Knowledge-Based Systems 207 (2020) 106339

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Visual Question Answering via Combining Inferential Attention and Semantic Space Mapping

Yun Liu [a], Xiaoming Zhang [b,c,*], Feiran Huang [d], Zhibo Zhou [a], Zhonghua Zhao [e], Zhoujun Li [f]

[a] Beijing Key Laboratory of Network Technology, Beihang University, Beijing, 100191, China
[b] School of Cyber Science and Technology, Beihang University, Beijing, 100191, China
[c] Beihang University Hefei Innovation Research Institute, Hefei, 230012, China
[d] College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou, 510632, China
[e] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029, China
[f] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China

ARTICLE INFO

Article history:
Received 29 December 2019
Received in revised form 21 June 2020
Accepted 28 July 2020
Available online 8 August 2020

Keywords:
Visual Question Answering
Inferential attention
Semantic space mapping

ABSTRACT

Visual Question Answering (VQA) has emerged and aroused widespread interest in recent years. Its purpose is to explore the close correlations between the image and question for answer inference. We have two observations about the VQA task: (1) the number of newly defined answers is ever-growing, which means that answer prediction on pre-defined labeled answers may lead to errors, as some unlabeled answers may be the right choice to the question–image pairs; (2) in the process of answering visual questions, the gradual change of human attention has an important guiding role in exploring the correlations between images and questions. Based on these observations, we propose a novel model for VQA, i.e., combining Inferential Attention and Semantic Space Mapping (IASSM). Specifically, our model has two salient aspects: (1) a semantic space shared by both the labeled and unlabeled answers is constructed to learn new answers, where the joint embedding of a question and the corresponding image is mapped and clustered around the answer exemplar; (2) a novel inferential attention model is designed to simulate the learning process of human attention to explore the correlations between the image and question. It focuses on the more important question words and image regions associated with the question. Both the inferential attention and the semantic space mapping modules are integrated into an end-to-end framework to infer the answer. Experiments performed on two public VQA datasets and our newly constructed dataset show the superiority of IASSM compared with existing methods.

1. Introduction

With the great development of natural language processing and computer vision, problems of combining vision and language in artificial intelligence are inspiring considerable research interests. A new task called Visual Question Answering (VQA) [1–5] has emerged as an promising but intractable research point. VQA requires the algorithms to output answers for natural language questions about the contents of the given images. Compared with conventional multi-modal tasks such as cross-modal retrieval [6–8] and image captioning [9–11], the VQA task demands a deep understanding of the input image and question sentence to infer the answer. VQA can be widely applied to many scenarios and plays a crucial role, e.g., early education, human-machine interaction, medical assistance, and automatic customer service [12].

A great variety of VQA methods sprung up in recent years are based on deep neural networks [13–16], which concentrates on learning an effective multi-modal joint embedding of the image and question to infer the answer. Most of the existing methods first employ a visual attention mechanism [17,18] to learn the joint embedding by exploring the correlations between the question words and image regions. Then, VQA is considered as a classification problem, and the learned joint embedding is fed into an answer classifier trained on a large number of labeled samples. The candidate answers of the classifier are the labeled answers in the training dataset. However, there are two drawbacks in the existing VQA methods. On one hand, a sufficient set of question–image pairs labeled with corresponding answers are usually unavailable in real-world applications. It results in

* Corresponding author.
E-mail address: yolixs@buaa.edu.cn (X. Zhang).

https://doi.org/10.1016/j.knosys.2020.106339
0950-7051/

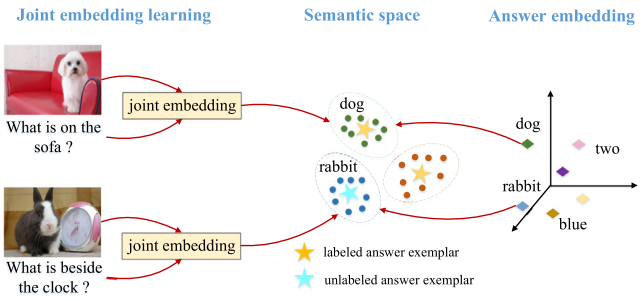


Fig. 1. Visualization of the mapping from question–image joint embedding to semantic space. Each answer exemplar in the semantic space corresponds to an answer embedding.

that the performance of VQA will be affected when the correct answers of the testing questions are unlabeled and existing outside the training dataset. Moreover, the number of newly defined answers is ever-growing, which means that training a specific model for each answer is unattainable. One the other hand, existing attention-based VQA methods implicitly explore the correlations between the visual content and textual sentence, which cannot capture the different importance of the terms in the process of answering questions. These attention-based methods lack a reasonable derivation process, which makes them inconsistent with the progressive changes in human attention as they answer visual questions. Therefore, there is an urgent need to design an explicit mechanism to solve these two shortcomings in the previous VQA methods for more accurate answer prediction.

In recent years, Zero-Shot Learning (ZSL) [19–21] has been proposed as an ambitious paradigm in the image recognition task to recognize the novel classes for which no training samples are provided. Inspired by it, we regard that new answers that are unlabeled in the training set can be predicted by treating them as recognizing the novel classes with ZSL. To predict the unlabeled answers, we introduce an intermediate semantic space that is shared between the labeled and unlabeled answers. In the semantic space, semantic information can be transferred from labeled answers to unlabeled answers. As the examples shown in Fig. 1, given the question "what is on the sofa?" and the corresponding image with the labeled answer such as "dog", a joint embedding of the question–image pair is first learned. Then, the joint embeddings which seek for the same labeled answer are mapped into the semantic space and clustered around the corresponding answer exemplar. The answer exemplar in this instance is the embedding of the labeled answer "dog" sought by the question–image pair. When testing with a new question "what is beside the clock?" which seeks for an unlabeled answer "rabbit", the joint embedding learned from the question–image pair is mapped into the semantic space. Since this embedding is located nearby the labeled answer exemplar "dog", the answer for the image and the corresponding question is obtained by searching the most matched answer exemplar around "dog". Then, the unlabeled answer "rabbit" is inferred.

As for the visual attention mechanism used to capture the correlations between the image and question, it can receive inspiration from the learning process of human attention. Take the image and the related question "what is beside the clock?" shown in Fig. 1 as an example, the inferential process of human attention can be elaborated as follows. First, we will focus on finding the image regions related to the noun word "clock". Then, these regions are combined with the question sentence to learn the question-related multi-modal information, i.e., "What is located beside the focused regions?". In the end, this multi-modal information will be utilized to attend on the image again, and the answer "rabbit" is inferred. Therefore, it is appropriate to design a visual attention mechanism that is consistent with the learning process of human attention to infer the answer.

In this paper, we propose to make the best of Zero-Shot Learning and the inferential process of human attention for visual question answering. In particular, we investigate: (1) how to transfer information from labeled answers to unlabeled answers; (2) how to effectively capture the close correlations between the image and question sentence to learn effective multi-modal joint embedding for answer inference. To solve these questions, we propose a novel VQA model, i.e., combining Inferential Attention and Semantic Space Mapping (IASSM). Our model mainly contains two components as shown in Fig. 2. Specifically, an inferential attention network is designed to imitate the learning process of human attention to capture more effective multi-modal correlations between the image and question. The inferential attention network attempts to learn more reasonable attention maps for the noun words and the question-related multi-modal information. In order to predict unlabeled answers, a semantic space shared by the labeled and unlabeled answers is designed. Each answer exemplar in the semantic space corresponds to the embedding of an answer. For the question–image pairs, the joint embeddings are mapped to the semantic space and clustered around the answer exemplars. Then, the unlabeled answers to the new questions can be inferred by seeking the most matched answer exemplars in the semantic space. The major contributions of this paper are concluded as follows:

- Unlike previous VQA works, we investigate the problem of predicting unlabeled answers. A semantic space is designed to transfer information from labeled answers to unlabeled answers.
- Different from existing attention-based approaches, we propose a novel inferential attention network to imitate the learning process of human attention to learn more effective multi-modal correlations between the image and question.
- We construct a zero-shot dataset for VQA from a public dataset. Extensive experiments conducted on our constructed dataset and two public VQA datasets confirm the favorable performance of our model compared with the baselines.

The remainder of this paper is organized as follows. We first review the related works of VQA and Zero-Shot Learning. Then, our model IASSM is introduced in detail. Next, the experimental results and further analysis are presented. At last, we summarize this paper and look forward to future work.

2. Related work

VQA To facilitate VQA research, several datasets are constructed and introduced in [1,22–24], which are automatically generated or manually labeled from image caption datasets. Based on these datasets, the VQA works emerged in recent years mainly adopt deep neural networks to lean a joint embedding of the image and question for answer inference. Early methods detailed in [25–27] directly combine the two embeddings learned from image and question as a joint embedding and feed it into an answer classifier. However, these methods cannot explore the close correlations between the textual sentence and visual content. To tackle this problem, the work detailed in [28] designs a stacked-attention network, which utilizes a question representation to query the corresponding image multiple times to find meaningful image regions related to the question. After that, various attention models including dual attention [29], co-attention [30], multi-level attention [13], and dynamic attention [31] are proposed to

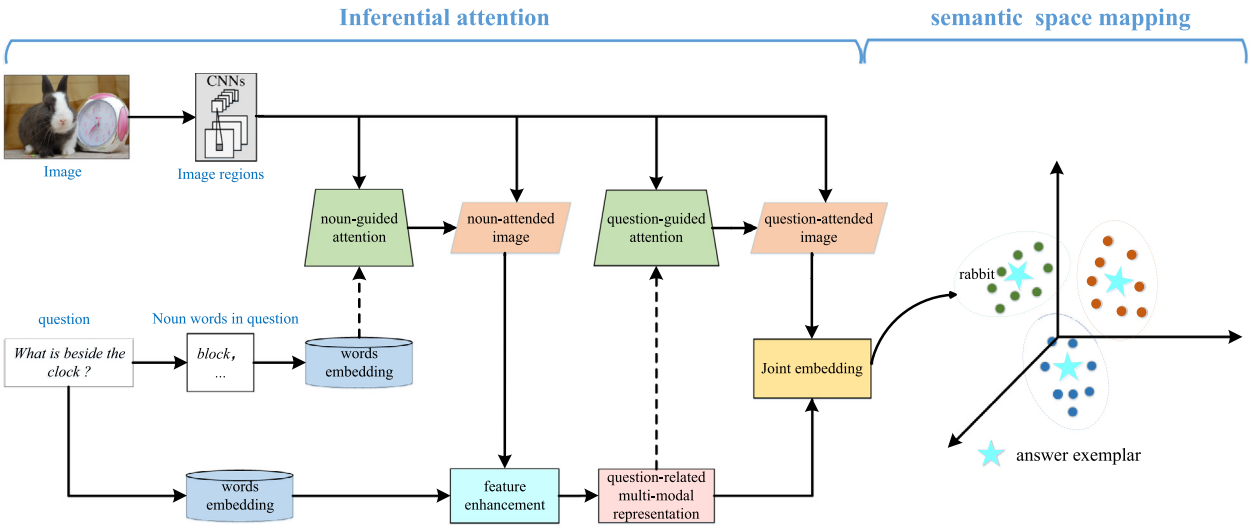


Fig. 2. The framework of the IASSM model, which mainly contains two components, i.e., inferential attention and semantic space mapping.

explore multi-modal correlations between the image and question from different views. Dynamic memory [32] and Dynamic parameter [33] networks are designed respectively, where the framework and weights are regulated adaptively according to the question. Some methods based on cross-modal feature fusion, such as MLB [34] and MCB [35], adopt bilinear models with the multi-modal pooling scheme to learn the joint embedding for answer inference. A generalized multimodal tensor-based tucker decomposition model is introduced in [36], which aims to efficiently parametrize bilinear interactions between visual and textual embeddings. An external knowledge base is applied in the works presented in [11,37,38], which is queried by the question and act as supplementary materials to provide richer information. Experiments show that adding external knowledge yields more useful features together with a better performance. A novel cubic visual attention model designed in [39] explores a particular spatial and channel attention on object regions from different dimensions. The work presented in [40] proposes an object difference attention model, which computes the attention distribution by realizing different operators between different image objects according to the question. A fusion scheme [41] based on differential networks that utilize a novel plug-and-play module to enable differences between pair-wise feature elements. The work presented in [42] enforces the correlation between the attention and the nonattention parts as a constraint for attention learning. Although these methods have achieved promising performance, they all use classifiers to make predictions on pre-defined labeled answers, but cannot predict unlabeled answers outside the training dataset, which will degrade the performance of answer inference.

Zero-Shot Learning Various methods for Zero-Shot Learning have been proposed in recent years. These methods mainly focus on image recognition. Most studies in this field inherently have a two-step process. First, an embedding function is learned to map the intermediate representations and visual features into the same semantic space. Second, a nearest-neighbor search in the mapped space is carried out to predict the class label. Recently, a deep visual semantic embedding method [19] is proposed, which uses the annotated image data and the semantic information extracted from the unlabeled text to identify the visual objects. The work presented in [20] proposes a new zero-shot learning scheme that makes full use of the clustering structures to gather similar semantic information in the semantic space. The method presented in [43] constructs a new framework to find a low-rank mapping that associates visual features with their semantic representations. A novel approach detailed in [21] employs the augmented semantics to the hinge loss functions to learn a robust mapping for zero-shot learning. The work introduced in [44] proposes a local relative distance metric to analyze the bias problem in generalized ZSL and designs a novel embedding model called co-representation network to solve it. In [45], ZSL is addressed as a verification problem and a deep extension paradigm is designed to enable previous and future ZSL works to benefit from deep models. Despite a certain success has been achieved by these methods, only the approach presented in [46] considers zero-shot learning in the VQA task. However, it requires external knowledge of web online images to learn the semantic information of the words unseen in the training dataset. Moreover, it does not discriminate the important words and image regions. This results in that the performance of new answers learning for VQA still lags behind satisfaction. In contrast, our model only uses the semantic space to infer new answers without external knowledge. In addition, we simulate human attention to distinguish the importance of different question words to the image regions, making it more effective to capture multi-modal correlations between the image and question.

3. IASSM for visual question answering

In this section, we first present an overview of IASSM. Then, each component of IASSM is introduced in detail.

3.1. Overview

Before the problem formulation, we define the notations used in this paper. For an original image I, multi-CNNs [47] are employed to encode the visual features of the image regions. We denote $R = \{r_1, r_2, \ldots, r_{m\times m}\} \in \mathbb{R}^{l\times m\times m}$ as the encoded regions of an image, where $m \times m$ is the number of regions and l is the dimensionality of the feature vector of each region. Let $Q = \{q_1, q_2, \ldots, q_t\}$ indicates a question sentence with t words, where $q_i \in \mathbb{R}^d$ is a word vector pre-trained on an external corpus and d is the dimensionality of the word vector. Additionally, we denote $A_l = \{a_{l1}, a_{l2}, \ldots, a_{ln}\}$ as the labeled answers and $A_u = \{a_{u1}, a_{u2}, \ldots, a_{uk}\}$ as the unlabeled answers in the answer space. $a \in A = \{A_l \cup A_u\}$ is an answer which corresponds to an answer exemplar in the semantic space.

Given a question Q and the corresponding image I, The VQA task aims to automatically generate the appropriate answer a to

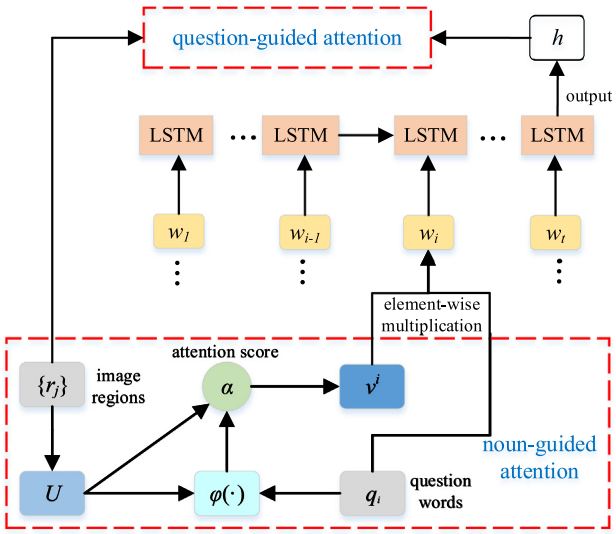


Fig. 3. Architecture of the inferential attention model, which mainly contains the noun-guided attention and the question-guided attention modules.

the question from the answer set A. For this goal, we focus on learning human-like attention to capture multi-modal correlations and predicting both labeled and unlabeled answers for more accurate answer inference. Our proposed model IASSM contains two components as shown in Fig. 2. The inferential attention network is built to capture the close correlations between the question and image to learn more effective multi-modal joint embedding. The semantic space mapping is proposed to infer the labeled and unlabeled answers for question–image pairs. The details of our model are described below.

3.2. Inferential attention

Given a question–image pair, the correlations between the image and question are expected to be captured to learn a multi-modal joint embedding for question answering. Previous works use attention-based methods to explore the close correlations between textual and visual contents. However, these approaches are ineffective in discovering the important words and regions to capture the fine-grained multi-modal correlations. In this section, we propose an inferential attention model to simulate the gradual learning process of human attention. It aims to learn more effective attention mapping by exploring the close correlations between image regions and the crucial sentence words, i.e., the noun words. The architecture of the inferential attention network is displayed in Fig. 3, which mainly contains two components including noun-guided attention and question-guided attention.

3.2.1. Noun-guided attention

Actually, the noun words play a crucial role in understanding the meaning of a question. When asked a question about the corresponding image, the noun words of the question sentence are first focused, and then the related regions in the image will be analyzed to infer the answer. Therefore, the correlations between the noun words and the image regions contribute greatly to infer the answer. We propose a noun-guided attention model to imitate this learning process.

For each word q_i in the question sentence Q, an attention score α_{ij} $(1 \le j \le m \times m)$ is allocated to each region r_j of the corresponding image according to its relevance with the content of r_j. Usually, a bilinear function is used to evaluate α_{ij}:

$$\alpha_{ij} \propto \varphi(q_i^T U r_j), \tag{1}$$

where the $\alpha_{i\cdot}$ are taken to normalize over all the image regions and $\propto$ is a proportional symbol. U is a learnable weight matrix and $\varphi(\cdot)$ is a smooth function, e.g., a softmax function. The attention scores are then utilized to adjust the intensity of attention on different image regions. The weighted sum of all the candidate regions are mapped from the visual feature space to the word space as follows:

$$v^i = \sum_{j=1}^{m\times m} \alpha_{ij}(U r_j), \tag{2}$$

Compared with the original visual features shared by all words, the weighted visual feature mapping v^i is more effective to represent the regions related to the current word q_i. Notice that, the attention score α_{ij} is designed for the noun words in the question sentence. If q_i is not a noun word, each image region has the same score for word q_i. Namely, there are no prominent features in the image correlated with it. For this reason, the following formula is designed to rewrite the attention weights:

$$\begin{cases} \alpha_{ij} \propto \varphi(q_i^T U r_j), & q_i \text{ is a noun word} \\ \alpha_{ij} = 1/(m \times m), & q_i \text{ is not a noun word.} \end{cases} \tag{3}$$

Based on this formula, we can assign more reasonable scores for the regions based on different words.

3.2.2. Question-guided attention

After obtaining the image regions related to the noun words, the regions will be combined with the question sentence to learn the question-related multi-modal information. Then, the multi-modal information is exploited to search the image again to infer the answer. To imitate this process, a question-guided attention model is proposed. It first exploits the correlations between textual words and visual regions to learn the question-related multi-modal representation. Then, the representation is utilized to attend the image again for more effective answer inference.

For each word vector q_i, it has the same dimension with the corresponding visual feature vector v^i after the processing of the noun-guided attention model. An enhanced feature vector w_i is designed as the joint representation of the multi-modal contents, which is obtained by element-wise multiplication of the two inputs q_i and v^i. L_2 normalization is used to constrain the magnitude of the representation as follows:

$$w_i = Norm_2(q_i \circ v^i), \tag{4}$$

where $\circ$ represents the operation of element-wise multiplication. We use LSTM to encode the sequence of the enhanced features $\{w_1, w_2, \ldots, w_t\}$. The output of the last LSTM cell acts as the question-related multi-modal representation h. It is used to attend the image again to infer the answer.

For further calculation, we first use the following formulas to transform h and image regions R into a c-dimensional common space:

$$h^{(c)} = tanh(W_h h + b_h), \qquad h^{(c)} \in \mathbb{R}^c, \tag{5}$$

$$R^{(c)} = tanh(W_r R + b_r), \qquad R^{(c)} \in \mathbb{R}^{c\times m\times m}, \tag{6}$$

where W_h and W_r are the trainable weight matrices, b_h and b_r are the bias terms. After that, $h^{(c)}$ is spatially replicated $m \times m$ times to form $H^{(c)} \in \mathbb{R}^{c\times m\times m}$ which matches the size of the transformed image regions $R^{(c)} = \{r_1^{(c)}, r_2^{(c)}, \ldots, r_m^{(c)}\}$.

The attention map of the multi-modal representation over the image regions is defined as follows:

$$M = Norm_2(H^{(c)} \circ R^{(c)}), \qquad C \in \mathbb{R}^{c\times m\times m}, \tag{7}$$

$$\beta = softmax(W_\beta * M + b_\beta), \qquad \beta \in \mathbb{R}^{m\times m}, \tag{8}$$

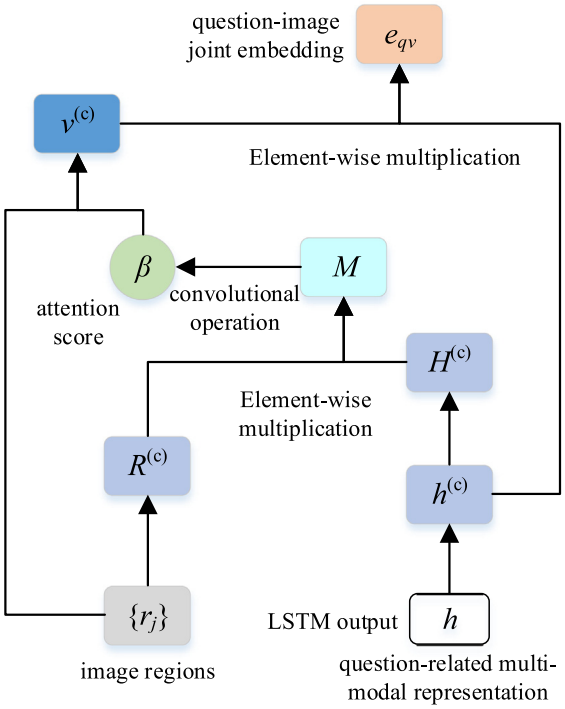


Fig. 4. An illustration of the question-guided attention network structure.

where $\circ$ represents the element-wise multiplication and $*$ is the convolutional operation. $W_\beta \in \mathbb{R}^{c\times1\times1}$ and $b_\beta \in \mathbb{R}^c$ are the trainable parameters. Based on the attention map β, a weighted sum of the transformed image regions is calculated as the attentive image feature $v^{(c)}$. Then, the transformed question-related multi-modal representation $h^{(c)}$ and the attentive image feature $v^{(c)}$ are jointly embedded by element-wise multiplication as follows:

$$v^{(c)} = \sum_{j}^{m\times m} \beta_j r_j^{(c)}, \qquad v^{(c)} \in \mathbb{R}^c, \tag{9}$$

$$e_{qv} = Norm_2(h^{(c)} \circ v^{(c)}), \qquad e_{qv} \in \mathbb{R}^c, \tag{10}$$

where e_{qv} is the joint embedding of the question and image, in which the question-related multi-modal semantic information and the correlations between the words and regions are encoded. The learning method of e_{qv} is consistent with the learning process of human attention, making it more effective to reflect the answer. The structure of the question-guided attention network is illustrated in Fig. 4.

3.3. Semantic space mapping

In order to predict unlabeled answers, a semantic space shared by both labeled and unlabeled answers is built for information transformation. That is, the joint embeddings of the question–image pairs are mapped into the semantic space in which the semantic information can be transferred from the labeled answers to unlabeled answers. Then, the answer for a question–image pair can be obtained by searching the most matched answer exemplar in the semantic space.

To facilitate comparison, a linear transformation is first used to map the question–image embedding e_{qv} into the d-dimensional semantic space to obtain $\phi(e_{qv})$ which has the same dimensionality with the answer exemplar $\psi(a)$. Notice that, $\psi(a)$ corresponds to the embedding of an answer a in the semantic space. A matching score is defined by the inner product between the mapped embedding of the question–image pair $\phi(e_{qv})$ and the answer exemplar $\psi(a)$ as follows:

$$s(e_{qv}, a) = \phi(e_{qv})\psi(a), \tag{11}$$

Like the classification score in previous VQA approaches [20, 21], the matching score is used to measure the matching degree between a question–image embedding and the answer exemplars. We use a standard softmax loss to define the matching loss as follows:

$$\mathcal{L}_{mat} = -\log \frac{exp(s(e_{qv}, a))}{\sum_{a'} exp(s(e_{qv}, a'))}, \quad a' \in A_l, \tag{12}$$

where A_l is the set of labeled answers.

Intuitively, the matching score between the mapped question-image embedding and the correct answer exemplar should be higher than the matching score between the embedding and a randomly -selected answer exemplar. For this goal, we use the matching score and the hinge rank loss to define the semantic similarity loss $\mathcal{L}_{sem}$ for per training example:

$$\mathcal{L}_{sem} = max(0, M_{sem} - \phi(e_{qv})\psi(a^+) + \phi(e_{qv})\psi(a^-)), \tag{13}$$

where a^+ is the vector of the correct answer exemplar for the provided question–image embedding. a^- is a vector of randomly selected wrong answer exemplar and M_{sem} is a hyper-parameter. This formula indicates that the matching score of the correct answer exemplar is at least greater than the margin M_{sem} compared with the randomly selected answer exemplar for the question–image embedding.

One specific insight of the semantic space is the existence of the clustering structure. When mapping to the semantic space, the question–image embeddings searching for the same answer will be located around the exemplar of that answer. Taking this into consideration, we propose to use a clustering loss to regulate the intra-answer and inter-answer distances among the mapped question–image embeddings. Formally, the clustering loss $\mathcal{L}_{clu}$ is defined as:

$$\mathcal{L}_{clu} = max(0, M_{clu} + d(\phi(e_{qv}), \phi(e_{qv}^+)) - d(\phi(e_{qv}), \phi(e_{qv}^-))), \tag{14}$$

In this formula, e_{qv}^+ represents a joint embedding learned from a randomly selected question–image pair, which has the same answer as e_{qv}. Oppositely, e_{qv}^- is another joint embedding of a randomly selected question–image pair, with a different answer as e_{qv}. M_{clu} is a hyper-parameter and $d(x, y)$ is used to calculate the squared Euclidean distance between x and y. Similarly, this formula indicates that the distance between two question–image embeddings with the same answer is at least smaller than the margin M_{clu} compared with the distance between two embeddings with different answers.

By combining the matching loss, the semantic similarity loss, and the clustering loss, we define the loss function of the proposed IASSM for each sample as follows:

$$\mathcal{L} = \mathcal{L}_{mat} + \lambda\mathcal{L}_{sem} + \eta\mathcal{L}_{clu}, \tag{15}$$

where λ and η represent the weight values of the corresponding items.

At the test stage, the answer a^* for a question–image pair can be inferred by simply selecting the most matched answer as follows:

$$a^* = \arg\max_{a\in A}(s(e_{qv}, a)) = \arg\max_{a\in A}(\phi(e_{qv})\psi(a)), \tag{16}$$

where A is the set of both labeled and unlabeled answers. In this way, we are able to predict the unlabeled answers for VQA via the semantic space mapping.

4. Experiments

4.1. Datasets and baselines

We conduct extensive experiments on the following three datasets to evaluate the performance of IASSM:

VQA v1.0 [1] is the most commonly used VQA dataset constructed from the image caption dataset MS-COCO. The questions are divided into three categories, including *yes/no*, *number*, and *other*. Each question corresponds to 10 answers created by crowsourced workers. This dataset consists of three splits, i.e., train (248,349 samples), val (121,512 samples), and test (244,302 samples). The test set contains two parts: test-std and test-dev. Additionally, there are two subtasks, i.e., Open-Ended (OE) and Multi-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.

COCO-QA [23] is another widely used VQA dataset automatically generated from MS-COCO dataset. Four question categories contained in this dataset, including *Object* (70%), *Number* (7%), *Color* (17%) and *Location* (6%). This dataset has a training set and a test set. It contains a total of 92,396 questions, 69 172 images, and 435 answers. Moreover, all answers are single-word.

ZSL-VQA Since there is no existing VQA dataset for zero-shot learning, we constructed such a dataset from the COCO-QA dataset and named it ZSL-VQA. We consider the samples to be zero-shot if the corresponding answers existing outside the training set. This dataset is formed by defining new splits of the training and test sets on the COCO-QA dataset. Specifically, no overlap between the answers of the training and test sets. To ensure that each question category is contained in the two splits of the datasets, 10% and 20% of the answers that have the least frequencies in the category of *object* and other categories are selected as the test set. The training set is composed of all remaining instances. Answers in the test set typically describe fine-grained categories and very specific concepts. An analysis of the resulted dataset is given in Table 1.

To analyze the performance of IASSM, the following approaches are used as the main baselines:

- **DPPnet** [33]: A model utilizes a dynamic parameter layer whose weights are determined adaptively according to the input sentence to learn a convolutional neural network for answer inference.
- **SAN** [28]: A model uses a question representation to query an image multiple times to find the regions associated with the answer.
- **HieCoAtt** [30]: A co-attention framework that jointly capture the textual and visual attention distribution in a hierarchical scheme through a 1-dimensional convolutional neural network.
- **ZSVQA** [46]: A novel model which is the first time of considering zero-shot in the VQA task, while it depends on web online images for each unseen word and ignores the correlations between image and question.
- **Dual-MFA** [48]: Multi-modal features learned from different visual attention mechanisms, i.e., free-form and detected-boxes, are fused using a novel embedding scheme to infer the answer.
- **CVA** [39]: A novel visual attention mechanism uses a particular channel and spatial attention on object areas of an image for answer prediction.
- **ODA** [40]: An object difference attention mechanism computes the attention distribution via realizing difference operators among different image objects in an image according to the questions.

Table 1
Training and test splits of the ZSL-VQA dataset.

	Training		Test	
Number of samples	99 947		17 737	
Number of answer categories	286		144	
Number of each question category				
(objects, number,	71 201	6595	7997	2053
color, location)	14 084	8067	5563	2124

- **DF** [41]: A fusion scheme based on differential networks which utilize a novel plug-and-play module to enable differences between pair-wise feature elements.

4.2. Experimental settings and evaluation methods

For question sentences, Stanford POS Tagger[1] is used to find the noun words before training. We employ 200-dimensional pre-trained GloVe [49] vectors to encode each word in the question sentence. As for the images, we embed the visual features using the VGG19 [50] networks pre-trained on the ImageNet 2012 dataset. Concretely, the output of the convolutional layer "conv5_4" is used as the region maps with the dimensionality of $512 \times 14 \times 14$ for visual attention computing. As for the proposed model, the size of LSTM and the dimensionality of c-dimensional space in the question-guided attention network are set to be 300 and 500, respectively. The dimensionality d of the semantic space is set to be 200. M_{sem} and M_{clu} are set to be 0.1 and 1.0, respectively. The activation functions used in this paper are fixed to $relu(\cdot)$ except special instruction. For the optimization algorithm, the adaptive moment estimation is employed with an initial learning rate of 3×10^{-3}, a weight decay 10^{-8} and a momentum 0.98. We fix the batch size to be 128 and the hyper-parameters λ and η are empirically set to be 0.4 and 0.06 respectively. Dropout with a value of 0.5 is used after each linear transformation to prevent overfitting for relatively good performance.

For VQA v1.0 dataset, the evaluation method detailed in [1] is used as the metric: *accuracy = min(#humans that provided that answer/3, (1)*. That is, an answer is considered as 100% accurate if no less than 3 workers provided the exact answer. As for COCO-QA and ZSL-VQA datasets, the commonly used classification accuracy is employed to evaluate the models. Moreover, WU–Palmer Similarity (WUPS) introduced in [53] is utilized as another metric to analyze the model performance as reported in [23]. WUPS calculates the similarity score between two words based on their common subsequence in a taxonomy tree. Specifically, 0.0 and 0.9 are chosen as the threshold to form metrics WUPS@0.0 and WUPS@0.9 for evaluation, respectively.

4.3. Experimental results and further analysis

We evaluate the performance of IASSM by comparing it with state-of-the-art baselines in both general VQA task and zero-shot VQA task.

4.3.1. General VQA task

Table 2 shows the experimental results on VQA v1.0 dataset. The models are trained on the train + val set and tested on both test-dev and test-std sets. On the test-dev set, it can be observed that IASSM improves the overall accuracy of the best baseline DF from 68.62% to 69.35% in Open-Ended task, and improves DF by 0.11% in Multi-Choice (MC) task. Furthermore,

[1] http://nlp.stanford.edu/software/tagger.shtml.

Table 2
Experimental results of comparing the proposed IASSM with state-of-the-art methods on the VQA v1.0 dataset. The best value on each category is highlighted.

Method	Test-dev					Test-std				
	Open-Ended				MC	Open-Ended				MC
	All	Yes/No	Number	Other	All	All	Yes/No	Number	Other	All
DPPnet [33]	57.22	80.71	37.24	41.69	62.48	57.36	80.28	36.92	42.24	62.69
SAN [28]	58.70	79.30	36.60	46.10	–	58.9	79.11	36.41	46.42	–
HieCoAtt [30]	61.80	79.70	38.70	51.70	65.8	62.1	79.95	38.22	51.95	66.07
MCB [35]	64.70	82.50	37.60	55.60	69.10	66.50	–	–	–	–
Dual-MFA [48]	65.41	83.59	40.18	56.34	70.04	67.09	83.37	40.39	56.89	69.97
CVA [39]	65.92	83.73	40.91	56.36	70.30	66.20	83.79	40.41	56.77	70.41
VKMN [51]	66.00	83.70	37.90	57.00	69.10	66.10	84.10	38.10	56.90	69.10
DCN [52]	66.89	84.61	42.35	57.31	–	67.02	85.04	42.34	56.98	–
ODA [40]	67.83	85.82	43.03	58.07	72.28	67.97	85.81	42.51	**58.24**	72.32
DF [41]	68.62	86.08	**43.52**	58.38	73.31	68.48	85.81	42.87	58.23	73.05
IASSM(ours)	**69.35**	**87.04**	42.86	**58.94**	**73.42**	**69.05**	**86.85**	**43.68**	58.17	**73.24**

Table 3
Experimental results on both COCO-QA and ZSL-VQA datasets. "–" indicates the data is unavailable. The best value on each category is highlighted.

Methods	COCO-QA			ZSL-VQA				
	Accuracy	WUPS@0.9	WUPS@0.0	Accuracy	Obj.	Num.	Col.	Loc.
VSE [23]	54.85	64.78	88.12	21.46	24.15	15.38	19.26	17.51
Img-CNN [25]	58.40	68.50	89.67	–	–	–	–	–
ZSVQA [46]	60.32	69.68	90.13	30.34	34.58	27.16	29.73	28.81
DPPnet [33]	61.19	70.84	90.61	–	–	–	–	–
SAN [28]	61.60	71.60	90.90	23.13	26.25	18.12	21.51	19.32
HieCoAtt [30]	65.40	75.10	92.00	–	–	–	–	–
Dual-MFA [48]	66.49	76.15	92.29	–	–	–	–	–
CVA [39]	67.51	76.70	92.41	–	–	–	–	–
ODA [40]	69.33	78.29	93.02	–	–	–	–	–
DF [41]	69.36	78.35	93.19	–	–	–	–	–
IASSM(ours)	**69.92**	**79.04**	**93.74**	**36.36**	**39.64**	**33.32**	**35.29**	**34.63**

IASSM outperforms HieCoAtt, Dual-MFA, CVA, DCN, and ODA on accuracy by 7.55%, 3.94%, 3.43%, 2.46%, and 1.52% respectively. As for the test-std set, IASSM improves the overall accuracy of the best baseline DF from 68.48% to 69.05% in the Open-Ended task. Similar improvements can also be seen on the three question categories.

From the comparison results on the COCO-QA dataset displayed in the left part of Table 3, one can see that IASSM achieves better performance compared with all the baselines on the metric of classification accuracy. Concretely, IASSM outperforms SAN, HieCoAtt, Dual-MFA, CVA, and ODA on accuracy by 8.32%, 4.52%, 3.43%, 2.41%, and 0.59%, respectively. Moreover, IASSM improves the accuracy of the best baseline DF from 69.36% to 69.92%. Similar improvements of our model compared with the baselines can also be seen on the metrics of WUPS@0.0 and WUPS@0.9.

The reason for the improvements is that state-of-the-art approaches explore the correlations between the image and question via various attention-based methods to infer the answer. However, these attention models do not consider the different importance of different words (especially the noun word) and image regions, which will affect the effectiveness of the learned multi-modal joint embedding. The improvements indicate that the inferential attention model which imitates the learning process of human attention is effective for VQA. It is reasonable as inferential attention can focus on the exact image regions for more effective embedding learning.

4.3.2. Zero-shot VQA task

The right part of Table 3 shows experimental results on the ZSL-VQA dataset. Three existing approaches VSE, ZSVQA, and SAN are selected as the baselines to compare with IASSM. VSE and SAN are common classification approaches, which do not consider predicting the unlabeled answers. ZSVQA is the only baseline considering zero-shot learning in VQA. However, it does not pay sufficient attention on the latent correlations between question words and image regions. Moreover, it relies on web online images to predict the unlabeled answers. From the results, one can see that our model IASSM improves the accuracy of ZSVQA from 30.34% to 36.36%. Similar improvements of IASSM compared with ZSVQA can also be seen on the four question categories, i.e., *Object*, *Number*, *Color* and *Location*. The results demonstrate that IASSM has better performance than ZSVQA. The reason is that IASSM can predict the unlabeled answers by exploring the semantic space mapping and the correlations between the image and question more effectively. Additionally, IASSM outperforms VSE and SAN on accuracy by 14.90% and 13.23%, respectively. It indicates that predicting the unlabeled answer is beneficial for the performance of VQA.

Table 4
Ablation study on both COCO-QA and ZSL-VQA datasets.

Methods	Accuracy	
	COCO-QA	ZSL-VQA
NG-Att	62.53	31.72
QG-Att	65.72	34.57
I-Att[a]	69.92	36.36
Mat-Loss	62.36	33.43
MatSem-Loss	65.45	35.84
MatClu-Loss	67.23	34.76
MatSemClu-loss[a]	69.92	36.36

[a]Represents the implementation of our model.

4.4. Ablation study

We further conduct ablation experiments to analyze the effectiveness of the individual components proposed in our approach

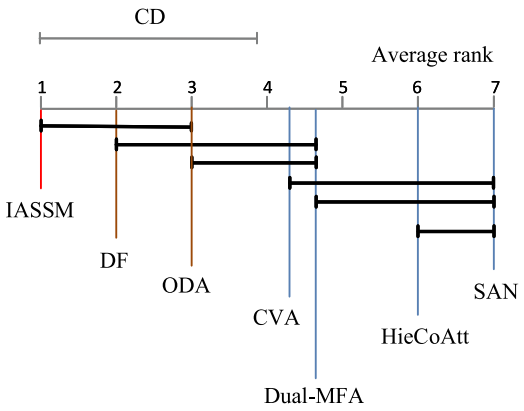


Fig. 5. Experimental results of the Friedman–Nemenyi test.

as shown in Table 4. These models are evaluated on both COCO-QA and ZSL-VQA datasets. The evaluation is performed by changing one component while fixing the other components of IASSM. The first part of Table 4 shows the experimental results of different attention models. One can see that our Inferential Attention model (I-Att) outperforms the Noun-Guided Attention (NG-Att) and the Question-Guided attention (QG-Att) by 7.39% and 4.20% in COCO-QA dataset, and 4.64% and 1.79% in ZSL-VQA dataset, respectively. It indicates that the proposed inferential attention network which imitates the learning process of human attention helps to improve the accuracy of question answering. The reason is that the inferential attention model pays more attention on the important words and image regions for in-depth semantic comprehension. The second part of Table 4 shows the importance of different loss functions designed in the semantic space mapping. One can see that the matching loss plus semantic similarity loss (MatSem-Loss) outperforms the single matching Loss (Mat-Loss) on accuracy by 3.09% and 2.41% in the two datasets, respectively. Similarly, the matching loss plus the clustering loss (MatClu-Loss) outperforms the single matching loss on accuracy by 4.87% and 1.33% in the two datasets, respectively. Moreover, the weighted sum of matching loss, semantic similarity loss, and clustering loss (MatSemClu-Loss) achieve the best performance. It demonstrates that the three loss functions play an indispensable role in IASSM for answer prediction.

4.5. Statistical significance test

In order to evaluate if there exists a significant difference in the performance of different VQA approaches, we conduct Friedman–Nemenyi test [54] with a 95% confidence level on VQA v1.0 Test-dev, VQA v1.0 Test-std, and COCO-QA datasets, based on the metric of accuracy. Friedman test is a nonparametric test method that uses rank to test the significant differences of multiple population distributions on several datasets. First, for each dataset, all models are ranked according to the performance of the models on that dataset. Then, calculating the average rank of each model on each dataset. Finally, the Friedman–Nemenyi test is employed to find the differences between the models. If the null hypothesis (all models are equivalent) is rejected, Nemenyi's post-doc test can be performed to compare the critical distance (CD) with the difference between the average rank of any two models. If the critical distance is less than the difference between the two models, it can be concluded that there is a significant difference between the two models.

The experiment results of Friedman–Nemenyi test are shown in Fig. 5. It can be seen that the proposed IASSM obtains the best average ranking performance in all the approaches. Fig. 5 also illustrates that the performance of IASSM significantly outperforms many baselines, including SAN, HieCoAtt, Dual-MFA, and CVA. The differences between IASSM and the two baselines (DF and ODA) are less significant. The reason is that DF and ODA can exploit the latent correlations between visual content and textual sentence in a sensible way for multi-modal joint embedding learning. However, despite DF and ODA achieve the second and third average rank respectively, they are not effective to capture the different importance between different words and image regions. This results in that they cannot learn accurate attention maps like IASSM which can imitate the learning process of human attention.

4.6. Parameter sensitivity

To analyze how the performance of IASSM is affected by parametrization, we present the results of answer classification accuracy with different parameter settings on the ZSL-VQA dataset. Concretely, we assess how the values of the balance parameters (λ and η) and the size of the d-dimensional semantic space affect the result.

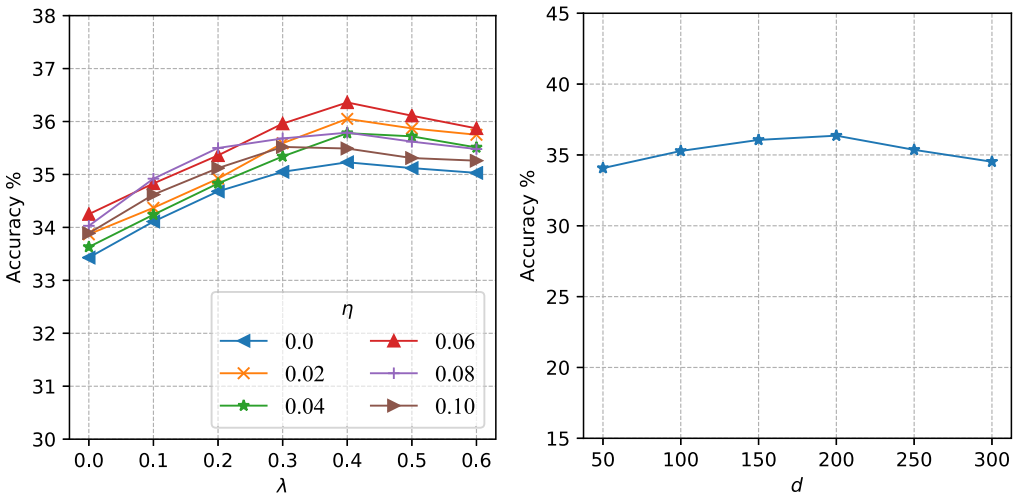
Balance parameters In the loss function of IASSM in Eq. (15), λ is used to balance the importance of the semantic similarity loss, and η is a trade-off parameter for the clustering loss. We first fix the dimension of the semantic space $d = 200$ and then test the performance with different values of λ and η. Based on the curves in Fig. 6, one can see that the semantic similarity between the question-image embedding and answer embedding contributes to the performance since the model achieves relatively low performance when $\lambda = 0$. However, a too large value of λ will result in overfitting. On the other side, when $\eta = 0$, a relatively low accuracy is achieved, which indicates that a trade-off term to make the joint embeddings to be clustered around the corresponding answer exemplar is important for answer inference. Additionally, compared with η, the value of the parameter λ has more impact on the performance. From the curve, it can be seen that the model achieves the highest accuracy when $\lambda = 0.4$ and $\eta = 0.06$.

Space dimension In the proposed model, the joint embeddings of the question–image pairs are mapped into the d-dimensional semantic space. The answer exemplars are also represented by d-dimensional feature vectors in the semantic space. Fig. 6(b) shows how the dimensionality of the semantic space affects the performance by fixing $\lambda = 0.4$ and $\eta = 0.06$. It can be seen that the accuracy rises initially and then starts to drop slowly as the size of the dimension increases. It is reasonable as 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 model performance. Overall, a reasonable size of the dimension is greatly helpful for the performance. From the curve, one can see that IASSM reaches the highest accuracy when $d = 200$.

4.7. Analysis of the attention model

For investigating the effect of the attention model in the process of answer inference, we visualize the attention maps of the widely used question-guided attention and the proposed inferential attention, respectively. Question-guided attention directly uses a question sentence to search important image regions for answer inference. The inferential attention model aims to inferentially simulate the learning process of human attention for more effective embedding learning.

Following [55], the weights of the two attention models are visualized using upsampling and Gaussian filtering. The visualization results are shown in Fig. 7. It can be seen that the question-guided attention model can effectively capture the main



(a) Balance parameters (b) Space dimension

Fig. 6. Parameter sensitivity study for balance parameters and space dimension.

correlations between the question and image regions, while it cannot accurately focus on the right regions related to the answers. For instance, it can be observed that the question-guided attention model aligns the important word "dogs" with the right regions in Fig. 7(b). However, it aligns "flowers" with the regions that are unrelated to the right answer in Fig. 7(d), which results in a wrong answer prediction. Notably, the inferential attention model can effectively learn the question-related multi-modal information and focus on the correct image regions associated with the answer, thus predicting a more confident or more accurate answer compared with the question-guided attention model. For instance, Fig. 7(a) shows that the inferential attention model pays more attention to the cap regions rather than the head and cap regions in the question-guided attention model. One can see that the inferential attention model obtains the correct answer "white", while the wrong answer "black" is inferred by the question-guided attention model. In Fig. 7(c), the football-related regions are focused by the inferential attention to infer the correct answer more confidently. The results indicate that the inferential attention model can learn more effective semantic information to capture the correlations between the question and image.

4.8. Analysis of new answers learning

To analyze the performance of new answers learning in the proposed model, we show the results of two test samples in Fig. 8. Specifically, the results of new answers learning on both SAN [28] and IASSM are displayed. The correct answers to the two question–image pairs are unlabeled and exist outside the training dataset. As can be seen from the results, the existing approach SAN does not consider new answers learning, and thus obtains wrong answers. For instance, as the answers "broccoli" and "apricot" are unlabeled in the training dataset, SAN will not consider them as the candidate answers in the answer inference process. However, IASSM can predict the accurate answers for the two question–image pairs although the answers exist outside the training dataset. The reason is that the semantic space mapping designed in IASSM can transfer information from the labeled answers to unlabeled answers. The results demonstrate that learning new answers can significantly improve VQA performance.

5. Conclusion and future work

This paper focuses on the Visual Question Answering (VQA) task that has emerged in recent years, and its research has important practical implications for early education, human-machine interaction, etc. Our goal is to explore human-like attention to capture more effective multi-modal correlations and predicting unlabeled answers for VQA. We propose a novel model for the VQA task, i.e., combining Inferential Attention and Semantic Space Mapping (IASSM). Specifically, an inferential attention model is designed to inferentially imitate the learning process of human attention, in which the important words and image regions are paid more attention. Besides, a semantic space shared by both labeled and unlabeled answers is designed to transfer semantic information from the labeled answers to unlabeled answers. Then, predicting unlabeled answers is considered as selecting the most matched answer exemplar in the semantic space. Extensive experiments performed on two public VQA datasets and the newly constructed dataset demonstrate the superiority of IASSM on both the general VQA task and the Zero-Shot VQA task. Our method is different from most of the existing works that are ineffective in inferring new answers. It certifies that learning new answers and distinguishing the different importance of words and regions like human attention to explore the close correlations between the image and question contribute to VQA performance.

In future work, we will explore free-form answers generation. Furthermore, it would be interesting to investigate the social information, like the micro-blogs published by the owner and the context information of the image, for visual question answering.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. U1636210 and U1636211) and Beijing Natural Science Foundation of China (No. 4182037).

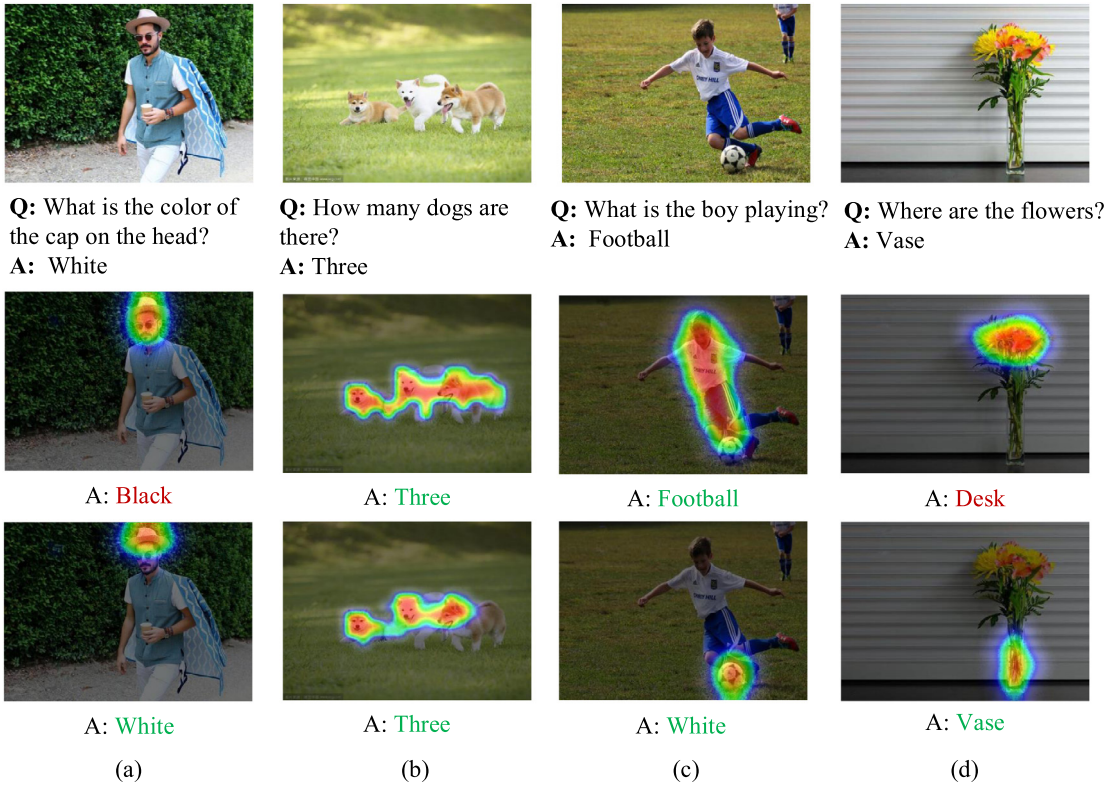


Fig. 7. Examples of attention visualization and the predicted answers. The first row contains the input images. The second row represents the question-guided attention maps. The last row indicates the inferential attention maps. Answers with green and red color represent correct and wrong prediction, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

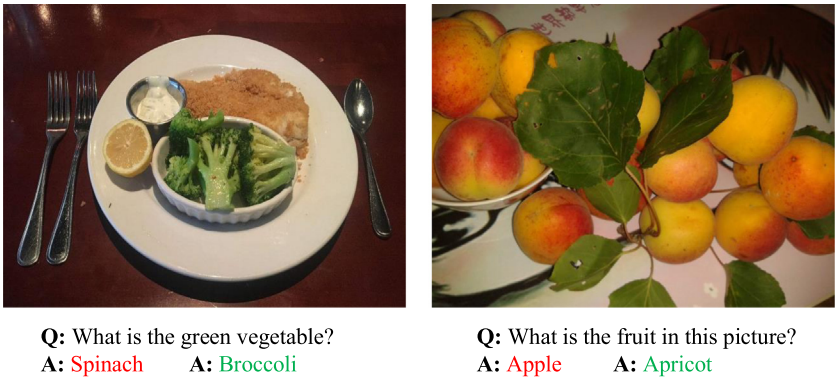


Fig. 8. Two qualitative results of new answers learning. Answers with red color (wrong answer) are given by SAN [28]. Answers with green color (correct answer) are predicted by the proposed IASSM model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[2] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel, Visual question answering: A survey of methods and datasets, Comput. Vis. Image Underst. 163 (2017) 21–40.

[3] Kushal Kafle, Christopher Kanan:, Answer-type prediction for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4976–4984.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh, Making the v in vqa matter: Elevating the role of image understanding in Visual Question Answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6904–6913.

[5] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, Devi Parikh, Probabilistic neural symbolic models for interpretable visual question answering, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 6428–6437.

[6] Liang Xie, Jialie Shen, Lei Zhu, Online cross-modal hashing for web image retrieval, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 294–300.

[7] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, Xiaogang Wang, Person search with natural language description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5187–5196.

[8] Po-Yao Huang, Vaibhav, Xiaojun Chang, Alexander G. Hauptmann, Improving what cross-modal retrieval models learn through object-oriented inter- and intra-modal attention networks, in: Proceedings of the International Conference on Multimedia Retrieval, 2019, pp. 244–252.

[9] Andrej Karpathy, Li Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[10] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.

[11] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1367–1381.

[12] Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, Jianyong Wang, R-vqa: learning visual relation facts with semantic attention for visual question answering, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, pp. 1880–1889.

[13] Dongfei Yu, Jianlong Fu, Tao Mei, Yong Rui, Multi-level attention networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 21–29.

[14] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, Anton van den Hengel, Image captioning and visual question answering based on attributes and external knowledge, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1367–1381.

[15] Tingting Qiao, Jianfeng Dong, Duanqing Xu, Exploring human-like attention supervision in visual question answering, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[16] Jongkwang Hong, Jianlong Fu, Youngjung Uh, Tao Mei, Hyeran Byun, Exploiting hierarchical visual features for visual question answering, Neurocomputing 351 (2019) 187–195.

[17] Badri Patro, Vinay P. Namboodiri, Differential attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7680–7688.

[18] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, Alexander Hauptmann, Focal visual-text attention for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6135–6143.

[19] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al., Devise: A deep visual-semantic embedding model, in: Proceedings of Advances in Neural Information Processing Systems, 2013, pp. 2121–2129.

[20] Soravit Changpinyo, Wei-Lun Chao, Fei Sha, Predicting visual exemplars of unseen classes for zero-shot learning, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 3496–3505.

[21] Bin Tong, Martin Klinkigt, Junwen Chen, Xiankun Cui, Quan Kong, Tomokazu Murakami, Yoshiyuki Kobayashi, Adversarial zero-shot learning with semantic augmentation, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 2476–2483.

[22] Mateusz Malinowski, Mario Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, in: Proceedings of Advances in Neural Information Processing Systems, 2014, pp. 1682–1690.

[23] Mengye Ren, Ryan Kiros, Richard S. Zemel, Exploring models and data for image question answering, in: Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 2953–2961.

[24] Yun Liu, Xiaoming Zhang, Feiran Huang, Xianghong Tang, Zhoujun Li, Visual question answering via attention-based syntactic structure tree-LSTM, Appl. Soft Comput. 82 (2019) 105584.

[25] Lin Ma, Zhengdong Lu, Hang Li, Learning to answer questions from image using convolutional neural network, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 3567–3573.

[26] Aiwen Jiang, Fang Wang, Fatih Porikli, Yi Li, Compositional memory for visual question answering, 2015, arXiv preprint arXiv:1511.05676.

[27] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, Wei Xu, Are you talking to a machine? dataset and methods for multilingual image question, in: Proceedings of Advances in Neural Information Processing Systems, 2015, pp. 2296–2304.

[28] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alexander J. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[29] Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim, Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2156–2164.

[30] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh, Hierarchical question-image co-attention for visual question answering, in: Proceedings of Advances in Neural Information Processing Systems, 2016, pp. 289–297.

[31] Yiyi Zhou, Rongrong Ji, Jinsong Su, Xiaoshuai Sun, Weiqiu Chen, Dynamic capsule attention for visual question answering, in: The Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 9324–9331.

[32] Caiming Xiong, Stephen Merity, Richard Socher, Dynamic memory networks for visual and textual question answering, in: Proceedings of International Conference on Machine Learning, 2016, pp. 2397–2406.

[33] Hyeonwoo Noh, Paul Hongsuck Seo, Bohyung Han, Image question answering using convolutional neural network with dynamic parameter prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 30–38.

[34] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, Byoung-Tak Zhang, Hadamard product for low-rank bilinear pooling, 2016, arXiv preprint arXiv:1610.04325.

[35] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: Proceedings of Conference on Empirical Methods on Natural Language Processing, 2016, pp. 457–468.

[36] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, Nicolas Thome, Mutan: Multimodal tucker fusion for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2631–2639.

[37] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, Anton van den Hengel, Ask me anything: Free-form visual question answering based on knowledge from external sources, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4622–4630.

[38] Yuke Zhu, Joseph J. Lim, Li FeiFei, Knowledge acquisition for visual question answering via iterative querying, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6146–6155.

[39] Jingkuan Song, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, From pixels to objects: Cubic visual attention for visual question answering, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 906–912.

[40] Chenfei Wu, Jinlai Liu, Xiaojie Wang, Xuan Dong, Object-difference attention: A simple relational attention for visual question answering, in: Proceedings of the ACM Multimedia Conference, ACM, 2018, pp. 519–527.

[41] Chenfei Wu, Jinlai Liu, Xiaojie Wang, Ruifan Li, Differential networks for visual question answering, in: The Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 8997–9004.

[42] Fei Liu, Jing Liu, Richang Hong, Hanqing Lu, Erasing-based attention learning for visual question answering, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1175–1183.

[43] Yang Liu, Quanxue Gao, Jin Li, Jungong Han, Ling Shao, Zero shot learning via low-rank embedded semantic autoencoder, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 2490–2496.

[44] Fei Zhang, Guangming Shi, Co-representation network for generalized zero-shot learning, in: Proceedings of the International Conference on Machine Learning, 2019, pp. 7434–7443.

[45] Haofeng Zhang, Yang Long, Yu Guan, Ling Shao, Triple verification network for generalized zero-shot learning, IEEE Trans. Image Process. 28 (1) (2018) 506–517.

[46] Damien Teney, Anton van den Hengel, Zero-shot visual question answering, 2016, arXiv preprint arXiv:1611.05546.

[47] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[48] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, Xiaogang Wang, 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, 2018.

[49] Jeffrey Pennington, Richard Socher, Christopher Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[50] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[51] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, Jianguo Li, Learning visual knowledge memory networks for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7736–7745.

[52] Duy-Kien Nguyen, Takayuki Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6087–6096.
[53] Zhibiao Wu, Martha Palmer, Verb semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, 1994, pp. 133–138.
[54] Milton Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of International Conference on Machine Learning, 2015, pp. 2048–2057.