Cascading Top-Down Attention for Visual Question Answering

Weidong Tian*, Rencai Zhou, Zhongqiu Zhao
Key Laboratory of Konwledge Engineering with Gig Data
Hefei University of Technology
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China.
* Corresponding author.
email: wdtian@hfut.edu.cn.

***Abstract*—For solving Visual Question Answering (VQA), we commonly employ images and questions simutansouly to predict answers. Some attention mechansims should be used to foucs on the most valuable information, because there are too much information extracted from images and questions. Top-Down Attention (TDA) is one of the famous attention mechansims. For standard TDA, only important regions of the image associated with the question are highlighted. In this work, we propose a Cascading Top-Down Attention (CTDA) model. CTDA highlights the most important information collected from images and questions by a cascading attention process. First, the key words of the question, associated with the image, are highlighted by using a Question Top-Down Attention (QTDA) . Then, important regions of the image, associated with the question, are highlighted by using of Image Top-Down Attention (ITDA), useless information of the images and questions can be ignored effectively. We evaluate our model on two popular VQA data sets. CTDA obtains better results than standard TDA and the other state of the art models.**

***Index Terms*—Cascading Top-Down Attention, Question Top-Down Attention, Image Top-Down Attention, Visual Question Answering**

I. Introduction

In recent years, VQA [3] has emerged as a prominent multidisciplinary research problem in academia and industry. VQA requires two forms of information: questions and images. The inputs for VQA are images and free-form, open-end natural language questions. VQA's goal is to produce a natural language answer about the inputs. In order to correctly answer a question over an image, the computer needs a deep image understanding through fine-grained analysis and even multiple steps of reasoning. At present, most image captioning methods and VQA are processed through neural networks with visual attention. Attention typically produces a spatial map to highlight the image regions associated with the question, which improves the performance of the overall framework [1], [3], [9], [26], [27].

Most of attention models for VQA in literature focus on the problem of identifying "where to look" or visual attention. Few models focus on "what to listen". Combining the processed images and questions to predict the answers is the most common method [1], [23], [26], [29], [30]. The processed images are more relevant to questions. In fact, the words associated with the answer in the question may be just a few key words. When original questions are combined with images, some irrelevant words may introduce noises and affect final results. Motivated by this observation, we want to highlight the question' s key words to reduce the effect.

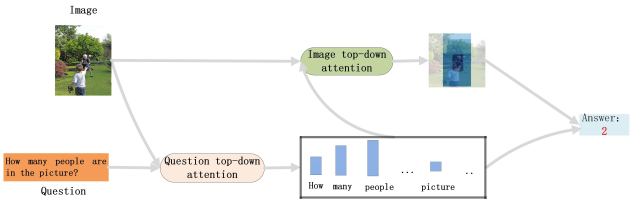

Fig. 1. CTDA sample diagram

In this paper, we have improved the model proposed by the winner of the 2017 VQA Challenge [23]. We propose CTDA to process this problem effectively(Fig 1). CTDA jointly reasons with ITDA and QTDA. Our model uses the pre-trained Glove vector [19] and Gated Recurrent Unit (GRU) [28] to generate the question embedding, and uses Faster Region-based Convolutional Neural Networks (Faster RCNN) to generate the object-centric features from the image. The image features and the question features are fed into attention module to create the embedded features. Finally, the embedded features are fed to a classifier to generate the final answer. In this sense, the image is used to guide the question attention and the question is used to guide image attention. The main contributions of this paper are:

a. We propose a novel model (CTDA) for VQA. CTAD based on question-guided image attention and image-guided question attention;

b. We evaluate our model in VQA v2.0 [9] and v1.0 [3]

