall model training. The joint embedding $\omega_{\mathbf{F}}$ is further used for question categorization and answer prediction.

5.2.4 Question Category and Answer Prediction

The question classifier subsystem is a single layered feed forward network $\text{FCNet}_{\text{qcs}}$ with *SoftMax* activation function at its output layer containing $\mid \mathscr{Q}_{\mathscr{C}} \mid = n_{qc}$ nodes. The fused embedding $\omega_{\mathbf{F}}$ is input to $\text{FCNet}_{\text{qcs}}$ to obtain the predicted question category vector $\hat{\mathbf{p}}_{\mathbf{q}} \in (0,1)^{n_{qc}}$ as output. The predicted question category $\hat{\text{q}}_{\text{c}}$ is selected by the winner-take-all strategy (Equations 5.19).

$$\hat{\mathbf{q}}_{\mathbf{pc}} = \text{FCNet}_{\text{qcs}}\left(\ \omega_{\mathbf{F}}\ \right) \tag{5.18}$$

$$\hat{\text{q}}_{\text{c}} = \underset{m=1,\ldots n_{qc}}{\arg\max}\ \hat{\mathbf{q}}_{\mathbf{pc}}[m] \tag{5.19}$$

This proposal has n_{qc} answer prediction subsystems (APS, henceforth). These APS are single layered feed-forward networks $\text{FCNet}_{\text{aps}}^{(\text{m})}$ ($m = 1, \ldots n_{qc}$). Each network has soft-max activation function at the output layer containing $\mid \mathscr{A}_m \mid$ nodes. The fused embedding $\omega_{\mathbf{F}}$ is input to the APS $\text{FCNet}_{\text{qcs}}^{(\hat{\text{q}}_{\text{c}})}$ corresponding to the predicted question category $\hat{\text{q}}_{\text{c}}$. This APS produces the predicted answer category vector $\hat{\mathbf{a}}^{(\hat{\text{q}}_{\text{c}})} \in (0,1)^{|\mathscr{A}_{\hat{\text{q}}_{\text{c}}}|}$. This is considered as the output answer probability vector $\hat{\mathbf{a}}$.

$$\hat{\mathbf{a}}_{\mathbf{p}}^{(\hat{\text{q}}_{\text{c}})} = \text{FCNet}_{\text{aps}}^{(\hat{\text{q}}_{\text{c}})}\left(\ \omega_{\mathbf{F}}\ \right) \tag{5.20}$$

The networks $\text{FCNet}_{\text{qcs}}$ and $\text{FCNet}_{\text{aps}}^{(\text{m})}$ ($m = 1, \ldots n_{qc}$) are trained using their associated loss functions and are described next.

5.2.5 Model Training

The VQA system has the following learnable components – **(a)** $LSTM_Q$ in textual feature extraction subsystem; **(b)** W_a^I, W_a^Q and W_a^A in attention on image subsystem; **(c)** U_a^Q, U_a^I, U_a^A and $LSTM_{aq}$ in attention on question subsystem; **(d)** V_f^I , V_f^Q in fusion subsystem; **(e)** $\text{FCNet}_{\text{qcs}}$ as question categorization subsystem; and **(f)** $\text{FCNet}_{\text{aps}}^{(\text{m})}$ ($m = 1, \ldots n_{qc}$) as answer prediction subsystems. The parameters of these VQA system components are learned in an end-to-end manner using the losses associated with question categorization and answer prediction.

An one-hot-encoded ground-truth question category vector $\tilde{\mathbf{q}}_{\mathbf{gc}} \in \{0,1\}^{n_{qc}}$ can be constructed using the knowledge of ground-truth category label q_c of input question q. The cross-entropy loss $\mathscr{L}_{QCS}$ is formulated as follows (Equation 5.21).

$$\mathscr{L}_{QCS} = -\sum_{m=1}^{n_{qc}} \tilde{\mathbf{q}}_{\mathbf{gc}}[l] \log\left(\hat{\mathbf{q}}_{\mathbf{pc}}[l]\right) \tag{5.21}$$

Let a be the ground-truth answer corresponding to the input image-question pair (I, q). The one-hot-encoded ground-truth answer vector $\tilde{\mathbf{a}}_g^{(m)}$ can be formed for every $\text{FCNet}_{\text{aps}}{}^{(m)}$ ($m = 1, \ldots, n_{qc}$). Note that, all elements of $\tilde{\mathbf{a}}_g^{(m)}$ are set to zero, if a $\notin \mathscr{A}_m$. Let, $\hat{\mathbf{a}}_{\mathbf{p}}^{(m)}$ be the answer-vector predicted by $\text{FCNet}_{\text{aps}}^{(\text{m})}$. The cross-entropy loss $\mathscr{L}_{APS}^{(m)}$ is formulated as follows.

$$\mathscr{L}_{APS}^{(m)} = -\sum_{l=1}^{|\mathscr{A}_m|} \tilde{\mathbf{a}}_{\mathbf{g}}^{(\mathbf{m})}[l] \log\left(\hat{\mathbf{a}}_{\mathbf{p}}^{(m)}[l]\right) \tag{5.22}$$

The total loss $\mathscr{L}_T$ is defined as the sum of the losses associated with the question categorization and answer prediction subsystems and is given by

$$\mathscr{L}_T = \mathscr{L}_{QCS} + \sum_{m=1}^{n_{qc}} \delta[m - \text{q}_\text{c}] \mathscr{L}_{APS}^{(m)} \tag{5.23}$$

The gradient of $\mathscr{L}_T$ is computed and backpropagated to learn the parameters of the DAQC-VQA components.

5.3 Experiment Design

The proposed approach of DAQC-VQA is benchmarked on the *VQA2.0* and *TDIUC* dataset. This section presents the baseline methods (Sub-section 5.3.1) and experimental setup details (Sub-section 5.3.2).

5.3.1 Baseline Methods

The performance of DAQC-VQA is compared against the following baseline methods.

