Multimodal Encoder-Decoder Attention Networks for Visual Question Answering

CHONGQING CHEN1, DEZHI HAN1, (Member, IEEE), AND JUN WANG2, (Senior Member, IEEE)
1School of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
2Department of ECE, University of Central Florida, Orlando, FL 32816-2362, USA
Corresponding author: Dezhi Han (dezhihan88@sina.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672338 and Grant 61873160.

ABSTRACT

Visual Question Answering (VQA) is a multimodal task involving Computer Vision (CV) and Natural Language Processing (NLP), the goal is to establish a high-efficiency VQA model. Learning a fine-grained and simultaneous understanding of both the visual content of images and the textual content of questions is the heart of VQA. In this paper, a novel Multimodal Encoder-Decoder Attention Networks (MEDAN) is proposed. The MEDAN consists of Multimodal Encoder-Decoder Attention (MEDA) layers cascaded in depth, and can capture rich and reasonable question features and image features by associating keywords in question with important object regions in image. Each MEDA layer contains an Encoder module modeling the self-attention of questions, as well as a Decoder module modeling the question-guided-attention and self-attention of images. Experimental evaluation results on the benchmark VQA-v2 dataset demonstrate that MEDAN achieves state-of-the-art VQA performance. With the Adam solver, our best single model delivers 71.01% overall accuracy on the test-std set, and with the AdamW solver, we achieve an overall accuracy of 70.76% on the test-dev set. Additionally, extensive ablation studies are conducted to explore the reasons for MEDAN’s effectiveness.

INDEX TERMS

Computer vision, encoder-decoder attention, multimodal task, natural language processing, question-guided-attention, self-attention, visual question answering.

I. INTRODUCTION

Visual Question Answer is a multimodal learning task that aims at automatically answering a natural language question related to the contents of a given picture. Compared to other multimodal learning tasks (e.g., image captioning [1], [6], [7], image-text matching [8], [12]), VQA is a more difficult and challenging task that requires to capture the high-level interactions between question and image, thus predicting an accurate answer. The performance of VQA has been substantially improved in recent years by designing better feature extractors, better attention mechanisms or better multimodal fusion approaches.

Despite being studied extensively, most existing VQA methods focus on learning fine-grained question features and image features to obtain richer multimodal feature representations. With the development of attention mechanisms in the field of deep learning, it has successfully been applied to the VQA task. Taking the question in Fig. 1 as an example, to correctly answer the question “How many people are playing”, the attention model should focus on specific regions of the image (i.e., the people around the Frisbee). The method based on visual attention was first proposed by Shih et al. [26], and it has become an integral part of the VQA model that requires fine-grained visual understanding [20], [29]. Besides the visual attention methods, researchers have successfully introduced co-attention approaches that focus on both the important regions of image and keywords of question to learn visual attention and text attention [27], [30]. However, the early co-attention methods mostly learn coarse interactions, ignoring the dense interactions between each question word and each image region, which leads to cannot infer the correlation between any question word and any image region. Therefore, the early co-attention methods have a severe limitation.

To overcome the above limitation, two dense co-attention models DCN [28] and BAN [12] have been proposed to model dense interactions between each question and image...
The main contributions of this paper are as follows:

- A novel Multimodal Encoder-Decoder Attention Networks (MEDAN) is proposed, which consists of Multimodal Encoder-Decoder Attention (MEDA) layers cascaded in depth, and can capture rich and reasonable question features and image features by associating keywords in question with important object regions in image.

- Evaluation results on the benchmark VQA-v2 dataset demonstrate that MEDAN achieves state-of-the-art VQA performance. With the Adam solver, our best single model delivers 71.01% overall accuracy on the test-std set, and with the AdamW solver, we achieve an overall accuracy of 70.76% on the test-dev set.

- Exploring the effectiveness of the proposed MEDAN through extensive ablation studies.

The remaining organizational structure of this paper is as follows: the second part introduces the research work related to VQA; the third part introduces the overall framework research and design of MEDAN; the fourth part verifies the effectiveness of MEDAN through some experiments; finally, take a summary.

