Multi-Channel Co-Attention Network for Visual Question Answering

Weidong Tian, Bin He, Nanxun Wang, Zhongqiu Zhao[1,2,*]
[1]*Key Laboratory of Knowledge Engineering with Big Data*
Hefei University of Technology
[2]*School of Computer Science and Information Engineering*
Hefei University of Technology
Hefei, Anhui, China.
*Corresponding author.
email: z.zhao@hfut.edu.cn.

***Abstract*—Visual Question Answering (VQA) is to reason out correct answers based on input questions and images. Significant progresses have been made by learning rich embedding features from images and questions by bilinear models. Attention mechanisms are widely used to focus on specific visual and textual information in VQA reasoning process. However, most state-of-the-art methods concentrate on fusing the global multi-modal features, while neglect local features. Besides, the dimension is reduced excessively (from $K\times$2048 to 2048) in general visual attention, which causes a mass of visual information loss. In this paper, we propose a novel multi-channel co-attention network (MC-CAN), which integrates multi-modal features from global level to local level. We design different multi-channel attention mechanisms separately for visual (from $K\times$2048 to $M\times$2048) and textual features at different level of integrations. Additionally, we further improve our proposed approach by combining it with the complementary modules such as the MLB and the Count modules. Experiments on benchmark datasets show that our approach achieves better VQA performance than other state-of-the-art methods.**

***Index Terms*—VQA, Multi-Channel Co-Attention Network, Multi-Hierarchical Fusion**

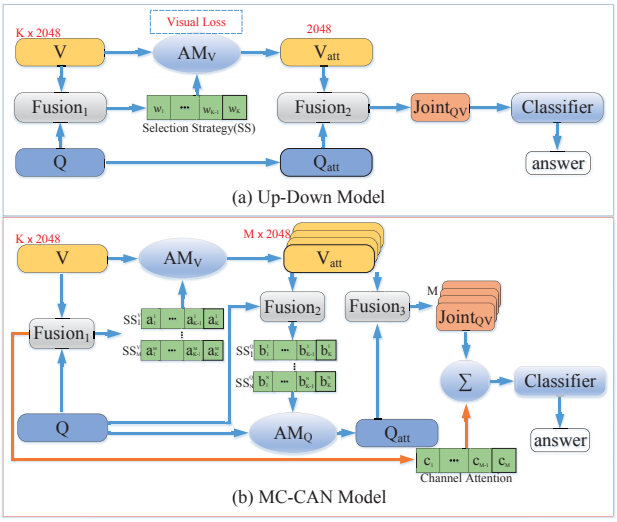


Fig. 1. The Up-Down model [4] **vs**. MC-CAN.

I. Introduction

The Visual Question Answering (VQA) [1] is a task to answer questions which are posed in natural language about images. The answers can be either selected from multiple pre-specified choices or generated by a model. Existing VQA approaches usually consist of three stages: feature representation, feature fusion and answer classification.

Feature representation. Pre-trained ResNet [2] and VGG [3] are commonly used in VQA visual feature extraction. The work in [4] shows that post-processing CNN with region-specific image features such as Faster R-CNN [5] can lead to an improvement of VQA performance. In contrast, Long Short-Term Memory (LSTM) and word embedding [6] are commonly used to generate textual features from either sentence-level or word-level. Lu et al. [7] further proposed to model the question from word-level, phrase-level, and entire question-level in a hierarchical fashion.

Feature fusion. At this stage, visual and textual features are aligned with each other. Except traditional methods such as concatenation, element-wise addition and element-wise product, a large number of new fusion methods (e.g., MCB [8], MLB [9], MUTAN [10], MFB [11]) have recently been proposed for finer-grained integration of multi-modal features. Additionally, attention mechanisms [12], [13] are adopted to force the system to look at informative regions in text or vision.

Answer classification. The most popular method for answer classification is to utilize the integrated image-question features to learn a multi-class classifier which can predict the best-matching answer.

With respect to multi-modal feature fusion, the visual attention mechanism has been widely used in VQA. Teney et al. [4] proposed a bottom-up and top-down (Up-Down) attention model (Fig.1(a)) which enables attention to be calculated at the level of objects and other salient image regions, and obtained the best results in the 2017 VQA Challenge. However, this model only concatenates visual and textual features but neglects the textual attention. Fortunately, bilinear pooling methods and co-attention mechanisms are complementary to this work [4].

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

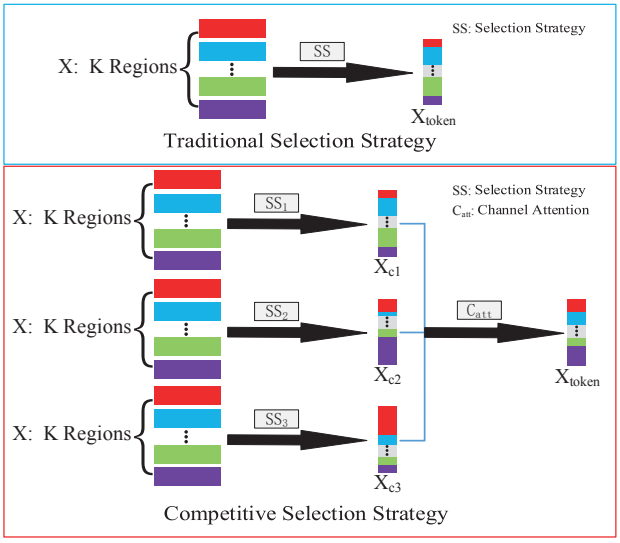


Fig. 2. Comparison between competitive selection (CS) strategy and traditional counterpart for image features screening. Note that X denotes global features, X_{ci} denotes local features, SS_i denotes global screening, and the process with C_{att} denotes local screening.

