Reasoning on the Relation: Enhancing Visual Representation for Visual Question Answering and Cross-Modal Retrieval

Jing Yu, Weifeng Zhang, Yuhang Lu, Zengchang Qin, Yue Hu, Jianlong Tan, and Qi Wu

***Abstract*—Cross-modal analysis has become a promising direction for artificial intelligence. Visual representation is crucial for various cross-modal analysis tasks that require visual content understanding. Visual features which contain semantical information can disentangle the underlying correlation between different modalities, thus benefiting the downstream tasks. In this paper, we propose a Visual Reasoning and Attention Network (VRANet) as a plug-and-play module to capture rich visual semantics and help to enhance the visual representation for improving cross-modal analysis. Our proposed VRANet is built based on the bilinear visual attention module which identifies the critical objects. We propose a novel Visual Relational Reasoning (VRR) module to reason about pair-wise and inner-group visual relationships among objects guided by the textual information. The two modules enhance the visual features at both relation level and object level. We demonstrate the effectiveness of the proposed VRANet by applying it to both Visual Question Answering (VQA) and Cross-Modal Information Retrieval (CMIR) tasks. Extensive experiments conducted on VQA 2.0, CLEVR, CMPlaces, and MS-COCO datasets indicate superior performance comparing with state-of-the-art work.**

***Index Terms*—Visual relational reasoning, visual attention, visual question answering, cross-modal information retrieval.**

I. Introduction

TRADITIONALLY, computer vision (CV) and natural language processing (NLP), which are both important research fields in Artificial Intelligence, have developed independently. Recently, with the advances in computer vision and natural language processing, researchers make a further step towards breaking the boundary of vision and natural language, such as automatic image annotation [1]–[3] image question answering [4]–[6], video question answering [7]–[13], cross-modal information retrieval (CMIR) [14]–[18], image captioning [19], etc. All these tasks require fine-grained visual understanding, or even content-specific visual reasoning to achieve semantic-rich visual representation, which is a top priority for AI to achieve human-level ability.

Despite the advances of deep networks in modeling both images and texts, the potential limitation for theses approaches [4], [15] in cross-modal analysis is that the visual and textual features are learnt independently without interactions with each other. Even for the effective co-attention [20] or dual attention [21] networks to connect two modalities via attention mechanism, these approaches are limited by treating each image region and textual fragment independently and ignoring the complex but semantically informative relational clues in the images. For example, to answer the VQA question in Fig. 1(a) *What is the woman doing?*, attention mechanism can merely help to focus on the question-relevant image regions, such as *woman*, *pot*, *cooker hood*, *etc.*, but can hardly infer the relationships among these regions, which are essential clues for predicting the woman's action of *cooking*. In the example of cross-modal retrieval in Fig. 1(b), the similarity between the image and the text can be accurately measured by analysing the visual regions and their relationships sharing common semantics with the textual content, including entities (e.g. *tanks*, *fish*, *people*) and relationships (e.g. *full of*, *stop to view*). Therefore, both object-level and relation-level semantics are essential for fine-grained cross-modal learning, which is challenging for existing approaches.

One of the recent advances in visual representation for cross-modal learning is visual relational reasoning. It aims to reason about the semantic relationships (i.e., *wearing*, *holding*, *riding*), positional relationships (i.e., *above*, *below*, *inside*, *around*), or even latent relationships between visual objects in an image. Such relationships are working in conjunction with deep neural networks to generate relation-aware visual representations. State-of-the-art work has proved that reasoning about visual relationships is crucial to improve the performance of VQA [22] and image captioning [19]. [22] proposed an approach to infer relationships between all the implicit object-like representation pairs via a plug-and-play MLP module for visual question answering.

Manuscript received June 19, 2019; revised November 2, 2019 and December 24, 2019; accepted January 31, 2020. Date of publication February 10, 2020; date of current version November 18, 2020. This work is supported by the National Key Research and Development Program under Grant 2017YFB0803301. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris. *(Corresponding author: Weifeng Zhang.)*

Jing Yu, Yue Hu, and Jianlong Tan are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China (e-mail: yujing02@iie.ac.cn; huyue@iie.ac.cn; jianglongtan@iie.ac.cn).

Weifeng Zhang is with the College of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314000, China (e-mail: 1574940537@qq.com).

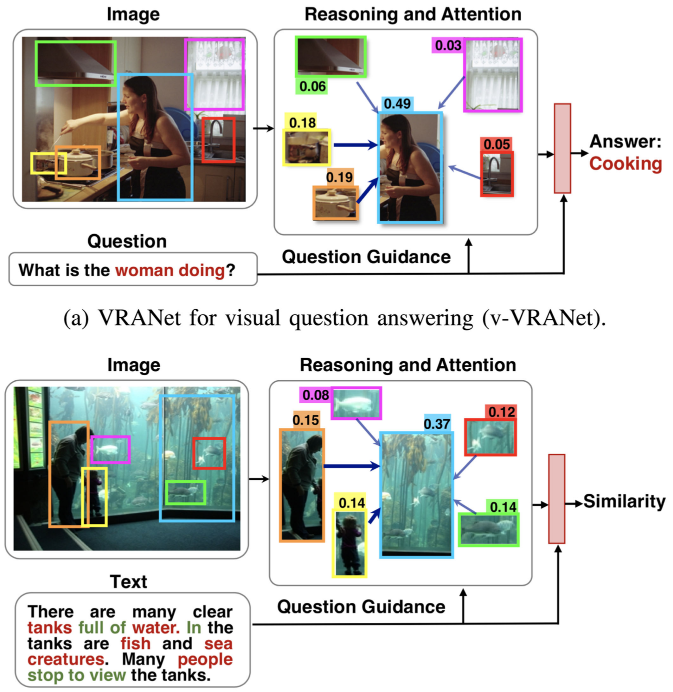
Yuhang Lu is with the Alibaba Group, Hangzhou 310052, China (e-mail: yuhanglu@iie.ac.cn).

Zengchang Qin is with the Intelligent Computing & Machine Learning Lab, School of ASEE, Beihang University, Beijing 100191, China (e-mail: zcqin@buaa.edu.cn).

Qi Wu is with the Australian Centre for Robotic Vision, The University of Adelaide, Adelaide, Australia 5005, Australia (e-mail: qi.wu01@adelaide.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2972830



(a) VRANet for visual question answering (v-VRANet).

(b) VRANet for cross-modal retrieval (c-VRANet).

Fig. 1. Overview of Visual Reasoning and Attention Networks (VRANet) for visual question answering and cross-modal retrieval. The darkness of the edges indicates different hidden relationships among image regions reasoned by the visual relational reasoning module. The probabilities of the image regions indicate their attention weights predicted by the bilinear visual attention module. In the question and the text, the red words indicate entities while the green words indicate the relationships between entities.

However, this method has been proven to only work well on synthesized 3D datasets [23] with geometrical objects. To address problems involving real-world objects, Wang *et al.* [24] proposed VQA-Machine, which can predict more accurate answers by using visual facts generated by existing vision algorithms. In order to enable the VQA-Machine to exploit a wide variety of available CV methods, they formulated the visual facts as triplets such as <*cat, eat, fish*>, which required prior knowledge from manualy labeled relational information to train their model. How to learn visual relationships without prior knowledge is a challenging problem and there still has room for further exploration.

Attention mechanism, first proposed by Bahdanau *et al.* [25] to improve machine translation, recently has also achieved great success in cross-modal learning to enhance visual representation and becomes an important component for VQA models [26]–[28] as well as CMIR models [14]–[16]. Attention networks in cross-modal learning models provide an effective way to selectively focus on visual information that is relevant to the textual content. However, these approaches are limited in that only leveraging attention mechanism cannot achieve relational reasoning since it treats each object individually and there is no interaction between objects.

Inspired by previous work, we believe that attending visual objects and reasoning the visual relationships, in other words, the correlations between salient regions/objects in an image, are important abilities for captureing fine-grained and semantic-rich visual features. In this paper, we go one step further in this field and propose a Visual Reasoning and Attention Network (VRANet) to enrich the visual representations with both relation-level and object-level clues for cross-modal learning. The VRANet mainly consists of two modules: the *visual relational reasoning module* reasons about both pair-wise and inner-group visual relationships among objects based on convolutional neural networks and enriches the representations of each object according to its relational importance; and the *bilinear visual attention module* identifies the critical objects according to the textual content based on the bilinear attention. The two modules process the original visual features in parallel to obtain the enriched relation-level and object-level features resepectively. The two kinds of features are then fused for downstream tasks. We evaluate our proposed VRANet on two representative cross-modal learning tasks: Visual Question Answering and Cross-Modal Information Retrieval. Extensive experimental results indicate the superior improvements by integrating our VRANet with the state-of-the-art methods, verifying the benefits of leveraging both of the relation-level and object-level visual semantics.

