Co-Attending Free-Form Regions and Detections with Multi-Modal Multiplicative Feature Embedding for Visual Question Answering

Pan Lu,[†] **Hongsheng Li,**[‡*] **Wei Zhang,**[♮] **Jianyong Wang,**[†] **Xiaogang Wang,**[‡*]

[†] Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science, Tsinghua University
[‡] Department of Electronic Engineering, The Chinese University of Hong Kong
[♮] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University
{lupantech, zhangwei.thu2011}@gmail.com,   {hsli, xgwang}@ee.cuhk.edu.hk,   jianyong@mail.tsinghua.edu.cn

Abstract

Recently, the Visual Question Answering (VQA) task has gained increasing attention in artificial intelligence. Existing VQA methods mainly adopt the visual attention mechanism to associate the input question with corresponding image regions for effective question answering. The free-form region based and the detection-based visual attention mechanisms are mostly investigated, with the former ones attending free-form image regions and the latter ones attending pre-specified detection-box regions. We argue that the two attention mechanisms are able to provide complementary information and should be effectively integrated to better solve the VQA problem. In this paper, we propose a novel deep neural network for VQA that integrates both attention mechanisms. Our proposed framework effectively fuses features from free-form image regions, detection boxes, and question representations via a multi-modal multiplicative feature embedding scheme to jointly attend question-related free-form image regions and detection boxes for more accurate question answering. The proposed method is extensively evaluated on two publicly available datasets, COCO-QA and VQA, and outperforms state-of-the-art approaches. Source code is available at https://github.com/lupantech/dual-mfa-vqa.

1 Introduction

In recent years, multi-modal learning for language and vision has gained much attention in artificial intelligence. Great progress has been achieved for different tasks including image captioning (Karpathy and Fei-Fei 2015), visual question generation (Mostafazadeh et al. 2016; Li et al. 2017b), video question answering (Ye et al. 2017) and text-to-image retrieval (Xie, Shen, and Zhu 2016; Li et al. 2017a). The Visual Question Answering (VQA) (Antol et al. 2015) task has recently emerged as a more challenging task. The algorithms are required to answer natural language questions about a given image's contents. Compared with the conventional visual-language tasks such as image captioning and text-to-image retrieval, the VQA task requires the algorithms to have a better understanding on both the input image and question in order to infer the answer.

State-of-the-art VQA approaches utilize visual attention mechanism to relate the question to meaningful image re-
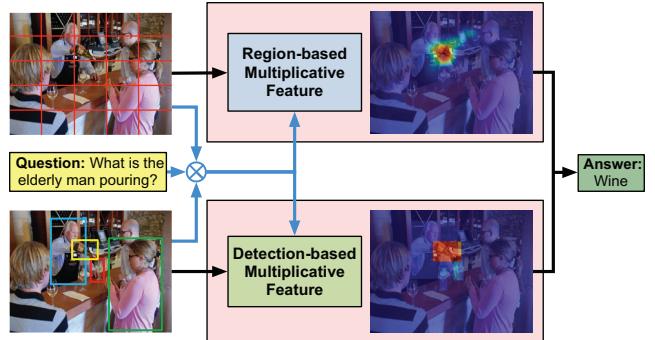

Figure 1: Co-attending free-form regions and detection boxes based on the question, the whole image, and detection boxes for better utilizing complementary information to solve the VQA task.

gions for accurate question answering. Most visual attention mechanisms in VQA can be categorized into free-form region based methods (Lu et al. 2016; Fukui et al. 2016) and detection-based methods (Li and Jia 2016; Shih, Singh, and Hoiem 2016). For the free-form region based methods, the question features learned by a Long Short-Term Memory (LSTM) network and the image features learned by a Convolutional Neural Network (CNN) are fused by either additive, multiplicative or concatenation operations at every image spatial location. The free-form attention map is obtained by applying a softmax non-linearity operation across the fused feature map. Since there is no restriction on the obtained attention map, the free-form attention region is able to attend both global visual context and specific foreground objects for inferring answers. However, since there is no restriction, the free-form attentive regions might focus on partial objects or irrelevant context sometimes. For instance, for an question, "*What animal do you see?*", a free-form region attention map might mistakenly focus on only a part of the foreground cat and generate an answer of "*dog*". On the other hand, for the detection-based attention methods, the attention mechanism is utilized for relating the question to pre-specified detection boxes (e.g., Edge boxes (Zitnick and Dollár 2014)). Instead of applying the softmax operation over all image spatial locations, the operation is calculated

---

over all detection boxes. Therefore, the attended regions are restricted to pre-specified detection-box regions and such question-related regions could be more effective for answering questions about foreground objects. However, such restrictions also cause challenges for other types of questions. For instance, for the question "*How is the weather today?*", there might not exist a detection box in the sky, resulting in a failure in answering this question.

