Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-08790-0

Object-difference drived graph convolutional networks for visual question answering

Xi Zhu[1,2] · Zhendong Mao[3] · Zhineng Chen[4] · Yangyang Li[5] · Zhaohui Wang[1,2] · Bin Wang[6]

Received: 23 April 2019 / Revised: 20 December 2019 / Accepted: 24 February 2020
Published online: 20 March 2020

Abstract
Visual Question Answering(VQA), an important task to evaluate the cross-modal understanding capability of an Artificial Intelligence model, has been a hot research topic in both computer vision and natural language processing communities. Recently, graph-based models have received growing interest in VQA, for its potential of modeling the relationships between objects as well as its formidable interpretability. Nonetheless, those solutions mainly define the similarity between objects as their semantical relationships, while largely ignoring the critical point that the difference between objects can provide more information for establishing the relationship between nodes in the graph. To achieve this, we propose an object-difference based graph learner, which learns question-adaptive semantic relations by calculating inter-object difference under the guidance of questions. With the learned relationships, the input image can be represented as an object graph encoded with structural dependencies between objects. In addition, existing graph-based models leverage the pre-extracted object boxes by the object detection model as node features for convenience, but they are suffering from the redundancy problem. To reduce the redundant objects, we introduce a soft-attention mechanism to magnify the question-related objects. Moreover, we incorporate our object-difference based graph learner into the soft-attention based Graph Convolutional Networks to capture question-specific objects and their interactions for answer prediction. Our experimental results on the VQA 2.0 dataset demonstrate that our model gives significantly better performance than baseline methods.

Keywords Visual question answering · Graph convolutional networks · Object-difference

✉ Zhendong Mao
maozhendong2008@gmail.com

Extended author information available on the last page of the article

1 Introduction

Visual Question Answering (VQA) has attracted extensive attention from both the computer vision and natural language processing communities in recent years, since it is regarded as an important attempt to Artificial Intelligence for heterogeneous data understanding and reasoning [5]. Given an image and a related question, the VQA task requires both image and language understanding to answer a question. The fundamental challenges in VQA lay in effectively extracting the visual representation of images and embedding the textual sentences of questions, as well as breaking the semantic gap between image and question representations.

Many efforts have been devoted to filling this gap, especially with the recent advances in deep learning. Conventional VQA approaches [1, 8, 10, 20] encode questions through Recurrent Neural Networks(RNNs) [6] and images by CNN [25] separately, then combine them into joint representations by utilizing attention mechanism [19, 21, 27, 30, 35, 37]or external knowledge bases [13, 14, 22, 31, 32, 34], followed by a fully-connected layers to predict answers. These attention or external knowledge based VQA models are stuck with concentrating on objects which are related to the question, while neglecting the relationships among these objects. Thus they underperform in answering complex or reasoning questions. Recently, to address the aforementioned challenge, graph-based models [23, 29, 40] are proposed, for its great potentiality in modeling visual relationships. Therefore, in this paper, we focus on graph-based VQA models.

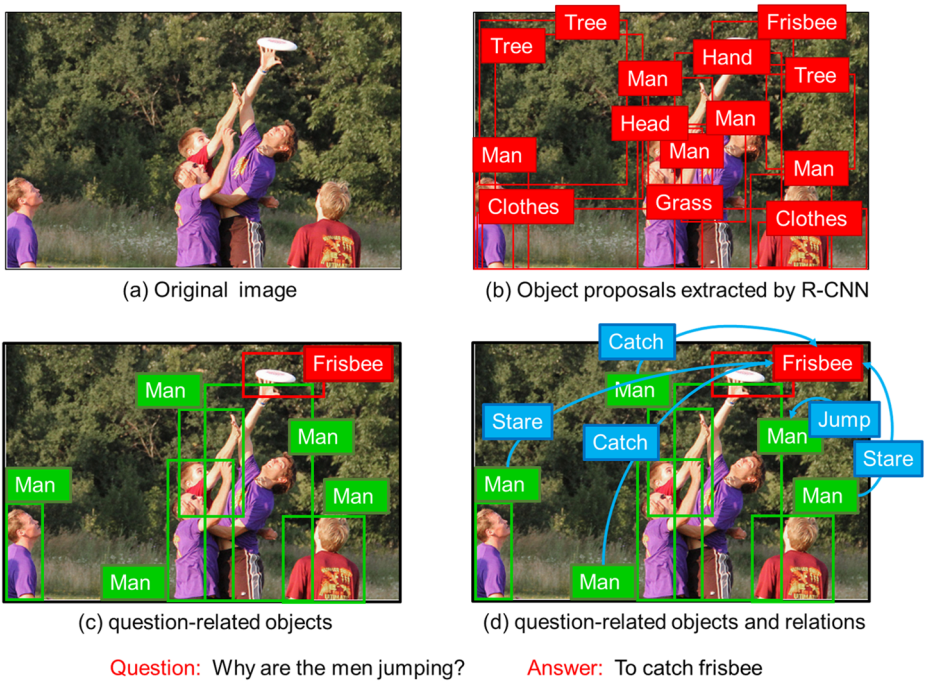

Fig. 1 The question-related objects and relations. Given a question "Why are the men jumping?" and an image as shown in the top left picture (**a**), our goal is to answer the question with "To catch frisbee". Picture (**b**) displays a part of object proposals extracted by Faster R-CNN [1], (**c**) is the question-related objects, (**d**) shows the question-specific objects as well as the relations. The red and green boxes indicate objects while the blue boxes are semantic relationships between objects

Although existing graph-based VQA models have achieved promising results, they still suffer from the following two limitations. Firstly, the graph nodes (object bounding boxes features) generated by existing object detection models may be redundant and not optimal for specific questions, as shown in Fig. 1(b). Secondly, the human-defined edges [40] or the automatically learned edges defined by the similarity between objects fail to model the actual semantic relationships contained in specific questions, and even introduce noisy information. In practice, human process of predicting the answer of a given question for an image typically consists of two steps: 1) identifying and localizing question-related objects regions as shown in Fig. 1(c) and 2) comparing these identified objects one by one to extract question-adaptive relationships between them, as shown in Fig. 1(d), while this property is often neglected by previous graph-based methods. From this perspective, an ideal input graph of the graph-based model should be question-specific, which is able to accurately focus on the question-related objects as well as question-adaptive relationships.