1. *Fusion based Methods* – Fusion of text and image features play an important role in VQA performance. The state-of-art VQA models like MCB [35], MLB [37], MFH [39], MUTAN [13], and BLOCK [40] are chosen as baseline as they primarily contribute towards the fusion of two modalities.
2. *Attention based Methods* – The performance of DAQC-VQA against baseline methods chosen from the following three attention-based categories.
   - *Visual Attention* – SAN [11], BAN [61], BTUP [12], BAN2-CTI [62], CTDA [116], DoG for VQA [115], QTA [16], and QAA [59] are chosen as visual attention based baseline VQA methods.
   - *Co-attention* – MUTAN [13] is chosen as the baseline method under the co-attention category.
   - *Dense attention* – DFAF [14], and MLIN [15] are chosen as dense attention based baseline VQA methods.

All the chosen baseline methods have been discussed in Chapter 2.

5.3.2 Implementation Details

For all experiments (TDIUC & VQA2.0), $n_v = 36$ region proposals are used for visual feature extraction. Each region is represented using ResNet-101 embeddings of $d_v = 2048$ dimensions. Question length is set to $n_w = 14$ words by trimming or padding (as necessary). Dimension of pretrained GloVe word embedding is kept as $d_w = 300$. For obtaining LSTM encoding of question, hidden and output layer dimensions of $LSTM$ are set to 1024, i.e., $d_q = 1024$. All hidden layer dimensions in attention modules are kept as 1024, i.e., $d_{hi} = d_{hq} = 1024$. Dimension of fused embedding is also set to $d_f = 1024$. Number of question categories for TDIUC and VQA2.0 are respectively set to $n_{qc} = 12$ and $n_{qc} = 3$. The model is trained for 17 epochs with a batchsize of 512. The Adamax optimizer [117] is used with a decaying step learning rate. The initial learning rate is set to 0.002 with a decay factor of 0.1 after each 5 epochs. The code of the present implementation is available at - https://github.com/akkkb/DAQC_VQA.

5.4 Results and Discussions

This section presents the performance comparison of the proposed DAQC-VQA system with the baseline models (Sub-section 5.4.1). Basic analysis for dataset size

and error propagation is presented in Sub-section 5.4.1. An ablation analysis is also performed to investigate the importance of some of the components of DAQC-VQA (Sub-section 5.4.1).

Table 5.1 Performance comparison of DAQC-VQA with state-of-the-art in terms of *Category-wise Performance*, *Overall Accuracy*, *Arithmetic-MPT* and *Harmonic-MPT* on TDIUC dataset.

Question Type	**MCB [35]**	**SAN [11]**	**RAU [1]**	**BAN [61]**	**QTA [16]**	**DAQC-VQA**
Scene Recognition	93.06	92.3	93.96	93.1	93.80	**94.18**
Sport Recognition	92.77	95.5	93.47	95.7	**95.55**	95.49
Color Attributes	68.54	60.9	66.86	67.5	60.16	**74.27**
Other Attributes	56.72	46.2	56.49	53.2	54.36	**61.00**
Activity Recognition	52.35	51.4	51.60	54.0	60.10	**60.22**
Positional Reasoning	35.40	27.9	35.26	27.9	34.71	**41.44**
Object Recognition	85.54	87.5	86.11	87.5	86.98	**88.41**
Absurd	96.08	84.82	93.4	96.08	**100.0**	**100.0**
Utility & Affordance	35.09	26.3	31.58	24.0	31.48	**35.67**
Object Presence	93.64	92.4	94.38	95.1	94.55	**95.53**
Counting	51.01	52.1	48.43	53.9	53.25	**57.85**
Sentiment Und.	66.25	53.6	60.09	58.7	64.38	**68.14**
Overall Accuracy	81.86	82.3	84.26	85.5	85.03	**87.84**
Arithmetic-MPT	67.90	65.0	67.81	67.4	69.11	**72.68**
Harmonic-MPT	60.47	53.7	59.00	54.9	60.08	**65.40**

5.4.1 Quantitative Results

Question category wise comparison on TDIUC dataset – Table 5.1 compares the performance of DAQC-VQA on the TDIUC dataset with the other baseline models. Here we have only compared with the models for which question category wise results are available on the TDIUC dataset. The first 12 rows tabulate the class-wise accuracy values for the respective 12 question categories. The last three rows present the performance in terms of *Overall Accuracy*, *Arithmetic-MPT* and *Harmonic-MPT* [1].

It can be observed that DAQC-VQA outperforms all the chosen baseline models on all three evaluation metrics. QTA [16] and MCB [35] demonstrated the best performance among all baseline models in terms of AMPT and HMPT, respectively. DAQC-VQA obtains relative performance improvements of 5.16% and 8.15% compared to QTA and MCB, respectively.

DAQC-VQA also outperforms other methods in multiple category-wise accuracy values. For example, notable performance gains of 8.36% and 17.06% are witnessed for *'Color'* and *'Positional Reasoning'* question categories respectively. It is noteworthy to mention that most of the baseline models (under comparison) exploit complex

Table 5.2 Comparison of *Overall Accuracy* of DAQC-VQA with other state-of-the-art models on TDIUC dataset.

Category	Methods	Overall Accuracy
FUSION	**MLB**[37]	83.10
	MUTAN[13]	82.70
	MFH[39]	84.30
	BLOCK[40]	85.96
VISUAL ATTENTION	**BTUP**[12]	82.91
	QCG[71]	82.05
	RN[114]	84.61
	RAMEN[31]	86.86
CO-ATTENTION	**BAN2-CTI**[62]	87.0
	QAA[59]	84.60
DENSE ATTENTION	**DFAF**[14]	85.55
	MLIN⋆[15]	87.60
CO-ATTENTION	**DAQC-VQA**	**87.84**

Table 5.3 Comparison for VQA 2.0 validation split

Category	Methods	Yes / No	Number	Other	Overall
FUSION	**MCB**[35]	77.37	36.66	51.23	59.14
	MLB[37]	81.89	42.97	53.89	62.98
	MUTAN[13]	81.09	41.87	54.69	62.71
	MFH[39]				61.60
VISUAL ATTENTION	**SAN**[11]	78.40	40.71	54.36	61.70
	RN[114]	80.51	41.92	54.75	62.74
	BTUP[12]	80.34	42.80	55.80	63.20
CO-ATTENTION	**BAN**[61]	–	–	–	66.0
	BAN2-CTI[62]	–	–	–	66.00
	DoG[115]	**82.16**	**45.45**	55.70	64.29
	CTDA[116]	81.26	43.24	55.67	63.65
	QAA[59]	–	–	–	60.5
DENSE ATTENTION	**DFAF**[14]	–	–	–	**66.21**
	MLIN⋆[15]	–	–	–	66.18
CO-ATTENTION	**DAQC-VQA**	82.15	43.57	**56.39**	64.51

and deeper attention networks, while DAQC-VQA employs a comparatively simpler attention network.