The rest of this paper is organized as follows: In Section II, we review progress in several domains strongly associated with our work. Then we introduce our proposed VRANet in Section III, mainly including the visual relational reasoning module and the bilinear visual attention module. Two novel models integrated with VRANet are proposed for cross-modal reasoning and retrieval, and applied to VQA in Section IV and CMIR in Section V. The experimental setup and results are given in Section VI. We conclude our work in Section VII.

II. Related Work

A. *Visual Question Answering*

VQA aims to answer a question in natural language according to a natural image or a clip of video. It is quite a challenging task since it requires understanding and reasoning over both visual and textual content. A typical solution for VQA is to fuse visual and textual features for a joint representation and infer the answer based on the fused representation. The most typical methods for feature fusion are element-wise summation/multiplication and direct concatenation. Zhou *et al.* [4] proposed a typical baseline to predict the answer from the concatenation of the image features extracted from pre-trained CNN and question features represented by bag-of-words. Besides straightforward solutions, existing work applies bilinear pooling [29], [30] or more complex fusion methods [31]. Noh *et al.* [31] exploited a novel CNN model for feature fusion with a dynamic parameter layer whose weights are learned adaptively according to the question.

However, the above approaches are based on global features of the image. Such global features fail to capture fine-grained information and may possibly introduce noise. Some research has studied the attention mechanism to focus on semantically relevant image regions or video frames regarding a given question. Yang *et al.* [5] performed visual attention multiple times via stacked attention networks, and Anderson *et al.* [6] proposed a top-down attention on pre-detected salient regions. These models exploited the fine-grained correlations between visual and textual content and eliminate noisy information. Zhang *et al.* [12] developed a hierarchical convolutional self-attention encoder to model long-form video contents and a multi-scale attentive decoder to generate open-ended answers. Recently, visual relational reasoning has been introduced into VQA and achieved better performance for problems that require a logical understanding of the question and the image [22], [24], [32]. Jin *et al.* [11] proposed a multi-interaction network with exploiting relations between objects which achieved state-of-the-art on video question answering task. These approaches mimic human thinking, which has not been thoroughly studied yet. In this paper, we argue that visual representation plays an important role in various cross-modal analysing tasks, such as VQA and CMIR. Different from the existing work, we go one step further to leverage both visual relational reasoning and visual attention for enriching visual representations. Visual attention mechanism helps our models to filter out irrelevant visual features and visual relational reasoning mines relationships between visual objects. By combining these two modules together, fine-grained and semantic-rich visual representation can be effectively captured.

B. Cross-Modal Information Retrieval

CMIR is a task to enable queries from one modality to retrieve information from another modality. The typical solution for CMIR is to map the data from different modalities into a common semantic space to directly compare their semantic similarity. Several statistical methods are based on Canonical Correlation Analysis (CCA) [33], [34] to maximize the pair-wise correlations. However, these methods ignored high-level semantic priority and could be hard to extend to large-scale data [35]. Another research trend is based on deep learning [14]–[16], [36], leveraging existing techniques to provide rich semantics by nonlinear transformations. Typically, [15] proposed a two-branch neural networks with two layers of nonlinearities on top of visual and textual features. Instead of MLP for feature mapping, [16] proposed a 2-stage CNN-LSTM network to generate and refine the cross-modal features progressively. [36] leveraged attention mechanism to focus on essential image regions and words for correlation learning. Recently, [14] explored the relationships between words and prove the effectiveness for representing texts and eventually improve the CMIR accuracy. In addition, Zhu *et al.* [18] introduced a cross-modal interaction network to explore the potential relations of video and query contents, thus effectively localize the most relevant moment in an video according to the query.

To the best of our knowledge, we are the first to model the visual relationships for the CMIR task. Our model automatically detects and reasons about visual relationships for more informative visual representation, which supports fine-grained image-text joint embedding for more accurate cross-modal similarity measurement.

III. Methodology

In this section, we detail our proposed Visual Reasoning and Attention Network (VRANet), which leverages the textual part to guide the relational reasoning and essential region selection over the visual part. The framework of VRANet is illustrated in Fig. 2. Our proposed VRANet is built based on the bilinear visual attention module which identifies the critital objects. Driven by the textual information, we add a novel Visual Relational Reasoning (VRR) module to futher reason about the relationships between two objects as well as the relationships among multiple objects. We obtain fine-grained visual representations by integrating the relational reasoning module and visual attention module, which provide rich cross-modal correlated clues for downstream tasks. In this section, we first introduce the bilinear visual attention module as the fundamental part of our model. The novel Visual Relational Reasoning will be introduced in details in the second part.

A. Bilinear Visual Attention Module

Given the visual region features $V = [v_1, v_2, \ldots, v_K]^T \in \mathcal{R}^{K \times d_V}$ and the text features $Q \in \mathcal{R}^{d_Q}$, the bilinear visual attention module aims to enhance the visual features by focusing on the text-relevant regions. In our model, in order to reduce the computational cost as well as the risk of over-fitting, we leverage the low-rank bilinear model [37] to obtain the attention map.

The illustration of our bilinear attention module is shown in Fig. 3. The attention weight ω_i for the image region i can be calculated as follows:

$$\omega_i = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \tag{1}$$

where z_i in equation (1) is formally defined as:

$$z_i = P^T(H^T v_i \circ U^T Q) \tag{2}$$

where $P \in \mathcal{R}^d$ is a learnable vector. To reduce parameters, following [30], [38], we use the same projection matrix $H \in \mathcal{R}^{d_V \times d}$ and $U \in \mathcal{R}^{d_Q \times d}$ for all the image regions. Finally, the attended visual representation $V_{att} \in \mathcal{R}^{d_V}$ of all the regions in an image can be the weighted sum of all the region features:

$$V_{att} = \mathcal{A}^T \cdot V \tag{3}$$

where $\mathcal{A} = [\omega_1, \omega_2, \ldots, \omega_K]^T$ is the attention map.

B. Visual Relational Reasoning Module

As shown in Fig. 4, our proposed visual relational reasoning module is composed of four parts, including dimension reduction, pair-wise combination, pair-wise relational reasoning and inner-group relational reasoning. Given the visual features $V = [v_1, v_2, \ldots, v_K]^T \in \mathcal{R}^{K \times d_V}$ and the corresponding geometric features $G \in \mathcal{R}^{K \times d_G}$ of the K image regions, and the text features $Q \in \mathcal{R}^{d_Q}$, our visual relational reasoning module aims to capture visual relationships between image regions and generate relation-aware visual represnetations under the guidance of the text.

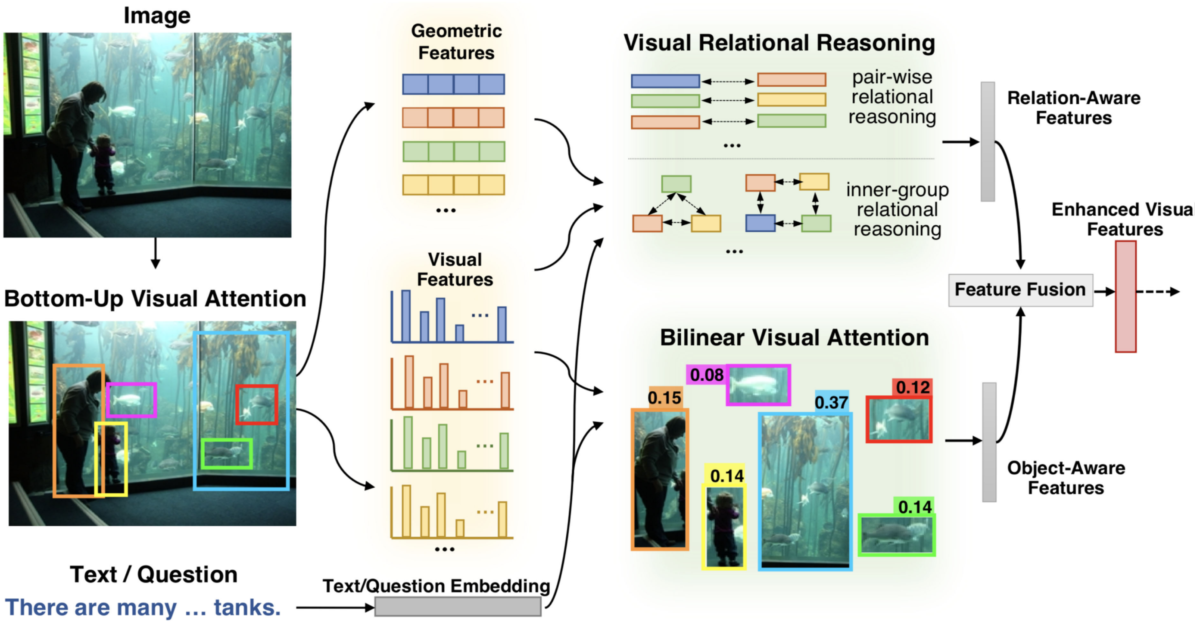


Fig. 2. The framework of the proposed Visual Reasoning and Attention Network (VRANet).

