CQ-VQA: Visual Question Answering on Categorized Questions

Aakansha Mishra
Dept. of CSE
IIT Guwahati
Guwahati, India
ak.kkb@iitg.ac.in

Ashish Anand
Dept. of CSE
IIT Guwahati
Guwahati, India
anand.ashish@iitg.ac.in

Prithwijit Guha
Dept. of EEE
IIT Guwahati
Guwahati, India
pguha@iitg.ac.in

***Abstract*—This paper proposes *CQ-VQA*, a novel two-level hierarchical but end-to-end model to solve the task of visual question answering (VQA). The first level of CQ-VQA, referred to as Question Categorizer (QC), classifies questions to reduce the potential answer search space. The QC uses attended and fused features of the input question and image. The second level, referred to as Answer Predictor (AP), comprises of a set of distinct classifiers corresponding to each question category. Depending on the question category predicted by QC, only one of the classifiers of AP remains active. The loss functions of QC and AP are aggregated together to make it an end-to-end model. The proposed model (CQ-VQA) is evaluated on the TDIUC dataset and is benchmarked against state-of-the-art approaches. Results indicate a competitive or better performance of CQ-VQA.**

***Index Terms*—VQA, CQ-VQA, Attention Network**

I. Introduction

The objective of a Visual Question Answering (VQA) system [1], [2] is to generate a natural language answer to a natural language question asked about a given image. VQA has gained wide attention for several reasons. First, it has got many real-life applications, e.g., scene interpretation for assistance to visually impaired persons, interactive robotic systems, etc. Second, it is a challenging AI problem as it requires a simultaneous understanding of two modalities – image and text, and reasoning over the relations among the modalities. This wide attention has naturally led to the development of a plethora of methods.

The early approaches of VQA primarily focused on feature fusion of two modalities, where image- and text-based features are fused using simple techniques like addition, concatenation, or element-wise products [1], [3]. Later, improved feature fusion mechanisms such as bilinear pooling [4] and its variants MCB [4], MFB [5], MLB [6] and MUTAN [7] were proposed.

Another class of methods focus on identifying 'relevant' image regions for answering the given question. Attention-based methods [8]–[12] fall into this category. These methods aim to assign higher weights (attention scores) to the image regions pertinent to answer the given question while providing relatively negligible attention to other regions. It is noteworthy to mention that such methods do fuse features of the different modalities. However, performance improvement significantly depends on the extent of the information obtained by exploiting attention in different modalities. For example, studies in [11], [13] have shown that along with question guided attention on image, attention from image to questions allow better information flow and interaction between the two modalities, resulting in improved performance.

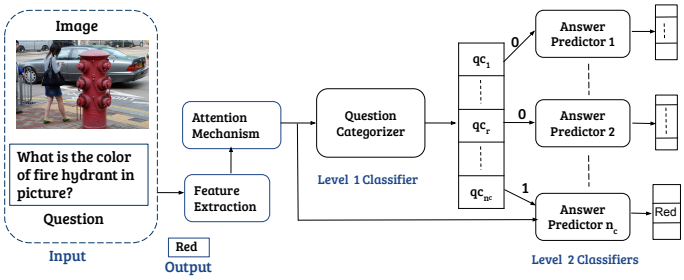

Fig. 1: An overview of the proposed framework of CQ-VQA. Features extracted from input question about an image are fused with image features obtained through an attention mechanism. The hierarchical structure of CQ-VQA first categorizes the input question (level 1 classifier) and accordingly selects an answer predictor for identifying the output answer.

This paper proposes a hierarchical model, referred to as *CQ-VQA*. CQ-VQA hierarchically solves the VQA task by breaking it into two sub-problems. Figure 1 illustrates the motivation and working principle of the CQ-VQA. As illustrated, a question "*What is the color of fire-hydrant in picture?*" is asked about the given image. As a human, we immediately understand that the question is about the color of an object, and the answer must be one of the colors. CQ-VQA mimics this intuition in a two-level hierarchical classification model. At the first level, a single classifier identifies the question category based on the fused features of the given question and image. Based on the selected question category by the first level classifier, the CQ-VQA model sends fused features to the corresponding classifiers of the second level. The second level contains a set of distinct classifiers, one for each question category and output of each classifier is a set of answers belongs to that category. In contrast to the existing VQA models, where they need to explore the entire search space of answers, CQ-VQA focuses on smaller answer search spaces in the final classification stage.

