Question-Driven Multiple Attention(DQMA) Model for Visual Question Answer

Jinmeng Wu
Hubei Key Laboratory of Optical
Information and Pattern Recognition,
Wuhan Institute of Technology.
jinmeng@wit.edu.cnm

Fulin Ge
Hubei Key Laboratory of Optical
Information and Pattern Recognition,
Wuhan Institute of Technology.
gefulin220@163.com

Pengcheng Shu
Hubei Key Laboratory of Optical
Information and Pattern Recognition,
Wuhan Institute of Technology.
lazycat7676@163.com

Lei Ma
Hubei Key Laboratory of Optical
Information and Pattern Recognition,
Wuhan Institute of Technology.
leima@wit.edu.cn

Yanbin Hao
College of Information Science and
Technology, University of Science and
Technology of China
haoyanbin@ustc.edu.cn

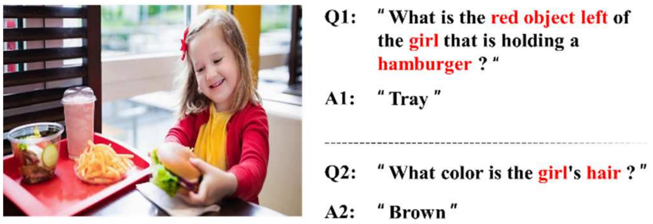
2022 International Conference on Artificial Intelligence and Computer Information Technology (AICIT) | 978-1-6654-5087-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/AICIT55386.2022.9930294

***Abstract*—Visual Question and Answer (VQA) refers to a typical multimodal problem in the fields of computer vision and natural language processing, which aims to give an open-ended question about an image that can be answered accurately. The currently existing visual question answer models inevitably introduce redundant and inaccurate visual information when exploring the rich interaction between complex image targets and texts, and they also fail to focus effectively on the targets in the scene. To address this problem, the Question-Driven Multiple Attention Model (QDMA) is proposed. Firstly, Faster R-CNN and LSTM are used to extract visual features of images and textual features of questions. Then we design a question-driven attention network to obtain question regions of interest in images so that the model can accurately target relevant targets in complex scenes. To establish intensive interaction between the image region of interest and the question word, the co-attentive network consisting of self-attentive and guided-attentive units is introduced. Finally, the correct answer is obtained by inputting question features and image features into an answer prediction module consisting of two-layer Multi-Layer Perceptron. On the VQA2.0 dataset, the suggested method is empirically compared with other methods. The results reveal that the model outperforms other methods, demonstrating the usefulness of the framework.**

***Keywords*—visual question and answer, attentional mecha nisms, feature fusion, multimodal fusion**

I. Introduction

Different ways of perceiving things can capture their complementary information in a variety of ways. Therefore multimodality builds a bridge between vision and language processing, and multimodal learning is widely used in visual language tasks, including image description[1], image-text matching[2], and visual question answer[3,4,5]. However, VQA appears to be more challenging compared to other multimodal tasks, which emphasizes the interaction between vision and language. Early VQA models[6] extensively adopted entity-level alignment methods, which refer to the simple fusion between textual features and target features, and these attempts aimed at cross-modal fusion but lacked understanding of images and texts. For example, given an image of a girl holding a cup to answer the question "What is on the right hand of the girl in the image?", the model loses critical positional information in the image and semantic information in the question context, thus failing to determine the interrelationships between objects in the local region. To better improve the cross-modal integration between vision and language and to make the results more accurate, introducing a visual attention mechanism is an effective way to optimize the VQA task.



Q1: " What is the red object left of the girl that is holding a hamburger ? "
A1: " Tray "
Q2: " What color is the girl's hair ? "
A2: " Brown "

Fig. 1. This is a case where different questions focus on the same image differently, and different types of questions focus on different areas of the image.

The attentional mechanism is actually an integral part of almost all VQA methods [7,8]. It is proposed to introduce attention to the question paper with the image, which could improve the accuracy of identifying the image region associated with the question. But single attention does not correctly focus on the region where the answer is located.

To solve this problem, sense co-attention networks [9,10] have been proposed, and these methods can simply interact with each target region in the image with the words in the question. The results show that such models focuses more on targets in scenes related to textual contexts. However, a large amount of irrelevant information is introduced at the same time, as shown in Figure 1. The first question involves multiple target objects in the image as well as the positional and semantic information between the objects, but the second question only needs to focus on the semantic information about the girl's hair.

