Dual Attention and Question Categorization-Based Visual Question Answering

Aakansha Mishra, Ashish Anand, and Prithwijit Guha, *Member, IEEE*

***Abstract*—Visual question answering (VQA) aims at predicting an answer to a natural language question associated with an image. This work focuses on two important issues pertaining to VQA, which is a complex multimodal AI task: First, the task of answer prediction in a large output answer space, and second, to obtain enriched representation through cross-modality interactions. This work aims to address these two issues by proposing a dual attention (DA) and question categorization (QC)-based visual question answering model (DAQC-VQA). DAQC-VQA has three main network modules: First, a novel dual attention mechanism that helps toward the objective of obtaining an enriched cross-domain representation of the two modalities; second, a question classifier subsystem for identifying input (natural language) question category. The second module of question categorizer helps in reducing the answer search space; and third, a subsystem for predicting 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. The performance of DAQC-VQA is evaluated on two widely used VQA datasets, viz., TDIUC and VQA2.0. Experimental results demonstrate competitive performance of DAQC-VQA against the recent state-of-the-art VQA models. An ablation analysis indicates that the enriched representation obtained using the proposed dual-attention mechanism helps improve performance.**

***Impact Statement*—Visual Question Answering (VQA) is a challenging multi-modal Artificial Intelligence task involving computer vision, natural language processing and commonsense reasoning. It has vast potential applications for several human-computer interaction tasks, including assisting visually impaired individuals, AI-based personal assistants etc. Furthermore, it is also considered an AI-complete task. This work aims to enhance the performance of VQA models by overcoming two challenges – cross-domain interaction and reasoning in large answer space. This work proposes a VQA model consisting of a Dual Attention mechanism and Question Categorizer. The dual attention mechanism allows the VQA model to obtain improved cross-domain (image and text-domains) semantic representation. Furthermore, question type identification for answer space reduction coupled with a dual attention mechanism improved or obtained competitive performance compared to state-of-art models on two VQA datasets.**

***Index Terms*—Attention networks, classification networks, dual attention, multimodal fusion, visual question answering.**

Manuscript received 4 October 2021; revised 21 January 2022; accepted 11 March 2022. Date of publication 17 March 2022; date of current version 20 January 2023. This work was supported by the Department of Biotechnology, Govt. of India under Grant BT/COE/34/SP28408/2018. This article was recommended for publication by Associate Editor Sanghamitra Bandyopadhyay upon evaluation of the reviewers' comments. *(Corresponding author: Aakansha Mishra.)*

Aakansha Mishra and Ashish Anand are with the Department of Computer Science & Engineering, Guwahati 781039, India (e-mail: ak.kkb@iitg.ac.in).

Prithwijit Guha is with the Department of Electronics and Electrical Engineering, Guwahati 781039, India, and also with the Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Guwahati, Guwahati 781039, India (e-mail: anand.ashish@iitg.ac.in; pguha@iitg.ac.in).

Digital Object Identifier 10.1109/TAI.2022.3160418

I. Introduction

VISUAL question answering (VQA) [1], [2] systems aim to predict or synthesize an answer to a natural language question related to a given image. VQA systems have received wide attention due to their applicability in various real-life scenarios such as interactive robotic systems, smart home management, and personal assistance. Last few years have witnessed increasing research interest in VQA, leading to significant progress in the performance of proposed models. However, the intrinsic multimodality and necessity of reasoning using both textual and visual features continue to pose challenges to VQA models.

Attention mechanism has played a significant role in improving the VQA model performance. Initial VQA models adopting attention mechanism [2]–[11] focused on finding relevant regions in image pertaining to the given question. Attention on image (AoI) regions using textual features has become a default component of VQA models. However, recent studies [12]–[15] have indicated that image-conditioned attention on question (AoQ) further helps models to 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 [9], [16] 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 behavior 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 article proposes the dual attention and question categorization-based visual question answering system (DAQC-VQA). DAQC-VQA combines subsystems for *dual attention*, *question categorization*, and *answer prediction*. Fig. 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.

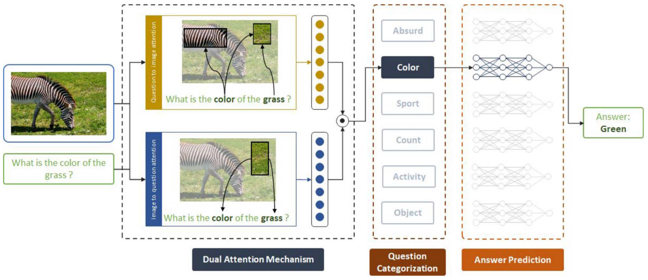

What is the color of the grass ?
Question to image attention
What is the color of the grass ?
Image to question attention
What is the color of the grass ?
Absurd
Color
Sport
Count
Activity
Object
Answer:
Green
Dual Attention Mechanism
Question Categorization
Answer Prediction

Fig. 1. An overview of the proposed DAQC-VQA system.

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

- *Dual Attention*—AoI and 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.
- *End-to-end formulation*—An end-to-end trainable model and associated loss functions for the seamless combination of the *dual-attention mechanism* and the *question categorizer*.
- *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.

II. Related Work

The existing and related works in VQA can be broadly divided into two major categories based on their primary themes. These are *fusion* schemes and *attention*-based models. Attention-based VQA models do use one of the fusion methods to obtain joint representation of the image and text modalities.

A. Fusion Schemes

The fusion schemes can be subdivided into two major categories—*Simple* and *Bilinear* fusion schemes. The simple fusion schemes involve elementwise addition [3], [13], [17], elementwise multiplication [1], [4], and concatenation [9], [18] of features of the two modalities. These methods are relatively computationally inexpensive. However, they fail to capture complex interaction between the two modalities [19].

Bilinear fusion schemes try to obtain expressive representation of the two modalities by capturing complex interaction between them. The complex interaction of the two modalities is captured by allowing interaction of each element of one modality with the other. Fukui et al. [19] introduced the first bilinear fusion model MCB by using outer product-based interaction between the elements of the two modalities. Further, to avoid the high computational complexity of outer product operation, MCB uses convolution operation to approximate outer product [20]. MCB outperforms the simple fusion mechanisms at the cost of increased computation and resource requirements.