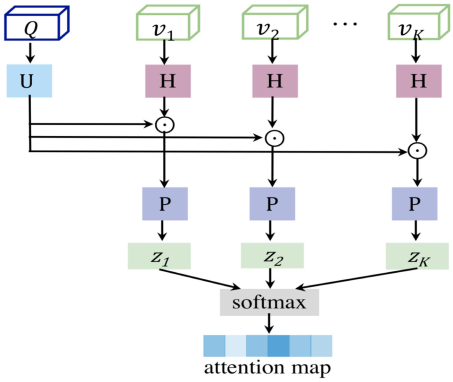


Fig. 3. The flowchart of bilinear visual attention module.

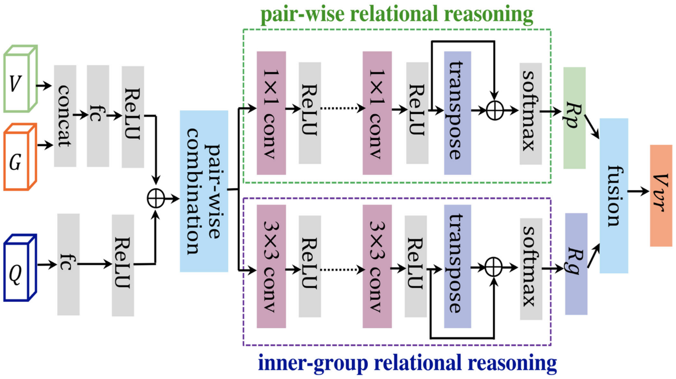


Fig. 4. The flowchart of visual relational reasoning module.

Dimension reduction: For each image region, the visual features and geometric features are first concatenated together [39] to generate object features. To reduce the memory usage of the following operations, we nonlinearly project the region features and the question features into a subspace with lower dimension and fuse their features to obtain the text-enhanced region representations as follows:

$$V_{reg} = ReLU(W_V \cdot [V, G] + b_V) + ReLU(W_Q \cdot Q + b_Q) \quad (4)$$

where W_V and W_Q are the learnable weights, b_V and b_Q are the biases. $V_{reg} = [v_{reg_1}, \ldots v_{reg_K}]^T \in \mathcal{R}^{K\times d_s}$, where d_s is the dimension of the subspace and v_{reg_i} is the representation of the image region i which combines the visual feature, geometric feature and question feature.

Pair-wise combination: The pair-wise combination aims to combine all the image region representations in pairs to preliminarily construct the interactions between each two regions. The process can be formulated as follows:

$$V_{pc} = \tilde{V}_{pc} + \tilde{V}_{pc}^T \quad (5)$$

where $\tilde{V}_{pc} = [V_{reg}, \ldots, V_{reg}]_K$, which means that we repeat V_{reg} for K times to generate $\tilde{V}_{pc}$. We get the pair-wise combination, denoted as $V_{pc} \in \mathcal{R}^{K\times K\times d_s}$, between each two image regions by combining $\tilde{V}_{pc}$ and its transposition $\tilde{V}_{pc}^T$.

Inspired by [22], our proposed visual relational reasoning module captures the relationships between image regions by constraining the functional form of a neural network, just as the capacity to reason about spatial, translation invariant properties is built-in to CNNs. Different from [22] which only reasons about the pair-wise visual relationships, we not only mine the pair-wise visual relationships, but also reason about the inner-group visual relationships among multiple image regions. To achieve this aim, we design two parallel streams of CNNs, one with stacked 1×1 convolutional layers for inferring pair-wise relationships Rp, and the other with stacked dilated convolutional layers for inferring inner-group relationships Rg.

Pair-wise relational reasoning: As shown in the green dashed box in Fig. 4, on the top of the pair-wise combination region features V_{pc}, we reason about the pair-wise relationships by building N layers of 1×1 convolution operations each followed by a ReLU activation layer. The output of the last convolutional layer is $V_p \in \mathcal{R}^{K \times K}$. Finally, we compute the importance of each image region according to its role in the pair-wise relationships. Formally,

$$Rp = \text{softmax}(V_p + V_p^T) \tag{6}$$

where both $Rp_{i,j}$ and $Rp_{j,i}$ indicate the importance of the relationship between image region i and region j. Specifically, we apply $softmax$ over the sum of V_p and its transposition to make Rp a symmetric matrix,.

Inner-group relational reasoning: Inner-group relationships consider the relational information among multiple image regions. Compared to the pair-wise relational reasoning, a simple solution to reason inner-group relationships is to conduct $n \times n$ convolution operations on the pair-wise combination region features. The scope of the receptive field linearly depends on the depth of the convolution layers. In other words, larger receptive field requires more convolutional layers, which empirically depends on more trainning data and computational cost for better performance. In our model, we follow the idea in [40] to employ dilated convolutions that enable an exponentially large receptive field [41] while reducing the computational cost. As shown in the purple dashed box in Fig. 4, we build 3×3 dilated convolutional layers each followed by a ReLU activation layer on top of V_{pc}. Finally, we apply $softmax$ over the output of the last convolution denoted as $V_g \in \mathcal{R}^{K \times K}$to generate the importance distribution according to the inner-group relationships:

$$Rg = softmax(V_g + V_g^T) \tag{7}$$

Given the above pair-wise relationships and inner-group relationships, the visual feature of each image region can be updated as follows:

$$\tilde{v}_i = \sum_{j=1}^{K} (Rp_{i,j} \cdot v_j + Rg_{i,j} \cdot v_j) \tag{8}$$

In this way, the representation of each image region aggregates the information of itself as well as the relation-relevant regions. Finally, the relation-aware visual representation of the whole image, denoted as V_{vr}, is obtained by the sum of all the image region features:

$$V_{vr} = \sum_{i=1}^{K} \tilde{v}_i \tag{9}$$

IV. v-VRANet for Visual Question Answering

Visual question answering is a task that requires joint reasoning over the questions in natural language and the images. To achieve this with human-like abilities, we aim to learn a question-guided visual reprensentation which satisfies the following propoerties. First, the inter-object relationships mentioned in the question are embedded in the visual representation. Moreover, the visual representation simultanously encodes the relevant visual objects according to the question, so that it can infer more accurate answer based on both object-level and relationship-level clues.

Our v-VRANet model plugs the the proposed VRANet into the typical CNN+RNN architecture. As shown in Fig. 5(a), it treats visual question answering as a classification task and mainly contains three parts, including question modeling, image modeling and image-question joint embedding for answer prediction.

In the question modeling path (bottom in Fig. 5(a)), the model learns the question features $Q \in \mathcal{R}^{d_Q}$ by GRU based on word embeddings. In the image modeling path (top in Fig. 5(a)), we first detect K region proposals for use in bilinear visual attention and visual relational reasoning. According to [6], we use Faster R-CNN [42] in conjunction with the ResNet-101 to detect all the salient region proposals in an image. Then we take the output of the model and perform non-maximum suppression for each object category based on an IoU threshold. We then select all the regions, where any predicted category probability exceeds a confidence threshold. Finally, the top-K region proposals with highest category probabilities are selected for the use of v-VRANet. Then we represent these region proposals by their visual features $V \in \mathcal{R}^{K \times d_V}$ and geometirc features $G = [g_1, g_2, \ldots, g_K]^T \in \mathcal{R}^{K \times d_G}$. where $g_i = [\frac{x_i}{w}, \frac{y_i}{h}, \frac{w_i}{w}, \frac{h_i}{h}]$. (x_i, y_i), w_i, and h_i are the coordinates, width, and height of the selected region i respectively. w and h are the width and height of the input image respectively. Then the question features Q, visual features V and geometric features G are fed into the visual relational reasoning module to obtain the relation-aware visual features V_{vr}. Meanwhile, the question features Q and the visual features V are fed into the bilinear visual attention module for the question-attended visual features V_{att}.