Overall performance comparison on TDIUC dataset – Table 5.2 compares overall performance of DAQC-VQA with the other baseline models for which question category-wise results are not available. It can be observed that DAQC-VQA again obtains the best performance across all the baseline models. Its performance with MLIN [15] is comparable. However, MLIN incorporates 100 object region proposals from Faster-RCNN, while DAQC-VQA uses only 36. The DAQC-VQA system has $31M$ trainable parameters for TDIUC dataset.

Comparative analysis on VQA 2.0 – Table 5.3 compares the performance of DAQC-VQA with state-of-the-art models on validation split of VQA2.0 dataset. For this dataset, models employing dense and complex attention mechanisms such as, DFAF [14], MLIN [15], BAN [61] and BAN2-CTI [62] have obtained the best overall performance. However, the performance of DAQC-VQA is comparable with other state-of-the-art models on VQA 2.0 dataset even using significantly lesser number of parameters than those methods. Table 5.5 compares the number of trainable parameters with that of some of the existing methods.

Performance of Question Categorization with ACA-VQA & CSCA-VQA – A new series of experiments were conducted to investigate the impact of question categorization on multistage attention. In order to achieve this, the fused embedding ($\omega_{\mathbf{F}}$) for the question categorization subsystem ($\text{FCNet}_{\text{qcs}}$) and answer prediction subsystem ($\text{FCNet}_{\text{aps}}$) was obtained from the "Aggregated Co-Attention VQA" (ACA-VQA) model, which was introduced in Chapter 3. The ACA-VQA model with three stages, stage-wise loss, and unshared answer predictor parameters was found to be the best-performing model. The embedding was obtained from the highest-performing ACA-VQA model, and the performance on the TDIUC dataset for three evaluation metrics is shown in the first row of Table 5.4. Furthermore, question categorization-based classification was analyzed for the "Cascaded Self- and Co-Attention VQA" (CSCA-VQA) model (Chapter 4). The results for TDIUC with four blocks of attention are presented in the second row of Table 5.4.

Table 5.4 Model performance with multistage attention models ACA-VQA 3 and CSCA-VQA 4 for TDIUC dataset.

Input	Overall Accuracy	Arithmetic-MPT	Harmonic-MPT
ACA-VQA [3]	87.32	70.59	60.20
CSCA-VQA [4]	87.50	70.29	58.16
DAQC-VQA	**87.84**	**72.68**	**65.40**

The results shown in Table 5.4 indicate that the performance of the question categorization subsystem decreases with the use of multistage attention models. This suggests that as the attention mechanism becomes more complex and effective, the categorizer's performance worsens. In contrast, simpler models with a single stage attention mechanism tend to perform well with a question categorizer. Based on this observation, we conducted thorough experiments and analysis on a dual-attention-based single-stage model, called DAQC-VQA, to assess its effectiveness. By focusing on a simpler model with only one stage, it is expected to achieve a better balance between attention and categorization performance.

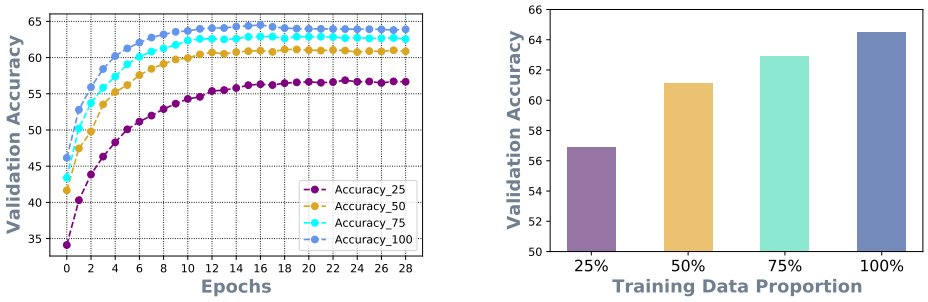
Table 5.5 Computational Complexity in terms of model parameter count for VQA2.0 dataset.

Model	MCB [35]	MLB [37]	MUTAN [13]	MFH [39]	BAN [61]	DAQC-VQA (Ours)
Parameter Count (in Millions)	63	25	62	62	76	34
Accuracy	59.14	62.98	63.61	61.6	66.0	64.51

5.4.2 Basic Analysis

Impact of Training Data Size on Performance – An additional experiment was performed with varying amount of training instances to obtain the learning curves. This experiment aimed to analyse the importance of the amount of training examples in the proposed model. In particular, four datasets were created using 25%, 50%, 75% and 100% of training samples from the VQA2.0 dataset. In each training dataset, the original proportions of class-wise answers were maintained. The experimental results are summarized in Figure 5.5. Figure 5.5a shows the performance on validation set over the epochs during training. It can be seen that the performance curves are similar for all the four cases. However, as expected, the performance improves with increasing training dataset size. The same is reflected in Figure 5.5b in terms of overall accuracy.