Most existing VQA approaches, including Up-Down model, just integrate the global original features of images and questions once in visual attention, while neglect the fusion of local features. However, directly using global integrated image-question features may introduce noisy information and lose detail information for given questions. Another problem is that the dimension of image features, with visual attention employed, is reduced excessively (from $K\times2048$ to 2048) (Fig.1(a)), which results in most visual detail information losing.

Thereby, we propose a novel competitive selection (CS) strategy for screening features from different modalities. As illustrated as Fig.2, we compare CS strategy with traditional counterpart in visual features screening. Given image features of K regions X, in contrast to single selection strategy such as region attention used in traditional methods, our proposed CS strategy is divided into two steps, namely global screening and partial screening separately. At first step, a variety of different selection strategies are utilized to select image features from different perspectives. Note that the generated features X_{ci} by i-th selection strategies are similar to the representation of an image channel, so we call X_{ci} i-channel representation of X. At second step, all channel features $\{X_{c1}, X_{c2}, \cdots\}$ are further screened by channel attention C_{att} to obtain the final normalized features X_{token}. Analogously, the CS strategy can also be used to screen textual features.

Based on the CS strategy and co-attention mechanism, we separate the whole multi-modal feature fusion into multiple processes to strengthen the correlation between image and question gradually by different level of selections from global to local. Different from traditional simple global screening for multi-modal features (Fig.1(a)), we firstly implement multiple global screenings, then implement multiple partial screenings, and finally calculate the correlation weights of each part to obtain the normalized fused features for the given question. Furthermore, we propose the multi-channel co-attention network (MC-CAN) (Fig.1(b)), which integrates multi-modal features from global level to local level. For feature screening, we design different multi-channel attention mechanisms separately for visual (from $K\times2048$ to $M\times2048$) and textual features at different level of integrations, which can not only reduce the loss of detail information, but also make the multi-modal integration finer-grained. Note that the "channel" means one of different representations of an object. For instance, we use multiple projections of text features to get multi-channel features.

In summary, the contributions of this paper are as follows:

- We propose a novel competitive selection (CS) strategy for screening multi-modal features, and further propose the multi-channel co-attention network (MC-CAN) for VQA, which integrates multi-modal features from global level to local level.
- We design different multi-channel attention architectures separately for visual and textual features at different level of integrations.
- We further improve our proposed MC-CAN by combining it with the state-of-the-art VQA methods (e.g., MLB [9], Count [14]) and achieve better performance on VQA v2 and VQA-CP v2 benchmarks.

II. Related Work

A. Multi-modal Bilinear Models

Multi-modal feature fusion plays a crucial role in VQA. In early studies, the common frameworks employed simple linear fusion methods such as concatenation, element-wise addition and element-wise product to integrate the multi-modal features. Since the distributions of multi-modal features might vary dramatically, the integrated image-question representations obtained by such linear models can not be sufficiently expressive to fully capture complex associations between visual and textual modalities.

Fukui et al. [8] first introduced the compact bilinear pooling method [15] called Multi-modal Compact Bilinear (MCB) pooling for the multi-modal feature fusion in VQA. To break through the bottleneck of high-dimensional features (e.g. 16000-d) obtained by MCB for computationally complex models, Kim et al. [9] proposed the Multi-modal Low-rank Bilinear (MLB) pooling method, employing Hadamard product of two feature vectors to generate feature vectors with low dimensions, and thereby to produce deep models with fewer parameters. Based on MLB, Yu et al. [11] further proposed the Multi-modal Factorized Bilinear (MFB) pooling, which computes a fused feature with a matrix factorization trick to reduce the number of parameters and to improve the convergence rate. At the same time, Ben-younes et al. [10] proposed the Multi-modal Tucker Fusion (MUTAN), which unifies MCB and MLB into the same framework, and decomposes the weight tensor for bilinear pooling according to the

Tucker decomposition. MUTAN achieves better performance than MLB and MCB with fewer parameters.

To summarize, these bilinear models have greatly improved the interactions between visual and textual modalities. For the same purpose, we design the multi-hierarchical fusion of multi-modal features from global to local. Our fusion model is expected to be complementary to these bilinear methods, And by integrating with them, our method could further boost the VQA performance.

B. Attention Mechanism

Attention Mechanism (AM) was originally proposed to solve language-related tasks, and then became popular in image captioning [12] and VQA [1]. The AM assumes that specific parts of the input (image or question) are more effective than others for VQA reasoning.

Earlier studies mainly considered question-guided attention on image regions. Yang et al. [13] developed stacked attention networks (SANs) which use semantic representation of a question as query to search for the regions in an image which are related to the answer. Kim et al. [16] extended the SANs by incorporating the network into a residual architecture to produce better attention information. The works in [17] and [18] calculate the correlation score for visual features of each bounding box according to textual features. Moreover, in [19], the cross-region relation of image is encoded for properly answering questions which involve complex inter-region relations.

Later studies paid more attention on the opposite orientation called image-guided attention on question features. In [7], a co-attention mechanism was proposed, which employs attention both on image regions and on question words. Considering the sequential consistency of textual words instead of treating each word in a sentence independently, Chao et al. [20] proposed a question attention scheme on the output embeddings, which are generated by a bi-directional LSTM.

Another type of attention is question-guided attention on question features. Yu et al. [21] introduced a multi-level attention network which can reduce the semantic gap by semantic attention and simultaneously benefit fine-grained spatial inference by visual attention. Further, Yu et al. [11] combined this mechanism with a novel multi-modal feature fusion (MFB) of image and question.