For the answer prediction, we jointly embed the question features Q and the two types of visual features, i.e. V_{vr} and V_{att}, as follows:

$$f_{vqa} = (W_r V_{vr} \circ W_a V_{att}) \circ W_q Q \tag{10}$$

where $W_r \in \mathcal{R}^{d_f \times d_V}$, $W_a \in \mathcal{R}^{d_f \times d_V}$, and $W_q \in \mathcal{R}^{d_f \times d_Q}$ are learnable weight matrices and the bias terms are omitted without loss of generality. d_f is the dimension of the fused vector. Given the fused image-question features, we compute the probability of answer a_i using a simple two-layer MLP with ReLU nonlinearity in its hidden layer and Sigmoid activation in the last layer:

$$P(a_i|I, q) = \sigma(MLP(f_{vqa}))_i \tag{11}$$

We choose the answer with the maximum probability from all of the candidates as the final prediction. Accordingly, we use the cross entropy loss to train the model.

V. c-VRANet for Cross-Modal Information Retrieval

Cross-modal information retrieval between images and texts requires comparing the similarity across these two modalities on semantic level. The typical solution is to map the image features and text features into a common semantic space for similarity measurement. However, semantically relevant data

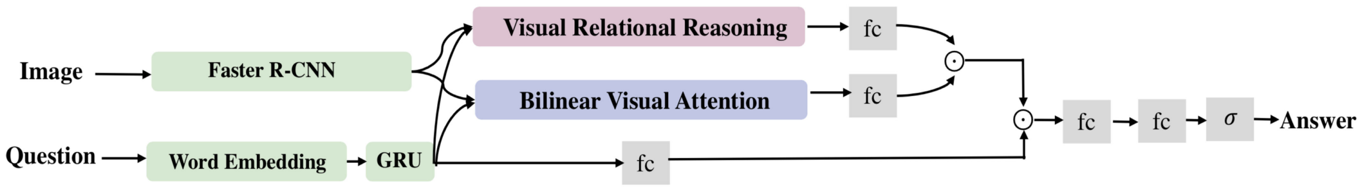


Fig. 5. v-VRANet: The network architecture for VQA based on our proposed VRANet.

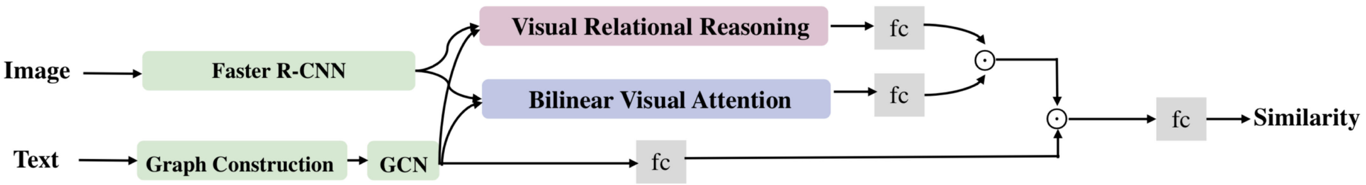


Fig. 6. c-VRANet: The network architecture for CMIR based on our proposed VRANet.

from different modalities usually contians imbalanced information. Aligning all the modalities in the same space will weaken the modality-specific semantics and introduce unexpected noise. An alternative solution is based on feature fusion that the cross-modal similarity is directly predicted by analyzing the fine-grained correlations between the image features and the text features. The key of this solution is to enhance visual representation and textual representation and then fuse them together.

As shown in Fig. 6, our c-VRANet especially enhances visual representation using our proposed VRANet, while extracts textual representation using Graph Convolutional Networks (GCN) to explore the structural semantics in the text for image-text retrieval. And then our model fuses these representations from different modalities to obtain the image-text joint embedding to capture their shared relationships and concepts for similarity measurement. As shown in Fig. 6, we use pre-trained faster R-CNN in conjection with ResNet-101 to detect and represent K objects in an image, which is same as the image modelling path in the v-VRANet model. Then, under the guidance of the text, the visual features are enhanced via the visual relational reasoning module to encode the visual relationships occuring in the text. Meanwhile, the visual regions highly relevant to the text are emphasied in the visual features via the bilinear visual attention module. In contrast to the v-VRANet which uses GRU to encode the question features, the c-VRANet leverages the Graph Convolutional Networks (GCN) to encode the text features in the same way as the state-of-the-art CMIR model [14] for fair comparison.

We jointly embed the text features Q and the two type of visual features, i.e. V_{vr} and V_{att}, as follows:

$$f_{cmir} = (W_r V_{vr} \circ W_a V_{att}) \circ W_q Q \tag{12}$$

where $W_r \in \mathcal{R}^{d_f \times d_V}$, $W_a \in \mathcal{R}^{d_f \times d_V}$, and $W_q \in \mathcal{R}^{d_f \times d_Q}$ are learnable weight matrices. Given the fused image-text features, we compute the cross-modal similarity score between the image Img and the text Txt using a simple fully-connected layer with Sigmoid nonlinearity:

$$Score(Img, Txt) = \sigma(FC(f_{cmir})) \tag{13}$$

We adopt pair-wise similarity loss function in [43] as our optimization objective. Specifically, we maximize the mean similarity score u_p between matching image-text 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 σ_n^2 pairs, is added to the loss function to accelerate convergence. The loss function is defined as:

$$loss = (\sigma_p^2 + \sigma_n^2) + \lambda \max(0, m - (u_p - u_n)) \tag{14}$$

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.

VI. Experiments

A. Evaluation on Visual Question Answering

1) Datasets and Evaluation Metrics: We mainly evaluate our VQA models on VQA 2.0 [44] dataset. VQA 2.0 is based on Microsoft COCO image data [45]. The dataset contains 443,757 train questions, 214,354 validation question, 447,793 test questions (named as $test\text{-}standard$), generated on 123,287 images. There is also a 25% subset of the $test\text{-}standard$ referred to as $test\text{-}dev$. The questions are classified into three categories: *yes/no*, *number* and *other*. About 50% of the questions fall into the *other* category. 10 free-response answers are generated for each question by different voters. We report our results on the challenging Open-Ended task.

We choose the answers appearing more than 9 times in the training set to form the set of candidate answers, resulting in 3129 candidate answers. Following previous work, we train our model on $train{+}val$ splits and report the $test\text{-}dev$ and $test\text{-}standard$ results from the VQA evaluation server (except for the ablation study). We use the tools provided by [44] to evaluate the accuracy of the predicted answer a:

$$Accuracy(a) = \min\left(\frac{count(a)}{3}, 1\right) \tag{15}$$

where $count(a)$ is the count of the answer a voted by different annotators.

We also test our model on the CLEVR dataset [23], which is a diagnostic dataset that tests a range of visual reasoning abilities. CLEVR contains 100,000 images of 3D-rendered objects, such as spheres and cylinders. 699,989 questions-answer pairs and about 70,000 images are used as training set, 149,991 question-answer pairs and about 15,000 images are used as validation set, and 14,988 question-answer pairs and about 15,000 images are used as test set. Different from VQA 2.0, each question in CLEVR has only one ground-truth answer over 28 candidate answers in the training set. Typically, the classification accuracy [46]–[48] is used as the evaluation criteria.

2) Experimental Setup: For both VQA 2.0 and CLEVR datasets, we use the Adamax solver with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Gradient clipping techniques are used. The batch size is set to be 512. For encoding questions, each word is embedded into a 300-dimensional vector. The hidden state of GRU is set to be a 1024-dimensional vector. For the VQA 2.0 dataset, we set the number of region proposals K to 36 according to the settings in [6]. For the CLEVR dataset, we follow the settings in [22] since there is no real-world objects in the images of CLEVR dataset. Instead of relying on the visual features of the region proposals extractd by Faster R-CNN, we train our VQA model end-to-end together with a relatively small CNN composed of 4 layers of convolutions. Each convolution layer contains 128 kernels with the kernal size 3×3 and strides 2, followed by ReLU activation and batch-norm. Thus each image is represented as a $8 \times 8 \times 128$ tensor. Each word is embedded by a 64-dimensional vector and fed into a single layer of GRU whose hidden state dimension is set to be 128. The other settings follow [22].[1]

3) Ablation Study: The architecture of our full VQA model v-VRANet is composed of multiple essential modules and has several important hyper-parameters to be set. In this experiment, we conduct ablation tests to evaluate the contribution of each module to the final prediction accuracy. Using VQA 2.0, we evaluate several variant versions of our v-VRANet model by training them on the train split and reporting the performance on the val split. The variant versions of our v-VRANet model include:

- **CNN+GRU** model: we remove both the visual relational reasoning module and the visual attention module from the full model in Fig. 5.
- **CNN+GRU+ATT** model: we only remove the visual relational reasoning module from the full model in Fig. 5.
- **CNN+GRU+VRR** model: we only remove the visual attention module from the full model in Fig. 5.
- **v-VRANet** model: the proposed full VQA model in Fig. 5.