Cascading Error Analysis – The proposed model categorizes the input question first and then selects one (through winner-take-all strategy) of several answer classifier sub-systems to predict the answer. Thus, question misclassification in the first stage may lead to wrong answer prediction. The effect of this cascading error is analyzed (using VQA2.0 dataset) by computing the percentages of answers correctly or wrongly predicted corresponding to correct (first row) or incorrect (second row) question categorization. The results of this analysis are shown in Figure 5.6. The accurate answer prediction rate for for wrong question categorization is 0.17%. This low rate correct answer prediction is attributed to the limited number of answers that overlap between different question categories. For example, the question “What material is the necklace made from?” falls under the category of “attribute” while the question “What color is the border of the clock?” falls under the category of “color”. In both cases, the correct answer is “Gold”. It is observed that even with incorrect question categorization, only 0.69% of the answers are wrongly predicted. This is quite low compared to the answer prediction error even with correct question categorization. Thus, in proposed



(a) (b)

Fig. 5.5 Illustrating the learning curves on training datasets formed with different amounts of instances from VQA2.0. (a) Performance on validation set of VQA2.0 with respect to number of epochs. (b) Overall accuracy for VQA2.0 dataset with different proportion of training data.

		Answer Prediction	
		Correct	Incorrect
Question Categorization	Correct	64.51%	34.63%
	Incorrect	0.17%	0.69%

Fig. 5.6 *Question Categorizer – Answer Predictor* cascade error analysis.

model, the cascading error induced by first stage of question categorization is very low.

5.4.3 Ablation Analysis

DAQC-VQA employs a dual attention mechanism to encode question and image representations. These representations of the two modalities can be combined together

Table 5.6 Ablation Analysis I – Comparison of model performance with different variants of input to *Question Classifier* for TDIUC dataset.

Input	Overall Accuracy	Arithmetic-MPT	Harmonic-MPT
Q	87.18	69.36	60.97
Q_A	87.54	70.91	60.30
$I_A \otimes Q$	87.52	72.08	64.45
$I_A \otimes Q_A$	**87.84**	**72.68**	**65.40**

Table 5.7 Ablation Analysis II – Comparison of model performance with different variants of input to *Question Classifier* for VQA2.0 dataset.

Input	Yes / No	Number	Other	Overall Accuracy
Q	81.93	42.74	55.78	63.89
Q_A	81.97	42.09	56.01	63.95
$I_A \otimes Q$	81.99	43.38	56.09	64.15
$I_A \otimes Q_A$	**82.15**	**43.57**	**56.39**	**64.51**

Table 5.8 Ablation Analysis III – Performance Analysis by training *Without 'Absurd'* category.

Input	Overall Accuracy	Arithmetic-MPT	Harmonic-MPT
Q	84.07	68.16	60.29
Q_A	84.13	68.76	59.47
$I_A \otimes Q$	83.46	68.69	**61.44**
$I_A \otimes Q_A$	**84.21**	**69.05**	59.59

Table 5.9 Evaluating model performance on VQA2.0 dataset to investigate the effect of *Dual Attention* and Question Categorization.

DA	QC	Yes / No	Number	Other	Overall Accuracy	Parameter (in Millions)
✗	✗	79.08	40.75	49.96	59.69	15
✗	✓	79.15	41.73	50.10	59.92	18
✓	✗	78.67	40.95	48.30	58.74	28
✓	✓	**82.15**	**43.57**	**56.39**	**64.51**	**34**

in various ways and subsequently provided as input to the question classifier and answer prediction subsystems. An ablation analysis is performed to identify a proper choice of input embedding to question classifier subsystem from the following four variants.

The First and simplest variant considers only the LSTM encoding of Question and is indicated by Q. The second variant Q_A utilizes image to question attention and considers the question representation encoded with attended word embeddings in the context of the image.

Both the third and fourth variants use fusion mechanism. In the third variant ($I_A \otimes Q$), the embedding is obtained by fusion of LSTM encoding of question and image attended question. Fourth variant, indicated by $I_A \otimes Q_A$, is obtained by the fusion of attended image and attended question encoding. The results of this ablation analysis on TDIUC and VQA2.0 datasets are reported in Tables 5.6 and 5.7 respectively.

It is observed that in both cases, the best performances are obtained by using $I_A \otimes Q_A$ as an input embedding to the question classifier subsystem. However, the impact of the choice of input embedding is more prominent in the TDIUC dataset than in VQA2.0.

The effect of language prior is a major issue in VQA where the answer prediction is dictated by the language bias present in training data [1][6] but not by the visual content. This phenomenon motivated the second set of ablation analysis experiments involving the TDIUC dataset. The *Absurd* category of questions (having no relation with the image) helps in identifying the language induced bias. These experiments involve the DAQC-VQA system training *Without Absurd* category. The results indicate

a drop in the overall model performance when trained without the Absurd category (Table 5.8). This implies that model learning with the *Absurd* category indeed helps in reducing language prior bias.

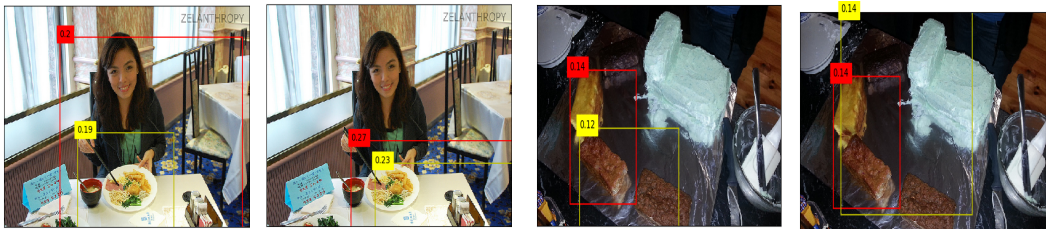
Another set of ablation analysis is performed to investigate the role of our key contributions, namely, *Dual Attention* and *Question Categorization*. Here, experiments are performed with models having different combinations of dual attention and question categorization components and the results are presented in Table 5.9. The performance of primary model neither deploying dual attention nor question categorization is shown in first row. This model simply fuses the features of two modalities and feeds resulting embedding to the classifier for answer prediction. Here, the model performance is relatively poor. The impact of the question categorization component was assessed by performing question categorization before passing them to the answer predictor networks. This was done instead of directly feeding the fused embeddings to the answer classifier network. An improvement of 1.59% is obtained in overall performance, thereby showing the efficacy of question categorization module (second row). Another experiment (third row) presents the role of dual attention (only) in the model. A fusion based approach is not able to capture the interaction of two modalities. On the other hand, dual attention provides the features that capture the interaction of correlated elements of both question and image in a better way. The incorporation of dual attention gives a significant improvement in the performance. Last row of table demonstrates the model with both question categorization and dual attention incorporated in the system. This combination outperforms all the models in terms of overall accuracy as well as individual class-wise accuracies.

