Adversarial Learning With Multi-Modal Attention for Visual Question Answering

Yun Liu, Xiaoming Zhang, Feiran Huang, *Member, IEEE*, Lei Cheng, and Zhoujun Li, *Member, IEEE*

***Abstract*— Visual question answering (VQA) has been proposed as a challenging task and attracted extensive research attention. It aims to learn a joint representation of the question–image pair for answer inference. Most of the existing methods focus on exploring the multi-modal correlation between the question and image to learn the joint representation. However, the answer-related information is not fully captured by these methods, which results that the learned representation is ineffective to reflect the answer of the question. To tackle this problem, we propose a novel model, i.e., adversarial learning with multi-modal attention (ALMA), for VQA. An adversarial learning-based framework is proposed to learn the joint representation to effectively reflect the answer-related information. Specifically, multi-modal attention with the Siamese similarity learning method is designed to build two embedding generators, i.e., question–image embedding and question–answer embedding. Then, adversarial learning is conducted as an interplay between the two embedding generators and an embedding discriminator. The generators have the purpose of generating two modality-invariant representations for the question–image and question–answer pairs, whereas the embedding discriminator aims to discriminate the two representations. Both the multi-modal attention module and the adversarial networks are integrated into an end-to-end unified framework to infer the answer. Experiments performed on three benchmark data sets confirm the favorable performance of ALMA compared with state-of-the-art approaches.**

***Index Terms*— Adversarial learning, multi-modal attention, visual question answering (VQA).**

I. Introduction

THE recent advancement of natural language processing and computer vision continues to promote the development of artificial intelligence. Meanwhile, the problems of combining image and language understanding steadily inspire considerable research attention [1]. One of the current research hotspots of question answering system has been extended from pure natural language to cross-modal research. A new multi-modal learning task named visual question answering (VQA) [2]–[4] has been considered as a promising but intractable research point. VQA is expected to answer arbitrary natural language questions about the given images. Similar to other multi-modal tasks such as cross-modal retrieval [5], [6] and image captioning [7], [8], VQA requires an in-depth understanding on both the input image and question. VQA can be easily expanded to other tasks and plays a significant role in various applications, including early education, automatic customer service, human–machine interaction, and so on.

Despite VQA has attracted widespread attention from academia and industry, it still faces some challenges to infer the answer. First, there are various manifestations of multi-modal data, such as the textual sentence of the question and the visual content of the image. The features of different modalities are heterogeneous and correlated. Therefore, the VQA method should be effective to bridge the gap between different modalities. Second, VQA aims to learn an effective joint embedding of the question–image pair and input it to a classifier for answer prediction. Therefore, the answer-related information should be effectively reflected in the learned question–image join embedding. However, there is no explicit mechanism to extract the answer-related features from the image directly. The previous methods mainly focus on exploring the correlation between image and question to learn a joint representation based on the information backpropagated from the answer classifier. However, the answer-related information expected to be learned is not effectively captured by the learned joint representation, which results in undesirable performance.

Actually, the heterogeneity and relation among the triplet ⟨image, question, and answer⟩ provide useful clues for answer inference. On the one side, there exist close correlations between question words and image regions. As the examples shown in Fig. 1, some textual words in the question sentence have direct correspondences with some regions in the image. If the alignments between the semantic words and the corresponding regions are captured, the multi-modal contents can

Manuscript received February 7, 2019; revised November 18, 2019, March 13, 2020, and June 23, 2020; accepted August 8, 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 the Shenzhen Research Institute of Big Data under Grant 2019ORF01012. *(Corresponding author: Xiaoming Zhang.)*

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

Xiaoming Zhang is 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 (e-mail: yolixs@buaa.edu.cn).

Feiran Huang is with the College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou 510632, China.

Lei Cheng is with the 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 http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3016083

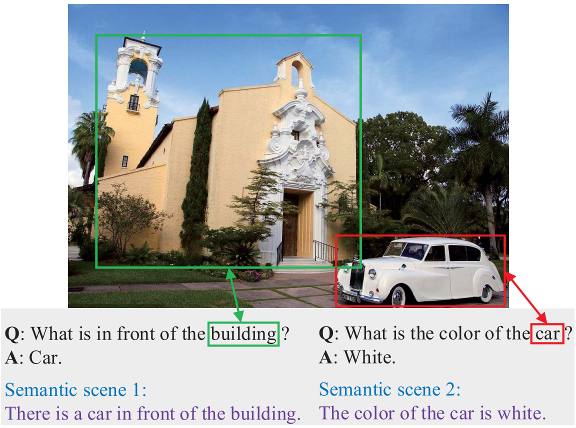

Q: What is in front of the building?
A: Car.
Semantic scene 1:
There is a car in front of the building.
Q: What is the color of the car?
A: White.
Semantic scene 2:
The color of the car is white.

Fig. 1. Two examples of VQA. (1) There exist fine-grained correlations between question words and image regions. The words "building" and "car" are correlated with the specific image regions. (2) A question–answer pair corresponds to a specific semantic scene existing inside the image. The semantic scene is unique for each question–answer pair and important to specify the answer for the corresponding question–image pair.

be modeled jointly to learn a more effective joint embedding to infer the answer. On the other side, each question–answer pair corresponds to a specific semantic scene existing in the image. The specific semantic scene is related to the answer and expected to be embedded in the question–image joint embedding. As the examples shown in Fig. 1, the question "What is in front of the building?" and its corresponding answer "Car" correspond to a semantic scene "There is a car in front of the building" in the image. Similarly, the question "What is the color of the car?" and its corresponding answer "White" correspond to another semantic scene "The color of the car is white" in the image. The two examples indicate that each image contains different answers for different questions, and the answer is determined by both of the question sentence and image, which is reflected by the specific semantic scene existing in the image. Based on the analysis, it can be considered that learning the semantic scene and encoding it in the joint embedding of the question–image pair is crucial for answer inference. Meanwhile, the answer-related information, i.e., semantic scene, also exists in the question–answer pair and is unique for each question–image pair to specify the answer. Therefore, it is highly desirable to match the question–image joint embedding with the answer-related question–answer embedding. Based on these clues, a more effective embedding can be learned to reflect the answer.

Although it seems evident to discover the answer-related semantic information, it is still difficult for the existing approaches to encode it for answer inference. First, most of the existing approaches [9]–[11] directly model the correlation between the image and question to learn a joint embedding. However, the question generated for the image is not specified, and thus, the learned embedding is difficult to reflect the answer-related information. For instance, the question "What is in front of the building?" can be generated as a question for images about a church, theater, or mansion. Therefore, inferring the answer by exploiting the multi-modal correlation between question and image directly may concentrate on some wrong regions that are unrelated to the answer. The reason is that the learning process lacks the guidance of answer-related information. On the other side, many existing approaches [12]–[14] utilize the feedback information from the supervised VQA classifier to learn the multi-modal correlation. However, it mainly focuses on improving classification accuracy rather than learning the answer-related features, which is not consistent with the process of answer inference.

The recently proposed multi-view adversarially learned inference model [15], which relies on an adversarial learning network for joint distribution matching and achieves promising performance in various multi-media tasks. Motivated by this work, we propose to take the advantage of adversarial learning to match the question–image joint embedding with the answer-related information. Especially, we expect to solve: 1) how to exploit the multi-modal correlation between the question words and image regions to learn an effective question–image joint embedding and 2) how to capture the answer-related information and embed it in the question–image joint embedding to infer the answer. The solutions result in a novel end-to-end model for VQA, i.e., adversarial learning with multi-modal attention (ALMA). The framework of ALMA mainly consists of four subnetworks as shown in Fig. 2. The two embedding generators, i.e., question–image embedding and question–answer embedding, are designed to generate modality-invariant representations for the items from the triplet ⟨image, question, and answer⟩. In the question–image embedding network, multi-modal attention with the Siamese similarity learning method is employed to capture the correlation between image and question to learn a more effective question–image joint embedding. The two embedding generators try to confuse the embedding discriminator, i.e., the embedding discrimination network, which acts as an adversary and has the objective to distinguish the representations learned from the two generators. Through the adversarial learning between the generators and the discriminator, the learned question–image joint embedding can capture the answer-related information existing in the question–answer pair. When the adversarial learning converges, the question–image joint embedding is optimal to reflect the answer to the question. In addition, an answer prediction network is designed by dismantling the question features from the learned answer-related question–image joint embedding to infer the answer.

The main contributions are summarized as follows.

1) We investigate the problem of VQA by utilizing the adversarial learning method to explore the answer-related information, by which a more effective and answer reflected joint embedding can be learned for the question–image pair to improve the performance of VQA.
2) A multi-modal attention model with the Siamese similarity learning method is proposed to explore the alignment between image regions and question words for fine-grained correlation learning.
3) A novel end-to-end model ALMA is built for VQA, and the extensive experiments conducted on three benchmark data sets confirm the superiority of ALMA compared with state-of-the-art approaches.