Subsequent to MCB, several methods were proposed to take care of the two primary issues of the bilinear fusion schemes. These are high-dimensional feature representations, and high computational cost. Kim et al. (MLB, [21]) used the Hadamard product of the two modalities represented by two low-rank projection matrices. Although MLB managed to reduce the computational cost along with a low-dimensional representation, empirical results indicated that it converges slowly.

Fusion schemes presented in MFB [22], MUTAN [6], and BLOCK [23] also obtain low-dimensional representations having similar robust, expressive representations as in MCB and also have faster convergence.

B. Attention-Based Models

To answer a question related to an image, one should focus on image regions that are relevant to the question. Attention models help in identifying such regions. Existing works in VQA have used three types of attention strategies, namely, *Visual attention*, *Co-attention*, and *Dense attention*.

VQA models with *visual attention* mechanism focus on image based on the given question. The first such method was proposed by Shih et al. [18] and was used by several other VQA models like [3], [4], [7], [8], [24], [25], etc. Initial models focused on attention on grids of feature maps obtained from pretrained convolutional neural network (CNN) [18], [19], [21], [26]. Each grid region is assigned a weight based on its relevance to the question. Yang et al. [3] introduced stacked attention for VQA. It was observed that multiple iterations over image help identifying relevant image regions in the context of the given question. Bottom-up top-down attention mechanism [4] applies attention on most relevant image regions (top-k) instead of each image grid region. The top-k image regions are extracted from a pretrained faster-RCNN [27]. Ding et al. [11] have proposed two attention mechanisms, namely, *stimulus-driven* and *concept-driven*, which are inspired from the human psychology for image caption generation task. Xi et al. [10] have introduced a VQA model based on multiobjective visual relationship detection, where relevant image regions are extracted from question-guided attention. Further, 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. Farazi et al. [28] have 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. Question conditioned graph (QCG) [29] is a network-based method for VQA. Nodes of graph correspond to salient image regions, and edges represent the interaction of regions in the context of the given question.

Coattention mechanism includes attention on text features in context of image along with visual attention [12]. Kim et al. [24] proposed a multimodality attention mechanism based on bilinear interaction between the two modalities. Do et al. [25] have proposed the attention mechanism comprising trilinear interaction of image, question, and answer. As answers are not available during the test phase, knowledge distillation is used to transfer knowledge from the trilinear model to the

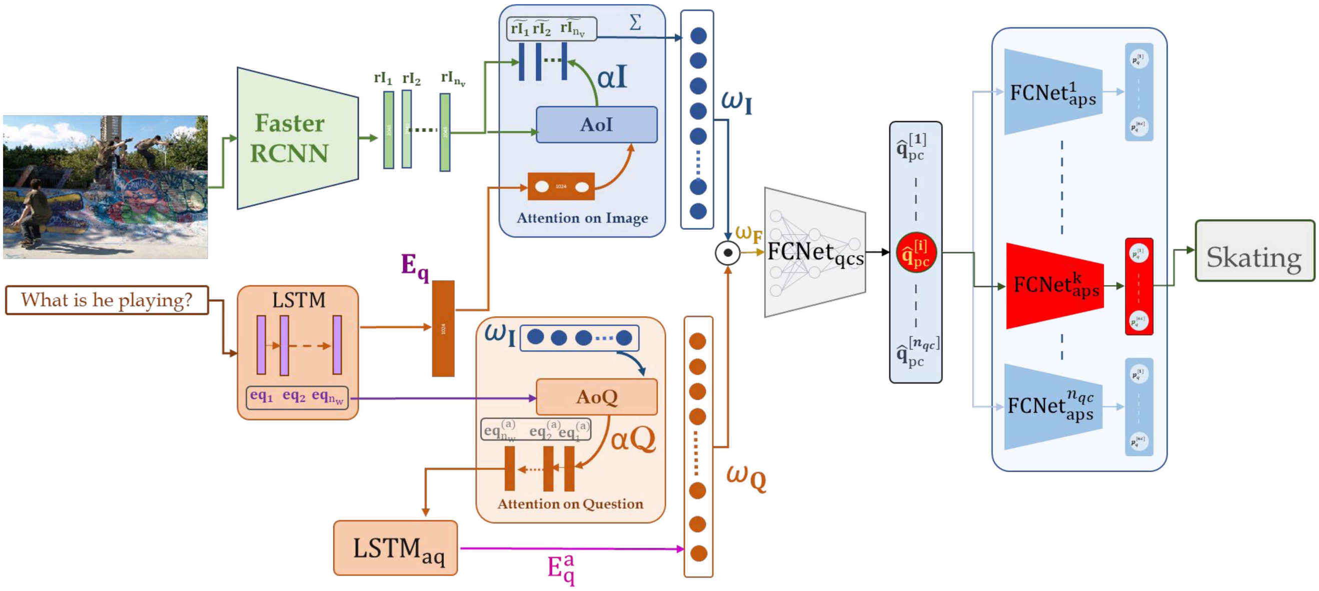

Faster RCNN
rI_1 rI_2 rI_{n_v}
Σ
αI
AoI
Attention on Image
ω_I
ω_F
$\text{FCNet}_{\text{qcs}}$
$\hat{q}_{pc}^{[1]}$
$\hat{q}_{pc}^{[i]}$
$\hat{q}_{pc}^{[n_{qc}]}$
$\text{FCNet}_{\text{aps}}^{1}$
$\text{FCNet}_{\text{aps}}^{k}$
$\text{FCNet}_{\text{aps}}^{n_{qc}}$
Skating
What is he playing?
LSTM
eq_1 eq_2 eq_{n_w}
E_q
ω_I
AoQ
αQ
Attention on Question
LSTM_{aq}
E_q^a
ω_Q

Fig. 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 other modality. AoI represents the Attention-on-Image module based on LSTM based question encoding. AoQ shows the Attention-on-Question module guided by attended visual representation. $\text{FCNet}_{\text{qcs}}$ is the question category classifier. And, $\text{FCNet}_{\text{aps}}^{(l)}$ represents the l^{th} answer prediction sub-module, which performs the final answer prediction.

bilinear model. The coattention mechanism is also referred as dual attention in the literature [14], [15], [30]. Xu and Saenko [30] have proposed dual-attention network-based VQA model. It selects the salient visual regions and relevant question words as features for two modalities. These features are fused by a recurrent network for answer classification. Mishra et al. [15] have proposed a dual-attention-based VQA model that puts attention on two modalities at multiple stages. Tian et al. [31] exploit the coattention mechanism in a cascaded manner to produce better feature representation of two modalities. Shrestha et al. [32] have proposed a coattention-based unified model (RAMEN) for the VQA task that performs well on datasets comprising images from different domains (synthetic and real-world images).