5.4.4 Qualitative Results

The qualitative evaluation of the DAQC-VQA system is performed through the visualization of results (Figure 5.7). The results show the top-2 salient regions and their attention scores generated by a baseline approach and the *AoI* module of the DAQC-VQA system. Here, the visual attention based bottom-up top-down model [12] is used as baseline. The baseline uses $I_A \otimes Q$ fused embedding for answer classification. These results are demonstrated in figure 5.7 on TDIUC dataset for different categories like *object recognition*, *activity recognition*, *object presence* and *others*.

In Figure 5.7a, the visual attention of baseline model focuses on a region that captures the girl with a higher score. DAQC-VQA has top-2 scores corresponding to the region that includes the girl's hands and the food. Similarly, for 5.7c, DAQC-VQA focuses on the complete picture of food to infer the answer.

However, DAQC-VQA also gets confused for some samples, leading to wrong inference. For example, in Figure 5.8a, DAQC-VQA assigns highest score to a *knife*

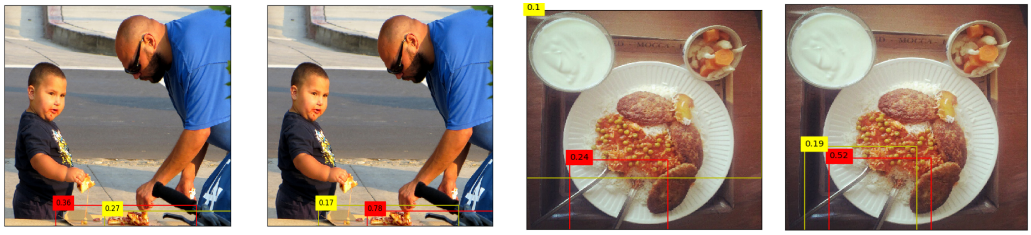


(a) **Q.** What is the girl doing?
Ans: Sitting ✗ (0.2, 0.19)
GT: Eating

(b) **Q.** What is the girl doing?
Ans: Eating ✔ (0.27, 0.23)
GT: Eating

(c) **Q.** What food is shown in the photo?
Ans: Sandwich✗(0.14, 0.12)
GT: Cake

(d) **Q.** What food is shown in the photo?
Ans: Cake ✔ (0.14, 0.14)
GT: Cake



(e) **Q.** What food is shown in the picture?
Ans: Cake ✗ (0.36, 0.27)
GT: Pizza

(f) **Q.** What food is shown in the picture?
Ans: Pizza ✔ (0.78, 0.17)
GT: Pizza

(g) **Q.** Are there any knives in the photo?
Ans: Yes ✗ (0.24, 0.1)
GT: No

(h) **Q.** Are there any knives in the photo?
Ans: No ✔ (0.52, 0.19)
GT: No

Fig. 5.7 Qualitative results were obtained from DAQC-VQA. The salient regions corresponding to the Top-2 attention scores are presented (top1, top2). The left image corresponds to the baseline model (without question classification), and the right image presents the DAQC-VQA model.

(a) **Q.** What food is in the photo ?
Ans: Carrot ✗ (0.44, 0.19)
GT: Cake

(b) **Q.** What vehicle is in the picture ?
Ans: Truck ✗ (0.75, 0.09)
GT: Car

Fig. 5.8 Poor performance cases of DAQC-VQA due to failure in capturing relevant relations.

that resembles a *carrot*. Thus, the answer to the question of inferring the food is wrongly predicted as *carrot*. Similarly, in Figure. 5.8b the question is on identifying a vehicle in the scene. However, DAQC-VQA focuses on a *truck* while the ground-truth answer is *car*.

5.5 Discussions

This chapter introduced a novel approach to Visual Question Answering by using Dual Attention and Question Categorization (DAQC-VQA). The dual attention mechanism is used to extract richer feature representations for both image and text modalities by allowing cross-modal interactions. This means that the model attends to both the image and the question at the same time, allowing for more informative representations to be extracted. The question categorizer was introduced further to reduce the search space for answer prediction. By categorizing the input question, the system can identify the relevant information and focus on answering only the necessary components of the question. The proposed DAQC-VQA was realized through an end-to-end trainable system with a joint loss function. This means that the entire model can be trained in a single step, while improving the overall efficiency of the VQA system. The DAQC-VQA is validated on two benchmark VQA datasets (TDIUC and VQA2.0) and compared against several state-of-the-art approaches. The quantitative and qualitative evaluations demonstrated the competitive and often better performance of DAQC-VQA compared to the baseline models. Also, it was observed that the question categorizer performed better with simpler single stage dual attention model compared to the ones equipped with multiple stages of dense attention mechanism.

Chapter 6

Conclusions and Future Work

This thesis has proposed VQA methods that improve the attention module and thus extract richer representation for multimodal features. These improved feature representations are then fused and fed to perform the task of answer classification. Extensive experiments and analysis are performed on two widely used publicly available VQA datasets i.e., TDIUC and VQA2.0.

There are three main contributions of the thesis. In the first contribution (presented in Chapter 3), a multistage co-attention based VQA (ACA) model is proposed. Attention on visual and textual modality is applied in alternate manner in context of each other. *Attention aggregation* is performed to preserve the attention for corresponding modality from each stage. Results and analysis demonstrated the effectiveness of utilizing multistage attention, aggregation and stage-wise loss in the model.

Chapter 4 presents the second contribution, where the self-attention mechanism based architecture (CSCA) is proposed along with cross-modality attention to encode the fine grained information from dual modality. It was observed that cascading of self- and co-attention blocks multiple times helped in facilitating model's ability to extract informative features from the input data.