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

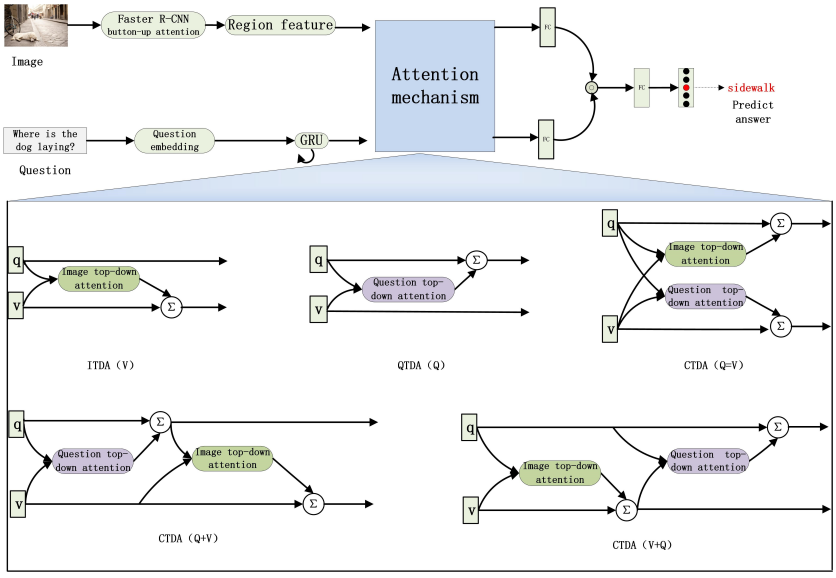

Fig. 2. Variants of the attention in TDA model

data sets. We obtain comparable or even better results than the current state of the art models;

c. We perform contrast experiments and ablation experments to quantify the roles of different components in our proposed model.

II. Related Work

In recent researches, mainstream models use neural networks to combine images and questions for VQA [1], [3], [27]. Typically, these models can be characterized as top-down approaches, where context is provided by a representation of the question in the case of VQA [2], [6], [19], [20], [23], [30], [31]. Anderson et al. [1] firstly used TDA for VQA. In Fast-RCNN [1], [20] and GRU [28], TDA is used to obtain image features and question features respectively.

In previous work, few people paid attention to questions in VQA, but there are also work related to natural language processing (NLP) that has begun to model language attention. Moritz et al. [11] proposed a model based on the attention mechanism to avoid the understanding bottleneck caused by fixed-width hidden vectors in text reading. A more granular attention mechanism was proposed by Tim et al. [21], and the author used a verbatim neural network attention to reason over the implications of the two sentences. Santos et al. [7] proposed a two-way attention to predict and then input the results of the prediction pairing into the public representation space.

The popular VQA models mainly use word embeddings for questions. The features are used to represent words and can be pre-trained on language models [18], [19]. The benets of this pre-training is that one can perform unsupervised pre-training in large corpora and even includes words which not necessarily exist in training questions or answers [8], [12], [22], [24]. VQA is closely related to image captioning [1], [6]. Oriol et al. [26] firstly proposed to extract advanced image feature vectors from the GoogleNet system and generate captions. Oriol et al. [28] proposed bi-directional GRU and Xu et al. [29] proposed bi-directional Long-Short Term Memory (LSTM).

Before image features are inputted to the VQA system, a large amount of other data is used to pre-train the feature extraction models. In this way, when images are inputted to VQA, the features are extracted directly by the pre-training model. Most of VQA models do not use the actual output of the classifier, but use its hidden state to finally represent the image as a group of identified properties and objects to operate on. This is the current method of processing the mainstream of images [8], [17], [29], [31]. Yang et al. [31] proposed a stacked attention network, which ran multiple hops to infer the answer progressively. Kevin et al. [22]generated image regions with object proposals and then selected regions relevant to questions.

The attention in VQA is a method that is consistent with human thinking. Most of questions are only related to certain regions of images. Answering such questions generally requires only local information in images. Excessive image information can cause noise. So most of attentions use question to pay attention to image regions.

Neural network attention has been widely used in different fields of computer vision and natural language processing [6], [27], [28]. Most of methods use the soft attention [4] which was firstly proposed. Soft attention adds a network layer to the network structure for generating soft weights and then use them to calculate weighted averages. The difference between the two main types of soft attention is in input features and candidate features. The first type uses an alignment function based on the input and the ”connection” of each candidate. The second type uses an alignment function based on the input dot product and each dot product.

People are paying more and more attention to the research

of attention, in order to achieve the purpose of multi-step effective reasoning and further filtering redundant information. Yang et al. [31] linearly superimposed multiple attention, and the latter attention is based on the output of the previous attention to achieve multi-step reasoning. Fukui et al. [8] applied two parallel attention to images, which proves that the model can focus on multiple regions of images.