II. RELATED WORK

With the development and application of deep learning, the research on VQA is more extensive, and the performance of VQA is much better. Here several research methods of VQA are introduced.

A. MULTIMODAL FEATURE REPRESENTATION FOR VQA

The basis of the VQA task is to extract visual features of the image and textual features of the question accurately and effectively. In early VQA methods, the VGG [13] network was commonly used to extract visual features. With the ResNet network proposed by Kaiming et al. [14]. The researchers gradually shift to the ResNet network that performs better than VGG in visual feature extraction. Recently, the bottom-up attention network [1] derived from Faster R-CNN [31] outperforms better than ResNet, which won the VQA challenge 2017. For the extraction of textual features, the LSTM network [15] and the GRU network [16] are commonly used. Additionally, pre-training methods on GloVe [17] or Bert [18] are also applied to obtain better features. Multimodal feature representation plays an important role in improving the VQA performance.

B. MULTIMODAL FEATURE FUSION FOR VQA

VQA task requires an understanding of image and question contents as well as the relation between them. For this reason, in the early stage of VQA methods, some simple joint representation methods [33] between the two modalities are used for multimodal features fusion. To capture the high-level interactions between image and question. A feature fusion method MCB based on bilinear pooling was applied to VQA by Fukui et al. [20]. Subsequently, Kim et al. [34] proposed the MLB method, whose performance is equal to MCB but...
with less weighted parameters. With better performance but less weighted parameters than MCB, The MFB method was proposed by Yu et al. [21]. Then, Yu et al. [30] proposed the MFH method that outperforms the aforementioned methods. Additionally, the method MUTAN was proposed by Ben-Younes et al. [22], indicating that both MCB and MLB are the special cases of this method. Multimodal feature fusion is key to improve VQA performance.

**C. ATTENTION MECHANISM FOR VQA**

The application of attention mechanism enables VQA to automatically ignore the image regions irrelevant to the given question, and choose to focus on the important image regions, thus predicting a correct answer. In the early stage, the attention mechanism was mainly used to focus on regions of the image, which is called visual attention. Yang et al. [32] proposed Stacked Attention Networks with learning the attention on image regions through multiple iterations. By using object regions instead of spatial grids to represent the image region features, the network BUTD was proposed by Anderson et al. [1] to learn the attention on object regions of the image. Later, the co-attention method, which not only requires to learn visual attention on the image but also needs to learn textual attention on the question, was introduced to VQA. Lu et al. [27] proposed a co-attention learning method that alternately learns image attention and question attention. Yu et al. [30] simplified co-attention into two steps, one is self-attention learning of the question, and the other is question-guided-attention learning of the image. The aforementioned co-attention models cannot infer the correlation between any question word and any image region because of ignoring the dense interactions between them. To overcome this problem, two dense co-attention models DCN and BAN have been proposed to model complete interactions between each question and each image region by Kim et al. [12] and Nguyen et al. [28] respectively. But, the disadvantage of DCN and BAN is that their deep models are not much better than their shallow model. Against this disadvantage, several new models based on deep co-attention were proposed by Gao et al. [3], [5] and Zhou et al. [2], [4], which achieve better performance on VQA task.

**III. MULTIMODAL ENCODER-DECODER ATTENTION NETWORKS FOR VQA**

Fig. 2 is the overall flowchart of our proposed MEDAN. Next will specifically introduce three modules of it.

**A. IMAGE AND QUESTION REPRESENTATION**

Fig. 2 (a) is image and question representation.

1) **IMAGE REPRESENTATION**

Following [1], the input image is represented as a set of image region features. These features are extracted from a Faster R-CNN [31] model pre-trained on the Visual Genome dataset [36]. Give an image I, the image features of I are represented as $V \in \mathbb{R}^{k \times 2048}$, where $k \in [10, 100]$ is the number of object regions. For the $i$-th object, it is represented as a feature $v_i \in \mathbb{R}^{2048}$.