Finally, in Chapter 5, a novel architecture (DAQC) is presented that solves the VQA problem by splitting it into two sub-problems. First, the input question category is identified and accordingly an answer prediction sub-system is activated. This splitting helps in reducing the answer search space for final answer classification. Along-with this a *dual attention* mechanism is proposed to obtain better feature representation of visual and textual features. Both qualitative and quantitative results on TDIUC and VQA2.0 datasets demonstrated the competitive and often better performance of DAQC-VQA compared to the baseline models.

Table 6.1 presents the performance of the three models, namely ACA-VQA, CSCA-VQA, and DAQC-VQA, on the *VQA2.0* and *TDIUC* datasets. Additionally, the table

shows the parameter count for each model. The final version of the ACA-VQA model for *VQA2.0* consists of *two* stages, while for *TDIUC*, it has *three* stages. Each stage has its own loss, shared linear transformation parameters, and unshared answer predictors. The final CSCA model for both datasets has *four* stages and performs the best among all three contributed models, outperforming the others on two datasets.

Table 6.1 ACA-VQA, CSCA-VQA and DAQC-VQA model performance for VQA2.0 and TDIUC dataset

Methods	**VQA2.0**		**TDIUC**			
	Overall Accuracy	**Param Count (in Millions)**	**A-MPT**	**H-MPT**	**Overall Accuracy**	**Param Count (in Millions)**
ACA-VQA [3]	64.95	32	72.40	66.10	86.82	30
CSCA-VQA [4]	**67.36**	43	**73.34**	**67.05**	**88.12**	29
DAQC-VQA [5]	64.51	34	72.68	65.40	87.84	31

Based on the results obtained across the models, the thesis makes the following conclusions:

- Co-attention applied in multiple stages helps in obtaining improved feature representation.
- Co-attention coupled with self-attention further helps in refining feature representation which in turn leads to improved performance for the VQA task.
- Finally, a relatively simpler model relying on partitioned answer spaces using question-category information and a single stage co-attention can also achieve significant performance.

6.1 Scope of Thesis

In this section a discussion is presented to encompass the evolving research landscape, which notably includes the dynamic realm of LLMs and and VLMs.

This thesis takes on a crucial role by focusing on the enhancement of Visual Question Answering (VQA) as a classification task. Leveraging attention mechanisms across dual modalities, this work contributes innovative approaches to VQA, aiming to augment the performance of existing models. The implications of this thesis extend beyond the realm of VQA. Attention mechanisms serve as a powerful tool for modeling interactions between visual and textual data. The novel attention mechanisms introduced in this thesis hold the potential for adaptation in a wide range of cross-modal tasks, including image captioning, content-based image retrieval, and autonomous navigation systems. Additionally, the thesis addresses the nuances of specific question types, such as recognition, counting, and sentiment understanding, shedding light on

the challenges posed by these questions. The methodologies developed here can serve as valuable reference points for researchers seeking to improve the performance of VLMs in handling diverse and complex queries.

VLMs [124][125][126][127][128][129] are known for their impressive performance for multimodal tasks, but they come with significant complexity as well as are trained on billions of data samples. It requires huge computational resources to train these models. As the current research direction leans towards developing LLMs and VLMs, these models have grown substantially in size and complexity, which poses a challenge when it comes to deploying them on edge devices due to their high computational demands.

To address this challenge, the methods proposed in this thesis can be leveraged to transfer the knowledge embedded in VLMs to smaller, more lightweight models. This transfer of knowledge has the potential to significantly enhance the efficiency of incorporating VQA models into real-world applications. By doing so, we can bridge the gap between the powerful capabilities of VLMs and the practical constraints of deploying models on resource-constrained edge devices, making advanced VQA systems more accessible and usable in a wide range of real-life scenarios.

6.2 Potential Future Research Works

Within the scope of this thesis, the goal revolves around solving the VQA as a classification task. To this end, significant contributions are made through the introduction of various attention mechanisms that operate across dual modalities. Further these mechanisms are unified with VQA pipeline in a way that they result in a better architecture leading to competitive performance with similar class of state of the art methods.

However, the proposed approaches faced certain challenges and have limitations. Following discussion highlights some of the prominent challenges, that can be addressed in the future work.

- The methodologies put forth in the thesis exhibit performance limitations when confronted with questions that necessitate advanced *reasoning capabilities*. The current methods lack the explicit means to engage in relation reasoning, which, if incorporated, has the potential to significantly augment the efficiency and accuracy of the models. This is particularly required in scenarios where questions demand complex relational understanding, such as identifying intricate interactions between entities or attributes.
- One drawback of the current model is its *susceptibility to performance degradation in the presence of noisy input data*. The absence of mechanisms to

effectively handle noisy data impedes the model's ability to maintain consistent performance levels across various real-world scenarios. So, improving the model's ability to deal with noisy information is a crucial way to make it better. By equipping the model with robust preprocessing techniques, noise detection algorithms, or adaptive learning strategies, its capacity to filter out and mitigate the impact of noisy inputs can be significantly bolstered.

- Another scope for improvement could involve the *incorporation of contextual embeddings and domain-specific knowledge to enhance the model's understanding and performance*. The current model might lack the capacity to effectively leverage contextual cues and domain-specific information present in the questions, answers, and images. By integrating pre-trained contextual embeddings, such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer), the model can better grasp the nuanced semantics of the text and images. These embeddings capture contextual relationships between words and phrases, enabling the model to better understand the implicit connections in complex questions and answers.
- Harnessing the power of VLMs: This works did not exploit VLMs. Future works can leverage VLMs to make smaller and lightweight models. That will significantly enhance the efficiency of incorporating VQA models into real-world applications.

In addition to the above-mentioned future work directions, the following problems can also be considered. These problems were not particularly addressed in this work.

1. **Multilingual Datasets in Indian Context:** Existing and widely available datasets for VQA are in English language. Also, the associated images are mostly cast in the western context. To make it more relevant for real life purpose in Indian context, it would be better to develop multilingual datatsets with images cast in Indian context.