The performance of CQ-VQA is evaluated on the TDIUC

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

dataset [14] containing 12 explicitly defined question categories. The other commonly used VQA datasets do not have question categories explicitly available. Compared to state-of-the-art models, the experimental results on this dataset have shown the competitive or better performance of CQ-VQA. The primary contributions of this work are as follows:

- A novel hierarchical model for decomposing the VQA task into two sub-problems – question categorization and answer prediction.
- End-to-end model training model by combining the two loss functions of the two sub-problems.
- Comprehensive overall and question category-wise performance analysis and comparison with state-of-art VQA models.

The rest of the paper is organized as follows. A brief review of VQA literature is presented in Section II. Section III discusses the necessary details of the proposed approach. The experimental results are presented and discussed in Sections IV and V respectively. Finally, we conclude in Section VI and sketch the extensions of the present proposal.

II. Related Work

Existing works in VQA can be broadly divided into three categories. These are (a) feature fusion-based approaches, (b) attention-based methods, (c) reasoning based techniques. This proposal uses attention models for visual and question feature fusion. Accordingly, the existing works in the first two categories are briefly reviewed next.

A. VQA: Feature Fusion

These approaches project both visual and question embeddings to a common space to predict the answer. The embeddings of the visual modality are obtained using pre-trained CNNs. These networks are learned from large image data sets dealing with different classification problems [15]–[17]. The questions are represented in two ways. The first class of approaches have used Bag-of-Words (BoW) representations for questions [1], [3], [18]. The second group of methods represent questions as sequences of pre-trained word embeddings [19], [20]. These embedding sequences are further input to Recurrent Neural Networks (RNNs) for obtaining question embeddings [19], [20]. The third group of approaches represent questions using pre-trained CNN features [21], [22]. However, most existing works use the second method involving pre-trained word embedding sequence and RNN.

The Neural-Image-QA [23] system uses VGG-Net image features [24] and one-hot-encoded word representations are given as input to Long short term memory (LSTM) network for generating question features. Authors in [1], [2] have fused extracted image features (VGG-Net) and LSTM encoded question vector by element-wise multiplication. The 4096-dimensional image features in [25] are transformed into a vector (of same size as word embedding dimension). The modified and combined embeddings are given as input to LSTM for generating an answer. In [4], authors have proposed the fusion of multi-modal features through the outer product (Bilinear pooling) as it provides multiplicative interaction (rich representation) between all elements of modalities. Bilinear pooling based fusion achieves superior performance, but seems to be a less efficient solution as a large number of parameters are needed for the projection of outer product to obtain a joint representation of both modalities. However, later works in [6], [26] have proposed Multimodal Compact Bilinear Pooling (MCB) and Multimodal Low-rank Bilinear (MLB) pooling, respectively for efficient use of bilinear pooling.

B. Attended Feature Fusion

Attention-based models [8]–[12] focus on the image region(s) that is (are) most relevant to the task (question). In VQA, attention models aim to interpret "where to look" in the image for answering the question. Existing works have used attention in different ways. The attention can be on image [9], on question [12], or on both (Co-attention) [11]. For example, [8] proposed a model that predicts the answer by selecting an image region which is most relevant to question text.

A multi-step attention based method is proposed in [9] that allows reasoning over fine-grained information asked in a question. Question embeddings used to generate attention distribution over image regions. The attention score obtained from the weighted sum of image region embeddings is used as a visual feature for the next step. The attention mechanism is used with outer product based fusion of image and question embeddings [26]. Multimodal Factorized Bilinear (MFB) [5] pooling has been introduced to efficiently and effectively combine multi-modal features on top of low-rank bilinear pooling technique [6]. The usage of a stack of dense co-attention layers is proposed in [13]. Here, each word of a question interacts with each region proposal in an image and vice-versa. A combination of top-down and bottom-up attention models is proposed in [27]. The bottom-up model detects salient regions extracted using Faster-RCNN [28], while the top-down mechanism uses task-specific context to predict attention score of the salient image regions.

A Question-Conditioned Graph (QCG) is processed for VQA in [29]. Here, the objects proposed from faster-RCNN act as nodes and edges define the interaction between regions conditioned on the question. For each node, a set of nodes is chosen from the neighborhood using the strongest connection criterion. This leads to a question specific graph structure. Bilinear Attention Network (BAN) [30] fuses both the modalities by the interaction of each region proposal with each word of the question and uses residual connections to provide multiple attention glimpses. In Relation Network (RN) [31], every pair of object proposal embeddings are aggregated (summed up) and it is found that the resulting vector encodes the relationship between different regions, thereby enabling compositional reasoning. In Question Type guided Attention (QTA) [32], the semantics of question category are used with both bottom-up, top-down and residual features. A recurrent deep neural network with an attention mechanism is proposed in [33], where each network is capable of predicting the answer.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Dynamic Fusion With intra-and inter-modality Attention Flow (DFAF) [34] is a stacked network that uses inter-modality and intra-modality information for fusing features. Here, the use of average pooled features can dynamically change intra-modality information flow. The Multimodal Latent Interaction (MLIN) is proposed in [35] that realizes multi-modal reasoning through the process of summarization, interaction and aggregation. A generalized algorithm RAMEN is proposed in [36] to deal with VQA datasets containing either only synthetic or real-world images.