Similar to observing images, humans often focus on some of key words when they understand questions. While performing attention on images, some people have also proposed to pay attention to questions [17]. Through an alternating or parallel attention between questions and images, these models enable more detailed information filtering. Lu et al. [17] proposed draw attention to the image at the three levels of the word, phrase and sentence of the question.

III. Model

In our model (Fig 2), we use the well-known joint multi-model embedding of question and image [12], [25], [26]. ITDA on image features to highlight important images regions, which obtain new image features. QTDA on question features to highlight question's key words, which obtain new question features. Finally, by combining the new image features and the new question features, the score is predicted from the candidate answers.

A. Question Embedding

Whether it's training process or validation process, questions and images are inputted. In order to improve the computational efficiency, the maximum length of questions is adjusted to 14 words, because only about 0.25% of the questions in VQA v2.0 data set exceed 14 words [23]. The given question is encoded as q.

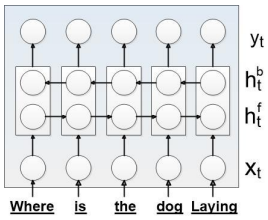

Fig. 3. GRU

$$q = \{q_1, q_2, q_3 ..., q_t\}, q_i \in \mathbb{R}^D \tag{1}$$

where D is the representation vector, the dimension is setted to 300; t is the number of words in the question. After the word is embedded, a sequence size of 14×300 vector is obtained. Words over the length are discarded directly, and those below the length are directly filled with the zero vector. Then bi-directional GRU [28] processes the 14 words. The corresponding hidden state with dimension 1024 is obtained, which is inputted to VQA system as question feature (q).

The GRU (Fig 3) only has two gates, including update gate and reset gate. The update gate is used to control the range, and the status information of the previous moment can enter the current status. The larger the threshold of update gate, the more state information is currently introduced. The reset gate is used to control the degree of ignoring the status information of the previous moment. The smaller the threshold of reset gate, the more it is ignored. The forward propagation of the GRU model can be summarized as the following formula:

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{2}$$

$$\tilde{h_t} = tanh(W_{\tilde{h_t}}[r_t * h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h_t} \tag{4}$$

$$y_t = \sigma(W_o h_t) \tag{5}$$

where [] indicates that two vectors are spliced, W_r,W_z,$W_{\tilde{h}}$,W_o are the weight to be learned, $*$ represents matrix element multiplication.

B. Image Features

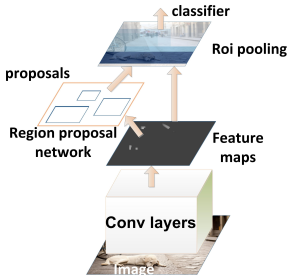

Fig. 4. Faster RCNN

Image features are extracted by Faster RCNN. Faster RCNN is divided into two phases to detect objects (Fig 4). In the first phase, feature map is described as a Region Proposal Network (RPN), which is used to pre-discover the possible locations of the targets in the graph and provided candidate regions for the predicted objects. In the second phase, regions of interest (RoI) are merged and feature maps of each suggestion box are extracted. These feature maps are inputted together to the final layer of the Convolutional Neural Networks (CNN). The output of the final model includes the softmax distribution on the class label and the refinement of each particular bounding box.

Faster R-CNN uses pre-trained ResNet-101 [10] for classification on ImageNet. To learn a better representation of features, model adds an additional training output to predict the attribute class (except the object class). To predict the properties of the region, it connects the average merged convolution features to the learning embedding of the instance object class and feeds it to another output layer. The output layer defines a softmax distribution for each attribute class and "no attribute" class. The input images through bottom-up attention to obtain feature vector (v) of size $K \times 2048$, where K is a plurality of image positions, and then extract feature vector (v) as image features centered on image K object. We select a fixed $K = 36$ in our experimental evaluation.