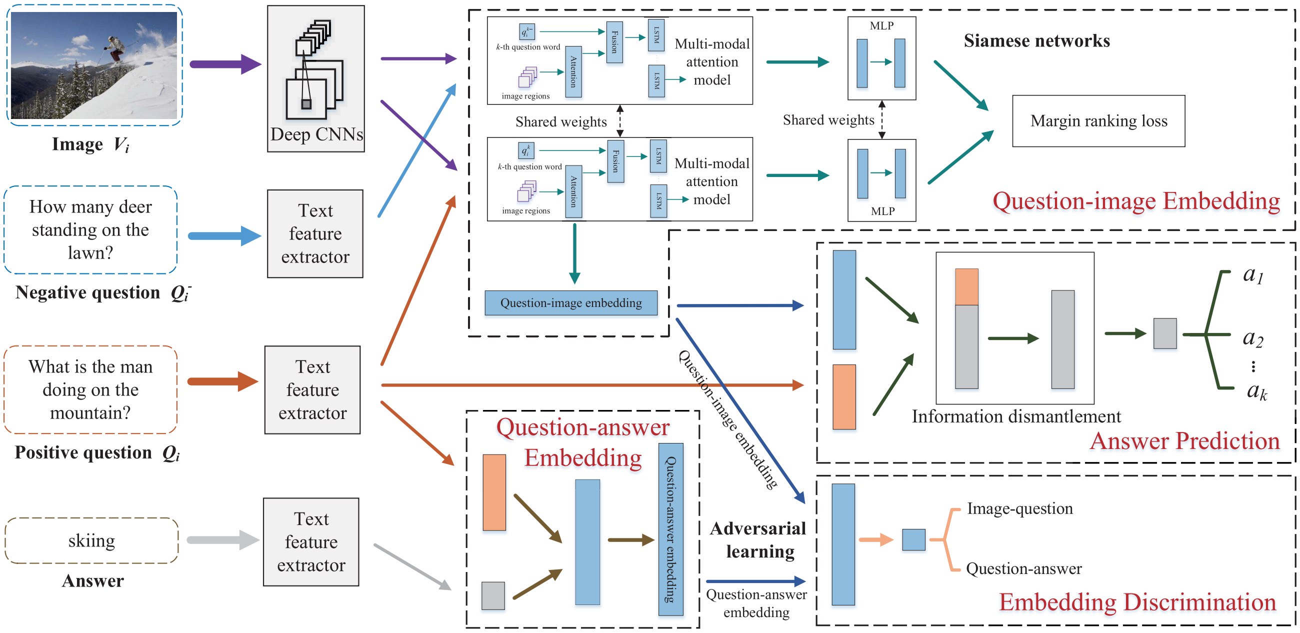


Fig. 2. Framework of ALMA, which mainly consists of four subnetworks as the four dotted boxes highlighted with the red font. Specifically, the two embedding generators, i.e., question–image embedding and question–answer embedding, generate two modality-invariant representations for items from the triplet ⟨image, question, answer⟩. For question–image embedding, a multi-modal attention model with the Siamese networks is utilized to explore the fine-grained correlation between image regions and question words. The architecture of the Siamese networks is made up of two identical parameter-shared components which are input with the positive and negative question–image pairs, respectively. The embedding discrimination network plays as a discriminator that tries to distinguish the two representations, by which the question–image representation is reinforced to learn the answer-related information. The answer prediction network dismantles question features from the question–image joint embedding to infer the answer.

This article extends its preliminary version ALARR [16] in terms of both technique and performance evaluation. First, we improve ALMA by employing multi-modal attention with the Siamese similarity learning method to explore the fine-grained correlation between image regions and question words. It can encode the fine-grained visual–textual correlation for more effective embedding learning. Second, another data set and more experiments are added to analyze the performance of ALMA more comprehensively. Finally, more detailed information about the subnetworks designed in ALMA is discussed.

The rest of this article is organized as follows. We first summarize the related existing works in Section II. Then, the proposed ALMA is detailed in Section III. Section IV describes the experiments and analysis of the results. Finally, the conclusion and future work are introduced in Section V.

II. Related Work

A. Visual Question Answering

Most of the existing works [17]–[20] on VQA are based on deep neural networks. These methods can be roughly divided into three categories. The first type of method mainly uses the visual attention mechanism [10], [21], [22] to find the crucial image regions related to the question words for visual–textual correlation learning. The approach presented in [14] utilizes a multiquery method to gradually explore the meaningful image regions, which is the first time to employ multi-modal attention in VQA task. After that, the methods detailed in [11], [23], and [24] model the correlation between question and the corresponding image using multilevel attention mechanisms and get a promising performance. The second type of method relies on an external knowledge base [25]–[27], which expects to find useful information from outside to enable the model containing richer information to improve VQA performance. For instance, an external knowledge DBpedia acting as the auxiliary knowledge base is employed in [25] to capture more meaningful information related to the answer. The last type of method is based on multi-modal fusion, which concentrates on learning high-level interactions between visual and linguistic features. Multimodal low-rank bilinear attention network (MLB) [13] and multimodal compact bilinear pooling (MCB) [28] are the famous models, in which multi-modal pooling methods realized by bilinear models are proposed to learn the question–image joint representation. A generalized multi-modal pooling approach MUTAN [29] unifies MCB and MLB into the same framework with fewer parameters and better performance. Although these types of methods have achieved certain success, they learn the joint embedding mainly from the combination of the question and image features directly, which cannot effectively exploit the answer-specific information existing inside the image. In comparison to these works, the model uses multi-modal attention with the Siamese similarity learning method to explore the fine-grained correlation between question words and image regions. Meanwhile, adversarial learning is used to embed the answer-specific information into the joint embedding to improve the performance of VQA.

B. Adversarial Learning

Generative adversarial networks (GANs) [30] have been proved successful in image generation [31], [32], dialog system [33]–[35], and information retrieval [5], [6], [36]. The principle of GAN is the interplay between two networks, i.e., the generator network and the discriminator network, conducted as a minimax game. On the one side, the discriminator is tasked with distinguishing real data samples from the fake ones generated by the generator. On the other side, the generator constantly generates data samples to confuse the discriminator and make the discriminator's task harder. The work presented in [31] describes an information-theoretic extension called infoGAN which can learn disentangled representations in a completely unsupervised manner. An adversarial training approach detailed in [33] achieves a good performance on the dialog generation task. This method employs adversarial learning to make the generated sequences existing clear distinction with the human-generated dialog. The approach presented in [36] proposes a cross-modal retrieval model, which utilizes the adversarial network to enable flexible retrieval experience across different modalities in a common subspace. An adversarial attention network introduced in [37] tries to regularize the generated multi-modal representation by matching the posterior distribution of the representation to the given priors. The method presented in [38] focuses on the problem of nonidentifiability in bidirectional adversarial networks. An unified perspective of considering GAN models as joint matching is provided to tackle this problem. Although some existing works [39]–[42] adopt adversarial learning in the VQA task, the question–image joint representation is directly learned by the generator from the multi-modal attention, ignoring the semantic scene in the question–answer pair, which makes it difficult to capture the answer-related information. This model, on the contrary, uses the adversarial learning to match the question–image joint embedding with the answer-specific information, which makes the joint embedding more effective to reflect the answer.

III. ALMA FOR VISUAL QUESTION ANSWERING

A. Problem Statement

Before the problem formulation, we define the notations used in this article. Without losing of generality, two modalities of data are commonly seen in the VQA task, i.e., visual image and textual sentence. Let $\mathcal{Q} = \{\mathcal{Q}_1, \ldots, \mathcal{Q}_i, \ldots, \mathcal{Q}_n\}$ and $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_i, \ldots, \mathcal{V}_n\}$ denote n samples of questions and images, respectively. For the image $\mathcal{V}_i$, the deep convolutional neural networks (CNNs), i.e., the VGG19 network [43] pretrained on ImageNet 2012 classification challenge data set, are used to extract the image regions represented as $\mathcal{R}_i = \{r_{i,1}, \ldots, r_{i,j}, \ldots, r_{i,m}\} \in \mathbb{R}^{c \times m}$, where $r_{i,j}$ represents the embedding of jth region, c is the dimension of a region, and m is the number of the regions. Let $\mathcal{Q}_i = \{q_i^1, \ldots, q_i^t, \ldots, q_i^l\} \in \mathbb{R}^{e \times l}$ denotes the question embedding with length l, where q_i^t with dimension e is a word vector encoded by Glove [44]. $\mathcal{A} = \{a_1, \ldots, a_i, \ldots, a_k\}$ represent the predefined answer set, where k is the number of the candidate answers and $a_i \in \mathbb{R}^e$ is usually a single word. $\mathcal{Y} = \{y_1, \ldots, y_i, \ldots, y_n\}$ denotes the ground-truth answer labels for the questions, where $y_i \in \mathbb{R}^k$ is a one-hot vector.

We treat the VQA task as a maximum likelihood estimation problem. Given a question $\mathcal{Q}_i$ and the corresponding image $\mathcal{V}_i$, the VQA problem can be solved by computing the likelihood probability distribution p_{vqa}. For each answer a' in the answer set $\mathcal{A}$, the model outputs the probability of which answer being the correct one. It is formulated as follows:

$$\hat{a} = \underset{a' \in \mathcal{A}}{\arg\max}\ p_{\text{vqa}}\big(a' | \mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{vqa}}\big) \tag{1}$$

where $\hat{a}$ is the predicted answer, and θ_{vqa} are the model parameters. The framework of ALMA is shown in Fig. 2, which mainly contains four subnetworks, as detailed as follows.

B. Question–Image Embedding

Attention mechanism has been proved to be an effective approach to learn the correlation between different modalities as presented in [45] and [46]. Different from these works, we propose a multi-modal attention model with the Siamese learning method to capture the fine-grained correlation between image regions and question words. It can facilitate more effective learning of the question–image joint embedding.

Given a question–image pair $(\mathcal{Q}_i, \mathcal{V}_i)$, the words in the question sentence are usually reflected by the corresponding regions in the image. In the attention network, for each word q_i^t, a score $\alpha_{i,j}^t$ between 0 and 1 is assigned to each region $r_{i,j}$ based on its relevance with the content of q_i^t. The softmax function is employed to evaluate $\alpha_{i,j}^t$ as follows:

$$\alpha_{i,j}^t = \frac{\exp\left(s_{i,j}^t\right)}{\sum_{j=1}^m \exp\left(s_{i,j}^t\right)} \tag{2}$$

where

$$s_{i,j}^t = \varphi\left(\left(q_i^t\right)^T \mathcal{W}_s r_{i,j} + b_s\right) \tag{3}$$

is the unnormalized attention score which reflects the closeness of the relationship between the word q_i^t and the image region $r_{i,j}$. $\mathcal{W}_s$ and b_s are the weight matrix and the bias term to be learned, respectively. $\varphi(\cdot)$ is a nonlinear activation function tanh. $\alpha_{i,\cdot}^t$ is used to normalize the attention of word q_i^t over all the regions $\{r_{i,j}\}_{1 \leq j \leq m}$. Then, the attention scores are utilized to regulate the intensity of attention on different regions. The weighted sum of all candidate regions is mapped from visual feature space to the input word space q_i^t:

$$u_i^t = \sum_{j=1}^m \alpha_{i,j}^t r_{i,j}. \tag{4}$$

Compared with the original visual features shared by all words, the weighted visual feature mapping u_i^t is more effective to represent the regions related to the current word q_i^t. The element-wise multiplication is applied on the two inputs u_i^t and q_i^t to obtain a joint feature vector o_i^t as follows:

$$o_i^t = u_i^t \circ q_i^t \tag{5}$$

where $\circ$ indicates the operation of element-wise multiplication. Since the question sentence is a sequence of words, the sequential model long short-term memory (LSTM) is used to handle the sequence input $\{o_i^1, o_i^2, \ldots, o_i^t, \ldots, o_i^l\}$. Namely, o_i^t is an input to the LSTM memory cell at time step t. The output of the last cell of LSTM acts as the joint embedding $\mathcal{E}_{\text{qv}}^i \in \mathbb{R}^d$ of the question–image pair. To intuitively present the

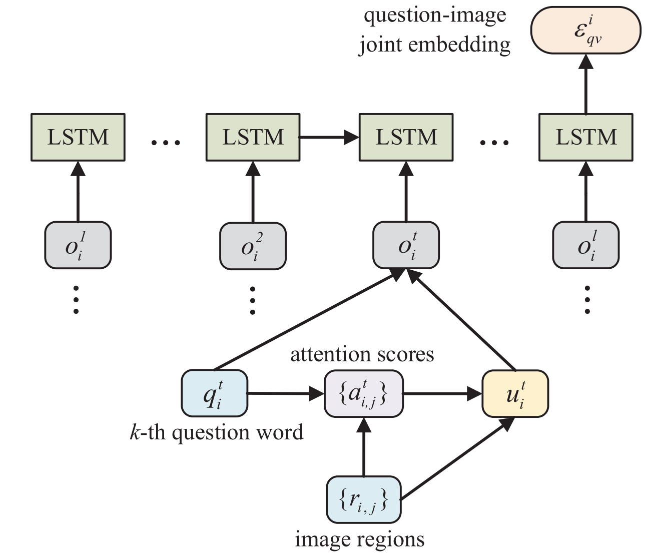


Fig. 3. Architecture of the multi-modal attention network.

calculation process, the whole procedure can be pipelined as a vector generation function

$$\mathcal{E}^i_{\text{qv}} = g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big) \tag{6}$$

where θ_{qv} is the set of the parameters in the multi-modal attention module. $\mathcal{E}^i_{\text{qv}}$ represents the joint embedding learned from the ith question–image pair. Fig. 3 shows the whole architecture of the multi-modal attention network.

It is expected that the multi-modal attention method can automatically assign the appropriate weights to the regions for each word. That is, the spatially structured image regions should be aligned with the semantic words correctly. However, there is no explicit knowledge to learn the alignment. Motivated by the visual-semantic embedding method [47], [48] used for multi-modal correlation learning, we employ Siamese similarity as an auxiliary knowledge to pilot the learning of the attention model. The Siamese networks measure the similarity between two inputs by mapping the inputs to the target space and calculating the distance (European distance, etc.). We take the full advantage of Siamese networks to learn the alignment between image regions and sentence words. Specifically, for a question–image pair $(\mathcal{Q}_i, \mathcal{V}_i)$, the image is first treated as the anchor, and a sentence unrelated to the image is sampled as the negative question $\mathcal{Q}_i^-$. Then, both $(\mathcal{V}_i, \mathcal{Q}_i)$ and $(\mathcal{V}_i, \mathcal{Q}_i^-)$ are input to the multi-modal attention model to obtain two question–image joint embeddings. The pairwise margin ranking loss is used to learn the matching scores between the two embeddings

$$\begin{aligned}\mathcal{L}_{\text{qv}}\big(\mathcal{Q}, \mathcal{V}, \mathcal{Q}^-; \theta_{\text{qv}}, \theta_h\big) &= \sum_{i=1}^{n} \max\Big(0, M - h\big(g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big); \theta_h\big) \\ &\quad + h\big(g\big(\mathcal{Q}_i^-, \mathcal{V}_i; \theta_{\text{qv}}\big); \theta_h\big)\Big)\end{aligned} \tag{7}$$

where function $h(\cdot)$ is used to learn the matching scores from the embeddings $g(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}})$ and $g(\mathcal{Q}_i^-, \mathcal{V}_i; \theta_{\text{qv}})$, which is conducted by a multi-layer perception (MLP). θ_h is the parameter set in the MLP networks. The weight parameters θ_{qv} and θ_h are shared by both of the positive and negative samples. The loss function aims to ensure that the matching score for the positive pair $(\mathcal{Q}_i, \mathcal{V}_i)$ is at least greater than the margin M compared to the negative pair $(\mathcal{Q}_i^-, \mathcal{V}_i)$.

For the negative pair $(\mathcal{Q}_i^-, \mathcal{V}_i)$, the question $\mathcal{Q}_i^-$ is mostly unrelated to the image $\mathcal{V}_i$. Therefore, there is no prominent alignment between the question words and image regions. Then, the formulations (2) are inappropriate to learn the negative attention weights. Namely, the attention weights defined in (2) could result in incorrect matching scores. For this reason, (2) is rewritten as the following formula to calculate the attention weights:

$$\alpha^t_{i,j} = \begin{cases} \dfrac{\exp\Big(s^t_{i,j}\Big)}{\sum_{j=1}^{m} \exp\Big(s^t_{i,j}\Big)}, & \text{input} : (\mathcal{Q}_i, \mathcal{V}_i) \\ 1/m, & \text{input:}\big(\mathcal{Q}_i^-, \mathcal{V}_i\big). \end{cases} \tag{8}$$

By this formula, the positive alignment can be assigned with a reasonable value to $\alpha^t_{i,j}$, whereas the attention in the negative pair is ignored.

C. Question–Answer Embedding

Actually, the information of a question–answer pair is related to an answer-specific semantic scene existing inside the image, which is expected to be embedded in the question–image joint embedding. Namely, capturing the answer-specific information residing in the question–answer pair can contribute to infer the answer. In this subsection, we describe an embedding method to encode the semantic scene inside the question–answer pair.

For the question–answer pair $(\mathcal{Q}_i, \mathcal{A}_i)$, LSTM is employed to learn the sequential correlation between the question words and the answer. In particular, the answer is treated as a word and placed at the beginning and end of the sequence of words in the question. Then, the entire sequence words are encoded using LSTM. Finally, the output of the last cell of LSTM acts as the question–answer joint embedding $\mathcal{E}^i_{\text{qa}} \in \mathbb{R}^d$. Note that the number of the LSTM cells is set the same as the number of LSTM cells d in the question–image embedding model. In this way, the question–answer embedding $\mathcal{E}^i_{\text{qa}}$ has the same dimension with the question–image embedding $\mathcal{E}^i_{\text{qv}}$.

We pipeline the whole procedure of question–answer embedding as a vector generation function

$$\mathcal{E}^i_{\text{qa}} = f\big(\mathcal{Q}_i, \mathcal{A}_i; \theta_{\text{qa}}\big) \tag{9}$$

where θ_{qa} is the parameter set in the question–answer embedding module. and $\mathcal{E}^i_{\text{qa}}$ is the question–answer joint embedding learned from the ith sample. Obviously, $\mathcal{E}^i_{\text{qa}}$ contains the answer-specific information as it is learned from the answer directly.

D. Embedding Discrimination

For the purpose of matching the question–image joint embedding with the answer-related information, the embeddings $\mathcal{E}^i_{\text{qv}}$ and $\mathcal{E}^i_{\text{qa}}$ learned from the triplet ⟨image, question, and answer⟩ are expected to be as similar as possible. To achieve this goal, we conduct adversarial learning between the two embedding generators (question–image embedding and question–answer embedding) and the embedding discriminator. When the adversarial model converges, the joint representation learned from the question–image pair is consistent with the representation learned from the question–answer pair. Then, the question–image joint embedding is optimal to reflect the answer.

To distinguish the embeddings as reliable as possible, an embedding discriminator is constructed. It is implemented by a two-layer MLP activated by softmax in the last layer. The target of the discrimination contains two classes, i.e., question–image and question–answer

$$d(x; \theta_{\text{dis}}) = \text{softmax}(\mathcal{W}_{d2}(\delta(\mathcal{W}_{d1}x + b_{d1})) + b_{d2}) \quad (10)$$

where $\mathcal{W}_{d1}$ and $\mathcal{W}_{d2}$ are weight matrix parameters, b_{d1} and b_{d2} are bias parameters in the MLP, δ is the nonlinear activation function relu, and x is an embedding vector either $\mathcal{E}^i_{\text{qv}}$ or $\mathcal{E}^i_{\text{qa}}$. In addition, θ_{dis} is the set of parameters in the embedding discriminator.

In the proposed ALMA model, the embedding discriminator is designed to play as the "discriminator" in GAN [30]. In order to train the discriminator, the joint embeddings $\mathcal{E}^i_{\text{qa}}$ and $\mathcal{E}^i_{\text{qv}}$ are assigned with the label of $\overline{01}$. The discriminator tries to classify the two embeddings $\mathcal{E}^i_{\text{qv}}$ and $\mathcal{E}^i_{\text{qa}}$ correctly. It aims to minimize the discriminating loss $\mathcal{L}_{\text{dis}}$ as follows:

$$\begin{aligned}&\mathcal{L}_{\text{dis}}(\mathcal{Q}, \mathcal{V}, \mathcal{A}; \theta_{\text{dis}})\\ &= -\frac{1}{n}\sum_{i=1}^{n}\Big(\log\big(d\big(f\big(\mathcal{Q}_i, \mathcal{A}_i; \theta_{\text{qa}}\big); \theta_{\text{dis}}\big)\big)\\ &\qquad + \log\big(1 - d\big(g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big); \theta_{\text{dis}}\big)\big)\Big) \quad (11)\end{aligned}$$

where the discriminating loss $\mathcal{L}_{\text{dis}}$ is essentially the cross-entropy loss of the embedding modality classification.

On contrary to the discriminator, the embedding generators generate modality-invariant embeddings and try to confuse the embedding discriminator in the training process. In other words, the question–image embedding module aims to make the discriminator misclassify $\mathcal{E}^i_{\text{qv}}$ as $\mathcal{E}^i_{\text{qa}}$, whereas the question–answer embedding module tries to make the discriminator misclassify $\mathcal{E}^i_{\text{qa}}$ as $\mathcal{E}^i_{\text{qv}}$. Thus, the joint embeddings $\mathcal{E}^i_{\text{qa}}$ and $\mathcal{E}^i_{\text{qv}}$ are assigned with the label of $\overline{10}$. The embedding generators focus on minimizing the embedding loss $\mathcal{L}_{\text{emb}}$ as follows:

$$\begin{aligned}&\mathcal{L}_{\text{emb}}\big(\mathcal{Q}, \mathcal{V}, \mathcal{A}; \theta_{\text{qv}}, \theta_{\text{qa}}\big)\\ &= -\frac{1}{n}\sum_{i=1}^{n}\Big(\log\big(1 - d\big(f\big(\mathcal{Q}_i, \mathcal{A}_i; \theta_{\text{qa}}\big); \theta_{\text{dis}}\big)\big)\\ &\qquad + \log\big(d\big(g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big); \theta_{\text{dis}}\big)\big)\Big). \quad (12)\end{aligned}$$

The same as the discriminating loss in (11), the embedding loss is essentially the cross-entropy loss of embedding misclassification.

E. Answer Prediction

Similar to the existing approaches [10], [19], [29], we treat VQA as a multi-class classification problem and, hence, learn a classifier to infer the answer. As discussed above, through the adversarial learning in the embedding discrimination, $\mathcal{E}^i_{\text{qv}}$ can learn the answer-related information as similar as $\mathcal{E}^i_{\text{qa}}$ and, hence, is optimal to reflect the answer. We design an answer prediction network to extract the answer features from the question–image joint embedding $\mathcal{E}^i_{\text{qv}}$ for answer inference.

In the question–answer embedding module, the joint embedding $\mathcal{E}^i_{\text{qa}}$ is obtained by integrating the information from the question features and the answer features. Through the adversarial learning, $\mathcal{E}^i_{\text{qv}}$ is consistent with $\mathcal{E}^i_{\text{qa}}$. Then, we can consider that $\mathcal{E}^i_{\text{qv}}$ also contains both the question and the answer features. Contrary to the learning process of the question–answer embedding, the answer-specific features are extracted by dismantling the question features h_i from $\mathcal{E}^i_{\text{qv}}$ through element-wise division as follows:

$$z_i = \mathcal{E}^i_{\text{qv}} \oslash h_i \quad (13)$$

where $\oslash$ represents the element-wise division between two vectors. $h_i \in \mathbb{R}^d$ is the output of the last cell of LSTM used to encode the words of the question sentence $\mathcal{Q}_i$. Therefore, z_i can be considered as the answer-specific features derived from the image. For the implementation of the answer classifier, a two-layer MLP network and softmax function are utilized to calculate the probability for each candidate answer item

$$p_i = \text{softmax}\big(\mathcal{W}_{p2}\big(\delta\big(\mathcal{W}_{p1}z_i + b_{p1}\big)\big) + b_{p2}\big) \quad (14)$$

where $p_i \in \mathbb{R}^k$, and k is the number of the predefined candidate answers. $\mathcal{W}_{p1}$ and $\mathcal{W}_{p2}$ are two weight matrixes, b_{p1} and b_{p2} are bias terms, and δ is the nonlinear activation function relu. Similarly, we pipeline the whole answer prediction procedure as a generation function

$$p_i = a\big(g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big), \mathcal{Q}_i; \theta_{\text{ans}}\big) \quad (15)$$

where θ_{ans} is the parameter set in the answer classifier.

To train the answer classifier, the cross-entropy is also used to define the answer classification loss $\mathcal{L}_{\text{ans}}$:

$$\begin{aligned}&\mathcal{L}_{\text{ans}}\big(\mathcal{Q}, \mathcal{V}; \theta_{\text{qv}}, \theta_{\text{ans}}\big)\\ &\qquad = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log\big(a\big(g\big(\mathcal{Q}_i, \mathcal{V}_i; \theta_{\text{qv}}\big), \mathcal{Q}_i; \theta_{\text{ans}}\big)\big) \quad (16)\end{aligned}$$

where y_i is a one-hot vector representing the ground-truth answer label for the ith question–image instance.

F. Adversarial Learning: Optimization

The structure of GAN includes two modules, i.e., the generator and the discriminator. The optimization process of the GAN involves two loss functions, generator loss and discriminator loss, playing as a minimax game. The generator and discriminator are iteratively trained until the model converges. Namely, the two losses reach an agreement with only a little vibration. The difference between standard GAN and the proposed ALMA is that there are two generators in ALMA. Therefore, we need to combine the loss functions from the two generators together for model optimization. Since the two generators generate question–image embedding and question–answer embedding, the loss functions related to the two embeddings are treated as one part of the generator loss. Moreover, the answer classification loss is considered as another part of the generator loss as the question–image joint embedding is input to the answer classifier for answer inference. Then, the margin ranking loss in (7), the embedding learning loss in (12), and the answer classification loss in (16) are simultaneously minimized by rewriting the loss function for the generator

$$\begin{aligned}&\mathcal{L}_g\big(\mathcal{Q},\mathcal{V},\mathcal{Q}^-,\mathcal{A};\theta_{\mathrm{qv}},\theta_h,\theta_{\mathrm{qa}},\theta_{\mathrm{ans}}\big)\\ &\quad=\mathcal{L}_{\mathrm{ans}}\big(\mathcal{Q},\mathcal{V};\theta_{\mathrm{qv}},\theta_{\mathrm{ans}}\big)+\alpha\cdot\mathcal{L}_{\mathrm{emb}}\big(\mathcal{Q},\mathcal{V},\mathcal{A};\theta_{\mathrm{qv}},\theta_{\mathrm{qa}}\big)\\ &\qquad+\beta\cdot\mathcal{L}_{\mathrm{qv}}\big(\mathcal{Q},\mathcal{V},\mathcal{Q}^-;\theta_{\mathrm{qv}},\theta_h\big)+\mathcal{L}_{L_2}\big(\theta_{\mathrm{qv}},\theta_h,\theta_{\mathrm{qa}},\theta_{\mathrm{ans}}\big)\end{aligned}\tag{17}$$

where α and β are the hyperparameters, and $\mathcal{L}_{L_2}$ is a L2 regularization term to prevent overfitting. The discriminating loss in (11) is rewritten to add a L2-norm regularizer for the discriminator

$$\mathcal{L}_d(\mathcal{Q},\mathcal{V},\mathcal{A};\theta_{\mathrm{dis}})=\mathcal{L}_{\mathrm{dis}}(\mathcal{Q},\mathcal{V},\mathcal{A};\theta_{\mathrm{dis}})+\mathcal{L}_{L_2}(\theta_{\mathrm{dis}}).\tag{18}$$

The whole procedure of exploring the answer-related information to infer the answer is realized by the adversarial network. It is implemented by alternatively training the generator and discriminator with the optimization algorithm of stochastic gradient descent (SGD) over the shuffled minibatches. In the learning step of generator, parameters $\theta_{\mathrm{qv}},\theta_h,\theta_{\mathrm{qa}}$, and θ_{ans} are learned by minimizing the loss function of (17). For the discriminator, the parameters θ_{dis} are updated according to the loss function of (18). Algorithm 1 shows the details of the training procedure of ALMA.

Algorithm 1 Pseudocode of Optimizing ALMA With Step Size μ, Using Mini-Batch SGD Algorithm. g-Steps and d-Steps Are Hyperparameters

```
Input: min-batch images $\mathcal{V}$, question $\mathcal{Q}$ and the corresponding answer $\mathcal{A}$. Negative sampled mini-batch question $\mathcal{Q}^-$, mini-batch answer label $\mathcal{Y}$, hyperparameters $\alpha$ and $\beta$
Output: min-batch answer probability vectors for the corresponding question–image pairs.
repeat
    for g-steps do
        update parameters $\theta_{vq}$, $\theta_h$, $\theta_{\mathrm{qa}}$ and $\theta_{ans}$ by descending their stochastic gradients for the generator:
        $\theta_{vq} \longleftarrow \theta_{vq} - \mu\cdot\nabla_{\theta_{vq}}\mathcal{L}_g(\mathcal{Q},\mathcal{V},\mathcal{Q}^-,\mathcal{A};\theta_{vq})$
        $\theta_h \longleftarrow \theta_h - \mu\cdot\nabla_{\theta_h}\mathcal{L}_g(\mathcal{Q},\mathcal{V},\mathcal{Q}^-,\mathcal{A};\theta_h)$
        $\theta_{\mathrm{qa}} \longleftarrow \theta_{\mathrm{qa}} - \mu\cdot\nabla_{\theta_{\mathrm{qa}}}\mathcal{L}_g(\mathcal{Q},\mathcal{V},\mathcal{Q}^-,\mathcal{A};\theta_{\mathrm{qa}})$
        $\theta_{ans} \longleftarrow \theta_{ans} - \mu\cdot\nabla_{\theta_{ans}}\mathcal{L}_g(\mathcal{Q},\mathcal{V},\mathcal{Q}^-,\mathcal{A};\theta_{ans})$
    end for
    for d-steps do
        update parameters $\theta_{dis}$ by descending its stochastic gradients for the discriminator:
        $\theta_{dis} \longleftarrow \theta_{dis} - \mu\cdot\nabla_{\theta_{dis}}\mathcal{L}_d(\mathcal{Q},\mathcal{V},\mathcal{A};\theta_{dis})$
    end for
until ALMA converges
```

IV. Experiments

To evaluate the performance of ALMA, extensive experiments are performed on three VQA benchmark data sets by comparing it with state-of-the-art approaches.

A. Data Sets and Baselines

The data sets detailed as follows are used to evaluate the performance of the method.

VQA v1.0 [49] is one of the most popular data sets used in the VQA task. A total of 123 287 images contained in this data set are created from the MS-COCO data set. For each question, ten answers are generated from different crow-sourced workers. There are three categories of questions divided in this data set, i.e., yes/no, number, and other. In addition, the data set contains three splits, i.e., train (248 349 questions), val (121 512 questions), and test (244 302 questions). The test set is divided into two splits: test-std and test-dev. Moreover, there exist two subtasks: open-ended (OE) and multiple-choice (MC). In the experiments, the top 1000 frequent answers are employed as the possible outputs, and these answers cover 82.7% of the total answers.

VQA v2.0 [50] is constructed to minimize the effectiveness of learning data set priors. Compared with the VQA v1.0 data set, this data set is more balanced, and no subtasks exist in the VQA v2.0 data set. For each question of the data set, there exist a couple of similar images which results in two different answers to the question. In total, the data set consists of 204 721 images, 328 120 questions, and 22 523 answers. Moreover, it is about twice as large as the VQA v.10 data set.

COCO-QA [3] is also a widely used VQA data set, which is automatically created from the captions of the images existing in the MS-COCO data set. This data set is divided into a training set and a test set. In the training set, there are 78 763 samples generated from 8000 images. In addition, 4000 images are used to generate 38 948 samples in the test set. Four categories of question including object (70%), number (7%), color (17%), and location (6%) are contained in this data set. Especially, only 435 unique answers exist in the data set, and all of the answers are single word.

The following approaches are used as the main baselines.

1) **SAN** [14]: A stacked attention network uses the embedding of the question as a query to search for the meaningful visual regions.
2) **HieCoAtt** [9]: A novel coattention model that jointly reasons about image and question attention in a hierarchical fashion.
3) **VQA-AA** [41]: An adversarial learning-based method uses a generative attention model to produce more diverse visual attention maps for question answering.
4) **Dual-MFA** [12]: A model jointly attends on the detected boxes and free-form image regions related to the question for more accurate VQA.
5) **DCN** [51]: A dense coattention network proposed to enable improved fusion of embeddings encoded from image regions and sentence words.
6) **MUTAN** [29]: An approach focuses on the multi-modal fusion between visual and textual representations, which designs a multi-modal tensor-based Tuker decomposition with a low-rank matrix constraint to control the full bilinear interaction's complexity.
7) **ALARR** [16]: The preliminary version of this article is that it leverages a three-level attention model with adversarial learning method for answer inference.

TABLE I

EXPERIMENTAL RESULTS ON THE VQA V1.0 DATA SET. "–" REPRESENTS THAT THE DATA ARE UNAVAILABLE. THE BEST VALUES ARE HIGHLIGHTED BY BOLD

Methods	VQA v1.0 Test-dev					VQA v1.0 Test-std				
	Open-Ended				MC	Open-Ended				MC
	All	Yes/No	Number	Other	All	All	Yes/No	Number	Other	All
VSE [3]										
-Q+I	52.62	75.53	33.65	37.34	-	-	-	-	-	-
-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 [19]	57.22	80.71	37.24	41.69	62.48	57.36	80.28	36.92	42.24	62.69
SAN [14]	58.7	79.3	36.6	46.1	-	58.9	79.11	36.41	46.42	-
DMN [18]	60.3	80.5	36.8	48.3	-	60.36	80.43	36.82	48.3	-
HieCoAtt [9]	61.8	79.7	38.7	51.7	65.8	62.1	79.95	38.22	51.95	66.07
MCB [28]	66.7	83.4	39.8	58.5	69.10	66.5	-	-	-	-
MLB [13]	66.77	84.57	39.21	57.81	-	66.89	84.02	37.9	54.77	68.89
VQA-AA [41]	-	-	-	-	-	66.9	84.0	38.1	56.5	69.8
DCN [51]	66.89	84.61	**42.35**	57.31	-	67.02	**85.04**	42.34	56.98	68.89
Dual-MFA [12]	67.11	84.63	39.48	58.14	70.04	67.09	83.37	40.39	56.89	69.97
MUTAN [29]	67.42	85.14	39.81	58.52	-	67.36	-	-	-	-
ALARR [16]	68.61	85.08	41.12	59.64	71.57	68.43	83.97	42.06	57.89	71.28
ALMA (ours)	**68.94**	**85.49**	42.09	**59.97**	**71.96**	**68.76**	84.11	**42.59**	**58.06**	**71.42**

B. Experiment Preparation

To encode the textual language, the 200-D word vectors pretrained on Glove [44] are employed. The images with channel RGB and size 224 × 224 are used as the visual input, and deep CNNs are utilized to extract visual features. We employ VGG19 networks [43] pretrained on ImageNet 2012 classification challenge data set as the deep CNN model. The region representation is expressed as the output of the layer "conv5_4," and the dimension is 512 × 14 × 14. This indicates that there are 196 candidate regions of each image.

For the hyperparameters in this model, we use the validation set in each data set to select the optimal values. Since the optimal values of some parameters are slightly different on different data sets, we choose the appropriate values to make the model perform as well as possible on different data sets. This also improves the generalization ability of the model. Specifically, in the question–image embedding network, the LSTM cells are set to have 256 hidden neurons. This means that the dimension d of the question–image embedding $\mathcal{E}^i_{\text{qv}}$ and question–answer embedding $\mathcal{E}^i_{\text{qa}}$ is 256. The network structure of MLP in the Siamese learning networks is set as $128 \rightarrow 64 \rightarrow 1$. The margin M in (7) is empirically set to be 0.5. For the embedding discriminator, the two-layer MLP is designed with a structure of $128 \rightarrow 2$. In the answer prediction network, the size of the first layer in the two-layer MLP is set to be 256, and the size of the last layer is set to be the number of the predefined candidate answers. All of the activation functions are fixed to relu(·) except special instructions. During the adversarial training process, we set g-steps = 5 and d-steps = 1 to make the model converge relatively fast and stable. The batch size is fixed to 128. The SGD algorithm with a momentum of 0.99 and a weight decay of 10^{-8} is used for optimization. To prevent overfitting, dropout with a value of 0.5 is used after each linear transformation, and L_2 regularization term is added to both of the generator loss and discriminator loss. The hyperparameters α and β are selected from an empirical range (α from 0 to 1 and β from 0 to 0.1) using the grid search method. The reported results on VQA v1.0 data set with different values of α and β are shown in Fig. 6(a).

Following [2], [3], and [49], we employ the classification accuracy as the evaluation metric in this article. Furthermore, WUPS [52] is another metric commonly used to measure the performance. Following [17] and [14], 0.0 and 0.9 are chosen as the thresholds to form WUPS@0.0 and WUPS@0.9 for evaluation, respectively. Notably, the evaluation of data sets VQA v1.0 and VQA v2.0 is different from the COCO-QA data set since each question has ten answer labels in the two data sets. We follow [49] to change the metric of accuracy: accuracy = min(#workers that provided that answer/3, 1). In other words, it is considered 100% accurate if at least three workers provide the exact answer.

C. Comparison With State-of-the-Art Approaches

In this subsection, we compare ALMA with state-of-the-art approaches on VQA v1.0, VQA v2.0, and COCO-QA data sets.

First, Table I shows the comparison results on the VQA v1.0 data set. One can see that ALMA achieves the best performance and significantly outperforms the preliminary version ALARR. On the Test-dev data set, ALMA outperforms the best baseline MUTAN by 1.52% on the overall accuracy in the OE task and achieves the highest accuracy 71.96% in the MC task. As for the Test-std data set, ALMA outperforms MUTAN by 1.4% on the overall accuracy in the OE task and obtains the highest accuracy in the MC task. Similar improvements can be seen in both Test-dev and Test-std data sets, where ALMA achieves better performance on each

TABLE II

EXPERIMENTAL RESULTS ON THE COCO-QA DATA SET. "−" REPRESENTS THE DATA IS UNAVAILABLE. THE BEST VALUES ARE HIGHLIGHTED BY BOLD

Methods	Accuracy	Object	Number	Color	Location	WUPS@0.9	WUPS@0.0
VSE [3]							
-VIS+LSTM	53.24	56.50	46.10	45.90	45.50	63.84	87.95
-2VIS+BLSTM	54.85	58.20	44.80	49.50	47.30	64.78	88.12
Img-CNN [17]	58.40	-	-	-	-	68.50	89.67
DPPnet [19]	61.19	-	-	-	-	70.84	90.61
SAN [14]	61.60	65.40	48.60	57.90	54.00	71.60	90.90
HieCoAtt [9]	65.40	68.00	51.00	62.90	58.80	75.10	92.0
Dual-MFA [12]	66.62	68.86	51.32	**65.89**	58.92	76.15	92.29
ALARR [16]	67.84	69.67	52.14	65.62	60.07	77.43	92.83
ALMA (ours)	**68.27**	**70.22**	**52.48**	65.81	**60.35**	**78.14**	**93.06**

TABLE III

EXPERIMENTAL RESULTS ON THE VQA V2.0 DATA SET. THE BEST VALUES ARE HIGHLIGHTED BY BOLD

Methods	VQA v2.0 Test-std			
	All	Yes/No	Number	Other
ATM(single)[53]	62.27	79.32	39.77	52.59
ATM(single+bottom-up)[53]	65.67	82.2	43.9	56.26
VKMN(single) [20]	64.36	83.7	37.9	57.79
VKMN(Ensemble) [20]	66.67	82.88	43.17	57.95
DCN(16) [51]	66.97	83.59	46.98	57.09
DCN(17) [51]	67.04	83.85	**47.19**	56.95
DCN(18) [51]	67.00	83.89	46.93	56.90
ALARR [16]	67.76	84.12	46.89	57.64
ALMA (ours)	**68.12**	**84.62**	47.08	**58.14**

TABLE IV

COMPARING ALMA WITH THE PRELIMINARY VERSION ALARR

Method	Accuracy	
	VQA v1.0	COCO-QA
ALARR	60.15	67.84
ALMA$_{lstm}$	60.65	67.96
ALMA$_{Siamses}$	60.94	68.08
ALMA$_{Siamses+lstm}$	61.42	68.27

question category. Notably, ALMA achieves a higher accuracy compared with VQA-AA which is the existing VQA method based on adversarial learning.

Second, Table II shows the comparison results on the COCO-QA data set. One can see that ALMA improves the overall accuracy of the preliminary version ALARR from 67.84% to 68.27%. Moreover, ALMA outperforms Dual-MFA and HieCoAtt by 1.65% and 2.87%. Similar improvements can be seen on the metrics of WUPS@0.9 and WUPS@0.0. In addition, compared to the baselines shown in this table, the proposed model ALMA obtains the highest accuracy on the three question types including object, number, and location, except the question type Color.

Third, Table III shows the comparison results on the VQA v2.0 data set, which shows the superiority of ALMA compared with the preliminary version ALARR. In addition, ALMA outperforms the representative methods VKM (Ensemble) and DCN (17) by 1.45% and 1.08% on the overall accuracy, respectively. As for the experimental results on each question category, ALMA achieves the highest accuracy on both Yes/No and other categories. For the number category, ALMA also obtains relatively high accuracy.

In summary, the proposed model ALMA obtains the best experimental results compared with the baselines on all the three data sets. It demonstrates that the adversarial network and multi-modal attention model play a crucial role in the VQA task. Meanwhile, ALMA achieves higher accuracy than VQA-AA, which manifests that using the semantic scene existing in the question–answer pair to guide the question–image embedding learning outperforms the generative attention model used in VQA-AA. Moreover, ALMA obtains more promising performance than the preliminary version ALARR, which indicates that the multi-modal attention with Siamese similarity methods is effective for VQA.

D. Comparison With the Preliminary Version

ALMA extends its preliminary version ALARR [16] in two aspects. First, ALMA uses multi-modal attention with the Siamese similarity learning method to replace the multi-level attention module designed in ALARR for question–image joint embedding learning. Second, LSTM is used to embed the sequential structure of the question and answer features to learn the question–answer embedding instead of the feature combination method in ALARR. Based on ALARR, the two aspects are improved respectively, and two models ALMA$_{\text{Siamses}}$ and ALMA$_{\text{lstm}}$ are obtained. Table IV shows the results of comparing ALMA with ALARR on two data sets. For VQA v1.0, the training and validation sets are used to train and test the models, respectively. The test set is not employed due to the restrictions of the online submission. For COCO-QA, both training and test sets are used to train and test the models, respectively. It can be observed that both ALMA$_{\text{Siamese}}$ and ALMA$_{\text{lstm}}$ achieve better performance than ALARR. In addition, ALMA$_{\text{Siamese+lsm}}$ shows more impressive performance than the other three models. This experiment indicates that the improvements on both question–image

TABLE V

Ablation Study on VQA v1.0 and COCO-QA Data Sets. "*" Denotes Variant Implementations of The Model

Methods	Accuracy	
	VQA v1.0	COCO-QA
No-Att	45.56	54.78
MM-Att	54.74	61.65
MM-Att + Sia	59.63	64.32
MM-Att + Sia + AL*	61.41	68.26
H-LSTM	58.62	65.91
E-LSTM	59.21	66.74
HE-LSTM*	61.36	68.24
Ans-Rep	59.84	65.73
Inf-Dis-Sub	60.71	67.75
Inf-Dis-Div*	61.35	68.26

embedding and question–answer embedding are significant and effective compared to the original version ALARR.

E. Ablation Study

We further conduct ablation experiments to study the effectiveness of the individual subnetworks designed in ALMA. Table V shows the evaluation results on data sets VQA v1.0 and COCO-QA. The first part of Table V displays the experimental results of some auxiliary methods used in multi-modal attention. One can see that the proposed multi-modal attention (MM-Att) model which learns the fine-grained visual–textual correlation improves the accuracy remarkably compared to the approach (No-Att) without attention model. MM-Att with Siamese similarity learning (MM-Att + Sia) explores the alignment between image regions and question words, outperforming MM-Att by 4.89% and 2.67% of accuracy on VQA v1.0 and COCO-QA data sets, respectively. The method MM-Att + Sia + AL employs adversarial learning to guide the MM-Att + Sia model to learn the answer-related information. It obtains the highest accuracy of 61.41% and 68.26% on the two data sets, respectively. The improvement demonstrates that Siamese similarity and adversarial learning have a prominent contribution to VQA. The second part of Table V shows the experimental results of utilizing different methods to embed the question–answer pairs. Compared with the methods which use the answer features as the input to the head of LSTM (H-LSTM) or the end of LSTM (E-LSTM), the method which inputs the answer features to the head and end of the LSTM (HE-LSTM) achieves a better performance. The last part of Table V compares the different methods of information dismantlement in the answer prediction network. Compared with the method that directly inputs the answer-related representation (Ans-Rep) into the answer classifier, information dismantlement using element-wise subtraction (Inf-Dis-Sub) and element-wise division (Inf-Dis-Div) obtains relatively higher accuracy. Meanwhile, Inf-Dis-Div outperforms Inf-Dis-Sub by 0.64% and 0.51% on VQA v1.0 and COCO-QA data sets, respectively. It certifies that employing the element-wise division method to learn the answer features from the question–image joint embedding is effective.

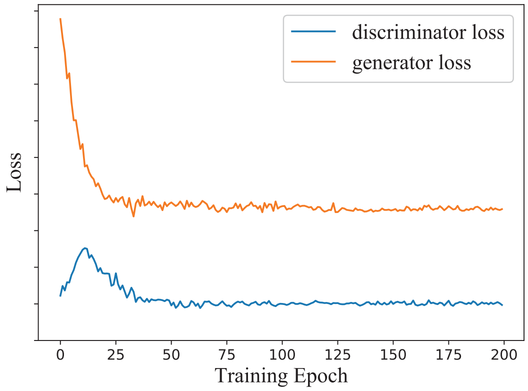


Fig. 4. Convergence of the adversarial learning between generator and discriminator.

F. Effect of Adversarial Learning

In the training process, the adversarial learning of ALMA is implemented by jointly optimizing the loss functions of the generator and discriminator alternatively. To further explore the effectiveness of adversarial learning in ALMA, we record the loss values of the generator and discriminator on data set VQA v1.0 from epoch 1 to 200 and show them in Fig. 4. It shows that the loss of the generator decreases quickly with certain vibration in the first 30 epochs. Relatively, the loss of the discriminator increases in the first 15 epochs and then also decreases with a certain vibration. The vibration indicates that the two players, i.e., the generator and the discriminator, compete fiercely during the period of the first 50 epochs. Then, the loss of both the generator and the discriminator decreases slowly and vibrates less. After about 100 epochs, the competition between the two players reaches an agreement and only a little vibration exists, which means that the model has converged to learn an answer-related representation. Inversely, if the loss of the discriminator explodes, the model will fail to impose the answer-related information existing inside the question–answer embedding to match the joint embedding of the question–image pair. On the other side, if the discriminator dominates the optimization process, the generator would fail to generate an answer-related question–image joint embedding.

G. Evaluation of Siamese Networks

To reveal the effect of Siamese networks used in this model, we perform a group of experiments to evaluate the validity of the Siamese networks on the three data sets.

The loss function of the Siamese networks plays a crucial role in learning the alignment between image regions and sentence words. Three candidate loss functions are considered in this model, i.e., cross-entropy loss, contrastive loss [48], and margin ranking loss. Fig. 5(a) shows the performance of ALMA with different loss functions in the Siamese networks. It can be observed that, compared with cross-entropy loss, contrastive loss and margin ranking loss can achieve higher accuracy on all the three data sets. Cross-entropy acting as a classification loss function to discriminate similar and

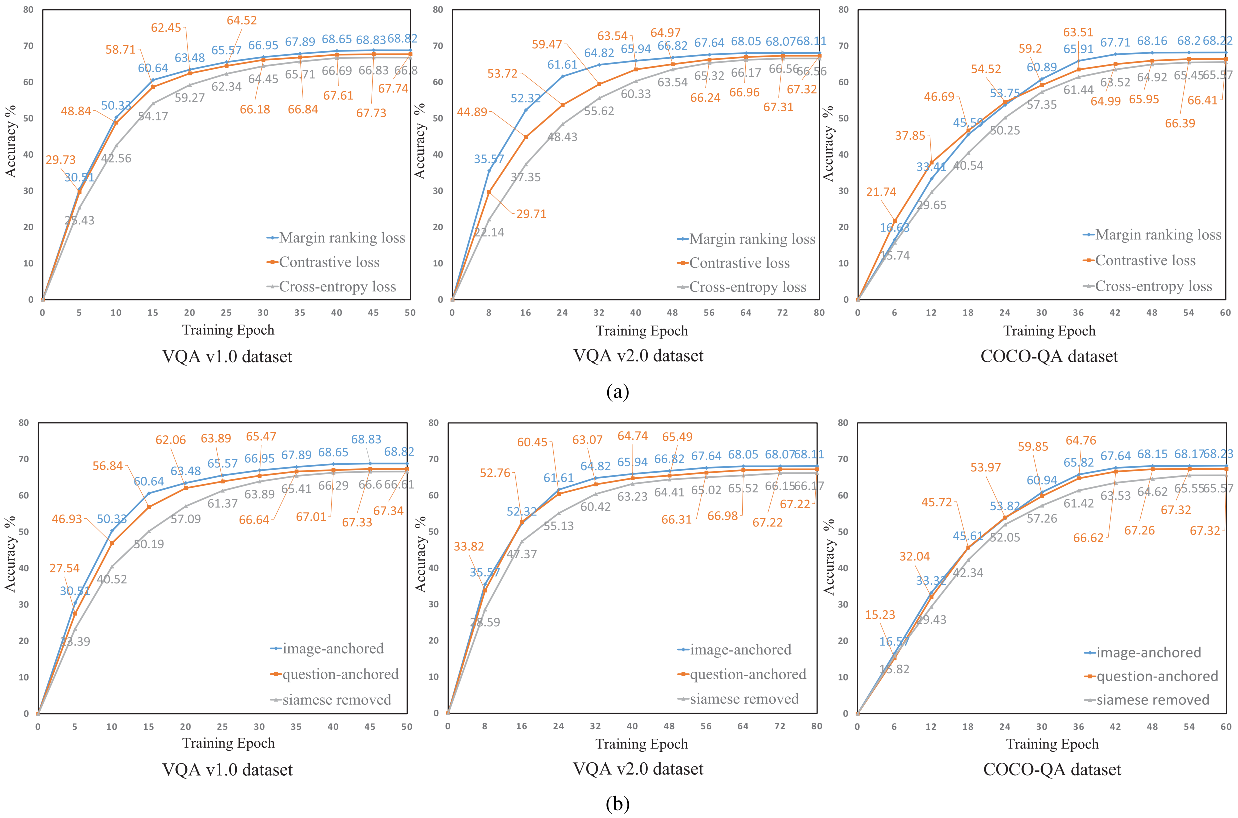


Fig. 5. Evaluation the performance of Siamese networks designed in the proposed model ALMA on the three data sets. (a) Performance of ALMA with different loss functions in the Siamese networks. (b) Performance of ALMA with different modalities acting as the anchor in the Siamese networks.

dissimilar input samples is obviously not enough in Siamese networks. Margin ranking loss and contrastive loss can make the input distance of two different categories as large as possible, thus achieving better performance. In addition, margin ranking loss achieves more promising performance than contrastive loss. Therefore, margin ranking loss is chosen as the final loss function in the Siamese networks.

In the ALMA model, the input of the Siamese networks contains an anchor, a positive and a negative sample about the anchor. We compare the performance of the Siamese networks with different modalities acting as the anchor. In addition, ALMA removed that the Siamese networks in the question–image embedding component is also considered for comparison. Fig. 5(b) shows the experimental results. One can see that Siamese networks significantly improve the performance of question answering, i.e., both image-anchored and question-anchored Siamese networks outperform the model without Siamese networks on all the three data sets. It can also be observed that the image-anchored Siamese networks show more promising performance than the question-anchored Siamese networks. The reason might be that the image contains more various patterns of information than the information existing in the language. The question-anchored Siamese networks contain two images for each question in the input, which increases the complexity of the model and may enable the model to focus on useless image regions.

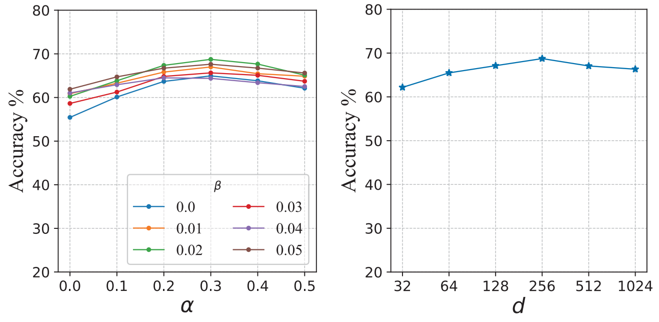
Typically, Siamese networks are used when the input samples are all of the same modality. This experiment demonstrates that Siamese networks are also suitable for multi-modal inputs and have a good performance in the VQA task.

H. Discussion of Model Collapse and Convergence

Model collapse and convergence are the common problems existing in GAN applications, which are also encountered in the training process. On the one hand, the generator always generates a small family of question–image embeddings, which results in similar answer predictions for many of the input samples. Obviously, it is a clear manifestation of the model collapse. We solve this problem using the following strategies according to the methods detailed in [54]. First, the softmax activation function in the last layer of the discriminator is removed. Second, the loss functions of the generator and discriminator do not take the form of logarithmic loss. Finally, when the discriminator parameters are updated, their absolute values are truncated to be not greater than a fixed constant. On the other hand, the model initially shows no convergence or even instability. To address this problem, we add regularization to the generator and discriminator losses to avoid overconfidence and overfitting. Moreover, a smaller batch size is selected for preliminary training to make the model jump out of the local extreme value, and then, a larger batch size is used to make the model converge.

I. Parameter Sensitivity

To analyze how the performance of ALMA is affected by parametrization, we evaluate how the value of the balance



(a) (b)

Fig. 6. Parameter sensitivity study for balance parameters and embedding dimensions on the VQA v1.0 data set. (a) Balance parameters. (b) Embedding dimension.

parameters (α and β) and the size of the answer-related embedding dimension (d) affect the result.

1) Balance Parameters: In the loss function (17) of the generator, α is used to balance the importance of the embedding loss, and β is a tradeoff parameter of the margin ranking loss. We fix the dimension of the answer-related representation $d = 256$ and use the grid search method to test α and β from an empirical range (α from 0 to 1 and β from 0 to 0.1). Based on the curves in Fig. 6(a), one can see that the embedding loss contributes to the performance since the model achieves relatively low accuracy when $\alpha = 0$. However, a too large value of α will result in overfitting. On the other side, when $\beta = 0$, relatively low performance is obtained, which indicates that a tradeoff term to learn the visual–textual correlation with the margin ranking loss is important for answer inference. In addition, compared with β, the value of α has more effect on the answer classification accuracy. It can be observed that the model achieves the highest accuracy when $\alpha = 0.3$ and $\beta = 0.02$.

2) Embedding Dimension: In ALMA, the dimension of the answer-related representation is the dimension d of both the question–image joint embedding $\mathcal{E}^i_{\text{qv}}$ and the question–answer joint embedding $\mathcal{E}^i_{\text{qa}}$. Fig. 6(b) shows how the dimension affects the performance by fixing $\alpha = 0.3$ and $\beta = 0.02$. It can be seen that the accuracy rises initially and then starts to drop slowly when the size of the dimension increases. It is reasonable that a higher dimension of the representation can embed more useful features for answer inference. However, a too large size of the dimension will bring in noise, which may decline the performance. Overall, a reasonable size of the dimension is greatly helpful for the performance, and ALMA reaches the highest accuracy when $d = 256$.

J. Visualization of Attention Model

To evaluate the effectiveness of the multi-modal attention and how the adversarial learning affects the attention, we visualize some attention maps generated by the multi-modal attention (MM-Att) and the attention guided by adversarial learning (MM-Att + AL). Following [55], Gaussian filtering and upsampling are utilized to visualize the weights of the two attention models MM-Att and MM-Att + AL. Examples in the VQA v1.0 test set are chosen for illustration. Given a question and the corresponding image, the visual attention weights are displayed over the image regions. Fig. 7 shows the visualization results of five examples. It can be observed that both MM-Att and MM-Att + AL can attend on the important regions for answer inference. Especially, MM-Att + AL is more effective to focus on the answer-related image regions as shown in Fig. 7(a), (d), and (e). In Fig. 7(b), MM-Att pays some attention on the unrelated regions of the answer, whereas MM-Att + AL gives the main attention to the "play badminton" regions related to the answer. In Fig. 7(c), MM-Att cannot recognize the little deer behind the larger deer and thus obtains a wrong answer, whereas MM-Att + AL accurately attend on the regions covering the head of the little deer and a correct answer is inferred. The results demonstrate that the designed adversarial learning model plays a crucial role in specifying the answer in the image.