Different from other co-attention models, the attention weights of our model are generated by different level integrated features, and we design different multi-channel attention mechanisms for visual and textual features separately. It should be noticed that our attention mechanism is compatible with other attention mechanisms, since our attention is imposed on the channel of the input (image or text), instead of on image regions or question words.

III. Proposed Method

The VQA task requires to provide an answer when given an image $\boldsymbol{v} \in \mathcal{V}$ and a corresponding question $\boldsymbol{q} \in \mathcal{Q}$. Most previous works regard the open-ended VQA as a classification task:

$$arg\max_{\boldsymbol{a}_i \in \mathcal{A}} p_\theta\left(\boldsymbol{a}_i | \boldsymbol{q}, \boldsymbol{v}\right) \tag{1}$$

where θ means the whole set of parameters of the model, and $\mathcal{A}$ denotes the set of candidate answers. These works mainly rely on designing effective strategies to fuse multimodal features, and then learning a MLP classfier to get the predicted answer. However, almost all of existing methods utilize a single global fusion of multi-modal features in attention mechanisms (AM), while overlook the significance of local feature fusion, which has a significant contribution to finer-grained feature integration.

Thereby, we propose a novel competitive selection (CS) strategy for screening multi-modal features, and further propose the multi-channel co-attention network (MC-CAN) for VQA, which integrates different modalities from global level to local level. Fig.3 provides an overview of the architecture of our model. The inputs to our model contain a question and a corresponding image. The visual and textual features are extracted by a convolutional neural network (CNN) and a Gated Recurrent Unit (GRU) respectively. Then our MC-CAN disposes the process of multi-modal feature fusion. Note that visual and textual features are integrated three times at different levels. Additionally, we design different multi-channel attention mechanisms separately for visual and textual features at different levels.

A. Multi-Hierarchical Fusion for VQA

The goal of multi-modal fusion is to explore the interaction between two modalities (vision and text). This projection from different unimodal spaces to a multi-modal one is supposed to extract the relevant correlations between two independent spaces. Besides, a powerful model is expected to have ability to understand full scene, and to focus its attention on relevant visual regions while discarding the useless information regarding the question.

As illustrated in Fig.3, we divide the whole integration process of multi-modal features into three levels (Global, Image Local, Image-Question Local). At the Global level, the original visual and textual features ($\mathcal{V}$ for image, $\mathcal{Q}$ for question) are integrated to produce region attention weights of M channels with the dimension ($K \times M$), namely $\boldsymbol{R}_{att}$ (Fig.5(a)), which is imposed on K image regions M times to obtain the M representations (M channels $\boldsymbol{V}^M$) of region-normalized visual features. At the Image Local level, we firstly calculate the multi-channel textual features $\boldsymbol{Q}^N$ (Fig.5(b)) by projecting $\mathcal{Q}$ N times. Then $\boldsymbol{V}^M$ and $\mathcal{Q}$ are integrated to produce the attention weights of N channels $\boldsymbol{Q}_{att}$ (Fig.5(b)), which are inposed on $\boldsymbol{Q}^N$ to obtain channel-normalized textual features $\widehat{\boldsymbol{Q}}$. At the Image-Question Local level, already provided with attentions imposed on visual regions and on textual channels at the former two levels, we thereby obtain the fused features of M channels $\boldsymbol{f}_{qv}^3$ by integrating $\boldsymbol{V}^M$ and $\widehat{\boldsymbol{Q}}$. Finally, the attention weights of M channels $\boldsymbol{C}_{att}$, which are calculated by $\boldsymbol{R}_{att}$ from the Global level fusion, are further multiplied to the

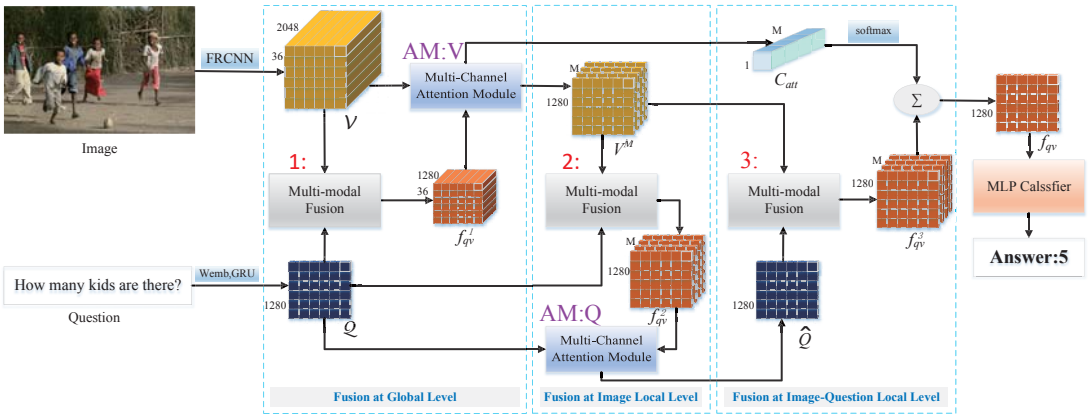
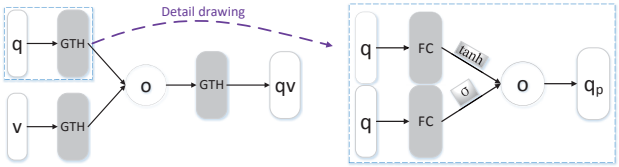


Fig. 3. The architecture diagram of our MC-CAN, where $\mathcal{V}$ and $\mathcal{Q}$ denote the original visual and textual features, $\boldsymbol{f}_{qv}^1$, $\boldsymbol{f}_{qv}^2$, $\boldsymbol{f}_{qv}^3$ denote integrated features at different fusion lecvels respectively, $\boldsymbol{V}^M$ denotes the region-normalized features of M channels, $\widehat{\boldsymbol{Q}}$ denotes the channel-normalized features, and $\boldsymbol{C}_{att}$ denotes the attention weights of M channels, which is imposed on $\boldsymbol{f}_{qv}^3$ to obtain the final channel-normalized fused features $\boldsymbol{f}_{qv}$. Additionally, $AM{:}V$ and $AM{:}Q$ denote the visual and the textual attention modules, which are detailed in Fig.5.