This proposal uses attention score obtained from top-down attention mechanism for fusing image and question embeddings. The answer space is decomposed into smaller subspaces based on specific question categories. A two-stage hierarchical process is followed to predict answers (stage-2) based on predicted question category (stage-1). Our proposal is discussed next.

III. Proposed Approach

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

$$\hat{\mathbf{a}} = \underset{\mathbf{a} \in \mathbf{A}}{\arg\max}\, P(\ \mathbf{a} \mid \mathcal{S}_{VQA}(\ \mathbf{f}(\mathbf{q})\ ,\ \mathbf{g}(I)\ )\ ) \tag{1}$$

This proposal approaches VQA using a hierarchical architecture (Figure 2) involving different answer prediction subsystems corresponding to distinct question categories. This requires suitable deep networks for computing question and image features (vector space embeddings). These features of different modalities are fused using attention information. The process of feature extraction (Subsections III-A and III-B) and attention score guided feature fusion (Subsection III-C) are described next.

A. Visual Feature Extraction

The *Visual Features* of images are extracted as embeddings by using a pre-trained deep network. Existing works [27], [30], [36] in VQA have mostly used *Faster-RCNN* [28] for visual feature extraction. This model employs *ResNet-101* [16] as its base network and uses top-k region proposals ($\mathbf{R}_i;\ i = 1, \ldots k$) for visual feature extraction. Let $\mathbf{v}_i$ ($\mathbf{v}_i \in \mathcal{R}^{d_v}$) be the ResNet-101 feature extracted from $\mathbf{R}_i$. The image I is represented by the set of visual features $\mathbf{G}(I) = \{\mathbf{v}_i; i = 1, \ldots k\}$. Experimental results have shown that a higher value of k leads to a better representation at the expense of significantly higher computations. This proposal also uses the Faster-RCNN model with $k = 36$ [27], [36]. This is followed by question feature extraction and is described next.

B. Question Feature Extraction

The *Question Features* are computed by using word embeddings obtained from pre-trained deep netwroks. All questions are padded or truncated to obtain word sequences of a fixed length (n_w, say). The pre-trained GloVe embedding [20] is used to convert a question $\mathbf{q}$ to an ordered sequence of word embeddings $\mathbf{E_w}(\mathbf{q}) = \{\mathbf{ew}_j : \mathbf{ew}_j \in \mathcal{R}^{d_w}; j = 1, \ldots n_w\}$. This obtained sequence of word embeddings are fed to a LSTM network $\mathbf{Q_{LSTM}}$ to generate the question embedding $\mathbf{f}(\mathbf{q})$. The j^{th} hidden state embedding of $\mathbf{Q_{LSTM}}$ is obtained for each input word embedding $\mathbf{ew}_j$. The question embedding is obtained as the output of the final hidden state of $\mathbf{Q_{LSTM}}$ as $\mathbf{f}(\mathbf{q}) = \mathbf{Q_{LSTM}}(\mathbf{q})$ ($\mathbf{f}(\mathbf{q}) \in \mathcal{R}^{d_q}$). The architecture of $\mathbf{Q_{LSTM}}$ is adopted from the LSTM network used in [37].

The features extracted from visual (image) and text (question) modalities are fused using scores obtained from a top-down attention model. This attention mechanism is described next.

C. Attention Mechanism

Attention plays a key role in fusing visual and question features. Attention guided fusion of visual and textual features is well explored in several existing works (Sub-section II-B). Only a few among top-k region proposals (identified during visual extraction) are relevant with respect to an input question $\mathbf{q}$. An attention network provides different scores to these region proposals using $\mathbf{f}(\mathbf{q})$ and $\mathbf{G}(I)$. Attention score guided feature fusion is performed to obtain the embedding $\mathbf{h_a}(\mathbf{q}, I)$. This process is described next.