Dense-attention is a complex attention mechanism that works with both intra- and intermodality attention. Gao et al. [7] have combined intra- and intermodalities attention and built a dense attention model to capture the interaction of the two modalities robustly. Gao et al. [8] have proposed a multimodal latent interaction (MLIN) module for modeling the interaction between the two modalities and within each modality in latent space.

Question category information is still underutilized for the VQA task. Question type-guided attention (QTA) [9] is one of the methods that exploits semantic information of question category for visual attention. CQ-VQA [16] also exploits question type for redirecting the answer classifier module to focus on a limited answer space. However, CQ-VQA only uses a visual attention mechanism.

In this article, DAQC-VQA extends the idea of CQ-VQA by using the dual-attention mechanism. In contrast to other similar VQA models, DAQC-VQA uses a top-down attention mechanism along with an elementwise multiplication-based fusion scheme.

III. Proposed Method

The VQA system outputs an answer $\hat{a} \in \mathcal{A}$ in response to an input image $I \in \mathcal{I}$ and an associated natural language question $q \in \mathcal{Q}$. Following recent works [1], [4], [5], [33], this proposal formulates the VQA task as a classification problem. Here, the answer $\hat{a} \in \mathcal{A}$ is predicted using features of the inputs (I, q).

A pretrained object proposal network (faster-RCNN [27]) is used to capture the most prominent regions of input image I. These region proposals are further processed by the pretrained ResNet-101 network [26] for visual feature extraction. Similarly, the GloVe embeddings [34] of the words in q are processed by an long short term memory (LSTM) network for computation of question encoding. Detailed descriptions of visual and textual feature extraction are provided in Sections III-A and III-B.

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 Section III-C. These attended visual and textual features are then fused by elementwise multiplication to obtain a (multimodal) joint embedding (Section III-D). This joint embedding is used further for answer prediction.

Classification of large number of categories (here, the number of answers $n_a = \|\mathcal{A}\|$) is often considered a hard problem. However, the answer set $\mathcal{A}$ 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 $\{\text{yes}, \text{no}\}$. Similarly, other question categories will have their corresponding answer sets of lesser size.

Let $\mathrm{q_c} \in \mathcal{Q_C}$ ($\|\mathcal{Q_C}\| = n_{qc}$) 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 $\mathcal{A}_m$ ($\cup_{m=1}^{n_{qc}}\mathcal{A}_m = \mathcal{A}$). The question categorization stage leads to the selection of one from n_{qc} independent answer prediction subsystems. The final predicted answer $\hat{\mathrm{a}} \in \mathcal{A}_m$ ($m = 1, \ldots n_{qc}$) 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 $(\mathrm{I}, \mathrm{q}) \in \mathcal{I} \times \mathcal{Q}$ and corresponding ground-truth pairs $(\mathrm{q_c}, \mathrm{a}) \in \mathcal{Q_C} \times \mathcal{A}$. Fig. 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.

A. *Visual Feature Extraction*

Following the existing approaches in VQA [4], [7]–[9], [24], this proposal uses the pretrained faster-RCNN [27] to select the *top*-n_v regions of I. The d_v-dimensional visual features of these n_v regions are extracted using the pretrained ResNet-101 network [26]. This provides us with the initial (unattended) visual representation $\mathbf{rI} \in \mathbb{R}^{d_v \times n_v}$ of I, and is given by

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

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

B. *Textual Feature Extraction*

All input questions are padded or trimmed to the same length to contain n_w words. These words are embedded in a d_w-dimensional space using pretrained GloVe embeddings [34]. This provides us with the initial unattended question representation $\mathbf{E_q} \in \mathbb{R}^{d_w \times n_w}$ as

$$\mathbf{E_q} = [\mathbf{eq}_0, \ldots \mathbf{eq}_j, \ldots \mathbf{eq}_{n_w}] \;\&\; \forall \mathbf{eq}_j \in \mathbb{R}^{d_w} \tag{2}$$

A question encoding $\mathbf{q_e} \in \mathbb{R}^{d_q}$ is further obtained by processing $\mathbf{E_q}$ by an LSTM network $LSTM_Q$.

C. *Attention Mechanism*

DAQC-VQA exploits dual attention on the following two modalities. These are AoI and AoQ, and are discussed as follows.

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 ith 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 nonlinear transformation [see (3) and (4)].

$$\tilde{\mathbf{r}}_i = \sigma\left(W_a^I \mathbf{r}_i\right) \tag{3}$$

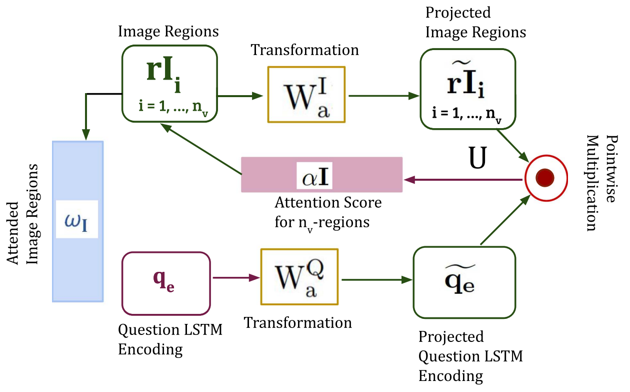


Fig. 3. Block diagram for demonstrating the attention-on-image guided by question.

$$\widetilde{\mathbf{q_e}} = \sigma\left(W_a^Q \mathbf{q_e}\right). \tag{4}$$

Here $\tilde{\mathbf{r}}_i \in \mathbb{R}^{d_{hi}}$ ($i = 1, \ldots n_v$) are the transformed visual features, $\widetilde{\mathbf{q_e}} \in \mathbb{R}^{d_{hi}}$ 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 elementwise multiplied (designated by $\odot$) with $\widetilde{\mathbf{q_e}}$ 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 $W_a^A \in \mathbb{R}^{1 \times d_{hi}}$ [see (5) and (6)].

$$\mathbf{u}_i = \widetilde{\mathbf{rI}}_i \odot \widetilde{\mathbf{q_e}} \tag{5}$$

$$\alpha\mathbf{I} = \mathrm{SoftMax}\left(W_a^A \mathbf{U}\right). \tag{6}$$

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 n_v attended visual embeddings [see (7)]. The functional block diagram for obtaining attended visual representation is presented in Fig. 3.