In this paper, we propose a novel graph-based approach for VQA by localizing question-related objects and modeling question-adaptive relationships. To reduce the object redundancy problem mentioned in the first limitation, we explicitly incorporate a soft-attention layer in the graph convolution process to highlight those question-related objects. Our attention mechanism can perform soft selections of objects depending on the question. Notably, here we do not directly apply a hard-selection strategy to filter the question-unrelated objects, since the question-unrelated objects may be answer-related. As for the second limitation, unlike previous approaches [23, 38] which build an edge between two objects according to their similarity in the vector space, we argue that the difference between objects should be taken into account when generating adjacency matrix for the following two reasons: 1) the influence of question-related objects can be boosted by comparing the differences between one object and the others under the given question. 2) the semantic relationship between two objects can be understood as a transformation of the difference between two objects, as suggested in TransE [3]. To this end, we propose a difference-based graph learner, in which the semantics of question and the difference between objects are taken into account to establish the question-adaptive connections between nodes in the graph.

Based on the soft-attention mechanism and the learned graph, the graph convolutional networks(GCN) [12] is utilized to encode the input image into a high-level representation by summarizing the structural dependencies of objects under the guidance of the target question. Experimental results show our model outperforms other graph-based model by a significant margin on the public benchmark VQA 2.0 dataset and is proved to be more interpretable. In summary, our contributions can be summarized as follows:

- We propose a novel graph-based approach for visual question answering, which is capable of modeling inter-object relationships by comparing objects explicitly under the guidance of questions.
- To handle the problem of object redundancy, we explicitly incorporate a soft-attention layer in the graph convolution process to signify question-specific objects.
- Our experimental results on the VQA 2.0 dataset demonstrate that our approach gives significantly better performance than baseline methods. Moreover, we conduct qualitative analysis to understand how the GCN module works with the help of attention mechanism, as well as the evaluation of the question-specific graph.

The rest of the paper is organized as follows. In Section 2, we introduce the related work. In Section 3, we present the fundamental idea of our model by showing how to construct a

question-specific graph and how the attention mechanism works. In Section 4, we present the experimental results and discuss the ablation study. Finally, we conclude the paper with remarks on our future direction in Section 5.

2 Related work

2.1 Visual question answering

Visual Question Answering(VQA) has aroused hot interest in both computer vision and natural language processing community since it involves not only natural language questions but also visual images. This task is comprised of multi sub-problems over multimodal data, including object detection, question understanding, visual relationships discovery, and cross-modal fusion and classification [4, 15, 28]. Existing works [2, 8–10, 14, 16, 18, 20, 32, 33, 35, 37] regard the answer prediction process as a classification task with three modules, in which the first two modules are designed to learn question and image representations by LSTM and CNN respectively, and then the third module fuses these two representations into a single multimodal vector followed by a dense layer for answer prediction.

Yang et al. [37] first attempt to incorporate attention layers into the VQA task, aiming to focus on the salient regions rather than the whole image for more accurate prediction. Following, a large variety of attention-based variations [1, 19, 21, 27, 30, 35] including fine-grained attention and co-attention have been proposed, and these attention based models have dramatically improved the performance of visual question answering.

Recently, owing to the flourish of the knowledge graph, the knowledge-based models [13, 14, 22, 31, 32, 34] using external knowledge beyond standard VQA datasets begin to emerge, which have achieved success in dealing with common-sense questions. They not not rely on the given images and questions, but also require external common-sense knowledge.

2.2 Graph-based visual question answering

Although the existing attention-based and knowledge-based VQA models have achieved promising performance on many standard datasets, these models do not resolve all the problems, especially in handling complex questions such as the “counting” and “reasoning” questions. To this end, Teney et al. [29] firstly use a graph-based representation for image and question, in which the image is initially presented as a scene-graph while question as a parse tree. Then the scene-graph is embedded into an image feature vector containing structural information, and the sparse tree is transformed into a question feature vector with syntactic information. However, it is only valid on small “abstract scenes” dataset and it is highly engineered. Zhang et al. [40] firstly utilize graph model aiming to solve the “counting” problems by comparing the inter and intra-class features, but this model also highly relies on the engineered relations.

Recently, graph convolution networks (GCN) is widely used in modeling non-structured irregular data for node classification or link prediction in many domains, such as social networks [12], 3D mesh modeling [17], visual relationsip understanding [39] and even biological networks [26]. Inspired by the properties of modeling entities and their relationships in a graph, Narasimhan et al. [22] firstly try to use GCN to answer factual questions based on the knowledge graph. This model mainly concentrate on the entity-relation graph extracted from the image and knowledge graph. However, it heavily relies on the domain-dependent

external knowledge graph. Norcliffe-Brown et al. [23] directly use GCN to infer answers based on the graph constructed by the bounding boxes extracted by Bottom-up [1], and it is claimed that they could model the relationships between objects. However, the graph they constructed is just based on the similarity between bounding boxes, which is not question-specific, and ignores the diversity of relationships between image objects. Different from them, we build the object relation graph by taking the difference between objects into consideration and assign different importance of each object based on its correlation to the question. Unlike v-AGCN [38] who uses pre-extracted visual relationships as prior knowledge to model the objects as well as their interactions, our framework can directly model objects as well as their relationships without any prior knowledge or pre-training, and our question-adaptive object relation graph can be learned end-to-end.

3 Method

Given an input image and a pertinent question, the VQA system is to predict the probability distribution of each candidate answer and assign the maximum one to the question as the final answer. In this paper, we propose a novel graph-based model for visual question answering, in which an object-difference based graph learner is incorporated into Graph Convolution Networks to capture the relationships between objects under the guidance of question. The overview of our approach is illustrated in Fig. 2. First, the input images and questions are encoded into high-level representation by Faster R-CNN [25] and GRU [6] in the data embedding module, respectively. Second, the encoded images will be transformed into question-adaptive object relation graphs under the guidance of questions in our object-difference based graph learner module. Third, the learned graph will be fed into the attention graph convolution layer to update the representation of objects and visual relationships. Finally, we make answer predictions based on the learned graph representation. The proposed

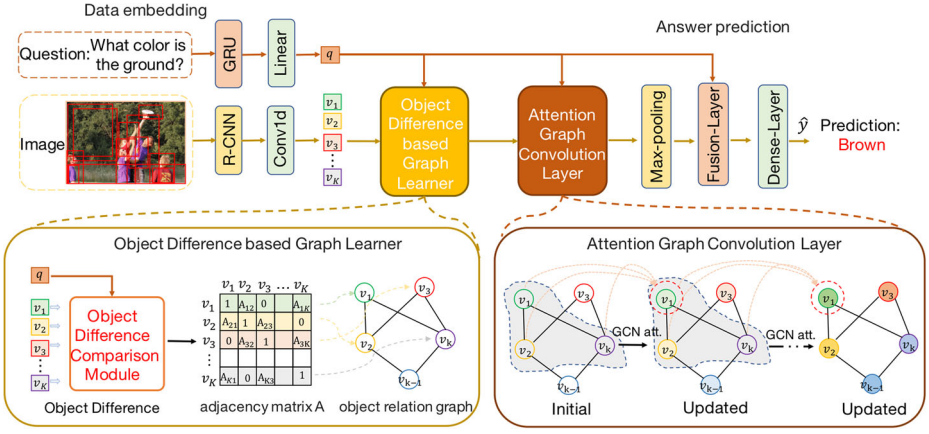

Fig. 2 The framework of our proposed model mainly consists of four components: Data embedding, Object-Difference based Graph Learner, Attention Graph Convolution Layer, and Answer Prediction. Faster R-CNN and GRU are adopted to encode image and question into high-level feature embeddings, Object-Difference based Graph Learner aims to model question-specific relations among objects, the goal of Attention Graph Convolution Layer is to reason on the learned graph to get the question-specific representation which will be fed into Answer Prediction module to predict answers

model is optimized to convergence by minimizing the multi-class cross-entropy loss. In the following, we will describe each of these components in details.

3.1 Data embedding

Concerning a standard visual question answering system, the input data includes a natural language question q and a visual image I. Correspondingly, the representation used to predict answer should contain question and image information. Follow existing general method, we first learn the image representation and sentence embedding separately and then project them into a shared space.

For image, we extract image features from Bottom-Up & Top-Down attention model [1]. Specifically, we choose the top-K identified objects with its associated region visual features $\mathbf{V}^f = \mathbf{v}_{i\,i=1}^{K}$ and their spatial feature $\mathbf{C} = \mathbf{c}_{i\,i=1}^{K}$, where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ denotes the feature vector of the i^{th} object and $\mathbf{c}_i = [x_i, y_i, w_i, h_i]$ corresponds to the spatial position of the i^{th} object, in which (x_i, y_i) is the bottom left point's coordinate of the i^{th} bounding box, while w_i and h_i are the width and height respectively. Note that the object feature vectors are fixed while training our model.

For question, we adopt GLOVE word embedding [24] to transform each word into a vector and feed these word vectors to GRU [6] to obtain question representation $\mathbf{q}^f \in \mathbb{R}^{d_q}$. What is notable, all the questions are padded and truncated to the same length.

To map visual features and question features into a shared space, we employ a fully-connected layer and a one-dimensional convolution layer to transform question features and image features respectively as shown in (1)~(2).

$$\mathbf{V} = \mathcal{R}elu\left(Conv1d\left(\mathbf{V}^f\right)\right) \tag{1}$$

$$\mathbf{q} = \mathcal{R}elu\left(Linear\left(\mathbf{q}^f\right)\right) \tag{2}$$

where $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{K} \in \mathbb{R}^{K \times d}$ can be viewed as the visual feature of K objects and $\mathbf{q} \in \mathbb{R}^d$ is question feature.

3.2 Object-difference based graph learner

Graph model, as an essential instrument in interpreting complex relationships between entities, and reasoning about potential facts, has stimulated a mass of researches in different fields, such as social networks, knowledge graph and recommendation system. Due to these inherent properties of the graph model, it is capable of helping visual question answering model to answer or infer complex questions. Although the graph model has been used by [23] to improve the interpretability of the VQA model, they only consider the similarity between two objects when building an adjacency matrix, whose relation is monotonous and not question-specific.

Motivated by the multi-relational embedding model TransE [3] in knowledge bases, we can model the semantic relationship between two objects based on their difference. As suggested in TransE, the relationships between entities are presented as translations in a low-dimensional embedding space, and its primary assumption is the embedding of tail entity should be close to the embedding of the head entity plus their relation's vector. That means the relationship between two entities can be presented as the difference vector between the head and tail entity.

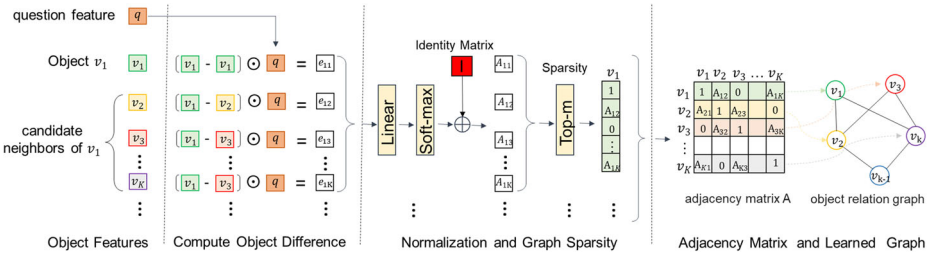

Fig. 3 The details of the object-difference graph learner. Each object $\mathbf{v}_i$ is compared with all the other objects $\mathbf{v}_j \in \mathbf{v}_{k=1}^{K} \backslash \mathbf{v}_i$ under the guidance of question. $\odot$ denotes element-wise multiplication. What's notable, we plus an identity matrix $\mathbf{I}$ on the adjacency matrix after softmax to ensure that each object $\mathbf{v}_i$ has the strongest correlation with itself. Finally, we select the top-m neighbors of each object $\mathbf{v}_i$ for sparsification

To this end, we propose a novel object-graph construction strategy enhanced by object difference comparison, to generate a graphic representation of the input image conditioned on the specific question. The learned graphic structure will be passed into the graph convolution layer in Section 3.3, in which each node gathers and compiles information from its neighboring nodes according to the adjacency matrix.

An ideal object graph should be question-specific, and its edges should accurately stand for the interactions between nodes while nodes represent objects. Thus, we aim to construct a weighted directed question-specific graph $G = \mathcal{V}, E$, whose nodes $\mathcal{V} = \mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K$ are the feature vectors of the objects in the input image detected by Faster R-CNN in Section 3.1, while the edges E represent relationships between objects as well as their relevance to the given question. The details of the object-difference based graph leaner are exhibited in Fig. 3. Inspired by [35], the question specific relation between two objects can be learned by comparing them under the guidance of the question $\mathbf{q}$. Specifically, each object $\mathbf{v}_i$ is compared with all the other objects $\mathbf{v}_j \in \mathbf{v}_{k=1}^{K} \backslash \mathbf{v}_i$ under the guidance of question $\mathbf{q}$. Therefore, the adjacency matrix $\mathbf{A}$ which models the relationships of objects can be defined as (3).

$$\mathbf{A}_{ij} = \left[\left(\mathbf{v}_i - \mathbf{v}_j \right) \odot \mathbf{q} \right]_d \mathbf{W}_p \tag{3}$$

where each element $\mathbf{A}_{ij}$ in the adjacency matrix indicates the relationships between i-th object $\mathbf{v}_i$ and the j-th object $\mathbf{v}_j$ under the guidance of $\mathbf{q}$. $\odot$ denotes element-wise multiplication, $\mathbf{W}_p \in \mathbb{R}^{d \times 1}$ is a learnable parameter vector that transforms the comparison result $(\mathbf{v}_i - \mathbf{v}_j) \odot \mathbf{q}$ into scalars, notably, $\mathbf{W}_p$ is shared with other object pairs.

As we can see in (3), the diagonal elements of adjacency matrix $\mathbf{A}$ are zeros, which indicates the relation to the object itself have been neglected in such definition. For that purpose, we add an identity matrix $\mathbf{I}$ after the softmax operation on the original adjacency matrix $\mathbf{A}$ as shown in (4), since the object keeps the strongest correlation with itself comparing with other objects.

$$\mathbf{A} = softmax(\mathbf{A}) + \mathbf{I} \tag{4}$$

We also impose restrictions on the graph sparsity follow [23], since a visual question is related to a small subset of all the objects and relationships in a visual scene. Thus, the question-specific graph should be extreme sparse which only focus on the most relevant objects and relationships. Specifically, for each node i, we only choose the top-m most relevant objects as well as its neighbors $N(i) = Top m(\mathbf{A}_i)$, and reserve the corresponding relationships.

3.3 Soft-attention based graph convolution layer

When people asked to answer a question, they are likely to pay more attention to the salient objects related to the question rather than all the objects in the image, and then interpret the interactions among these objects. Motivated by this, we introduce a graph convolution layer enhanced by a soft attention mechanism as shown in Fig. 4 to produce a question-related graphic representation of the image, where the question-irrelevant objects in the graph would be assigned negligibly small weights. Consequently, these question-irrelevant objects in the graph will be nearly neglected as the layers of graph convolution deepens. We define the object-question attention as in (6), in which the contribution weight α_i of the i^{th} object is calculated by dot product between image objects vector $\mathbf{v}_i$ and question vector $\mathbf{q}$ after normalized by softmax.

$$\alpha_i = \frac{exp\left(\mathbf{v_i^T q}\right)}{\sum_i exp\left(\mathbf{v_i^T q}\right)} \tag{5}$$

where $\mathbf{v}_i$ indicates i^{th} object feature vector, $\mathbf{q}$ indicates question embedding.

Based on the question-specific object graph constructed in Section 3.2 and the attention mechanism mentioned above, we can learn object representations by spatial graph convolution networks [36]. As discussed in Section 3.2, the graph nodes correspond to objects in the image, and the edges are the relationships between the connected objects. Same to [23], we also inject object position information into graph node in the form of polar coordinate vector generated by a pseudo-coordinate function $\mathbf{u}(\mathbf{c}_i,\mathbf{c}_j)$. The polar coordinate vector describes the relative spatial positions of object i and object j.

We use a set of N Gaussian kernels with learnable means and covariances as patch operator describing the influence of each neighborhoods, where the mean is interpretable as a direction and distance in pseudo coordinates. Unlike [23], we additionally consider the edge weight and the attention weight when summarizing neighborhood information for a node. For each Gaussian kernel n, the patch operator for node i is defined as:

$$f_n(i) = \sum_{j \in N(i)} w_n\left(\mathbf{u}\left(\mathbf{c}_i, \mathbf{c}_j\right)\right)\mathbf{v}_j A_{ij} \alpha_j, \qquad n = 1, 2, ..N \tag{6}$$

where $w_n(\mathbf{u}(\mathbf{c}_i,\mathbf{c}_j))$ is the n-th Gaussian kernel weight (relative spatial position information) of the j-th neighbor for node i, and $f_n(i) \in \mathbb{R}^d$ is the aggregated information of node i from its

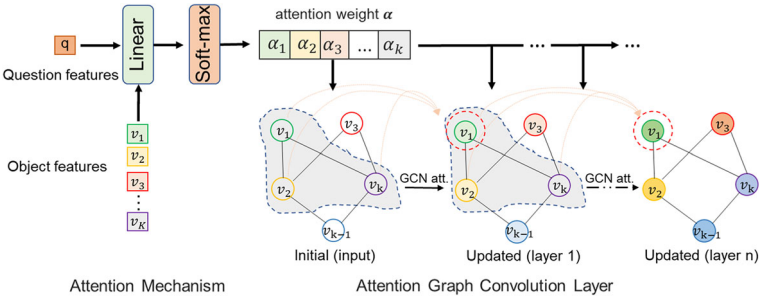

Fig. 4 The details of the soft-attention mechanism in graph convolution layer. For each object $\mathbf{v}_i$, we first compute its attention weight α_i based on the question q. Then these attention weights will be assigned to the object relation graph, and direct the graph convolution operation

neighbors including itself in the n-th kernel. $\mathbf{v}_j$ is the representation of the j-th neighbor of node i, $\mathbf{A}_{ij}$ is the edge weight between node i and node j, and α_j is the attention weight of object j. We then concatenate all patch operator outputs $\{f_1(i),f_2(i),...,f_N(i)\}$ of all the N kernels as the convolution operation result $f(i)$ for node i. Consequently, we can say that we not only incorporate the relationship between objects but also taking the object attention weight into consideration.

3.4 Answer prediction

The visual question answer prediction can be regarded as a multi-class classification problem, where each class corresponds to one of the most common answers in the training set. We use max-pooling on the graph nodes to get final image representation $f^{'}$, and then fuse the question feature and image representation by element-wise multiplication followed by two fully-connected layers to predict the score $\mathbf{y}$ on candidate answers as shown in (7).

$$\mathbf{y} = \delta\left(\left[f^{'} \odot \mathbf{q}\right]\mathbf{W}_c\right) \tag{7}$$

Finally, cross-entropy loss function is adopted as the objective function for training the VQA system as below:

$$L(\mathbf{t},\mathbf{y}) \quad = \sum_i \left(\mathbf{t}_i log\frac{1}{1+exp(-\mathbf{y}_i)} + (1-\mathbf{t}_i) log\frac{exp(-\mathbf{y}_i)}{1+exp(-\mathbf{y}_i)}\right) \tag{8}$$

where $\mathbf{t}_i = \frac{number\ of\ votes\ for\ answer\ i}{n}$ is the ground truth soft-score of each class, n is the valid provide answers associated with each question.

4 Experiment

4.1 Dataset and baselines

Dataset: The proposed method is evaluated on the commonly used benchmark dataset VQA 2.0 [7]. This dataset has 1105904 samples consisting of 204721 images, 332793 questions, and 29332 answers. Each sample includes an image, a question, and 10 human annotated answers. Specifically, for each question in VQA 2.0, there are two complementary images with similar visual scene and different answers. The VQA 2.0 dataset is divided into three splits: 40.1%,19.4% and 40.5% for training, validation, and test respectively.

Evaluation Metrics: We use the official evaluation metric proposed in [2] to evaluate our model as denoted below:

$$Acc(ans) = min\left\{\frac{\#\ number\ of\ ans\ human\ voted}{3}, 1\right\} \tag{9}$$

Baselines: We compare our object-difference based VQA models with two kind of VQA approaches: non-graph based models which use attention mechanism and graph based models which primarily depend on graph structure. In the

following list, the first three methods are non-graph based methods which use conventional CNN or attetion mechanism to predict answers, and the latter two are based on graph representation. All the baselines are listed as follows:

1) ReasonNet [8] learns to reason over a learned multifaceted image representation conditioned on text data. 2) Bottom-Up [1]is the classical visual question answering model with attention and the pre-trained model to extract bounding boxes, which has won the 2017 VQA challenge 3) Learn-Count [40] focuses on optimising counting questions based on human-defined object relation graph. 4) Graph-Learner [23] directly constructs object relation graph based on their simlarities, and then employ spatial graph convolution networks to reason about answers. 5) v-AGCN [23] employs pre-extracted visual relation as prior knowledge to transform input image into graph and then use an anisotropic graph convolution module for relational reasoning.

4.2 Implementation details

Following the popular choices of most previous works, all questions are padded and truncated to the same length 14 and we consider the 3000 most common answers in the train set as possible answers for our network to predict. The input images are encoded as a set of 36 object bounding boxes with corresponding 2048-dimensional feature vectors as described in Bottom-Up & Top-Down [1]. These strategies are also used by the methods that we compare against. We use the 300 dimension Glove embeddings [24] to initialize word embeddings. For question encoder, we use a dynamic Gated Recurrent Unit (GRU) with a hidden state size of 1024. Our graph learner, which learns the adjacency matrix, comprises two dense linear layers of size 1024. For GCN, we set the hidden dimension size to 1024 and use two-layer graph convolution layers, both layers have 8 Gaussian kernels. All dense layers and convolutional layers are activated using ReLU. Dropout with $p = 0.5$ is used after the input layer and before the classifier layer to prevent the proposed networks from overfitting. The training samples are shuffled randomly at each epoch, and batch size set as 64. Our model is trained using Adam optimizer [11] for 35 epochs with the initial learning rate of 0.0001 which is halved after the 30th epoch. All the hyper-parameters are tuned on the validation set.

4.3 Quantitative results

Table 1 shows the comparison results on the test-dev and test-standard splits of VQA 2.0 dataset. The top three rows show the performance of non-graph based models which do not involve relational reasoning mechanism, and the middle two rows display the results of graph based models specialized in dealing with complex or inference questions. And the next four rows correspond to the four variants of our model with different object feature combination mode in graph learner. Specifically, Sim-GCN computes the edge of two nodes depending on the similarity between two nodes, Cat-GCN means concatenation, Add-GCN denotes addition, and Ele-GCN performs elementwise multiplication. All the varieties adopt soft-attention

mechanism in spatial graph convolutional layer. The last row is our final model ODA-GCN which uses object defference based graph learner.

From the results on the test-standard set, we observe that: 1)Our model achieves the highest accuracy compared with other graph-based methods, which demonstrates the advantages of our model in relational reasoning. 2)As evident from the 0.25% increase in accuracy about "Other" category , our ODA-GCN model has been equipped with stronger ability in answering complex questions comparing with other graph-based models. That confirms the ability of the ODA-GCN to learn complex question-specific graph and complex relationships or interactions of the image objects reasoning. 3)There is also an improvement of 1.51% in counting question and 1.19% in "Yes/No" question, which serves as evidence of the contribution of the attention mechanism over GCN. 4)Our model achieves comparable performance with other non-graph based VQA models. What's notable, the model Learn-Count [40] is highly dependent on the engineered graph in their counting module and they have used 100 objects proposals in training, but our model is trained end-to-end without any engineered features and 36 objects proposals.

4.4 Ablation study

In this work, we have introduced two components for our ODA-GCN model: the object-difference based graph learner and soft-attention mechanism in the spatial graph convolution layer. To assess the contribution of each component, we perform an ablation study with different equipment of components on the validation set. In this section, we designed three extra models: 1)We first directly use Graph-Learner [23] as our baseline model denoted as Baseline since our model is predominately based on the backbone of this model but with a different graph learner and an additional attention mechanism. 2)We also design another model OD-GCN only utilizing object-difference based graph learner as the graph learner module, while other modules are identical to Baseline. 3)The last model OD-GCN reuses the graph learner and other settings mentioned in Baseline, except the soft-attention mechanism in spatial graph convolution layers. For a fair comparison, all the hyperparameters are identical to Baseline, and we train these four models on the training set and report the results on the validation set in our ablation study.

From the results shown in Table 2, we find that either using object-difference based graph learner(OD-GCN) or soft-attention mechanism in spatial graph convolution layers(A-GCN)

Table 1 Compared with state-of-the-art VQA models on the test-stand & test-dev splits of VQA 2.0 dataset

Method	Test-dev				Test-stand			
	Y/N	Num.	Others	All	Y/N	Num.	Others	All
ReasonNet [8]	–	–	–	–	78.86	41.98	57.39	64.61
Bottom-Up [1]	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
Learn-Count [40]	83.14	51.62	58.97	68.09	83.56	51.39	59.11	68.41
Graph-Learner [23]	–	–	–	–	82.39	45.77	56.14	65.77
v-AGCN [38]	82.39	45.93	56.46	65.94	82.58	45.12	56.71	66.17
Sim-GCN(baseline)	83.56	45.54	56.02	66.17	–	–	–	–
Cat-GCN(baseline)	83.51	47.06	56.52	66.56	–	–	–	–
Add-GCN(baseline)	83.49	46.58	56.57	66.53	–	–	–	–
Ele-GCN(baseline)	83.57	47.30	56.60	66.65	–	–	–	–
ODA-GCN(ours)	83.73	47.02	56.57	66.67	83.77	47.28	56.96	66.87

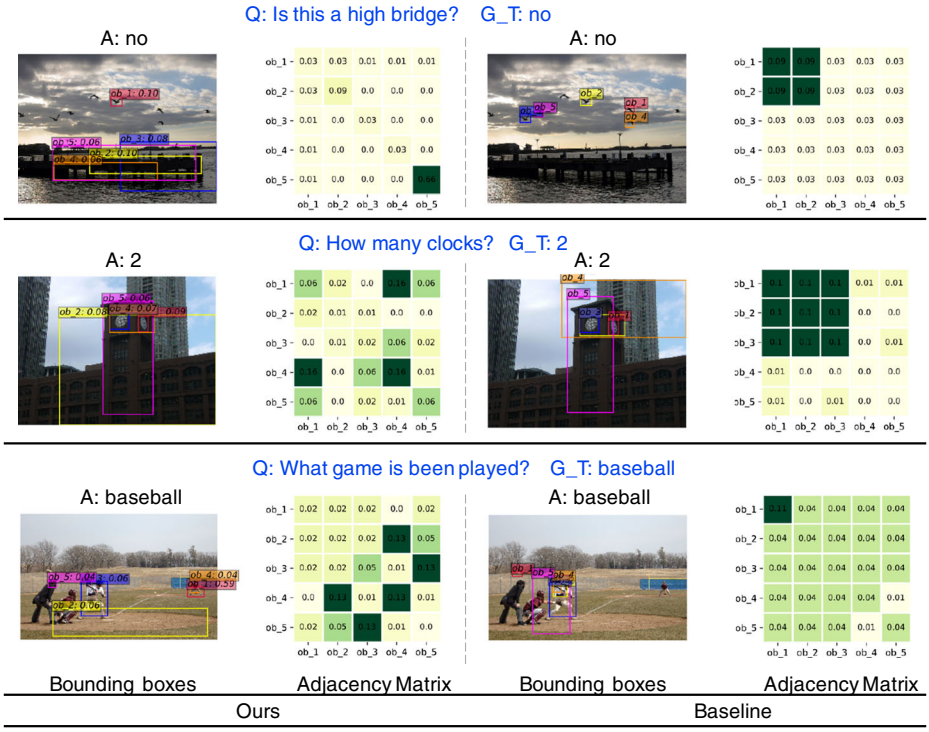

Fig. 5 Visualization of the adjacency matrix learned by our ODA-GCN model and Baseline. The columns from left to right are the learned question-related objects(with attention weights) by our ODA-GCN model, the adjacency matrix of these objects selected by ODA-GCN, the learned question-related objects by Baseline, the adjacency matrix of these objects signified by baselines. What's notable, the numbers in adjacency matrix are relation weights corresponding to object pairs. G_T is ground truth

can significantly outperform the baseline model on validation split, which means these two new modules are valid. What's more, our full model ODA-GCN simultaneously adopting both object-difference based graph learner and soft-attention mechanism, also achieved higher accuracy compared with OD-GCN and A-GCN, which demonstrates these two modules can reinforce each other.

4.5 Visualization analysis

In this subsection, we present visualization analysis to demonstrate that using our proposed object-difference graph learner and soft-attention mechanism can lead to more interpretability

Table 2 Performance on VQA 2.0 validation set for ablation study (Attent.: attention mechanism. Obj-Diff: Object-Difference Graph Learner)

Model	Attent.	Obj-Diff	Validation Acc
Baseline	No	No	63.44
A-GCN	Yes	No	63.67
OD-GCN	No	Yes	64.17
ODA-GCN	Yes	Yes	64.23

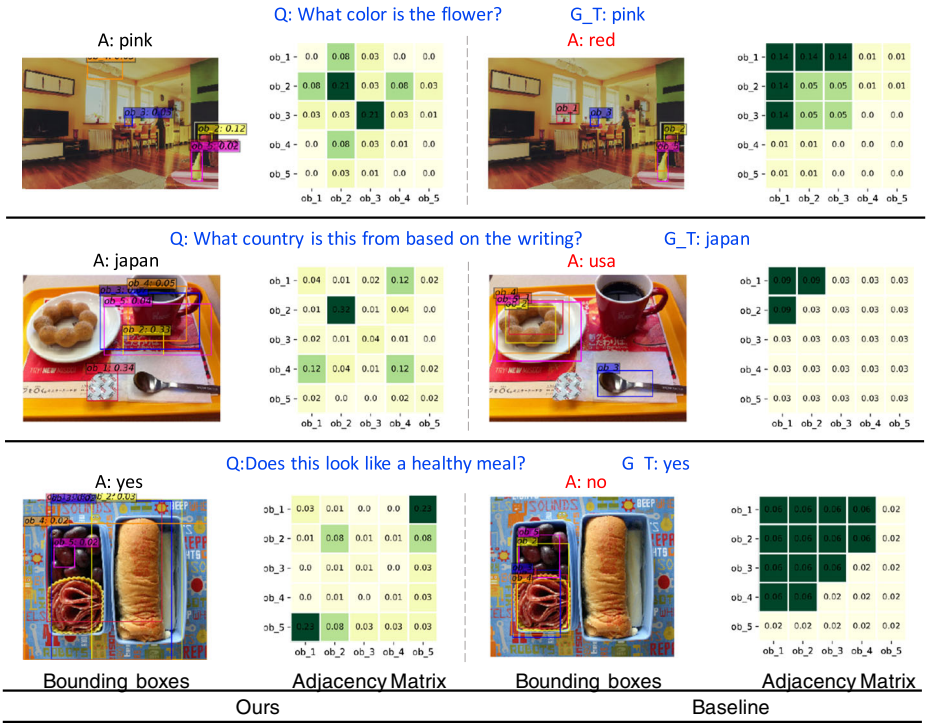

Fig. 6 Visualiztion of adjacency matrix. This figure shows the instances that our model predicted correctly while baseline failed

details that are relevant to the potential answers. We achieved this by visualizing the adjacency matrixes as well as the attention weights over each object learned by our object-difference based model, and we compared our model with Baseline mentioned in Section 4.4. We display the adjacency matrixes and attention weights learned by these two models on the validation split of VQA 2.0 dataset. Note the adjacency matrix of each image is of size 36 × 36, and the attention probability distribution is 36. For the sake of presentation here, we only display top-5 question-attended objects as well as their adjacency matrix down-sampled from the full adjacency matrix of each instance according to the attention weights.

Figure 5 presents some successful examples of both models, and Fig. 6 shows the instances that our model have predicted correctly, but Baseline failed. These instances cover all the three kinds of questions, including "Yes/No", "Num." and "Other". Specifically, the columns from left to right are the five objects most relevant to the question learned by our ODA-GCN model, the adjacency matrix of these objects, the top-5 question-related objects learned by Baseline, and the adjacency matrix of these objects signified by Baseline. As we have mentioned, each element in the learned adjacency matrixes denotes the relation weight between its two corresponding objects. And we have observed that the diagonal elements are bigger than other elements that because we have added self-loop in our adjacency matrixes. Intuitively, the diagonal elements indicate the dependence on the object itself when answering a question.

For the "Yes/No" questions, we take the first row of Fig. 5 as an example. Although both models have given a correct answer, our model can not only focus on the question-related objects "high bridge", but also the relation "high" in the adjacency matrix, while the Baseline model only

focuses on the partial keyword "high" and neglecting relations. For the "counting" questions, as shown in the second row of Fig. 5, our model can model objects "ob_1", "ob_4", "ob_5" and their relations as shown in the adjacency matrix. For the "Other" question, the instance in the second row of Fig. 5 shows that both models have attended on the objects, but the relations learned by our model is more accurate than Baseline does. What's more, the adjacency matrixes generated by our model are sparser than Baseline did in all instances. Therefore, we can say our model can learn more interpretable question-adaptive relations and localize salient objects.

5 Conclusion

In this paper, we propose a novel graph-based approach for Visual Question Answering. It aims at constructing a question-specific relational graph by reducing redundant objects and revealing more accurate relations. To alleviate the influence of object redundancy, we explicitly incorporate a soft-attention layer in the graph convolution process to signify question-specific objects. Furthermore, comparing objects explicitly by the difference operator enables our model to identify which relation is exactly we want. Experimental results show that our model significantly outperforms baseline graph-based methods. In the future, we will research how to utilize more detailed relation types for specific questions reasoning.

Acknowledgements Zhendong Mao is the corresponding author. This work was supported by the National Key Research and Development Program of China (grant No. 2016QY03D0505) and the National Natural Science Foundation of China (grant No. U19A2057).

References

1. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
2. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D (2015) Vqa: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
3. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems, pp 2787–2795
4. Cheng Z, Ding Y, He X, Zhu L, Song X, Kankanhalli MS (2018) Aˆ 3ncf: an adaptive aspect attention model for rating prediction. In: IJCAI, pp 3748–3754
5. Cheng Z, Chang X, Zhu L, Kanjirathinkal RC, Kankanhalli M (2019) Mmalfm: explainable recommendation by leveraging reviews and images. ACM Trans Inform Syst (TOIS) 37(2):16
6. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder app.roaches. arXiv:https://arxiv.org/abs/14091259
7. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) 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, pp 6904–6913
8. Ilievski I, Feng J (2017) Multimodal learning and reasoning for visual question answering. In: Advances in neural information processing systems, pp 551–562
9. Kazemi V, Elqursh A (2017) Show, ask, attend, and answer: a strong baseline for visual question answering. arXiv:https://arxiv.org/abs/170403162
10. Kim JH, Lee SW, Kwak D, Heo MO, Kim J, Ha JW, Zhang BT (2016) Multimodal residual learning for visual qa. In: Advances in neural information processing systems, pp 361–369
11. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:https://arxiv.org/abs/14126980
12. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv: https://arxiv.org/abs/160902907