The visual and question features are of different dimensions. Two fully connected networks $\mathbf{VQ_{fcn}}$ ($\mathbf{VQ_{fcn}} : \mathcal{R}^{d_v} \rightarrow \mathcal{R}^{d_f}$) and $\mathbf{QQ_{fcn}}$ ($\mathbf{QQ_{fcn}} : \mathcal{R}^{d_q} \rightarrow \mathcal{R}^{d_f}$) are used to map both visual and question features to vectors of size d_f. Both $\mathbf{VQ_{fcn}}$ and $\mathbf{QQ_{fcn}}$ are fully connected networks where the input and output layers are directly connected without any intermediate hidden layer. These two networks are used to map both visual and question embeddings to $\mathcal{R}^{d_f}$ as:

$$\tilde{\mathbf{v}}_i = \mathbf{VQ_{fcn}}(\mathbf{v}_i) \tag{2}$$

$$\tilde{\mathbf{f}}_\mathbf{q} = \mathbf{QQ_{fcn}}(\mathbf{f}(\mathbf{q})) \tag{3}$$

These networks ($\mathbf{VQ_{fcn}}$ and $\mathbf{QQ_{fcn}}$) provide us with $\tilde{\mathbf{G}}(I) = \{\tilde{\mathbf{v}}_i; i = 1, \ldots k\}$ and $\tilde{\mathbf{f}}_\mathbf{q}$ respectively. Let $\mathbf{u}_i = \tilde{\mathbf{v}}_i \otimes \tilde{\mathbf{f}}_\mathbf{q}$ be the element-wise product of $\tilde{\mathbf{v}}_i$ and $\tilde{\mathbf{f}}_\mathbf{q}$. The vector $\mathbf{u}_i$ ($\mathbf{u}_i \in \mathcal{R}^{d_f}$) is input to the attention network $\mathbf{NN_{att}}$ to obtain the attention score s_i corresponding to region proposal $\mathbf{R}_i$ ($i = 1, \ldots k$). The attention network $\mathbf{NN_{att}}$ is a fully connected network ($\mathbf{NN_{att}} : \mathcal{R}^{d_f} \rightarrow (0, 1)$) that directly connects the input to a single-valued output without any intermediate hidden layer. Finally, the attention score weighted feature fusion is performed to obtain $\mathbf{h_a}(\mathbf{q}, I)$ as:

$$\mathbf{h_a}(\mathbf{q}, I) = \tilde{\mathbf{f}}_\mathbf{q} \otimes \left(\sum_{i=1}^{k} s_i \tilde{\mathbf{v}}_i \right) \tag{4}$$

where $\mathbf{h_a}(\mathbf{q}, I) \in \mathcal{R}^{d_f}$. The process of attention score guided feature fusion is illustrated in Figure 3. The value of $\mathbf{h_a}(\mathbf{q}, I)$