$$\omega_\mathbf{I} = \sum_{i=1}^{n_v} \alpha_I^{(i)} \times \mathbf{r}_i. \tag{7}$$

The **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 nonlinearly transformed to a d_{hq}-dimensional space [see (8) and (9)].

$$\overline{\mathbf{eq}}_j = \sigma\left(U_a^Q \mathbf{eq}_j\right) \tag{8}$$

$$\overline{\omega} = \sigma\left(U_a^I \omega_\mathbf{I}\right). \tag{9}$$

Here, $\overline{\mathbf{eq}}_j \in \mathbb{R}^{d_{hq}}$ ($j = 1, \ldots n_w$) are the transformed word embeddings, $\overline{\omega} \in \mathbb{R}^{d_{hq}}$ is the transformed visual representation,

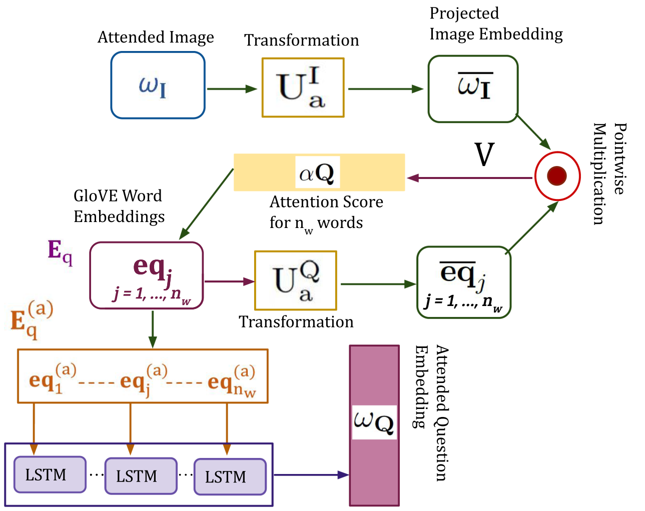

Attended Image
$\omega_{\mathbf{I}}$
Transformation
$\mathrm{U}_{\mathbf{a}}^{\mathbf{I}}$
Projected Image Embedding
$\overline{\omega_{\mathbf{I}}}$
Pointwise Multiplication
V
$\alpha\mathbf{Q}$
Attention Score for $\mathrm{n_w}$ words
GloVE Word Embeddings
$\mathbf{E}_\mathrm{q}$
$\mathbf{eq}_{\boldsymbol{j}}$ $j = 1, ..., n_w$
$\mathrm{U}_{\mathbf{a}}^{\mathrm{Q}}$
Transformation
$\overline{\mathbf{eq}}_{j}$ $j = 1, ..., n_w$
$\mathbf{E}_\mathrm{q}^{(\mathrm{a})}$
$\mathbf{eq}_1^{(\mathrm{a})}$ ---- $\mathbf{eq}_\mathrm{j}^{(\mathrm{a})}$ ---- $\mathbf{eq}_{\mathrm{nw}}^{(\mathrm{a})}$
LSTM
LSTM
LSTM
$\omega_{\mathbf{Q}}$
Attended Question Embedding

Fig. 4. Block diagram for demonstrating the AoQ guided by attended 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 elementwise multiplied to obtain the intermediate embeddings $\mathbf{V} = [\mathbf{v}_1, \ldots \mathbf{v}_j, \ldots \mathbf{v}_{n_w}]$. These are further transformed to obtain the word-wise attention scores, say $\alpha\mathbf{Q}$, $\alpha\mathbf{Q} = [\alpha_Q^{(1)}, \ldots \alpha_Q^{(j)}, \ldots \alpha_Q^{(n_w)}]$ using the parameter vector $U_a^A \in \mathbb{R}^{1\times n_w}$ [see (10) and (11)].

$$\mathbf{v}_j = \overline{\mathbf{eq}}_j \odot \overline{\omega} \tag{10}$$

$$\alpha\mathbf{Q} = \mathrm{SoftMax}\left(U_a^A \mathbf{V}\right). \tag{11}$$

The word embeddings $\mathbf{eq}_j$ are multiplied with their corresponding attention scores $\alpha_Q^{(j)}$ to obtain the attended embeddings $\mathbf{eq^{(a)}}_j \in \mathbb{R}^{d_w}$ $(j = 1, \ldots n_w)$. These attended embeddings $\mathbf{E_q^a} \in \mathbb{R}^{d_w\times n_w}$ are input to the LSTM $LSTM_{aq}$ to obtain the final attended question encoding $\omega_{\mathbf{Q}} \in \mathbb{R}^{d_q}$ [see(12)–(14)].

$$\mathbf{eq^{(a)}}_j = \alpha_Q^{(j)} \times \mathbf{eq}_j \tag{12}$$

$$\mathbf{E_q^a} = \left[\mathbf{eq^{(a)}}_1, \ldots \mathbf{eq^{(a)}}_j, \ldots \mathbf{eq^{(a)}}_{n_w}\right] \tag{13}$$

$$\omega_{\mathbf{Q}} = LSTM_{aq}\left(\mathbf{E_q^a}\right). \tag{14}$$

The parameters of U_a^Q, U_a^I, U_a^A, and $LSTM_{aq}$ are learned during overall model training. The functional block diagram to obtain attended question encoding is shown in Fig. 4.

D. 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 $\omega_{\mathbf{F}}$ [see (17)].

$$\overline{\omega_{\mathbf{I}}} = V_f^I \omega_{\mathbf{I}} \tag{15}$$

$$\overline{\omega_{\mathbf{Q}}} = V_f^Q \omega_{\mathbf{Q}} \tag{16}$$

$$\omega_{\mathbf{F}} = \overline{\omega_{\mathbf{I}}} \odot \overline{\omega_{\mathbf{Q}}}. \tag{17}$$

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.