For better understanding the question and image contents and their relations, a good VQA algorithm needs to identify global scene attributes, locate objects, identify object attributes, quantity and categories to make accurate inference. We argue that the above mentioned two types of attended mechanisms provide complementary information and should be effectively integrated in a unified framework to take advantages of both the attended free-form regions and attended detection regions. Take the above mentioned two questions as examples, the question about animal could be more effectively answered with detection-based attention maps, while the question about the weather can be better answered with the free-form region based attention maps.

In this paper, we propose a novel dual-branch deep neural network for solving the VQA problem that combines free-form region based and detection-based attention mechanisms (see Figure 1). The overall framework consists of two attention branches, each of which associates the question with the most relevant free-form regions or with the most relevant detection regions in the input image. For obtaining more question-related attention weights for both types of regions, we propose to learn the joint feature representation of the input question, the whole image, and the detection boxes with a multiplicative feature embedding scheme. Such a multiplicative scheme does not share parameters between the two branches and is shown to result in more robust answering performance than existing methods.

The main contributions of our work can be summarized as twofold.

- We propose a novel dual-branch deep neural network that effectively integrates the free-form region based and detection-based attention mechanisms in a unified framework;

- In order to better fuse features from different modalities, a novel multiplicative feature embedding scheme is proposed to learn joint feature representations from the question, the whole image, and the detection boxes.

2 Related work

State-of-the-art VQA algorithms are mainly based on deep neural networks for learning visual-question features and for predicting the final answers. Due to the establishment of the VQA dataset and the online evaluation platform by (Antol et al. 2015), there is an increasing number of VQA algorithms proposed every year.

Multi-modal feature embedding for VQA. The existing joint feature embedding methods for VQA typically combine the visual and question features learned by deep neural networks and then solve the task as a multi-class classification problem. (Zhou et al. 2015) proposed a simple baseline,

which learns image features with CNN and question features from LSTM, and concatenated these two features to predict the answer. Instead of using LSTM for learning question representations, (Noh, Hongsuck Seo, and Han 2016) used GRU (Cho et al. 2014) and (Ma, Lu, and Li 2016) trained CNN for question embedding. Different from the above mentioned methods addressing the VQA task as a classification problem, the work by (Malinowski, Rohrbach, and Fritz 2015) fed both image CNN features and question representations into an LSTM to generate the answer by sequence-to-sequence learning. There are also high-order approaches for multi-modal feature embedding. MCB (Fukui et al. 2016) and MLB (Kim et al. 2017) both designed bilinear pooling approaches to learn multi-modal feature embedding for VQA. In (Ben-Younes et al. 2017), a generalized multi-modal pooling framework is proposed, which shows that MCB and MLB are its special cases. Inspired by ResNet (He et al. 2016), (Kim et al. 2016) proposed element-wise multiplication for the joint residual mappings.

Attention mechanism for VQA. A large quantity of recent works focused on incorporating attention mechanisms for solving the VQA task. In (Xu and Saenko 2016), the attention weights over different image regions are calculated based on the semantic similarity between the question and image regions, and the updated question features are obtained as the weighted sum of the different regions. (Yang et al. 2016) proposed a multi-stage attention framework, which stacks the attention modules to search question-related image regions by iterative feature fusion. The attention mechanism is not limited to image modality, (Lu et al. 2016) proposed a co-attention mechanism that simultaneously attends both the question and image with joint visual-question feature representations.

Since some of the questions are related to objects in the images, object detection results are explored to replace the visual features obtained from the whole-image region. (Shih, Singh, and Hoiem 2016) utilized the attention mechanism to generate visual features from 100 detection boxes for the VQA task. Similarly, (Li and Jia 2016) proposed a framework that updates the joint visual-question feature embedding by iteratively attending the top detection boxes.

However, all the attention-based methods focus on one type of image regions for question-image association (i.e., either free-form image regions or detection boxes), and each of them have limitations on solving certain types of questions. In contrast, in order to better utilize the complementary information from both types of image regions, our proposed approach effectively integrates both attention mechanisms in a unified framework.

3 Method

The overall structure of our proposed deep neural network is illustrated in Figure 2, which takes the question, the whole image, and the detection boxes as inputs, and learns to associate questions to free-form image regions and detection boxes simultaneously to infer the answer. Our proposed model for VQA consists of two co-attention branches for free-form image regions and for detection boxes, respectively. Before calculating the attention weights for each type
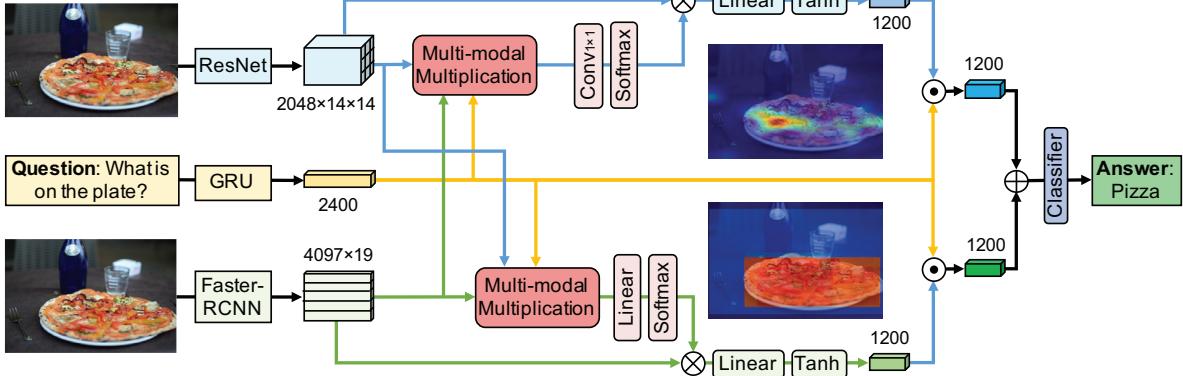
Figure 2: Illustration of the overall network structure for solving the VQA task. The network has two attention branches with the proposed multiplicative featucre embedding scheme, where one branch attends free-form image regions and another branch attends detection boxes for encoding question-related visual features.

of regions, the joint feature representations of the question, the whole-image visual features, and the detection box visual features are obtained via a multiplicative approach. In this way, the attention weights for each type of regions are learned from all the three inputs. Importantly, the joint multimodal feature embedding has different parameters for each branch, which leads to better answering accuracy. The resulting visual features from the two branches are fused with the question representation and then added to obtain the final question-image embedding. A multi-class linear classifier is adopted for obtaining the final predicted answer.

The input feature encoding will be introduced in Section 3.1. Section 3.2 introduces the branch with visual attention mechanism for associating free-form image regions to the input questions with our multiplicative feature embedding scheme, and Section 3.3 describes the another branch with detection based attention mechanism. The final answer prediction will be introduced in Section 3.4.

3.1 Input feature encoding

Encoding whole-image features. We utilize an ImageNet pre-trained ResNet-152 (He et al. 2016) for learning image visual features. The input image to ResNet is resized to 448×448 and the $2048 \times 14 \times 14$ output feature map of the last convolution layer is used to encode the whole-image visual features $R \in \mathcal{R}^{2048 \times 14 \times 14}$, where 14×14 is its spatial size and 2048 is the number of visual feature channels.

Encoding detection-box features. We adopt the Faster-RCNN (Ren et al. 2015) framework to obtain object detection boxes in the input image. For all the object proposals and their associated detection scores generated by Faster-RCNN, the non-maximum suppression with an intersection over union (IoU) 0.3 is applied and the top-ranked 19 detection boxes are chosen as the detection-box features for our overall framework. The 4096-dimensional visual features of Faster-RCNN's fc7 layer concatenated with boxes' detection scores are utilized to encode the visual feature of each detection box. Let $D = [f_1, \cdots, f_{19}] \in \mathcal{R}^{4097 \times 19}$ denote the visual features of the top-ranked 19 detection boxes.

Encoding question features. The Gated Recurrent Unit (GRU) (Cho et al. 2014) is adopted to encode the question features, which show its effectiveness in recent VQA methods (Lu et al. 2016; Kim et al. 2017). A GRU cell consists of an update gate z and a reset gate r. Given a question $q = [q_1, ..., q_T]$, where q_t is the one-hot vector of at position t, and T is the length of the question. We convert each word q_t into a feature vector via a linear transformation $x_t = W_e q_t$. At each time step, the word feature vector q_t is sequentially fed into the GRU to encode the input question. At each step, the GRU updates the update gate z_t and reset gate r_t and outputs a hidden state h_t. The GRU update process operates as

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{1}$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \tag{2}$$
$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h), \tag{3}$$
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t, \tag{4}$$

where σ represents the sigmoid activation function. The weight matrices W, U and bias vector b are learnable parameters of the GRU.

We take the final hidden state h_T as the question embedding, i.e., $Q = h_T \in \mathcal{R}^k$, where k denotes the embedding length. Following (Li and Jia 2016; Kim et al. 2017), we utilize the pre-trained skip-thought model (Kiros et al. 2015) to initialize the embedding matrix W_e in our question language model. Since the skip-thought model is previously trained on large text corpus, we are able to transfer external language-based knowledge to our VQA task by fine-tuning the GRU from the initial point.

3.2 Learning to attend free-form region visual features with multiplicative embedding

Our proposed framework has two branches, one for attending question-related free-form image regions to learn whole-image features and the other for attending detection-box regions to learn question-related detection features. For each attention branch, it takes question feature Q, whole-image

feature R and detection-box feature D as inputs, and outputs the question-attended visual features v_1 and v_2 for question answering.

Our free-form region attention branch tries to associate the input question to relevant regions of the input image. There is no restriction to the attended regions, which are of free-form and are able to capture the global visual context and attributes of the image. Unlike existing VQA methods that fuse question and image features via simple concatenation (Shih, Singh, and Hoiem 2016) or addition (Lu et al. 2016) to guide the calculation of the attention weights, each of our attention branch fuses all three types of input features via a multiplicative feature embedding scheme for utilizing full information from the inputs. The network structure for attending visual features with our multiplicative feature embedding scheme is show in Figure 3(a).

Given the question embedding $Q \in \mathcal{R}^k$, the whole-image representation $R \in \mathcal{R}^{2048 \times 14 \times 14}$, and the detection representation $D \in \mathcal{R}^{4097 \times 19}$, we first embed them to a 1200-dimensional common space via the following equations,

$$R_1 = \tanh(W_{r_1} R + b_{r_1}), \tag{5}$$

$$D_1 = \frac{1}{19} \cdot \mathbf{1}(\tanh(W_{d_1} D + b_{d_1})^T), \tag{6}$$

$$Q_1 = \tanh(W_{q_1} Q + b_{q_1}), \tag{7}$$

where $W_{r_1} \in \mathcal{R}^{1200 \times 2408}$, $W_{d_1} \in \mathcal{R}^{1200 \times 4097}$, $W_{q_1} \in \mathcal{R}^{1200 \times k}$ are the learnable weight parameters, $b_{r_1}, b_{d_1}, b_{q_1} \in \mathcal{R}^{1200}$ are the bias parameters, $\mathbf{1} \in \mathcal{R}^{19}$ represents an all-1 vector, and $\tanh$ is the hyperbolic tangent function. For learning the attention weights for the whole-image feature R, the transformed detection features are averaged across all detection boxes following Eq. (6) before feature fusion. After mapping all input features into the 1200-dimensional common space, the detection feature $D_1 \in \mathcal{R}^{1200}$ and question feature $Q_1 \in \mathcal{R}^{1200}$, are spatially replicated to a 14×14 grid to form $\tilde{D}_1$ and $\tilde{R}_1$, which match the spatial size of the whole-image feature $R_1 \in \mathcal{R}^{1200 \times 14 \times 14}$.

The joint feature representation C_1 of the three inputs is obtained by element-wise multiplication (Hadamard product) of $\tilde{Q}_1$, R_1 and $\tilde{D}_1$, and followed by a L_2 normalization to constrain the magnitude of the representation,

$$C_1 = \text{Norm}_2(\tilde{Q}_1 \circ R_1 \circ \tilde{D}_1), \tag{8}$$

where $\circ$ indicates element-wise multiplication. The free-form attention map is then obtained by convolving the joint feature representation C_1 with a 1×1 convolution followed by a softmax operation over the 14×14 grid,

$$a_1 = \text{softmax}(W_{c_1} * C_1 + b_{c_1}), \tag{9}$$

where $W_{c_1} \in \mathcal{R}^{1200 \times 1 \times 1}$ and $b_{c_1} \in \mathcal{R}^{1200}$ are the learnable convolution kernel parameters. The attended whole-image feature over all spatial locations can be calculated by

$$v_1 = \sum_i^{14 \times 14} a_1(i) R_1(i), \tag{10}$$

which can represent the attended whole-image visual feature that are most related to the input question.

(a) Region Attended Visual Feature   (b) Detection Attended Visual Feature
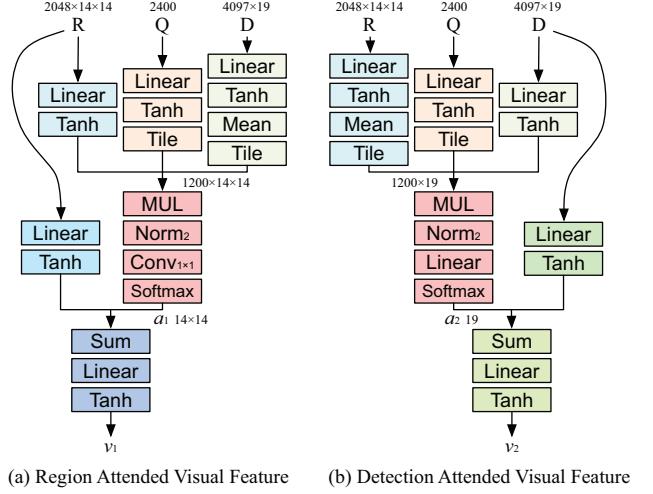
Figure 3: Learning to attend visual features with multimodal multiplicative feature embedding for (a) free-form image regions and for (b) detection boxes.

3.3 Learning to attend detection-box visual features with multiplicative embedding

Our second branch focuses on learning question-related detection features by attending the detection-box features with joint input feature representation. Similar to the first branch, we fuse the question representation, the whole-image features, and the detection-box features for learning the attention weights for detection boxes. Unlike previous detection based attention methods (Shih, Singh, and Hoiem 2016; Li and Jia 2016), our proposed attention mechanism integrates whole-image features for better understanding the overall image contents. The structure for learning the joint feature representation is shown in Figure 3(b).

Similar to the joint representation for free-form image region attention, given the question embedding $Q \in \mathcal{R}^k$, the image representation $R \in \mathcal{R}^{2048 \times 14 \times 14}$, the detection representation $D \in \mathcal{R}^{4097 \times 19}$, we transform each representation to a common semantic space, and then obtain the 1200-dimensional joint feature representation C_2 as

$$D_2 = \tanh(W_{d_2} D + b_{d_2}), \tag{11}$$

$$R_2 = \frac{1}{196} \cdot \mathbf{1}(\tanh(W_{r_2} R + b_{r_2})^T), \tag{12}$$

$$Q_2 = \tanh(W_{q_2} Q + b_{q_2}), \tag{13}$$

$$C_2 = \text{Norm}_2(\tilde{Q}_2 \circ \tilde{R}_2 \circ D_2), \tag{14}$$

where W_{d_2}, W_{r_2}, W_{q_2}, b_{d_1}, b_{r_1}, b_{q_2} are the learnable parameters for linear transformation. The transformed question feature Q_2 and whole-image feature R_2 are replicated across the number dimension to match the dimension of the transformed detection features $D_2 \in \mathcal{R}^{1200 \times 19}$ before calculating the joint representation C_2 following Eq. (14).

The final attended detection representation v_2 over all de-

tection boxes is defined as

$$a_2 = \text{softmax}(W_{c_2}C_2 + b_{c_2}), \tag{15}$$

$$v_2 = \sum_{i}^{19} a_2(i)D_2(i), \tag{16}$$

where $W_{c_2} \in \mathcal{R}^{19 \times 1200}$ and $b_{c_2} \in \mathcal{R}^{19}$ are the learnable parameters of learning the attention weights a_2 for the detection boxes.

3.4 Learning for answer prediction

Similar to existing VQA approaches (Antol et al. 2015; Lu et al. 2016; Fukui et al. 2016), we model VQA as a multi-class classification problem. Given the attended whole-image feature v_1, attended detection feature v_2 with the input question feature Q, question-image joint encodings are obtained by element-wise multiplication of the transformed question features and the attended features from both branch,

$$h_r = v_1 \circ \tanh(W_{hr}Q + b_{hr}), \tag{17}$$
$$h_d = v_2 \circ \tanh(W_{hd}Q + b_{hd}), \tag{18}$$

where W_{hr}, W_{hd}, b_{hr}, b_{hd} are the learnable parameters for transforming the input question feature Q. The reason why we choose different question transformation parameters for the two branches is that the attended visual features from the two different branches captures different information from the input image. One is from attended free-form region features and is able to capture the global context and attributes of the scene. The another is from attended detection features and is able to extract information about foreground objects.

After merging question-image encodings from the two branches via addition, a linear classifier is trained for final answer prediction,

$$p_{ans} = \text{softmax}(W_p(h_r + h_d) + b_p), \tag{19}$$

where W_p and b_p are the classifier parameters, and p_{ans} represents the probability of the final answer prediction.

4 Experiments

4.1 Datasets and evaluation metrics

We evaluate the proposed model on two public datasets, the VQA (Antol et al. 2015) and COCO-QA (Ren, Kiros, and Zemel 2015) datasets.

VQA dataset (Antol et al. 2015) is based on Microsoft COCO image data (Lin et al. 2014). The dataset consists of 248,349 training questions, 121,512 validation questions and 244,302 testing questions, generated on a total of 123,287 images. There are three types of questions including *yes/no*, *number* and *other*. For each question, 10 free-response answers are provided. We take the top 2000 most frequent answers as the candidate outputs for learning the classification model similar to (Kim et al. 2016), which covers 90.45% answers in the training and validation sets.

COCO-QA dataset (Ren, Kiros, and Zemel 2015) is created based on the caption annotations of Microsoft COCO dataset (Lin et al. 2014). There are 78,736 training samples and 38,948 testing samples, respectively. It contains four

types of questions, *object*, *number*, *color* and *location*. All of the answers are single-words and considered as the valid answers, namely 430 answers, which are used for the possible answer classification.

Since we formulate VQA as a classification task, we use the accuracy metric to measure the performances of different models on both two datasets. In particular, for the VQA dataset, a predicted answer is regarded as correct if it matches more than three in ten ground truth answers. WUPS calculates the similarity between two words based on their common subsequence in the taxonomy tree. In addition, Wu-Palmer similarity (WUPS) (Wu and Palmer 1994) is also reported for the COCO-QA dataset. Same as the previous work (Lu et al. 2016; Li and Jia 2016), we report the WUPS scores with the thresholds of 0.9 and 0.0.

4.2 Implementation details

For encoding questions, the length of questions is fixed to 26 and each word embedding is a vector of size 620. The hidden state of GRU is set to 2400. Given the question, image and detection representations, the joint feature embedding of these inputs h is set as 1200. Following (Kim et al. 2017), we take two glimpses of each attention map, that is to say, for each branch, another set of attention weights a_1 is trained for whole-image features R_1 in the first branch and another set of a_2 for detection features D_2 in the second branch. The two sets of attended features are concatenated in each branch as the final feature.

We implement our model with the Torch library. The RMSProp method is used for training our network with an initial learning rate of 3×10^{-4}, a momentum of 0.99 and a weight-decay of 10^{-8}. The batch size is set to 300, and is trained for 250,000 iterations. The validation process is performed every 10,000 iterations with early stopping when validation accuracy stops improving for more than 5 validations. Dropout is applied after every linear transformation and gradient clipping techniques are used for regularization in GRU training. Multi-GPU parallel technology is adopted to accelerate the training process.

4.3 Comparison with state-of-the-arts

Table 1 shows results on the VQA test set for both open-ended and multiple-choice tasks by our proposed approach and compared methods. The approaches shown in Table 1 are trained on the train+val split of the VQA dataset and evaluated on the test split, where the *test-dev* set is normally used for validation and the *test-std* set for standard testing. The models in the first part of Table 1 are based on simple joint question-image feature embedding methods. Models in the second part of Table 1 employ detection-based attention mechanism and models in the third part use free-form region-based attention mechanism. We only compare results of the single models since most approaches do not adopt the model ensemble strategy. We can see that our final model (denoted as Dual-MFA, where MFA stands for Multiplicative Feature Attention) improves the state-of-the-art MLB approach (Kim et al. 2017) from 65.07% to 66.09% for the open-ended task, and from 68.89% to 69.97% for the multiple-choice task on the test-std set. Specifically, in the

| Method | Test-dev | | | | | Test-std | | | | |
| | Open-Ended | | | | MC | Open-Ended | | | | MC |
	All	Y/N	Num.	Other	All	All	Y/N	Num.	Other	All
LSTM Q+I (Antol et al. 2015)	53.74	78.94	35.24	36.42	57.17	54.06	79.01	35.55	36.80	57.57
iBOWING (Zhou et al. 2015)	55.72	76.55	35.03	42.62	61.68	55.89	76.76	34.98	42.62	61.97
DPPnet (Noh, Hongsuck Seo, and Han 2016)	57.22	80.71	37.24	41.69	62.48	57.36	80.28	36.92	42.24	62.69
FDA (Ilievski, Yan, and Feng 2016)	59.24	81.14	36.16	45.77	64.01	59.54	81.34	35.67	46.10	64.18
DMN+ (Xiong, Merity, and Socher 2016)	60.30	80.50	36.80	48.30	-	60.36	80.43	36.82	48.33	-
Region Sel. (Shih, Singh, and Hoiem 2016)	-	-	-	-	62.44	-	-	-	-	62.43
QRU (Li and Jia 2016)	60.72	82.29	37.02	47.67	65.43	60.76	-	-	-	65.43
SMem (Xu and Saenko 2016)	57.99	80.87	37.32	43.12	-	58.24	80.80	37.53	43.48	-
SAN (Yang et al. 2016)	58.70	79.30	36.60	46.10	-	58.85	79.11	36.41	46.42	-
MRN (Kim et al. 2016)	61.68	82.28	38.82	49.25	66.15	61.84	82.39	38.23	49.41	66.33
HieCoAtt (Lu et al. 2016)	61.80	79.70	38.70	51.70	65.80	62.06	79.95	38.22	51.95	66.07
VQA-Machine (Wang et al. 2017)	63.10	81.50	38.40	53.00	67.70	63.30	81.40	38.20	53.20	67.80
MCB (Fukui et al. 2016)	64.70	82.50	37.60	55.60	69.10	-	-	-	-	-
MLB (Kim et al. 2017)	64.89	**84.13**	37.85	54.57	-	65.07	**84.02**	37.90	54.77	68.89
Dual-MLB (our baseline)	65.12	83.32	39.96	55.26	69.55	-	-	-	-	-
Dual-MFA (ours)	**66.01**	83.59	**40.18**	**56.84**	**70.04**	**66.09**	83.37	**40.39**	**56.89**	**69.97**

Table 1: Evaluation results by our proposed method and compared methods on the VQA dataset.

question types of *number* and *other*, our proposed approach brings 2.49% and 2.12% improvements on the test-std set. QRU (Li and Jia 2016) is the state-of-the-art detection-based attention method, and our model significantly outperforms it by 5.33% on the test-std set. We also compare our approach with a baseline model Dual-MLB, which consists of two attention branches based on the MLB tensor fusion module. The baseline model fuses the question and free-form image region embedding by MLB in one branch, and fuses question and image detection embedding by MLB in another branch. The baseline Dual-MLB also outperforms MLB by 0.23% on the test-dev set, which shows that our improvements are not only from the integration of the two attention mechanisms, but are also caused by effective joint feature embedding of question, image, and detection features.

Table 3 compares our approach with the state-of-the-art approaches on the COCO-QA test set. Our final model Dual-MFA improves the state-of-the-art HieCoAtt (Lu et al. 2016) from 65.40% to 66.49%. In particular, our model achieves an improvement of 2.99% for the question type *color*. Similar to the results on the VQA dataset, our model significantly outperforms the state-of-the-art detection-based attention method QRU by 3.49%.

4.4 Ablation study

In this section, we conduct ablation experiments to study effectiveness of individual component designs in our model. Table 2 shows the results of baseline models replacing different components in our model, which are trained on the VQA training set and tested on the validation set. Following other compared approaches, the test set is not used in the study due to online submission restrictions. Specifically, we compare different multi-modal feature embedding designs and investigate the roles of the two attention mechanisms.

Method	Validation
MFA-MUL*	59.01
MFA-ADD	56.34
MFA-Norm*	59.18
MFA w/o Norm	59.01
MFA-Power	58.93
Dual-MFA-ADD*	59.82
Dual-MFA-MUL	58.35
Dual-MFA-CAT	58.53
MFA-D*	55.57
QRU-D (Li and Jia 2016)	53.99
MLB-D (Kim et al. 2017)	54.89
MFA-R*	59.18
MLB-R (Kim et al. 2017)	57.40
MUTAN (Ben-Younes et al. 2017)	58.16
Dual-MLB	59.07
Dual-MFA* (full model)	**59.82**

Table 2: Ablation study on the VQA dataset, where "*" denotes our model's design.

The first part of Table 2 shows different element-wise operations used for joint embedding of three input features. Element-wise multiplication (denoted as MFA-MUL) in Eq. (8) performs better than element-wise addition (denoted as MFA-ADD) by 3.67%. The second part of Table 2 shows L_2 normalization (denoted as MFA-Norm) in joint feature embedding works better than the model with unsigned power operation of $1/3$ (denoted as MFA-Power) and the model without normalization (denoted as MFA w/o Norm). For fusing outputs h_r and h_d from the two attention branches, element-wise addition in Eq. (19) achieves better perfor-

Method	All	Obj.	Num.	Color	Loc.	WUPS0.9	WUPS0.0
2VIS+BLSTM (Ren, Kiros, and Zemel 2015)	55.09	58.17	44.79	49.53	47.34	65.34	88.64
IMG-CNN (Ma, Lu, and Li 2016)	58.40	-	-	-	-	68.50	89.67
DDPnet (Noh, Hongsuck Seo, and Han 2016)	61.16	-	-	-	-	70.84	90.61
SAN (Yang et al. 2016)	61.60	65.40	48.60	57.90	54.00	71.60	90.90
QRU (Li and Jia 2016)	62.50	65.06	46.90	60.50	56.99	72.58	91.62
HieCoAtt (Lu et al. 2016)	65.40	68.00	51.00	62.90	58.80	75.10	92.00
Dual-MFA (ours)	**66.49**	**68.86**	**51.32**	**65.89**	**58.92**	**76.15**	**92.29**

Table 3: Evaluation results by our proposed method and compared methods on the COCO QA dataset.

Q: What sport is this? A: tennis
Q: What is the color of the surfboard? A: white
Q: Is it a sunny day? A: yes
Q: How many giraffes are there? A: 5
Q: What does the red circle sign mean? A: no parking
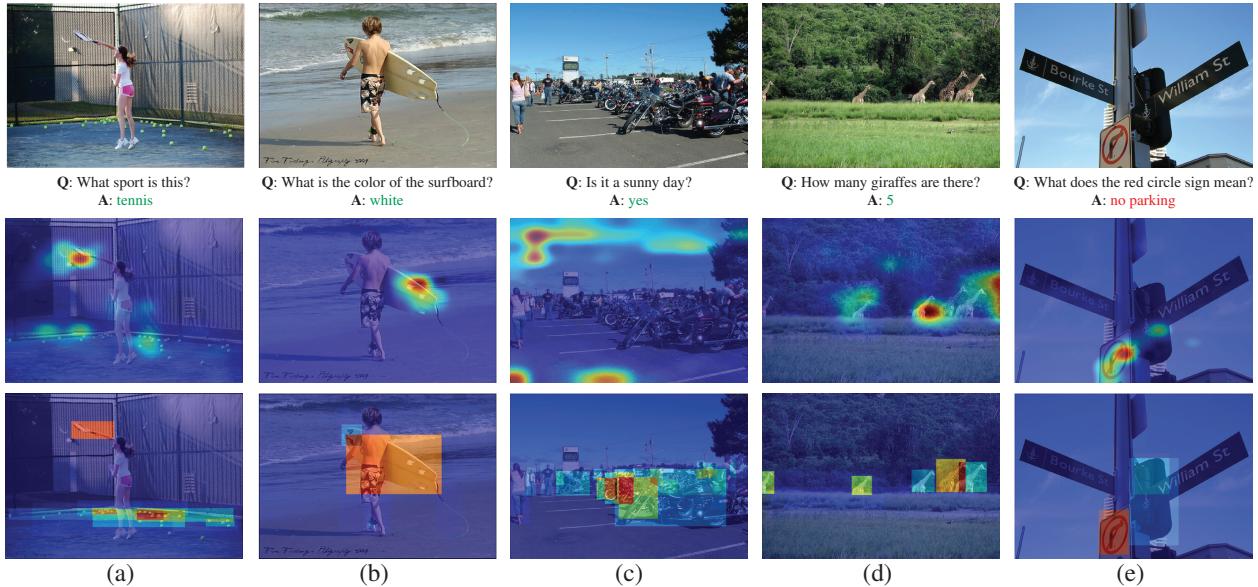
(a)   (b)   (c)   (d)   (e)

Figure 4: Visualization examples on the VQA test-dev set. (First row) input images. (Second row) free-form region based attention maps. (Third row) detection-based attention maps.

mance than element-wise multiplication (denoted as Dual-MFA-MUL) and concatenation of two vectors (denoted as Dual-MFA-CAT).

The third and fourth parts of Table 2 compare our multimodal multiplicative feature embedding for a single attention mechanism (MFA-R or MFA-D) with detection-based attention baseline models (QRU-D and MLB-D) and region-based attention baseline models (MUTAN and MLB-R), which replace our multiplicative feature embedding with QRU (Li and Jia 2016) or MLB (Kim et al. 2017) respectively. The MUTAN (Ben-Younes et al. 2017) approach is also used for comparison. Results show that our models with a single attention branch outperform all compared detection-based and region-based baselines. Finally, we compare our final model Dual-MFA with the baseline model Dual-MLB. Our Dual-MFA model benefits from its effective multi-modal multiplicative feature embedding scheme and achieves an improvement of 0.75%.

The parameter size of our final full model is 82.6M, while the best region-based model (MFA-R) is 61.7M, and the best detection-based model (MFA-D) is 56.8M. As the parameters in the language model and the classifier are shared

by two attention branches, parameters of our full model increase in a minor degree.

4.5 Qualitative evaluation

We visualize some co-attention maps generated by our model in Figure 4 and present five examples from the VQA test set. Figures 4(a) and (b) show that our model attends to the corresponding image regions with two attention branches, which leads to correct answers with higher confidence. There are also cases where only one attention branch is able to attend the correct image region to obtain the correct answer. In Figure 4(c), the free-form region based attention map is able to attend the blue sky and dry ground, while the detection-based attention map fails as there are no pre-given detection boxes covering these image regions. In Figure 4(d), the detection-based attention map has higher weights on all five giraffes to generate the correct answer, while the free-form attention map attends incorrect regions. A failure case is also presented in Figure 4(e). The model fails to generate the correct answer as the classification label "no turning right" does not exit in the training set despite attending the correct image region.

5 Conclusion

In this paper, we propose a novel deep neural network with co-attention mechanism for visual question answering. The deep model contains two branches for visual attention which aim to select the free-form image regions and detection boxes most related to the input question. We generate visual attention weights by a novel multiplicative embedding scheme, which fuses the question, whole-image and detection-box features effectively. Ablation study demonstrates the effectiveness of individual components of our proposed model. Experimental results on two large VQA datasets show that our proposed model outperforms state-of-the-art approaches.

Acknowledgement

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Ben-Younes, H.; Cadène, R.; Thome, N.; and Cord, M. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*.

Cho, K.; van Merriënboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Ilievski, I.; Yan, S.; and Feng, J. 2016. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Kim, J.-H.; Lee, S.-W.; Kwak, D.; Heo, M.-O.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Multimodal residual learning for visual qa. In *NIPS*.

Kim, J.-H.; On, K.-W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

Kiros, R.; Zhu, Y.; Salakhutdinov, R. R.; Zemel, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NIPS*.

Li, R., and Jia, J. 2016. Visual question answering with question representation update (qru). In *NIPS*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017a. Person search with natural language description. In *CVPR*.

Li, Y.; Duan, N.; Zhou, B.; Chu, X.; Ouyang, W.; and Wang, X. 2017b. Visual question generation as dual task of visual question answering. *arXiv preprint arXiv:1709.07192*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*.

Ma, L.; Lu, Z.; and Li, H. 2016. Learning to answer questions from image using convolutional neural network. In *AAAI*.

Malinowski, M.; Rohrbach, M.; and Fritz, M. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*.

Mostafazadeh, N.; Misra, I.; Devlin, J.; Mitchell, M.; He, X.; and Vanderwende, L. 2016. Generating natural questions about an image. In *ACL*.

Noh, H.; Hongsuck Seo, P.; and Han, B. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Ren, M.; Kiros, R.; and Zemel, R. 2015. Exploring models and data for image question answering. In *NIPS*.

Shih, K. J.; Singh, S.; and Hoiem, D. 2016. Where to look: Focus regions for visual question answering. In *CVPR*.

Wang, P.; Wu, Q.; Shen, C.; and Hengel, A. v. d. 2017. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *CVPR*.

Wu, Z., and Palmer, M. 1994. Verbs semantics and lexical selection. In *ACL*.

Xie, L.; Shen, J.; and Zhu, L. 2016. Online cross-modal hashing for web image retrieval. In *AAAI*.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.

Xu, H., and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*.

Ye, Y.; Zhao, Z.; Li, Y.; Chen, L.; Xiao, J.; and Zhuang, Y. 2017. Video question answering via attribute-augmented attention network learning. In *ACM Multimedia*.

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

Zitnick, C. L., and Dollár, P. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*.