C. Non-linear layers

Our model uses multiple non-linear transformation layers. In the implementation process, each non-linear transformation layer is activated using gated hyperbolic tangent, and functions are implemented in these layers $f : x \in \mathbb{R}^m \rightarrow y \in \mathbb{R}^n$. The specific parameters are defined as follows:

$$\hat{y} = tanh(w_x + b) \tag{6}$$

$$g = \sigma(w_x^{'} + b^{'}) \tag{7}$$

$$y = \hat{y} \circ g \tag{8}$$

where σ is the sigmoid activation function, $w, w^{'} \in \mathbb{R}^{n \times m}$ and the deviation $b, b^{'} \in \mathbb{R}^n$ are the weights that needs to be trained, $\circ$ is the Hadamard (element-by-element) product g multiplication as the gate that activates y in the middle.

D. CTDA

1) QTDA: Image features (v) are fused with question features $(q_i, i = 1, 2, ..., t, t = 14)$ by element-wise multiplication. They use the non-linear (f_x) and linear layers to obtain scalar attention weights (α) associated with words of the question (Fig 5). Formally:

$$a_i = w_a f_1(f_q(q_i) \circ f_v(v))) \tag{9}$$

$$\alpha = softmax(a) \tag{10}$$

$$\hat{q} = \sum_{i=1}^{t} \alpha_i q_i \tag{11}$$

where w_a needs to be trained, f_x is that the given non-linear transformation. Softmax function is used to normalize the attention weights for all words. It is weighted and summed by normalized values to obtain an attention-grabbing question $(\hat{q})$.

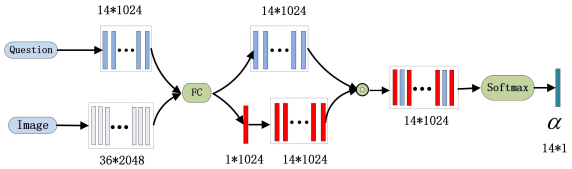

Fig. 5. QTDA

2) ITDA: Question features (q) are fused with image features $(v_i, i = 1, 2, ..., k, k = 36)$ by element-wise multiplication. They pass through the non-linear (f_x) and linear layers to obtain attention weights (β) associated with regions of the image (Fig 6). Formally:

$$b_i = w_b f_2(f_v^{'}(v_i) \circ f_q^{'}(q)) \tag{12}$$

$$\beta = softmax(b) \tag{13}$$

$$\hat{v} = \sum_{i=1}^{k} \beta_i v_i \tag{14}$$

where w_b needs to be trained. Softmax function is used to normalize the attention weights for all locations. It is weighted and summed by normalized values to obtain an attention-grabbing image features$(\hat{v})$.

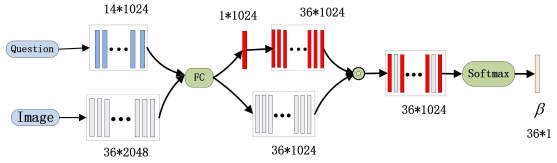

Fig. 6. ITDA

E. Attention Fusion

The image features $(\hat{v})$ are obtained by QTDA and the question features $(\hat{q})$ obtained by QTDA. They are respectively passed through the non-linear layer, and then connected by a simple element-by-element multiplication method:

$$h = f(\hat{q}) \circ f(\hat{v}) \tag{15}$$

The fusion features (h) are called the joint embedding of the question and the image. They are sent to the output classifier for predicting answer.

F. Output Classifier

While the fusion features are classified by the output classifier, VQA is considered as a multi-label classification task. In the output classifier,output vocabulary has a set of candidate answers. In fact,each question is associated with one or more answers. The precision mark in each answer is in range [0,1]. Our multi-label classifier passes the joint embedding the h through a non-linear layer (f), then through a linear mapping (w_o) predicts scores $\hat{s}$ for candidates answers:

$$\hat{s} = \sigma(w_o f(h)) \tag{16}$$

where $w_o \in \mathbb{R}^{N \times 1024}$ is the weight matrix that needs to be trained; σ is the activation function, which is used to normalize the final score to [0, 1]. By using a loss similar to binary cross entropy, we obtain a soft target score. Final stage can be regarded as a logistic regression to predict the correctness of each candidate answer. Our objective function is:

$$L = \sum_{i=1}^{M} \sum_{j=1}^{N} s_{ij} log(\hat{s}_{ij}) - (1 - s_{ij}) log(1 - s_{ij}) \tag{17}$$

The indexes i and j correspond to questions(M) and candidate answers (N). The ground-truth scores (s) are the aforementioned soft accuracies of ground truth answers. The above formulation proved to be much more effective than a softmax classifier as commonly used in other VQA models.