E. Question Category and Answer Prediction

The question classifier subsystem is a single-layered feedforward network $\mathrm{FCNet_{qcs}}$ with soft-max activation function at its output layer containing $\|\mathcal{Q_C}\| = n_{qc}$ nodes. The fused embedding $\omega_{\mathbf{F}}$ is input to $\mathrm{FCNet_{qcs}}$ to obtain the predicted question category vector $\hat{\mathbf{p}}_{\mathbf{q}} \in (0, 1)^{n_{qc}}$ as output. The predicted question category $\hat{\mathrm{q}}_\mathrm{c}$ is selected by the winner-take-all strategy [see (19)].

$$\hat{\mathbf{q}}_{\mathbf{pc}} = \mathrm{FCNet_{qcs}}\left(\omega_{\mathbf{F}}\right) \tag{18}$$

$$\hat{\mathrm{q}}_\mathrm{c} = \arg\max_{m=1\ldots n_{qc}} \hat{\mathbf{q}}_{\mathbf{pc}}[m] \tag{19}$$

This proposal has n_{qc} answer prediction subsystems (APS, henceforth). These APS are single-layered feedforward networks $\mathrm{FCNet_{aps}}^{(m)}$ $(m = 1, \ldots n_{qc})$. Each network has soft-max activation function at the output layer containing $\| \mathcal{A}_m \|$ nodes. The fused embedding $\omega_{\mathbf{F}}$ is input to the APS $\mathrm{FCNet_{qcs}}^{(\hat{\mathrm{q}}_\mathrm{c})}$ corresponding to the predicted question category $\hat{\mathrm{q}}_\mathrm{c}$. Thes APS produce the predicted answer category vector $\hat{\mathbf{a}}^{(\hat{\mathrm{q}}_\mathrm{c})} \in (0, 1)^{\|\mathcal{A}_{\hat{\mathrm{q}}_\mathrm{c}}\|}$. The predicted answer $\hat{\mathrm{a}}$ is selected by the winner-take-all strategy [see (21)].

$$\hat{\mathbf{a}}_{\mathbf{p}}^{(\hat{\mathrm{q}}_\mathrm{c})} = \mathrm{FCNet_{aps}}^{(\hat{\mathrm{q}}_\mathrm{c})}\left(\omega_{\mathbf{F}}\right) \tag{20}$$

$$\hat{\mathrm{a}} = \arg\max_{l=1,\ldots\|\mathcal{A}_{\hat{\mathrm{q}}_\mathrm{c}}\|} \hat{\mathbf{a}}_{\mathbf{p}}^{(\hat{\mathrm{q}}_\mathrm{c})}[l]. \tag{21}$$

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

F. Model Training

The VQA system has the following learnable components: 1) $LSTM_Q$ in textual feature extraction subsystem; 2) W_a^I, W_a^Q, and W_a^A in AoI subsystem; 3) U_a^Q, U_a^I, U_a^A, and $LSTM_{aq}$ in AoQ subsystem; 4) V_f^I, V_f^Q in fusion subsystem; 5) $\mathrm{FCNet_{qcs}}$ as question categorization subsystem; and 6) $\mathrm{FCNet_{aps}}^{(m)}$ $(m = 1, \ldots n_{qc})$ 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 $\tilde{\mathbf{q_{gc}}} \in \{0, 1\}^{n_{qc}}$ can be constructed using the knowledge of ground-truth category label $\mathrm{q_c}$ of input question q. The cross-entropy loss $\mathcal{L}_{\mathrm{QCS}}$ is formulated as follows:

$$\mathcal{L}_{\mathrm{QCS}} = -\sum_{m=1}^{n_{qc}} \tilde{\mathbf{q_{gc}}}[l] \log\left(\hat{\mathbf{q}}_{\mathbf{pc}}[l]\right). \tag{22}$$

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

$$\mathcal{L}_{\text{APS}}^{(m)} = -\sum_{l=1}^{\|\mathcal{A}_m\|} \mathbf{a_g}^{(m)}[l] \log\left(\hat{\mathbf{a}}_{\mathbf{p}}^{(m)}[l]\right). \tag{23}$$

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

$$\mathcal{L}_T = \mathcal{L}_{\text{QCS}} + \sum_{m=1}^{n_{qc}} \delta[m - \mathrm{q_c}]\mathcal{L}_{\text{APS}}^{(m)}. \tag{24}$$

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

IV. Experiment Details

The performance of the DAQC-VQA is evaluated against baseline methods on the two benchmark VQA datasets. The two datasets are 1) Task Directed Image Understanding Challenge (TDIUC) dataset [35], and 2) Visual Question Answering Version 2.0 dataset VQA2.0 [2]. The two datasets and the baseline methods are discussed in Sections IV-A and IV-B, respectively. Section IV-D summarizes the evaluation metrics used in this study.

A. VQA Datasets

The *TDIUC* dataset [35] is one of the largest available VQA datasets with real images taken from MSCOCO real image dataset. It consists of a total of $1.65\mathrm{M}$ question–image pairs. The TDIUC dataset has 12 explicitly defined category labels for questions. This distinguishes it from VQA 2.0. The question–image pairs are divided into train and validation sets. Fig. 5 shows question category-wise sample counts for TDIUC in train and validation splits.

The *VQA2.0* dataset is one of the widely used VQA datasets of real images with a total of $0.7\mathrm{M}$ question–image pairs partitioned into train, validation, and test sets. For each question–image pair, ten human-annotated answers are given. Fig. 5(b) presents the category-wise frequencies of questions in VQA2.0 dataset.

B. Baseline Methods

The performance of DAQC-VQA is compared against the following baseline methods. All the chosen baseline methods are discussed in Section II.

1) *Fusion-Based Methods*: Fusion of text and image features plays an important role in VQA performance. The state-of-the-art VQA models like MCB [19], MLB [21], MFH [36], MUTAN [6], and BLOCK [23] are chosen as baseline as they primarily contribute toward the fusion of two modalities.
2) *Attention-Based Methods*: The performance of DAQC-VQA against baseline methods is chosen from the following three attention-based categories.
   - *Visual Attention*: SAN [3], BAN [24], BTUP [4], BAN2-CTI [25], CTDA [31], DoG for VQA [37], QTA [9], and QAA [28] are chosen as visual attention-based baseline VQA methods.
   - *Co-attention*: MUTAN [6] is chosen as the baseline method under the coattention category.
   - *Dense attention*: DFAF [7] and MLIN [8] are chosen as dense attention-based baseline VQA methods.