(a) Feature fusion (b) GTH layer

Fig. 4. The fusion strategy for multi-modal features.

M-channel fused features $\boldsymbol{f}_{qv}^3$ to obtain the final normalized integrated features $\boldsymbol{f}_{qv}$.

Feature fusion strategy. We use the MLB as the basis of our fusion strategy, as illustrated in Fig.4(a). Different from the MLB, we utilize the *gated tanh* (GTH) layer (Fig.4(b)), which is proposed in [4], to replace the traditional FC layer, since the GTH has excellent performance of projecting different modal features into common feature space. The GTH implements a function f_a: $\boldsymbol{x} \in \mathbb{R}^m \rightarrow \boldsymbol{y} \in \mathbb{R}^n$ with parameters $a = \left\{\boldsymbol{W}, \boldsymbol{W}^{'}, \boldsymbol{b}, \boldsymbol{b}^{'}\right\}$ defined as follows:

$$\tilde{\boldsymbol{y}} = tanh(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{2}$$

$$\boldsymbol{g} = \sigma(\boldsymbol{W}^{'}\boldsymbol{x} + \boldsymbol{b}^{'}) \tag{3}$$

$$\boldsymbol{y} = \tilde{\boldsymbol{y}} \circ \boldsymbol{g} \tag{4}$$

where σ is the sigmoid activation function, $\boldsymbol{W}, \boldsymbol{W}^{'} \in \mathbb{R}^{n \times m}$ are learned weights, $\boldsymbol{b}, \boldsymbol{b}^{'} \in \mathbb{R}^n$ are learned biases, and $\circ$ denotes the Hadamard (element-wise) product. The vector $\boldsymbol{g}$ acts multiplicatively as a gate on the intermediate activation $\tilde{\boldsymbol{y}}$.

B. Multi-Channel Attention Mechanism

As illustrated conceptually in Fig.5, we design different multi-channel attention mechanisms separately for visual and textual features at different level of integrations. Fig.5(a) shows the module of visual attention for image, and Fig.5(b) shows the module of textual attention for question.

Multi-Channel Visual Attention for Image. Given textual features $\mathcal{Q} \in \mathbb{R}^H$ and a set of spatial visual features $\mathcal{V} \in \mathbb{R}^{K \times 2048}$, the representation of global integrated features at the Global level is given by:

$$\boldsymbol{q}_{prj} = GTH(\mathcal{Q}), \boldsymbol{v}_{prj} = GTH(\mathcal{V}) \tag{5}$$

$$\boldsymbol{f}_{qv_1}^1 = GTH(\boldsymbol{q}_{prj} \circ \boldsymbol{v}_{prj}) \tag{6}$$

where the operation $GTH(X)$ denotes the *gated tanh* layer in Fig.4, which projects X to the specified common feature space for further integration. $\boldsymbol{q}_{prj} \in \mathbb{R}^H$ and $\boldsymbol{v}_{prj} \in \mathbb{R}^{K \times H}$ are the projections of textual and visual features in common feature space respectively. $\boldsymbol{f}_{qv}^1 \in \mathbb{R}^{K \times H}$ is the vector of integrated features, and $\circ$ denotes the Hadamard product. As illustrated in Fig.5(a), the attention weights of K image regions for different visual channels $\boldsymbol{R}_{att} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_M)$ are given by:

$$\boldsymbol{R}_{att} = softmax(FC(\boldsymbol{f}_{qv}^1)) \tag{7}$$

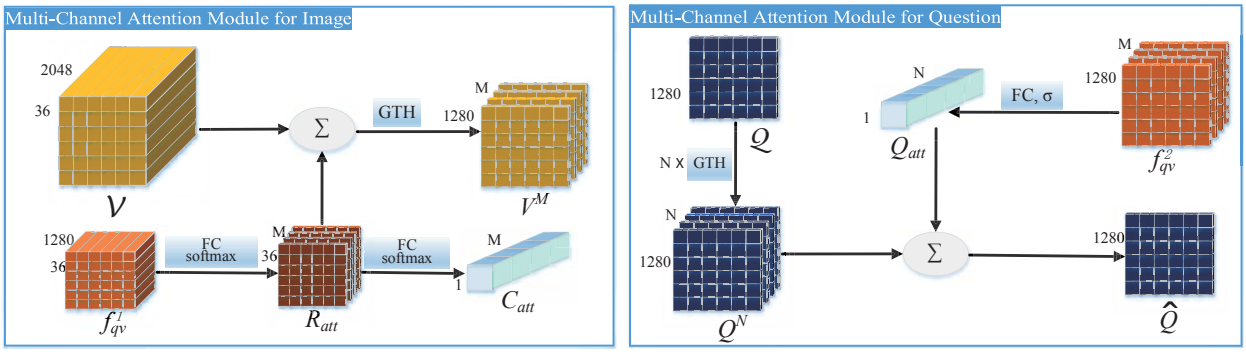
where $\boldsymbol{R}_{att} \in \mathbb{R}^{K \times M}$, M is the number of visual feature channels. The *ith-channel* region-normalized visual features $\boldsymbol{v}_i \in (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M)$, are the multiplication of *ith-channel* visual region attention weights $\boldsymbol{c}_i = (a_{i_1}^r, \cdots, a_{i_K}^r) \in \mathbb{R}^K$ and the original image features $\mathcal{V} \in \mathbb{R}^{K \times 2048}$, with a $GTH(.)$ projecting, as shown by:

$$\boldsymbol{v}_i = GTH(\boldsymbol{c}_i \mathcal{V}), \boldsymbol{V}^M = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M)^T \tag{8}$$

where $\boldsymbol{v}_i \in \mathbb{R}^H, \boldsymbol{V}^M \in \mathbb{R}^{M \times H}$. Additionally, considering the final multi-channel feature integration at the Image-Question Local level, the attention weight of each channel $\boldsymbol{C}_{att} = (a_1^c, \cdots, a_M^c)$ is further calculated by:

$$\boldsymbol{C}_{att} = softmax(FC(\boldsymbol{R}_{att})) \tag{9}$$

where $\boldsymbol{C}_{att} \in \mathbb{R}^M$, which is in charge of focusing on the M channels of the final multi-channel integrated features at the last fusion level.



(a) Attention for image

(b) Attention for question

Fig. 5. Illustration of our multi-channel attention mechanism. Note that the visual attention in (a) and textual attention in (b) are implemented at different fusion levels(refer to Fig.3).

Multi-Channel Textual Attention for Question. Given textual features $\mathcal{Q} \in \mathbb{R}^H$ and region-normalized visual features of M channels $\boldsymbol{V}^M \in \mathbb{R}^{M\times H}$, by Eq.5 and 6, we can get the integrated features $\boldsymbol{f}_{qv}^2 \in \mathbb{R}^{M\times H}$ at the Image Local level. And the attention weights $\boldsymbol{Q}_{att} = (a_1^q, \cdots, a_N^q) \in \mathbb{R}^N$ of N question channels are further obtained by:

$$\boldsymbol{Q}_{att} = \sigma(FC(\boldsymbol{f}_{qv}^2)) \tag{10}$$

where σ is the sigmoid activation function. Generally, the VQA question consists of several words no longer than maxlength (we set it as 14). Though there may be some words useless for VQA inference such as "can", "the", "be", etc, we still argue that these useless words are important parts of the whole textual semantics. Moreover, the attention weight of each question word is difficult to predict. Therefore, we try another way (Fig.5(b)), which projects the question features as N different representations (N-channel question features) $\boldsymbol{Q}^N$ by N $GTH(.)$s, defined as follows:

$$\boldsymbol{Q}^N = (GTH_1(\mathcal{Q}), \cdots, GTH_N(\mathcal{Q}))^T \tag{11}$$

where $\boldsymbol{Q}^N \in \mathbb{R}^{N\times H}$. Then the channel-normalized question representation $\widehat{\boldsymbol{Q}} \in \mathbb{R}^H$ is obtained by:

$$\widehat{\boldsymbol{Q}} = \boldsymbol{Q}_{att}\boldsymbol{Q}^N = \sum_{i=1}^{N} a_i^q GTH_i(\mathcal{Q}) \tag{12}$$

Final Multi-Channel Feature Fusion. Given channel-normalized textual features $\widehat{\boldsymbol{Q}} \in \mathbb{R}^H$ and region-normalized visual features of M channels $\boldsymbol{V}^M \in \mathbb{R}^{M\times H}$, similarly, the integrated features of M channels $\boldsymbol{f}_{qv}^3 = (\boldsymbol{f}_1^3, \ldots, \boldsymbol{f}_M^3)^T \in \mathbb{R}^{M\times H}$ can be calculated at the Image-Question Local level by Eq.5 and 6. As shown in Fig.3, the final normalized fused features $\boldsymbol{f}_{qv}$ can be obtained by:

$$\boldsymbol{f}_{qv} = \boldsymbol{C}_{att}\boldsymbol{f}_{qv}^3 = \sum_{i=1}^{M} a_i^c \boldsymbol{f}_i^3 \tag{13}$$

where $\boldsymbol{f}_i^3$, $\boldsymbol{f}_{qv} \in \mathbb{R}^H, \boldsymbol{C}_{att} = (a_1^c, \cdots, a_M^c)$ is the attention weight of each channel computed by Eq.9. Then the $\boldsymbol{f}_{qv}$ is fed to an MLP classifier including two layers(FC(300)-ReLU-Dropout(0.2)-FC(3129)) to predict the final answer $\boldsymbol{p}_a$:

$$\boldsymbol{p}_a = arg\max_{\boldsymbol{a}_i \in \mathcal{A}}(Classifier(\boldsymbol{f}_{qv})) \tag{14}$$

where $\mathcal{A}$ denotes the set of candidate answers. During the classification procedure, we just select the candidate answer $\boldsymbol{a}_i$ with highest relevance score as the final predicted answer $\boldsymbol{p}_a$.

IV. Experiments

In this Section, we evaluate our proposed approach and compare it with other state-of-the-art methods on two public datasets: the VQA v2 [22] dataset and VQA-CP v2 [23] dataset.

A. Datasets and evaluation metrics

VQA v2. The VQA v2 is the most popular dataset, which balances the answers to each question to minimize the VQA v1 [1] dataset priors. It consists of 1,105,904 questions (443,757 train, 214,354 val and 447,793 test), related to 204,721 images (82,783 train, 40,504 val and 81,434 test). There are three types of questions including Yes/No, $Number$ and $Other$.