2. **Model Compression for Edge Devices:** For real life applications, VQA models should be compressed so that it could be deployed on edge devices like smartphones, tablets or other low power embedded processors. Accordingly, the models need to be compressed to enable them to infer on such devices.

3. **Multimodal Dialog:** One round of communication may not suffice the requirements of real life applications. Accordingly, datasets and models should be developed to enable text or multimodal dialog around images and videos.

4. **Unseen Data Inference:** Availability of labeled data is always a concern for ML tasks. Models should be proposed that could perform well even there is unavailability of annotated data in hand. To develop the VQA model that is capable to infer answer for unseen objects or question could make VQA ready for more general purpose.

List of Publications

Manuscripts Published

1. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, CQ-VQA: Visual Question Answering on Categorized Questions. In Proceedings of International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–8.

2. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, Multistage Attention based Visual Question Answering. In Proceedings of 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 9407–9414.

3. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, Dual Attention and Question Categorization based Visual Question Answering. IEEE Transaction on Artificial Intelligence (TAI), vol. 4(1), pp. 81–91, 2022.

4. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, ACA-VQA: Aggregated Co-attention based Visual Question Answering [Accepted for Publication at ICVGIP, 2023]]

Manuscripts Under Review

1. **Aakansha Mishra**, Ashish Anand and Prithwijit Guha, CSCA: VQA with Cascade of Self- and Co-Attention Blocks. [Under Review]

References

[1] K. Kafle and C. Kanan, "An Analysis of Visual Question Answering Algorithms," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1965–1973.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[3] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[5] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the Role of Image Understanding in Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[7] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[8] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 804–813.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[13] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal Tucker fusion for Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2612–2620.

[14] P. Gao, Z. Jiang, H. You, P. Lu, S. C. Hoi, X. Wang, and H. Li, "Dynamic Fusion With Intra-and Inter-Modality Attention Flow for Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6639–6648.

[15] P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality Latent Interaction Network for Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5825–5835.

[16] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question Type Guided Attention in Visual Question Answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 151–166.

[17] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual Question Answering Model based on Visual Relationship Detection," *Signal Processing: Image Communication*, vol. 80, p. 115648, 2020.

[18] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and Concept-driven Analysis for Image Caption Generation," *Neurocomputing*, vol. 398, pp. 520–530, 2020.

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

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] L. Ma, Z. Lu, and H. Li, "Learning to Answer Questions from Image using Convolutional Neural Network," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[22] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," in *Advances in Neural Information Processing Systems*, 2015, pp. 2296–2304.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[24] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1–9.

[25] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," *arXiv preprint arXiv:1802.05766*, 2018.

[26] C. Wu, J. Liu, X. Wang, and X. Dong, "Object-difference attention: A simple relational attention for visual question answering," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 519–527.

[27] A. Jabri, A. Joulin, and L. Van Der Maaten, "Revisiting Visual Question Answering Baselines," in *European Conference on Computer Vision*. Springer, 2016, pp. 727–739.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[29] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] R. Shrestha, K. Kafle, and C. Kanan, "Answer Them All! Toward Universal Visual Question Answering Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 472–10 481.

[32] M. Ren, R. Kiros, and R. Zemel, "Exploring Models and Data for Image Question Answering," in *Advances in Neural Information Processing Systems*, 2015, pp. 2953–2961.

[33] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask Your Neurons: A Neural-based Approach to Answering Questions bout Images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1–9.

[34] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S. C. Hoi, and X. Wang, "Question-guided Hybrid Convolution for Visual Question Answering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–485.

[35] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 457–468.

[36] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact Bilinear Pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.

[37] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard Product for Low-rank Bilinear Pooling," *arXiv preprint arXiv:1610.04325*, 2016.

[38] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1821–1830.

[39] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond Bilinear: Generalized Multimodal Factorized High-order Pooling for Visual Question Answering," *IEEE Transactions on Neural Networks and Learning Systems*, no. 99, pp. 1–13, 2018.

[40] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8102–8109.

[41] Z. Fang, J. Liu, X. Liu, Q. Tang, Y. Li, and H. Lu, "BTDP: Toward sparse fusion with block term decomposition pooling for Visual Question Answering," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 2s, pp. 1–21, 2019.

[42] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms—part i: Lemmas for partitioned matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1022–1032, 2008.

[43] L. De Lathauwer , "Decompositions of a higher-order tensor in block terms—part ii: Definitions and uniqueness," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1033–1066, 2008.

[44] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms—part iii: Alternating least squares algorithms," *SIAM journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1067–1083, 2008.

[45] W. Zhang, J. Yu, W. Zhao, and C. Ran, "Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation," *Information Fusion*, vol. 72, pp. 70–79, 2021.

[46] M. Lao, Y. Guo, N. Pu, W. Chen, Y. Liu, and M. S. Lew, "Multi-stage hybrid embedding fusion network for visual question answering," *Neurocomputing*, vol. 423, pp. 541–550, 2021.

[47] K. J. Shih, S. Singh, and D. Hoiem, "Where to Look: Focus Regions for Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4613–4621.

[48] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *arXiv preprint arXiv:1704.03162*, 2017.

[49] H. Xu and K. Saenko, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 451–466.

[50] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7680–7688.

[51] Y. Lin, Z. Pang, D. Wang, and Y. Zhuang, "Feature enhancement in attention for visual question answering." in *IJCAI*, 2018, pp. 4216–4222.

[52] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[53] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[54] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-attention for Visual Question Answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[55] L. Gao, L. Cao, X. Xu, J. Shao, and J. Song, "Question-led object attention for visual question answering," *Neurocomputing*, vol. 391, pp. 227–233, 2020.

[56] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, and P. Lu, "Multimodal feature-wise co-attention method for visual question answering," *Information Fusion*, vol. 73, pp. 1–10, 2021.