2) **QUESTION REPRESENTATION**

Firstly, give a question $Q$ that is tokenized into words. Then, each word of the question is transformed into a vector, and pre-trained on a large-scale corpus by using the 300-dimensional Glove word embeddings [17] to get the final size of words $u \times 300$, where $u \in [1, 14]$ is the number of words in question. Finally, the word embeddings are passed through a one-layer and 512-dimensional LSTM network to obtain the question features $E \in \mathbb{R}^{u \times 512}$. For the $j$-th word, it is represented as a feature $e_j \in \mathbb{R}^{512}$.

$V$ and $E$ can be given as follow:

$$V = \text{FRCNN} (I)$$

$$E = \text{LSTM} (\text{GloVe} (Q))$$

where $\text{FRCNN} (\cdot)$ represents extraction of image features through a Faster R-CNN model, $\text{GloVe} (\cdot)$ represents pre-training by GloVe word embeddings, $\text{LSTM} (\cdot)$ represents extraction of question features by the LSTM network.

In practice, to handle the different number of object regions $k$ and the variable number of words $u$, following [2], both of $V$ and $E$ are filled to their maximum sizes (i.e., $k = 100$, $u = 14$) by zero-padding. Especially, we conduct a linear transformation of $V$ to make its dimension consistent with
the question features. As shown in (3):

$$ V = \text{Linear} (V) $$

(3)

B. MULTIMODAL ENCODER-DECODER ATTENTION

Fig. 2 (b) is Multimodal Encoder-Decoder Attention (MEDA). MEDA is mainly composed of the Encoder module and Decoder module. As shown in Fig. 3. The core of Encoder is \( \text{SA}_E \) unit, and the core of Decoder is \( \text{GA}_V \) unit and \( \text{SA}_V \) unit. No matter \( \text{SA}_E \), \( \text{GA}_E \) or \( \text{SA}_V \), all of them are based on Scaled Dot-Product Attention and Multi-Head Attention in [25].

1) SCALED DOT-PRODUCT ATTENTION AND MULTI-HEAD ATTENTION

As shown in Fig. 4 (a) is Scaled Dot-Product Attention. The input includes queries and keys of dimension \( d_k \) and values of dimension \( d_v \). For the sake of simplicity, both of \( d_k \) and \( d_v \) are set to the same number \( d_{\text{model}} \). Firstly, we compute the dot products of the query and all keys, and divide each by \( \sqrt{d_{\text{model}}} \). Then, applying a softmax function to get the weights on the values.

In practice, to compute the attention function of queries with the number of \( m \) synchronously, we pack them into a matrix \( Q \in \mathbb{R}^{m \times d_{\text{model}}} \). And \( t \) key-value pairs are packed into a matrix \( K \in \mathbb{R}^{t \times d_{\text{model}}} \) and a matrix \( V \in \mathbb{R}^{t \times d_{\text{model}}} \). The attended features \( F \) is obtained as follows:

$$ W = \text{Score}(Q, K) = \text{softmax} \left( \frac{Q K^T}{\sqrt{d_{\text{model}}}} \right) $$

$$ F = \text{Att}(Q, K, V) = W V $$

(4)

where \( \text{softmax} (\cdot) \) represents normalization by a softmax function, \( W \in \mathbb{R}^{m \times t} \) is the weights matrix, \( F \in \mathbb{R}^{m \times d_{\text{model}}} \) represents the attended features.

As shown in Fig. 4 (b) is Multi-Head Attention. To improve the representation capacity of attended features, it allows the model to jointly attend to information from diverse representation subspaces at different positions. The Multi-Head Attention includes \( h \) paralleled heads, and each head is equivalent to an independent Scale Dot-Product Attention operation. We compute the matrix of output features as follow:

$$ F = \text{MHAtt} (Q, K, V) $$

$$ = \text{concat} (h_{\text{head}}_1, h_{\text{head}}_2, \ldots, h_{\text{head}}_h) W^O $$

$$ \text{head}_i = \text{Att} (Q W^Q_i, K W^K_i, V W^V_i) $$

(5)

where \( W^O \in \mathbb{R}^{h_{\text{head}} \times d_{\text{model}}} \) is the projection matrix for all heads, \( \text{concat} (\cdot) \) represents concatenation of all heads, \( W^Q_i, W^K_i, W^V_i \in \mathbb{R}^{d_{\text{model}} \times d_{h}} \) are the projection matrices for \( i \)-th head. Usually, we make \( d_h = d_{\text{model}}/h \).