IV. Experiments

A. Evaluation indicators

We use the accuracy metric provided by the VQA Challenge, which is highly reliable for interpersonal differences in expressing answers:

$$Acc(a) = \frac{1}{K} \sum_{k=1}^{K} min(\frac{\sum_{j=1}^{K} {}_{j \neq k} \mathbb{I}(a = a_j), 1}{3}) \tag{18}$$

Among them $(a_1, a_2, a_3, ..., a_k)$ is the correct answer provided by the user,$k = 10$. If there are more than three annotators agreeing on the answer, the answer is believed correct,which is the most intuitive explanation.

B. Data set

TABLE I
OUR RESULTS ON VQA 1.0 VALIDATION SET

Method	Y/N	Num	Other	All
LSTM Q+I [17]	79.8	32.9	40.7	54.3
Image Atten [17]	79.8	33.9	43.6	55.9
HieCoAtt [17]	79.6	35.0	45.7	57.0
CTDA (Q+V)(ours)	82.89	39.08	54.94	63.41

TABLE II
OUR RESULTS ON VQA 2.0 VALIDATION SET

Method	Y/N	Num	Other	All
HieCoAtt [17]	71.80	36.53	46.25	54.57
MCB [17]	77.37	36.66	51.23	59.14
SAA [13]	77.45	38.46	51.76	59.67
FE [15]	80.46	42.80	53.57	62.26
CTDA (Q+V)(ours)	81.26	43.24	55.67	63.65

We evaluate our model on two data sets, in VQA v1.0 and v2.0 data set. VQA data set contains manually annotated questions and answers about the Microsoft COCO data set. There are three subcategories according to the type of answer (including yes/no, number and other). Each question has 10 free-response answers.

VQA v1.0 has 82,783 images, 248,349 questions and 2,483,490 answers for training set; 40,504 images, 121,512 questions and 1,215,120 for validation set. VQA v2.0 has 82,783 images, 443,757 questions and 4,437,570 answers for training set; 40,504 images, 214,354 questions and 2,143,540 answers for validation set.

C. Setup

We build a dictionary by combining all the answers in the training and validation sets. Remove the number of occurrences less than 9 times. Then check if the standard answers to all questions are covered. If not,join them. Finally, an answer dictionary is obtained. For each question, we match all answers in the dictionary with ten answers. If a answer in the dictionary is matched with the annotated answer successfully, we count how often the answer appears in all standard answers. According to the number of occurrences, all an swers in the dictionary are scored by the above evaluation indicators.

We train our model on training data set, report results from the validation set and the test-dev, the test-standard results from the 2019 VQA challenge evaluation server. After that, the weight of the candidate answer to questions is obtained, and then the candidate answer with the highest weight is selected as the predicted answer. Then we get answer and score for each question.

D. Results and analysis

From table I and table II, CTDA (Q+V) can get a level that is comparable or even better than the current popular model [8], [13], [16], [17], [23] in VQA v1.0 and v2.0. In VQA v1.0 data set, the accuracy of CTDA (Q+V) is 9.11%, 7.51%, and 6.41% higher than LSTM Q + I, Image Atten, and HieCoAtt respectively; In VQA v2.0 data set, the accuracy of CTDA(Q+V) is 9.08%, 4.51%, 3.98%, and 1.39% higher than HieCoAtt, MCB, SAA, and FE respectively. Therefore, v the accuracy of CTDA (Q+V) proposed in this paper is far better than these models in VQA v1.0 and v2.0.

Table III and table IV show the performance of our model in VQA v2.0 test-dev set and test-standard set. We trained our model on train set and validation set and tested the performance on test-standard set and test-dev set. On the test-dev, we compared the results of six models: MFB, MFH, FE, MUTAN, MLB and V. Our CTDA (Q+V) model has higher accuracy than the worst MFB model among them 2.47%, higher accuracy than the best FE model among them 1.05%. On the test-standard, we compared the results of ten models: MFB, MFH, FE, MUTAN, MLB, ITDA (V), Prior, Language-only,d-LSTM + nI, and LV-NUS. Our CTDA (Q+V) has higher accuracy than the worst Prior model among them 41.78%, higher accuracy than the best LV-NUS model among them 0.99%.