Therefore, in order to overcome the information redundancy brought by the complex scene images when answering different questions, suppress irrelevant information and reduce the influence of noise from areas of the regions that are not relevant to the text, we propose the Question-Driven Multiple Attention Model. Firstly, multimodal visual and textual features are generated by Faster R-CNN[11] and LSTM [12] respectively, and then these two features are input into the attention module. The question representation from LSTM is used to predict the attention distribution of different parts of the image. This module selects the regions of interest that are relevant to the question in the image of complex scenes, which can diminish the significance of the visual representation that is irrelevant to the answer and reduce its impact on the subsequent fusion inference. Then, two units of self-attention (SA) and guided attention (GA) are introduced [13], and the modular combination of the two units enables the intensive interaction between regions of interest and question words. Finally, the visual features and questions of relational perception are

978-1-6654-5087-4/22/$31.00©2022 IEEE

embedded into the multi-modal fusion module to get the final answer.

The VQA2.0 dataset is used to train and evaluate our model. Experiments on the QDMA answer generation model demonstrate that our method outperforms several recent challenging baselines. This study proves the effectiveness of the framework.

II. Method

Given a picture and a collection of related questions, a VQA model needs to understand the content associated with different questions in the question collection and obtain answers by reasoning about the semantic relationships between individual objects. At this stage, the number of answers in VQA datasets is often limited, and most of the studies generate answers by solving the task as a classification task. Thus, the visual quiz task can be described as:

$$\hat{y}=\operatorname{argmax}_{y\in Y} P(y \mid I,Q,\vartheta) \tag{1}$$

where $\hat{y}$ denotes the answer predicted by the model and Y is a dictionary of candidate answers. I,Q,ϑ denote the image, question set and model parameters respectively and P representing the VQA model. We illustrate DQMA model in Fig. 2.

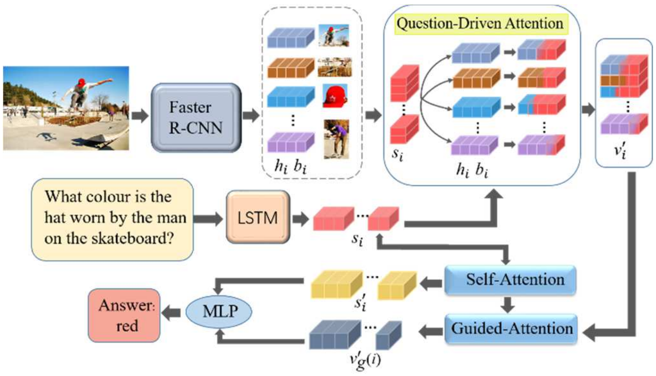


Fig. 2. Illustration of overall architecture of proposed Question-Driven Multiple Attention Model (DQMA). The more red gradients in the Query-Driven Attention model, the greater the importance of the target in the graph to the question.

A. Question and Image Feature Extraction

We are given the question tokens $[\omega_1,\omega_2,...,\omega_m]$, where m is the length of question in words. The question feature extraction module uses Glove[14] pre-training model for word vector embedding to obtain the initial embedding vector, which is then input to the LSTM network for encoding to obtain the feature vector of the question, as in (1).

$$\boldsymbol{S}=[\mathrm{s}_1,s_2,...,s_m]=LSTM(\omega_1,\omega_2,...,\omega_m) \tag{2}$$

The visual feature extraction module applies Faster R-CNN to detect the target objects in the image and to extract the visual features of the image's salient regions. So the object proposal features $\boldsymbol{H}=[h_1,h_2,...,h_K]$ of the input image can be obtained, where $h_i \in R^{d_h}$, d_h represents the visual feature dimension of the candidate frames, and K represent the number of region proposals in each image. The spatial location coordinates of each candidate frame are $\boldsymbol{B}=[b_1,b_2,...,b_K]$, where $b_i=[x_i,y_i,w_i,h_i]\in R^4$. Thus, the visual feature representation $v_i \in R^{d'}$ consists of object proposal features and corresponding bounding box coordinates. Question features $s_i \in R^d$ and visual features $v_i \in R^{d'}$ for subsequent attention module.

B. Question-Driven Attention Model

After receiving the visual features v_i and the textual features s_i in the previous section, due to the complexity of the scenes, not all the target objects in the images are relevant to the question, so the model uses the attention mechanism to obtain the regions of interest in the problem in order to capture the target more precisely in the complex scenes.