As in [4], we choose correct answers appearing more than 8 times in the train set to form the set of candidate answers. But we don't use the additional question and answers from Visual Genome (VG) dataset.

VQA-CP v2. The VQA-CP (Visual Question Answering under Changing Priors) v2 uses the same data from VQA v2, but re-organizes VQA v2 such that the answers to each type of questions have different distributions for train and test sets. For example, "white" might be the most frequent answer to "What color..." questions in the train set, but "black" is the most frequent in the test set. Therefore, a model might perform badly in the test set if it has paid more attention on the distribution of answers to each type of questions in the train set, since there are completely different answer biases between these two splits.

Evaluation metrics. On average, in both VQA v2 and VQA-CP v2, each image is associated with 3 questions, and each

TABLE I

ABLATION STUDIES ON THE NUMBER OF IMAGE CHANNELS FOR VQA. WHERE Avg_{top5} DENOTES THE AVERAGE OF TOP 5 ACCURACIES FOR THE All TYPE IN A SINGLE EXPERIMENT. THE BOLDED MODEL DENOTES THAT ITS CHANNEL NUMBER IS SELECTED AS THE FINAL SETTING (E.G, MC-CAN$(5,1)$ MEANS THAT WE CHOOSE $M=5$ AS THE FINAL SETTING). FOR EACH QUESTION TYPE, THE BEST RESULT IS BOLDED.

MC-CAN(M,N)	VQA v2 val set				
	Yes/No	Num	$Other$	All	Avg_{top5}
Up-Down	81.915	43.827	57.042	64.646	64.621
MC-CAN(1,1)	82.375	44.227	**57.445**	65.074	65.056
MC-CAN(2,1)	82.562	44.326	57.393	65.131	65.112
MC-CAN(3,1)	82.626	**44.875**	57.373	**65.218**	**65.201**
MC-CAN(4,1)	82.690	44.727	57.319	65.196	65.175
MC-CAN(5,1)	82.894	44.503	57.230	65.199	65.193
MC-CAN(6,1)	82.832	44.336	57.222	65.150	65.108
MC-CAN(7,1)	82.764	44.646	57.160	65.134	65.107
MC-CAN(8,1)	82.585	44.428	57.302	65.108	65.096
MC-CAN(16,1)	82.893	44.402	57.128	65.135	65.125
MC-CAN(24,1)	82.822	44.117	57.212	65.112	65.069
MC-CAN(32,1)	82.738	44.322	57.115	65.060	65.037
MC-CAN(48,1)	**82.958**	44.509	56.983	65.102	65.076
MC-CAN(64,1)	82.774	44.562	57.048	65.072	65.049

question is labeled with 10 answers by human annotators. We conduct the evaluation using the index in [1] as:

$$Acc(\boldsymbol{p}_a) = \min(1, \frac{hunmans\ that\ provided\ \boldsymbol{p}_a}{3}) \quad (15)$$

The index indicates that if the predicted answer $\boldsymbol{p}_a$ appears more than or equal to 3 times in human labeled answer list, the accuracy achieves 1.

B. Experimental Settings

We use the recent Up-Down 2048-d features provided in [4] based on Faster R-CNN [5] to represent each image as a set of 36 localized regions. For the question, the words are represented as 300-d embeddings initialized with pre-trained GloVe vectors [6], which are then fed to a GRU to obtain a 1280-d question embedding. For computational efficiency, we restrict the maximum length of each question by selecting the first 14 words.

Our models are optimized with the Adamax optimizer [31] with a batch size of 512 and trained on a 2080Ti GPU. And we set the hidden dimension H of our MC-CAN as 1280. Additionally, we conduct ablation studies to evaluate the impact of channel number (M for image, N for question) on the validation set of VQA v2.

C. Ablation Studies

We compare the performance of our proposed method with the baseline [4] adopting a single global integration for attention. And we further explore the effect of different number of channels (M for image in Tab. I, N for question in Tab. II) on our MC-CAN and every incremental channel just increases model size by 3K. These models are trained on the train set and evaluated on the validation set of VQA v2. It is worth to note that the accuracy of our re-implemented baseline (Up-Down) with MLB fusion applied on the validation set of VQA v2 is 1.4% higher than the performance reported in [4], which

TABLE II

ABLATION STUDIES ON THE NUMBER OF QUESTION CHANNELS FOR VQA. WE SELECT $N=2$ AS THE FINAL SETTING.

MC-CAN(M,N)	VQA v2 val set				
	Yes/No	Num	$Other$	All	Avg_{top5}
MC-CAN(1,1)	82.375	44.227	**57.445**	65.074	65.056
MC-CAN(1,2)	82.674	**44.783**	57.313	**65.194**	**65.139**
MC-CAN(1,3)	82.651	44.161	57.301	65.098	65.039
MC-CAN(1,4)	82.629	44.276	57.297	65.103	65.078
MC-CAN(1,5)	82.548	44.370	57.235	65.054	65.035
MC-CAN(1,6)	82.647	44.361	57.280	65.112	65.085
MC-CAN(1,7)	**82.768**	44.425	57.136	65.095	65.087
MC-CAN(1,8)	82.566	44.004	57.264	65.027	64.975
MC-CAN(1,16)	82.523	44.018	57.182	64.972	64.957

reflects that the MLB fusion has more powerful performance than traditional concatenation. Besides, we average the top 5 accuracies of the All type to evaluate the stability of these results, which reflects the holistic capability of these models to some extent.