[57] C. Wu, J. Liu, X. Wang, and X. Dong, "Chain of Reasoning for Visual Question Answering," *Advances in Neural Information Processing Systems*, vol. 31, pp. 275–285, 2018.

[58] Q. Sun, B. Xie, and Y. Fu, "Second order enhanced multi-glimpse attention in visual question answering," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[59] M. Farazi, S. Khan, and N. Barnes, "Question-Agnostic Attention for Visual Question Answering," in *25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3542–3549.

[60] H. Noh and B. Han, "Training Recurrent Answering Units with Joint Loss Minimization for VQA," *arXiv preprint arXiv:1606.03647*, 2016.

[61] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear Attention Networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 1564–1574.

[62] T. Do, T.-T. Do, H. Tran, E. Tjiputra, and Q. D. Tran, "Compact Trilinear Interaction for Visual Question Answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 392–401.

[63] P. Huang, J. Huang, Y. Guo, M. Qiao, and Y. Zhu, "Multi-grained attention with object-level grounding for visual question answering," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3595–3600.

[64] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6087–6096.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[66] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[67] Y. Liu, X. Zhang, Q. Zhang, C. Li, F. Huang, X. Tang, and Z. Li, "Dual self-attention with Co-attention networks for Visual Question Answering," *Pattern Recognition*, vol. 117, p. 107956, 2021.

[68] H. Tan and M. Bansal, "LXMERT: Learning Cross-modality Encoder Representations from Transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[69] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *arXiv preprint arXiv:1908.02265*, 2019.

[70] M. Narasimhan, S. Lazebnik, and A. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[71] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning Conditioned Graph Structures for Interpretable Visual Question Answering," in *Advances in Neural Information Processing Systems*, 2018, pp. 8334–8343.

[72] P. Xiong, H. Zhan, X. Wang, B. Sinha, and Y. Wu, "Visual query answering by entity-attribute graph matching and reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8357–8366.

[73] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, "Murel: Multimodal relational reasoning for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1989–1998.

[74] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for Visual Question Answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 313–10 322.

[75] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[76] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web." in *International Joint Conference on Artificial Intelligence*, vol. 7, 2007, pp. 2670–2676.

[77] O. Etzioni, A. Fader, J. Christensen, S. Soderland *et al.*, "Open information extraction: The second generation," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[78] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[79] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.

[80] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[81] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "Yago2: A spatially and temporally enhanced knowledge base from wikipedia," *Artificial Intelligence*, vol. 194, pp. 28–61, 2013.

[82] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[83] N. Tandon, G. De Melo, F. Suchanek, and G. Weikum, "Webchild: Harvesting and organizing commonsense knowledge from the web," in *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014, pp. 523–532.

[84] N. Tandon, G. De Melo, and G. Weikum, "Acquiring comparative commonsense knowledge from the web," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[85] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[86] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv preprint arXiv:1511.02570*, 2015.

[87] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4622–4630.

[88] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2413–2427, 2018.

[89] Q. Wu, C. Shen, P. Wang, A. Dick, and A. v. d. Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367–1381, 2018.

[90] Y. Zhu, J. J. Lim, and L. Fei-Fei, "Knowledge acquisition for visual question answering via iterative querying," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1154–1163.

[91] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 451–468.

[92] D. Song, S. Ma, Z. Sun, S. Yang, and L. Liao, "Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning," *Knowledge-Based Systems*, vol. 230, p. 107408, 2021.

[93] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu, "Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 1097–1103.

[94] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," *Pattern Recognition*, vol. 108, p. 107563, 2020.

[95] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "KAT: A knowledge augmented transformer for vision-and-language," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jul. 2022.

[96] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[97] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "VQA: Visual Question Answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, May 2017.

[98] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4971–4980.

[99] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1682–1690.

[100] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[101] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[102] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[103] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2901–2910.

[104] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A Visual Question Answering benchmark requiring External Knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3195–3204.

[105] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.

[106] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Visual question answering: A survey of methods and datasets," *Computer Vision and Image Understanding*, vol. 163, pp. 21–40, 2017.

[107] A. A. Yusuf, F. Chong, and M. Xianling, "An analysis of graph convolutional networks and recent datasets for visual question answering," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6277–6300, 2022.

[108] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[109] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[110] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Making visual representations matter in vision-language models," *CVPR 2021*, 2021.

[111] D. Gao, R. Wang, S. Shan, and X. Chen, "Learning to Recognize Visual Concepts for Visual Question Answering With Structural Label Space," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 494–505, 2020.

[112] A. Mishra, A. Anand, and P. Guha, "Multi-stage Attention based Visual Question Answering," in *25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9407–9414.

[113] M. Farazi, S. Khan, and N. Barnes, "Question-Agnostic Attention for Visual Question Answering," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 3542–3549.

[114] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A Simple Neural Network Module for Relational Reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4967–4976.

[115] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In Defense of Grid Features for Visual Question Answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.

[116] W. Tian, R. Zhou, and Z. Zhao, "Cascading Top-Down Attention for Visual Question Answering," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[117] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[118] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.

[119] Y. Liu, X. Zhang, F. Huang, L. Cheng, and Z. Li, "Adversarial learning with multi-modal attention for visual question answering," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[120] X. Zhu, Z. Mao, Z. Chen, Y. Li, Z. Wang, and B. Wang, "Object-difference drived graph convolutional networks for visual question answering," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16 247–16 265, 2021.

[121] L. Peng, Y. Yang, Z. Wang, X. Wu, and Z. Huang, "CRA-Net: Composed Relation Attention Network for Visual Question Answering," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1202–1210.

[122] H. Nam, J.-W. Ha, and J. Kim, "Dual Attention Networks for Multimodal Reasoning and Matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.

[123] A. Mishra, A. Anand, and P. Guha, "CQ-VQA: Visual Question Answering on Categorized Questions," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[124] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[125] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[126] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[127] M. Shukor, C. Dancette, A. Rame, and M. Cord, "Unified model for image, video, audio and language tasks," *arXiv preprint arXiv:2307.16184*, 2023.

[128] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[129] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.