C. Implementation Details

For all experiments (TDIUC and VQA2.0), $n_v = 36$ region proposals are used for visual feature extraction. Each region is represented using ResNet embeddings of $d_v = 2048$ dimensions. Question length is set to 14, $n_w = 14$ words by trimming or padding (if required). 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 is $n_{qc} = 12$ and is $n_{qc} = 3$ for VQA2.0. The model is trained for 17 epochs with a batchsize of 512. Adamax optimizer [38] 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 5 epochs.[1]

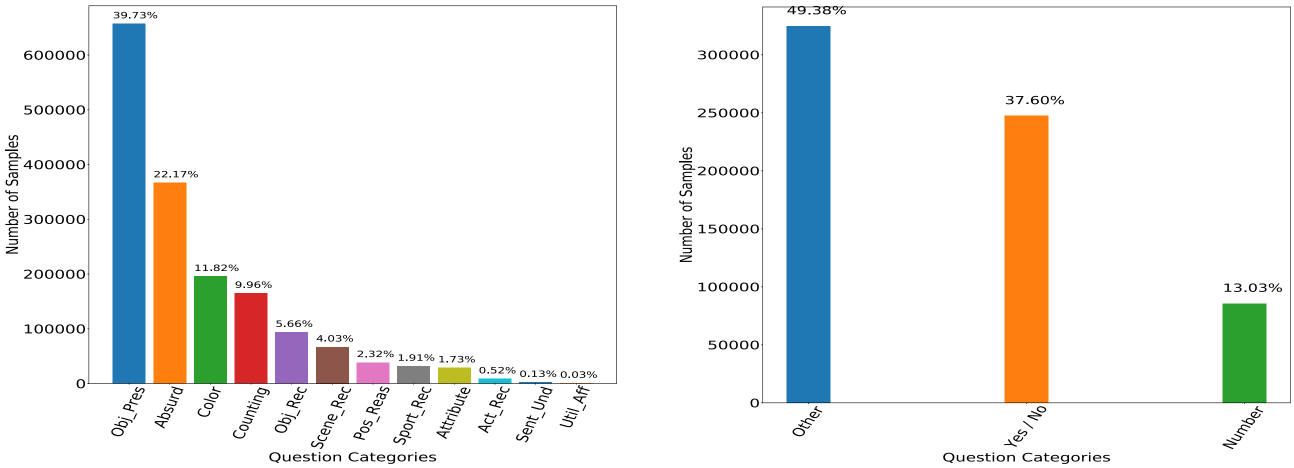
D. Evaluation Metrics

The VQA datasets are imbalanced as a few question categories have a very high number of samples compared to the others. Fig. 5(a) and (b) shows the respective distributions of question categories for the TDIUC and VQA2.0 datasets. Accuracy is not a fair metric for imbalanced data. To deal with the aforementioned issue, Kafle and Kanan [35] have defined two additional evaluation metrics. These are *Arithmetic-Mean Per Type (AMPT)* and *Harmonic-Mean Per Type (HMPT)* for the TDIUC dataset. *Overall accuracy* is also used for evaluation.

The *overall accuracy* is computed as the ratio of the number of correctly predicted samples to the total number of samples. The *AMPT* and *HMPT* are computed as the arithmetic mean and harmonic mean of the answer prediction accuracy values of each question category, respectively. Each question category is assigned a uniform weight, thus making the evaluation unbiased. Unlike AMPT, the HMPT measures the ability of a model to have a high score across all question categories.

To evaluate model performance on VQA2.0, Goyal et al. [2] have used the approach proposed in [1]. The evaluation metric *Accuracy* uses the responses from ten human annotators.

[1] Code is [Online]. Available: https://github.com/akkkb/DAQC_VQA



(a)

(b)

Fig. 5. Proportion of each question category for TDIUC & VQA2.0 datasets. (a) Distribution of 12 categories of TDIUC questions [35] (train + val). (b) Distribution of three categories of VQA2.0 Dataset questions [2] (train + val).

A model-predicted answer $\hat{\mathbf{a}}$ is evaluated as follows:

$$\mathbf{Accuracy}(\hat{\mathbf{a}}) = \min\left\{\frac{\#\mathbf{humans\ that\ said}\ \hat{\mathbf{a}}}{\mathbf{3}}, \mathbf{1}\right\}. \tag{25}$$

This implies that the model-predicted answer $\hat{\mathbf{a}}$ is considered correct if at least three out of ten human annotators have also responded to the input question–image pair with the ground-truth answer $\mathbf{a} = \hat{\mathbf{a}}$.

V. Results and Discussions

This section presents the performance comparison of the proposed DAQC-VQA system with the baseline models. An ablation analysis is also performed to investigate the importance of some of the components of DAQC-VQA.

A. *Quantitative Results*

Question Category-Wise Comparison on TDIUC Dataset. Table I 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* [35].

It can be observed that DAQC-VQA outperforms all the chosen baseline models on all three evaluation metrics. QTA [9] and MCB [19] 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.

TABLE I
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 [19]**	**SAN [3]**	**RAU [35]**	**BAN [24]**	**QTA [9]**	**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**

The bold numeric values highlight the best performances.

It is noteworthy to mention that most of the baseline models (under comparison) exploit complex and deeper attention networks, while DAQC-VQA employs a comparatively simpler attention network.

Overall Performance Comparison on TDIUC Dataset: Table II 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 [8] 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 III compares the performance of DAQC-VQA with state-of-the-art models on validation split of VQA2.0 dataset. For this dataset, models

TABLE II
COMPARISON OF OVERALL ACCURACY OF DAQC-VQA WITH OTHER STATE-OF-THE-ART MODELS ON TDIUC DATASET

Category	Methods	Overall Accuracy
FUSION	MLB[21]	83.10
	MUTAN[6]	82.70
	MFH[36]	84.30
	BLOCK[23]	85.96
VISUAL ATTENTION	BTUP[4]	82.91
	QCG[29]	82.05
	RN[39]	84.61
	RAMEN[32]	86.86
CO-ATTENTION	BAN2-CTI[25]	87.0
	QAA[28]	84.60
DENSE ATTENTION	DFAF[7]	85.55
	MLIN⋆[8]	87.60
CO-ATTENTION	DAQC-VQA	**87.84**

TABLE III
COMPARISON FOR VQA 2.0 VALIDATION SPLIT

Category	Methods	Yes / No	Number	Other	Overall
FUSION	MCB[19]	77.37	36.66	51.23	59.14
	MLB[21]	81.89	42.97	53.89	62.98
	MUTAN[6]	81.09	41.87	54.69	62.71
	MFH[36]				61.60
VISUAL ATTENTION	SAN[3]	78.40	40.71	54.36	61.70
	RN[39]	80.51	41.92	54.75	62.74
	BTUP[4]	80.34	42.80	55.80	63.20
CO-ATTENTION	BAN[24]	–	–	–	66.0
	BAN2-CTI[25]	–	–	–	66.00
	DoG[37]	**82.16**	**45.45**	55.70	64.29
	CTDA[31]	81.26	43.24	55.67	63.65
	QAA[28]	–	–	–	60.5
DENSE ATTENTION	DFAF[7]	–	–	–	**66.21**
	MLIN⋆[8]	–	–	–	66.18
CO-ATTENTION	DAQC-VQA	82.15	43.57	**56.39**	64.51

The bold numeric values highlight the best performances.