The comparison results are shown in Tab.I and Tab.II. For fair comparison, all other variables of these models are controlled to be consistent except the channel number M or N, which is set as 1 when another one is being evaluated.

From Tab.I we can find that:

(1) MC-CAN$(1,1)$ gives better results than the baseline, where $(M=1, N=1)$ means no channel attention. This phenomenon indicates that multi-hierarchical fusion is more effective for the finer-grained interaction of multi-modal features.

Analysis. For Yes/No, larger M and N can both greatly enhance its accuracy such as $(M=48, accuracy=82.958\%)$ and $(N=7, accuracy=82.768\%)$. The reason is that the stacked multi-channel attentions with incremental channels can provide finer-grained information. For Num, a proper M or N benefits its performance such as $(M=3, accuracy=44.875\%)$ and $(N=2, accuracy=44.783\%)$, but larger M or N $(M>3, N>2)$ weakens the effectiveness of our MC-CAN. The reason may be that excessively finer-grained information, which is generally called noise, from excessive channels makes the Num questions reasoning more difficult. For $Other$, our MC-CAN performs worse with the larger M or N, and the model with a single image channel and a single question channel achieves the best accuracy $(M=1, N=1, accuracy=57.445\%)$.

Considering a variety of factors, including different question types, model complexity and stability, and additional validation experiments, we choose $M=5$ and $N=2$ as the final setting, and compare its performance with other state-of-the-art methods.

(2) The multi-channel attention for image improves the performance of MC-CAN, since almost all MA-CAN$(M,1)$ with $M>1$ achieve better performance than MC-CAN$(1,1)$.

(3) A suitable number of image channels can greatly improve the accuracy of VQA (e.g., M from 2 to 8), but an excessively large channel number might be useless (e.g., $M \in \{24, 32, 48, 64\}$).

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON TEST-STANDARD AND TEST-DEV SET OF VQA V2. THE MODELS WITH "†" ARE RE-IMPLEMENTED VERSIONS FROM [24]. THE "-" INDICATES THE RESULT IS NOT AVAILABLE. ALL THE REPORTED RESULTS ARE OBTAINED WITH A SINGLE MODEL WITHOUT MODEL ENSEMBLING.

Model	*test-std*				*test-dev*			
	Yes/No	Num	$Other$	All	Yes/No	Num	$Other$	All
ReasonNet [25]	78.86	41.98	57.39	64.61	-	-	-	-
MUTAN† [10]	83.06	44.28	56.91	66.38	82.88	44.54	56.50	66.01
MLB† [9]	83.96	44.77	56.52	66.62	83.58	44.92	56.34	66.27
QGHC [26]	-	-	-	65.90	83.54	38.06	57.10	65.89
DA-NTN [24]	84.60	47.13	58.20	67.94	84.29	47.14	57.92	67.56
Up-Down [4]	82.20	43.90	56.26	65.67	81.82	44.21	56.05	65.32
VQA-E [27]	83.22	43.58	56.79	66.31	-	-	-	-
QCG [28]	82.91	47.13	56.22	66.18	-	-	-	-
VCT_{REE} [29]	84.55	47.36	59.34	68.49	84.28	47.78	59.11	68.19
MuRel [30]	-	-	-	68.41	84.77	**49.84**	57.85	68.03
MC-CAN(ours)	**85.58**	**47.44**	**60.07**	**69.27**	**85.43**	48.26	**60.21**	**69.24**
MC-CAN+count(ours)	**85.40**	**51.09**	**60.21**	**69.66**	**85.28**	**50.71**	**60.17**	**69.44**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE VALIDATION SET OF VQA V2. THE MODELS WITH "†" WERE TRAINED BY [24]. THE "-" INDICATES THE RESULT IS NOT AVAILABLE. FOR EACH QUESTION TYPE, THE BEST RESULT IS BOLDED.

Model	Yes/No	Num	$Other$	All
ReasonNet [25]	73.78	36.98	54.81	60.60
MUTAN† [10]	81.09	42.25	54.41	62.84
MLB† [9]	81.89	42.97	53.89	62.98
QGHC [26]	-	-	-	60.64
DA-NTN [24]	**83.09**	44.88	55.71	64.58
Up-Down [4]	80.3	42.8	55.8	63.2
VQA-E [27]	80.85	43.02	54.16	63.51
VCT_{REE} [29]	82.6	**45.1**	57.1	65.1
MuRel [30]	-	-	-	65.14
MC-CAN(ours)	82.92	44.93	**57.24**	**65.27**
MC-CAN+count(ours)	83.00	**47.82**	**57.24**	**65.68**

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE VALIDATION SET OF VQA-CP V2. THE MODELS WITH "†" WERE TRAINED BY [32].

Model	$Accuracy(\%)$
MuRel [30]	39.54
Up-Down† [4]	38.01
QCG† [28]	38.32
BAN† [33]	39.31
MAC† [34]	31.96
RN† [35]	36.70
Ans Them All [32]	39.21
MC-CAN(ours)	**40.78**
MC-CAN+count(ours)	**40.66**

Similarly, Tab. II reports the ablation studies on the number of question channels N. Compared with image channels, our MC-CAN is more sensitive to the number of question channels. The reason may be that the structure of question features (no more than 14 words) is simpler than image features (36 bounding boxes), which makes the question features more sensitive to noises with the same number of channels.