We also analyse several important hyper-parameters in our models, including the dimension of the final fused image-question representation denoted as d_f, the dimension of the subspace in our visual relational reasoning module denoted as d_s, and the second dimension of the projection matrix H and U in our visual attention module denoted as d. The prediction accuracy and the model size are shown in Table I.

TABLE I
ABLATION STUDY ON EACH MODULE AND HYPER-PARAMETERS OF OUR MODEL USING THE VAL SPLIT OF THE OPEN-ENDED TASK (VQA 2.0). * INDICATES THE HYPER-PARAMETERS USED IN THE FINAL MODEL. THE MODEL SIZE ILLUMINATES THE TOTAL NUMBER OF MODEL'S LEARNABLE PARAMETERS. M MEANS MILLION

Model	Acc.(%)	Model size(M)
CNN+GRU		
- d_f=512	56.54	14.6
- d_f=1024*	59.00	20.7
- d_f=2048	58.33	47.7
CNN+GRU+ATT ($d_f = 1024$)		
- d=1024	63.71	23.7
- d=2048	64.02	26.7
- d=3072*	64.18	29.7
- d=4096	63.95	32.7
CNN+GRU+VRR ($d_f = 1024$)		
- d_s=128	63.12	21.2
- d_s=192	63.50	21.5
- d_s=256*	63.61	21.8
- d_s=384	63.57	22.7
CNN+GRU+VRR ($d_f = 1024$, $d_s = 256$)		
- only pair-wise relations	63.41	21.5
- only inner-group relations (3x3)	63.47	21.7
- only inner-group relations (5x5)	63.00	22.5
- only inner-group relations (7x7)	61.25	23.4
v-VRANet ($d_f = 1024$, $d_s = 256$, $d = 3072$)	64.42	32.8

The first block in Table I shows the performance of our CNN+GRU model with different settings of the dimension of the final fused image-question representation denoted as d_f. It is obvious that d_f has significant impact on model performance and size. We have tried $d_f = 512$, $d_f = 1024$, and $d_f = 2048$ and found that $d_f = 1024$ results in the best performance.

The second block in Table I indicates the impacts of bilinear visual attention module. When $d = 4096$, the performance has achived saturation. Compared with the CNN+GRU model, the best performance of CNN+GRU+ATT significantly improves the prediction accuracy from 59.00% to 64.18%, which proves that the question-driven visual attention provides informative semantics for the answer prediction.

The third block in Table I demonstrates that the subspace dimension in the visual relational reasoning module slightly affects the prediction performance as well as the size of the model. $d_s = 256$ results in best performance and we choose this setting in the following experiments. Compared with the CNN+GRU model, we can see that our proposed visual relational reasoning module significantly improves the prediction accuracy from 59.00% to 63.61%, with the increase of only 1.1 M learnable parameters.

To evaluate the impact the pair-wise relational reasoning component and the inner-group relational reasoning component in the visual relational reasoning module, we design two variants of CNN+GRU+VRR model and the results are shown the fouth block in Table I. We tried 3×3, 5×5, and 7×7 kernel sizes. The results show that: (1) 3×3 convolutional kernel seems most effective to model inner-group relations. (2)

[1] [Online]. Available: http://github.com/rosinality/relation-networks-pytorch

TABLE II
COMPARISON OF PERFORMANCE OF OUR v-VRANet WITH THE STATE-OF-THE-ART MODELS ON VQA 2.0. ALL OF THESE RESULTS ARE OBTAINED BY SINGLE MODEL TRAINED WITH VQA 2.0 TRAIN AND VAL SPLIT

Model	test-dev				test-standard			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
Prior[49]	-	-	-	-	26.98	01.17	00.36	61.20
Language only [49]	-	-	-	-	44.26	27.37	31.55	67.01
LSTM+CNN[49]	-	-	-	-	54.22	41.83	35.18	73.46
MCB reported in [49]	-	-	-	-	62.27	53.36	38.28	78.82
MLB [30]	-	-	-	-	65.07	54.77	37.90	**84.02**
MFB [50]	64.98	-	-	-	-	-	-	-
MFH [38]	65.80	-	-	-	-	-	-	-
SLA[51]	63.89	55.86	40.35	79.95	64.06	55.82	40.63	80.01
ACMN[52]	63.81	53.07	44.18	81.59	64.05	53.22	43.80	81.83
DCN [20]	66.60	56.72	**46.60**	**83.50**	67.00	56.90	**46.93**	83.89
Bottom-up [6]	65.32	56.05	44.21	81.82	65.67	56.26	43.90	82.20
DRA [53]	66.45	57.40	44.78	82.85	66.85	57.63	44.37	83.35
Bottom-up + MLB	65.60	56.41	44.63	81.95	65.97	56.76	44.42	82.22
Bottom-up + MFB	65.75	56.53	44.88	82.12	66.09	56.78	44.78	82.40
Bottom-up + MFH	65.99	56.71	45.11	82.42	66.20	56.80	44.60	82.68
Bottom-up + ATT + MFH	67.18	57.98	45.98	83.23	67.27	57.68	44.85	83.40
Bottom-up + VRR	66.05	56.96	45.40	82.22	66.26	57.07	44.77	82.34
v-VRANet	**67.20**	**58.41**	45.51	83.31	**67.34**	**58.49**	44.96	83.39

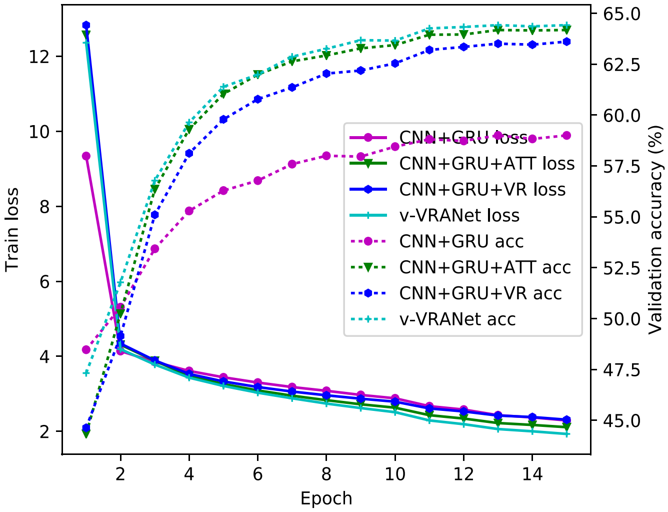


Fig. 7. The training loss and validation accuracy vs. epochs of our VQA models. Cross-Entropy loss is used for all the methods.

Simultaneously modeling pair-wise and inner-group relations can promote the model's performance.

Based on the above observations, we set $d_f = 1024$, $d_s = 256$, $d = 3072$ in our v-VRANet model and the following experiments are conducted with these settings. Compared with the CNN+GRU model, v-VRANet remarkably increases the prediction accuracy by 5.42% as shown in the last block in Table I. We also shows the he training loss and validation accuracy of all the above models in the Fig. 7.

4) Comparison With State-of-the-Art Models: In Table II, we compare our results on the VQA 2.0 dataset to the state-of-the-art models. All the results are obtained by single model trained on the $train$+val split. This table is splitted into three blocks. The first block shows several models without object detection. In the middle block, all the models are designed on top of the pre-trained Faster R-CNN for object detection. Thereinto, Bottom-up model [6] is the winner of VQA challenge 2017. DRA [53] is a recently proposed model using Faster R-CNN features without visual relational reasoning. We also implement three models based on the popular feature fusion approaches, i.e. MLB [30], MFB [38] and MFH [50], on top of the same bottom-up image features [6]. We name these models as Bottom-up+MLB, Bottom-up+MFB, Bottom-up+MFH respectively. The model Bottom-up in Table II utilizes the simple element-wise sum to fuse all the region features. It is worth to note that the MFB and MFH model in the first block use complicated attention mechanism such as question self attention and image-question co-attention. Their results are given in the second block. We also add bilinear attention to Bottom-up+MFH to generate Bottom-up+ATT+MFH model. We show the results of our proposed models in the third block.