2) ENCODER AND DECODER

a: ENCODER

Following Fig. 3 (a), each Encoder consists of an \( \text{SA}_E \) unit that is equivalent to a Multi-Head Attention function, and a Feed Forward (FF) network. We regard Encoder as a model to extract fine-grained question features by self-attention learning. The input of question features \( E = [e_1, e_2, \ldots, e_d] \in \mathbb{R}^{d \times 512} \) is further transformed into three matrices: queries matrix \( Q_E \in \mathbb{R}^{d \times 512} \), keys matrix \( K_E \in \mathbb{R}^{d \times 512} \), and values matrix \( V_E \in \mathbb{R}^{d \times 512} \). The calculation of them is as follows:

$$ Q_E = \text{Linear} (E) $$

$$ K_E = \text{Linear} (E) $$

$$ V_E = \text{Linear} (E) $$

(6)

The \( \text{SA}_E \) unit is used to learn the correlation between each word pair \( < e_i, e_j > \) in question, and outputs the question features through self-attention. Especially, residual connection [14] followed by layer normalization [37] is applied to the outputs as follows:

$$ F_{E \rightarrow E} = \text{LN} \left( E + \text{MHAtt} (Q_E, K_E, V_E) \right) $$

(7)

where \( F_{E \rightarrow E} \in \mathbb{R}^{d \times 512} \) is the output features of question, \( \text{LN} (\cdot) \) represents layer normalization.

FF network is composed of two full connection layers with ReLu [37] function, and Dropout [38] is applied to prevent overfitting. Denoting the input for the FF network as \( F_{E \rightarrow E} \)
obtained by $\text{SA}_E$, we get the output $F'_{E \leftarrow E} \in {\mathbb R}^{n \times 512}$. Then using the residual connection and layer normalization to optimize. The attended question features $E' = \text{Encoder } (E) \in R^{n \times 512}$ is given by:

$$F'_{E \leftarrow E} = FF (F_{E \leftarrow E}) = \max (0, F_{E \leftarrow E} W_1 + b_1) W_2 + b_2$$

$$E' = \text{Encoder } (E) = LN (F_{E \leftarrow E} + F'_{E \leftarrow E}) \quad (8)$$

where max (·) represents choosing a larger number between two. The input dimension, hidden dimension and output dimension of the FF network are 512, 2048, and 512, $W_1 \in R^{512 \times 2048}$, $W_2 \in R^{2048 \times 512}$, $b_1 \in R^{n \times 2048}$, and $b_2 \in R^{n \times 512}$ are four parameter matrices.

As shown in Fig. 3, $N$ is the number of layers that Encoder and Decoder can be stacked. We define $\text{MEDA}^n$ (one-layer MEDA) as follows: Firstly, taking the original question features $E^0 = E$ as the input, and output $E^1$ through a one-layer encoder. Then, taking $E^1$ and original image features $V^0 = V$ as the two inputs, and output $V^1$ through a one-layer decoder. Finally, we get question–image features pair $(E^1, V^1)$.