TABLE III
OUR RESULTS ON VQA 2.0 TEST-DEV SET

Method	Y/N	Num	Other	All
MFB [32]	-	-	-	64.98
MFH [33]	-	-	-	65.80
MUTAN [5]	82.88	44.21	56.50	66.01
FE [15]	82.50	45.80	57.34	66.40
V [23]	81.82	44.21	56.05	65.32
MLB [14]	83.58	44.92	56.34	66.27
CTDA (Q+V) (ours)	83.68	46.48	58.37	67.45

TABLE IV
OUR RESULTS ON VQA 2.0 TEST-STANDARD SET

Method	Y/N	Num	Other	All
MCB [8]	78.82	38.28	53.36	62.27
V [23]	82.20	43.90	56.26	65.67
Prior [9]	61.20	0.36	1.17	25.98
Language-only [9]	67.01	31.55	27.37	44.26
d-LSTM+n-I [9]	73.46	35.18	41.83	54.22
LV-NUS [23]	81.89	46.29	58.30	66.77
UPMC-LIP6 [5]	82.07	41.06	57.12	65.71
FE [15]	82.44	44.93	57.60	66.52
MUTAN [5]	83.06	44.28	56.91	66.38
MLB [14]	83.96	44.77	56.52	66.62
CTDA (Q+V)(ours)	83.91	46.66	58.53	67.76

From the results of the above four tables, it can be seen that our CTDA (Q+V) can reach a level that is comparable to or better than the current mainstream models on both in VQA v1.0 and v2.0. Specifically, we analyze and compare CTDA (Q+V) and the V model. CTDA (Q+V) is a model improved on the basis of ITDA (V). ITDA (V) is the winner of

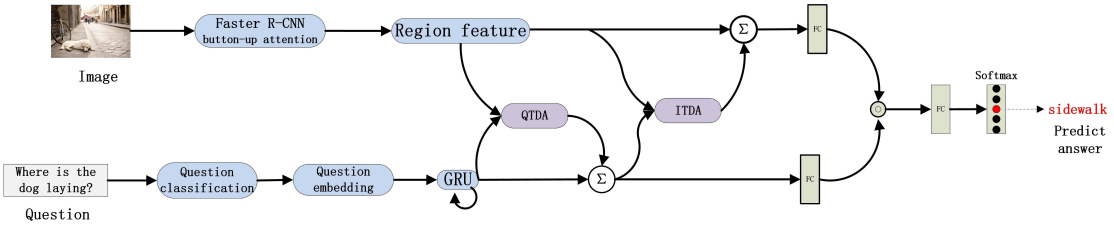

Fig. 7. CTDA (Q+V)

the 2017 VQA Challenge. The model uses quite good results. But the model only focuses on the image on the object that pays attention, ignoring the question on VQA's importance, no consideration is given to the possibility of noise in the input question, which has an impact on the entire VQA.

CTDA (Q=V) is based on the above considerations. In order to prevent some invalid or noisy information from being input into VQA, paying attention to questions and highlighting key words in questions; then paying attention to questions and highlight important regions in images. Remove invalid or noisy information.

In the specific implementation process, we consider that questions are related to images, so we use images to pay attention to questions. From experimental results, the results of CTDA (Q=V) are significantly improved than the results of V, which are 2.13% and 2.09% higher in VQA v2.0 test-dev and test-standard data sets, respectively. The accuracy of CTDA (Q=V) idea is verified. Our model achieves an overall single model accuracy of 67.45% on the test-dev set and 67.76% on the test-standard set, outperforming the best previously reported results by 1.05% and 1.14%.

E. Ablation study

We perform ablation studies to quantify the role of each component in our model. Specifically, we re-train our approach by ablating certain components:

ITDA (V): Applying TDA to images alonely, where no question attention is performed. The model only uses question features to guide images' attention, not pays attention to image-guided question.

QTDA (Q): Applying TDA to questions alonely, where no image attention is performed. The model only uses images features to guide questions' attention, not pays attention to question-guided image.

These models are paid CTDA, include image attention and question attention,but there is a little difference: In CTDA (Q=V), when images and question are paid TDA, images features and question features are original features to guide images' attention and questions' attention.

