Adversarial Learning of Answer-Related Representation for Visual Question Answering

Yun Liu
School of Computer Science and Engineering,
Beihang University, Beijing, China
gz_liuyun@buaa.edu.cn

Xiaoming Zhang*
School of Cyber Science and Technology,
Beihang University, Beijing, China
yolixs@buaa.edu.cn

Feiran Huang
School of Computer Science and Engineering,
Beihang University, Beijing, China
huangfr@buaa.edu.cn

Zhoujun Li
State Key Laboratory of Software Development
Environment, Beihang University, Beijing, China
lizj@buaa.edu.cn

ABSTRACT

Visual Question Answering (VQA) aims to learn a joint embedding of the question sentence and the corresponding image to infer the answer. Existing approaches learn the joint embedding don't consider the answer-related information, which results in that the learned representation is not effective to reflect the answer of the question. To address this problem, this paper proposes a novel method, i.e., Adversarial Learning of Answer-Related Representation (ALARR) for visual question answering, which seeks an effective answer-related representation for the question-image pair based on adversarial learning between two processes. The embedding learning process aims to generate modality-invariant joint representations for the question-image and question-answer pairs, respectively. Meanwhile, it tries to confuse the other process, embedding discriminator, which tries to discriminate the two representations from different modalities of pairs. Specifically, the joint embedding of the question-image pair is learned by a three-level attention model, and the joint representation of the question-answer pair is learned by a semantic integration model. Through the adversarial leaning, the answer-related representation are better preserved. Then an answer predictor is proposed to infer the answer from the answer-related representation. Experiments conducted on two widely used VQA benchmark datasets demonstrate that the proposed model outperforms the state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Question answering**; **Multimedia and multimodal retrieval**;

KEYWORDS

visual question answering, adversarial learning, representation

*Corresponding author: Xiaoming Zhang(yolixs@buaa.edu.cn).

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271765

ACM Reference Format:
Yun Liu, Xiaoming Zhang, Feiran Huang, and Zhoujun Li. 2018. Adversarial Learning of Answer-Related Representation for Visual Question Answering. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271765

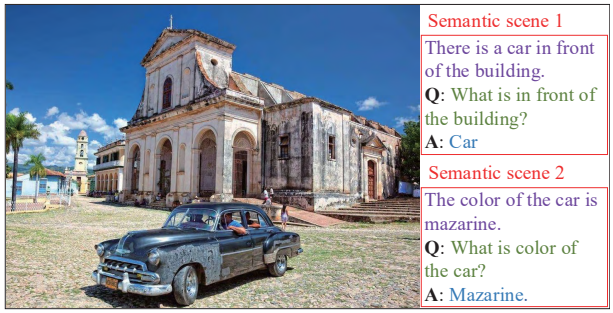

Figure 1: Two question-answer pairs and their corresponding semantic scenes in the image.

1 INTRODUCTION

Recently, multi-modal learning has gained much attention in artificial intelligence. A new task named Visual Question Answering (VQA) [15, 17] has attracted extensive research attention. Different from text-based QA system [3, 4] in natural language processing, VQA takes an image and a free-form, open-ended, natural-language question sentence about the image as the input. A natural-language answer is produced as the output [1]. Compared with other visual-language tasks such as image captioning [9] and text-to-image retrieval [27], VQA requires a better understanding on both the input question and image to infer the answer. VQA can significantly benefit a variety of applications, such as automatic customer service, early education, and so on.

Exiting works on VQA are mainly based on deep neural network [12, 16, 32], which commonly use Convolutional Neural Networks (CNN) to embed the image and word vectors along with Long Short-Term Memory (LSTM) to embed the question [12, 34]. They can be roughly divided into three categories. The first type of methods directly combine or concatenate the visual and textual embeddings

to derive the answer [16, 17]. The second type of methods aim to discover the important image regions for the corresponding question using the visual attention model [6, 13, 30]. The last type of methods [2, 7, 10] focus on multi-modal feature embedding and resort to a bilinear pooling for VQA. These methods make great efforts to learn a joint representation of the question-image pair based on VQA classifier, while which is not effective to encode the answer-related information in the learned representation.

Usually, each question-answer pair corresponds to a specific semantic scene of the image, and this specific semantic scene is related to the answer and also covered by the question-image pair. For example, for the image in Figure 1, there are two question-answer pairs, which correspond to two semantic scenes existing in the image. For the first question-answer pair, 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. As for the second question-answer pair, the question "What is color of the car?" and its corresponding answer "Mazarine" correspond to another semantic scene "The color of the car is mazarine" in the image. On the other side, given a question and the corresponding image, the semantic scene related to the answer can be also specified. Therefore, the target of VQA can be considered to learn the answer-related semantic information and encode it in the joint embedding of the question-image pair for answer inferring. Specifically, the semantic scene is unique for each question-answer pair and is important to specify the answer. Therefore, it is vitally important to utilize the answer-related semantic information in the question-answer pair to guide the learning of the joint embedding of the question-image pair, which can make the joint embedding more effective to reflect the answer.

Despite the answer-related semantic scene seems easy for a person to recognize from the question-image pair, it is still a great challenge for previous works to effectively exploit it for VQA. First, some existing works [2, 13, 32] mainly focus on exploiting the correlation between image and question to infer the answer. However, the question is not specified to the corresponding image, and it can be generated as a question for any images. For example, the question " What is in front of the building" can also be generated for other images, such as image about school or church. Therefore, exploiting the correlation between question and image directly may focus on some wrong regions that are unrelated to the answer, since it lacks the guidance of answer comprehension in the learning process. On the other side, some works [10, 14, 30] learn the joint embedding of question-image guided by the supervised VQA classifier, which mainly concentrates on the performance of classification rather than learn the answer-related semantic information in the image. Obviously, this strategy is not well consistent with the natural process of answer inferring. Therefore, an explicit mechanism to learn the answer-related semantic information should be included when learning the joint embeddings of question-image pairs.

To tackle these challenges, we propose to take advantage of the answer-related semantic scenes contained in the question-answer pairs and fuse this information to learn the joint embedding of the question-image pair. In particular, we investigate: (1) How to exploit the answer-related semantic scenes for joint embedding learning; (2) How to exploit the discriminative features from the learned representation to infer the answer. Our solutions to these questions result in a novel model, i.e., Adversarial Learning of Answer-Related Representation (ALARR), for VQA. The framework of ALARR is presented in Figure 2. Specifically, our framework mainly consists of four components. The two embedding components, i.e., question-image embedding and question-answer embedding, are proposed to generate the modality-invariant representations for items from question-image pairs and question-answer pairs in the common subspace. They have the objective to confuse the embedding classifier, i.e., the embedding discrimination, which acts as an adversary. The embedding discrimination tries to distinguish the representations learned from question-image pairs and question-answer pairs, which steers the learning of representations. Through bringing the embedding classifier in the adversary role, it is expected that the modality invariance of the learned representations is reached more efficiently. The question-image representation is optimal for answer inferring when the process converges, namely when the embedding discrimination fails to discriminate it from the representation learned from the question-answer pair. In this way, the representation learned from the question-image pair can more effectively reflect the answer-related semantic information in the question-answer pair. To effectively infer the answer, an answer prediction model is designed by dismantling question features from the answer-related representation to derive the answer. The main contributions are summarized as follows:

- Different from previous methods, we investigate the problem by exploiting the answer-related information for joint embedding learning of question and image. This process is more effective to encode the answer-related information in the learned representation and hence can improve the performance of question inferring.
- We propose a novel model ALARR which is built around the concept of adversarial learning for VQA. Two opposite processes are designed in our approach to learning a more effective joint representation for VQA.
- The proposed model ALARR is evaluated on two benchmark datasets. Experimental results show that ALARR significantly outperforms the state-of-the-art approaches.

The rest of the paper is structured as follows. In Section 2, we position our approach in the context of the related existing works. Then, the proposed method ALARR is detailed in Section 3. Next, we describe the experiments and analysis of the results in Section 4, and the paper is concluded in Section 5.

2 RELATED WORK

Most of the existing approaches on VQA mainly use deep neural networks to learn the visual-textual features for answer classification. Approaches presented in [12, 16, 17] embed the image and question sentence by CNN and LSTM, respectively. Then, the two embeddings are combined as the joint representation input to the VQA classifier to infer the answer. The work presented in [30] uses the multi-query method with the stacked attention model to find the fine-grained features in the image, which is the first work of using attention model in VQA task. After that, [13, 18] employ a multi-level attention model to discover the important regions of the image. They concatenate the weighted visual features with the textual question features to derive the answer. Dynamic memory

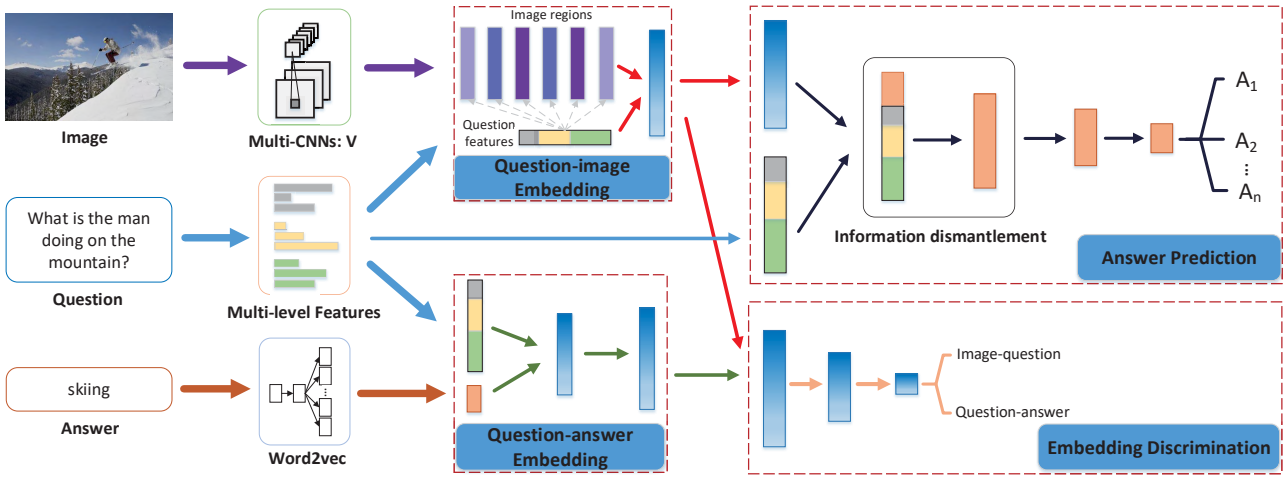

Figure 2: Framework of the proposed ALARR model for VQA, which mainly consists of four components, i.e., question-image embedding, question-answer embedding, embedding discrimination and answer prediction.

network [28] and dynamic parameter [19] are also proposed for VQA, the structure or weights of the models are determined adaptively based on the questions. Moreover, external knowledge bases [24, 35] used in VQA also achieves some improvement of the performance. In [25], a large-scale knowledge base DBpedia is used as the supplementary information which doesn't appear in the question images. Some other approaches focus on multi-modal feature embedding for VQA. MCB [7] and MLB [10] both design bilinear models with the multi-modal pooling method to learn multi-modal feature embedding. In MUTAN [2], a generalized multi-modal pooling framework is proposed, which shows that MCB and MLB are its special cases. However, most of the existing approaches mainly learn the multimodal representation from the question-image pairs directly. The representations learned by these approaches are not effective to reflect the answer-specific information in the image.

Our approach is motivated by the adversarial learning [8], which has enjoyed great success in computer vision [5, 20], information retrieval [22, 23], and sequence generation [11, 33]. The core of the adversarial learning framework is the interplay between two processes, a feature generator and a discriminatory classifier, conducted as a minimax game. For the feature generator, it has the objective to confuse discriminatory classifier that acts as an adversary. On the other side, discriminatory classifier tries to distinguish the items in terms of their features and in this way guides the learning of the feature generator [22]. [5] employs a generative adversarial network to maximize the mutual information between a small subset of the latent variables and the observation. [11] utilizes adversarial training for open-domain dialogue generation. The model is trained to produce sequences that are indistinguishable from human-generated dialogue utterances. An adversarial cross-Modal retrieval method is proposed in [22], which try to minimize the gap among the representations of all items from different modalities with same semantic labels. To the best of our knowledge, though the adversarial learning approach has been proved successful in many areas, it has not been effectively explored in VQA task.

3 VQA BUILT ON ADVERSARIAL LEARNING

3.1 Problem Statement

In this paper, we focus on learning answer-related representation for visual question answering. Given an image, one or more questions about the image can be generated, and the corresponding answer is inferred for each question. We denote each triple instance of image, question, and answer as $O = (\mathcal{V}, Q, \mathcal{A})$. $\mathcal{V} \in \mathbb{R}^{d_v \times m \times m}$ represent the image regions extracted by multi-CNNs, where $m \times m$ is the number of regions and d_v is the dimensionality of the feature vector of each region. Q denotes the question sentence. $\mathcal{A} \in \mathbb{R}^{d_w}$ is the answer corresponding to the question Q and the image $\mathcal{V}$, which is usually a single word.

Specific to ALARR proposed in this paper, in order to make the two modalities of the pairs, i.e., question-answer and question-image, directly comparable, we aim at finding a common subspace $\mathcal{S}$. Then features of the question-answer and question-image pairs can be embedded into $\mathcal{S}_{Q\mathcal{A}} = f_{Q\mathcal{A}}(Q, \mathcal{A}; \theta_{qa})$ and $\mathcal{S}_{\mathcal{V}Q} = f_{\mathcal{V}Q}(\mathcal{V}, Q; \theta_{vq})$. Here, $f_{Q\mathcal{A}}$, $f_{\mathcal{V}Q}$ are the mapping functions and $\mathcal{S}_{Q\mathcal{A}}$, $\mathcal{S}_{\mathcal{V}Q}$ is the transformed features in the common space $\mathcal{S}$, respectively. We expect the distributions of $\mathcal{S}_{Q\mathcal{A}}$ and $\mathcal{S}_{\mathcal{V}Q}$ to be modality-invariant and semantically discriminative, but also to better preserve the underlying semantic similarity structure in data. Then, the learned representation of the question-image pair is more similar to the representation learned for the corresponding question-answer pair, and thus it can be more effective to reflect the answer. In the following subsections, we will introduce how these requirements are met.

3.2 Question-Image Embedding

Given a question sentence and the associated image, the VQA task aims to automatically find the correlation between the question sentence and the image to infer the answer. To effectively exploit the correlation at different semantic level, a three-level of attention model on the question, i.e., word-level, phrase-level, and sentence-level, is proposed to learn a joint representation for the question

and image. Inspired by [14], our three-level attention model fuses the input features via a multiplicative embedding scheme to utilize the full information from the inputs. The network structure of the three-level attention model is shown in Figure 3.

At the word-level, we embed the words in the question sentence Q to a vector space using the end-to-end learning method to obtain $Q^w = \{q_1^w, q_2^w, ...q_t^w\} \in \mathbb{R}^{d_w \times t}$, where q_i^w represents the word vector of the i-th word in the question, d_w is the dimensionality of the word vector. At the phrase-level, 1-dimensional CNN are used to encode the information contained in unigram, bigram, and trigram of the word embedding vectors. Concretely, at each word location, the inner product of the word vectors is calculated with filters of the corresponding three window sizes. For the i-th word, the convolution output with window size s is calculated as follows:

$$\hat{q}_{s,i}^w = tanh(\mathcal{W}_c^s * q_{i:i+s-1}^w), \qquad s \in \{1, 2, 3\} \tag{1}$$

where $*$ represent the convolutional operation, and $\mathcal{W}_c^s$ is the convolutional weight parameters. The word-level features Q^w are appropriately 0-padded before feeding into bigram and trigram convolutions. Given the convolution results, the max-pooling is applied across all the$n$-grams to obtain phrase-level features:

$$q_i^p = max(\hat{q}_{1,i}^p, \hat{q}_{2,i}^p, \hat{q}_{3,i}^p), \qquad i \in \{1, 2, ..., t\} \tag{2}$$

Through this pooling method, our approach can adaptively select different gram features at each time step. It can extract the important features of the phrase-level automatically, while reserving the original sequence length and order. Then, we can obtain $Q^p = \{q_1^p, q_2^p, ..., q_t^p\} \in \mathbb{R}^{d_h \times t}$, where d_h is the size of the hidden layer in $\mathcal{W}^s$. At the sentence-level, LSTM is used to encode the sequence Q^p, and the LSTM hidden vector at time step i acts as the sentence-level features q_i^s. Similarly, we can obtain the sentence-level features $Q^s = \{q_1^s, q_2^s, ..., q_t^s\} \in \mathbb{R}^{d_l \times t}$, where d_l is the hidden size of LSTM. For making the size of the features in different levels less affected by the length of the question, we employ max-pooling across different time steps to obtain multi-level features as follow:

$$q^k = max(q_1^k, q_2^k, ..., q_t^k), \qquad k \in \{w, p, s\} \tag{3}$$

After that, we obtain three types of features correspond to the three levels, i.e., $q^w \in \mathbb{R}^{d_w}$, $q^p \in \mathbb{R}^{d_h}$, and $q^s \in \mathbb{R}^{d_l}$. Figure 3 (a) shows the encoding method of the multi-level question features.

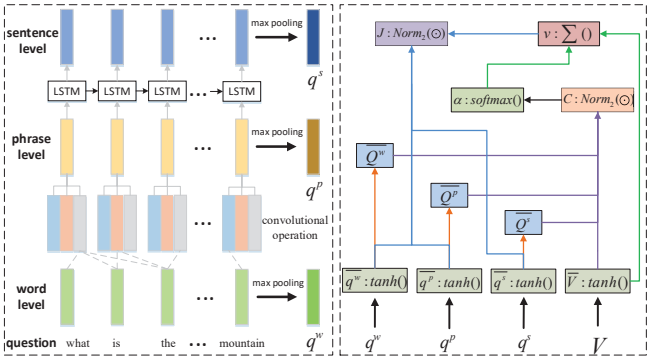

Figure 3: (a) Encoding method of three-level question features; (b) The framework of the three-level attention model.

To fuse the visual features of image regions $\mathcal{V}$ with the word, phrase, and sentence level features q^w, q^p, and q^s, we first embed them into the $\mathcal{S}$-dimensional common space as follows:

$$\overline{q}^w = tanh(\mathcal{W}_w q^w + b_w) \tag{4}$$

$$\overline{q}^p = tanh(\mathcal{W}_p q^p + b_p) \tag{5}$$

$$\overline{q}^s = tanh(\mathcal{W}_s q^s + b_s) \tag{6}$$

$$\overline{\mathcal{V}} = tanh(\mathcal{W}_v \mathcal{V} + b_v) \tag{7}$$

where $\mathcal{W}_v \in \mathbb{R}^{\mathcal{S} \times d_v}$, $\mathcal{W}_w \in \mathbb{R}^{\mathcal{S} \times d_w}$, $\mathcal{W}_p \in \mathbb{R}^{\mathcal{S} \times d_h}$, $\mathcal{W}_s \in \mathbb{R}^{\mathcal{S} \times d_l}$ are the learnable weight parameters, and b_v, b_w, b_p, $b_s \in \mathbb{R}^{\mathcal{S}}$ are the bias parameters. After mapping the multi-level features into the $\mathcal{S}$-dimensional common space, $\overline{q}^w$, $\overline{q}^p$ and $\overline{q}^s$ are spatially replicated $m \times m$ times to form $\overline{Q}^w$, $\overline{Q}^p$ and $\overline{Q}^s$, respectively, which matches the spatial size of the entire image $\overline{\mathcal{V}} \in \mathbb{R}^{\mathcal{S} \times m \times m}$.

The joint representation C of the four inputs is obtained by element-wise multiplication of $\overline{\mathcal{V}}$, $\overline{Q}^w$, $\overline{Q}^p$ and $\overline{Q}^s$. A L_2 normalization is used to constrain the magnitude of the representation:

$$C = Norm_2(\overline{\mathcal{V}} \odot \overline{Q}^w \odot \overline{Q}^p \odot \overline{Q}^s) \tag{8}$$

where $C \in \mathbb{R}^{\mathcal{S} \times m \times m}$ and $\odot$ denotes element-wise multiplication. The attention map is then obtained by convolving the joint representation C with 1×1 filters followed by a $softmax$ operation over the $m \times m$ grids as follow:

$$\alpha = softmax(\mathcal{W}_c * C + b_c) \tag{9}$$

where $*$ represents the convolutional operation, $\mathcal{W}_c \in \mathbb{R}^{\mathcal{S} \times 1 \times 1}$ and $b_c \in \mathbb{R}^{\mathcal{S}}$ are the learnable convolution kernel parameters. Then, a weighted sum of the image features over all spatial locations can be calculated as follow:

$$v = \sum_j^{m \times m} \alpha_j \overline{\mathcal{V}}_j \tag{10}$$

Compared with the originally independent visual features shared by the questions, the visual feature mapping v is more effective to reflect the image regions related to the current question Q. After that, the question and image are jointly embedded by element-wise multiplication of the transformed multi-level features and the attended visual features as follow:

$$\mathcal{S}_{\mathcal{V}Q}(\theta_{vq}) = Norm_2(v \odot \overline{q}^w \odot \overline{q}^p \odot \overline{q}^s) \tag{11}$$

where θ_{vq} denotes the whole parameters of the three-level attention model. The framework of the three-level attention model is shown in Figure 3 (b).

3.3 Question-Answer Embedding

In reality, each question-answer pair points to the specific semantic scene related to the answer in the image, which is also covered by the question-image pair. Obviously, the answer-related semantic information resides in the question-answer pair is helpful for better learning of the joint embedding of the question-image pair. Given a triple instance of a question, image, and answer, we expect the representations learned from the question-image pair and question-answer pair should be similar in the common subspace. That is, the learned representations should be modality-invariant. In this paper, a question-answer embedding model is proposed to learn

the representation of question-answer pair through multiplying the multi-level question features by the corresponding answer features.

Given a question $\mathcal{Q}$ and its corresponding answer $\mathcal{A} \in \mathbb{R}^{d_w}$, we first embed $\mathcal{A}$ into an $\mathcal{S}$-dimensional common space via the following equation:

$$\overline{\mathcal{A}} = tanh(\mathcal{W}_a\mathcal{A} + b_a) \tag{12}$$

where $\mathcal{W}_a \in \mathbb{R}^{\mathcal{S}\times d_w}$, $b_a \in \mathbb{R}^{\mathcal{S}}$ are the learnable weight and bias parameters, respectively. For the question $\mathcal{Q}$, we leverage the $\mathcal{S}$-dimensional word-level features $\overline{q}^w$, phrase-level feature $\overline{q}^p$, and sentence-level features $\overline{q}^s$ extracted in Eq. (1-6) as the question features similarly. After that, the semantic information of the question-answer pair is encoded by element-wise multiplication of the answer features with the multi-level question features, using the constraint of L_2 normalization as following:

$$\mathcal{S}_{\mathcal{QA}}(\theta_{qa}) = Norm_2(\overline{\mathcal{A}} \odot \overline{q}^w \odot \overline{q}^p \odot \overline{q}^s) \tag{13}$$

where θ_{qa} is the parameters in the question-answer embedding model. $\mathcal{S}_{\mathcal{QA}} \in \mathbb{R}^S$ represents the representation learned from the question-answer pair. It can be considered that $\mathcal{S}_{\mathcal{QA}}$ is specific for the answer, since it is learned from the answer directly. Therefore, it can be used to guide the representation learning for the question-image pair.

3.4 Embedding Discrimination

We expect the joint representations learned from the question-image and question-answer pairs are consistent in the common subspace. That is, the two modalities of representations $\mathcal{S}_{\mathcal{QA}}$ and $\mathcal{S}_{\mathcal{VQ}}$ learned from a triple of instances should both contain the answer and be similar as more as possible in the learned space. In this way, the joint embedding $\mathcal{S}_{\mathcal{QA}}$ of the question-image pair is learned to be more effective for answer inferring. On the other side, the embedding discrimination component aims to distinguish the representation as reliably as possible given an unknown embedding.

To distinguish an item of representation, an embedding classifier is defined to act as the "discriminator" in GAN (Generative Adversarial Networks) [8]. To construct a classifier, we use a 2-layer feed-forward Full Connected (FC) layers activated by $softmax$ in the last layer to obtain the probability distribution over classes as following:

$$\mathcal{D}(x) = softmax(\mathcal{W}_{d2}(\delta(\mathcal{W}_{d1}x + b_{d1})) + b_{d2}) \tag{14}$$

where $\mathcal{W}_{d1}, \mathcal{W}_{d2}$ are the weight parameters, and b_{d1}, b_{d2} are the bias parameters of the FC layers. δ is the activation function and x is either $\mathcal{S}_{\mathcal{VQ}}$ or $\mathcal{S}_{\mathcal{QA}}$.

To train the classifier, the joint representation $\mathcal{S}_{\mathcal{VQ}}$ of the question-image pair is assigned the label $\overline{01}$, while the joint representation $\mathcal{S}_{\mathcal{QA}}$ corresponding to the question-answer pair is assigned the label $\overline{10}$. For a pair of $\mathcal{S}_{\mathcal{VQ}}$ and $\mathcal{S}_{\mathcal{QA}}$, the classification loss is considered in the training process as to minimize the adversarial loss, which can be defined as:

$$\mathcal{L}_{adv}(\theta_{dis}) = m \cdot (log\mathcal{D}(\mathcal{S}_{\mathcal{QA}}) + log(1 - \mathcal{D}(\mathcal{S}_{\mathcal{VQ}}))) \tag{15}$$

where θ_{dis} is the parameters in the embedding discrimination. Essentially, $\mathcal{L}_{adv}$ denotes the cross-entropy loss of answer modality classification of all instances. Furthermore, m is the ground-truth representation modality label of the given instance, expressed as a one-hot vector, while $\mathcal{D}(\cdot)$ is the generated representation modality probability per item.

Inversely, in the adversarial learning process, the embedding learning processes of $\mathcal{S}_{\mathcal{VQ}}$ and $\mathcal{S}_{\mathcal{QA}}$ try to confuse the embedding classifier. That is, the three-level attention model want the classifier to misclassify $\mathcal{S}_{\mathcal{VQ}}$ as $\mathcal{S}_{\mathcal{QA}}$, while the question-answer integration model try to make the classifier to misclassify $\mathcal{S}_{\mathcal{QA}}$ as $\mathcal{S}_{\mathcal{VQ}}$. Therefore, they aim to minimize the representation modality learning loss:

$$\mathcal{L}_{rep}(\theta_{vq}, \theta_{qa}) = m \cdot (log(1 - \mathcal{D}(\mathcal{S}_{\mathcal{QA}})) + log(\mathcal{D}(\mathcal{S}_{\mathcal{VQ}}))) \tag{16}$$

Similar to the adversarial loss in Eq. 15, $\mathcal{L}_{rep}$ denotes the cross-entropy of the representation learning for $\mathcal{S}_{\mathcal{VQ}}$ and $\mathcal{S}_{\mathcal{QA}}$.

3.5 Answer Prediction

Similar to existing VQA works [2, 19, 32], we consider VQA as a multi-class classification problem. As shown in Figure 2, the joint embedding $\mathcal{S}_{\mathcal{VQ}}$ of the question-image pair is input to two components, i.e, the embedding discrimination to improve the effectiveness of the joint representation, and answer prediction to infer the answer. Through the adversarial learning, $\mathcal{S}_{\mathcal{VQ}}$ is more similar to $\mathcal{S}_{\mathcal{QA}}$ and hence is more effective to reflect the answer. Then, the discriminative features are dismantled from the joint representation $\mathcal{S}_{\mathcal{VQ}}$ for answer inferring.

In the question-answer embedding model, the joint representation $\mathcal{S}_{\mathcal{QA}}$ is obtained by element-wise multiplication of the answer features with the three-level question features. Inversely, the answer inferring from the joint embedding of the question-image pair is performed by element-wise division between $\mathcal{S}_{\mathcal{VQ}}$ and the three-level question features, i.e., word-level features $\overline{q^w}$, phrase-level features $\overline{q^p}$, and sentence-level features $\overline{q^s}$. Meanwhile, the L_2 normalization is used to constrain the magnitude of the representation as follow:

$$\widehat{\mathcal{A}} = Norm_2(\mathcal{S}_{\mathcal{VQ}} \oslash (\overline{q}^w \odot \overline{q}^p \odot \overline{q}^s)) \tag{17}$$

where $\oslash$ represents the element-wise division operation and $\odot$ denotes the element-wise multiplication operation, respectively. $\widehat{\mathcal{A}} \in \mathbb{R}^{\mathcal{S}}$ is regarded as the answer-specific features derived from the image. For the answer prediction classifier, we use a 2-layer feed-forward FC layers and $softmax(\cdot)$ activation function to calculate the probability of each answer item as follow:

$$\mathcal{P} = softmax(\mathcal{W}_{p2}(\delta(\mathcal{W}_{p1}\widehat{\mathcal{A}} + b_{p1})) + b_{p2}) \tag{18}$$

where $\mathcal{P} \in \mathbb{R}^k$, and k is the number of the predefined candidate answers. $\mathcal{W}_{p1}$ and $\mathcal{W}_{p2}$ are the weight parameters, and b_{p1} and b_{p2} are the bias parameters of the FC layers. δ is the activation function. To train the classifier, we also employ the cross-entropy to define the answer classification loss $\mathcal{L}_{ans}$ as following:

$$\mathcal{L}_{ans}(\theta_{vq}, \theta_{pre}) = -\frac{1}{k}\sum_{i=1}^{k}[y_i \cdot log\mathcal{P}_i + (1 - y_i) \cdot log(1 - \mathcal{P}_i)] \tag{19}$$

θ_{pre} is the parameters of the classifier, and y_i is the ground-truth answer label of the given question-image instance, which is either 1 or 0.

3.6 Adversarial Learning: Optimization

In the embedding process, we introduce the $\mathcal{L}_2$ normalization term to prevent overfitting of the learned parameters as following:

$$\mathcal{L}_{nor} = \sum_{l=1}^{L}(||\mathcal{W}_{vq}^{l}||^2 + ||\mathcal{W}_{qa}^{l}||^2) \tag{20}$$

where $\mathcal{W}_{vq}^{l}$ and $\mathcal{W}_{qa}^{l}$ denote the layer-wise parameters in the embedding of both the question-image and question-answer pairs respectively. Based on the above analysis, the loss function of the whole process of answer inferring is referred to as embedding loss, which is formulated as the combination of the representation modality learning loss (Eq. 16), the answer classification loss (Eq. 19), and the normalization item as following:

$$\mathcal{L}_{emb}(\theta_{vq}, \theta_{qa}, \theta_{pre}) = \alpha \cdot \mathcal{L}_{rep} + \beta \cdot \mathcal{L}_{ans} + \mathcal{L}_{nor} \tag{21}$$

where the hyper-parameters α and β control the contributions of the two terms.

Then, the whole process of learning the optimal answer-related representation to infer answer is conducted by jointly minimizing the adversarial loss and embedding loss, as obtained in Eq. 15 and Eq. 21, respectively. Since the optimization goals of the two objective functions are opposite in adversarial learning, the process runs as a **minimax game** [8] of the two concurrent sub-processes:

$$(\widehat{\theta}_{vq}, \widehat{\theta}_{qa}, \widehat{\theta}_{pre}) = \underset{\theta_{vq}, \theta_{qa}, \theta_{pre}}{\arg\min} (\mathcal{L}_{emb}(\theta_{vq}, \theta_{qa}, \theta_{pre}) - \mathcal{L}_{adv}(\widehat{\theta}_{dis})) \tag{22}$$

$$\widehat{\theta}_{dis} = \underset{\theta_{dis}}{\arg\max}(\mathcal{L}_{emb}(\widehat{\theta}_{vq}, \widehat{\theta}_{qa}, \widehat{\theta}_{pre}) - \mathcal{L}_{adv}(\theta_{dis})) \tag{23}$$

Usually, the stochastic gradient optimization algorithms can be used to implement this minimax game. Specifically, the minimax optimization can be performed efficiently by incorporating Gradient Reversal Layer (GRL) [31]. This process is transparent when forward-propagating, but multiplying its value by -1 when back-propagating. If the Gradient Reversal layer is added before the first layer of the answer modality classifier, the minimax optimization can be performed simultaneously, as shown in the Algorithm 1.

Algorithm 1 Pseudocode of optimizing our ALARR Method. g-steps applied to the semantic scene embedding, is a hyperparameter.

Initialization: Image features for current batch $\mathbb{V} = \{\mathcal{V}_1, ..., \mathcal{V}_n\}$
Question features for current batch $\mathbb{Q} = \{\mathcal{Q}_1, ..., \mathcal{Q}_n\}$
Answer features for current batch $\mathbb{A} = \{\mathcal{A}_1, ..., \mathcal{A}_n\}$
Corresponding labels for current batch $\mathbb{Y} = \{y_1, ..., y_n\}$;
Number of the answer terms: k;
Hyperparameters: $\mathcal{S}$, α, β.
training process:

```
for number of training iterations do
    for g-steps do
        update parameters θ_vq, θ_qa and θ_pre by descending
        their stochastic gradients:
        θ_vq ⟵ θ_vq − μ · ∇_{θ_vq}(L_emb − L_adv)
        θ_qa ⟵ θ_qa − μ · ∇_{θ_qa}(L_emb − L_adv)
        θ_pre ⟵ θ_pre − μ · ∇_{θ_pre}(L_emb − L_adv)
    end for
    update parameters of semantic scene classifier θ_dis by
    ascending its stochastic gradients through Gradient
    Reversal Layer
    θ_dis ⟵ θ_dis + μ · ∇_{θ_dis}(L_emb − L_adv)
    return currently learned parameters in the ALARR.
end for
```

4 EXPERIMENTS

In this section, we first conduct a set of experiments to analyze the effectiveness of ALARR by comparing its performance with multiple state-of-the-art approaches on two datasets. Then we conduct additional evaluations to investigate the performance of ALARR in more detail.

4.1 Datasets and Baselines

We evaluate our model ALARR on two public datasets named **VQA** and **COCO-QA**, and they are detailed as follows:

VQA [1] is created from the MS-COCO image caption dataset and labeled by human annotators as part of the visual question answering challenge. It has become the most widely used dataset in VQA task. There are three types of questions, including *yes/no*, *number*, and *other*. For each question, ten labeled free-response answers are provided. Answers are typically a word or a short phrase. Approximately 40% of the questions have a *yes/no* answer. Totally, there are 248,349 training questions, 121512 validation questions, and 244302 testing question, generated on a total of 123287 images. The top 1000 most frequent answers are used as the test set which covers 82.7% of the total answers.

COCO-QA [17] is another dataset based on MS-COCO dataset. The caption of the image is first parsed with an off-the-shelf parser, then the key components in the caption are replaced with the question words to form questions. Both of the question and answer are generated automatically from image captions in MS-COCO. It contains four categories, i.e., *Object*, *Number*, *Color* and *Location*. These categories possess 70%, 7%, 17%, and 6% of the whole dataset respectively. There are 78,736 training samples and 38,948 test samples in the dataset, and their questions are generated from a total of 123,287 images. Additionally, each image corresponds to only one question and all answers are single-word.

The following baselines are used in the experiments:

- **VSE** [17]: Image features are input to the head and the end of the question sentence respectively. A bidirectional LSTM is used to encode the compositive features.
- **Img-CNN** [12]: Three individual end-to-end CNNs are used to encode image, represent the question and learn their joint representation for answer classification, respectively.
- **DPPnet** [19]: A model learning a convolutional neural network with a dynamic parameter layer whose weights are determined adaptively based on questions.
- **SAN** [30]: A multiple-layer stacked attention neural networks model, which query an image multiple times to infer the answer progressively.
- **DMN** [28]: A dual attention networks which jointly leverage visual and textual attention mechanisms to capture fine-grained interplay between vision and language.

- **HieCoAtt** [13]: A hierarchical architecture that co-attends to the question-image pair at three levels, i.e., word-level, phrase-level, and question-level.
- **MCB** [7]: A model utilizing Multimodal Compact Bilinear pooling to combine the multi-modal features.
- **MLB** [10]: A low-rank bilinear pooling using Hadamard product to obtain an attention mechanism for multimodal learning of VQA.
- **Dual-MFA** [14]: A framework which fuses features from free-from image regions, detection boxes, and question representations via a multi-modal feature embedding scheme.
- **MUTAN** [2]: A multimodal tensor-based Tucker decomposition to parametrize bilinear interactions between visual and textual representations.

These approaches briefly introduced above cover the main methods of the three categories of the existing works on VQA, i.e., feature combination methods, attention-based methods, and multi-modal feature embedding methods. Additionally, another representative method named dynamic parameters DPPnet [19], also used as a baseline in the experiments.

4.2 Experimental Preparation

For the question sentence and corresponding answer, each word is embedded into a 200-dimensional GloVe feature vector. For the image, we employ VGG19 [21] network pretrained on ImageNet 2012 classification challenge dataset to extract visual features. The output of the convolutional layer "conv5_4" is used as the region features, of which the dimension is $512 \times 14 \times 14$. This means that each image has a total of 196 candidate regions. In the three-level attention networks, the hidden layer size d_h of the convolutional operation and the output size d_l of LSTM are set to be 200 and 300, respectively. Additionally, the dimension of the common space $\mathcal{S}$ is set to be 500. All of the activation functions are set with $relu$ except special instructions. Moreover, in order to prevent overfitting, dropout technique is applied after each linear transformation in the training process. For the dataset VQA, since it is larger than other datasets, we double the sizes of these dimension parameters in the three-level attention networks and the answer predictor. In the training process, the batch size is fixed to 128, g-steps is empirically set to be 5, and the adaptive moment estimation (Adam) with a momentum of 0.99 and a weight decay of 10^{-8} are used. The best learning rate and the model parameters α and β are picked using grid search.

For the two datasets VQA and COCO-QA, we formulate the VQA task as a classification problem similar to the existing works, due to most of the answers only contain single-word. Following [1, 15, 17], we evaluate all the methods using the metric of classification accuracy. Additionally, WU-Palmer Similarity (WUPS) reported in [26] is also used to measure the performance. WUPS calculates the similarity between two words based on their common subsequence in a taxonomy tree. A threshold can be set for WUPS, such as if the similarity is less than the threshold, it is zeroed out. Following [12, 30], WUPS@0.9 and WUPS@0.0 are used as the metrics besides the classification accuracy. Since the dataset VQA has ten answer labels for each question, its evaluation is different from the dataset of COCO-QA, we follow [1] to define the metric: $accuracy = min($# *humans that provided that answer*$/3, 1)$, namely an answer is deemed 100% accurate if at least 3 workers provided exact answer.

4.3 Results and Analysis

The performances comparison on the two datasets VQA and COCO-QA are illustrated in Table 1 and Table 3, where "-" represents the data is unavailable. Table 1 shows the results of the dataset VQA for both of the tasks Open-Ended and Multi-Choice (MC). The approaches are trained on the training set split from the dataset and evaluated on the test set, where the data set *test-dev* is normally used for validation and the data set *test-std* is used for standard testing. From table 1, one can see that our model ALARR improves the performance of the best baseline MUTAN from 67.42% to 68.61% on the open-ended task, and from 70.04% to 71.57% on the MC task of the *test-dev* set. Specifically, for the question types of *Number* and *Other*, ALARR obtains improvements of 1.31% and 1.12% compared with MUTAN. Similar improvement on MUTAN can also be observed in the data set *test-std*. Additionally, comparing with other state-of-the-art approaches, ALARR significantly outperforms the models of LSTM+Q+I, DPPnet, SAN, HieCoAtt, MLB, and Dual-MFA by 14.47%, 11.07%, 9.53%, 6.33%, 1.54%, and 1.34% absolutely in the metric of accuracy on the data set *test-dev*.

There are several reasons for the improvement. First, the embedding of question-image in ALARR fully exploits the answer-related information by using the question-answer pair to help the embedding learning. Therefore, the learned representation for the question-image pair is more specific to reflect the answer. Second, the adversarial learning used in ALARR can make that the embedding of question-image pair is more consistent with the question-answer pair to cover the answer, and in turn improves the embedding of question-answer pair. On the other side, VSE simply concatenates the features of image and question to infer the answer, and DPPnet resorts dynamic parameters determined adaptively based on the questions. The two methods can't learn the fine-granularity correlations between image and question, which results in that the performance is affected by the noisy regions and words. SAN, DMN, HieCoatt, and Dual-MFA employ stacked, hierarchical or dual attention network to learn a joint embedding from the question, free-form image regions or detected boxes for answer inferring. However, the joint embedding of these methods is mainly learned for answer classification, in which the learning process doesn't consider the answer-related information. MCB, MLB, and MUTAN focus on multi-modal feature embedding and achieve certain success. However, these methods mainly consider the question-image pair in the learning process, which results in that the learned representation may be less specific to the answer. The improvement against the state-of-the-art approaches proves that adversarial learning is effective to learn a more reasonable representation for VQA.

Moreover, Table 3 compares our method with the state-of-the-art approaches on the dataset COCO-QA. One can observe that our model ALARR also improves the best baseline Dual-MFA from 66.62% to 67.84% in accuracy. Additionally, ALARR significantly outperforms Img-CNN, DPPnet, SAN and HieCoAtt by 9.44%, 6.65%, 6.24% and 2.44% in accuracy, respectively. Among these methods, Img-CNN leverage three individual end-to-end CNNs to encode

Methods	Test-dev					Test-std				
	Open-Ended				MC	Open-Ended				MC
	All	Yes/No	Number	Other	All	All	Yes/No	Number	Other	All
VSE:										
-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:	57.22	80.71	37.24	41.69	62.48	57.36	80.28	36.92	42.24	62.69
SAN:	58.7	79.3	36.6	46.1	-	58.9	79.11	36.41	46.42	-
DMN:	60.3	80.5	36.8	48.3	-	60.36	80.43	36.82	48.3	-
HieCoAtt:	61.8	79.7	38.7	51.7	65.8	62.1	79.95	38.22	51.95	66.07
MCB:	66.7	83.4	39.8	58.5	69.10	66.5	-	-	-	-
MLB:	66.77	84.57	39.21	57.81	-	66.89	**84.02**	37.9	54.77	68.89
Dual-MFA:	67.11	84.63	39.48	58.14	70.04	67.09	83.37	40.39	56.89	69.97
MUTAN:	67.42	**85.14**	39.81	58.52	-	67.36	-	-	-	-
ALARR:	**68.61**	85.08	**41.12**	**59.64**	**71.57**	**68.43**	83.97	**42.06**	**57.89**	**71.28**

Table 1: Performance comparison of the dataset VQA on official server

the image, question, and their joint representation. This method also neglects the fine-granularity correlations between image and question, and it also doesn't consider the answer-related information in the learning process. Comparing with the two models of VSE, i.e., VIS+LSTM and 2VIS+BLSTM, ALARR improves the accuracy by 14.60% and 12.99%. Similar improvements are observed in the metrics of WUPS@0.9 and WUPS@0.0. In order to study the effectiveness of ALARR in detail, we report the performance of different types of questions on the dataset COCO-QA in Table 3. Compared to the best baseline Dual-MFA, our model ALARR improves the accuracy by 1.11% for question type of *Object*, 1.22% for the question type *Number*, 1.57% for the question type *Color*, and 1.05% for the question type *Location*, respectively. These prominent improvements are similar to results on the dataset VQA for the four types of the questions. Table 3 shows that our proposed model ALARR achieves significant improvement in VQA task, which again proves that the joint representation learned with the adversarial network is more effective for VQA.

Method	Validation
TLA-Mul*	59.45
HieCoAtt	58.33
SenAtt	56.27
EW-MUL*	59.12
EW-ADD	57.73
Fea-CAT	58.05
Mul-Norm*	59.32
Mul-noNorm	58.79
TLA-Full	59.75
ALARR*	**61.81**

Table 2: Ablation study on VQA dataset, where "*" denotes variant implementations of our model

4.4 Further Analysis

4.4.1 Ablation study. We further conduct ablation experiments to analyze the effectiveness of some individual components designed in our method. Table 2 shows the results of some variant implementations of our model. These approaches are trained on the training set and tested on the validation set of the dataset of VQA. Following other compared approaches, the test set is not used in the study due to the restrictions of the online submission.

The first part of Table 2 shows the performance of different attention models used for learning the joint embedding of image and question. Our three-level attention model implemented by multiplication in a common space (denoted as TLA-Mul) outperforms the hierarchical co-attend architecture [13] and the sentence-level attention on image (denoted as SenAtt) by 1.12% and 3.18% respectively. This demonstrates that the multi-level features of the questions-attending image by multiplication in a common space is effective for VQA. The second part shows that the element-wise multiplications (denoted as EW-MUL)in Eq. 8, Eq. 11, and Eq. 13 work better than element-wise addition (denoted as EW-ADD) and feature concatenation (denoted as Fea-CAT). The third part in Table 2 shows that L_2 normalization used in the multiplication equations (denoted as Mul-Norm) perform better than the method without normalization restriction (denoted as Mul-noNorm). Additionally, the last part of Table 2 compare ALARR with the full implementation of the three-level attention model (denoted as TLA-Full) without adversarial learning. Results show that ALARR significantly outperforms TLA-Full by 2.06%. It demonstrates that, with the answer-related information in the question-answer pair, the embedding learning of question-image pair is more effective for VQA.

Methods	Accuracy	WUPS@0.9	WUPS@0.0	All	Obj.	Num.	Col.	Loc.
VSE:								
-VIS+LSTM	53.24	63.84	87.95	-	56.5	46.1	45.9	45.5
-2VIS+BLSTM	54.85	64.78	88.12	-	58.2	44.8	49.5	47.3
Img-CNN:	58.40	68.50	89.67	58.4	-	-	-	-
DPPnet:	61.19	70.84	90.61	61.16	-	-	-	-
SAN:	61.6	71.6	90.9	61.6	65.4	48.6	57.9	54.0
HieCoAtt:	65.4	75.1	92.0	65.4	68.0	51.0	62.9	58.8
Dual-MFA:	66.62	76.15	92.29	66.49	68.86	51.32	**65.89**	58.92
ALARR:	**67.84**	**77.43**	**92.83**	**67.56**	**69.67**	**52.14**	65.62	**60.07**

Table 3: Performances on COCO-QA dataset

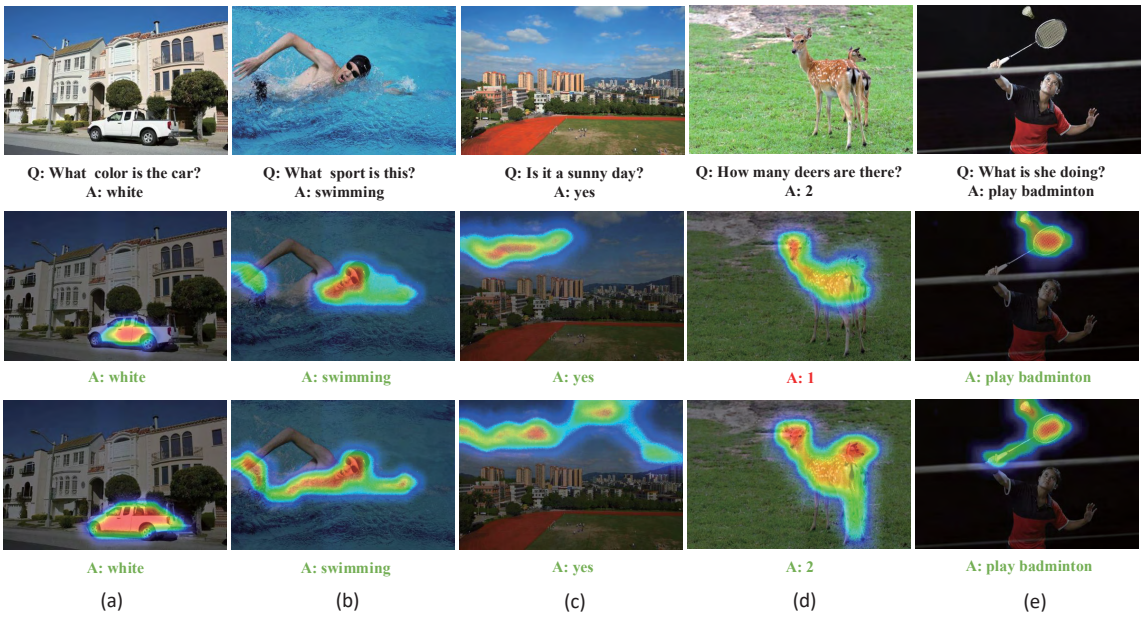

Figure 4: Examples of attention visualization. The first row presents the input images, the second row presents the three-level attention maps, and the third row presents the three-level attention maps guided by adversarial learning of answer-related representation. The green color indicates that the answer is correct, and the red color indicates that the answer is wrong.

4.4.2 Visualization of Attention model. To analyze the effectiveness of the attention model in ALARR, we visualize some attention maps generated by the three-level attention model (TLA) and ALARR which has an adversarial learning network on TLA. Following [29], we employ upsampling and Gaussion filtering to visualize the attention weights of the two models TLA and ALARR. Examples chosen from the widely used dataset VQA are used for illustration. Given a question, we show the visual attention over the image regions. Figure 4 presents five examples of attention visualization for TLA and ALARR, respectively. From the results, it can be seen that both TLA and ALARR are effective to learn the attention. Specifically, ALARR can learn more accurate attention than TLA, and thus it can produce a more accurate or more confident answer. For instance, Figure 4(a), Figure 4(b), Figure 4(c) and Figure 4(e) show that ALARR pays more reasonable attention on the regions corresponding to the question words than TLA, which makes ALARR more effective to infer the answer. In Figure 4(d), TLA can't identify the little deer behind the larger deer and a wrong answer is inferred, while ALARR effectively attends the head of the little deer with the right regions and obtains a correct answer. The results indicate that the adversarial learning network used in ALARR is helpful to the attention model to learn more specific information of the answer, and hence it helps the model to infer answer more effectively.

5 CONCLUSION

In this paper, we explore to learn a more effective embedding of the question-image pair by exploiting the answer-related information

in the question-answer pair. Specifically, a novel model, i.e., Adversarial Learning of Answer-Related Representation (ALARR), is proposed for VQA. The joint embedding of the question-image pair is learned through a three-level attention network, and the embedding of the question-answer pair is learned by a question-answer embedding model. To more effectively reflect the answer-related information for the question-image embedding, an adversarial network is proposed to make the representations learned for the question-image and question-answer pairs are more consistent. Then, an answer prediction model is proposed to infer the answer from the answer-related embedding. Experimental results demonstrate that the proposed model significantly outperforms state-of-the-art approaches on two widely used VQA datasets. Different from existing approaches that learn the embedding from the question-image pairs directly, our approach exploits the answer-related information in the question-answer pair to learn a more effective embedding and make more effective answer inferring. It certifies that the answer-related representation learning via adversarial network proposed in this paper is beneficial for VQA.

6 ACKNOWLEDGEMENT

This work was supported by Beijing Natural Science Foundation of China (No. 4182037), the National Natural Science Foundation of China (Nos. U1636210, U1636211, 61370126, 61672081, and 61602237), and in part by the Fund of the State Key Labo-ratory of Software Development Environment (No. SKLSDE-2017ZX-19).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of IEEE International Conference on Computer Vision*. 2425–2433.

[2] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *Proceedings of IEEE International Conference on Computer Vision*. 2631–2639.

[3] Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question Answering with Subgraph Embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 615–620.

[4] Antoine Bordes, Jason Weston, and Nicolas Usunier. [n. d.]. Open Question Answering with Weakly Supervised Embedding Models. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 165–180.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of Advances In Neural Information Processing Systems*. 2172–2180.

[6] eng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. 2017. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3909–3918.

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 457–468.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*. 2672–2680.

[9] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[10] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *Proceedings of 5th International Conference on Learning Representations*.

[11] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2157–2169.

[12] Hang Li Lin Ma, Zhengdong Lu. 2016. Learning to Answer Questions from Image Using Convolutional Neural Network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 3567–3573.

[13] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *proceedings of Advances in Neural Information Processing Systems 29*. 289–297.

[14] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. 2018. Co-Attending Free-Form Regions and Detections With Multi-Modal Multiplicative Feature Embedding for Visual Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

[15] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Proceedings of the Neural Information Processing Systems Conference*. 1682–1690.

[16] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images. In *IEEE International Conference on Computer Vision*. 1–9.

[17] Richard S. Zemel Mengye Ren, Ryan Kiros. 2015. Exploring Models and Data for Image Question Answering. In *Proceedings of the Neural Information Processing Systems Conference*. 2953–2961.

[18] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2156–2164.

[19] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 30–38.

[20] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Proceedings of Advances In Neural Information Processing Systems*. 2226–2234.

[21] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*. arXiv:1409.1556.

[22] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial Cross-Modal Retrieval. In *Proceedings of the ACM on Multimedia Conference*. 154–162.

[23] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models. In *Proceedings of the 40th International ACM SIGIR*. 515–524.

[24] Senzhang Wang, Xia Hu, Philip S Yu, and Zhoujun Li. 2014. MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD*. 1246–1255.

[25] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2016. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.

[26] Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 133–138.

[27] Liang Xie, Jialie Shen, and Lei Zhu. 2016. Online Cross-Modal Hashing for Web Image Retrieval. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 294–300.

[28] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic Memory Networks for Visual and Textual Question Answering. In *Proceedings of the 33nd International Conference on Machine Learning*. 2397–2406.

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio:. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*. 2048–2057.

[30] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[31] Victor S. Lempitsky Yaroslav Ganin. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*. 1180–1189.

[32] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level Attention Networks for Visual Question Answering. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[33] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2852–2858.

[34] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 4995–5004.

[35] Yuke Zhu, Joseph J. Lim, and Li FeiFei. 2017. Knowledge Acquisition for Visual Question Answering via Iterative Querying. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 6146–6155.