Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal feature-wise co-attention method for visual question answering

Sheng Zhang [a], Min Chen [c], Jincai Chen [a,b,d,*], Fuhao Zou [d], Yuan-Fang Li [e], Ping Lu [a,b,d]

[a] *Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China*
[b] *Key Laboratory of Information Storage System, Engineering Research Center of Data Storage Systems and Technology, Ministry of Education of China, Wuhan 430074, China*
[c] *Embedded and Pervasive Computing (EPIC) Lab, Huazhong University of Science and Technology, Wuhan 430074, China*
[d] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[e] *Faculty of Information Technology, Monash University, Clayton 3800, Australia*

ARTICLE INFO

Keywords:
Feature-wise attention learning
Deep learning
Multimodal feature fusion
Visual question answering (VQA)

ABSTRACT

VQA attracts lots of researchers in recent years. It could be potentially applied to the remote consultation of COVID-19. Attention mechanisms provide an effective way of utilizing visual and question information selectively in visual question and answering (VQA). The attention methods of existing VQA models generally focus on spatial dimension. In other words, the attention is modeled as spatial probabilities that re-weights the image region or word token features. However, feature-wise attention cannot be ignored, as image and question representations are organized in both spatial and feature-wise modes. Taking the question "What is the color of the woman's hair" for example, identifying the hair color attribute feature is as important as focusing on the hair region. In this paper, we propose a novel neural network module named "multimodal feature-wise attention module" (MulFA) to model the feature-wise attention. Extensive experiments show that MulFA is capable of filtering representations for feature refinement and leads to improved performance. By introducing MulFA modules, we construct an effective union feature-wise and spatial co-attention network (UFSCAN) model for VQA. Our evaluation on two large-scale VQA datasets, VQA 1.0 and VQA 2.0, shows that UFSCAN achieves performance competitive with state-of-the-art models.

1. Introduction

The development of deep learning has accelerated the advancement of computer vision and natural language processing. Visual question answering (VQA) [1,2] requires simultaneous comprehension of visual images and natural language questions, it has become one of the most active research areas in artificial intelligence.

VQA is a complex multimodal task that aims at automatically answering a textual question related to the content of a given image. It is useful to help visually impaired subjects to realize the visual environment and can be applied in entertainment, education. VQA also can support remote consultation and cross-modal queries of medical images. For example, an unprofessional doctor could provide a chest X-ray of a patient, and ask the VQA expert system about the condition of the patient online. These can prevent spread and improve the diagnosis efficiency in COVID-19. VQA requires a fine-grained understanding of the semantics of both images and questions. Specifically, to produce a correct answer, visual information relevant to a question and textual information related to the content of an image need to be extracted, which is referred to as vision-language cross-grounding problems.

For tackling these problems, attention-based models [3–5] have been extensively explored for VQA, where the visual attention mechanism typically produces a spatial map highlighting image regions relevant to the question. Likewise, the question attention mechanism, which focuses on core words of a sentence, has been considered along with the visual attention [6–8]. However, these existing methods only consider spatial attention, i.e., *where to look* or *where to read*. In this paper, we argue that semantic feature attention (i.e., *what to look at* and *what to read*) is equally important. For instance, to answer the questions in Fig. 1 correctly, a model not only needs to locate the woman's hair or the ground in the image, but also needs to focus on the color feature attribute of them.

How to assign more attention to important attributes? First, we discuss how image and question representations reflect semantics attributes. In VQA, image representation is generally obtained by using CNN [5,7,9] (or Faster R-CNN [10–13]). Some methods use the feature map with the size of $W \times H \times C$ from the last pooling layer of CNN, which can retain spatial information of the original images. $W \times H$

[*] Corresponding author.
E-mail addresses: zhangmonkey@hust.edu.cn (S. Zhang), minchen2012@hust.edu.cn (M. Chen), jcchen@hust.edu.cn (J. Chen), fuhao_zou@hust.edu.cn (F. Zou), yuanfang.li@monash.edu (Y.-F. Li), luping06@hust.edu.cn (P. Lu).

https://doi.org/10.1016/j.inffus.2021.02.022
Received 20 September 2020; Received in revised form 24 November 2020; Accepted 21 February 2021
Available online 1 March 2021

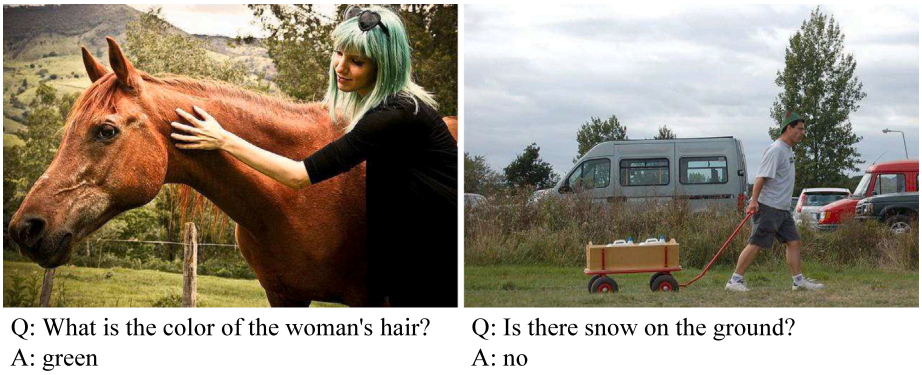

Q: What is the color of the woman's hair?
A: green
Q: Is there snow on the ground?
A: no

Fig. 1. Examples of visual question answering. The questions are related to attributes of objects in images in these examples. Specifically, the first question focuses on the color attribute of the hair. The second question is related to the color attribute of the ground.

is the resolution of regions in the image representation (i.e. spatial dimension). Each region includes C feature values, which mean an implied semantics (i.e. feature-wise dimension). For a question, we generally apply Long Short-Term Memory [14] (LSTM) or Gated Recurrent Unit [15] (GRU) to extract its word-level representation, which contains T feature vectors of length N. T is the number of words (i.e. spatial dimension). N is the number of features for each word and it depends on the number of LSTM (or GRU) cells (i.e. feature-wise dimension).

Essentially, a feature is an activation response of neurons in a convolution kernel or an LSTM cell. Each neuron actually detects an implied semantic pattern. Therefore, a feature describes a certain semantic. We can model feature-wise attention via assigning different weights to these features. The feature-wise attention can highlight significant semantic attributes and give less emphasis to unimportant ones [16], so it can complement the spatial attention mechanism.

Motivated by the above observation, we construct the multimodal feature-wise attention module (MulFA) for images and questions modalities respectively. MulFA can perform both question-guided image and image-guided question feature-wise attention. With simultaneous image and question MulFA modules at its core, we introduce our union feature-wise and spatial co-attention network (UFSCAN) model for VQA. UFSCAN comprises a number of additional modules. It includes a novel multimodal feature-wise co-attention module (referred to as “MulFCoA”) to model feature-wise attention of image and question modalities. We also design a visual spatial attention module (VSA) to highlight image regions related to the question. We finally construct a multimodal residual module (MulRM), which is a variant of multimodal residual networks [17], to effectively fuse visual information from spatial attention with textual information.