In CTDA (V+Q), when images and question are paid TDA, the original questions features are used to guide images' attention, then obtain new images features after the attention is applied. The new images features are used to guide questions' attention, then obtain new questions features after the attention is applied.

In CTDA (Q+V), when images and question are paid TDA, the original images features are used to guide questions' attention, then obtain new questions' features after the attention is applied. The new questions features are used to guide images' attention ,then obtain new images features after the attention is applied.

TABLE V
OUR RESULTS ON VQA 1.0 VALIDATION SET

Method	Y/N	Num	Other	All
QTDA(Q)	81.73	38.47	48.13	59.49
ITDA(V) [23]	82.70	38.43	54.03	62.80
CTDA(Q=V)	82.76	36.00	54.57	62.79
CTDA(V+Q)	83.08	37.81	54.47	63.08
CTDA(Q+V)	82.89	39.08	54.94	63.41

TABLE VI
OUR RESULTS ON VQA 2.0 VALIDATION SET

Method	Y/N	Num	Other	All
QTDA(Q)	78.26	41.06	47.99	58.45
ITDA(V) [23]	79.92	42.06	54.68	62.50
CTDA(Q=V)	80.82	41.69	55.21	63.06
CTDA(V+Q)	81.14	42.14	55.41	63.33
CTDA(Q+V)	81.26	43.24	55.67	63.65

Table V shows results on the v1.0 validation set and table VI shows results in VQA v2.0 validation set. It can be seen from the two tables that results of QTDA(Q) and ITDA(V) are relatively poor in both data sets, so only paying attention to images or questions, the characteristics of the other object still contain invalid and noisy information. Among them, since image features information is more than questions features information, it is better to eliminate only noise in images features than to remove only noise in questions features. In v1.0, the accuracy of ITDA (V) is 3.31% higher than QTDA(Q); in VQA v2.0, the accuracy of ITDA(V) is 4.05% higher than QTDA(Q).

Compared with V, On v1.0 validation set, CTDA (Q=V)'s score is reduced 0.01%, the results shows that our model and V model have comparable effects on v1.0.; CTDA (V+Q)'s score is improved 0.28%; CTDA (Q+V)'s score is improved to 0.61%. In VQA v2.0 validation set, CTDA (Q=V)'s score is improved to 0.56%; CTDA (V+Q)'s score is improved 0.83%; CTDA (Q+V)'s score is improved 1.15%. Whether the model is in VQA v1.0 validation set or in VQA v2.0 validation set, the results of CTDA (Q+V) are the best (Fig 7). But the effect

is better on v2.0 than v1.0, because the data set distribution of v2.0 is more reasonable than v1.0.

Images contain more information than questions. When CTDA is paid, the score can be improved a lot. But the order of image attention and question attention influence the final results. For CTDA(Q=V), images guide questions and questions guide images are all using the original data. There is no prior attention to images or questions, and the effect is relatively poor. For CTDA (V+Q), this is equivalent to paying attention to images firstly, due to images contains more content, By paying attention to images, there are residual information. For CTDA (Q+V), this is equivalent to paying attention to questions, which is equivalent to denoising the question first. Due to questions contain less content, the denoising effect is better. The above experimental results prove that this is better than other methods.

The experimental results show that for both images and questions, if only one aspect of the analysis is understood, the accuracy of prediction results will not be high. CTDA understands and analyzes the cascading superposition, and can make more reasonable reasoning. It has a more comprehensive understanding of questions and images, thus makes results more accurate.

V. Conclusion

In this paper, we propose a new VQA model. Our model is evaluated in VQA data set and obtained better results than other models [8], [13], [17], [23]. The studies show that highlighting questions' key words can improve VQA's accuracy, which also provides new ideas for future research in VQA field.

VI. Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos.61672203 & 61976079), Anhui Natural Science Funds for Distinguished Young Scholar (No.170808J08) and AnHui Province Key Laboratory of Dynamic Digital Publishing of Educational Resources.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler,et al, 'Bottom-up and top-down attention for image captioning and visual question answering', in CVPR, (2018).

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell,et al, 'Deep compositional question answering with neural module networks', in CoRR, volume abs/1511.02799, (2015).