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

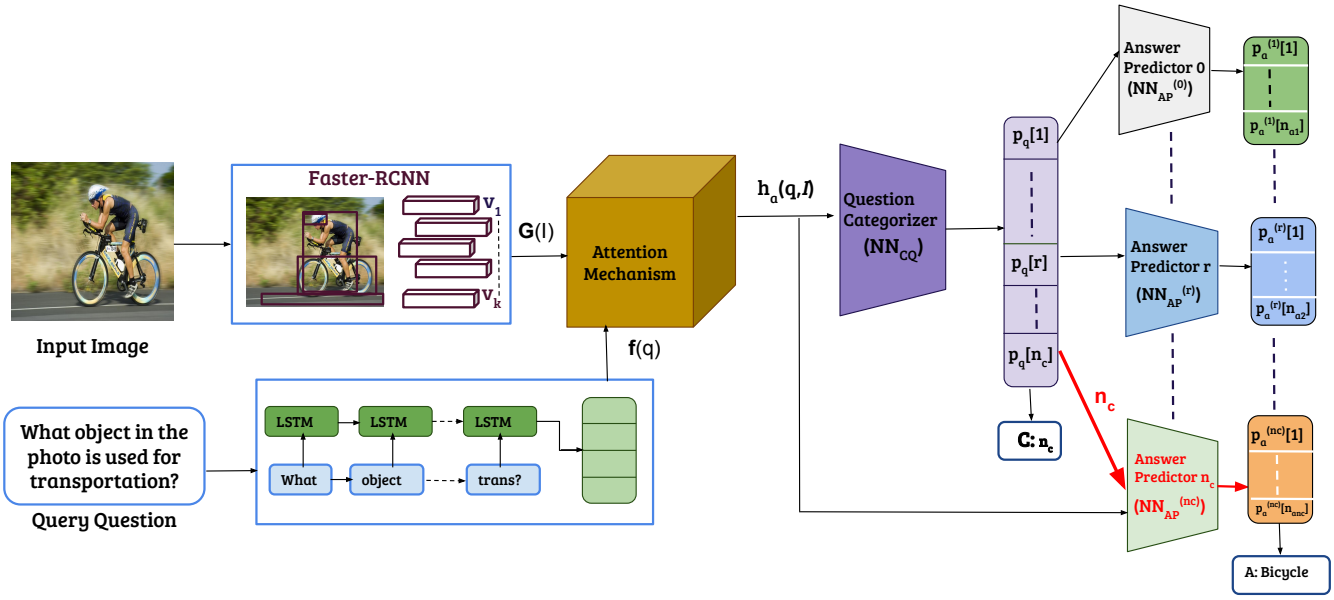

Fig. 2: Illustrating the proposed approach. ResNet-101 features of regions proposed by faster-RCNN are extracted for visual representation. The question is encoded using a LSTM. Features of both modalities are fused by using region scores from a top-down attention model. The fused embedding is input to the **Question Categorizer** which selects one **Answer Predictor** (from multiple classifiers) to identify the output answer. For illustration, the n_c^{th} category got highest score from category selection network. hence, the classifier corresponding to n_c^{th} category will be active (shown in red) for final answer prediction.

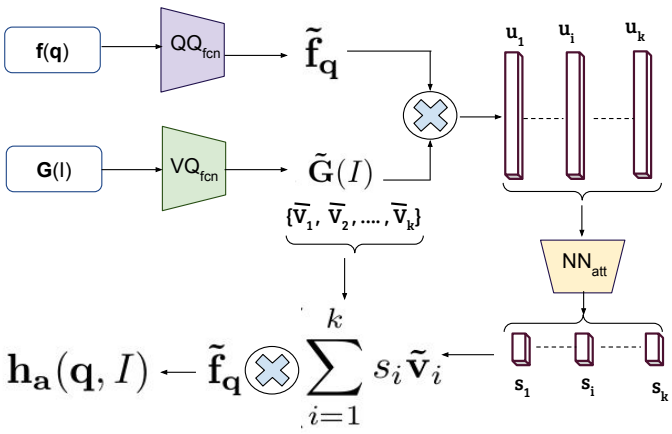

Fig. 3: The functional block diagram of top-down attention network score guided fusion of visual and question features.

depends on the parameters of $\mathbf{Q_{LSTM}}$, $\mathbf{NN_{att}}$, $\mathbf{VQ_{fcn}}$ and $\mathbf{QQ_{fcn}}$. The parameters of aforementioned networks are tuned by minimizing the net loss (equation 9) defined for the proposed hierarchical model CQ-VQA. The upcoming subsection presents the CQ-VQA model and the associated loss functions.

D. CQ-VQA: Learning the Model

This work proposes a hierarchical model for visual question answering. This hierarchical model has two levels. The first level, takes the attention guided fused feature $\mathbf{h_a(q, I)}$ to classify into one of n_c question categories. Note that n_c depends on the dataset under consideration. For example TDIUC (Section IV-A) has $n_c = 12$ question categories. The first level uses a single hidden layer feedforward network $\mathbf{NN_{CQ}}$ ($\mathbf{NN_{CQ}} : \mathcal{R}^{d_q} \rightarrow (0,1)^{n_c}$) to perform the task of question category classification.

Let $\mathbf{t_q}$ be the one-hot-encoded target vector representing the ground truth question category $\mathbf{q_c}$. Let $\mathbf{p_q}$ be the output of $\mathbf{NN_{CQ}}$. The question classification loss is defined as:

$$\mathcal{L}_Q(\mathbf{q}, I, \mathbf{q_c}) = -\sum_{r=1}^{n_c} \mathbf{t_q}[r] \log(\mathbf{p_q}[r]) \tag{5}$$

The second level of the hierarchy in CQ-VQA predicts the answers based on input question and image. Generally, the answer search space is large. This proposal decomposes the answer set $\mathbf{A}$ into n_c subsets $\mathbf{A}_r$ according to the question categories. Thus, $\mathbf{A}_r \subset \mathbf{A}$ $(r = 1, \ldots n_c)$ and $\cup_{r=1}^{n_c} \mathbf{A}_r = \mathbf{A}$. The question classification network $\mathbf{NN_{CQ}}$ acts as a switch for selecting one of n_c answer prediction sub-systems. Each answer prediction sub-system is a VQA system capable of predicting one from a subset of $\mathbf{A}$ based on the question category. We believe that this answer search space decomposition makes the task of VQA easier by reducing the number of outputs for each answer predictor. For example, questions of the form "Is there a bird in the image?" are of the binary answer (yes/no) category and the corresponding answer prediction sub-system has only two outputs. Similarly, questions asking for "What color is the bird?" has only a small number of answers (colors) to choose from a small subset of $\mathbf{A}$.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Let $n_a^{(r)}$ be the number of possible answers for the r^{th} question category. The target answer $\mathbf{a}$ is one-hot-encoded through the $n_a^{(r)}$ dimensional vector $\mathbf{t_a^{(r)}}$. The attention guided fused feature $\mathbf{h_a}(\mathbf{q}, I)$ is input to the r^{th} answer prediction sub-system $\mathbf{NN_{AP}^{(r)}}$ for predicting the answer probability vector $\mathbf{p_a}^{(r)}$, ($\mathbf{p_a}^{(r)} \in (0,1)^{n_a^{(r)}}$). The answer prediction networks are fully connected networks with single hidden layer. The loss $\mathcal{L}_A^{(r)}$ for training $\mathbf{NN_{AP}^{(r)}}$ is defined as:

$$\mathcal{L}_A^{(r)}(\mathbf{q}, I, \mathbf{a}) = -\sum_{j=1}^{n_a^{(r)}} \mathbf{t_a}^{(r)}[j] \log(\mathbf{p_a}^{(r)}[j]) \tag{6}$$

The net loss at the second level is defined as

$$\mathcal{L}_{AA}(\mathbf{q}, I, \mathbf{a}) = \sum_{r=1}^{n_c} \delta[r-\rho]\mathcal{L}_A^{(r)}(\mathbf{q}, I, \mathbf{a}) \tag{7}$$

$$\rho = \underset{l=1,\ldots n_c}{\arg\max}\, \mathbf{p_q}[l] \tag{8}$$

where $\delta[i-j]$ is the Kronecker delta function. The overall loss of CQ-VQA for input question $\mathbf{q}$, its category $\mathbf{q_c}$, associated image I and ground-truth answer $\mathbf{a}$ is given by:

$$\mathcal{L}_{CQVQA}(\mathbf{q}, \mathbf{q_c}, I, \mathbf{a}) = \mathcal{L}_Q(\mathbf{q}, I, \mathbf{q_c}) + \mathcal{L}_{AA}(\mathbf{q}, I, \mathbf{a}) \tag{9}$$

This proposal minimizes the loss $\mathcal{L}_{CQVQA}(\mathbf{q}, \mathbf{q_c}, I, \mathbf{a})$ for all question-image-answer combinations $(\mathbf{q}, I, \mathbf{a}) \in \mathbf{Q} \times \mathbf{I} \times \mathbf{A}$. The gradients computed by using this net loss (equation 9) are back-propagated for end-to-end training of $\mathbf{Q_{LSTM}}$, $\mathbf{VQ_{fcn}}$, $\mathbf{QQ_{fcn}}$, $\mathbf{NN_{att}}$, $\mathbf{NN_{CQ}}$ and $\mathbf{NN_{AP}^{(r)}}$ $(r = 1, \ldots n_c)$.

IV. Experiments

This section briefly discusses the dataset, evaluation metrics, and implementation details.

A. Dataset: TDIUC

We select Task-Directed Image Understanding Challenge (TDIUC) dataset [14] in our experiments. The TDIUC dataset provides categories of questions associated with images explicitly and is ideal for evaluation of the CQ-VQA. The other VQA datasets do not explicitly provide such information, so this paper does not consider them for evaluation.

TDIUC [14] is the largest available VQA dataset of real images. TDIUC consists of $16,54,167$ open-ended questions of 12 categories associated with $1,67,437$ images. 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 4 shows the category-wise sample distribution of questions. The largest 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. The 'Absurd' is an exceptional category consisting of questions having no semantic relation with an associated image input. Such questions have a single answer, and that is 'Does-Not-Apply' [14]. Researchers have observed the

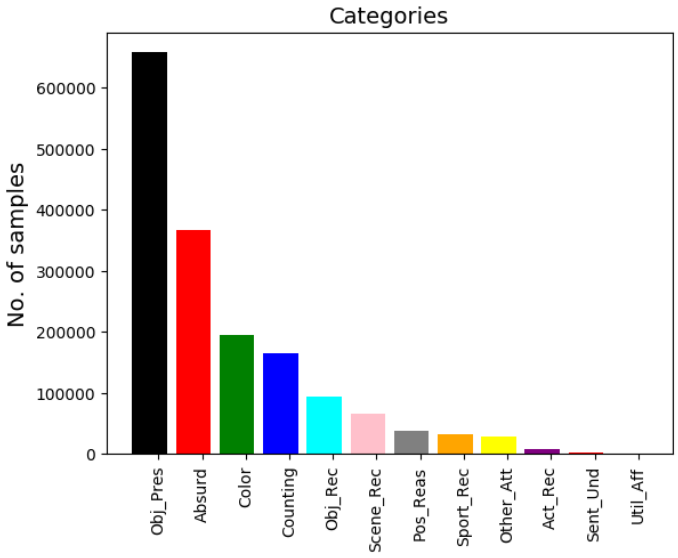

Fig. 4: Distribution of 12 Categories of TDIUC Questions [14].

phenomenon of VQA model bias towards language priors. The introduction of the 'Absurd' forces the model to learn proper relations between the question(s) and the visual contents of the image(s).