The major target regions in the image are highlighted by feeding the visual and linguistic aspects into the attention model. We first calculate the attention weights of the problem and visual features, which will generate the attention scores of different image regions generated by the problem variation, and then the attention scores can be used to focus more on the most relevant regions for a given different problem, denoted as:

$$a_i=\boldsymbol{W}_a^T\tanh(\boldsymbol{W}_1 s_i+\boldsymbol{W}_2 v_i) \tag{3}$$

Where $\boldsymbol{W}_a^T\in R^{1\times f},\boldsymbol{W}_1\in R^{f\times d},\boldsymbol{W}_2\in R^{f\times d'}$ are learnable weight matrix. a_i represents the attention scores generated by the question for the image features. After calculating the relevance score a_i, they are normalized to obtain the attention weights $a_i\in[0,1]$, and the target region features v_i' of the adaptive contextual information can be calculated from the weighted average of the input features v_i and attention weights a_i, as in (4).

$$v_i'=v_i a_i'=softmax(a_i)v_i=\frac{exp(a_i)}{\sum_{k=1}^{K}exp(a_i)}v_i \tag{4}$$

In this module, attention mechanisms are used in directing visual attention to task-relevant regions as input to subsequent collaborative attentive units SA and GA.

C. Collaborative Attention Networks

The co-attentive module [13] consists of a self-attentive unit SA and a guided attention unit GA, which are used to process the multimodal input features of VQA. The calculation method of attention weight is shown in Eq. 5

$$Att(q,\boldsymbol{K},\boldsymbol{V})=soft\max(\frac{q\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{5}$$

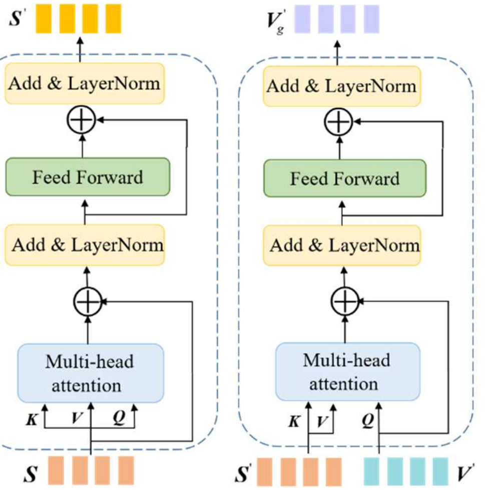
Where, a query $\boldsymbol{q}\in R^{1\times d}$, a key matrix $\boldsymbol{K}\in R^{n\times d}$ and a value matrix $\boldsymbol{V}\in R^{n\times d}$. Multi-headed attention is used to determine whether the model pays attention to distinct representation subspaces at different points.

$$head_i=Att(q\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{KW}_i^{\boldsymbol{K}},\boldsymbol{VW}_i^{V}) \tag{6}$$

$$multihead(q,\boldsymbol{K},\boldsymbol{V})=\boldsymbol{W}_0[head_1,head_2,...,head_h] \tag{7}$$

Where, $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K, \boldsymbol{W}_i^V \in R^{d\times d_h}$ denotes the h head mapping matrix. $\boldsymbol{W}_0 \in R^{h\times d\times d_h}$ is the weight matrix of the spliced attention features.

The question features are fed to SA to capture long-range dependencies, and the essential words of the issue are localized by learning the internal feature linkages of the sentences to generate a differentiated question context representation $\boldsymbol{S}' = [s_1', s_2', ..., s_m'] \in R^{m\times d}$, as shown in Figure 3(a).



(a) Self-Attention (SA) (b) Guided-Attention (GA)

Fig. 3. C. Collaborative Attention Network. (a) Self Attention block (SA), and (b) Guided Attention block (GA).

The basic construction of the GA unit is identical to that of the SA unit, except the input consists of picture features $\boldsymbol{V}'$ and question features $\boldsymbol{S}'$ of varying size, as shown in Figure 3(b). The sample is learned by $\boldsymbol{S}'$ guiding $\boldsymbol{V}'$ to obtain $\boldsymbol{V}_g'$ to enhance the interaction between the context and the corresponding region and to improve the closeness between the two modalities.

D. Answer Prediction Module

After learning attention, the output image features and textual features contain attentional information about the question words and visual areas, the probability y of the candidate answer is calculated using a fully connec ted layer and a softmax activation layer consisting of Multilayer Perceptron (MLP).

$$y = soft\max(MLP(v_g', s_i')) \qquad (8)$$

Where y is the predicted answer probability, from which the model selects the label with the highest probability as the final predicted answer.