$\text{MEDA}^n$ can be obtained by MEDA that cascaded in depth, where $n$ represents the number of layers that Encoder-Decoder stacked. Denoting the input features for $\text{MEDA}^n$ as $E^{n-1}$ and $V^{n-1}$ respectively, their output features are denoted by $E^n$ and $V^n$, which are further fed to the $\text{MEDA}^{n+1}$ as its inputs in a recursive manner. The formula is as follow:

$$(E^{n-1}, V^{n-1}) \xrightarrow{\text{MEDA}^n} (E^n, V^n) \quad (14)$$

C. IMAGE–QUESTION FEATURE FUSION AND OUTPUT CLASSIFIER

As shown in Fig. 2 (c), after $\text{MEDA}^n$, we get the question features $E^n = [e^n_1, e^n_2, \ldots, e^n_u] \in R^{u \times 512}$ and image features $V^n = [v^n_1, v^n_2, \ldots, v^n_v] \in R^{v \times 512}$. To fusion multimodal features simply, the question features $E^n$ and image features $V^n$ require compressing into $R^{512}$. Firstly, we design an MLP layer that consists of two full connection layers to transform $E^n$ and $V^n$. Then, we compute the attention weights $\alpha ^E$ (or $\alpha ^V$) of question word features $e^n_j$ (or image region features $v^n_j$) through a softmax function. Eventually, the attended question features $\bar{e}^n \in R^{512}$ (or image features $\bar{v}^n \in R^{512}$) can be obtained by summarizing the multiplication of both the attention weights $\alpha ^E$ (or $\alpha ^V$) and relevant question word features $e^n_j$ (or image region features $v^n_j$). Taking $E^n$ as an example, the attended features $\bar{e}^n$ is obtained as follows:

$$\bar{e}^n = \sum_j \alpha ^E_j e^n_j \quad (16)$$

where $\text{MLP} (·)$ represents a linear transformation operation, and its dimensions of the input layer, hidden layer, and output layer are 512, 512, and 1, $\alpha ^E = [\alpha ^E_1, \alpha ^E_2, \ldots, \alpha ^E_u] \in R^u$ are the learned attention weights of words in question. Similarly, we can get the attended image features $\bar{v}^n$ by $V^n$.

After obtaining $\bar{e}^n$ and $\bar{v}^n$, we design a linear multimodal fusion method to get the fusion feature $f \in R^{1024}$. Then, the fusion feature $f$ is projected into a vector $s \in R^{k}$, where L represents the number of most frequency answers in the training set. Finally, a sigmoid function [35] is used to obtain the final classification result $A \in R^2$ as follows:

$$f = LN \left( W^T_e \bar{e}^n + W^T_v \bar{v}^n \right) \quad (17)$$
\[ s = Linear(f) \] (18) 
\[ A = \text{sigmoid}(s) \] (19)

where \( W_f \in R^{512 \times 1024} \) and \( W_v \in R^{512 \times 1024} \) are two linear projection matrixes, and each linear projection operation is equivalent to passing through a linear layer with input dimension of 512 and output dimension of 1024, \( LN(\cdot) \) is used to optimize training. And \( \text{sigmoid}(\cdot) \) is used for classification.

In particular, following [39], we use Binary cross-entropy (BCE) as a loss function to train the classification of \( L \) answers.

Especially, \( \text{Linear}(\cdot) \) that appeared in the aforementioned equations represents a linear layer, which is used for linear transformation. The input and output dimensions of all \( \text{Linear}(\cdot) \) are 512 except for that in (3) with 2048-dim input and 512-dim output, and that in (18) with 1024-dim input and \( L \)-dim output.

IV. EXPERIMENTS

In section IV, we will verify the performance of the proposed new model MEDAN through a series of experiments. In section A, the benchmark dataset VQA-v2 used in our experiment is introduced. In section B, we give a specific description of some hyper-parameters in experiments. In section C, extensive ablation studies for some optional hyper-parameters described in section B (e.g., the number of \( N \) that Encoder and Decoder can be stacked, the number of heads \( h \) in Multi-Head Attention) are conducted. In section D we explore the effectiveness and shortcomings of our proposed model MEDAN through attention visualization. In section E, the performance of our model MEDAN is compared with some State-of-the-Art models.