It is obvious that v-VRANet outperforms all the state-of-the-art models, which proves the effectiveness of our proposed relational reasoning module. Comparing the results in the first and the second blocks, we can see that models on top of the pre-trained Faster R-CNN features obtain general improvement compared to the models without object detection. Therefore, we leverage the pre-trained Faster R-CNN networks to extract salient visual features in v-VRANet.

Despite the benefit from the aforementioned Faster R-CNN features, our model achieves superior performance due to the proposed two modules by compareing the results in the second and the third blocks. It's obvious that v-VRANet obtains further improvements over all the question types, i.e. *Other* with 4.2%, *Number* with 2.9%, and *Yes/No* with 1.8%. Compared to DCN, our model predicts more accurate answers for *Other* questions, which frequently begin with "*what*," "*where*," "*which*," "*why*," or "*who*". Answering the *Other* questions requires reasoning ability of the VQA models. About half of the questions in VQA 2.0 fall into the *Other* type. Since our model is capable of reasoning, it achieves the best overall accuracy, though its performance is inferior to the DCN model on the *Number* or *Yes/No*

TABLE III
COMPARISON OF PERFORMANCE OF OUR v-VRANet WITH THE STATE-OF-THE-ART MODELS ON CLEVR. ALL OF THESE RESULTS ARE OBTAINED BY USING SINGLE MODEL. * DENOTES USE OF EXTRA SUPERVISORY INFORMATION THROUGH PROGRAM LABELS

Model	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
CNN+LSTM [54]	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA+MLP [23]	73.2	59.7	77.9	75.1	80.9	70.8
PG+EE (700K prog.)* [54]	96.9	92.7	97.1	98.7	98.1	98.9
N2NMN* [55]	83.7	68.5	85.7	84.9	90.0	88.7
N2NMN* [55]	83.7	68.5	85.7	84.9	90.0	88.7
Simple RN [22]	95.5	90.1	97.8	93.6	97.9	97.1
FiLM [32]	97.6	94.5	99.2	93.8	99.2	99.0
MAC [56]	**98.9**	**97.2**	**99.5**	**99.4**	**99.3**	**99.5**
v-VRANet	96.1	93.4	97.1	95.1	97.4	96.8

question types. For the DCN model, the proposed multi-level visual features and stacked dense co-attention mechanism have great benefit for answering the *Number* and *Yes/No* questions. Inspired by DCN, we believe that exploiting more informative visual representations as well as multi-step reasoning can gain further improvement on the VQA task.

In addition, to further demonstrate the efficiency of our VRR module, we design Bottom-up+VRR model without bilinear attention. First, we compare the performance of Bottom-up+MFH with Bottom-up+VRR, which are differing in that Bottom-up+VRR uses our proposed VRR module and uses element-wise multiplication to implement visual-textual feature fusion while Bottom-up+MFH uses MFH for visual-textual feature fusion and without VRR. We can see that Bottom-up+VRR achieves slightly better performance compared with Bottom-up+MFH while the model size of Bottom-up+VRR is much smaller than that of Bottom-up+MFH. Bottom-up+VRR only needs 21.7M parameters while Bottom-up+MFH requires 54.5 M parameters. It proves that our VRR module can achieve competitive performance with MFH with much smaller network size. Second, we add a new baseline Bottom-up+ATT+MFH to use bilinear visual attention to further enhance the bottom-up features. From the results in Table II we can also obtain the same conclusion that v-VRANet achieves slight improvement with significant smaller model size compared with Bottom-up + ATT + MFH.

Finally, we compare the performance of v-VRANet on the CLEVR dataset with state-of-the-art work in Table III. Our model achieves superior performance compared with most of the existing approaches. It is worth to note that simple RN [22] which realizes pair-wise computational model, achieves 95.5% performance on the test set. Compared with the simple RN, our model achieves better performance, achieving 96.1% overall accuracy on the test set. These results demonstrate the generality of our model. It is shown that our model is inferior to MAC [56] and FiLM [32]. Since CLEVR dataset contains highly compositional questions requiring progressive reasoning processes. MAC gains such abilities by a stacked attention-based reasoning model which decomposes the question into a series of reasoning steps. FiLM enables recurrent neural network over the input question to conduct convolution operations over the image based on conditional information, which achieves multi-step reasoning. Hence, stacking our VRANet module to achieve multi-step visual relational reasoning is one of our future work.

TABLE IV
COMPARISON OF THE COMPUTATIONAL TIME AND MODEL SIZE ON THE VQA 2.0 DATASET. THE UNIT OF TRAINING AND TESTING TIME IS MILLISECOND PER ITERATION. THE UNIT OF MODEL SIZE IS MILLION

Model	Training time (ms)	Testing time (ms)	Model size (M)
Bottom-up	419	250	26.6
Bottom-up+MLB	422	257	30.1
Bottom-up+MFB	435	264	38.1
Bottom-up+MFH	444	277	54.5
v-VRANet	429	260	32.8

5) Computational Cost: In Table IV, we compare the computational complexity and model size of our v-VRANet model with several fusion-based VQA models, including Bottom-up, Bottom-up+MLB, Bottom-up+MFB, Bottom-up+MFH, which are introduced in Section VI-A4. In both training and testing process, we record the running times for each iteration and report the mean time in millisecond per iteration (each iteration contains 512 samples). We also report the total number of each model's learnable parameters. This experiment is conducted on a machine with two Intel Xeon E5 CPUs and two Nvidia Telsa V100 GPUs. From Table IV and Table II, we can see that our v-VRANet requires comparable computational complexity and model size with the popular fusion-based VQA models, but achieves best performance on VQA task.

6) Qualitative Analysis: Typical examples of the image-question pairs along with the predicted answers are shown in Fig. 8. Only using bilinear visual attention module results in the highest attention weight to the regions in red box according to the question, but fails to attend to the other essential regions (green boxes), such as *"kite"* in Fig. 8(a) and *"cissor"* in 8(b) since the questions do not mention these words. Thus, the final visual representation has few information of these important regions, leading to incorrect predictions of our CNN+GRU+ATT model. Our v-VRANet model, equipped with both visual relational reasoning module and visual attention module, reasons about the relationships between image regions. We show the top four related regions of the most attended region by v-VRANet for each example. According to equation (8), our v-VRANet model generates the features of the region in red box by fusing all these regions' visual features and their relationships, leading to higher confidence on the correct answers.

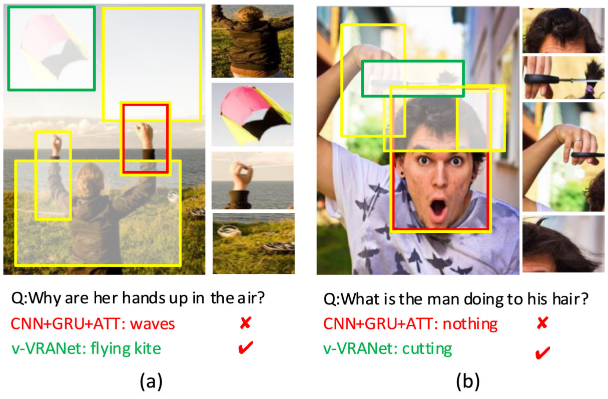


Fig. 8. Visualization of two typical VQA examples. On the left side of each example, the yellow boxes are the top four attended regions and the red box shows the region with highest attention weight based on the CNN+GRU+ATT model. The green boxes shows the essential region for answer prediction but obtaining low attention weights by CNN+GRU+ATT. Our v-VRANet has relational reasoning ability and we shows the top four relevant regions to the most attended region on the right side of each whole image. The bottom of each image lists the corresponding question and answers.

B. Evaluation on Cross-Modal Information Retrieval

1) Datasets and Evaluation Metrics: In this section, we test our models on two popular CMIR benchmark datasets. The first is Cross-Modal Places [57] (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 the score for text-queries Q_T and the score for image-queries Q_I.

We also test our model on the Microsoft COCO (MS-COCO) dataset [58]. Each image in MS-COCO has 5 captions. We use the splits of [59] which contains 82783 train images, 5000 validation images and 5000 test images. However there are also 30504 images originally in the validation set but have been left out in this split which is referred as *rV*. Following the typical settings in recent proposed models [60], [61], the results on 1 K test images are reported by averaging the results over 5 folds of 1 K test images. For the 5 K test images, the results are reported on the full 5 K test images. Following existing approaches [60], we also use *rV* to augment our training set. To compare our model with state-of-the-arts, we use the same evaluation metrics as those in [60]–[66], i.e., R@K, defined as the percentage of queries in which the ground-truth matchings are contained in the first K retrieved results. Higher value of R@K stands for better performance. We also use Med *r*, which is the median rank of the first retrieved ground-truth text or image. Lower value of Med *r* means better performance.