B. Evaluation Metrics

This proposal employs three commonly used evaluation metrics 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 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)* [14] are generally used to achieve unbiased evaluation. *Arithmetic-MPT* computes the arithmetic mean of the individual accuracies of each question category. This evaluation metric assigns uniform weight to each question category. *Harmonic-MPT* reports the harmonic mean of individual question category accuracies. Unlike Arithmetic-MPT, the Harmonic-MPT measures the ability of a model to have a high score across all question categories.

C. Implementation Details

The top-36 ($k = 36$) region proposals of ResNet-101 are used to compute $d_v = 2048$ dimensional visual feature vectors. The length of each question is set to $n_w = 14$ words. Questions with more than 14 words are truncated and lesser than that are padded with zero embedding vectors. The pre-trained GloVe network is used to generate word embeddings of size $d_w = 300$. A sequence of these word embeddings are input to a LSTM ($\mathbf{Q_{LSTM}}$, Subsection III-B) for question embedding generation. The sizes of hidden and output layer of $\mathbf{Q_{LSTM}}$ are both set to 1024. Thus, the question embeddings

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I: Category-wise performance comparison with state-of-the-art methods on TDIUC dataset

Question Type	NMN [38]	RAU [33]	MCB [4]	QTA [32]	CQ-VQA
Scene Recognition	91.88	93.96	93.06	93.80	**94.05**
Sport Recognition	89.99	93.47	92.77	**95.55**	95.39
Color Attributes	54.91	66.86	68.54	60.16	**73.35**
Other Attributes	47.66	56.49	56.72	54.36	**59.24**
Activity Recognition	44.26	51.60	52.35	60.10	**61.19**
Positional Reasoning	27.92	35.26	35.40	34.71	**40.40**
Object Recognition	82.02	86.11	85.54	86.98	**88.13**
Absurd	87.51	96.08	84.82	**100.0**	**100.0**
Utility & Affordance	25.15	31.58	**35.09**	31.48	34.50
Object Presence	92.50	94.38	93.64	94.55	**95.41**
Counting	49.21	48.43	51.01	53.25	**56.78**
Sentiment Und.	58.04	60.09	66.25	64.38	**66.56**
Overall Accuracy	79.56	84.26	81.86	85.03	**87.52**
Arithmetic-MPT	62.59	67.81	67.90	69.11	**72.08**
Harmonic-MPT	51.87	59.00	60.47	60.08	**64.45**

TABLE II: Comparing *Overall Accurary* of CQ-VQA and other state-of-art models. CQ-VQA outperforms all models except MLIN. The higher accuracy of MLIN (marked with $\star$) can be attributed to its usage of top 100 region proposals for visual feature extraction, while all other models (including CQ-VQA) have used only top-36 regions.

Model	Overall Accuracy
BTUP [27]	82.91
QCG [29]	82.05
BAN [30]	84.81
RN [31]	84.61
DFAF [34]	85.55
RAMEN [36]	86.86
MLIN$\star$ [35]	**87.60**
CQ-VQA	87.52

are of size $d_q = 1024$. For attention module, both visual features $\mathbf{v}_i$ $(i = 1, \dots k)$ and question features $\mathbf{f}(\mathbf{q})$) are projected to 1024 dimensional space. These $d_f = 1024$ dimensional vectors are further processed for attention score weighted feature fusion (Subsection III-C).

The TDIUC dataset contains 12 question categories. Thus, the question categorization network $\mathbf{NN_{CQ}}$ predicts the vector $\mathbf{p_q}$ of size $n_c = 12$. Accordingly, one network $\mathbf{NN_{AP}}^{(r)}$ (from $n_c = 12$) is selected to predict the answer $\mathbf{a}$ using $d_f = 1024$ dimensional fused feature $\mathbf{h_a}(\mathbf{q}, I)$. The complete model is trained in an end-to-end manner for 17 epochs with a batch size of 512. The Adamax optimizer [39] 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.

V. Results & Discussion

This section discusses a comparative performance analysis of CQ-VQA and other state-of-art methods (Subsection V-A). We perform an ablation analysis to understand the effectiveness of the proposed model (CQ-VQA). Subsection V-B discusses the results of this analysis.