A. DATASETS

In our experiments, we use the benchmark dataset VQA-v2 [40] for training and evaluation. It consists of images from the Microsoft COCO dataset [41] and human-annotated question-answer pairs for images. Compared to the dataset VQA-v1 [9], VQA-v2 has more annotations and better balance. The dataset VQA-v2 is split into train set (82783 pictures and 443757 question-answer pairs), validation set (40504 pictures and 214354 question-answer pairs) and test set (81434 pictures and 447793 question-answer pairs). Particularly, the test set is divided into a test-dev set of 25% and a test-std set of 75%. Each image contains 3 questions and each question has 10 answers. The answers with the highest frequency are regarded as an accurate answer. Additionally, all questions are divided into three types: Yes/No, Number, and Other.

B. EXPERIMENTAL DETAILS

During our experiments, we define the following hyper-parameters. The input and output dimensions of the Multi-Head Attention are 512, and the number of heads \( h \in \{1, 2, 4, 8\} \). Following the strategy in [39], we set the number of answers with the higher frequency \( L = 3129 \). The number of layers that Encoder and Decoder stacked using the number of layers before \( N \in \{1, 2, 4, 6, 8\} \). Additionally, the dropout in all full connection layers is 0.1. And the Bacthsize for training is set to 64.

To train the MEDAN model, we use two strategies. One is using the Adam solver [23] with parameters \( (\beta_1 = 0.9, \beta_2 = 0.98) \) following [2], and the other is using AdamW solver [24] with same \( \beta_1 \) and \( \beta_2 \). Besides, following [2], we set the same strategy of learning rate. All the models in our experiments are trained up to 14 epochs. In particular, if the accuracy is evaluated on the validation set, only the train set is used for training. If testing on the test set, we use the train set, validation set, and vg set for training, where the vg set comes from the Visual Genome dataset [36]. By using the additional VQA dataset for training, the performance can be improved. The implementation of the proposed model MEDAN is based on Pytorch. The hardware and software used in our experiments are shown in Table 1.

| Hardware                      | NVIDIA Quadra P4000 |
|-------------------------------|---------------------|
| Graphic Processing Unit(GPU)  | 8GB                 |
| RAM                           | 128GB               |
| Operating System              | Linux               |
| Software                       | Anaconda(pytorch3.6) |
| Programming Language          | CUDA10.1            |
|                               | CUDA11.0            |

C. ABLATION STUDIES OF MEDAN

We conduct extensive ablation studies with AdamW solver on the VQA-v2 dataset to explore the effectiveness of variable \( N \) and \( h \) in our model. The results are shown in Table 2. Additionally, we use two different solvers (i.e., AdamW and Adam) to evaluate the performance on the validation set as shown in Fig. 5.

1) DEPTH OF MEDA

Firstly, as shown in Table 2. (a), we explore the effectiveness of the layer numbers \( N \) on a basis of \( h = 8 \). In our best single model, the default number of layers \( N \) is set to 6. During ablation studies, we evaluate the performance of different \( N \in \{1, 2, 4, 6, 8\} \), and the results demonstrate that increasing the number of layers before \( N = 6 \) can steadily improve the performance of our proposed MEDAN. The reason for the improvement may be the use of residual connections [14]. When \( N = 8 \), there is a subtle improvement, but the time of training and parameters will increase a lot. Therefore, considering the overall performance of the model, we choose \( N = 6 \) in our best model.
TABLE 2. Ablation studies of our proposed MEDAN on VQA-v2 Validation set.

| N  | Y/N | Number | Other  | All   |
|----|-----|--------|--------|-------|
| 1  | 82.86 | 47.18  | 57.13  | 65.49 |
| 2  | 83.48 | 48.75  | 57.85  | 66.29 |
| 4  | 84.45 | 48.88  | 58.41  | 66.94 |
| 6  | 84.63 | 49.16  | 58.47  | 67.08 |
| 8  | 84.76 | 49.41  | 58.44  | 67.14 |

2) NUMBER OF HEADS
Secondly, as shown in Table 2. (b), we research the effectiveness of the number of parallel heads $h$ in Multi-Head Attention on a basis of $N = 1$. The default number of heads $h$ is set to 8 in our best model. On the premise that the overall dimension of Multi-Head Attention is 512, we evaluate different $h \in \{1, 2, 4, 8\}$ on the validation set. As the results showed, when $h = 8$ or $h = 16$, we get the best performance. Considering the time of training, we choose $h = 8$ in our best model.