III. EXPERIMENT

A. Datasets and Metrics

The most used dataset for VQA tasks is the vQA2.0 dataset. There are 204,721 images in the dataset, in total. 123,287 of these images are used for training and validation of the model, and 81,434 of them are used for the test set. Additionally, there are a total of 369,861 problems in the training and validation sets, and 244,302 problems in the test set, with an average of three problems per image. An average of three questions were created for each image, and they were split into four categories: "Overall," "Yes/No," "Number," and "other". In the VQA tasks, an answer is considered to be correctly predicted as long as the predicted answer is the same as 3 or more of the 10 humans provided answers:

$$Acc(ans) = (1, \frac{\# humans\ provided\ ans}{3}) \qquad (9)$$

B. Training Detail and Parameter Settings

To assess the efficacy of our methodology, we train the DQMA model according to the following experimental protocol. The object proposals and visual features are generated using Faster-RCNN. The maximum value K of the number of target areas is 100. A single-layer LSTM network is used to acquire the question features after the input issue has been broken down into tokens and up to 14 words have been intercepted. Eight heads make up the multi-head attention system and the multi-head have a 512-dimensional latent space. Our network is trained with a maximum of 30 epochs and a batch size of 64. Additionally, we configure the learning rate to vary with epoch by setting it to $\min(2.5te^{-5}, e^{-4})$, where T is the value of epoch at the time of the current training. When 10⩽T, the learning rate begins to gradually decrease by a factor of 0.5 every two epochs.

C. Experimental Results and analysis

To evaluate the validity of the model, the QDMA model in this paper is compared with other VQA models on the VQA2.0 Test-dev set. The selected models include the baseline model and other representative models in recent years, which are the Bottom-up model [15], the BAN model [7], the DFAF model [16], and the MCAoA model [13]. The experimental results are shown in Table 1.

Table 1: Experimental results of comparing the model proposed in the text with other representative models on Test-dev.

Methods	Overall	Other	Yes/No	Number
Bottom-up[15]	65.22	56.01	81.72	44.15
BAN[7]	68.76	59.71	85.36	50.92
DFAF[15]	69.32	59.77	86.77	53.06
MCAoA[13]	70.78	60.78	86.98	53.54
Ours(DQMA)	**71.59**	**61.65**	**87.02**	**53.79**

The experimental results produced by the model suggested in this paper have, as can be shown in Table 1, a more significant improvement than those achieved by other models. The accuracy of the model in this paper on the Test-dev set is 71.59%, 87.02%, 53.79%, 61.65%, respectively, for the four categories: "Overall", "Yes/No", "Number", and "other". The overall accuracy of the DQMA model is increased by 5.64% when compared to the 2017 baseline model Bottom-up; by 1.57%–2.13% when compared to the subsequent BAN and DFAF model that entered the attention mechanism; and by 0.81% when compared to the recently proposed baseline model.

D. Datasets and Metrics

In this section, we visualize the qualitative resultIn this section, we visualize the qualitative results of predicted answers and attention distribution employing our DQMA model on the VQA2.0 dataset, as shown in Figure 4. Each row in the figure corresponds to the prediction results of the DQMA model for different questions on the same image. Effective attention mechanism enables the DQMA model to comprehend difficult questions, locate the target item in the picture that is relevant to the question, and then properly answer the question.

IV. CONCLUSIONS

We propose the Question-Driven Multiple Attention Model (QDMA). The network consists of question and image feature extractor, question-guided attention module, co-attentive module, and MLP answer classifier. On VQA 2.0 datasets, our model successfully uses the Multiple Attention module to filter out the essential information in the pictures and lessen the interference of irrelevant information. Our QDMA's present design, meanwhile, is still rudimentary. We intend to more carefully mine the correlation data between questions and images in the network and better match the data between various patterns.

REFERENCES

[1] Chen L, Jiang Z, Xiao J, et al. Human-like controllable image captioning with verb-specific semantic roles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 16846-16856.