TABLE IV
COMPUTATIONAL COMPLEXITY IN TERMS OF MODEL PARAMETER COUNT FOR VQA2.0 DATASET

Model	MCB	MLB	MUTAN	MFH	BAN	DAQC-VQA
	[19]	[21]	[6]	[36]	[24]	(Ours)
Parameter Count (in Millions)	63	25	62	62	76	34
Accuracy	59.14	62.98	63.61	61.6	66.0	64.51

TABLE V
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**

employing dense and complex attention mechanisms such as DFAF [7], MLIN [8], BAN [24], and BAN2-CTI [25] 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 IV compares the number of trainable parameters with that of some of the existing methods.

B. Basic Analysis

Impact of Training Data Size on Performance: To analyze the importance of the amount of training examples in the proposed model, an additional experiment was performed with varying amount of training instances to obtain the learning curves. 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 Fig. 6. Fig. 6(b) 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 Fig. 6(b) 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 subsystems 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 VQA-V2 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 Fig. 7. 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 model, the cascading error induced by first stage of question categorization is very low.

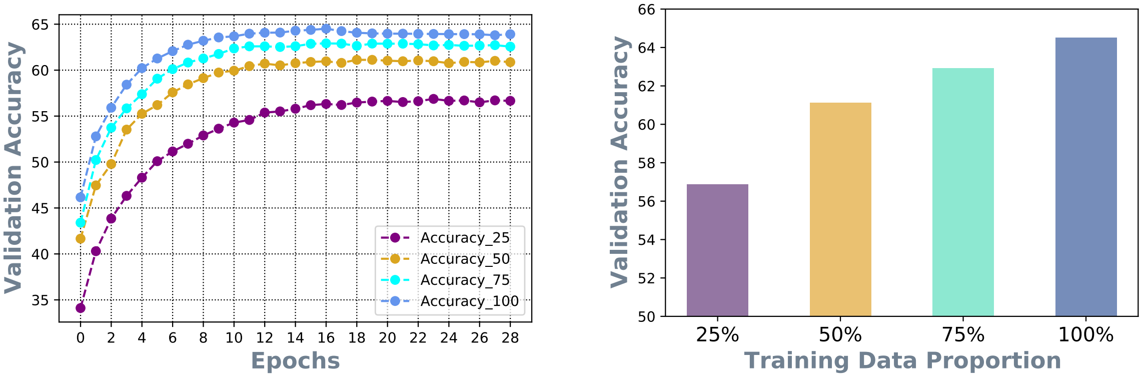
C. Ablation Analysis

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

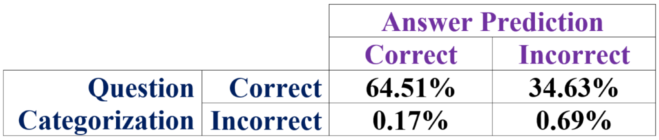
It is observed that in both cases, the best performances are obtained by using $I_A \otimes Q_A$ as an input embedding to the question classifier subsystem. However, the impact of the choice of



(a)

(b)

Fig. 6. Illustration of 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. 7. *Question Categorizer – Answer Predictor* cascade error analysis.

TABLE VI
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 VII
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

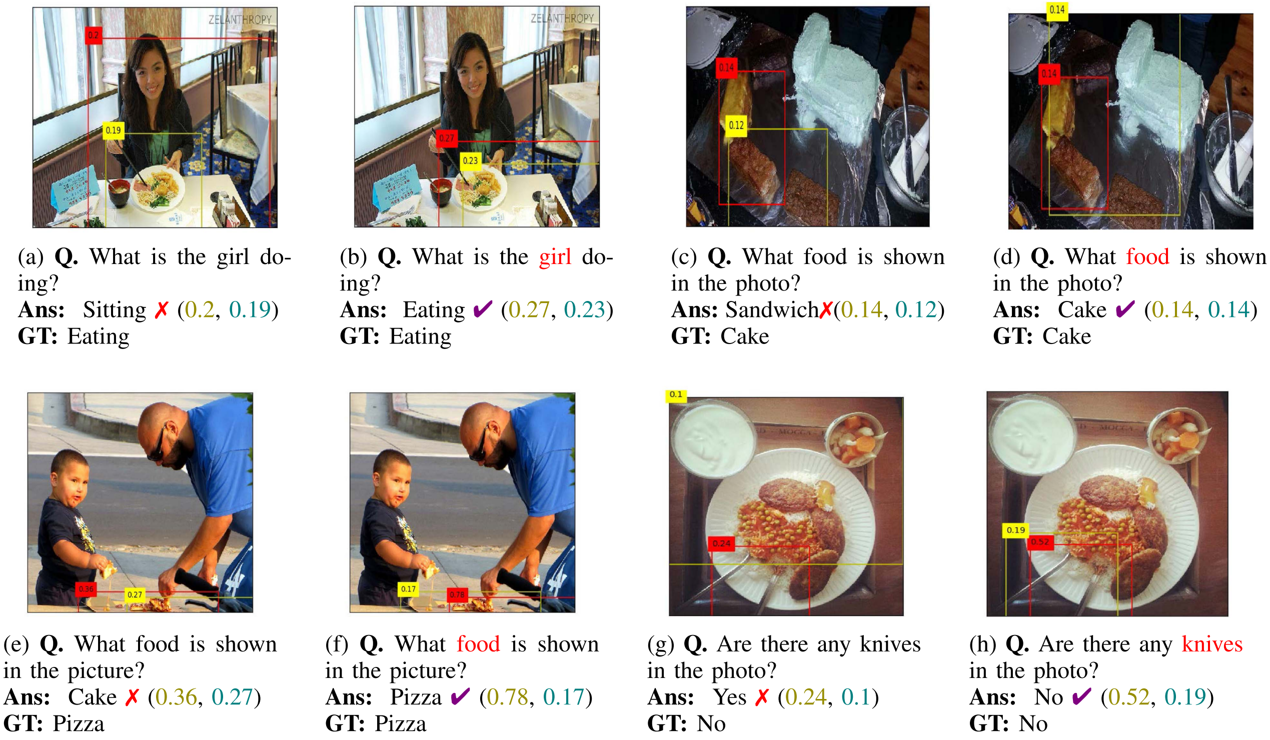
TABLE VIII
EVALUATING MODEL PERFORMANCE ON VQA2.0 DATASET TO INVESTIGATE THE EFFECT OF DUAL ATTENTION AND QUESTION CATEGORIZATION