3) TWO DIFFERENT SOLVERS
In our experiments, we use two different solvers to train the model. As shown in Fig. 5 (a), (b), (c), and (d) represent the per-type and overall accuracies respectively. Although the results on the validation set show that the effect of AdamW is slightly better than Adam for training the model, we still used two different solvers to evaluate the performance on the test set, and the final results demonstrate both of them are helpful for our proposed model MEDAN.

D. ATTENTION VISUALIZATION
As shown in Fig. 6, we visualize the final question and image attention learned by (15). Here are four typical examples (i.e., two correctly predicted examples and two incorrectly predicted examples). Each example includes two parts: the image, question (Q), and answer (A) are presented in the left. The learned image attention, question attention, and
prediction (P) are presented in the right. The brightness of regions represents a correlation between themselves and question, and the darkness of words represents their importance in the attention weights. For the correct samples, we can find that the learned question attention is more focused on keywords, and the learned image attention has better a correlation between keywords and relevant image regions. For the incorrect samples, there is the uncertainty of attention learning, sometimes it cannot focus on the real keywords (e.g., ignoring the word ‘looking up’ in the third example), thus focusing on irrelevant image regions, and leading to predict incorrectly. These visualization results will help us make further improvement to the model in the future.

TABLE 3. Results of single model compared with the state-of-the-art models on Zhe test-dev and test-std set.

| Model           | Test-dev | Test-std |
|-----------------|----------|----------|
|                 | Y/N  | Num  | Other | All   | All   |                  |
| BUTD[1]         | 81.82 | 44.21 | 56.05  | 65.32 | 65.67 |                  |
| MFH[33]         | 85.31 | 49.56 | 59.89  | 68.76 | -     |                  |
| BAN[12]         | 85.42 | 50.93 | 60.26  | 69.52 | -     |                  |
| BAN+Counter[12] | 85.42 | 54.04 | 60.52  | 70.04 | 70.35 |                  |
| DFAF[3]         | 86.09 | 53.32 | 60.49  | 70.22 | 70.34 |                  |
| MCAN[2]         | 86.82 | 53.26 | 60.72  | 70.63 | 70.90 |                  |
| MUAN[4]         | 86.77 | 54.40 | 60.89  | 70.82 | 71.10 |                  |
| MEDAN(Adam)     | 87.10 | 52.69 | 60.56  | 70.60 | 71.01 |                  |
| MEDAN(AdamW)    | 87.02 | 53.57 | 60.77  | 70.76 | 70.98 |                  |

E. COMPARISON WITH THE STATE-OF-THE-ART

In Table 3, we compare our MEDAN model with the current state-of-the-art models on the VQA-v2 dataset, where all models are based on a single model. Among them, BUTD [1] is the winner of the VQA challenge 2017. This model proposed the Bottom-Up attention method based on Faster R-CNN for the extraction of visual features. The method is used by all subsequent models. MFH [30] is a better bilinear pooling method at present. The model BAN [12] achieves good performance by stacking 12 layers, and with additional counting module [42], the object counting performance is improved significantly. DFAF [3] and MCAN [2] are the best models based on deep co-attention. Our proposed model MEDAN is also based on deep co-attention. As shown in Table 3, with Adam solver for training, our model is 0.56 and 0.66 points higher than DFAF on test-dev and test-std. Although it is 0.03 points lower than MCAN on test-dev, it is 0.11 points higher on test-std. And with AdamW solver for training, our model is 0.54 and 0.16 points higher than DFAF and MCAN on test-dev. And on test-std, MEDAN is 0.64 and 0.08 points higher than DFAF and MCAN. By using Gated Self-Attention instead of Self-Attention, the model MUAN that based on deep co-attention achieves the highest accuracy. Although the accuracy of our model is a little lower than MUAN, the advantage of our model is still obvious. And we guess that if combing the two models, the accuracy may be further improved, which is very enlightening for my subsequent research. In a word, these comparisons illustrate the adequate effectiveness of our proposed MEDAN.