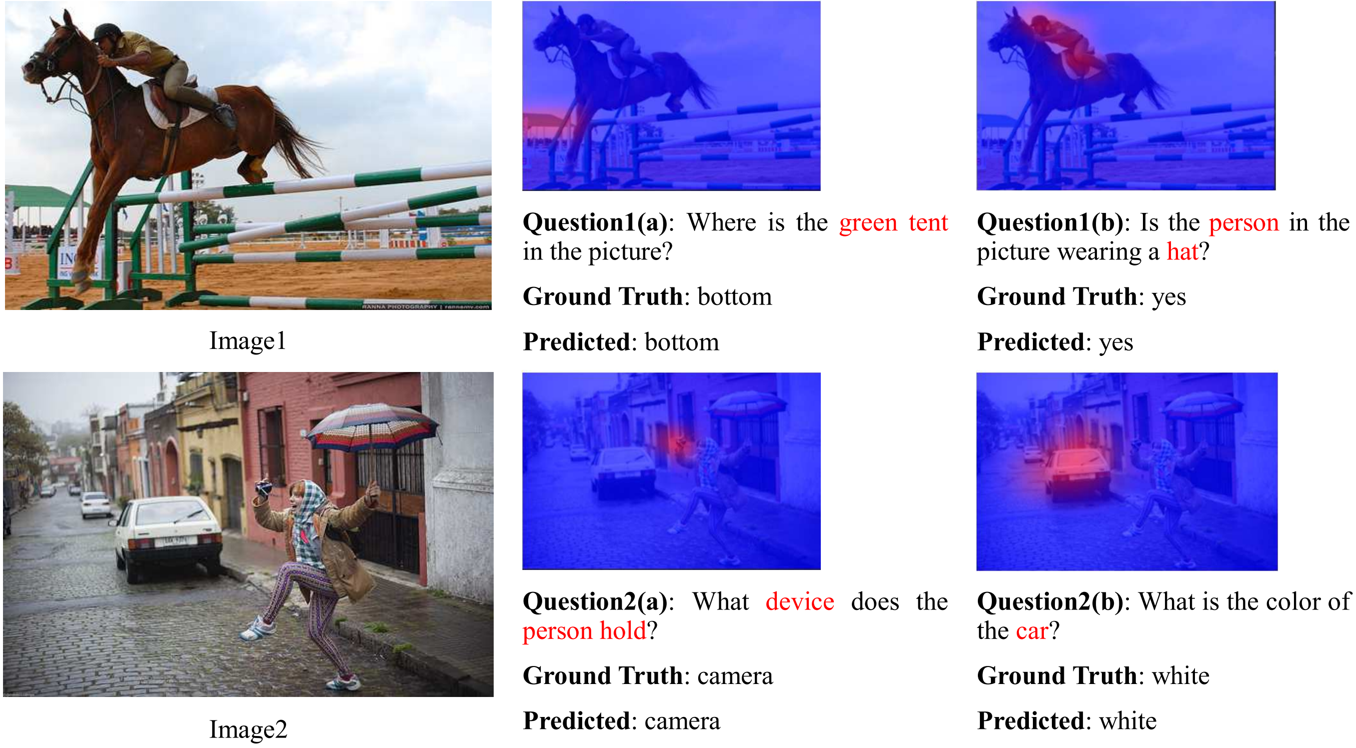


Image1

Question1(a): Where is the green tent in the picture?

Ground Truth: bottom

Predicted: bottom

Question1(b): Is the person in the picture wearing a hat?

Ground Truth: yes

Predicted: yes

Image2

Question2(a): What device does the person hold?

Ground Truth: camera

Predicted: camera

Question2(b): What is the color of the car?

Ground Truth: white

Predicted: white

Fig. 4. Qualitative results of our DQMA network on GQA dataset. The figure demonstrate the distribution of attention when answering two different questions for each of the two pictures.

[2] Litman R, Anschel O, Tsiper S, et al. Scatter: selective context attentional scene text recognizer. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2020: 11962-11972.

[3] Zhu X, Mao Z, Chen Z, et al. Object-difference drived graph convolutional networks for visual question answering. Multimedia Tools and Applications, 2021, 80(11): 16247-16265.

[4] Lu J, Goswami V, Rohrbach M, et al. 12-in-1: Multi-task vision and language representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 10437-10446.

[5] Vickers P, Aletras N, Monti E, et al. In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 468-475.

[6] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4223–4232, 2018.

[7] Kim J H, Jun J, Zhang B T. Bilinear attention networks. Advances in neural information processing systems (NIPS), 2018, 31.

[8] Nguyen D K, Okatani T. 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: 6087-6096.

[9] He S, Han D. An effective dense co-attention networks for visual question answering. Sensors, 2020, 20(17): 4897.

[10] Rahman T, Chou S H, Sigal L, et al. An improved attention for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1653-1662.

[11] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015, 28.

[12] Greff K, Srivastava R K, Koutník J, et al. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[13] Rahman T, Chou S H, Sigal L, et al. An improved attention for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 1653-1662.

[14] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[15] Gao P, Jiang Z, You H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6639-6648.