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**

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 [35], [2] 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 VII). 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 VIII. The performance of primary model neither deploying dual attention nor question categorization is shown in the 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. To analyze the effect of question categorization component (only), on top of primary model, instead of feeding fused embedding to answer classifier network, question categorization is performed followed by answer predictor networks. 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. The last row of Table VIII 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.



(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

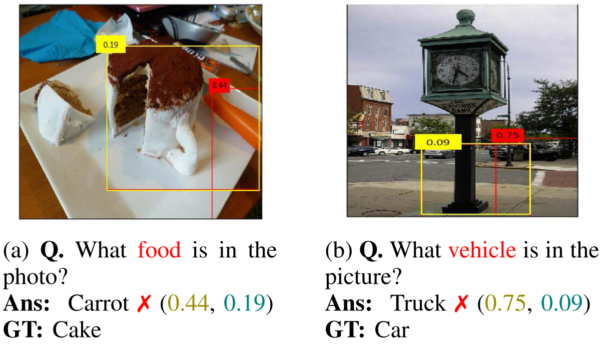
(e) **Q.** What food is shown in the picture?
Ans: Cake ✗ (0.36, 0.27)
GT: Pizza

(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. 8. 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. 9. Poor performance cases of DAQC-VQA due to failure in capturing relevant relations.

D. Qualitative Results

The qualitative evaluation of the DAQC-VQA system is performed through the visualization of results (Fig. 8). 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 [4] is used as baseline. The baseline uses $\mathrm{I_A} \otimes \mathrm{Q}$ fused embedding for answer classification. These results are demonstrated on TDIUC dataset for different categories like *object recognition*, *activity recognition*, etc.

In Fig. 8(a), 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 8(c), 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 Fig. 9(a), DAQC-VQA assigns the highest score to a *knife* that resembles *carrot*. Thus, the answer to the question of inferring the food is wrongly predicted as *carrot*. Similarly, in Fig. 9(b), the question is on identifying a vehicle in the scene. However, DAQC-VQA focuses on a *truck* while the ground-truth answer is *car*.

VI. Conclusion

This article proposed an approach for DAQC-VQA. The dual attention mechanism extracts richer feature representations for both image and text modalities through cross-modal interactions. Categorizing the input question reduces the answer search space for the answer prediction subsystem. Further, this combination of dual attention and question categorization for VQA is realized through an end-to-end trainable system with a joint loss function. DAQC-VQA is validated on two benchmark VQA datasets (TDIUC and VQA2.0) against several state-of-the-art approaches. Quantitative and qualitative evaluation demonstrate the competitive (and often better) performance of DAQC-VQA with respect to the baseline models.

The present work can be extended in the following directions. First, DAQC-VQA employs simpler attention mechanism and feature fusion schemes. Thus, transformer-based attention models and more complex fusion schemes can be employed for further performance improvement. Second, explicit logical reasoning and interobject relation predictions can be appended to improve answer prediction accuracy. Third, it is witnessed that a larger training set improves accuracy. Thus, data augmentation (for both image and question) techniques can be explored to generate larger training datasets.

REFERENCES

[1] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[2] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6904–6913.

[3] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 21–29.

[4] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[5] A. Agrawal *et al.*, "VQA: Visual question answering," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 4–31, May 2017.

[6] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2612–2620.

[7] P. Gao *et al.*, "Dynamic fusion with intra-and inter-modality attention flow for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6639–6648.

[8] P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5825–5835.

[9] Y. Shi, T. Furlanello, S. Zha, and A. Anandkumar, "Question type guided attention in visual question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 151–166.

[10] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Process. Image Commun.*, vol. 80, 2020, Art. no. 115648.

[11] 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.

[12] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[13] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Eur. Conf. Comput. Vis.*, 2016, pp. 451–466.

[14] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.

[15] A. Mishra, A. Anand, and P. Guha, "Multi-stage attention based visual question answering," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 9407–9414.

[16] A. Mishra, A. Anand, and P. Guha, "CQ-VQA: Visual question answering on categorized questions," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[17] J. Song, P. Zeng, L. Gao, and H. T. Shen, "From pixels to objects: Cubic visual attention for visual question answering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 906–912.

[18] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4613–4621.

[19] 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 *Proc. 2016 Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.

[20] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2013, pp. 239–247.

[21] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. Int Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=r1rhWnZkg

[22] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1821–1830.

[23] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8102–8109.

[24] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.

[25] T. Do, T.-T. Do, H. Tran, E. Tjiputra, and Q. D. Tran, "Compact trilinear interaction for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 392–401.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[28] M. Farazi, S. Khan, and N. Barnes, "Question-agnostic attention for visual question answering," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3542–3549.

[29] W. Norcliffe-Brown, S. Vafeias, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8334–8343.

[30] H. Xu and K. Saenko, "Dual attention network for visual question answering," in *Proc. Eur. Conf. Comput. Vis., 2nd Workshop Storytelling Images Videos*, 2016. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-46604-0

[31] W. Tian, R. Zhou, and Z. Zhao, "Cascading top-down attention for visual question answering," in *Proc. Int. Joint Conf. Neural Netw.*, 2020, pp. 1–7.

[32] R. Shrestha, K. Kafle, and C. Kanan, "Answer them all! Toward universal visual question answering models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10472–10481.

[33] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4971–4980.

[34] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[35] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1965–1973.

[36] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, no. 99, pp. 1–13, Dec. 2018.

[37] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10267–10276.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[39] A. Santoro *et al.*, "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.