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, et al'Vqa: Visual question answering', in ICCV, pp. 2425–2433, (2015).

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', in CoRR, volume abs/1409.0473, (2014).

[5] Hedi Ben-Younes, Rmi Cadene, Matthieu Cord, and Nicolas Thome, 'Mutan: Multimodal tucker fusion for visual question answering', (2017).

[6] Xinlei Chen and C. Lawrence Zitnick, 'Learning a recurrent visual representation for image caption generation', in CoRR, volume abs/1411.5654, (2014).

[7] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou, 'Attentive pooling networks', in CoRR, volume abs/1602.03609, (2016).

[8] Akira Fukui, Dong Huk Park, Daylen Yang, et al'Multimodal compact bilinear pooling for visual question answering and visual grounding', in EMNLP, (2016).

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, 'Making the v in vqa matter: Elevating the role of image understanding in visual question answering', in CVPR, pp. 6325–6334, (2017).

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in CVPR, pp. 770–778, (2016).

[11] Karl Moritz Hermann and Tom, 'Teaching machines to read and comprehend', in NIPS, (2015).

[12] Allan Jabri, Armand Joulin, and Laurens van der Maaten, 'Revisiting visual question answering baselines', in ECCV, (2016).

[13] Vahid Kazemi and Ali Elqursh, 'Show, ask, attend, and answer: A strong baseline for visual question answering', in CoRR, volume abs/1704.03162, (2017).

[14] Jin Hwa Kim, Kyoung Woon On, Jeonghee Kim, Jung Woo Ha, and Byoung Tak Zhang, 'Hadamard product for low-rank bilinear pooling', in ICLR, (2016).

[15] Yuetan Lin, Zhangyang Pang, Donghui Wang, and Yueting Zhuang, 'Feature enhancement in attention for visual question answering', in IJCAI, (2018).

[16] Yuetan Lin, Zhangyang Pang, Donghui Wang, and Yueting Zhuang, 'Feature enhancement in attention for visual question answering', in IJCAI, (2018).

[17] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, 'Hierarchical question-image co-attention for visual question answering', in NIPS, (2016).

[18] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', in CoRR, volume abs/1301.3781, (2013).

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'Glove: Global vectors for word representation', in EMNLP, (2014).

[20] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', in IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, pp. 1137–1149, (2015).

[21] Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tom¨ as´ Kocisky, and Phil Blunsom, 'Reasoning about entailment with neural´ attention', in CoRR, volume abs/1509.06664, (2015).

[22] Kevin J. Shih, Saurabh Singh, and Derek Hoiem, 'Where to look: Focus regions for visual question answering', in CVPR, pp. 4613–4621, (2016).

[23] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel, 'Tips and tricks for visual question answering: Learnings from the 2017 challenge', in CVPR, (2018).

[24] Damien Teney, Lingqiao Liu, and Anton van den Hengel,'Graphstructured representations for visual question answering', in CVPR, pp. 3233–3241, (2017).

[25] Damien Teney and Anton van den Hengel, 'Zero-shot visual question answering', in CoRR, volume abs/1611.05546, (2016).

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 'Show and tell: A neural image caption generator', in CVPR, pp. 3156–3164, (2015).

[27] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengeln, 'Visual question answering: A survey of methods and datasets', in Computer Vision and Image Understanding, volume 163, pp. 21–40, (2017).

[28] Caiming Xiong, Stephen Merity, and Richard Socher,'Dynamic memory networks for visual and textual question answering', in ICML, (2016).

[29] Huijuan Xu and Kate Saenko,'Ask, attend and answer: Exploring question-guided spatial attention for visual question answering', in ECCV, (2016).

[30] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, 'Show, attend and tell: Neural image caption generation with visual attention', in ICML, (2015).

[31] Zichao Yang, Xiaodong He, Jianfeng Gao, et al 'Stacked attention networks for image question answering', in CVPR, pp. 21–29, (2016).

[32] Yu Zhou, Jun Yu, Jianping Fan, and Dacheng Tao, 'Multi-modal factorized bilinear pooling with co-attention learning for visual question answering', in IJCAI, (2017).

[33] Yu Zhou, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao, 'Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering', in IEEE Transactions on Neural Networks Learning Systems, pp. 1–13, (2017).