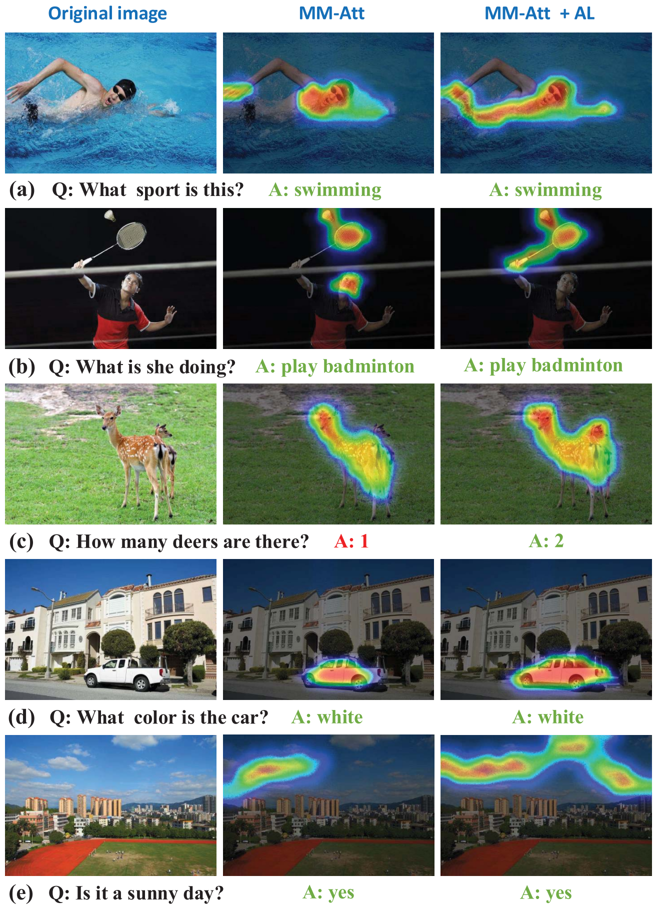
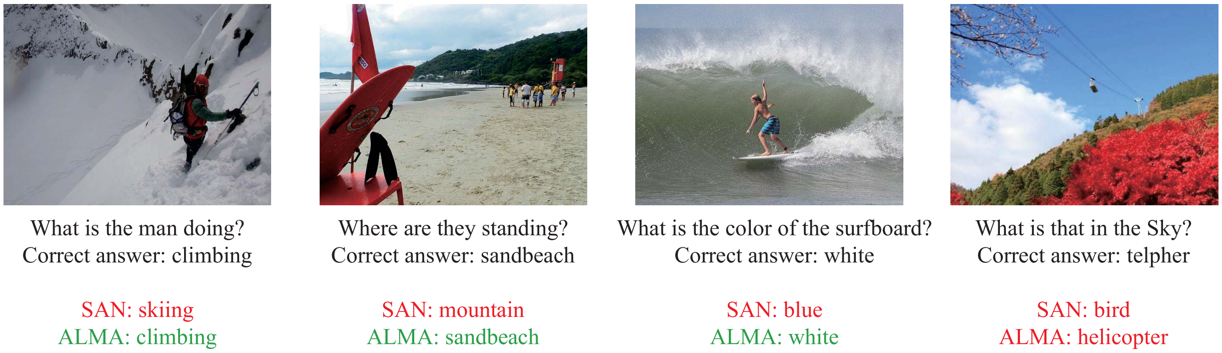

Original image
MM-Att
MM-Att + AL
(a) Q: What sport is this? A: swimming A: swimming
(b) Q: What is she doing? A: play badminton A: play badminton
(c) Q: How many deers are there? A: 1 A: 2
(d) Q: What color is the car? A: white A: white
(e) Q: Is it a sunny day? A: yes A: yes

Fig. 7. Visualization examples on the VQA v1.0 data set. The first column displays the original images. The second column shows the multi-modal attention (MM-Att) maps, and the last column is the multi-modal attention guided by adversarial learning (MM-Att + AL). The answers with green color indicate that they are correct, whereas red color represents wrong ones.

K. Qualitative Evaluation

We provide some example cases to evaluate the answer prediction results between ALMA and the previous approaches. Specifically, SAN [14] is selected as a representative of the baselines. The qualitative experiment results are shown in Fig. 8, where the examples are chosen from the VQA v1.0



What is the man doing?
Correct answer: climbing

SAN: skiing
ALMA: climbing

Where are they standing?
Correct answer: sandbeach

SAN: mountain
ALMA: sandbeach

What is the color of the surfboard?
Correct answer: white

SAN: blue
ALMA: white

What is that in the Sky?
Correct answer: telpher

SAN: bird
ALMA: helicopter

Fig. 8. Four example cases with the answers predicted by the proposed ALMA and the baseline model SAN. The green and red colors represent correct and wrong answers, respectively.

test set. It can be observed that ALMA predicts correct answers, whereas SAN obtains the wrong answers in the first three instances. Since there are few correlations between the question words and the image regions in the first instance, SAN obtaining the wrong answer might simply rely on the correlation between the man and the snow. However, ALMA can capture the fine-grained correlation between the man and the mountain to infer the correct answer. The last instance shows a failure case predicted by both SAN and ALMA. The reason might be that the telpher in the image is too small to identify compared with the samples in the training set. Moreover, fewer samples related to this answer existing in the training set might be another reason.

V. Conclusion and Future Work

In this article, we propose a novel VQA model, i.e., ALMA. Specifically, multi-modal attention with the Siamese similarity learning method is employed to learn the question–image joint embedding. Concurrently, a question–answer embedding model is designed to encode the question–answer pair. Then, adversarial learning between the question–image embedding and question–answer embedding is conducted to make the learned question–image embedding containing the answer-related information. Finally, the question–image embedding is input to an answer classifier for answer inference. Experiments performed on three widely used VQA data sets demonstrate the superiority of ALMA. Different from the existing approaches which learn the question–image joint embedding from visual regions and textual words directly, the approach focuses on learning the answer-related information to reflect the answer and thus achieves more promising performance.

There are many potential future extensions of this work. It would be interesting to investigate generative answers and zero-shot learning in the VQA problem.

References

[1] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, p. 6.

[2] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Neural Inf. Process. Syst. Conf.*, 2014, pp. 1682–1690.

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

[4] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.

[5] H. Li, C. Bai, L. Huang, Y. Jiang, and S. Chen, "Instance image retrieval with generative adversarial training," in *Proc. 26th Int. Conf. MultiMedia Modeling*, 2020, pp. 381–392.

[6] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[7] S. Yan, Y. Xie, F. Wu, J. S. Smith, W. Lu, and B. Zhang, "Image captioning via hierarchical attention mechanism and policy gradient optimization," *Signal Process.*, vol. 167, Feb. 2020, Art. no. 107329.

[8] M. Yang *et al.*, "Multitask learning for cross-domain image captioning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1047–1061, Apr. 2019.

[9] 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.

[10] 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. (CVPR)*, Jul. 2017, pp. 21–29.

[11] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2156–2164.

[12] 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. 1–8.

[13] J.-H. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

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

[15] 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. Discovery Data Mining*, Jul. 2018, pp. 1348–1357.

[16] 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. Manage.*, Oct. 2018, pp. 1013–1022.

[17] 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.

[18] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. 33nd Int. Conf. Mach. Learn.*, 2016, pp. 2397–2406.

[19] 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. (CVPR)*, Jun. 2016, pp. 30–38.

[20] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7736–7745.

[21] J. Song, P. Zeng, L. Gao, and H. T. Shen, "From pixels to objects: Cubic visual attention for visual question answering," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 906–912.

[22] B. Du *et al.*, "Deep irregular convolutional residual LSTM for urban traffic passenger flows prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 972–985, Mar. 2020.

[23] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7680–7688.

[24] 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. Multimedia Conf. MM*, 2018, pp. 519–527.

[25] Y. Zhu, J. J. Lim, and L. Fei-Fei, "Knowledge acquisition for visual question answering via iterative querying," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6146–6155.

[26] Q. Wu, C. Shen, A. van den Hengel, P. Wang, and A. R. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," 2016, *arXiv:1603.02814*. [Online]. Available: https://arxiv.org/abs/1603.02814

[27] Q. Wu, P. Wang, C. Shen, A. 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. (CVPR)*, Jun. 2016, pp. 4622–4630.

[28] 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.

[29] 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. (ICCV)*, Oct. 2017, pp. 2631–2639.

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

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[32] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training Gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2226–2234.

[33] 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.

[34] 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.

[35] J. Tang, J. Wang, Z. Li, J. Fu, and T. Mei, "Show, reward, and tell: Adversarial visual story generation," *Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, pp. 54:1–54:20, 2019.

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

[37] F. Huang, X. Zhang, and Z. Li, "Learning joint multimodal representation with adversarial attention networks," in *Proc. ACM Multimedia Conf. Multimedia Conf. MM*, 2018, pp. 1874–1882.

[38] C. Li *et al.*, "Alice: Towards understanding adversarial learning for joint distribution matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5495–5503.

[39] Q. Wu, P. Wang, C. Shen, I. Reid, and A. V. D. Hengel, "Are you talking to me? Reasoned visual dialog generation through adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6106–6115.

[40] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1548–1558.

[41] I. Ilievski and J. Feng, "Generative attention model with adversarial self-learning for visual question answering," in *Proc. Thematic Workshops ACM Multimedia Thematic Workshops*, 2017, pp. 415–423.

[42] Y. Zhang, S. Wang, B. Chen, J. Cao, and Z. Huang, "TrafficGAN: Network-scale deep traffic prediction with generative adversarial nets," *IEEE Trans. Intell. Transp. Syst.*, early access, Dec. 17, 2019, doi: 10.1109/TITS.2019.2955794.

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

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

[45] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[46] 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.*, Jun. 2016, pp. 4613–4621.

[47] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2623–2631.

[48] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[49] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[50] 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. (CVPR)*, Jul. 2017, pp. 6325–6334.

[51] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6087–6096.

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

[53] D. Teney, P. Anderson, X. He, and A. V. D. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4223–4232.

[54] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[55] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2048–2057.

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

His research interests include social media analysis, multi-modal data analysis, and data mining.

Xiaoming Zhang received the B.Sc. and M.Sc. degrees from the National University of Defense 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. 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*, *Signal Processing*, and international conferences such as ACM MM, AAAI, IJCAI, CIKM, ICMR, SDM, and EMNLP. His current research interests include social media analysis, image tagging, and text mining.

Feiran Huang (Member, IEEE) received the B.Sc. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2018.

He is currently with the College of Cyber Security/College of Information Science and Technology, Jinan University, where he has been a Lecturer since 2018. He has published over ten papers, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, *Knowledge-Based Systems*, and international conferences such as ACM MM, CIKM, and ICMR. His research interests include social media analysis, multi-modal 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, Pok Fu Lam, Hong Kong, in 2018.

Currently, he is 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 Defense 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 the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the 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.

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