V. CONCLUSION

In this paper, we proposed a novel model Multimodal Encoder-Decoder Attention Networks (MEDAN) for VQA. The core idea is to design a Multimodal Encoder-Decoder Attention (MEDA) layer, which consists of the Encoder module and Decoder module that can be stacked in depth.
The core of Encoder is a question self-attention (SA$_E$) unit, which is used to model fine-grained question features. Decoder mainly contains a question-guided-attention (GA$_V$) unit and an image self-attention (SA$_I$) unit, aiming to model the fine-grained image region features. By stacking the layers of Encoder-Decoder, our proposed model MEDAN achieves a new state-of-the-art performance of VQA.

REFERENCES

[1] 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 *Proc. CVPR*, Salt Lake, UT, USA, 2018, pp. 6077–6086.

[2] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 6281–6290.

[3] S. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, “Dynamic fusion with Intra- and inter-modality attention flow for visual question answering,” in *Proc. CVPR*, Long Beach, CA, USA, Jun. 2019, pp. 6639–6648.

[4] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, “Multimodal unified attention networks for Vision-and-Language interactions,” 2019, arXiv:1908.04107. [Online]. Available: http://arxiv.org/abs/1908.04107

[5] P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, “Multi-modality latent interaction network for visual question answering,” 2019, arXiv:1908.04289. [Online]. Available: http://arxiv.org/abs/1908.04289

[6] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[7] K. Xu, J. Ba, R. Kiros, K. Cho, C. Aaron Courville, R. Salakhutdinov, S. Richard Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. ICML*, Lille, France, 2015, pp. 2048–2057.

[8] H. Nam, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proc. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 2156–2165.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proc. ICCV*, Santiago, Chile, 2015, pp. 2425–2433.

[10] M. Malinowski and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input,” in *Proc. NIPS*, Montreal, QC, Canada, 2014, pp. 1682–1690.

[11] Z. Zhao, Z. Zhang, S. Xiao, Z. Yu, J. Yu, D. Cai, F. Wu, and Y. Zhang, “Open-ended long-form video question answering via adaptive hierarchical reinforced networks,” in *Proc. IJCAI*, Stockholm, Sweden, Jul. 2018, pp. 3683–3689.

[12] J. H. Kim, J. Jun, and B. T. Zhang “Bilinear attention networks,” in *Proc. NIPS*, Montreal, QC, Canada, 2018, pp. 1567–1574.

[13] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 777–786.

[15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Sep. 1997.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555

[17] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1532–1543.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019.

[19] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, “Compact bilinear pooling,” in *Proc. CVPR*, Boston, MA, USA, Jun. 2016, pp. 317–326.
CHONGQING CHEN is currently pursuing the M.S. degree with Shanghai Maritime University. His research interests include visual question answering and machine learning.

DEZHI HAN (Member, IEEE) received the Ph.D. degree from the Huazhong University of Science and Technology. He is currently a Professor of computer science and engineering with Shanghai Maritime University. His research interests include visual question answering, network security, cloud computing, mobile networking, wireless communication, and cloud security.

JUN WANG (Senior Member, IEEE) received the B.E. degree in computer engineering from Wuhan University, Wuhan, China, in 1993, the M.E. degree in computer engineering from the Huazhong University of Science and Technology, Wuhan, in 1996, and the Ph.D. degree in computer science and engineering from the University of Cincinnati, Cincinnati, OH, USA, in 2002. He is currently a Full Professor of computer science and engineering and the Director of the Computer Architecture and Storage Systems Laboratory, University of Central Florida, Orlando, FL, USA. He has authored over 80 publications in premier journals, such as the IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and leading HPC and systems conferences, such as VLDB, HPDC, EuroSys, ICS, Middleware, FAST, and IPDPS.

Prof. Wang was a recipient of the National Science Foundation Early Career Award, in 2009, and the Department of Energy Early Career Principal Investigator Award, in 2005.