ALSA: Adversarial Learning of Supervised Attentions for Visual Question Answering

Yun Liu, Xiaoming Zhang, Zhiyun Zhao, Bo Zhang, Lei Cheng, and Zhoujun Li, *Member, IEEE*

***Abstract*—Visual question answering (VQA) has gained increasing attention in both natural language processing and computer vision. The attention mechanism plays a crucial role in relating the question to meaningful image regions for answer inference. However, most existing VQA methods: 1) learn the attention distribution either from free-form regions or detection boxes in the image, which is intractable in answering questions about the foreground object and background form, respectively and 2) neglect the prior knowledge of human attention and learn the attention distribution with an unguided strategy. To fully exploit the advantages of attention, the learned attention distribution should focus more on the question-related image regions, such as human attention for both the questions, about the foreground object and background form. To achieve this, this article proposes a novel VQA model, called adversarial learning of supervised attentions (ALSAs). Specifically, two supervised attention modules: 1) free form-based and 2) detection-based, are designed to exploit the prior knowledge for attention distribution learning. To effectively learn the correlations between the question and image from different views, that is, free-form regions and detection boxes, an adversarial learning mechanism is implemented as an interplay between two supervised attention modules. The adversarial learning reinforces the two attention modules mutually to make the learned multiview features more effective for answer inference. The experiments performed on three commonly used VQA datasets confirm the favorable performance of ALSA.**

***Index Terms*—Adversarial learning, supervised attention, visual question answering (VQA).**

Manuscript received May 16, 2020; revised July 31, 2020; accepted October 2, 2020. This work was supported in part by the Beijing Natural Science Foundation of China under Grant 4182037; in part by the National Natural Science Foundation of China under Grant U1636210 and Grant U1636211; and in part by the Open Research Fund from Shenzhen Research Institute of Big Data under Grant 2019ORF01012. This article was recommended by Associate Editor J. Han. *(Corresponding author: Zhiyun Zhao.)*

Yun Liu is with the Beijing Key Laboratory of Network Technology, Beihang University, Beijing 100191, China.

Xiaoming Zhang and Bo Zhang are with the Key Laboratory of Aerospace Network Security, Ministry of Industry and Information Technology, School of Cyberspace Science and Technology, Beihang University, Beijing 100191, China.

Zhiyun Zhao is with the Cyber Security Institute, National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China (e-mail: zzy@cert.org.cn).

Lei Cheng is with the Intelligent Application Technology Research Center, Shenzhen Research Institute of Big Data, Shenzhen 518000, China.

Zhoujun Li is with the State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2020.3029423

I. Introduction

RECENTLY, with the rapid development of computer vision and natural language processing, problems of combining both visual image and textual language have steadily inspired considerable research attention. Visual question answering (VQA) [1]–[3] has emerged as a challenging task and attracted broad interest in both industry and academia. VQA requires the algorithms to infer answers for natural language questions about the contents of the given images. Similar to other visual-language tasks, such as cross-modal retrieval [4]–[6] and image captioning [7]–[9], the VQA task requires a deep understanding of both the visual image and the textual question. VQA plays a significant role in various applications, e.g., human–machine interaction, medical assistance, and automatic customer service [10].

Existing VQA methods mainly utilize visual attention mechanisms to explore the question-related image regions for answer inference [11]–[13]. However, there are two challenges in the existing VQA methods. On the one hand, most visual attention mechanisms in VQA can be categorized into free form-based methods [14]–[16] and detection-based methods [17]–[19]. In the free form-based methods, the image is evenly divided into many regions and the attention is assigned across these regions. Although it can freely map the attention to regions of any size, it might focus on partial objects or irrelevant context. Take the question "What is the animal on the grass?" shown in Fig. 1(a) as an example, the free form-based attention focuses on only a part of the foreground dog and then generates the wrong answer of "cat." It indicates that free form-based attention is intractable in answering questions about a foreground object. In addition, the detection-based methods aim to learn the attention distribution over the prespecified detection boxes of the image. However, it is not effective in answering questions about the background form. For the other instance "Is this a sunny day?" shown in Fig. 1(a), exact boxes about the background sky might not exist in the image, which will lead to a failure answer. Apparently, it is a major challenge to learn more effective attention distribution for both questions about the foreground object and background form. On the other hand, attention distribution is the normalized importance of each region or detected box in an image. Most of the current VQA approaches learn the attention distribution by considering only the correlations between the question and image with an unsupervised strategy. These approaches have achieved certain success when the correlations between question words and image regions

or boxes are explicit. However, it is difficult to learn the attention distribution when the correlations are obscure. As the example shown in Fig. 1(b), the implicitly generated attention map focuses on the image regions unrelated to the object of the question and then obtains a wrong answer. However, the human being concentrates on the other regions related to the right answer. Therefore, learning effective attention distribution, like human attention, becomes another challenge.

To tackle these challenges, some efforts have been made in previous VQA methods. For the first challenge, the method presented in [11] linearly combines the free form-based attention module and the detection-based attention module to infer the answer. However, it is ineffective in capturing the internal relation and the complementary nature of the two types of attention mechanisms. For the second challenge, various attention models, including dual attention [16], multilevel attention [12], and cubic attention [20] are explored to learn the correlations in the question–image pairs. The weakness of these methods is that they learn the correlations between the question and image through an automatic exploratory method without supervision, which might fail to focus on the question-related part of the image.

Different from previous VQA methods, we aim to address the problems from the following two aspects. First, to effectively answer both of the questions about the foreground object and background form, it is intuitive to learn the image features which can capture both views, that is, foreground object and background form. As discussed above, it has been demonstrated that the detection-based and free form-based attention modules are effective to learn the features from the two views, respectively. Thus, by reinforcing the two attention modules to learn the complementary knowledge from each other, more powerful attention modules can be obtained to capture the features to effectively reflect both of the questions about the foreground object and background form. The recently proposed adversarially learned inference models [21], [22] have achieved great success to learn mutually coherent inference, which relies on adversarial networks to reinforce the generator to preserve underlying cross-view semantic structure in data. Therefore, we argue that treating the two attention modules as two generators in adversarial networks helps to effectively integrate them for answer inference. Second, to infer the answer for a given question–image pair, the learned attention distribution should be as consistent as possible with the human attention map. Namely, human attention can be considered as prior knowledge to learn effective attention distribution. By exploiting the human-annotated attention maps on the images, the learned attention distribution can be more effective to refect the answer.

In this article, we take full advantage of both multiview correlations and the prior knowledge of human-annotated attention maps for VQA. In particular, we investigate: 1) how to effectively learn the correlations in question–image pairs from different views, that is, free-form regions and detection boxes and 2) how to exploit the prior knowledge of human attention to learn more effective attention distribution on the image. Our solutions to these questions result in a new model, that is, adversarial learning of supervised attentions (ALSAs), which fuses two types of attention modules with

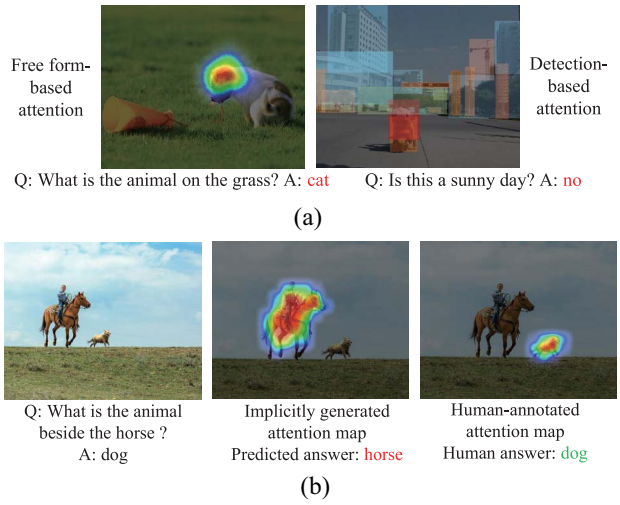

Fig. 1. Some examples about the previous attention-based VQA methods. (a) Free form-based and detection-based attention models. (b) Different attention maps and the corresponding answers.

adversarial networks for VQA. First, free form-based attention and detection-based attention modules are pretrained on an attention-annotated dataset to learn the prior knowledge of human attention. Second, to effectively learn the correlations between image and question from different views, adversarial networks are conducted between the free form-based attention and the detection-based attention modules to mutually reinforce them to learn the complementary knowledge. Concretely, a discriminator is designed to distinguish the multiview features learned from the two types of attention modules. The two attention modules act as two generators that try to confuse the discriminator that the multiview features are generated from the other module. In this way, the two attention modules are reinforced to learn more effective multiview features from different views. The major contributions are concluded as follows.

1) Unlike existing attention-based methods, we employ prior knowledge of human-annotated attention maps to learn two supervised attention modules for more effective attention distribution leaning.
2) Adversarial networks are constructed between two supervised attention modules learned from different views to reinforce them to effectively answer both of the questions about the foreground object and background form.
3) A novel VQA model ALSA is proposed and the fusion scheme to adopt adversarial learning on different attention models can be flexibly extended to more VQA models and other scenarios for better performance.

The remainder of this article is organized as follows. We first review the related works of VQA and adversarial learning. Then, the approach ALSA is introduced in detail. Next, the experiment results and further analysis are presented, and this article is concluded in the end.

II. Related Work

A. VQA

Most of the existing VQA approaches are based on deep neural networks [23]–[26]. Among these approaches, two

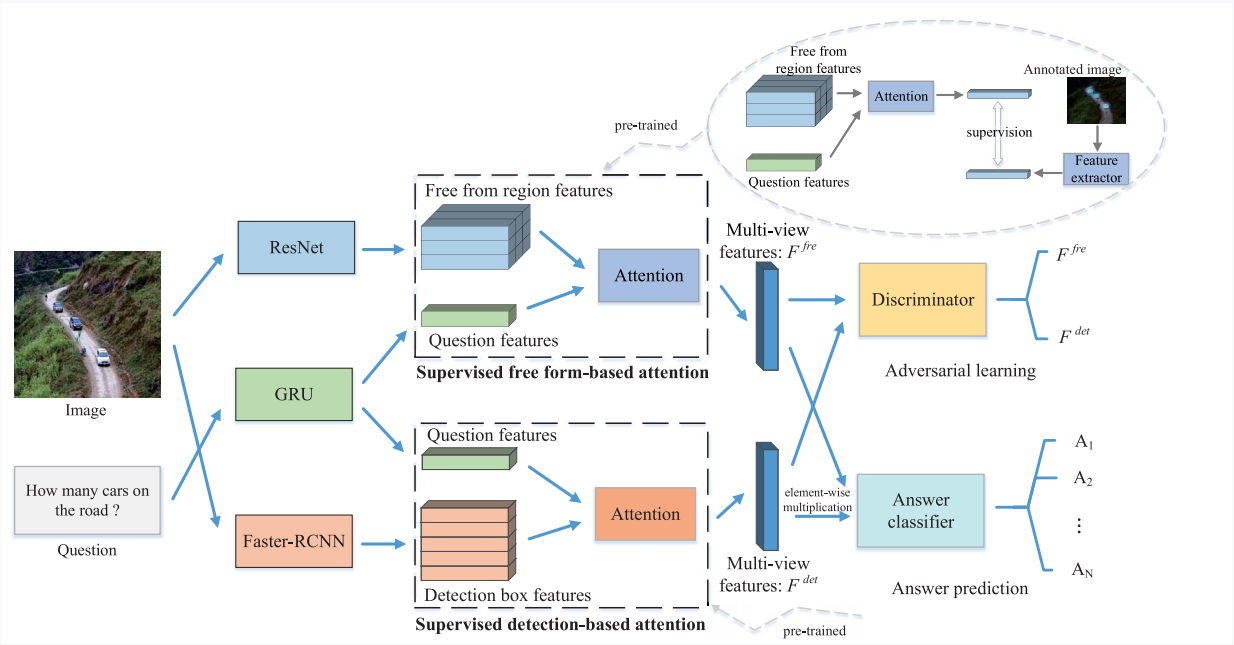

Fig. 2. Framework of the proposed ALSA for VQA. The supervised free-form based attention and the supervised detection-based attention modules are pretrained on the attention-annotated dataset. An illustration of the pretrained supervised attention module is shown in the oval dashed box. Adversarial learning is conducted between the multiview features learned from the supervised attention modules and the discriminator. The two multiview features are combined by elementwise multiplication before input to the answer classifier for answer inference.

types of attention mechanisms play an important role in answer inference. The first type of mechanism is free form-based attention method [14], [15], [27]. The works presented in [15], [16], and [28] utilize dual attention or multilevel attention networks to relate the question to meaningful image regions multiple times. The approach introduced in [11] fuses multiview features learned from different attention models by a multimodal embedding scheme. A visual attention mechanism designed in [20] applies a particular channel and spatial attention on object regions. The work presented in [29] proposes an object-difference attention model to compare objects explicitly using different operators to calculate the attention distribution. The second type of mechanism is the detection-based attention method [17], [19], [30]. The work presented in [31] encodes the fine-grained relations and the more sophisticated trinary relations between detection objects using two relation attention modules. The approach detailed in [32] proposes bilinear attention networks that find bilinear attention distributions to exploit vision-language information. A combined bottom-up and top-down attention mechanism is proposed in [18], which calculates attention weights based on the objects and other prominent image regions. In addition, some extension methods on the VQA task also achieve promising performance. MCB [33] and MLB [34] both adopt bilinear models with multimodal pooling scheme to learn the multimodal embedding for answer prediction. In MUTAN [35], a generalized multimodal pooling framework is proposed, which shows that MCB and MLB are its special cases. External knowledge bases, such as DBpedia, are used as the supplementary information of the image for more effective answer inference [36]–[39]. Different from these approaches, our method focuses on exploiting the prior knowledge of human attention and learning the attention distribution from different views to capture more effective multiview correlations in the question–image pairs.

B. Adversarial Learning

Recently, adversarial network [40] has achieved great success in representation learning [41], [42], cross-modal retrieval [43]–[45], and sequence generation [46]–[48]. The core of adversarial learning is the interaction between the feature generator and the discriminant classifier, conducted as a minimax game. On the one hand, the feature generator strives to generate new features with the ability to confuse the discriminant classifier. On the other hand, the discriminant classifier attempts to correctly distinguish the features generated by the feature generator and in this way guides the continuous learning of the feature generator. A cross-modal retrieval model proposed in [43] uses adversarial networks to learn more effective representations of different modalities for better similarity comparison between cross-modal items. The framework of IRGAN is proposed in [49], which provides a principled training environment that combines two kinds of retrieval models, that is, generative models and discriminative models. The approach detailed in [21] displays a multiview adversarially learned inference model. It relies on shared representations of cross-domain data to generate an arbitrary number of paired faking samples, based on which very few paired samples are enough to learn good mapping. A sequence generation method SeqGAN [50] is designed to train a generative adversarial network for structured sequences generation via policy gradient. Unlike previous approaches [11], [51], [52] that linearly fuse multimodal features from different attentional modules, our method focuses on capturing the

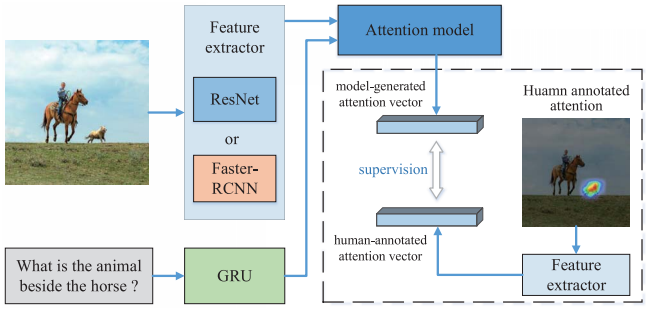

Fig. 3. Architecture of the supervised attention model.

complementarity of different attentional modules with the use of adversarial learning to reinforce the model's learning capacity. To the best of our knowledge, although adversarial learning has been proven successful in many scenarios, it has not been fully exploited in the attention-based VQA approach.

III. ALSA for Visual Question Answering

A. Problem Statement

Before the problem formulation, we define several notations used in this article. There are two modalities of data to be considered, that is, visual image and textual sentence. Let $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ denote the images and $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_n\}$ denote the corresponding questions, where n is the number of samples. Moreover, $\mathcal{A} = \{A_1, A_2, \ldots, A_t\}$ is denoted as the answers, where t is the number of the answer classes in the experiment dataset. Our work is to train a neural network that can predict the correct answer for a given question–image pair.

Fig. 2 illustrates the framework of ALSA. Specifically, two supervised attention modules, free form-based and detection-based, are first designed and pretrained on the attention-annotated dataset VQA-HAT (human attention) [53] using a supervised learning method. Then, the two attention modules initialized with the learned weights are used to capture the correlations between the question and image from different views to learn effective attention distribution. After that, adversarial networks are employed to reinforce the two attention modules mutually to produce more effective multiview features. Finally, the learned multiview features are fed into an answer classifier for answer prediction.

B. Supervised Attention Models

Existing VQA approaches mainly employ attention mechanisms to automatically capture the correlations between the image and question. However, these attention models learn the attention distribution using an unsupervised strategy without any prior knowledge, which results in that many unrelated image regions are focused. To address this problem, we argue that the attention distribution learned by the attention models should be as close to the human-annotated attention map as possible. For this goal, we design two supervised attention modules, that is, free form-based and detection-based, by pretraining them on the attention-annotated VQA-HAT [53] dataset to learn more effective multiview correlations between the question and image. The architecture of the supervised attention model is shown in Fig. 3.

1) Free Form-Based Attention: Given an image V_i, the pretrained deep networks ResNet [54] is applied to extract the features of free-form regions as $R_i \in \mathbb{R}^{d\times m\times m}$, where $m \times m$ is the number of regions and d is the dimensionality of the feature vector of each region. For question Q_i, the pretrained word embeddings are used to encode each word and feed them into a GRU [55]. The output of the last cell is used as the question representation $h_i \in \mathbb{R}^e$.

To exploit the correlations between image region features R_i and question features h_i, we first embed them into a c-dimensional space

$$R_i^{(c)} = \tanh(W_r R_i + b_r), \quad R_i^{(c)} \in \mathbb{R}^{c\times m\times m} \tag{1}$$

$$h_i^{(c)} = \tanh(W_{h1} h_i + b_{h1}), \quad h_i^{(c)} \in \mathbb{R}^c \tag{2}$$

where $W_r \in \mathbb{R}^{c\times d}$ and $W_{h1} \in \mathbb{R}^{c\times e}$ are the trainable weight matrix, and b_r and b_{h1} are bias parameters. Then, h_i^c is spatially replicated to form $H_i^{(c)} \in \mathbb{R}^{c\times m\times m}$, which matches the spatial size of $R_i^{(c)}$.

The elementwise multiplication is used to fuse $R_i^{(c)}$ and $H_i^{(c)}$. Then, the attention map is calculated by convolving the fused representation with 1×1 kernel followed by a softmax function over the $m \times m$ regions as follows:

$$M_i^{(c)} = R_i^{(c)} \odot H_i^{(c)}, \quad M_i^{(c)} \in \mathbb{R}^{c\times m\times m} \tag{3}$$

$$\alpha_i = \text{softmax}\left(W_\alpha * M_i^{(c)} + b_\alpha\right), \quad \alpha_i \in \mathbb{R}^{m\times m} \tag{4}$$

where $\odot$ represents the elementwise multiplication, and $*$ denotes the convolutional operation. $W_\alpha \in \mathbb{R}^{c\times 1\times 1}$ and $b_\alpha \in \mathbb{R}^c$ are the trainable parameters. After that, the attention scores are used to modulate the strength of the attention on different regions

$$\hat{R}_i = \sum_j^{m\times m} \alpha_{i,j} R_{i,j}, \quad \hat{R}_i \in \mathbb{R}^c. \tag{5}$$

Compared with the original image region features R_i, the attentive region feature vector $\hat{R}_i$ is more effective to reflect the relevance to the corresponding question Q_i. Then, $\hat{R}_i$ and $h_i^{(c)}$ are further fed into a multilayer perception (MLP) to learn a multiview feature vector F_i^{fre} in an f-dimensional latent space. We pipeline the whole procedure of learning F_i^{fre} as a vector generation function

$$F_i^{\text{fre}} = g_{fre}\left(Q_i, V_i; \theta_{fre}\right), \quad F_i^{\text{fre}} \in \mathbb{R}^f \tag{6}$$

where $g_{fre}(\cdot)$ simulates the whole free form-based attention networks to obtain F_i^{fre}, and θ_{fre} is the parameter set.

2) Detection-Based Attention: Given an image V_i, the pretrained Faster-RCNN [56] is used to obtain object detection boxes. The proposals of detected boxes are represented as $D_i \in \mathbb{R}^{l\times k}$, where k is a hyperparameter to represent the number of the detected boxes and l is the dimensionality of the feature vector of each box. For question Q_i, the pretrained word embeddings and GRU are used to encode the question into a vector $h_i \in \mathbb{R}^e$.

Similar to free form-based attention, we first embed D_i and h_i into a s-dimensional space

$$D_i^{(s)} = \tanh(W_d D_i + b_d), \quad D_i^{(s)} \in \mathbb{R}^{s\times k} \tag{7}$$

$$h_i^{(s)} = \tanh(W_{h2} h_i + b_{h2}), \quad h_i^{(s)} \in \mathbb{R}^{s} \tag{8}$$

where $W_d \in \mathbb{R}^{s\times l}$ and $W_{h2} \in \mathbb{R}^{s\times e}$ are the trainable weight matrices. b_d and b_{h2} are the bias parameters. Then, $h_i^{(s)}$ is spatially replicated to form $H_i^{(s)} \in \mathbb{R}^{s\times k}$, which matches the spatial size of $D_i^{(s)}$.

The following functions are applied to calculate the question-related attentive detection features:

$$M_i^{(s)} = D_i^{(s)} \odot H_i^{(s)}, \quad M_i^{(s)} \in \mathbb{R}^{s\times k} \tag{9}$$

$$\beta_i = \text{softmax}\Big(W_\beta M_i^{(s)} + b_\beta\Big), \quad \beta \in \mathbb{R}^{k} \tag{10}$$

$$\hat{D}_i = \sum_j^k \beta_{i,j} D_{i,j}, \quad \hat{D}_i \in \mathbb{R}^{s} \tag{11}$$

where $W_\beta \in \mathbb{R}^{1\times s}$ and b_β are the trainable matrix and bias term, respectively. Then, $\hat{D}_i$ and $h_i^{(s)}$ are further fed into a MLP to learn a multiview feature vector F_i^{det} in the same f-dimensional latent space as the space in the free-form based attention model. The whole process to generate F_i^{det} from the question–image pair (Q_i, V_i) can be denoted as follows:

$$F_i^{\text{det}} = g_{\text{det}}(Q_i, V_i; \theta_{\text{det}}), \quad F_i^{\text{det}} \in \mathbb{R}^{f} \tag{12}$$

where $g_{\text{det}}(\cdot)$ simulates the whole detection-based attention networks to obtain F_i^{det}, and θ_{det} is the parameter set.

3) Supervised Attention Model: To learn more effective attention distribution, the attention maps in the attention-annotated VQA dataset VQA-HAT [53] are used as prior knowledge to guide the learning of the two types of attention modules. In the VQA-HAT dataset, human attention maps labeled by 800 annotators are collected for more than 60k question–image pairs from the VQA v1.0 dataset [1]. For each question–image pair, the annotators are asked to sharpen the regions that will help them answer the question correctly. Specifically, each sample in the VQA-HAT dataset can be represented as a triplet of (Q_i, V_i, γ_i), where γ_i is the human-annotated attention map for the given question–image pair (Q_i, V_i). For each (Q_i, V_i), we, respectively, use (1)–(4) and (7)–(10) to extract the free form-based attention vector α_i and the detection-based attention vector β_i. Similar to the feature extractors used in the free form-based attention and detection-based attention modules, ResNet and Faster-RCNN are, respectively, used to extract the human-annotated attention vector $A_i(Q_i, V_i)$ from γ_i. Especially, the human-annotated attention score for each detection box in the detection-based attention module is the average score of the human-annotated regions covered by the box. Since both the supervised model-generated attention and the supervised human-annotated attention are represented as vectors, we can transfer the human-annotated attention vector to the model-generated attention vector using a supervised learning strategy. In this way, the prior knowledge of human attention can be encoded in the two types of attention modules to learn more effective attention distribution. The mean squared error (MSE) is used as the objective function to define the supervised attention loss $\mathcal{L}_{\text{att}}$

$$\mathcal{L}_{\text{att}}(Q_i, V_i; \theta) = \big\|A_i(Q_i, V_i) - A_i'(Q_i, V_i)\big\|^2 \tag{13}$$

where θ represents the parameter set, that is, either θ_{fre} or θ_{det}. $A_i'(Q_i, V_i)$ is the model-generated attention vector, which is either α_i or β_i. The two attention modules are trained by minimizing the attention loss $\mathcal{L}_{\text{att}}$ to exploit the human-annotated attention knowledge.

C. Adversarial Attention Learning

The pretrained free form-based and detection-based attention modules are used to generate two multiview feature vectors F_i^{fre} and F_i^{det}. Due to the drawbacks faced by the two types of attention modules in answering questions about the foreground object and the background form, respectively, it is necessary to capture the complementary nature of them for more effective answering. In this section, we introduce adversarial learning to enhance the interaction between the two attention modules to learn more effective multiview correlations. Since F_i^{fre} and F_i^{det} are outputs of the two attention modules, they reflect the correlations between image and question from different views, that is, free-form regions and detection boxes, respectively. Therefore, adversarial learning built between F_i^{fre} and F_i^{det} can make the two attention modules produce more effective multiview features from each other. In this way, the two attention modules are mutually reinforced to learn more effective multiview correlations.

First, MLPs activated by softmax in the last layer are used to construct a classifier $d(\cdot, \theta_{\text{dis}})$ as the discriminator of GAN [40]. The discriminator tries to discriminate F_i^{fre} and F_i^{det} generated by the free form-based attention module and the detection-based attention module, respectively. In the training process, F_i^{fre} and F_i^{det} are assigned with the label of $\overline{01}$ and input to the discriminator. Based on (6) and (12), the loss function of the discriminator is defined as

$$\begin{aligned}
&\mathcal{L}_{\text{dis}}(Q_i, V_i; \theta_{\text{dis}}) \\
&= -\log\Big(d\Big(F_i^{\text{fre}}; \theta_{\text{dis}}\Big)\Big) - \log\Big(1 - d\Big(F_i^{\text{det}}; \theta_{\text{dis}}\Big)\Big) \\
&= -\log\big(d\big(g_{fre}\big(Q_i, V_i, ; \theta_{fre}\big); \theta_{\text{dis}}\big)\big) \\
&\quad - \log(1 - d(g_{\text{det}}(Q_i, V_i, ; \theta_{\text{det}}); \theta_{\text{dis}})).
\end{aligned} \tag{14}$$

Notably, during the training of the discriminator, θ_{fre} and θ_{det} are kept unchanged.

Meanwhile, the two attention modules, free form-based attention and detection-based attention, act as two feature generators. Both of them try to confuse the discriminator that F_i^{fre} and F_i^{det} are learned from the other module, respectively. Therefore, the inputs F_i^{fre} and F_i^{det} are assigned with the label of $\overline{10}$. The loss function for the generators is formulated as

$$\begin{aligned}
&\mathcal{L}_{\text{gen}}\big(Q_i, V_i; \theta_{\text{pre}}, \theta_{\text{det}}\big) \\
&= -\log\Big(1 - d\Big(F_i^{\text{fre}}; \theta_{\text{dis}}\Big)\Big) - \log\Big(d\Big(F_i^{\text{det}}; \theta_{\text{dis}}\Big)\Big) \\
&= -\log\big(1 - d\big(g_{fre}\big(Q_i, V_i, ; \theta_{fre}\big); \theta_{\text{dis}}\big)\big) \\
&\quad - \log(d(g_{\text{det}}(Q_i, V_i, ; \theta_{\text{det}}); \theta_{\text{dis}}))
\end{aligned} \tag{15}$$

where θ_{dis} is kept unchanged and the parameters of the attention modules θ_{pre} and θ_{det} are fine tuned during the training of the generators.

D. Optimization for Answer Prediction

Multiview features F_i^{fre} learned from the free form-based attention module and F_i^{det} learned from the detection-based attention module are combined using the strategy of element-wise multiplication to infer the answer. Concretely, we build an answer classifier $a(\cdot, \theta_{\text{ans}})$ by using MLPs activated by softmax in the last layer to predict the answer. The cross-entropy loss is used to define the answer classification loss

$$\begin{aligned}\mathcal{L}_{\text{ans}}(Q_i, V_i, \theta_{\text{ans}}) &= -Y_i \cdot \log\Big(a\Big(F_i^{\text{fre}}, F_i^{\text{det}}; \theta_{\text{ans}}\Big)\Big) \\ &= -Y_i \cdot \log\big(a\big(g_{fre}\big(Q_i, V_i, ; \theta_{fre}\big) \\ &\qquad g_{\text{det}}(Q_i, V_i, ; \theta_{\text{det}}); \theta_{\text{ans}}\big)\big)\end{aligned} \tag{16}$$

where θ_{ans} is the set of parameters in the answer classifier, and Y_i is the ground-truth answer label for the ith question–image instance, which is a one-hot vector.

To optimize the attention networks and the adversarial learning for answer prediction, the loss functions of (15) and (16) are simultaneously minimized by rewriting the loss function of the generator as follows:

$$\begin{aligned}&\mathcal{L}_g\big(\mathcal{Q}, \mathcal{V}; \theta_{fre}, \theta_{\text{det}}, \theta_{\text{ans}}\big) \\ &= \sum_{i=1}^{n}\big(\mathcal{L}_{\text{ans}}(Q_i, V_i, \theta_{\text{ans}}) + \lambda \mathcal{L}_{\text{gen}}\big(Q_i, V_i; \theta_{\text{pre}}, \theta_{\text{det}}\big)\big)\end{aligned} \tag{17}$$

where λ is the balance hyperparameter to regulate the importance of the adversarial attention learning. The loss function of (14) is rewritten for the discriminator as follows:

$$\mathcal{L}_d(\mathcal{Q}, \mathcal{V}; \theta_{\text{dis}}) = \sum_{i=1}^{n} \lambda \mathcal{L}_{\text{dis}}(Q_i, V_i; \theta_{\text{dis}}). \tag{18}$$

The proposed model ALSA can be optimized by alternatively training the generator and the discriminator using backpropagation. In the learning process of the generator, parameters θ_{fre}, θ_{det}, and θ_{ans} are learned by minimizing the loss function of (17). For the discriminator, the parameter θ_{dis} is updated according to (18). Algorithm 1 shows the details of the training procedure of ALSA.

Algorithm 1 Pseudocode of Optimizing ALSA With Step Size μ on Mini-Batched Dataset. g-Steps and d-Steps Are Hyperparameters

Require: Min-batch question $\mathcal{Q}$ and the corresponding images $\mathcal{V}$.

Training Procedure:

repeat
  for g-steps **do**
    update parameters θ_{fre}, θ_{det}, and θ_{ans} for generator learning:
    $\theta_{fre} \longleftarrow \theta_{fre} - \mu \cdot \nabla_{\theta_{fre}} \mathcal{L}_g(\mathcal{Q}, \mathcal{V}; \theta_{fre}, \theta_{det}, \theta_{ans})$
    $\theta_{det} \longleftarrow \theta_{det} - \mu \cdot \nabla_{\theta_{det}} \mathcal{L}_g(\mathcal{Q}, \mathcal{V}; \theta_{fre}, \theta_{det}, \theta_{ans})$
    $\theta_{ans} \longleftarrow \theta_{ans} - \mu \cdot \nabla_{\theta_{ans}} \mathcal{L}_g(\mathcal{Q}, \mathcal{V}; \theta_{fre}, \theta_{det}, \theta_{ans})$
  end for
  for d-steps **do**
    update parameters θ_{dis} for discriminator learning:
    $\theta_{dis} \longleftarrow \theta_{dis} - \mu \cdot \nabla_{\theta_{dis}} \mathcal{L}_d(\mathcal{Q}, \mathcal{V}; \theta_{dis})$
  end for
until ALSA converges

IV. Experiments

A. Datasets and Baselines

Extensive experiments are conducted on the following three datasets to evaluate the performance of ALSA.

VQA v1.0 [1] has become the most widely used VQA dataset. It is created from the MS-COCO dataset and labeled by crowd-sourced workers. The questions are categorized into three types: 1) *yes/no*; 2) *number*; and 3) *other*. Ten free-response answers are labeled for each question and two subtasks contained in this dataset, that is, open-ended (OE) and multiple-choice (MC). This dataset is divided into three splits: 1) train; 2) val; and 3) test. Besides, the test set includes test-dev and test-std. Totally, it contains 328 120 questions, 204 721 images, and 22 523 answers. The top 1000 frequent answers are used as the possible outputs, which cover 82.7% of the total answers.

VQA v2.0 [58] attempts to minimize the effectiveness of learning dataset priors by balancing the answers to each question. This dataset is constructed based on the VQA v1.0 dataset, which collects additional images such that each question in the dataset has a pair of similar images but corresponds to different answers. It contains a total of 332 793 questions, 204 721 images, and 29 332 answers.

COCO-QA [59] is another widely used VQA dataset, which is also created from MS-COCO dataset. The questions are divided into four categories: 1) *Object*(70%); 2) *Number*(7%); 3) *Color*(17%); and 4) *Location*(6%). This dataset consists of a training set and a test set. There are a total of 92 396 questions, 69 172 images, and 435 answers. Furthermore, each image belongs to only one question and all answers are single word.

The following models are employed as the main baselines to compare with ALSA for evaluation.

1) *HieCoAtt [15]:* A co-attention model jointly explores the visual and textual attention distribution in a hierarchical scheme through a 1-D CNN.
2) *MLB [34]:* A low-rank bilinear pooling method uses Hadamard product for an effective attention mechanism of multimodal learning.
3) *Dual-MFA [11]:* A deep neural network linearly combining two branches of visual attention models aims to select the free-form image regions and detection boxes most related to the input question.
4) *VKMN [37]:* A visual knowledge memory network that seamlessly incorporates the structured human knowledge and deep visual features into a memory network in an end-to-end learning framework.
5) *ODA [29]:* A simple but effective object-difference attention model, which uses difference operator to

TABLE I
EXPERIMENTAL RESULTS OF ALSA AND THE COMPARED STATE-OF-THE-ART METHODS ON THE VQA V1.0 DATASET

Method	Test-dev					Test-std				
	Open-Ended				MC	Open-Ended				MC
	All	Yes/No	Number	Other	All	All	Yes/No	Number	Other	All
VSE [1]										
-LSTM+Q	48.75	78.15	35.64	26.68	-	-	-	-	-	-
-LSTM+Q+I	53.74	78.94	35.24	36.42	57.17	53.96	79.01	35.55	36.80	57.57
DPPnet [2]	57.22	80.71	37.24	41.69	62.48	57.36	80.28	36.92	42.24	62.69
SAN [14]	58.70	79.30	36.60	46.10	-	58.9	79.11	36.41	46.42	-
HieCoAtt [15]	61.80	79.70	38.70	51.70	65.8	62.10	79.95	38.22	51.95	66.07
MCB [33]	64.70	82.50	37.60	55.60	69.10	-	-	-	-	-
MLB [34]	64.89	84.13	37.85	54.57	-	65.07	84.02	37.90	54.77	68.89
CVA [20]	65.92	83.73	40.91	56.36	70.30	66.20	83.79	40.41	56.77	70.41
VKMN [37]	66.00	83.70	37.90	57.00	69.10	66.10	84.10	38.10	56.90	69.10
Dual-MFA [11]	66.01	83.59	40.18	56.34	70.04	66.09	83.37	40.39	56.89	69.97
DCN [58]	66.89	84.61	42.35	57.31	-	67.02	85.04	42.34	56.98	-
ODA [29]	67.83	85.82	**43.03**	58.07	72.28	67.97	85.81	42.51	**58.24**	72.32
ALSA (ours)	**69.52**	**87.12**	42.94	**59.06**	**73.61**	**69.32**	**86.94**	**43.84**	58.21	**73.67**

explicitly compare objects and calculate the attention distribution.

6) *CRA-Net [31]:* A composed relation attention network, which encodes the fine-grained and precise binary relations and the more sophisticated trinary relations between objects using two relation attention modules.

B. Experimental Settings and Evaluation Methods

We use ResNet-152 [54] pretrained on ImageNet to extract the free-form image regions. The output feature map of the last convolution layer is used to denote the whole image's visual features $R_i \in \mathbb{R}^{2048\times14\times14}$. The Faster-RCNN [56] is adopted to extract object detection boxes. For comparing with Dual-MFA [11], the top-ranked 19 object proposals are chosen as the detection boxes $D_i \in \mathbb{R}^{4096\times19}$ like Dual-MFA, where each object is represented by a 4096-D feature vector. For textual words, the 300-D pretrained GloVe [61] feature vector is used to embed each word of the question. The GRU is set to have 512 neural cells. The dimensionality of the common space c and s is set to be 256 and 512, respectively. The MLPs built for the discriminator and the answer classifier are set with the network structures of 1024-512-2 and 2048-1024-N, respectively, where N is the number of the answer classes. All of the activation functions are fixed to relu$(\cdot)$ except special instructions. g-steps and d-step in Algorithm 1 are empirically set to be 5 and 1, respectively, for relatively fast and stable convergence. As for the optimization algorithm, the adaptive moment estimation is employed with an initial learning rate of 3×10^{-3}, a momentum 0.98, and a weight decay 10^{-8}. We fix the batch size to 256. The dropout technique with probability 0.6 is adopted after each MLP to prevent over-fitting.

For VQA v1.0 and VQA v2.0 datasets, we use the metric defined in [1] to evaluate the performance, that is, *accuracy* $=$ *min* (*#humans that provided that answer*/3, 1). That is, if no fewer than three workers provide an exact answer, the answer is considered to be 100% accurate. As for the COCO-QA dataset, we evaluate the models using classification accuracy. Furthermore, the Wu-Palmer similarity (WUPS) proposed in [62] is also used as another metric according to [59]. Specifically, WUPS@0.0 and WUPS@0.9 are employed as metrics.

C. Results and Analysis

In the VQA v1.0 dataset, the models are trained on the train + val sets and tested on the test-dev and test-std sets. In the VQA v2.0 dataset, the models are trained on the train + val sets and tested on the test-dev set. For the COCO-QA dataset, the models are trained and tested on the training and test sets, respectively.

The results on VQA v1.0 are shown in Table I. On the test-dev set, it can be observed that ALSA improves the overall accuracy of the best baseline ODA from 67.83% to 69.52% in the OE task, and improves ODA by 1.33% in the multi-choice (MC) task. Furthermore, ALSA outperforms HieCoAtt, Dual-MFA, CVA, and DCN on accuracy by 7.72%, 3.41%, 3.60%, and 2.63%, respectively. As for the test-std set, ALSA improves the overall accuracy of the best baseline ODA by 1.35% in the OE task. A similar improvement can also be seen in the other three question categories.

Table II shows the experimental results on the COCO-QA dataset. It can be observed that ALSA outperforms all the baselines on the metric of accuracy. Concretely, ALSA improves the accuracy of DDPnet, SAN, HieCoAtt, DUAL-MFA, and CVA by 8.78%, 8.37%, 4.57%, 3.48%, and 2.46%, respectively. In addition, ALSA improves the accuracy of the best baseline ODA from 69.33% to 69.97%. The results on the metrics of WUPS@0.9 and WUPS@0.0 also display the improvements of our model compared with the baselines. As for the experimental results on each question category, ALSA also achieves relatively higher accuracy.

The comparison results on the VQA v2.0 dataset are shown in Table III. ALSA outperforms the representative methods of VKM (Ensemble), DCN (17), and ODA $\times$ 2 (100boxes) on accuracy by 2.54%, 2.17%, and 1.04%, respectively. Moreover, ALSA improves the overall accuracy of CRA-Net from 68.61% to 69.21%. The superiority of ALSA compared

TABLE II
EXPERIMENTAL RESULTS ON THE COCO-QA DATASET. "–" INDICATES THE DATA IS UNAVAILABLE

Methods	Accuracy	WUPS@0.9	WUPS@0.0	Object	Number	Color	Location
VSE [1]							
-VIS+LSTM	53.31	63.91	88.25	56.53	46.10	45.87	45.52
-2VIS+BLSTM	55.09	65.34	88.64	58.17	44.79	49.53	47.34
Img-CNN [61]	58.40	68.50	89.67	-	-	-	-
DPPnet [2]	61.19	70.84	90.61	-	-	-	-
SAN [14]	61.60	71.60	90.90	64.50	48.60	57.90	54.00
HieCoAtt [15]	65.40	75.10	92.00	68.00	51.00	62.90	58.80
Dual-MFA [11]	66.49	76.15	92.29	68.86	51.32	65.89	58.92
CVA [20]	67.51	76.70	92.41	69.55	50.76	68.96	59.93
ODA [29]	69.33	78.29	93.02	70.84	54.70	**74.17**	60.90
ALSA (ours)	**69.97**	**79.43**	**94.15**	**71.59**	**54.83**	72.74	**61.78**

TABLE III
EXPERIMENTAL RESULTS ON THE VQA V2.0 DATASET

Methods	VQA v2.0 Test-dev			
	All	Yes/No	Number	Other
VKMN (single) [37]	64.36	83.70	37.90	57.79
VKMN (Ensemble) [37]	66.67	82.88	43.17	57.95
DCN (16) [58]	66.97	83.59	46.98	57.09
DCN (17) [58]	67.04	83.85	47.19	56.95
DCN (18) [58]	67.00	83.89	46.93	56.90
ODA (36boxes) [29]	67.34	84.23	46.18	57.73
ODA x 2 (36boxes) [29]	67.52	84.30	46.62	57.96
ODA x 2 (100boxes) [29]	68.17	84.66	48.04	58.68
CRA-Net [31]	68.61	84.87	49.46	59.08
ALSA (ours)	69.21	85.73	48.98	59.17
BAN [32]	70.04	85.42	54.04	60.52
MCAN [30]	70.63	86.82	53.26	60.72

TABLE IV
ABLATION STUDY ON BOTH VQA V1.0 AND COCO-QA DATASETS

Methods	Accuracy	
	VQA v1.0	COCO-QA
F-Att	60.84	60.76
D-Att	61.25	61.46
FD-Att	63.21	63.52
VQS: F-Att + Sup	61.42	61.87
VQS: D-Att + Sup	62.05	62.91
VQS: FD-Att + Sup	65.71	66.35
VQA-HAT: F-Att + Sup	61.93	62.03
VQA-HAT: D-Att + Sup	62.63	63.12
VQA-HAT: FD-Att + Sup	66.52	67.14
FD-Att + AL	68.87	69.76
FD-Att + Sup + AL	**69.73**	**70.08**

with these baselines can also be seen in the three question categories. In addition, two state-of-the-art methods BAN and MCAN that outperform ALSA on this dataset are also displayed in Table III. The reason for the superiority of the two models is that both BAN and MCAN focus on using deep attention networks to learn the correlations between the question and image, while ALSA utilizes shallow attention modules to associate the question to meaningful image regions. However, the adversarial learning and supervised attention networks designed in ALSA can be built on any type of attention models, including BAN and MCAN, to reinforce them mutually from different views. Namely, ALSA can be applied to other excellent attention models, including BAN and MCAN to obtain better performance.

There are several reasons for the improvement. On the one side, existing attention approaches, such as SAN, HieCoAtt, CVA, and ODA implicitly explore the correlations between image and question without supervision, which might focus on some unrelated image regions and then affects the performance of question answering. The supervised attention modules which exploit the prior knowledge of human attention are more effective to attend on the meaningful image regions related to the question. On the other side, compared with Dual-MFA that linearly combines the multiview features learned from different attention modules, ALSA employs adversarial learning networks to reinforce the two types of attention modules mutually from different views. It can learn more effective multiview features from the attention modules for answer inference. The improvements indicate that the proposed supervised attention modules and adversarial learning used to reinforce different attention modules mutually are effective to improve the performance of VQA.

D. Ablation Study

We further conduct ablation experiments to analyze the effectiveness of individual components designed in ALSA, and the results are shown in Table IV. Note that the top-ranked 36 proposals are chosen as the object boxes in the detection-based attention model in this study. For the VQA v1.0 dataset, the models are trained on the training set and tested on the validation set. Due to the limitation of online submission, we do not use the test set. For the COCO-QA dataset, the models are trained on the training set and tested on the test set.

The first part of Table IV displays the experimental results of the free form-based attention (F-Att) and the detection-based attention (D-Att) models. It can be observed that D-Att outperforms F-Att by 0.41% and 0.70% on the VQA v1.0 dataset and COCO-QA dataset, respectively. FD-Att which combines F-Att and D-Att achieves better performance than both F-Att and D-Att on the two datasets. The second and third parts of Table IV show the effectiveness of the supervised attention models built on two attention-annotated datasets

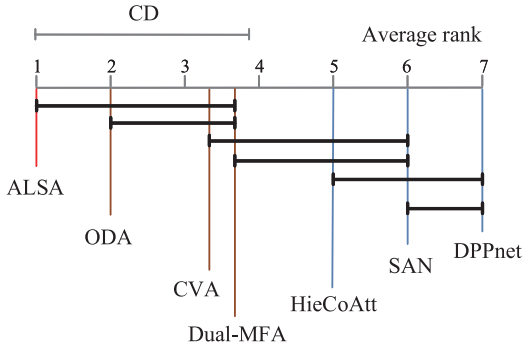

Fig. 4. Experimental results of the Friedman–Nemenyi test.

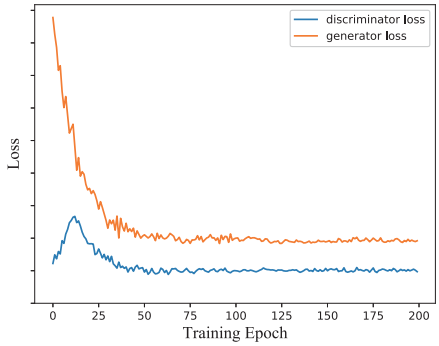

Fig. 5. Convergence of the adversarial learning between generator and discriminator.

VQS [63] and VQA-HAT, respectively. One can see that the supervised F-Att (F-Att + Sup) model, the supervised D-Att (D-Att + Sup) model, and the supervised FD-Att (FD-Att + Sup) model using either VQS or VQA-HAT datasets have better performance than F-Att, D-Att, and FD-Att, respectively. Additionally, the two parts of Table IV also show that the supervised attention models trained with the VQA-HAT dataset outperform the supervised attention models trained with the VQS dataset. Therefore, the VQA-HAT dataset is selected as the prior knowledge of human attention in our method. The last part of Table IV displays the superiority of using adversarial learning in our model. It can be observed that the model of adversarial learning built on FD-Att (FD-Att + AL) has better performance than F-Att, D-Att, and FD-Att. Moreover, adversarial learning built on FD-Att + Sup (FD-Att + Sup + AL) achieves the highest accuracy in all the models. It improves FD-Att + AL by 0.86% and 0.32% on VQA v1.0 and COCO-QA datasets, respectively. This table demonstrates the effectiveness of the supervised attention models and adversarial learning networks for VQA.

E. Statistical Significance Analysis

We employ Friedman–Nemenyi test [64] with 95% confidence level to analyze the difference between ALSA and the baselines. Due to the limitations of the baselines, the experiments are conducted on the VQA v1.0 Test-dev, VQA v1.0 Test-std, and COCO-QA datasets. The models are ranked according to the metric of overall accuracy on each dataset. If the null hypothesis (all models are equivalent) is rejected, Nemenyi's test can be conducted to compare the critical distance (CD) with the average rank difference between any two models. If the difference is greater than the corresponding CD, we can conclude that there is a significant difference between the performance of the two models.

Fig. 4 shows the results of Friedman–Nemenyi test. It can be observed that ALSA has significant improvement compared to many baselines, including HieCoAtt, SAN, and DPPnet. The difference between ALSA and the three baselines (ODA, CVA, and Dual-MFA) is less significant. The reason is that ODA, CVA, and Dual-MFA can exploit the latent correlations between visual regions and textual words in a sensible way for multimodal joint embedding learning. However, despite ODA, CVA, and Dual-MFA achieve the second, third, and fourth positions in average rank, respectively, they still neglect the prior knowledge of human attention and learn the attention distribution on either free-form regions or detected boxes independently. On the contrary, ALSA can use prior knowledge and adversarial networks to learn more effective attention distribution from different views, thereby achieving better performance.

F. Effect of Adversarial Learning

To investigate the effectiveness of adversarial learning in ALSA, we record the loss values of the generator and discriminator on the VQA v1.0 dataset from 1 to 200 epochs. As can be seen in the results shown in Fig. 5, the loss of the generator decreases rapidly with a certain vibration in the first 50 epochs. On the other side, the loss of the discriminator increases during the first 15 epochs and then also decreases with a certain vibration. The vibration indicates that there exists fierce competition between the generator and the discriminator in the first 50 epochs. Then, the losses of the generator and the discriminator gradually decrease, and the vibration becomes smaller. After about 100 epochs, the competition between the two players reaches an agreement, with only a small amount of vibration. It means that the model has converged to learn effective multimodal features. Conversely, if the generator or discriminator dominates the optimization process, the model will fail to generate a meaningful multiview representation for answer prediction.

G. Analysis of Supervised Attention Loss

In ALSA, the supervised attention loss in (13) is optimized to encode the prior knowledge for attention distribution learning. In this experiment, we consider three potential loss functions, that is, MSE, mean absolute error (MAE), and Huber loss. Fig. 6(a) and (b) shows the comparison of these loss functions used in the supervised free-form-based and the supervised detection-based attention models, respectively. Note that the hyperparameter δ in the Huber loss is set to 0.5 in our experiment. It can be observed that MSE converges faster and yields lower loss values in both of the two attention models. MAE has the slowest training speed and the worst convergence values. Huber loss is a compromise between MSE and MAE. The reason might be that the quality of the human-annotated attention maps in the VQA-HAT dataset is at a high

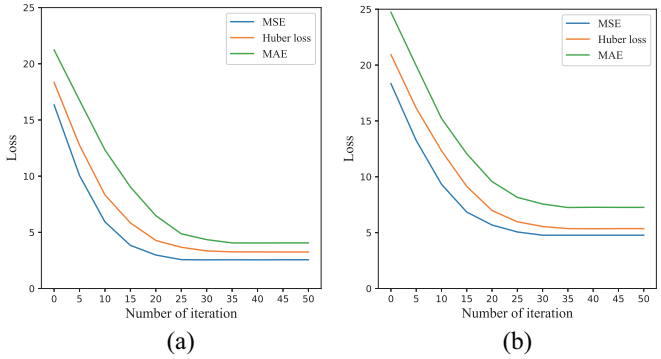

Fig. 6. Comparision of three potential loss functions used in the (a) supervised free form-based and (b) supervised detection-based attention models.

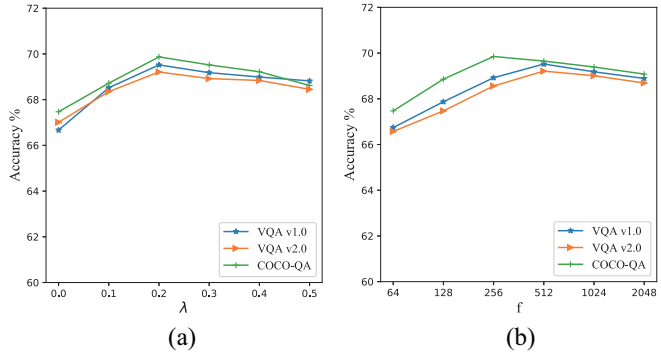

Fig. 7. Parameter sensitivity study for (a) balance parameters and (b) vector dimension of the multiview features.

level with fewer abnormal samples. MSE can continuously adjust the optimization rate to reduce the loss and bring it closer to the optimal value. Therefore, MSE is chosen as the objective function in the supervised attention models.

H. Parameter Sensitivity

To assess the impact of parameterization on the performance of ALSA, we present the experiment results with different values of the balance parameter λ and the dimension size f of the multiview feature vector.

Balance Parameter: In (17) and (18), λ is used to balance the importance of adversarial learning. We first fix the dimension of the multiview feature vector $f = 512$ on the VQA v1.0 and VQA v2.0 datasets, and fix $f = 256$ on the COCO-QA dataset, respectively. Then, the performance of ALSA corresponding to different values of λ is reported. Based on the curves in Fig. 7(a), one can see that adversarial learning contributes to the performance since the model achieves relatively low accuracy when $\lambda = 0$. However, a too large value of λ will result in overfitting. From the curve, it can be observed that the model obtains the best performance when $\lambda = 0.2$ on all of the three datasets.

Vector Dimension: In the proposed model, the multiview features learned from the free form-based attention and the detection-based attention are mapped to f-dimensional vectors F_i^{fre} and F_i^{det}, respectively. Fig. 7(b) shows how the dimensionality of the vector affects the performance by fixing $\lambda = 0.2$. It can be seen that the accuracy rises initially and then starts to drop slowly as the size of the dimension increases. The reason may be that a higher dimension of the vector can embed more useful features for answer prediction. However, a too large size of the dimension will bring in noise, which may decline the model performance. From the curve, one can see that ALSA reaches the highest accuracy when $f = 512$ on both VQA v1.0 and VQA v2.0 datasets, and $f = 256$ on the COCO-QA dataset.

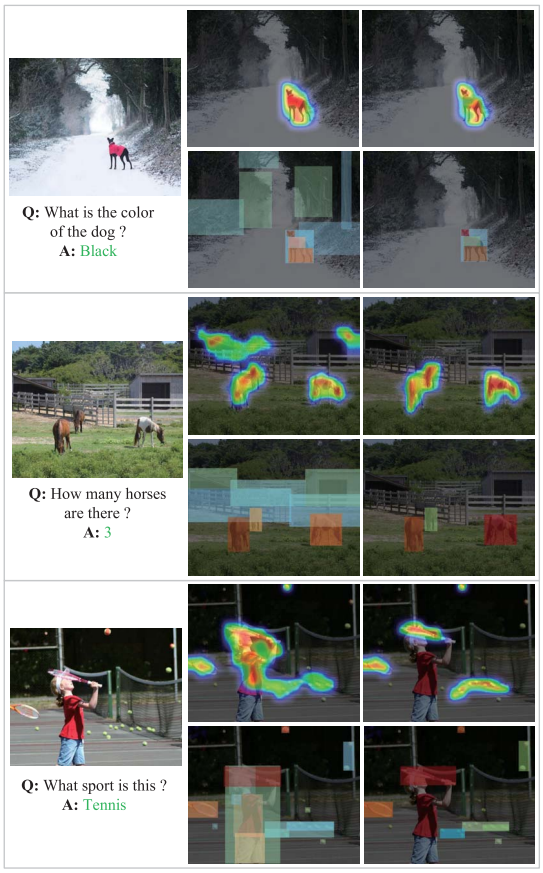

Fig. 8. Three visualization examples of the attention models. For each instance, the first column displays the original image and the question-answer pair. The second column shows two attention maps of the free-form-based attention and the detection-based attention, respectively. The last column illustrates the attention maps of the corresponding supervised attention models.

I. Visualization of the Attention Models

To analyze the effect of the supervised attention mechanisms, we visualize the attention maps of the supervised free form-based and the supervised detection-based attention models. Meanwhile, the attention maps of the unsupervised attention models are also visualized for comparison. Three examples are illustrated in Fig. 8. From the results, one can see that the supervised attention models are more effective to attend on the image regions related to the answer. For example, the first instance shows that the supervised free form-based attention model pays more attention on the regions related to the head and legs of the dog rather than the red cloth, which contributes to infer the answer of the question. This instance also shows that the supervised detection-based attention model focuses on more related boxes compared to the

Fig. 9. Four examples used for qualitative analysis of adversarial learning. For each instance, the first column shows the original image, the question-answer pair, and the answers predicted by four models. Note that the answers with green and red color represent correct and wrong answers, respectively. The second column displays two attention maps of the free form-based attention (F-Att) and the detection-based attention (D-Att), respectively. The third column shows the supervised attention maps of the combined F-Att and D-Att, that is, FD-Att + Sup. The last column illustrates the corresponding attention maps given by the model which conducts adversarial learning on the supervised attention, that is, FD-Att + Sup + AL.

detection-based attention model. One can also see the superiority of the supervised attention models in the last instance. Both of the supervised free form-based and the supervised detection-based attention models concentrate on the regions and boxes more related to the rackets and tennis balls compared with the attention models without supervision. These results indicate that the supervised attention models are more effective to capture the answer-related image content.

J. Qualitative Analysis of Adversarial Learning

We use several examples to analyze how the attention models are affected by adversarial learning. Since adversarial learning is performed on the supervised attention modules, we also show the attentional distribution map of the supervised attention. Fig. 9 presents four visual instances of four models, including the free form-based attention model (F-Att), the detection-based attention model (D-Att), the supervised attention on the combined F-Att and D-Att (FD-Att + Sup) model, and the model which conducts adversarial learning on the supervised attention (FD-Att + Sup + AL). Both F-Att and D-Att are pretrained on the VQA-HAT dataset. In the first instance, it can be observed that F-Att cannot separately recognize the regions related to the two white dogs and thus a wrong answer is obtained. D-Att pays more attention on the detection boxes related to the white dogs and then a correct answer is obtained. Although supervised attention uses prior knowledge to accurately focus on the detection boxes corresponding to the two white dogs, it still fails to distinguish the two white dogs on the free-form attention map. Therefore, FD-Att + Sup fails to infer the correct answer. Due to the effect of adversarial learning, the regions related to the white dogs are correctly covered by the free-form attention, and the detection boxes related to the white dogs are paid more attention. Therefore, FD-Att + Sup + AL obtains the correct answer. Like the first instance, D-Att and FD-Att + Sup +AL also infer the correct answer while F-Att infers the wrong answer in the third instance. FD-Att + Sup infers the correct answer due to the supervised attention maps cover the correct image regions and boxes. The second and the last instances show different results, where F-Att succeeds and D-Att fails to infer the correct answer. We can see that FD-Att + Sup + AL also infers the correct answer in the two instances. Take the last instance as an example, exact detection boxes about the sky do not exist, so D-Att cannot capture the weather conditions well, leading to the wrong answer. F-Att focuses on some regions in the sky and then answers the question correctly. Even with the help of supervised attention, FD-Att + Sup cannot accurately infer the correct answer. However, although only the free-form attention map focuses on the correct image regions related to the answer, FD-Att + Sup + AL conducting adversarial learning on the supervised attention model predicts the answer correctly. The reason is that adversarial learning is effective in reinforcing the free form-based and detection-based attention models mutually to improve the performance of VQA.

K. Model Generalization

In ALSA, the adversarial learning method used to fuse attention models from different views can be extended to more VQA models and other scenarios. Taking the two multimodal models, that is, Dual-MFA [11] in the VQA task and BDMLA [65] in the sentiment classification task, as the examples, both of they employ different attention mechanisms to correlate the textual language to the meaningful visual regions. For comparison, we use adversarial learning to fuse attention models from different views. Our methods are termed as Dual-MFA(AL) and BDMLA(AL), respectively. For Dual-MFA and Dual-MFA(AL), the models are trained on

TABLE V
MODEL GENERALIZATION OF ADVERSARIAL LEARNING USED IN FUSING ATTENTION MODELS FROM DIFFERENT VIEWS

Models	Accuracy
Dual-MFA	66.01
Dual-MFA(AL)	68.32
BDMLA	84.90
BDMLA(AL)	86.52

the VQA v1.0 train + val set and tested on the test-dev set. For BDMLA and BDMLA(AL), the Flickr dataset reported in [65] is used for model evaluation. The experiment results are shown in Table V. It can be seen that Dual-MFA(AL) improves Dual-MFA on accuracy by 2.31%. BDMLA(AL) outperforms BDMLA by 1.62%. These results indicate that the proposed fusion scheme of using adversarial learning to fuse different attention models is effective and can be flexibly applied to other models for better performance.

V. CONCLUSION

In this article, we investigated to explore the prior knowledge of human-annotated attention maps and learn the attention from different views for the VQA task. To effectively answer both the questions about the foreground object and background form, a novel model termed ALSAs is proposed. Specifically, two supervised attention modules, free form-based and detection-based, are designed and pretrained on the human-annotated VQA dataset to learn the correlations between the question and image from different views. In addition, adversarial learning networks are conducted to reinforce the two types of attention modules mutually to learn more effective multiview features. The extensive experiments performed on three public VQA datasets confirm the effectiveness and superiority of ALSA against the baselines. It also certifies that the proposed method can be extended to more VQA models and other scenarios for better performance.

This work has extensive research expansions in the future. It will be exciting to investigate free-form answers generation. Besides, it would be interesting to explore the social information, such as the microblogs published by the owner and the context information of the image, for VQA.

REFERENCES

[1] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 2425–2433.

[2] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 30–38.

[3] Y. Liu, X. Zhang, F. Huang, L. Cheng, and Z. Li, "Adversarial learning with multi-modal attention for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 24, 2020, doi: 10.1109/TNNLS.2020.3016083.

[4] L. Xie, J. Shen, and L. Zhu, "Online cross-modal hashing for Web image retrieval," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 294–300.

[5] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 5187–5196.

[6] A. Singh *et al.*, "Pythia-a platform for vision & language research," in *Proc. SysML Workshop NeurIPS*, 2018, p. 1.

[7] Y. Yang *et al.*, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.

[8] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, Jan. 2019.

[9] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.

[10] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, "R-VQA: Learning visual relation facts with semantic attention for visual question answering," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 1880–1889.

[11] P. Lu, H. Li, W. Zhang, J. Wang, and X. Wang, "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7218–7225.

[12] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 21–29.

[13] Y. Liu, X. Zhang, F. Huang, X. Tang, and Z. Li, "Visual question answering via attention-based syntactic structure tree-LSTM," *Appl. Soft Comput.*, vol. 82, 2019, Art. no. 105584.

[14] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 21–29.

[15] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.

[16] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2156–2164.

[17] R. Li and J. Jia, "Visual question answering with question representation update (QRU)," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4655–4663.

[18] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6077–6086.

[19] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4613–4621.

[20] J. Song, P. Zeng, L. Gao, and H. T. Shen, "From pixels to objects: Cubic visual attention for visual question answering," in *Proc. 37th Int. Joint Conf. Artif. Intell.*, 2018, pp. 906–912.

[21] C. Du, C. Du, X. Xie, C. Zhang, and H. Wang, "Multi-view adversarially learned inference for cross-domain joint distribution matching," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 1348–1357.

[22] V. Dumoulin *et al.*, "Adversarially learned inference," 2016. [Online]. Available: arXiv:1606.00704.

[23] A. Jiang, F. Wang, F. Porikli, and Y. Li, "Compositional memory for visual question answering," 2015. [Online]. Available: arXiv:1511.05676.

[24] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.

[25] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 30, 2019, doi: 10.1109/TPAMI.2019.2932058.

[26] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, early access, May 17, 2019, doi: 10.1109/TCYB.2019.2914351.

[27] Y. Liu, X. Zhang, F. Huang, and Z. Li, "Adversarial learning of answer-related representation for visual question answering," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, 2018, pp. 1013–1022.

[28] A. Osman and W. Samek, "DRAU: Dual recurrent attention units for visual question answering," *Comput. Vis. Image Understand.*, vol. 185, pp. 24–30, Aug. 2019.

[29] C. Wu, J. Liu, X. Wang, and X. Dong, "Object-difference attention: A simple relational attention for visual question answering," in *Proc. ACM Multimedia Conf.*, 2018, pp. 519–527.

[30] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 6281–6290.

[31] L. Peng, Y. Yang, Z. Wang, X. Wu, and Z. Huang, "CRA-Net: Composed relation attention network for visual question answering," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1202–1210.

[32] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.

[33] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.

[34] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016. [Online]. Available: arXiv:1610.04325.

[35] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2631–2639.

[36] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[37] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li, "Learning visual knowledge memory networks for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7736–7745.

[38] Q. Wu, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4622–4630.

[39] Y. Zhu, J. J. Lim, and L. FeiFei, "Knowledge acquisition for visual question answering via iterative querying," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6146–6155.

[40] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[41] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015. [Online]. Available: arXiv:1511.06434.

[42] H. Wang *et al.*, "GraphGAN: Graph representation learning with generative adversarial nets," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 2508–2515.

[43] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[44] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, and Z. Huang, "Modal-adversarial semantic learning network for extendable cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 46–54.

[45] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 4242–4251.

[46] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2157–2169.

[47] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6106–6115.

[48] X. Liu, X. Kong, L. Liu, and K. Chiang, "TreeGAN: Syntax-aware sequence generation with generative adversarial networks," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, Singapore, 2018, pp. 1140–1145.

[49] J. Wang *et al.*, "IRGAN: A minimax game for unifying generative and discriminative information retrieval models," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 515–524.

[50] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2852–2858.

[51] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Trans. Image Process.*, vol. 29, pp. 1204–1218, 2019.

[52] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 15, 2019, doi: 10.1109/TCSVT.2019.2947482.

[53] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[55] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014. [Online]. Available: arXiv:1406.1078.

[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[57] D. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6087–6096.

[58] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6904–6913.

[59] M. Ren, R. Kiros, and R. S. Zemel, "Exploring models and data for image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2953–2961.

[60] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 3567–3573.

[61] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[62] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguist.*, 1994, pp. 133–138.

[63] C. Gan, Y. Li, H. Li, C. Sun, and B. Gong, "VQS: Linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 1811–1820.

[64] M. Friedman, "A comparison of alternative tests of significance for the problem of *m* rankings," *Ann. Math. Stat.*, vol. 11, no. 1, pp. 86–92, 1940.

[65] J. Xu *et al.*, "Visual-textual sentiment classification with bi-directional multi-level attention networks," *Knowl. Based Syst.*, vol. 178, pp. 61–73, Aug. 2019.

Yun Liu received the B.Sc. degree in computer science and technology from Sichuan University, Chengdu, China, in 2015, and the M.Sc. degree in computer science and technology from Beihang University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering.

His research interests include social media analysis, multimodal data analysis, and data mining.

Xiaoming Zhang received the B.Sc. and M.Sc. degrees in computer science and technology from the National University of Defence Technology, Changsha, China, in 2003 and 2007, respectively, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2012.

He is currently with the School of Cyber Science and Technology, Beihang University, where he has been an Associate Professor since 2012. He has published over 40 papers, such as *ACM Transactions on Information Systems*, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, *World Wide Web Journal*, *Neurocomputing*, *Signal Processing*, ACM MM, AAAI, IJCAI, CIKM, ICMR, SDM, and EMNLP. His current research interests include social media analysis, image tagging, and text mining.

Zhiyun Zhao received the B.Sc. and M.Ac. degrees from Renmin University, Beijing, China, in 1996 and 2003, respectively.

He is currently working with the National Computer Emergency Technical Team/Coordination Center of China, Beijing. He has published more than ten papers, such as AIAAT 2020. His major interests are research on cyberspace strategy and international rules, social media analysis, and text mining.

Bo Zhang received the B.Sc. degree in computer science and technology from Jiangsu University, Zhenjiang, China, in 2017. He is currently pursuing the Ph.D. degree in cyberscience and technology at Beihang University, Beijing, China.

His research interests include social media analysis, multimodal data analysis, and data mining.

Lei Cheng received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2013, and the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2018.

He is currently a Research Scientist with the Shenzhen Research Institute of Big Data, Shenzhen, China. His research interests include tensor data analytics, statistical inference, and large-scale optimization.

Zhoujun Li (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the National University of Defence Technology, Changsha, China, in 1984 and 1999, respectively.

He is currently with the School of Computer Science and Engineering, Beihang University, Beijing, China, where he has been a Professor since 2001. He has published over 150 papers on international journals, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, *ACM Transactions on Information Systems*, *World Wide Web Journal*, and *Information Science*, and international conferences, such as SIGKDD, ACL, SIGIR, AAAI, IJCAI, MM, CIKM, EMNLP, SDM, and WSDM. His current research interests include data mining, information retrieval, and database.

Prof. Li was a PC Member of several international conferences, such as SDM 2015, CIKM 2013, WAIM 2012, and PRICAI 2012.