In summary, our contributions in this work are as follows:

1. We propose novel multimodal feature-wise attention mechanisms: (a) question-guided image feature-wise attention and (b) image-guided question feature-wise attention, and construct MulFA modules to learn cross feature-wise attentions between image and question modalities. These allow obtaining more discriminative representations for image and question modalities.
2. We construct an UFSCAN model for VQA, which simultaneously models feature-wise co-attention and spatial co-attention between image and question modalities, and adopts a MulRM to combine visual and textual information.
3. Finally, we evaluate our proposed UFSCAN on the large-scale and highly competitive datasets VQA 1.0 and VQA 2.0. Our model achieves performance competitive with state-of-the-art models on these VQA datasets. We also perform extensive ablation experiments and demonstrate the effectiveness of our proposed feature-wise attention mechanism.

2. Related work

VQA has aroused broad interests in recent years since the seminal work of Antol et al. [1]. The task of VQA is extremely challenging. VQA straddles the fields of computer vision and natural language processing. It requires a fine-grained understanding of the semantics of both images and questions, and reasoning over the combination of these two modalities, sometimes on the basis of external or common-sense knowledge. The VQA is closely related to image captioning [18–20], which is also a task involving visual images and natural language sentences.

To support the VQA task, there are generally the following three issues that need to be addressed. First, we need to extract discriminative representations for image and question information. Second, we need to combine visual and textual representations to generate the fused image–question features. Last, we need to train a multi-class classifier for predicting the best matching answer correctly using the fused image–question features. Furthermore, the attention mechanism is widely explored for VQA to highlight core words of a question and image regions relevant to the question. We next introduce related works of VQA.

Fusion strategies. To generate expressive image–question fusion features, high-level interactions between image and question representations need to be carefully encoded into the model. Due to the encoding of the full second-order interactions, the bilinear model is a powerful approach for the fusion problem in VQA. Given two feature vector $v \in \mathbb{R}^{n_1}$ and $q \in \mathbb{R}^{n_2}$ as input, the bilinear model can be formulated as

$$z = W[v \otimes q] \tag{1}$$

where $\otimes$ denotes outer product and $[]$ indicates linearizing the matrix in a vector, and W is the learned parameter. The main drawback of the bilinear model is that the number of parameters for W, $n_1 \times n_2$, is massive and thus intractable. [21] proposes multimodal low-rank bilinear pooling (MLB) to reduce the computational complexity of the original bilinear model. However, the low-rank tensor structure in MLB is equivalent to computing simple element-wise product between projection of visual and question representations. The multimodal Tucker fusion (MUTAN) method in [9] combines a Tucker decomposition with a low-rank matrix constraint as

$$\begin{aligned} z_r &= ((qW_q)M_r) \times ((vW_v)N_r) \\ z &= \sum_{r=1}^{R} z_r \end{aligned} \tag{2}$$

where $W_q \in \mathbb{R}^{n_2 \times t_q}$, $M_r \in \mathbb{R}^{t_q \times t_o}$, $W_v \in \mathbb{R}^{n_1 \times t_v}$ and $N_r \in \mathbb{R}^{t_v \times t_o}$ are learned parameters. R serves to balance the full bilinear interaction's complexity and accuracy.

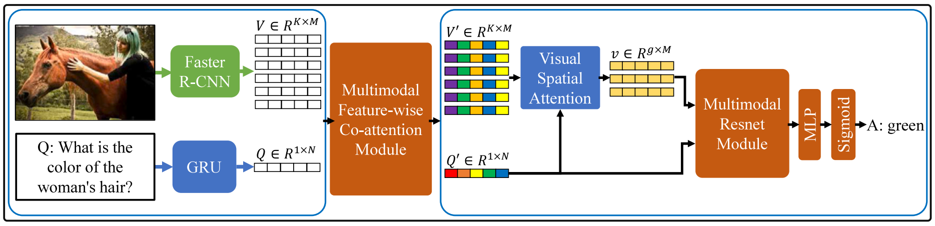


Fig. 2. Overview of the UFSCAN architecture for VQA. The image and question representations are extracted (Section 3). Then, MulFCoA (Section 3.2) is applied to provide feature-wise attention features. VSA (Section 3.3) is modeled to locate the image regions related to the question and generate spatial attention fine-grained features for the image. Finally, image features and question features are fused by MulRM (Section 3.4) and then passed through MLP to make answer predictions.

Attention. Many recent attention-based deep neural networks have been proposed for VQA. The attention methods proposed in [5,9,11,12,22] focus on solving the problem of visual spatial attention. Specifically, these methods produce a spatial probability distribution emphasizing image regions relevant to the question. Differently, in addition to reasoning about visual spatial attention, approaches proposed in [6,7,13] also involve question spatial attention. Similarly, question spatial attention highlights words including core meaning in a question by generating a spatial probability distribution on words.

However, only taking into account spatial attention is inadequate. It only addresses the problem of where need to focus on. In this work, we propose multimodal feature-wise attention modules (MulFA) to model feature-wise attention.

3. Network architectures for VQA

The goal of the VQA task is to answer a question according to the content of an image. The UFSCAN network architecture is illustrated in Fig. 2. Our model first harnesses Faster-RCNN [10] pre-trained using Visual Genome [23] to extract representations of an image. Therefore, we obtain K object feature vectors including M feature values. The resulting visual features can be represented as $V \in \mathbb{R}^{K\times M}$. The question representation is extracted by GRU [24] network with N hidden units. We do not consider an attention on words, following [11]. Thus, we get the sentence feature vector of questions, which can be defined as $Q \in \mathbb{R}^{1\times N}$.

These features are then fed into MulFCoA module described in Section 3.2, and it is composed of IMulFA and QMulFA modules in Section 3.1. The MulFCoA generates feature-wise attention features, where informative features are emphasized and less useful ones are suppressed. This makes the features more discriminate and increases the capability of fine-grained recognition. For image features including the spatial dimension (i.e., object dimension), we use VSA in Section 3.3 to find out multiple image regions related to the question. Finally, Section 3.4 shows how to fuse the region features and question features to generate images–question feature u_g. Thus, we call our network architecture as union feature-wise and spatial co-attention network (UFSCAN).

In VQA, most answers consist of a single word [1]. Therefore, existing methods generally treat VQA as a classification problem. Additionally, each question is associated with one or several answers. Thus, we treat VQA as a multi-label classification task like previous work [11,13]. We use the most frequent answers with more than eight occurrences as candidate. We finally obtain 3,129 candidate answers.

We pass the images–question feature from MulRM through a Multi-layer Perception (MLP) with a ReLU activation and then a *sigmoid* activation to generate scores $\hat{s}$ for the candidate answers:

$$\hat{s} = sigmoid(MLP(u_g)) \tag{3}$$

The sigmoid function normalizes the scores into $(0, 1)$. Because of multiple labels for a question, our objective function is a binary cross-entropy function in training phase, like

$$L = -\sum_{i}^{\mathbb{M}} \sum_{j}^{\mathbb{N}} s_{ij} log(\hat{s}_{ij}) - (1 - s_{ij}) log(1 - \hat{s}_{ij}) \tag{4}$$

where the indices i and j run respectively over the $\mathbb{M}$ samples and $\mathbb{N}$ candidate answers. In test phase, we select the candidate with the highest score as the predicted answer.

3.1. Multimodal feature-wise attention modules

For a given image–question pair, most of the existing methods have considered spatial cross-grounding. In other words, they predict relevance between each spatial object of an image and a question. They also learn the significance of each word in the question. However, these attention models only focus on learning the spatial attention, while completely ignore the attention of the feature channel dimension in the image and question representations. Some other tasks of computer vision, e.g. classification [25] and image caption [20], have demonstrated that incorporating the feature channel attention mechanism contributes to better performance, because it allows the model to effectively learn which feature channel is important.

In this section, we propose the Multimodal Feature-wise Attention (referred to as MulFA) modules. MulFA is aimed at generating attention weights to emphasize informative features and suppress less useful ones. To generate attention weights, MulFA requires performing interactions between visual and textual information. We apply a bilinear model (abbreviated as BM) to complete the interactions. We design different MulFA structures for image and question modalities, and describe them in Sections 3.1.1 and 3.1.2 respectively.

3.1.1. Image multimodal feature-wise attention module

The entire process of image multimodal feature-wise attention (abbreviated as IMulFA) module includes four steps: (1) squeezing image features, (2) fusing feature-wise statistics and question signal, (3) computing feature-wise attention weight, and (4) feature-wise re-weighting image features. Fig. 3 shows the high-level structure of IMulFA. We describe each step of IMulFA in detail below.

It has been observed in previous works [20,25] that the interdependencies among feature channels of the image features contribute to modeling a feature-wise attention. Thus, we should obtain feature channel descriptors of image features. The image features $V \in \mathbb{R}^{K\times M}$ includes K object feature vectors of M feature channels. Each of M feature channels has an implicit semantics. Accordingly, we take advantage of pooling operations to aggregate image features across the object dimension K (i.e., spatial dimension) to produce feature channel descriptors as follows:

$$z_{average,j} = \frac{1}{K} \sum_{i=1}^{K} V(i,j) \tag{5}$$

where $z_{average} \in \mathbb{R}^{M}$ denotes the average statistics of image features with M feature channels. Pooling across the object dimension of image features masks out spatial distribution information.

In VQA, the image feature-wise attention is dependent on the question, as the question decides which channel semantic attributes are more important. Hence, we use a bilinear model (referred to as BM) to combine the feature channel statistics and the question features:

$$f_{average} = BM(z_{average}, Q^{T}) \tag{6}$$

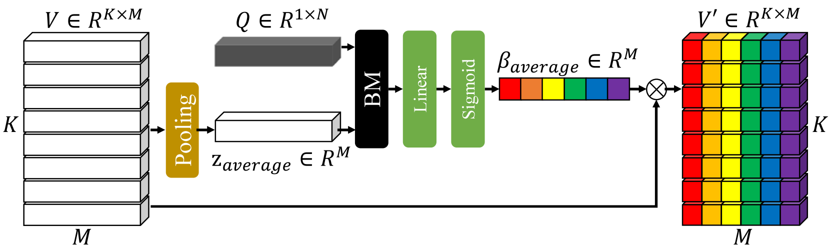


Fig. 3. The structure of the image multimodal feature-wise attention (IMulFA) module. The IMulFA includes four steps: (1) squeezing image features, (2) fusing the feature-wise statistics and question signals, (3) computing feature-wise attention weight, and (4) feature-wise re-weighting the image features.

where $Q \in \mathbb{R}^{1\times N}$ denotes the question features, and $f_{average} \in \mathbb{R}^C$ is the fusion features.

Next, we pass the fusion features through a single-layer linear network and then a *sigmoid* activation function to produce the feature-wise attention weight vector:

$$\beta_{average} = sigmoid(W_a^v f_{average}) \tag{7}$$

where $W_a^v \in \mathbb{R}^{M\times C}$ is a parameter matrix, and $\beta_{average} \in \mathbb{R}^M$ are the attention weight vector for feature channels of image features. The *sigmoid* activation normalizes each attention weight between 0 and 1, which could be interpreted as correlation between the feature semantic attributes and the question. An attention value close to 0 denotes a low degree of correlation and that close to 1 means a strong correlation.

Finally, the feature vector of each object is multiplied by the attention weight vector $\beta_{average}$ in an element-wise multiplication way to emphasize informative feature signals and suppress less useful ones:

$$V' = \beta_{average}^T \odot V \tag{8}$$

where $V' \in \mathbb{R}^{K\times M}$ is a feature-wise attention image features, and $\odot$ denotes broadcast element-wise multiplication between a vector and a matrix. Specifically, the element-wise multiplication is performed between each row of the matrix and the vector. We define this IMulFA as

$$V' = IMulFA(V, Q) \tag{9}$$

3.1.2. Question multimodal feature-wise attention module

The question multimodal feature-wise attention (abbreviated as QMulFA) module includes three main steps: (1) fusing multimodal information to generate feature-wise attention weight vectors, (2) squeezing attention weight vectors, and (3) recalibrating question feature. Fig. 4 illustrates the structure of the QMulFA.

Similar to IMulFA, we first should obtain the feature channel statistics of question features. Different from the image features, the question representation is a vector instead of a matrix. Consequently, the squeeze operation in spatial dimension for question features is unnecessary. In other words, the question features are treated as the feature channel statistics directly. Moreover, the question feature-wise attention is also relevant to the visual signal. We fuse the visual signal and the question feature channel statistics to produce the question feature-wise attention weight vectors. The ith weight vector h_i is generated by the ith object feature vector $V_i \in \mathbb{R}^M$ and the question features Q as follows:

$$\begin{aligned} f_i &= BM(V_i, Q^T) \\ h_i &= sigmoid(W_f^q f_i) \end{aligned} \tag{10}$$

where $h_i \in \mathbb{R}^N$, $W_f^q \in \mathbb{R}^{N\times C}$ is a parameter matrix of the single linear layer, and $f_i \in \mathbb{R}^C$ denotes the fusion feature obtained by a bilinear model.

The visual feature $V \in \mathbb{R}^{K\times M}$ includes K objects, each of which can guide the question feature-wise attention. Accordingly, we employ an average pooling operation to integrate the effect of all objects as follows:

$$\alpha = \frac{1}{K}\sum_{i=1}^{K} h_i \tag{11}$$

where $\alpha \in \mathbb{R}^N$ denotes the question feature-wise attention weight vector.

Finally, we combine Q and the attention by element-wise multiplication to recalibrate the question features as follows:

$$Q' = \alpha^T \times Q \tag{12}$$

where $Q' \in \mathbb{R}^{1\times N}$ is the feature-wise attention features. We define this QMulFA as

$$Q' = QMulFA(V, Q) \tag{13}$$

3.2. Multimodal feature-wise co-attention module for VQA

We have developed feature-wise attention learning modules to image and question modalities, as shown in Section 3.1.1 and Section 3.1.2 respectively. To combine them, we propose three co-attention mechanisms that differ in the order in which image and question feature-wise attention are performed. The first two mechanisms, which we call alternating co-attention, sequentially alternate between performing image and question feature-wise attention, as below

$$\begin{aligned} &V' = IMulFA(V, Q),\ \ Q' = QMulFA(V', Q) \\ &or \\ &Q' = QMulFA(V, Q),\ \ V' = IMulFA(V, Q') \end{aligned} \tag{14}$$

The third mechanism, which we call parallel co-attention, generates image and question attention simultaneously, defined as

$$\begin{aligned} V' &= IMulFA(V, Q) \\ Q' &= QMulFA(V, Q) \end{aligned} \tag{15}$$

We compare three different feature-wise co-attention mechanisms in the ablation study in Section 4.4.

3.3. Multimodal spatial attention module

In VQA, in order to answer the question correctly, we need to focus on the regions related to the question in an image. Thus, we construct a visual spatial attention module (abbreviated as VSA).

Our visual spatial attention module employs multiple spatial attention heads (a.k.a. glimpses) to filter out noises and highlight the regions that are highly relevant to the question. For each glimpse, we first fuse the visual feature $V' \in \mathbb{R}^{K\times M}$ and the question feature $Q' \in \mathbb{R}^{1\times N}$ computed by the bilinear model and then feed the fusion features to a softmax function to generate attention distributions over the regions of

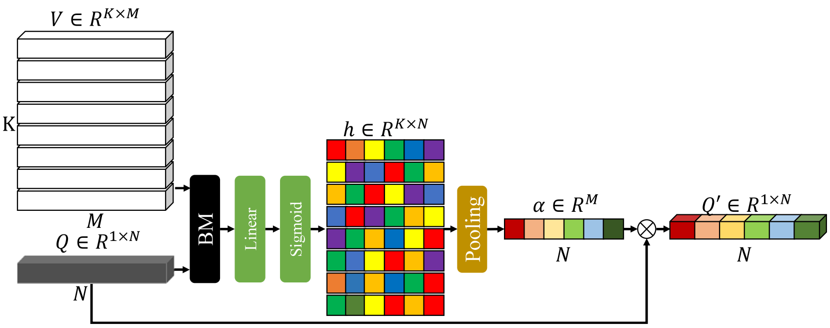


Fig. 4. The structure of the question multimodal feature-wise attention (QMulFA) module. The QMulFA includes three steps: (1) fusing the multimodal information to generate feature-wise attention weight vectors, (2) squeezing the attention weight vectors, and (3) adjusting the question features.

the image as follows:

$$\begin{aligned} h_i &= BM(V_i', Q'^T) \\ h &= [h_0, h_1, \dots, h_{K-1}] \\ p &= softmax(W_h^v h) \\ v_j &= \sum_{i=1}^{K} p_{j,i} V_i', \ \ j \in \{1, 2, \dots, g\} \end{aligned} \tag{16}$$

where V_i denotes the ith object feature, $h_i \in \mathbb{R}^C$ is the ith fusion feature, [.] is the stacking operation between vectors, hence $h \in \mathbb{R}^{C\times K}$. $W_h^v \in \mathbb{R}^{g\times C}$ are a parameter matrix. $p \in \mathbb{R}^{g\times K}$ indicates g image region attention distributions, where g is the number of glimpses. $v_j \in \mathbb{R}^M$ denotes the jth spatial attention visual feature.

3.4. Multimodal residual module

The visual spatial attention module generates multiple spatial attention features v_j where $j \in \{1, \dots, g\}$. And question features $Q' \in \mathbb{R}^{1\times N}$ are made up of a single feature vector. The traditional multimodal feature fusion methods, e.g. element-wise product, sum and concatenation, become invalid. So, it is difficult to effectively combine the visual features with question features. The Multimodal Residual Networks (MRN) [13,17] is inspired by residual structure to use shortcuts and residual mappings for handling multiple modalities. But BM can more effectively learn the multimodal representations than element-wise multiplication used in the original MRN. So we introduce BM into MRN and propose a multimodal residual module (abbreviated as MulRM) to integrate the joint representations from the multiple glimpses and question. The $j+1$th output of our MulRM is defined as follows:

$$u_{j+1} = BM_j(v_j, u_j) + W_j^m u_j, \quad j \in \{1, \dots, g\} \tag{17}$$

where $u_j \in \mathbb{R}^C$, $u_0 = Q'^T$ (if $N = C$), v_j is the attention feature vector of the jth glimpse, and W_j^m are parameters of a linear layer.

4. Experiments

In this section, we first describe the two datasets for evaluation, VQA 1.0 and VQA 2.0, in Section 4.1 and hyperparameter settings in Section 4.2. We present comparison results with the state-of-the-art models for VQA 1.0 and VQA 2.0 in Section 4.3. Finally, we report ablation results in Section 4.4.

4.1. Datasets

VQA 1.0. The VQA 1.0 dataset [1] contains over 204K images from the Microsoft Common Objects in Context (MS COCO) dataset [26], over 600K questions (at least 3 questions per image), and over 6 million answers (10 answers per question). The dataset is split into three subsets: train (80K images and 240K question–answer pairs), val (40K images and 120K question–answer pairs), and test (80K images and 240K question–answer pairs). We use the tools provided by Antol et al. [1] to evaluate the accuracy. Specifically, the VQA accuracy metric is $\min(1, \#humans\ that\ provided\ that\ answer/3)$

The test set is split into test-dev, test-standard, test-challenge, and test-reserve following the same test split strategy as the MC COCO dataset. The ground truth answers for the test set are unavailable, and one must submit their results to a remote evaluation server to acquire the testing scores.

VQA 2.0. The VQA 2.0 dataset [27] is an updated version of the VQA 1.0 dataset, which reduces the language bias and requires the VQA model to be equipped with stronger, fined-grained recognition capability. VQA 2.0 is also larger in scale, containing over 204K images from the MS COCO dataset, over 1 million question and over 11 million answers. The dataset is composed of 443,757 pairs (image, question, answer) for training, 214,354 for validation and 447,793 for testing. The evaluation metric is the same as that in VQA 1.0.

4.2. Hyperparameters and regularization

The size of image features and question embeddings is $M = 2048$ and $N = 1280$ respectively. Following the previous works (e.g. [12]), K is a variable number and varies from 10 to 100, depending on the number of objects in an image. The length of questions is $T = 14$. Questions shorter than 14 words are head-padded up to 14. The embedding layer is initialized using 300-dimensional pre-trained Wikipedia+Gigaword GloVe word embedding [28]. We also use the 300-dimensional semantically-closed mixture of these embeddings to enhance Glove word embedding as in [13]. So, the dimension of the word embedding is 600. We choose MUTAN, which is a bilinear model proposed in [9], as our multimodal feature fusion method. We choose all the projection dimensions to be equal to each other: $t_q = t_v = t_o = 1280$, and a rank $R = 3$ in MUTAN. Every linear mapping is regularized by Weight Normalization and Dropout [29] ($p = 0.2$, except for the classifier with $p = 0.5$). Our model and its several variants are end-to-end trained. The Adamax optimizer [30], a variant of Adam based on infinite norm, is used. The learning rate is gradually increased from 0.0005 to 0.002 in the first four epochs. After 10 epochs, the learning rate is decayed by $1/4$ for every 2 epochs up to 13 epochs (i.e., $5e^{-4}$ for the 11th epoch and $1.25e^{-4}$ for the 13th epoch). We clip the 2-norm of vectorized gradients to 0.25. Limited by computing resources, the batch size is always 256 for training and testing. Our model is trained in an end-to-end way. All experiments are implemented in the PyTorch framework [31] and performed on workstations with two NVIDIA TITAN Xp GPUs.

4.3. Results on VQA 1.0 and VQA 2.0

For easy reference, we briefly introduce several important models used for comparison in the beginning of this section. A more detailed discussion of related works can be found in Section 2.

- **Bottom-up** [11] applies the bottom-up attention mechanism (based on Faster R-CNN) proposed in [12], which enables regions related to a question to be inferred at the level of objects. Bottom-up is the basis of our UFSCAN model. The performance improvements over Bottom-up are detailed in the ablation study.
- **MLB** (Multimodal low-rank bilinear pooling) [21] is proposed to tackle the problem of the computational cost of the bilinear model, while harnessing its sufficient representation capacity.
- **MFH** (Multimodal factorized high-order pooling) [7] extends MLB with high-order pooling to fuse multimodal features. It uses convolutional image features (based on the 152-layer ResNet model [32]) and word-level question features (i.e., each word with a corresponding feature). Furthermore, MFH simultaneously considers keywords in a question and image regions related to question. They are identified by question self-attention and question-guided visual spatial attention respectively.
- **BAN** (Bilinear attention network) [13] adopts bottom-up attention on image features and word-level question features. BAN generates an attention map by calculating bilinear interactions among each pair of image and question features, which can be understood as an affinity matrix between multimodal features. It then fuses multimodal feature by the attention map to obtain the joint representation.
- **Counter** [33] is specially developed for dealing with counting questions, which are challenging and require a model to identify which classes of objects to count, find where they are and add them up. The Counter component is portable, and it is used in BAN to enhance the ability of counting.

Table 1 shows our evaluation results on the VQA 1.0 test set. We compare our models with the results of a number of the state-of-the-art models, including the MFH model that is the current champion of the VQA 1.0 Challenge. As shown in Table 1, Our model UFSCAN outperforms all the methods, including the VQA 1.0 challenge champion MFH [7]. It significantly outperforms all models except the three based on MFH. For a fair comparison, the latest model MFH+CoAtt+Glove (bottom-up) also uses the same bottom-up attention features as used in UFSCAN and is trained with the same train set and val set. Notably, MFH incorporates the question spatial attention mechanism and uses more question features than UFSCAN. Still, UFSCAN outperforms the best-performing model of MFH, MFH+CoAtt+Glove (bottom-up), demonstrating the advantages of our proposed MulFA. Moreover, with data augmentation using the Visual Genome, our model UFSCAN + VG achieves the overall best accuracy of 70.19% and 70.24% on the test-dev set and test-standard set respectively. Therefore, UFSCAN achieve a state-of-the-art performance on VQA 1.0.

Table 2 compares performance of our model on the VQA 2.0 dataset with the current state-of-the-art models. Zhang et al. [33] proposed a counting module, Counter, to focus on dealing with counting questions. Some works [13,33] have demonstrated that this module can significantly improve the accuracy of answering counting questions. For a clearer comparison, Table 2 is split into two parts: the first part summarizes the methods without the counting module and the second part contains the methods that incorporate the counting module. All the models are trained on the same training and validation splits, and use Visual Genome for data augmentation.

As can be seen in the first part of Table 2 (W/o Counter), our UFSCAN model outperforms all the strong baseline methods, and achieves a state-of-the-art result in the case without the counting module. Although the latest models, BAN and MFH+Bottom-Up, use the question representation including features from each time step of the RNN, UFSCAN still outperforms MFH+Bottom-Up by 1.07% and performs better than BAN in all but the “Yes/no” category. The questions with “Number” and “Other” types of answers require more powerful fine-grained recognition ability to answer correctly. For these two types, UFSCAN outperforms the state-of-the-art model BAN by 0.32% on “Number” and 0.48% on “Other”.

Table 1
Test-dev and test-standard accuracy (%) of single models on the VQA 1.0 test set (Section 4.1), compared to state-of-the-art models. “–” indicates the result is not available. “Att” indicates the visual spatial attention mechanism. “CoATT” indicates the question and visual co-attention mechanism. “GloVe” indicates that the word embedding method [28] is adopted. “VG” indicates that the model uses the Visual Genome for data augmentation.

Model	Test-dev	Test-standard
LSTM Q + I [1]	57.8	58.2
SMem [4]	58.0	58.2
SAN [5]	58.7	58.9
FDA [34]	59.2	59.5
DMN+ [35]	60.3	60.4
HieCoAtt [6]	61.8	62.1
RAU [36]	–	64.1
MCB + Att + GloVe + VG [22]	65.4	–
MLB + Att + StV + VG [21]	65.8	–
MFH + CoAtt + GloVe [7]	66.8	66.9
MFH + CoAtt + GloVe + VG [7]	67.70	67.5
MFH + CoAtt + GloVe (bottom-up)[a]	68.78	–
UFSCAN (ours)	**69.06**	**69.34**
UFSCAN + VG (ours)	**70.19**	**70.24**

[a] Detonets a model found in https://github.com/asdf0982/vqa-mfb.pytorch, which does not seem to have been published.

The second part of Table 2 shows the results of state-of-the-art models with the counting module incorporated. UFSCAN+counter improves accuracy of the counting questions (i.e., “Number” type) over UFSCAN by 4.01%. UFSCAN+counter significantly outperforms the Counter model by 2.37% and is better than Ban+counter by 0.42%. Notably, UFSCAN+counter performs better on the ‘Other’ and ‘Number’ types, outperforming BAN+counter by 0.82% and 0.95% respectively. Overall, UFSCAN achieves a comparable performance on VQA 2.0 test set.

4.4. Ablation study on VQA 2.0

In this section, we perform extensive ablation studies on the VQA 2.0 validation dataset to quantify the role of proposed components in our model. The results are shown in Tables 3 and 4.

The Bottom-Up [11] uses the same image and question representations as those in our model. In addition, Bottom-Up includes a visual spatial attention method with 1-glimpse. It uses multimodal low-rank bilinear pooling [21] to fuse multimodal features. Therefore, the bottom-up model can be regarded as the baseline of our model. For a fair comparison, we use the same training strategies and settings to retrain this model with the same word embeddings. The result is reported in first row of Table 3.

We design a set of experiments to evaluate the sensitivity of model performance to each design choice. Specifically, we compare performance of the following variants of our model:

- IMulFA alone, where we only add IMulFA to the baseline. The goal of this comparison is to verify the effectiveness of IMulFA.
- QMulFA, where we only add QMulFA to the baseline.
- parallel MulFCoA, where we parallelly use both IMulFA and QMulFA in the baseline.
- alternate MulFCoA$_1$, where we first use IMulFA and then QMulFA in the baseline.
- alternate MulFCoA$_2$, where we first use QMulFA and then IMulFA in the baseline.
- VSA-1, where we only use a 1-glimpse visual spatial attention module with a bilinear model in the baseline.
- MulRM, where we only add MulRM to the baseline.
- UFSCAN-1, the full model that is the baseline added parallel MulFCoA, VSA-1, and MulRM.

Table 2
Test-dev and test-standard accuracy of single-model on the VQA 2.0 dataset (Section 4.1). "–" indicates the result is not available.

	Model	Test-dev accuracy (%)				Test-standard accuracy (%)			
		All	Yes/no	Number	Other	All	Yes/no	Number	Other
W/o counter	Bottom-up [11]	65.32	81.82	44.21	56.05	65.67	82.2	43.9	56.26
	MFH [7]	66.12	–	–	–	–	–	–	–
	MFH+Bottom-up [7,13]	68.76	84.27	49.56	59.89	–	–	–	–
	BAN [13]	69.66	**85.46**	50.66	60.50	–	–	–	–
	UFSCAN (ours)	**69.83**	85.21	**50.98**	**60.98**	**70.09**	**85.51**	**50.21**	**61.22**
Counter	Counter [33]	68.09	83.14	51.62	58.97	68.41	83.56	51.39	59.11
	BAN + counter [13]	70.04	85.42	54.04	60.26	70.35	–	–	–
	UFSCAN + counter (ours)	**70.46**	**85.52**	**54.99**	**61.08**	**70.73**	**85.87**	**54.37**	**61.30**

Table 3
Ablations of a single network, evaluated on the VQA 2.0 validation set (Section 4.1). We train each model with three different random seeds and report the mean and standard deviation (±).

Component	Setting	Accuracy (%)
Baseline	Bottom-up[a] [11]	64.60 ± 0.10
	VSA-1 only	64.80 ± 0.01
	MulRM only	65.18 ± 0.03
Multimodal feature-wise attention	IMulFA only	65.28 ± 0.10
	QMulFA only	65.27 ± 0.05
Multimodal feature-wise co-attention	Parallel MulFCoA	65.70 ± 0.02
	Alternate MulFCoA$_1$	65.77 ± 0.07
	Alternate MulFCoA$_2$	65.73 ± 0.07
Full model	UFSCAN-1	66.23 ± 0.15

[a] Denotes the result of bottom-up is obtained using the same word embeddings, training strategies, and settings with our model.

Table 4
Validation scores on VQA 2.0 dataset for the state of the art and the different number of glimpses of our proposed UFSCAN. The standard deviations are reported after ± using three random initializations.

Model	Accuracy
Bottom-up [11]	63.37 ± 0.21
Counter [33]	65.42 ± 0.10
BAN-1 [13]	65.36 ± 0.14
BAN-12 [13]	66.04 ± 0.08
DCAF [37]	65.7
UFSCAN-1	66.23 ± 0.15
UFSCAN-2	66.58 ± 0.12
UFSCAN-3	**66.69 ± 0.11**
UFSCAN-4	66.50 ± 0.06
UFSCAN-2 (add)	65.90 ± 0.15
UFSCAN-2 (concat)	65.73 ± 0.04

Table 3 lists the comparison of our full model (UFSCAN) w.r.t these ablations on the VQA 2.0 validation set. As one must submit their test results to a remote evaluation server to acquire the testing scores, test set is not recommended being used for such experiments.

We first evaluate the effectiveness of VSA that uses a bilinear model. As shown in Table 3, VSA-1 improves the performance by 0.20%. This result demonstrates the advantage of the bilinear model on learning discriminative multimodal fusion features. This observation is consistent with previous work [9].

We compare the performance of MulRM and the baseline, in Table 3. It increases the performance by 0.58%. For multiple glimpses, we compare MulRM (used in UFSCAN-2) with other aggregation method (i.e., feature concatenation, element-wise summation) in the bottom of Table 4. The result of the concatenation method obviously declines, with a drop of 0.85%. The performance of the element-wise summation method is worse than the UFSCAN-2 model by 0.68%. These results demonstrate the advantage of MulRM on learning the joint representation from multimodal information.

We investigate the efficacy of the feature-wise attention mechanism. As shown in Table 3, IMulFA and QMulFA both outperform baseline, with a rise of 0.68% and 0.67% respectively. It illustrates that the multimodel feature-wise attention mechanism is effective for both of image and question modalities. We hypothesize the feature-wise attention mechanism can provide finer features to enhance fine-grained recognition ability of a model.

Next, we evaluate the MulFCoA. We compare our proposed three co-attention mechanisms, i.e. two alternating co-attention mechanisms and one parallel co-attention mechanism. Their performances are not much different, while all are better than IMulFA or QMulFA used alone by about 0.50%. These results indicate that the order of IMulFA and QMulFA has a small effect, and their efficacy can be superimposed. By adding MulFCoA to baseline, the performance can be improved by about 1.1%, which might be interpreted as that MulFCoA can simultaneously highlight informative features of image and question representations, and ignore unuseful ones. We select parallel MulFCoA in the full model, because it performs well and stable.

The last row of Table 3 shows the result of the full model. The model uses the parallel MulFCoA, 1-glimpse VSA and MulRM. Comparing with the baseline, the overall performance significantly increases by 1.63%. This improvement could be interpreted as the superposition of effect from above three modules. This result demonstrates our proposed modules are compatible with each other.

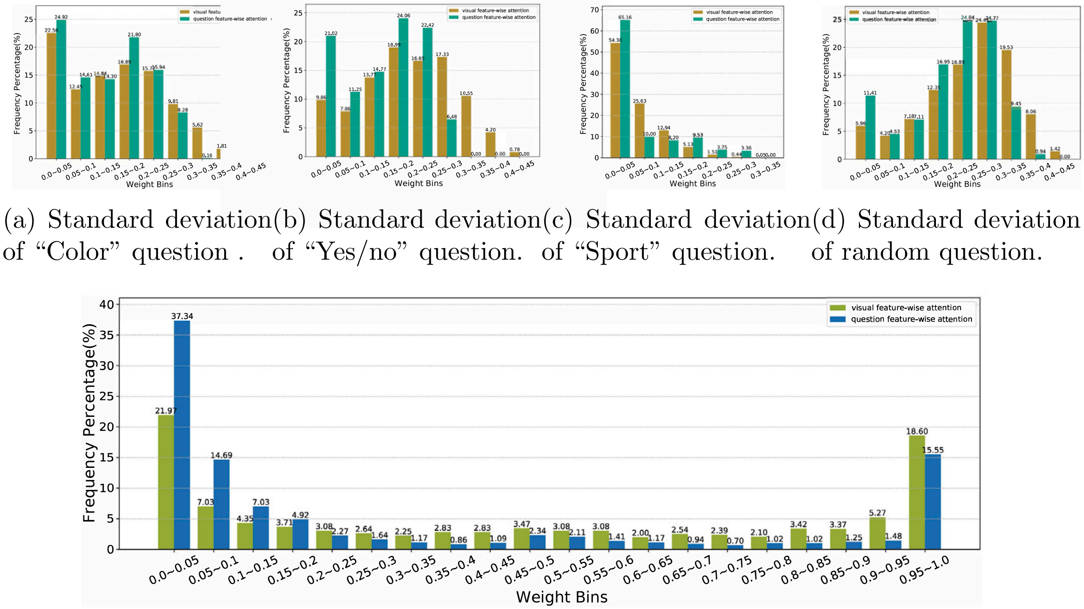
As shown in the middle part of Table 4, we investigate the influence of the number of glimpses (including 1, 2, 3, and 4). The results show the model with 3-glimpses performs best. We hypothesize that this is because a question is related to multiple objects in image, but overmuch glimpses introduce noise. Three glimpses embrace the best generalization ability.

The performance of a number of the state-of-the-art models on the validation set on VQA 2.0 is also summarized in the top of Table 2. The first row, Bottom-up, is the 2017 VQA Challenge winner architecture. UFSCAN-3 significantly outperforms it by 3.32% on average. The second row, Counter, is a counting module that focuses on dealing with counting question. UFSCAN-3 significantly outperforms this module by 1.27% on average. BAN-12 is the best-performing variant of BAN, compare it based on 12 glimpses, our UFSCAN-3 still outperforms BAN-12 by 0.65% on average. As can be seen, we achieve the best performance on the VQA 2.0 validation set.

4.5. Qualitative analysis

To qualitatively valuate effectiveness of the feature-wise attention mechanism, we calculate the mean and standard deviation of the attention weights for each feature channel of 100 question–image pairs from three typical types of questions: Yes/no, Sports and Color. As a comparison, we also calculate the mean and standard deviation for a number of randomly selected questions. Fig. 5 shows the percentage of the number of feature channels whose mean and standard deviation fall into each 0.05 interval respectively. From Fig. 5, we have the following observations:

1. The mean of feature attention weight of image and question features falls to various intervals in Fig. 5(e). This demonstrates



(a) Standard deviation of "Color" question .

(b) Standard deviation of "Yes/no" question.

(c) Standard deviation of "Sport" question.

(d) Standard deviation of random question.

(e) Mean of "Sport" type question.

Fig. 5. Frequency distribution histograms of the mean and standard deviation of visual and question feature-wise attention weights in a same feature channel. In each frequency distribution histogram, the *x*-axis corresponds to the channel attention weight value bins and the *y*-axis corresponds to the percentage of frequency. First four figures denote standard deviation histograms of feature-wise attention weight of a given question type. There are three question types from (a) to (c): "Color", "Yes/no" and "Sport". For clear comparing, we randomly choose questions from different types in (d). The fifth sub-figure shows mean histograms of attention weight of "Sport" type question. The figure of other two types of a question is similar. Best viewed in color.

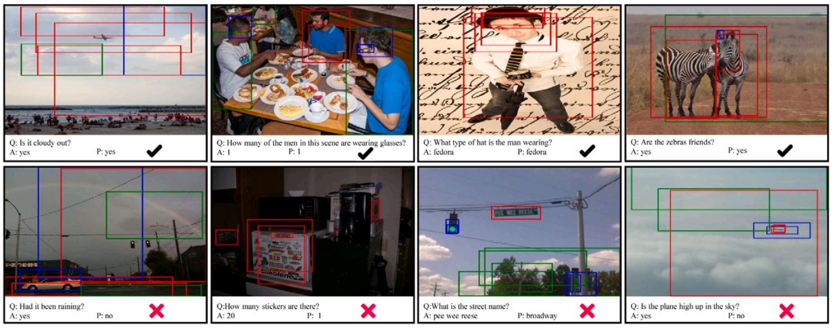

Q: Is it cloudy out?
A: yes P: yes ✓
Q: How many of the men in this scene are wearing glasses?
A: 1 P: 1 ✓
Q: What type of hat is the man wearing?
A: fedora P: fedora ✓
Q: Are the zebras friends?
A: yes P: yes ✓
Q: Had it been raining?
A: yes P: no ✗
Q:How many stickers are there?
A: 20 P: 1 ✗
Q:What is the street name?
A: pee wee reese P: broadway ✗
Q: Is the plane high up in the sky?
A: yes P: no ✗

Fig. 6. Examples of the visual attention of two-glimpse UFSCAN model on the VQA 2.0 validation data set. First row: four examples of correct predictions. Second row: four incorrect examples. In each example, the image, question (Q), ground truth (A), and prediction (P) are presented. The top 5 object bounding boxes of attention weight for each glimpse are drawn in each image. Red indicates the attention weight of the bounding box is relatively high in both of two glimpses. Blue and green denote the attention weight only is high at a certain glimpse. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different feature channels are assigned different weights. This shows that the feature-wise attention blocks have learned to emphasize informative feature signals and suppress less useful ones.

2. Comparing the four histograms of standard deviation distributions in Fig. 5(a)–(d), the standard deviation of the specific types of questions distributes across a smaller intervals, while that of the randomly selected questions falls in larger intervals. This demonstrates that the same feature channels for the same type of questions have similar weights, which in turn illustrates that a specific type of questions often entails some specific semantics.

To demonstrate the effects of visual spatial attention and explore the weakness of our approach, we visualize the learned visual spatial attention of some examples from validation data set in Fig. 6. The examples are randomly picked from each question type. The visual spatial attention generally can focus on the object relevant to the question. From the negative examples, we can find several limitations of our method. Firstly, our method fails to answer questions that require some commonsense knowledge (e.g., first and fourth examples at second row). Introducing an external knowledge base might improve complex reasoning and alleviate this problem. Secondly, the performance of our method is limited by the quality of the object detector (e.g., second examples at second row). Duplicated and missed detection would mislead the model to give an incorrect answer. Thirdly, since our method is not designed to read, it will perform poorly if the answer involves text in the image.

5. Conclusion

In this paper, we introduce a feature-wise attention mechanism for VQA to emphasize on informative features and suppress irrelevant ones, to extract more discriminative features for image and question representations. We design novel modules to model the question-guided image feature-wise attention and the image-guided question feature-wise attention, and combine the visual spatial attention with feature-wise attention to develop a new network model for VQA, the union feature-wise and spatial co-attention network (UFSCAN). The attention along multiple dimensions allows our VQA model to enjoy powerful fine-grained recognition ability. Our experimental results demonstrate that our method has achieved the state-of-the-art performance on two large-scale, real-world VQA datasets.

CRediT authorship contribution statement

Sheng Zhang: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Min Chen:** Conception and design of study, Writing - original draft, Writing - review & editing. **Jincai Chen:** Writing - original draft. **Fuhao Zou:** Conception and design of study, Analysis and/or interpretation of data. **Yuan-Fang Li:** Analysis and/or interpretation of data, Writing - original draft. **Ping Lu:** Acquisition of data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61672246, No. 61272068, No. 61672254, and Program for Hust Academic Frontier Youth Team, China. In addition, we gratefully acknowledge the support of NVIDIA Corporation (United States) with the donation of the Titan Xp GPU used for this research. All authors approved the version of the manuscript to be published.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual question answering, in: Proc. IEEE Int. Conf. Computer Vision (ICCV), 2015, pp. 2425–2433, http://dx.doi.org/10.1109/ICCV.2015.279.

[2] M. Malinowski, M. Fritz, Towards a visual turing challenge, 2014, CoRR, abs/1410.8027, URL: arXiv:1410.8027.

[3] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 4613–4621, http://dx.doi.org/10.1109/CVPR.2016.499.

[4] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VII, in: Lecture Notes in Computer Science, vol. 9911, Springer, 2016, pp. 451–466, http://dx.doi.org/10.1007/978-3-319-46478-7_28.

[5] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 289–297, URL: http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering.pdf.

[7] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learning Syst. 29 (12) (2018) 5947–5959, http://dx.doi.org/10.1109/TNNLS.2018.2817340.

[8] H. Nam, J. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2156–2164, http://dx.doi.org/10.1109/CVPR.2017.232.

[9] H. Ben-younes, R. Cadène, M. Cord, N. Thome, MUTAN: multimodal tucker fusion for visual question answering, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2631–2639, http://dx.doi.org/10.1109/ICCV.2017.285.

[10] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[11] D. Teney, P. Anderson, X. He, A. van den Hengel, Tips and tricks for visual question answering: Learnings from the 2017 challenge, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 4223–4232, http://dx.doi.org/10.1109/CVPR.2018.00444, URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Teney_Tips_and_Tricks_CVPR_2018_paper.html.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2018, pp. 6077–6086, http://dx.doi.org/10.1109/CVPR.2018.00636.

[13] J. Kim, J. Jun, B. Zhang, Bilinear attention networks, in: S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, MontrÉAl, Canada., 2018, pp. 1571–1581, URL: http://papers.nips.cc/paper/7429-bilinear-attention-networks.

[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735, URL: arXiv:https://doi.org/10.1162/neco.1997.9.8.1735.

[15] K. Cho, B. van Merrienboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in: D. Wu, M. Carpuat, X. Carreras, E.M. Vecchi (Eds.), Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, Association for Computational Linguistics, 2014, pp. 103–111, http://dx.doi.org/10.3115/v1/W14-4012, URL: https://www.aclweb.org/anthology/W14-4012/.

[16] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D.J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, in: Lecture Notes in Computer Science, vol. 8689, Springer, 2014, pp. 818–833, http://dx.doi.org/10.1007/978-3-319-10590-1_53.

[17] J. Kim, S. Lee, D. Kwak, M. Heo, J. Kim, J. Ha, B. Zhang, Multimodal residual learning for visual QA, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 361–369, URL: http://papers.nips.cc/paper/6446-multimodal-residual-learning-for-visual-qa.

[18] X. Chen, C.L. Zitnick, Learning a recurrent visual representation for image caption generation, 2014, CoRR, abs/1411.5654, URL: arXiv:1411.5654.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: F.R. Bach, D.M. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, in: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, 2015, pp. 2048–2057, URL: http://jmlr.org/proceedings/papers/v37/xuc15.html.

[20] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 6298–6306, http://dx.doi.org/10.1109/CVPR.2017.667.

[21] J. Kim, K.W. On, W. Lim, J. Kim, J. Ha, B. Zhang, Hadamard product for low-rank bilinear pooling, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017, URL: https://openreview.net/forum?id=r1rhWnZkg.

[22] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: J. Su, X. Carreras, K. Duh (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, The Association for Computational Linguistics, 2016, pp. 457–468, URL: http://aclweb.org/anthology/D/D16/D16-1044.pdf.

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73, http://dx.doi.org/10.1007/s11263-016-0981-7.

[24] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1724–1734, URL: http://aclweb.org/anthology/D/D14/D14-1179.pdf.

[25] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 7132–7141, http://dx.doi.org/10.1109/CVPR.2018.00745, URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

[26] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: D.J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, in: Lecture Notes in Computer Science, vol. 8693, Springer, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[27] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in VQA matter: Elevating the role of image understanding in visual question answering, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6325–6334, http://dx.doi.org/10.1109/CVPR.2017.670.

[28] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543, URL: http://aclweb.org/anthology/D/D14/D14-1162.pdf.

[29] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958, URL: http://dl.acm.org/citation.cfm?id=2670313.

[30] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, CoRR, abs/1412.6980, URL: arXiv:1412.6980.

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, 2017.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[33] Y. Zhang, J. Hare, A. Prügel-Bennett, Learning to count objects in natural images for visual question answering, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, URL: https://openreview.net/forum?id=B12Js_yRb.

[34] I. Ilievski, S. Yan, J. Feng, A focused dynamic attention model for visual question answering, 2016, CoRR, abs/1604.01485, URL: arXiv:1604.01485.

[35] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: M. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, in: JMLR Workshop and Conference Proceedings, vol. 48, JMLR.org, 2016, pp. 2397–2406, URL: http://jmlr.org/proceedings/papers/v48/xiong16.html.

[36] H. Noh, B. Han, Training recurrent answering units with joint loss minimization for VQA, 2016, CoRR, abs/1606.03647, URL: arXiv:1606.03647.

[37] F. Liu, J. Liu, Z. Fang, R. Hong, H. Lu, Densely connected attention flow for visual question answering, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 869–875, http://dx.doi.org/10.24963/ijcai.2019/122.