2) Experimental Setup: On CMPlaces dataset, we train our models by Adamax solver with 15 epochs with mini-batch size 256. We adopt the learning rate of 0.001 and dropout ratio of 0.5. m and λ in the loss of c-VRANet are set to 0.6 and 0.35, respectively. We set the number of region proposals K to 36 according to the settings in [6]. For the MS-COCO dataset, we apply GRU instead of GCN to extract the textural features, since GCN proposed in [14] is better for modelling long texts and has worse performance for the short captions in MS-COCO from our empirical study. We also train our model by Adamax solver, but with 30 epoches with mini-batch size 128 and initial learning rate of 0.0002. All of our experiments are conducted on a machine with two Intel Xeon E5 CPUs and two Nvidia Telsa V100 GPUs.

TABLE V
ABLATION STUDY ON EACH MODULE OF OUR C-VRANet MODEL ON THE CMPlaces DATASET

Model	Image Query	Text Query	Average
CNN+GCN	34.2	25.9	30.0
CNN+GCN+ATT	35.9	25.7	30.8
CNN+GCN+VRR	36.4	27.2	31.8
c-VRANet	**39.5**	**27.3**	**33.4**

3) Ablation Study: In this section, we conduct ablation experiments on CMPlaces dataset to evaluate the influence of the essential components in our proposed CMIR model on the retrieval accuracy. Similar to the ablation study for VQA, the ablated versions of our proposed CMIR model include:

- **CNN+GCN** model: we remove the visual relational reasoning module and the visual attention module from the full model as shown in Fig. 6. Specifically, the image features obtained by ResNet and the text features from GCN are directly fused by element-wise multiplication operation for the similarity measurement.
- **CNN+GCN+ATT** model: we only remove the visual relational reasoning module from the full model in Fig. 6.
- **CNN+GCN+VRR** model: we only remove the visual attention module from the full model in Fig. 6.
- **c-VRANet** model: the full CMIR model as shown in Fig. 6.

According to the best settings in the ablation study for VQA, we directly set the hyper-parameters $d_f = 1024$, $d_s = 256$, $d = 3072$ in the aforementioned ablated models. In all the models, the image features and text features are fused by element-wise multiplication.

Table V shows the ablation study results. The CNN+GCN model performs the worst since that it treats the image regions independently and ignores their relationships and visual importance. The CNN+GCN+ATT model gains some improvement over the CNN+GCN by applying the visual attention to focus on the potential image regions described by the text. The CNN+GCN+VRR model achieves about 1.8% improvement over the CNN+GCN by considering the relationships between image regions relevant to the text. c-VRANet is our full model. Obviously, it consistently outperforms the other incomplete solutions and results in a boost up to 3.3% compared to the CNN+GCN. It verifies that text-relevant region properties and inter-region relationships can provide more informative clues for image-text correlation learning and similarity measurement.

4) Comparison With State-of-the-Art Models: Table VII shows the comparison between our model and state-of-the-art

TABLE VI
COMPARISON OF PERFORMANCE OF OUR MODEL WITH THE STATE-OF-THE-ART METHODS ON MS-COCO DATASET

Model	Image Query				Text Query			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
1K Test Images								
DVSA [59]	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
GMM-FV [67]	39.4	67.9	80.9	2.0	25.1	59.8	76.6	4.0
m-CNN [62]	42.8	73.1	84.1	2.0	32.6	68.6	82.8	3.0
VQA-A [68]	50.5	80.1	89.7	-	37.0	70.9	82.9	-
HM-LSTM [63]	43.9	-	87.8	2.0	36.1	-	86.7	3.0
Order-embedding[61]	46.7	-	88.9	2.0	38.9	-	85.9	2.0
DSPE+FV[64]	50.1	79.7	89.2	-	39.6	75.2	86.9	-
sm-LSTM[65]	53.2	83.1	91.5	1.0	40.7	75.8	87.4	2.0
two-branch net[66]	54.9	84.0	92.2	-	43.3	76.4	87.5	-
CMPM(ResNet-152)[69]	56.1	86.3	92.9	-	44.6	78.8	89.0	-
VSE++(fine-tuned)[60]	57.2	-	93.3	1.0	45.9	-	89.1	2.0
c-VRANet(ours)	**58.1**	**86.9**	**93.4**	**1.0**	**50.4**	**83.6**	**92.3**	**1.0**
5K Test Images								
DVSA [59]	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0
GMM-FV [67]	17.3	39.0	50.2	10.0	10.8	28.3	40.1	17.0
VQA-A [68]	23.5	50.7	63.6	-	16.7	40.5	53.8	-
Order-embedding [61]	23.3	-	65.0	5.0	18.0	-	57.6	7.0
CMPM(ResNet-152)[69]	31.1	60.7	73.9	-	22.9	50.2	63.8	-
VSE++(fine-tuned)[60]	32.9	-	74.7	3.0	24.1	-	66.2	5.0
c-VRANet(ours)	**34.4**	**63.8**	**76.0**	3.0	**27.8**	**57.6**	**70.8**	**4.0**

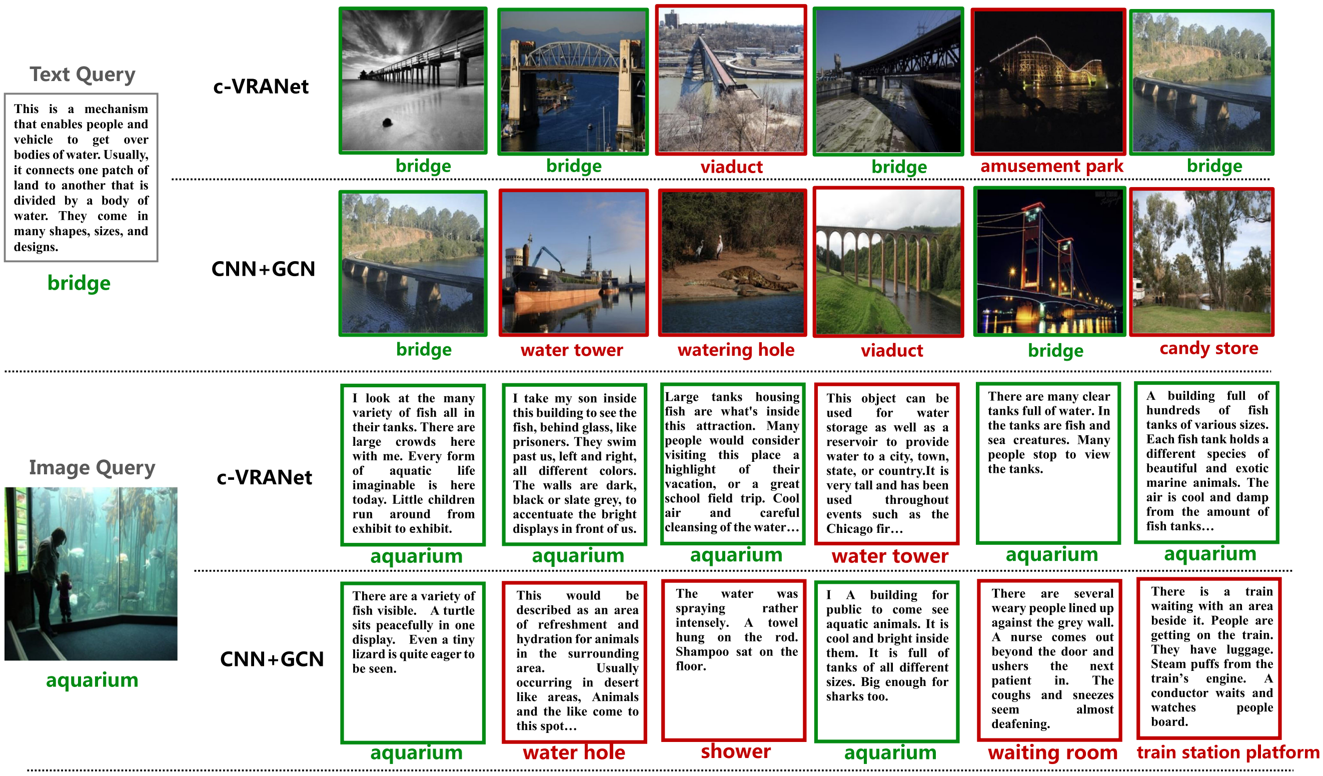
TABLE VII
MAP SCORE COMPARISION OF OUR MODELS WITH STATE-OF-THE-ART MODELS ON CMPlaces DATASET