A. Comparison with State-of-Art Methods

The performances of different VQA methods are compared under two settings. The first setting compares the overall accuracy of all models. There are VQA approaches for which, we do not have access to question category-wise results (not available in the literature). Such models are primarily compared in the first setting. Table II presents the accuracy of different methods. Results shown in bold represents the best performance among all models. The overall accuracy obtained by MLIN [35] and proposed CQ-VQA is comparable. However, it is noteworthy to mention that MLIN (marked with $\star$) has used top-100 regions to extract visual features, while all other models (including CQ-VQA) have used only top-36 regions. As discussed earlier (Subsection III-A), a higher number of region proposals (k) leads to improved performance at the cost of significantly higher computation.

Question category-wise VQA performance of models are compared in the second setting. Here, only those VQA approaches are considered for which such results are available in the literature. Table I shows the question category-wise accuracy of all methods compared in the study. The last three rows represent the comparisons of the three evaluation metrics for all VQA models under consideration.

Table I shows that CQ-VQA is the best performer on all three evaluation metrics. Further, at the category-wise performance, CQ-VQA is the best performer for 10 out of 12 classes. In the other two categories, *sport recognition* and *utility and affordance*, CQ-VQA is the second-best performer. For some question categories, significant performance improvement is obtained by CQ-VQA. For example, CQ-VQA obtains an improvement of 7% and 14% for 'color' and 'Positional Reasoning' categories, respectively.

B. Ablation Studies

The proposed approach leverages on question categories to solve the VQA problem. An ablation analysis is conducted to show the efficacy of the hierarchical approach of CQ-VQA. In this analysis, a baseline model is constructed by removing

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Question Categorization and *Answer Predictor* components of CQ-VQA. However, the baseline uses the same set of attended and fused features as CQ-VQA. Results shown in Table-III shows a relative improvement of 1.45% by CQ-VQA in terms of overall accuracy. CQ-VQA shows improved performance on the other two evaluation metrics as well.

The effect of language bias prior is commonly observed in VQA. This is analyzed next. In TDIUC dataset, the 'Absurd' category is introduced to test the effect of language prior biases in model performance. Our experiment compares the performance of CQ-VQA under two settings – with and without the 'Absurd' category of questions. Table IV shows a significant drop in performance, indicating that CQ-VQA is also affected by language prior biases.

TABLE III: Ablation analysis: Effect of removing hierarchy from CQ-VQA

Metrics	Baseline	CQ-VQA
Overall Accuracy	86.26	**87.52**
Arithmetic-MPT	70.71	**72.08**
Harmonic-MPT	63.37	**64.45**

TABLE IV: Ablation analysis: Performance of CQ-VQA on the data (except Absurd category samples) trained using with and without 'Absurd' Category samples

Without Absurd			
Metrics	**MCB**	**QTA**	**CQ-VQA**
Overall Accuracy	78.06	80.95	**83.46**
Arithmetic-MPT	66.07	66.88	**68.69**
Harmonic-MPT	55.43	58.82	**61.44**

VI. Conclusion & Future Work

This work presents a novel hierarchical end-to-end model CQ-VQA for the VQA task. CQ-VQA leverages over question categorization to reduce the potential answer search space. Empirical results on the TDIUC dataset indicate that the performance of CQ-VQA is competitive to state-of-art VQA methods.

One of the limitations of the proposed approach is the requirement of explicit question categories. We plan to extend the CQ-VQA model for datasets where ground-truths of question categories are not available. Furthermore, the performance of the CQ-VQA model can be enhanced by using better feature extractors, attention mechanisms, and complex question/answer prediction networks.

Acknowledgements: The authors acknowledge the Department of Biotechnology, Govt of India for the financial support for the project BT/COE/34/SP28408/2018, IITG for MHRD Fellowships to Aakansha, and the anonymous reviewers.

References

[1] 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.

[2] 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. [Online]. Available: https://doi.org/10.1007/s11263-016-0966-6

[3] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[16] 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.

[17] 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.

[18] 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.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[20] 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.

[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.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[22] Z. Wang and S. Ji, "Learning convolutional text representations for visual question answering," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 2018, pp. 594–602.
[23] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1–9.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[25] 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.
[26] 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," *arXiv preprint arXiv:1606.01847*, 2016.
[27] 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.
[28] 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.
[29] 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.
[30] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 1564–1574.
[31] 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.
[32] 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.
[33] H. Noh and B. Han, "Training recurrent answering units with joint loss minimization for vqa," *arXiv preprint arXiv:1606.03647*, 2016.
[34] 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.
[35] 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.
[36] 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.
[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[38] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," *CoRR*, vol. abs/1511.02799, 2015. [Online]. Available: http://arxiv.org/abs/1511.02799
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE