Knowledge-Based Systems 203 (2020) 106150

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Cross-modal learning with prior visual relation knowledge

Jing Yu [a,c], Weifeng Zhang [b,*], Zhuoqian Yang [d], Zengchang Qin [d], Yue Hu [a,c]

[a] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[b] College of Mathematics, Physics and Information Engineering, Jiaxing University, Zhejiang, China
[c] School of Cyber Security, University of Chinese Academy of Sciences, China
[d] Intelligent Computing & Machine Learning Lab, School of ASEE, Beihang University, China

ARTICLE INFO

Article history:
Received 29 January 2020
Received in revised form 11 May 2020
Accepted 14 June 2020
Available online 16 June 2020

Keywords:
Visual relation reasoning
Relation embedding
Anisotropic graph convolutional networks
Visual question answering
Cross-modal information retrieval

ABSTRACT

Visual relational reasoning is a central component in recent cross-modal analysis tasks, which aims at reasoning about the visual relationships between objects and their properties. These relationships provide rich semantics and help to enhance the visual representation for improving cross-modal learning. Previous works have succeeded in modeling latent visual relationships or rigid-categorized visual relationships. However, these kinds of methods leave out the problem of ambiguity inherent in the visual relationships because of the diverse relational semantics of different visual appearances. In this work, we explore to model the visual relationships by context-aware representations based on human prior knowledge. Based on such representations, we novelly propose a plug-and-play visual relational reasoning module to enhance image encoding. Specifically, we design an Anisotropic Graph Convolution to utilize the information of relation embeddings and relation directionality between objects for generating relation-aware image representations. We demonstrate the effectiveness of the relational reasoning module by applying it to both Visual Question Answering (VQA) and Cross-Modal Information Retrieval (CMIR) tasks. Extensive experiments are conducted on *VQA 2.0* and *CMPlaces* datasets and superior performance is reported when comparing with state-of-the-art works.

1. Introduction

Vision and natural language are most typical modalities for human to describe the real world. Large amount of information on the webpages, social networks, E-commerce website is conveyed by both visual and textual content. With the advances in Computer Vision (CV) and Natural Language Processing (NLP), researchers make a further step towards breaking the boundary of vision and natural language, such as visual question answering (VQA) [1–3], cross-modal information retrieval (CMIR) [4–6], image captioning [7,8], *etc.* All these tasks require fine-grained visual processing, or task-specific visual reasoning to obtain semantic-rich visual representation, which is a top priority for Artificial Intelligence (AI) to achieve human-like ability.

Much effort has been made in literature to enhance the visual representation for learning tasks. Research on heterogeneous transfer learning [9,10] has exploited to facilitate the representation in the visual domain by borrowing knowledge from the heterogeneous textual domain. Another important research direction leverages multi-view learning [11] to comprehensively describe the visual information by jointly considering different views, which brings significant performance improvement on cross-modal problems [12,13]. Both of the above research directions focus on enriching the global visual representations, which ignore the essential relational visual semantics. One of the recent advances in visual representation for cross-modal learning is visual relational reasoning [14]. It aims to reason about the interaction relationships (i.e., *wearing*, *holding*, *riding*), positional relationships (i.e., *above*, *below*, *inside*, *around*), or even latent relations between visual objects in an image. The relationships can be formally defined as triples $\langle subject, predicate, object \rangle$, i.e. $\langle woman, riding, horse \rangle$ or $\langle boy, kicking, ball \rangle$. Such visual relationships are working in conjunction with deep neural networks to generate relation-aware visual representation. State-of-the-art works have proved that visual relationships is crucial to improve the performance of cross-modal learning tasks [7,14] since a significant portion of cross-modal analysis requires visual relational reasoning ability as illustrated in Fig. 1.

However, modeling the relations can be a challenging problem. Previous work [7] primarily studies appearance-based models to detect visual relations categorically — they learn the relations as a classification task and output the rigid-divided category for each *predict* as the corresponding relation. Such relation information is typically integrated into deep architectures by using specific parameters for each category. Unfortunately, the visual appearances are too varied and semantic-rich to be modeled by rigid-divided categories. On one hand, the visual appearance of

* Corresponding author.
E-mail address: zhangweifeng@zjxu.edu.cn (W. Zhang).

https://doi.org/10.1016/j.knosys.2020.106150
0950-7051/

Fig. 1. Visual Question Answering (VQA) samples that requires visual relational reasoning, including (a) "YES/NO" question with comparison relationship, (b) "Number" question with interaction relationship, and (c) "Object" question with interaction relationship.

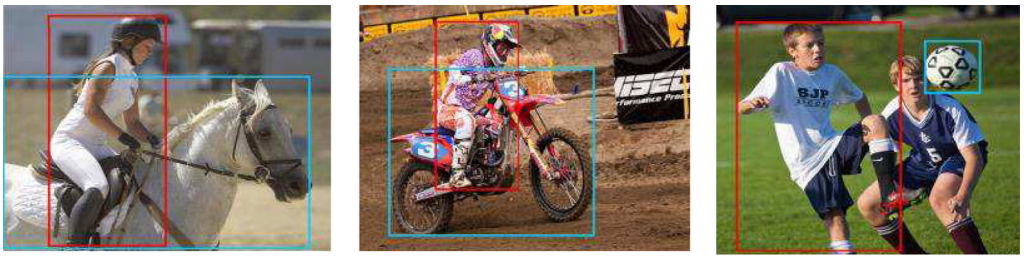

Fig. 2. Ambiguous relation samples with varied appearances. (a) and (b) are two samples with distinct visual semantics while belonging to the same relation "ride". (c) is the sample that implicates two relations, i.e. *kick* and *look at*, by the same appearance.

the same relation vary significantly because of different *object* and *subject* involved. For example, the appearance of the relation *ride* can be quite various in the following two scenes: "*a woman rides horse*" in Fig. 2(a) and "*a man rides motorcycle*" in Fig. 2(b). On the other hand, the same appearance could implicate diverse relations. Take Fig. 2(c) for example. There exists two kinds of relations, i.e. *kick* and *look at*, between the boy in white T-shirt and the football. In summary, the rigid *predict* categories can hardly convey the diverse relational semantics of different appearances.

In this paper, we bypass this challenge by describing each relation in an image as a relation embedding, conditioned on the objects involved in the relationships. Compared with previous work of relationship detection [7], our proposed relation embedding can be more effective and accurate to model the fine-grained semantics inherent in the *object*, *subject* and their interaction. Furthermore, we novelly propose a plug-and-play relation reasoning module based on graph convolutional networks, which explores the use of relation embedding to enhance image encoder for multimodal learning. As shown in Fig. 3, our basic design is to model the image as a directed semantic graph. Each node represents an object (salient region) detected by Faster R-CNN while each edge denotes the relation between two objects. The relation is represented by relation embedding obtained by relation encoder pre-trained on Visual Genome [15]. In our model, we apply graph convolution network to enrich the representation of each object by its relation-essential neighborhood. In this way, the representation of each object involves the contextual information and becomes more accurate to convey the semantics of the object in an image. Specifically, we propose an attention-based graph convolution module, named as Anisotropic Graph Convolution (AGConv). AGConv first examines all the relationships for an object to highlight the importance of different relationships to the object by attention weights. Then AGConv updates each object's representation based on the importance of the relationships by aggregating information from its neighborhood and the corresponding relationships. Finally the relation-aware object representations are injected into multimodal learning model for the downstream task. We evaluate our proposed relation reasoning module on two representative multimodal learning tasks: Visual Question Answering and Cross-Modal Information Retrieval. Experimental results demonstrate the superior improvements by integrating our module into the state-of-the-art methods, verifying the benefits of using the relation embedding and reasoning module.

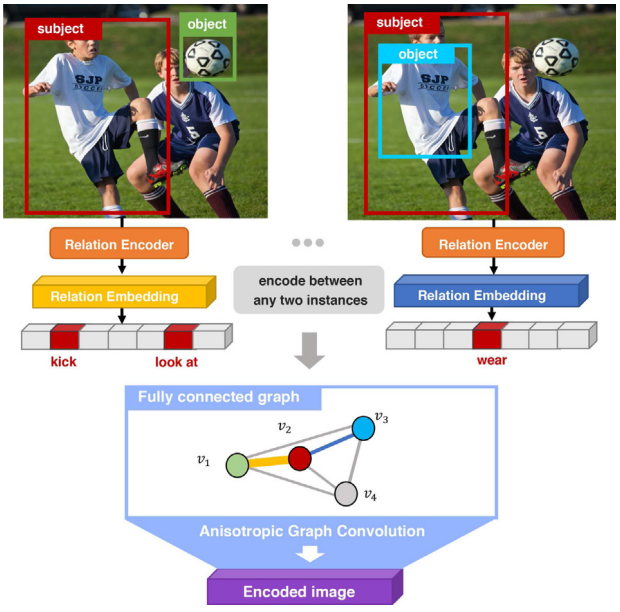

Fig. 3. Illustration of our module for visual relational reasoning. We first use an object detector to detect instances in images to form the nodes of the graph. Then a relation encoder is used to generate a relation embedding between each pair of instances. Finally, we use our anisotropic graph convolution to process information in the graph.

The rest of this paper is organized as follows: Section 2 briefly reviews the related work in visual relation reasoning, VQA and CMIR. Then we detail our proposed relation reasoning module, named as Anisotropic Graph Convolutional Network (AGConv), in Section 3. Two models integrated with AGConv are implemented for cross-modal reasoning and retrieval, and applied to VQA in Section 4 and CMIR in Section 5. We report the experimental results in Section 6 and conclude our work in Section 7.

2. Related work

2.1. Visual relation reasoning

Visual relation reasoning aims to represent and infer the relations among objects in an image, thus aggregating information from objects and their relations to enrich the image representation. It is a top priority for AI to achieve human-like intelligence [16]. How to model these relations is one of the major differences of the existing approaches. Early works [17] study the shallow geometric relations based on spatial information (e.g. *below*, *above*) to enhance visual segmentation. Later on, in [18], interactions (e.g. *wear*, *carry*) between paired objects are exploited, where visual relation reasoning is then formulated as a classification task. Afterwards, relationship are extended to richer definition [19], including geometric, comparative, composition, action, *etc*. The most recent works propose to model visual relations by scene graphs based on prior human knowledge [15] and effectively improve the performance of image captioning [20].

One limitation of the above approaches is that they merely represent the relations between two objects by rigid-categorized labels, failing to leverage the richer semantics inherent in the visual context and leading to ambiguity. For instance, Fig. 2(c) shows that multiple relationships (i.e. *kick* and *look at*) may exist between two objects and Fig. 2(a)(b) shows that even relationships in the same category (i.e. *ride*) may have diverse semantics. To solve this problem, quite a few works attempt to design deep models for inferring more complex relations among multiple objects. [14] proposes to infer relations between all the implicit object-like patch pairs via a plug-and-play MLP module for visual question answering. However, this method has only been proven to work on synthesized 3D datasets with primary geometrical objects, problems involving real-world objects and relations are yet to be better handled. [7] treats visual relations as labeled directional edges between two object nodes in the spatial and semantic graph and apply Graph Convolutional Networks (GCN) to reason about their implicit relations for image captioning. Although [7] makes relation reasoning sensitive to the relation types, it only applies different biases for "rigid-categorized" relation types and ignores the influence of the connected objects in the reasoning process. On the contrary, we attempt to learn relational embeddings to "softly" represent various relations and design a modified GCN module to support embedded relations for inferences seamlessly.

2.2. Visual Question Answering (VQA)

VQA aims to answer a question in natural language according to a natural image. A typical solution for VQA is to fuse visual and textual features for a joint representation and infer the answer based on the fused image-question representation. The most typical methods for feature fusion are element-wise summation/multiplication or direct concatenation. Besides straightforward solutions, several works apply bilinear pooling [21–23] or more complex fusion methods [24]. Noh et al. [24] explores a novel CNN model for feature fusion with a dynamic parameter layer whose weights are learned adaptively by the question.

However, the above approaches are based on global features of both images and questions, which fails to provide fine-grained information and possibly introduces noise. Several works adopt the attention mechanism to focus on semantically relevant image regions regarding a given question. Yang et al. [1] perform visual attention multiple times via stacked attention networks, and Anderson et al. [2] use a top-down attention on pre-detected salient regions. These models explore the fine-grained correlations between visual and textual content and eliminate noisy information. Recently, visual relation reasoning has been introduced into VQA and achieved better answers for questions that require a logical understanding of the question and the image [14]. These approaches mimic human thinking, which has not been thoroughly studied yet.

2.3. Cross-Modal Information Retrieval (CMIR)

CMIR is a task to enable queries from one modality to retrieve information in another modality. The typical solution for CMIR is to project the data from different modalities into a common semantic space to directly compare their similarity. Several statistical methods are based on Canonical Correlation Analysis (CCA) [25,26] to maximize the pairwise correlations. However, these methods ignore high-level semantic priority and could be hard to extend to large-scale data [27]. Another research trend is based on deep learning [4,5,28], leveraging existing techniques to provide rich semantics by nonlinear transformations. Typically, [5] proposes a two-branch neural network with two layers of nonlinearities on top of visual and textual features. [28] leverages attention mechanism to focus on essential image regions and words for correlation learning. Recently, [4] explores the relationship between words and prove the effectiveness for representing texts and eventually improve the CMIR accuracy.

In this paper, our method automatically detects and reasons visual relations for more informative image representations, then embeds them into the same semantic space with texts, where cross-modal similarity is directly computed.

3. Methodology

We propose a plug-and-play visual relational reasoning module, named as Anisotropic Graph Convolution (AGConv), for enhancing visual representations for cross-modal learning. To demonstrate the effectiveness of AGConv, we present two models equipped with AGConv for cross-modal learning: v-AGCN for visual question answering (Section 4) and c-AGCN for cross-modal information retrieval (Section 5). They share the common module of image modeling via AGConv but differ in their ways of associating visual and textual modalities. In this section, we introduce the architecture of AGConv. It mainly contains three parts: (1) Image representation (Section 3.1): each image is represented by a structured semantic graph with nodes indicating visual objects and edges indicating the visual relationships; (2) Relationship encoder (Section 3.2): the pre-trained relation encoder is leveraged to encode the relationship between each two objects as a relationship embedding, which is used to describe the semantics and directionality of edges in the constructed graph; (3) Anisotropic Graph Convolution (Section 3.3): this is the key module that reasons about the inner-group visual relationships among objects based on the ungraded graph attention networks and enrich the representation of each object according to its relationships with neighborhoods.

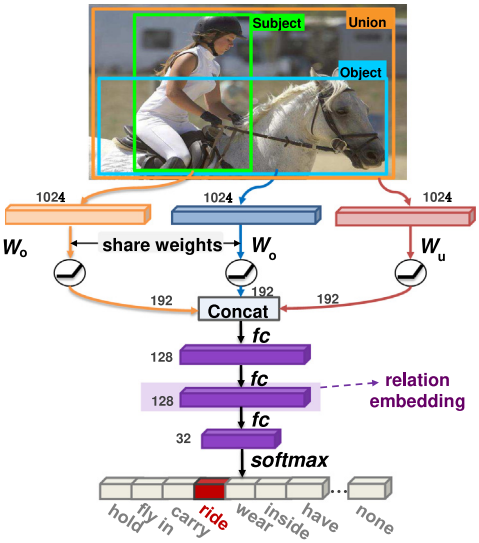

Fig. 4. Schematic illustration of the relation encoder.

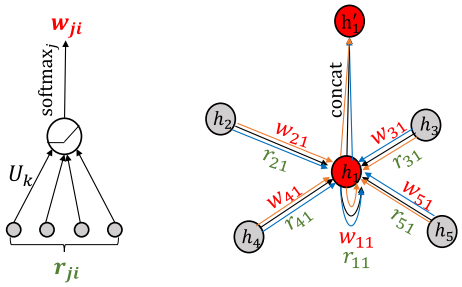

Fig. 5. Schematic illustration of the proposed anisotropic graph convolution with multiple attention heads. **Left**: the attention mechanism adopted in our model, where the attention distribution is calculated over all the relation embeddings r_{ji} ($j \in N$, $i \in N$) and parametrized by a weight vector U_k. **Right**: Illustration of multi-head anisotropic graph convolution of one central node over its neighborhood. Different arrow colors denotes distinct sets of attention parameters. The aggregated features from each attention head are concatenated to obtain the updated node feature h'_1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. Image representation

Each image is represented by a directed semantic graph. First, we use Faster R-CNN [29] in conjunction with the ResNet-101 [30] to detect salient image regions and output the region features, and we then take the final output of the model and perform non-maximum suppression for each object class using an IoU threshold. We then select all the regions, whose predicted category probability exceeding a confidence threshold. Finally we get the N (typically $N = 36$) object features denoted as $V = \{v_i\}^N$. This approach is similar to bottom-up attention, for more details please refer to [2]. Then each image is represented by a scene graph, where the nodes represent objects V while edges, denoted as $E = \{e_{ij}\}^{N\times N}$, represent the semantic visual relations predicted by our proposed relation encoder. Therefore, we construct a fully connected graph based on the object nodes and predicted relations to represent an image.

3.2. Relation encoder

To overcome relation ambiguity, we train a novel relation encoder to model contextual-sensitive relation embeddings, inspired by recent works in Visual Relationship Detection [31,32] and Scene Graph Generation [33].

The structure of the visual relation encoder is illustrated in Fig. 4. The relation encoder infers the type of the relationship between two objects based on three feature vectors extracted from three regions: the subject feature v_i, the object feature v_j and the relation (union of the two regions) feature v_{ij}. The three feature vectors are then projected into lower dimensions through a fully connected layer, where the subject feature and object feature share a set of parameters and the relation feature exclusively uses a set of parameters. Finally, projected feature vectors are concatenated and propagated through three dense fully connected layers to predict the posterior probabilities over all the relation categories.

The output embedding of the last second fully connected layer is chosen as the relation embedding, denoted as $\mathbf{r}_{ij}$ between the object i and the object j.

$$\mathbf{r}_{ij} = MLP([\sigma(\mathbf{W}_o \cdot v_i), \sigma(\mathbf{W}_o \cdot v_j), \sigma(\mathbf{W}_u \cdot v_{ij})]) \tag{1}$$

where $\sigma(\cdot)$ is ReLU activation function, $\mathbf{W}_o$ and $\mathbf{W}_u$ are learnable matrices. *MLP* is composed of two fully connected layers (fc).

Specifically, we choose the top 31 relations in Visual Genome and regard the "non-relation" as a special kind of relation, resulting in total $M = 32$ relations. The probability for relation m ($m \in [1, M]$) is $prob_m$. We choose CrossEntropy as the loss function. Formally, if the ground truth label is $Y = \{y_1, y_2, \ldots, y_M\}$, then the loss for a training sample will be,

$$L = -\sum_{m=1}^{M} y_m log(prob_m) \tag{2}$$

3.3. Anisotropic graph convolution

We propose a novel graph convolution approach, named as *Anisotropic Graph Convolution (AGConv)* , which aims to enrich the object representations with the contextual-sensitive relationships. Notably, "Anisotropic" here indicates that when we observe the relationship between subject v_i and object v_j from different views, v_j has different degrees of impact on v_i. We propose a novel relation-aware graph attention network and leverage multi-head attention mechanism to achieve "Anisotropic" in the graph convolution process. Different from Graph Attention Networks that only consider the visual features of subject and object to calculate the attention probability, our model leverages the relational information to remarkably highlights the impact of the object on the subject according to their relationships.

This Anisotropic Graph Convolution process is illustrated in Fig. 5. First, we compute K attention maps over the relation embedding between each pair of objects, corresponding to K attention heads in our network. The K attention heads then use K sets of independent parameters to aggregate information from each node's neighborhood. The attention weight for the k^{th} ($k \in K$) attention head between object v_i and object v_j is calculated as,

$$w^{(l)}_{ji,k} = \text{softmax}(\mathbf{U}^{(l)}_k \cdot \mathbf{r}_{ji} + \mathbf{b}^{(l)}_k) \tag{3}$$

where $\mathbf{U}^{(l)}_k$ and $\mathbf{b}^{(l)}_k$ are learnable weights. On top of the above attention weight, we define our anisotropic graph convolution on a single node as,

$$\mathbf{h}^{(l+1)}_i = \Big\Vert_{k=1}^{K} \sigma\Big(\sum_{j\in V} w^{(l)}_{ji,k}(\mathbf{W}^{(l)}_{A,k} \cdot [\mathbf{h}^{(l)}_j \parallel \mathbf{r}_{ji}] + \mathbf{b}^{(l)}_{A,k})\Big). \tag{4}$$

where $\mathbf{h}^{(l)}_i$ denote the hidden state of node i at the lth layer. $\mathbf{W}^{(l)}_{A,k}$ and $\mathbf{b}^{(l)}_{A,k}$ are learnable weights. Specially, $\mathbf{h}^{(0)}_i = v_i$. $\parallel$ denotes concatenation operation. After L layers of graph convolution, we

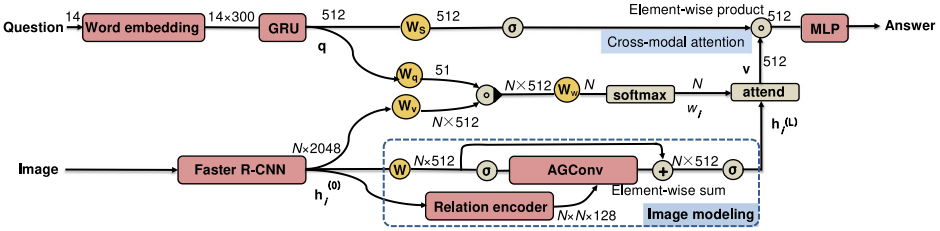

Fig. 6. The architecture of our visual-question-answering model. In this figure, Ⓦ denotes a linear transformation and σ denotes the ReLU activation. $\circ$ denotes element-wise product. The element-wise product used in "cross-modal attention" is slightly different in the way that it multiplies question embedding $\mathbf{q}$ with every object feature vector $\mathbf{v}_i$.

obtain the AGConv-enhanced object feature $\mathbf{h}_i^{(L)}$. In practice, we use only one layer ($L = 1$).

In our model, incorporating prior relation embeddings in Anisotropic Graph Convolution brings several advantages: (1) It introduce comprehensive relational information in image representation, which is ignored by monolithic visual features extracted by CNN or Faster R-CNN; (2) Compared with rigid-divided relation labels, our relation embeddings can efficiently keep fine-grained visual information, thus overcoming relation ambiguity to some extent; (3) Based on the prior relation embeddings, multi-head attention mechanism is capable to capture the relationships of each node's neighborhood from different perspectives, which provides more visual semantics compared to single-head attention.

4. v-AGCN for visual question answering

Visual question answering is a task that requires joint reasoning over questions in natural language and the images. We plug our AGConv module into a state-of-the-art VQA network by Anderson et al. [2] to prove the effects in visual relational reasoning. An overview of our proposed model, namely (v-AGCN), is shown in Fig. 6. It treats visual question answering as a classification problem and takes three types of data as inputs: N feature vectors of objects detected by the Faster R-CNN, $N \times N$ relation embeddings generated using the aforementioned relation encoder, and the word-level question embedding. The architecture of v-AGCN mainly contains three essential parts: (1) *fine-grained image modeling* (bottom branch in Fig. 6) obtains the relation-aware image representation via the AGConv module for relational reasoning; (2) *question modeling* (top branch in Fig. 6) leverages GRU followed by a non-linear layer to obtain the question representation q; (3) *cross-modal attention* (middle branch in Fig. 6) aims to capture the question-relevant image features via attention mechanism to enhance the image representation.

In the image modeling path, we first detect N objects using Faster R-CNN model to construct the scene graph. Then both the graph structure, object features, and relation features are fed into the AGConv module (in the blue dashed box in Fig. 6) for relational reasoning. Empirically, we add a residual connection from the input of AGConv to the feature combination after AGConv to provide enhanced feature alignment, that is, preserving enough object-appearance information and preventing the updated object features to drift drastically.

Cross-modal attention is then applied to focus on the question-relevant image features via cross-modal attention. A scalar attention weight w_i is obtained based on question features and object features. Formally,

$$w_i = \text{softmax}(\mathbf{W}_W[\mathbf{W}_v\mathbf{h}_i^{(L)} \circ \mathbf{W}_q\mathbf{q}]) \tag{5}$$

$$\mathbf{v}_i = w_i\mathbf{h}_i^{(L)} \tag{6}$$

where $\mathbf{W}_\omega$, $\mathbf{W}_v$ and $\mathbf{W}_q$ are all learnable weight matrices and the bias terms are omitted without loss of generality. $\mathbf{q}$ is the question embedding via GRU on top of word embeddings of the question. For the i^{th} object, $\mathbf{h}_i^{(L)}$ is the enhanced feature by L layer of AGConv introduced in Eq. (4), and $\mathbf{v}_i$ is the updated question-aware object feature.

As illustrated in Fig. 6, we fuse the question features and enhanced image features $\mathbf{v}_i$ by element-wise product. The decoder ranks all the answers from a set of candidates. We employ a two-layer MLP followed by ReLU activation over the fused question-image features to predict the probabilities of all the candidate answers. Formally,

$$\hat{\mathbf{p}} = \sigma(MLP(\mathbf{v} \circ \mathbf{W}_S\mathbf{q})) \tag{7}$$

where $\mathbf{W}_S$ is a learnable matrix. $\mathbf{v}$ is the comprehensive representation of the image which can be the sum of the objects, denoted as $\mathbf{v} = \sum_{i=1}^{N} \mathbf{v}_i$. $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{N_a}]$ is the predicted probability of all candidate answers. We denote the predicted probability for the i^{th} answer by $\hat{p}_i$ and the benchmark label by p_i. The cross-entropy loss function is used to train the model, formally defined as,

$$loss = -\sum_{j=1}^{B}\sum_{i=1}^{N_a}(p_i^{(j)}\log(\hat{p}_i^{(j)}) + (1 - p_i^{(j)})\log(1 - \hat{p}_i^{(j)})) \tag{8}$$

where B denotes the batch-size, j subscripts the index of the question-answer entry in training batches, and N_a is the size of the candidate answer set.

5. c-AGCN for cross-modal information retrieval

Cross-modal information retrieval typically projects textual and visual features into a common semantic space to measure their similarity directly. Following the idea of [5], we design a simple dual-path neural network by injecting AGConv module to learn multi-modal representations and compute their similarity by metric learning. We name the model as c-AGCN and show the illustration of the framework in Fig. 7. We model the texts by bag-of-word features followed by two layers of nonlinearities whose learnable matrices are W_1 and W_2. The image modeling is similar to that in v-AGCN and differs in the following two aspects.

First, cross-modal attention in v-AGCN is replaced by self-attention in c-AGCN to integrate AGConv-enhanced object features as the image representation. The Ⓜ in Fig. 7 denotes integration with self attention which is defined as,

$$\mathbf{v} = \Big\|_{k=1}^{K} \sigma\Big(\sum_{i \in V} w_{i,k}[\mathbf{W}_{B,k}^{(L)} \cdot \mathbf{h}_i^{(L)} + \mathbf{b}_{B,k}^{(L)}]\Big) \tag{9}$$

where $\mathbf{W}_{B,k}^{(L)}$ and $\mathbf{b}_{B,k}^{(L)}$ are learnable parameters. $\mathbf{h}_i^{(L)}$ is the enhanced feature of node i, which is the output of AGConv module. $\mathbf{v}$ is the integrated features served as the image representation. Attention weights are generated by a linear transformation of the nodes' hidden states as follows,

$$w_{i,k} = \text{softmax}(\mathbf{U}_{C,k}^{(L)}\mathbf{h}_i^{(L)} + \mathbf{b}_{C,k}). \tag{10}$$

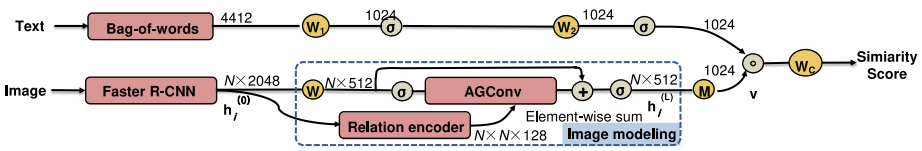

Fig. 7. The architecture of our cross-modal information retrieval model. In this figure, Ⓦ denotes a linear transformation and σ denotes the ReLU activation. $\circ$ denotes element-wise product. The Ⓜ denotes a self-attention mechanism explained Eqs. (9) and (10).

The reason that cross-modal attention is not used in c-AGCN in merging object features lies in the essential distinction between the CMIR and VQA tasks: the key challenge in cross-modal information retrieval is to align the representations of different modalities, which requires better representations of single-modal content to provide adequate semantics for cross-modal alignment.

Finally, we use a fully connected layer W_C to predict the cross-modal similarity based on the fused features from image and text paths by element-wise product. We adopt pairwise similarity loss function in [34] as our optimization objective. Specifically, we maximize the mean similarity score u_p between matching text-image pairs and minimize the mean similarity score u_n for non-matching pairs. Meanwhile, a variance loss, which minimizes the similarity variance of both matching σ_p^2 and non-matching pairs σ_n^2, is added to the loss function to accelerate convergence. The loss function is defined as,

$$loss = (\sigma_p^2 + \sigma_n^2) + \lambda \max(0, m - (u_p - u_n)) \tag{11}$$

where m is the margin between the mean distributions of matching and non-matching similarity and λ is used to balance the weight of the mean loss and variance loss.

6. Experiments

In this section, we first conduct qualitative visualization to evaluate that the relation embedding is capable to disambiguate the semantics of visual relations. Then we test our AGConv module on both VQA and CMIR tasks and fix the hyper-parameters for both v-AGCN and c-AGCN. The dimension of every hidden layer is shown in Figs. 6 and 7. We train all of our models by Adamax solver with 20 epochs with mini-batch size 256. We adopt the learning rate of 0.001 and dropout ratio of 0.5. The number of possible answers for VQA is set to 3,129 by filtering out answers that appear less than 9 times. m and λ in the loss of c-AGCN are set to 0.6 and 0.35, respectively. All the experiments are implemented with Tensorflow and conducted with NVIDIA Tesla V100 GPUs.

6.1. How to train the relation encoder?

To leverage the information in human prior knowledge, our relation encoder is trained on a large-scale vision-language dataset, i.e. Visual Genome [15]. Visual Genome originally provides relationship annotations of 4,017 categories, the distribution of the number of relation instances among which is very uneven, ranging from only a few to tens of thousands. The visual relationships are represented as triples $\langle subject, predicate, object \rangle$. The relation predicate can be a preposition, an action or a combination of both. Therefore, we manually selected and grouped 30 most frequent and representative relation categories and added two additional classes including a *none* relationship for any two objects that are not labeled with a relation predicate, and a *is-a* relationship for the self-connections of nodes. Thus, the relation encoder classifies 32 categories in total. In our implementation, the feature descriptors of image regions are obtained from the *Res4b22*[1] feature map of a *ResNet − 101*[30] model through *RoI Pooling*[29]. The relation embeddings generated by our relation encoder capture rich semantic information of the subject and object in the image. All the hyper-parameters are shown in Fig. 4.

6.2. What is captured by relation embeddings?

Since the relation embedding is designed for modeling fine-grained relations between instances and their interaction, it must has the ability to disambiguate the diverse relational semantics of different visual appearances, which is a challenging problem explained in Section 1. To examine this ability, we perform relation-embedding-based image retrieval to search for the k-nearest images containing the same relational semantics as the query image. Specifically, Faster R-CNN is first utilized to detect salient regions in 1,000 images randomly sampled from MSCOCO dataset. We then generate a relation embedding for each pair of salient regions in an image and predict its relation category accordingly. Afterwards, we randomly sample a set of 50,000 relation embeddings from all the obtained relation embeddings, each of which belongs to its predicted relation category. In this embedding set, we randomly select 32 sample embeddings as queries, belonging to the predefined 32 categories, respectively. For each query, we search for its 4-nearest-neighbors in the embedding set according to the Euclidean distance between embedding vectors, and visualize the images that containing the query and the retrieved relation embeddings.

Some samples of the retrieved images are shown in Fig. 8. In Fig. 8(a), we visualize four queries and their retrieved images in the category of *ride*. Though the four queries belongs to the same relation category, it is obvious that they respectively convey the semantic of *ride* from different views, such as *ride skateboard*, *ride horse* or *ride bike*. The results demonstrate that, based on the relation embeddings, we can accurately retrieve the relation-relevant images containing the same fine-grained relations as the queries. Similarly, Fig. 8(b) shows four queries and their retrieved results in the category of *on*. We come out with the same observation as the first four queries and this observation also exists in other categories. These experimental results demonstrate that our relation embeddings provide a large volume of information about relationships between objects and avoid the risk of losing useful relational information, that help our models effectively overcome the relationship ambiguity, which indicates the advantage of our model argued in 3.3.

6.3. Evaluation on visual question answering

6.3.1. Dataset and evaluation metrics

We evaluate v-AGCN on the VQA 2.0 [36] dataset, which is the upgraded version of VQA dataset. VQA 2.0 is more balanced in designing questions to relieve the problem of overfitting. Each question is corresponding to two different images with distinct

[1] *Res4b22* is the last convolutional layer in stage 4 of ResNet-101. This feature map achieves best detection results in [35].

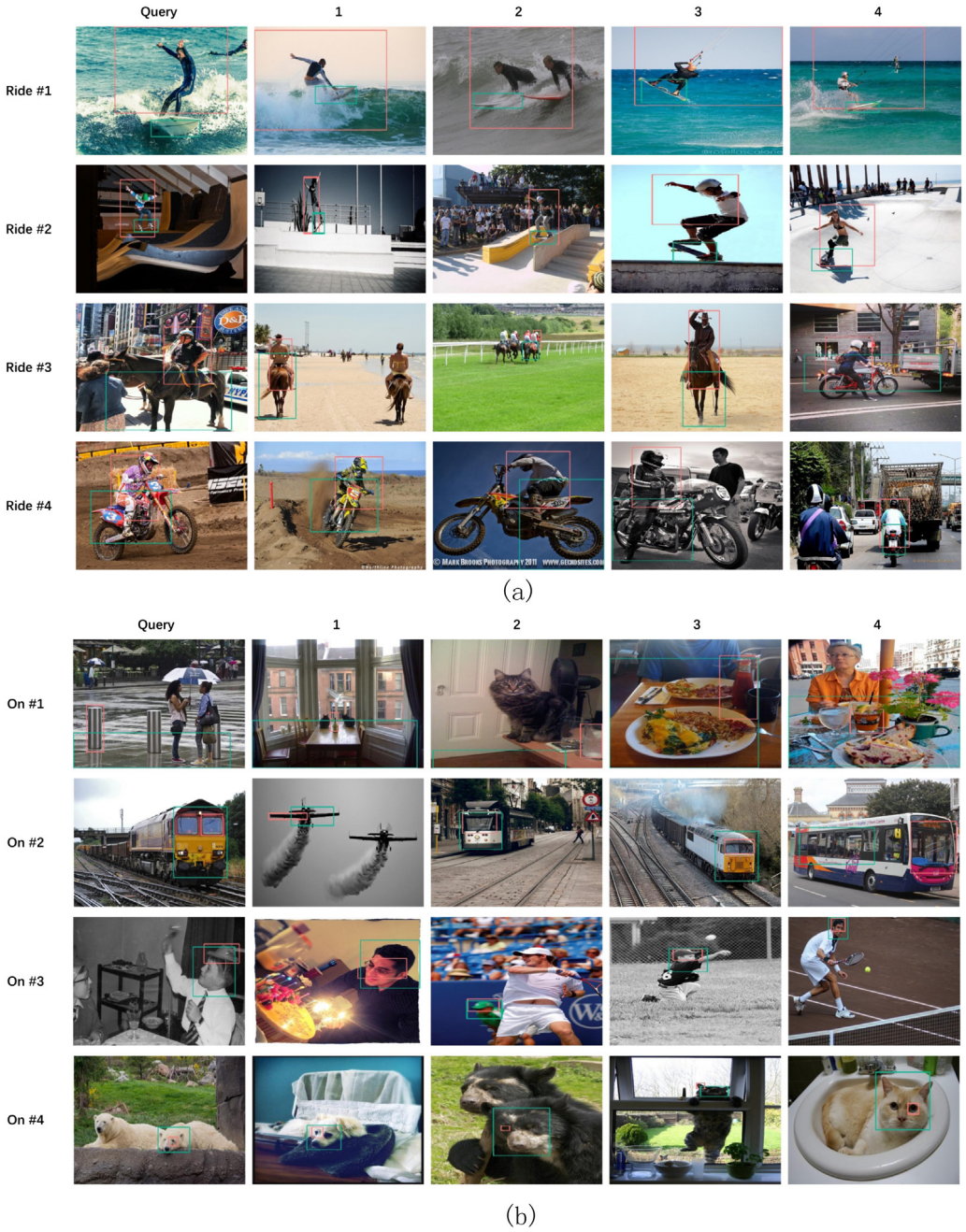

Fig. 8. Samples of relation-embedding-based image retrieval. Each row shows the retrieved results of top 5 nearest images containing the same relational semantics as the query image. In each image, the subject is framed with red box while the object is framed with green box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

annotated answers. Specifically, VQA 2.0 contains 204,721 images, each with at least three questions and ten ground truth answers per question. There are three types of questions: *yes/no*, *number* and *other*. The accuracy for each answer is based on a voting by annotators:

$$accuracy(a_i) = \min(1, \frac{n_i}{3}) \tag{12}$$

where n_i is the times that answer a_i is voted by different annotators. We follow the standard splitting of the dataset and use the tool provided by [37] to evaluate the accuracy. Only the training and validation sets are available for model optimization. Following previous work, we train our model on the training set and report the results on the validation set in our ablation study. We train our model on both the training and validation set and report the *test-dev* and *test-standard* results from the VQA evaluation server for the state-of-the-art comparison.

6.3.2. Ablation study

The architecture of v-AGCN is composed of several essential components. We conduct extensive ablation experiments to evaluate the contribution of each component. We first train our model and several ablated versions on the training set and compare their accuracy on the validation set. The variant models of v-AGCN include:

baseline model: we replace the image modeling component in Fig. 6 with a single linear layer, which is the model introduced in [2].

GCN model: In this model, a layer of traditional GCN [38] is utilized to conduct graph convolution instead of image modeling

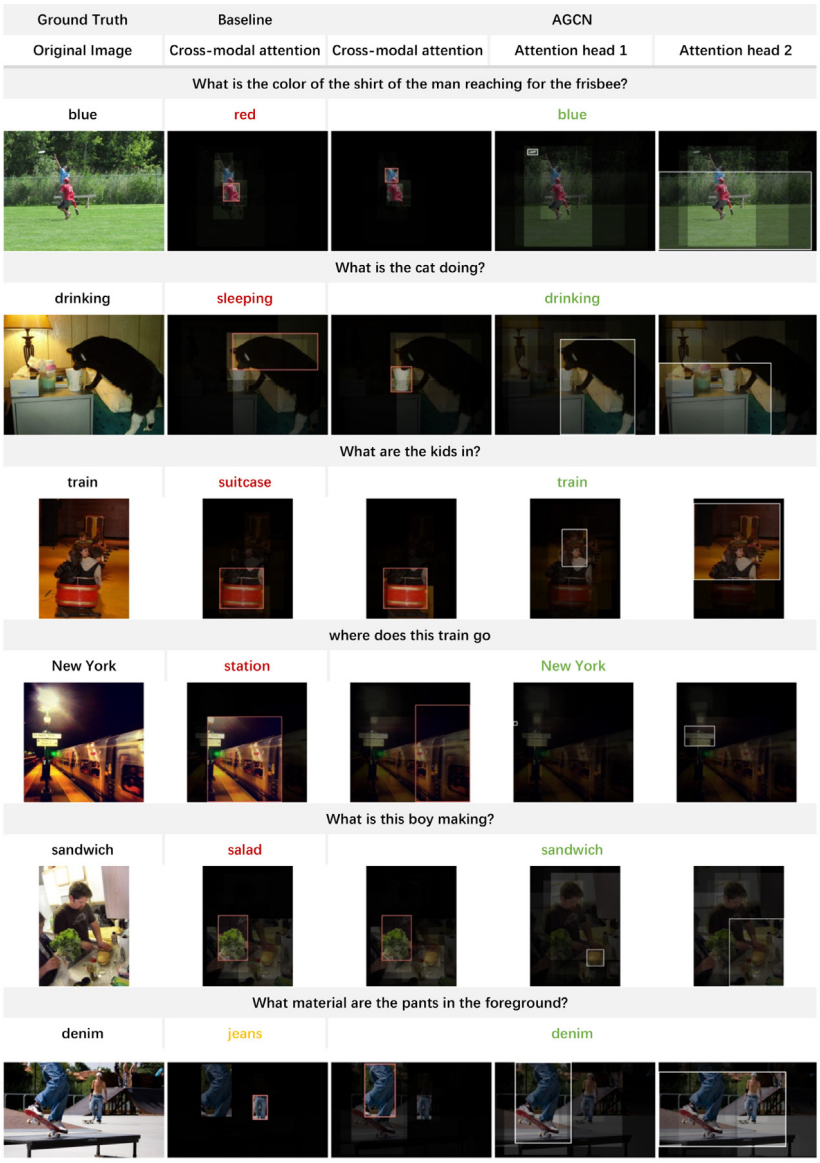

Fig. 9. Predicted results and the corresponding attention visualization of the visual question answering task on the VQA 2.0 dataset. Each row shows the results for a image-question pair. In each row, the first column shows groundtruth answers and original images. The second column visualizes the cross-modal attention map of the baseline model. The third column shows the cross-modal attention map of the v-AGCN model. The attention maps of the first and second attention heads in the v-AGCN model, which make correlation between the regions from different views, are respectively shown in column 4 and column 5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

component in Fig. 6. The model does not consider the relational information among objects.

GAT(single-head) model: this model applies the Graph Attention Network (GAT) [39] with single attention head to replace the image modeling component in Fig. 6. Compared with GCN, GAT(single-head) introduces the relationships of neighborhood by attention mechanism. Different from our AGConv using relation embedding to represent the relationships and compute the attention weights, GAT(single-head) uses the neighboring object to compute the attention weights: $\omega_{ji}^{(l)} = \text{softmax}(\mathbf{U}^{(l)} \cdot \mathbf{h}_j + \mathbf{b}^{(l)})$ is used instead of Eq. (3) in AGConv. Meanwhile, the graph convolution in Eq. (4) is replaced by: $\mathbf{h}_i^{(l+1)} = \sigma(\sum_{j\in V} \omega_{ji}^{(l)}(\mathbf{W}^{(l)} \cdot \mathbf{h}_j^{(l)} + \mathbf{b}_k^{(l)}))$.

GAT(single-head)+re model: compared with GAT (single-head), this model employs generates attention weights based on relation embeddings, formally defined as: $\omega_{ji}^{(l)} = \text{softmax}(\mathbf{U}^{(l)} \cdot \mathbf{r}_{ji} + \mathbf{b}^{(l)})$. Meanwhile, the graph convolution is replaced by: $\mathbf{h}_i^{(l+1)} = \sigma(\sum_{j\in V} \omega_{ji}^{(l)}(\mathbf{W}^{(l)} \cdot \mathbf{h}_j^{(l)} + \mathbf{b}_k^{(l)}))$.

v-AGCN: is our full model employs attention mechanism computed from relation embeddings and uses distinctive sets of parameters for multiple attention heads.

The experimental results are shown in Table 1. We can have the following observations:

First, v-AGCN obviously outperforms baseline, GCN, GAT (single-head), and GAT(single-head)+re. Compared with models without relation information, v-AGCN surpasses baseline and GCN by 1.5% and outperforms GAT(single-head) by 0.6%. It indicates the effectiveness of relation-aware visual representation in improving the final performance of VQA. Besides, by setting distinct weights for multiple attention heads in v-AGCN, we obtain about 1% improvement over GAT(single-head)+re with single attention head. We can infer that different attention heads could capture different views of correlations between visual regions based on the same relation embedding. In summary,

Table 1
Ablation study on each central component for the VQA tasks on the VQA 2.0 dataset.

Model	Validation accuracy
baseline	62.41
GCN	62.44
GAT(single-head)	63.32
GAT(single-head)+re	62.84
v-AGCN (ours)	
with 1-head	63.48
with 2-heads	**63.88**
with 3-heads	63.33

incorporating the relation embeddings and multi-head attention mechanism can significantly enrich the visual representation and thus promote the VQA accuracy.

Second, progressively equipping the baseline model with different components brings varying degrees of improvement. Though constructing the connections between visual regions, adding the typical GCN structure does not improve the results over baseline model. This is because that only introducing mutually connections without specific relation types or characterizations is not informative to enrich the visual representation. Equipping the GCN model with an attention mechanism based on object appearance results in the GAT(single-head) variant and raises the accuracy by 0.9%. Adding pre-trained relation embedding to GAT(single-head) results in performance decrease. GAT(single-head)+re uses the relation embedding to calculate the attention weights which is the same as v-AGCN, but does not use relation embedding when updating node representation. The performance decrease of GAT(single-head)+re demonstrates that, for VQA task, pre-trained relation embedding should be incorporated into both attention weight calculation and node update. Our v-AGCN properly uses pre-trained relation embedding and achieves the best validation accuracy of 63.9%.

Third, to evaluate the impact of hyper-parameter K (the number of attention heads in GCN), we also tried to set K to be 1, 2, and 3. The results are also reported in Table 1, which demonstrate that 2 attention heads lead to the best performance for our v-AGCN model. And we use this setting in all of the following experiments.

6.3.3. State-of-the-art comparison

Table 2 shows the model's performance on the VQA 2.0 dataset. We trained our model 10 times (initializing by different random seeds) and tested it on the test-dev and test-standard split. Our model performs steadily within 0.5% overall accuracy range in the statistical study. Since all the compared models report their best results on this dataset, for fair comparison, we also report our best results in Table 2. **Prior** and **Language-only** are two simple baselines from [36]. **MCB** [22] is the winning entry of VQA Challenge 2016 which uses a multimodal compact bilinear pooling mechanism, while **Adelaide** is the champion of VQA Challenge 2017 [40], which is the first approach using object-level features as image representation. **MFB** [3] is a recently proposed model which uses multimodal factorized bilinear pooling to fuse image and question. Following these models, we also use the combination of training and validation set to train v-AGCN and report the performance on the test set. It is shown that our method outperforms all the state-of-the-art methods on the overall accuracy.

It is obvious that the results of our single v-AGCN model significantly surpass all the top-ranked models and slightly outperform the 1st place model of Adelaide. The best accuracy of v-AGCN achieves 65.94% on the *test-dev* set and 66.17% on the *test-standard* set. It is worth to notice that our model has good performance especially for "Number" questions, which mainly dues to the following two reasons: (1) Most "Number" questions in VQA 2.0 need the model to count salient objects. Hence pre-trained object detector can help VQA model predict more correct answers for "Number" questions. Therefore, our model also uses pre-trained Faster-RCNN to detect salient objects, which is one reason for the good performance for the "Number" questions. (2) Some "Number" questions require relational information to predict the answers. For instance, the question in Fig. 1(b) "How many men cutting the cake?" requires the model to understand the interaction relationship "cutting" when counting the number of "men". In our model, we incorporate the pre-trained relation embedding in the visual reasoning process, which provides essential clues for reasoning the relationships between detected salient objects. Our model benefits from the pre-trained relation embedding and gets better result for "Number" questions compared with Adelaide model, which only leverages the advantages of salient regions.

6.3.4. Qualitative evaluation

To examine the behavior of the AGConv module, we visualize the attention maps that the cross-modal attention and inter-node attention in v-AGCN. Specifically, we show samples of input image-question pairs with their attention maps generated by v-AGCN and the baseline model. Examples are shown in Fig. 9. Each row contains visualization for a question. The first column in each row shows groundtruth answers and original images. The second column visualizes the cross-modal attention used in the baseline model. The third column shows the cross-modal attention in the v-AGCN model. The attention maps of the first and second attention heads in the v-AGCN model, which make correlation between the regions from different views, are respectively shown in column 4 and column 5. For each attention map, the red bounding box indicates the region with the maximum attention weight. The predicted answers are given above the attention maps accordingly.

Comparing with the results of baseline model, we observe that the cross-modal attention in v-AGCN focuses on more relevant image regions referred by the questions. When answering the question "*What material are the pants in the foreground?*", the agent should make a distinction between the two people, in which the v-AGCN model succeeds. Another property of our v-AGCN is that it is less prone to answering questions with partial

Table 2
Accuracy comparison of VQA tasks with state-of-the-art approaches on the VQA 2.0 dataset.

Model	test-dev				test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
Prior [36]	–	–	–	–	25.98	01.17	00.36	61.20
Language only [36]	–	–	–	–	44.26	27.37	31.55	67.01
LSTM+CNN [36]	–	–	–	–	54.22	41.83	35.18	73.46
MCB [22,36]	–	–	–	–	62.27	53.36	38.28	78.82
Adelaide [40]	65.32	56.05	44.21	81.82	65.67	56.26	43.90	82.20
MFB [3]	65.90	56.20	39.80	**84.00**	65.80	56.30	38.90	**83.80**
v-AGCN (ours)	**65.94**	**56.46**	**45.93**	82.39	**66.17**	**56.71**	**45.12**	82.58

information. When answering question "*What are the kids in?*", although the baseline model attended to a correct image region, the partial information in that region caused it to generate the answer *suitcase*, while the v-AGCN model observes other image regions and conclude that the kids are in fact in a *train*. Most importantly, it is shown that the v-AGCN model is capable of discovering and using visual relationships. For the question "*What is the color of the shirt of the man reaching for the frisbee*", v-AGCN correctly answers *blue* by finding a correlation between the man and the frisbee.

6.4. Evaluation on cross-modal information retrieval

6.4.1. Dataset and evaluation metrics

In this section, we test our models on the most recent CMIR benchmark datasets: Cross-Modal Places [41] (CMPlaces). CMPlaces is one of the largest cross-modal dataset providing weakly aligned data in five modalities divided into 205 categories. In our experiments, we utilize the natural images (about 1.5 million) and text descriptions (11,802) for evaluation. We randomly sample 250 images from each category and split the images for training, validation, and test with the proportion of 8:1:1. We also randomly split text descriptions for training, validation, and testing with the proportion of 14:3:3. As for the evaluation, MAP@100 is used to evaluate the query performance. We compute the overall MAP by averaging a score for text-queries Q_T and a score for image-queries Q_I.

6.4.2. Ablation study

The architecture of c-AGCN is composed of the same essential components as v-AGCN. We conduct several ablation experiments to evaluate the contribution of each component. We use 4 attention heads in AGConv here. We train our model and its ablated versions on the training set and compare their accuracy on the testing set. The ablation models of c-AGCN, including baseline, GCN, GAT(single-head), GAT(single-head)+re, are designed in the same way as the ablation models of v-AGCN. The only difference lies in the basic architecture that the ablation models in this section are based on the cross-modal information retrieval model illustrated in Section 5.

The MAP scores are shown in Table 3, including the results of text query(Q_T), image query (Q_I), and the average performance. From the observation, we come out with the same conclusion as the v-AGCN model. Our proposed c-AGCN model achieves the best overall performance. The baseline experiment shows the lowest text-query score and the highest image-query score, we hypothesize this is because dense fully-connected layers are prone to be overfitting. The traditional GCN model shows the lowest image query score. However, when adding the attention mechanism to GCN, it shows remarkable improvement for the image query task. By taking the advantages of relation embedding, GAT(single-head)+re gains another 2.1% improvement for the image query task and slight promotion in the text query task. To go one step further, when setting distinct weights for different attention heads, c-AGCN model gains another 2.3% promotion for the average performance.

6.4.3. State-of-the-art comparison

The comparison with state-of-the-art is also shown in Table 4. We compare our model with state-of-the-art approaches, including *GIN* [4] and Castrejon's models [41] proposed with the introduction of the CMPlaces dataset. Both *GIN* [4] and Castrejon 's models [41] use grid-structured CNN features to model images. The results for *GIN* are newly tested on the CMPlaces dataset in our work while the results for Castrejon et al. [41] are from their published paper. The MAP scores in Table 4 indicate that our model outperforms state-of-the-art approaches by considerable improvements. Specifically, the image queries benefit more from the anisotropic graph convolution than the text queries, since our AGConv module is mainly proposed for enhancing the expressive capacity and robustness of the visual representation.

Table 3
Ablation study on each components for the image-text retrieval on the CMPlaces dataset.

Method	Q_T	Q_I	Avg.
baseline	29.7	**40.5**	35.1
GCN	34.5	21.0	27.5
GAT(single-head)	34.5	35.7	35.1
GAT(single-head)+re	34.6	37.8	36.2
c-AGCN (ours)	**37.7**	39.3	**38.5**

Table 4
MAP score comparison of image-text retrieval on the CMPlaces dataset.

Method	Q_T	Q_I	Avg.
BL-Ind [41]	0.6	0.8	0.7
BL-ShFinal [41]	3.3	12.7	8.0
BL-ShAll [41]	0.6	0.8	0.7
Tune(Free) [41]	5.2	18.1	11.7
TuneStatReg [41]	15.1	22.1	18.6
GIN [4]	25.9	23.9	24.8
baseline	29.7	**40.5**	35.1
c-AGCN (ours)	**37.7**	39.3	**38.5**

6.4.4. Qualitative evaluation

For quality evaluation, two examples for image-query-text and text-query-image tasks are shown in Fig. 10. In the top of Fig. 10, the text query is a description of pantry. Compared with baseline model, more results retrieved by c-AGCN are images belonging to "pantry" and that they are semantically relevant to the textual content from fine-grain view. For example, the top five retrieved images all contain visual content highly related to the textual descriptions, including "...stores goods used for cooking..." and "...filled with non perishable goods such as canned food or jars". In contrast, the top results of the baseline model are quite similar to "pantry" in appearance, such as the first two images containing closet. However, the model does not reason about the relations between the closet and the goods and infer the purpose of the closet.

The bottom sample in Fig. 10 shows the top six retrieved texts given an image query belonging to the category of "train railway". Similar to the observation of text query, the top retrieved texts by c-AGCN are semantically relevant to both the category and the fine-grained visual content of the image query. However, the baseline model without visual relation reasoning is easy to confuse with the scenes, e.g. subway platform, train platform, railroad track, which are semantic similar to "train railway" to some extent. In summary, from all the qualitative analysis we can see that c-AGCN is effective in fine-grained cross-modal correlation learning by injected with visual relation information.

7. Conclusion

In this paper, we propose Anisotropic Graph Convolution (AGConv), which explores visual relation reasoning for improving cross-modal learning. Specifically, by incorporating prior knowledge in the visual knowledge base, we novelly propose to model the relations between visual regions in an image as context-aware embeddings to enrich the relation representations. Then the proposed Anisotropic Graph Convolution is applied to further reason about richer semantics based on the learnt visual relation embeddings for more informative image representation. To

Fig. 10. Retrieval samples of our proposed c-AGCN model and the baseline model on the CMPlaces dataset. Relevant results are highlighted with green boxes while irrelevant results are marked with red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

evaluate the effectiveness of our module, we inject AGConv into the models for Visual Question Answering and Cross-Modal Information Retrieval tasks. Extensive experiments prove that existing models could be enhanced by our proposed module and resulting in significant improvement compared with state-of-the-art approaches.

CRediT authorship contribution statement

Jing Yu: Methodology, Software, Writing - original draft. **Weifeng Zhang:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Zhuoqian Yang:** Software, Visualization. **Zengchang Qin:** Data curation, Validation. **Yue Hu:** Investigation, Writing - review & editing.

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.

Acknowledgment

This work is supported by the National Key Research and Development Program, China (Grant No. 2017YFB0803301).

References

[1] Z. Yang, X. He, J. Gao, Stacked attention networks for image question answering, in: CVPR, 2016, pp. 21–29.
[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: CVPR, pp. 6077–6086,
[3] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5947–5959.
[4] J. Yu, Y. Lu, Z. Qin, W. Zhang, Y. Liu, J. Tan, L. Guo, Modeling text with graph convolutional network for cross-modal information retrieval, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 223–234.
[5] L. Wang, Y. Li, S. Lazebnik, Learning deep structure-preserving image-text embeddings, in: CVPR, 2016, pp. 5005–5013.
[6] L. Jin, K. Li, Z. Li, F. Xiao, G. Qi, J. Tang, Deep semantic-preserving ordinal hashing for cross-modal similarity search, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2019) 1429–1440.
[7] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: ECCV, 2018, pp. 711–727.
[8] K. Fu, J. Li, J. Jin, C. Zhang, Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs, IEEE Trans. Neural Netw. Learn. Syst. 29 (12) (2018) 5910–5921.
[9] O. Day, T.M. Khoshgoftaar, A survey on heterogeneous transfer learning, J. Big Data 29 (2017) (2017).
[10] Y. Yao, Y. Zhang, X. Li, Y. Ye, Heterogeneous domain adaptation via soft transfer network, in: MM, 2019, pp. 1578–1586.
[11] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: recent progress and new challenges, Inf. Fusion 38 (2017) (2017) 43–54.
[12] G. Cao, A. Iosifidis, K. Chen, M. Gabbouj, Generalized multi-view embedding for visual recognition and cross-modal retrieval, IEEE Trans. Cybern. 48 (9) (2017) 2542–2555.
[13] J. Yu, J. Li, Z. Yu, Q. Huang, Multimodal transformer with multi-view visual representation for image captioning, IEEE Trans. Circuits Syst. Video Technol. (2019) http://dx.doi.org/10.1109/TCSVT.2019.2947482, to be published.
[14] A. Santoro, D. Raposo, D. Barrett, M. Malinowski, A simple neural network module for relational reasoning, in: NIPS, 2017, pp. 4967–4976.
[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, M. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2016) 32–73.
[16] P.W. Battaglia, J.B. Hamrick, V. Bapst, et al., Relational inductive biases, deep learning, and graph networks, 2018, arXiv:1806.01261.
[17] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, Int. J. Comput. Vis. 80 (3) (2008) 300–316.
[18] S.K. Divvala, A. Farhadi, C. Guestrin, Learning everything about anything: Webly-supervised visual concept learning, in: CVPR, 2014, pp. 3270–3277.
[19] L. Cewu, K. Ranjay, B. Michael, F. Li, Visual relationship detection with language priors, in: ECCV, 2016, pp. 852–869.
[20] A. Peter, F. Basura, J. Mark, G. Stephen, Spice: Semantic propositional image caption evaluation, in: ECCV, 2016, pp. 382–398.
[21] J. Tenebaum, W. Freeman, Separating style and content, in: NeurIPS, 1997, pp. 662–668.
[22] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrel, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, in: EMNLP, 2016, pp. 457–468.
[23] J. Kim, K. On, W. Lim, J. Kim, J. Ha, B. Zhang, Hadamard product for low-rank bilinear pooling, in: ICLR, 2017.
[24] H. Noh, P.H. Seo, H. Bohyung, Image question answering using convolutional neural network with dynamic parameter prediction, in: CVPR, 2016, pp. 30–38.

[25] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM MM, 2010, pp. 251–260.
[26] V. Ranjan, N. Rasiwasia, C.V. Jawahar, Multi-label cross-modal retrieval, in: ICCV, 2015, pp. 4094–4102.
[27] Z. Ma, Y. Lu, D. Foster, Finding Linear Structure in Large Datasets with Scalable Canonical Correlation Analysis, in: ICML, 2015, pp. 169–178.
[28] K. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: ECCV, 2018, pp. 212–228.
[29] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[31] B. Dai, Y. Zhang, D. Lin, Detecting visual relationships with deep relational networks, in: CVPR, 2017, pp. 3298–3308.
[32] S. Jae Hwang, S.N. Ravi, Z. Tao, H.J. Kim, M.D. Collins, V. Singh, Tensorize, factorize and regularize: Robust visual relationship learning, in: CVPR, 2018, pp. 1014–1023.
[33] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang, Scene graph generation from objects, phrases and region captions, in: ICCV, 2017, pp. 1270–1279.
[34] V.B. Kumar, G. Carneiro, I. Reid, Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions, in: CVPR, 2016, pp. 5385–5394.
[35] S. Ren, K. He, R. Girshick, X. Zhang, J. Sun, Object detection networks on convolutional feature maps, in: CVPR, 2017, 1476–1481.
[36] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering, in: CVPR, 2017, pp. 6325–6334.
[37] S. Antol, A. Agrawal, J. Lu, M. Mitchell, VQA: visual question answering, Int. J. Comput. Vis. 123 (1) (2017) 4–31.
[38] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, 2017.
[39] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, in: ICLR, 2018.
[40] D. Teney, P. Anderson, X. He, A. Hengel, Tips and tricks for visual question answering: learnings from the 2017 Challenge, in: CVPR, 2018, pp. 4223–4232.
[41] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, A. Torralba, Learning aligned cross-modal representations from weakly aligned data, in: CVPR, 2016, pp. 2940–2949.