13. Li G, Su H, Zhu W (2017) Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. arXiv:https://arxiv.org/abs/171200733
14. Liao L, Ma Y, He X, Hong R, Chua Ts (2018) Knowledge-aware multimodal dialogue systems. In: 2018 ACM Multimedia conference on multimedia conference. ACM, pp 801–809
15. Liu AA, Nie WZ, Gao Y, Su YT (2016) Multi-modal clique-graph matching for view-based 3d model retrieval. IEEE Trans Image Process 25(5):2103–2116
16. Liu AA, Su YT, Nie WZ, Kankanhalli M (2017) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 39(1):102–114
17. Liu AA, Nie WZ, Gao Y, Su YT (2018) View-based 3-d model retrieval: a benchmark. IEEE Trans Cybern 48(3):916–928
18. Liu J, Zhai G, Liu A, Yang X, Zhao X, Chen CW (2018) Ipad: intensity potential for adaptive de-quantization. IEEE Trans Image Process 27(10):4860–4872
19. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical co-attention for visual question answering. Advances in Neural Information Processing Systems (NIPS), 2
20. Malinowski M, Rohrbach M, Fritz M (2015) Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision, pp 1–9
21. Nam H, Ha JW, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299–307
22. Narasimhan M, Lazebnik S, Schwing A (2018) Out of the box: reasoning with graph convolution nets for factual visual question answering. In: Advances in neural information processing systems, pp 2654–2665
23. Norcliffe-Brown W, Vafeias S, Parisot S (2018) Learning conditioned graph structures for interpretable visual question answering. In: Advances in neural information processing systems, pp 8334–8343
24. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
25. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
26. Shang C, Liu Q, Chen KS, Sun J, Lu J, Yi J, Bi J (2018) Edge attention-based multi-relational graph convolutional networks. arXiv:https://arxiv.org/abs/180204944
27. Shih KJ, Singh S, Hoiem D (2016) Where to look: focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4613–4621
28. Tang S, Li Y, Deng L, Zhang Y (2017) Object localization based on proposal fusion. IEEE Trans Multimed 19(9):2105–2116
29. Teney D, Liu L, van den Hengel A (2017) Graph-structured representations for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
30. Teney D, Anderson P, He X, van den Hengel A (2018) Tips and tricks for visual question answering: learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4223–4232
31. Wang P, Wu Q, Shen C, Dick A, van den Hengel A (2018) Fvqa: fact-based visual question answering. IEEE Trans Pattern Anal Mach Intell 40(10):2413–2427
32. Wu Q, Wang P, Shen C, Dick A, van den Hengel A (2016) 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, pp 4622–4630
33. Wu Q, Teney D, Wang P, Shen C, Dick A, van den Hengel A (2017) Visual question answering: a survey of methods and datasets. Comput Vis Image Underst 163:21–40
34. Wu Q, Shen C, Wang P, Dick A, van den Hengel A (2018) Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans Pattern Anal Mach Intell 40(6):1367–1381
35. Wu C, Liu J, Wang X, Dong X (2018) Object-difference attention: a simple relational attention for visual question answering. In: 2018 ACM Multimedia conference on multimedia conference. ACM, pp 519–527
36. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI conference on artificial intelligence
37. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
38. Yang Z, Yu J, Yang C, Qin Z, Hu Y (2018) Multi-modal learning with prior visual relation reasoning. arXiv: https://arxiv.org/abs/181209681
39. Yang X, Zhang H, Cai J (2018) Shuffle-then-assemble: learning object-agnostic visual relationship features. In: Proceedings of the European conference on computer vision (ECCV), pp 36–52
40. Zhang Y, Hare J, Prügel-Bennett A (2018) Learning to count objects in natural images for visual question answering. arXiv:https://arxiv.org/abs/180205766

Xi Zhu received the B.E. degree in computer science from Minzu University of China, Beijing, China. She is currently pursuing her Ph.D. degree from Institute of Information Engineering, School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. Her research interests include cross-modal understanding and image-text retrieval.

Zhendong Mao received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, in 2014. He is currently a professor with School of Information Science and Technology, University of Science and Technology of China, Hefei. He was an associate professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, from 2014 to 2019. His research interests include computer vision, natural language processing, and artificial intelligence, and in particular, cross-modal representation and understanding.

Zhineng Chen received the M.Sc and B.Sc degrees in computer science from the College of Information Engineering, Xiangtan University, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2011. He is now an associate professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was a senior research associate with the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2012, and was an assistant professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, from 2012 to 2014. He has published over 40 academic papers in prestigious journals and conferences. His research interests include large-scale multimedia analytics, medical image analysis and pattern recognition.

Yangyang Li received his Ph.D. in Computer Science from Beijing University of Posts and Telecommunications in 2015, and his B.S. degree from Nanjing University of Information and Technology in 2009. He is now with Mobile Internet Development and Research Center, China Academy of Electronics and Information Technology. His research interest includes mobile internet, data center network, and edge computing.

Zhaohui Wang received the B.E. degree in communication engineering from University of Science and Technology Beijing, Beijing, China. She currently pursuing her Master degree from Institute of Information Engineering, School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. Her research interests include cross-modal understanding.

Bin Wang received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is currently the director of Xiaomi AI Lab and a guest professor in the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. His research interests include information retrieval and natural language processing.

Affiliations

Xi Zhu[1,2] · Zhendong Mao[3] · Zhineng Chen[4] · Yangyang Li[5] · Zhaohui Wang[1,2] · Bin Wang[6]

Xi Zhu
zhuxi@iie.ac.cn

Zhineng Chen
zhineng.chen@ia.ac.cn

Yangyang Li
yli@csdclab.net

Zhaohui Wang
wangzhaohui@iie.ac.cn

Bin Wang
wangbin11@xiaomi.com

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

[3] University of Science and Technology of China, Hefei, China

[4] Institute of Automation, Chinese Academy of Sciences, Beijing, China

[5] China Academy of Electronics and Information Technology, Beijing, China

[6] Xiaomi AI Lab, Xiaomi Inc., Beijing, China