D. Comparisons with State-of-the-arts

Tab.III reports the single-model performances of various state-of-the-art methods on both test-standard and test-dev sets of VQA v2 dataset. For fair comparison, all reported methods are trained on the combination of train set and validation set. We can find that our MC-CAN has stable improvements (3.6% for $test\text{-}std$, 3.92% for $test\text{-}dev$) than the baseline [4]. And our approach achieves the best accuracy in almost all question types except the Num type (48.26%, the best is 49.84%) on $test\text{-}dev$. Similarly, we further improve the accuracy of Num type (47.44% to 51.09% for $test\text{-}std$, 48.26% to 50.71% for $test\text{-}dev$) with Count conponent integrated.

Tab.IV reports the performance of our method with published results of the state-of-the-arts on VQA v2 validation set. We analyse Tab.IV as follows:

(1) Compared with the baseline (Up-Down) [4], our MC-CAN improves the performance from 63.2% to 65.27%. Specifically, our proposed approach brings 2.62%, 2.13% and 1.44% improvements in the question types of Yes/No, Num and $Other$, respectively.

(2) Compared with other state-of-the-art methods, our MC-CAN achieves the best score in overall questions (65.27%) and $Other$ (57.24%), and the accuracy of Yes/No and Num types approaches the best.

(3) We further improve our proposed approach by combining it with the Count component [14], which can greatly

improve the performance of Num type by introducing the spatial information of bounding boxes. Obviously, the Num accuracy is improved from 44.93% to 47.82% and the overall performance is further improved to 65.68%.

We also compare our proposed method with other state-of-the-art methods on VQA-CP v2 dataset in Tab.V. All models in Tab.V are trained on the train set and evaluated on the validation set. These results indicate that VQA-CP v2 is more difficult than VQA v2, though it is just re-organized from VQA v2. Obviously, our MC-CAN (40.78%) still outperforms other state-of-the-art methods at least 1.24%. Furthermore, we evaluate the combination of our method and Count component in the same way. However, the accuracy is decreased from 40.78% to 40.66%. The reason of this abnormal phenomenon may be that the Count component improves the Num score by utilizing biases of answer distributions to some extent, which is useless on VQA-CP dataset. Compared with the great improvement (0.41%) on the validation set of VQA v2 (Tab.IV), such a 0.12% decline still shows its powerful performance and stability.

V. Conclusion

In this paper, we propose a novel competitive selection (CS) strategy for screening multi-modal features, and further propose the multi-channel co-attention network (MC-CAN) for VQA. Our approach integrates multi-modal features from global level to local level, and we design different multi-channel attention architectures separately for visual and textual features at different level of integrations. We further improve our proposed approach by combining it with MLB fusion and Count modules. Experimental results on two large VQA datasets show that our proposed model outperforms existing state-of-the-art approaches.

Our future work will focus on combining our MC-CAN with other state-of-the-art methods such as bilinear pooling methods (MCB, MFB and MUTAN). We are also insterested in intergating our multi-channel AM with other AMs. Furthermore, our MC-CAN will be explored on other tasks related to vision and text such as image captioning.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos.61672203 & 61976079) and Anhui Natural Science Funds for Distinguished Young Scholar (No.170808J08)

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] 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 *CVPR*, 2018.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[6] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[7] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS*, 2016, pp. 289–297.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[9] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR*, 2017.

[10] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017.

[11] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[13] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016, pp. 21–29.

[14] Z. Yan, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," in *ICLR*, 2018.

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *CVPR*, 2016.

[16] J. H. Kim, S. W. Lee, D. H. Kwak, M. O. Heo, J. Kim, J. W. Ha, and B. T. Zhang, "Multimodal residual learning for visual qa," in *NIPS*, 2016, pp. 361–369.

[17] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *CVPR*, 2016, pp. 4613–4621.

[18] R. Li and J. Jia, "Visual question answering with question representation update (qru)," in *NIPS*, 2016, pp. 4655–4663.

[19] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *ICCV*, 2017, pp. 1291–1300.

[20] M. Chao, C. Shen, A. Dick, and A. V. D. Hengel, "Visual question answering with memory-augmented networks," in *CVPR*, 2018, pp. 6975–6984.

[21] D. Yu, J. Fu, M. Tao, and R. Yong, "Multi-level attention networks for visual question answering," in *CVPR*, 2017, pp. 4709–4717.

[22] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6904–6913.

[23] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *CVPR*, 2018, pp. 4971–4980.

[24] Y. Bai, J. Fu, T. Zhao, and T. Mei, "Deep attention neural tensor network for visual question answering," in *ECCV*, September 2018, pp. 20–35.

[25] I. Ilievski and J. Feng, "Multimodal learning and reasoning for visual question answering," in *NIPS*, 2017, pp. 551–562.

[26] G. Peng, H. Li, L. Shuang, L. Pan, Y. Li, S. C. H. Hoi, and X. Wang, "Question-guided hybrid convolution for visual question answering," in *ECCV*, 2018, pp. 469–485.

[27] Q. Li, Q. Tao, S. Joty, J. Cai, and J. Luo, "Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions," in *ECCV*, 2018, pp. 552–567.

[28] W. Norcliffe-Brown, E. Vafeais, and S. Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *NIPS*, 2018, pp. 8334–8343.

[29] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *CVPR*, 2019, pp. 6619–6628.

[30] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, "Murel: Multimodal relational reasoning for visual question answering," in *CVPR*, 2019, pp. 1989–1998.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] R. Shrestha, K. Kafle, and C. Kanan, "Answer them all! toward universal visual question answering models," in *CVPR*, 2019, pp. 10 472–10 481.

[33] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NIPS*, 2018, pp. 1564–1574.

[34] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *ICLR*, 2018.

[35] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *NIPS*, 2017, pp. 4967–4976.