Model	Image Query	Text Query	Average
BL-Ind [57]	6.0	8.0	7.0
BL-ShFinal [57]	12.7	3.3	8.0
BL-ShAll [57]	6.0	8.0	7.0
Tune(Free) [57]	18.1	5.2	11.7
Tune+StatReg(GMM) [57]	22.1	15.1	18.6
GIN [14]	23.9	25.9	24.8
c-VRANet	**39.5**	**27.3**	**33.4**

approaches, including *GIN* [14] and Castrejon's models [57] proposed as baseline models on the CMPlaces dataset. *GIN* achieves the state-of-the-art performance on several benchmark datsets and we test its performance on the CMPlaces dataset. The results of Castrejon *et al.* [57] are from their published paper. Both *GIN* [14] and Castrejon 's models [57] are relevant to our model in the sense that they use dual-path subnetworks to model images and texts. However, their approaches are limited by the "flat" image representations, which ignores the inherent fine-grained visual relationships and the guidance of the text. As we can see, our model significantly outperforms other approaches on all the situations. Specifically, the image queries benefit more from the visual reasoning and attention mechanism than the text queries, since our VRANet module is proposed mainly for enhancing the expressive capacity and robustness of the visual representation.

We also report the comparison of our model with the state-of-the-arts on the MS-COCO dataset in Table VI. We can see that our model outperforms all the existing models, which further demonstrates the advantages of our model. Compared with VSE++, our model achieves 4.5% and 2.3% improvement on R@1 for the text query on 1 K and 5 K test images, respectively. Meanwhile, our model also slightly outperforms VSE++ on R@1 for the image query. Despite the compared models, most recent work [36], [70] exploited fine-grained correlations between the images and the texts, such as cross-modal attention and cross-modal graph convolution, and achieved superior improvement on CMIR tasks. We believe that our relational reasoning module has complementary advantages with cross-model correlation learning, which will be a promising future work.

5) Qualitative Analysis: In contrast to the typical dual-path schemes that model images and texts independently, e.g. the CNN+GCN model introduced in Section VI-B3, our c-VRANet simultaneously considers the text-relevant visual objects and inter-object relationships. For qualitative comparison, Fig. 9 shows two examples from the CMPlaces dataset for text-query-image and image-query-text tasks. In the top example, to retrieve images of *"bridge"* as described in the text query, CMIR models need to identify the properties of *"bridge"* and *"water"* and their relationships of *"get over"*. As seen, compared with CNN+GCN model, c-VRANet retrieves more images belonging to *"bridge"* category and semantically relevant to the text content. The third incorrect image belongs to *"viaduct"*, which is a fine-grained category of *"bridge"* and that they have obvious overlap from semantic view. In contrast, the top retrieved results of the CNN+GCN model contain more incorrect ones. Though the model identifies some relevant visual content, e.g. *"water"*, *"vehicle"*, and *"land"*, it fails to reason about their relationships and leads to irrelevant results like *"water tower"* and *"water hole"*. In the bottom example, we come to the similar conclusion with the text query. To understand the semantics of *"aquarium"*, c-VRANet identifies both the objects like *"water"*, *"fish"*, *"people/crowds"* and their inter-object relationships. Furthermore, the CNN+GCN model without visual attention and relational reasoning is easily confused with other scenes, e.g. *"water hole"*, *"shower"*, etc. In these examples, we can see that c-VRANet is effective in fine-grained cross-modal correlation learning by incorporating with text-guided visual attention and visual relational reasoning.



This is a mechanism that enables people and vehicle to get over bodies of water. Usually, it connects one patch of land to another that is divided by a body of water. They come in many shapes, sizes, and designs.

I look at the many variety of fish all in their tanks. There are large crowds here with me. Every form of aquatic life imaginable is here today. Little children run around from exhibit to exhibit.

I take my son inside this building to see the fish, behind glass, like prisoners. They swim past us, left and right, all different colors. The walls are dark, black or slate grey, to accentuate the bright displays in front of us.

Large tanks housing fish are what's inside this attraction. Many people would consider visiting this place a highlight of their vacation, or a great school field trip. Cool air and careful cleansing of the water…

This object can be used for water storage as well as a reservoir to provide water to a city, town, state, or country.It is very tall and has been used throughout events such as the Chicago fir…

There are many clear tanks full of water. In the tanks are fish and sea creatures. Many people stop to view the tanks.

A building full of hundreds of fish tanks of various sizes. Each fish tank holds a different species of beautiful and exotic marine animals. The air is cool and damp from the amount of fish tanks…

There are a variety of fish visible. A turtle sits peacefully in one display. Even a tiny lizard is quite eager to be seen.

This would be described as an area of refreshment and hydration for animals in the surrounding area. Usually occurring in desert like areas, Animals and the like come to this spot…

The water was spraying rather intensely. A towel hung on the rod. Shampoo sat on the floor.

I A building for public to come see aquatic animals. It is cool and bright inside them. It is full of tanks of all different sizes. Big enough for sharks too.

There are several weary people lined up against the grey wall. A nurse comes out beyond the door and ushers the next patient in. The coughs and sneezes seem almost deafening.

There is a train waiting with an area beside it. People are getting on the train. They have luggage. Steam puffs from the train's engine. A conductor waits and watches people board.

Fig. 9. Cross-modal retrieval examples of our proposed c-VRANet model and the CNN+GCN model on the CMPlaces dataset. Relevant results are highlighted with green boxes while irrelevant results are marked with red boxes.

VII. Conclusion

In this paper, we propose a Visual Reasoning and Attention Network (VRANet) to enhance the visual representations for cross-modal reasoning and retrieval. The VRANet satisfies the following properties. First, the text-guided pair-wise and inner-group visual relationships are encoded via the visual relational reasoning module. Meanwhile, the text-relevant essential visual region information is embedded in the enhanced visual features via the bilinear visual attention module. To evaluate the effectiveness of our modules, we inject VRANet into the models for Visual Question Answering and Cross-Modal Information Retrieval tasks. Extensive experiments prove that baseline models are effectively enhanced by our proposed modules, resulting in significant improvement compared with state-of-the-art approaches. In the future work, we will go further to study the explicit reasoning schemes to enhance the interpretability of the model.

References

[1] C. Chaudhary, P. Goyal, D. Prasad, and Y. Chen, "Enhancing the quality of image tagging using a visio-textual knowledge base," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 1372–1382, Aug. 2018.

[2] X. Ke, J. Zou, and Y. Niu, "End-to-end automatic image annotation based on deep CNN and multi-label data augmentation," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2093–2106, Aug. 2019.

[3] W. Zhang, H. Hu, and H. Hu, "Training visual-semantic embedding networks for boosting automatic image annotation," *Neural Process. Lett.*, vol. 48, no. 3, pp. 1503–1519, 2018.

[4] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," in 2015, *arXiv:1512.02167v2*.

[5] Z. Yang, X. He, and J. Gao, "Stacked attention networks for image question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 21–29.

[6] P. Anderson, X. He, C. Buehler, and D. Teney, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6077–6086.

[7] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for visual question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6135–6143.

[8] J. Lei, L. Yu, M. Bansal, and T. Berg, "TVQA: Localized, compositional video question answering," in *Proc. EMNLP*, 2018, pp. 1369–1379.

[9] J. Liang *et al.*, "Focal visual-text attention for memex question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1893–1908, Aug. 2019.

[10] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," in 2019, *arxiv: 1904.11574*.

[11] W. Jin *et al.*, "Multi-interaction network with object relation for video question answering," in *Proc. ACM MM*, 2019, pp. 1193–1201.

[12] Z. Zhang, Z. Zhao, Z. Lin, J. Song, and X. H, "Open-ended long-form video question answering via hierarchical convolutional self-attention networks," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4383–4389.

[13] Z. Zhao, Z. Zhang, X. Jiang, and D. Cai, "Multi-turn video question answering via hierarchical attention context reinforced networks," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3860–3872, Aug. 2019.

[14] J. Yu *et al.*, "Modeling text with graph convolutional network for cross-modal information retrieval," in *Proc. Pacific Rim Conf. Multimedia.*, 2018, pp. 223–234.

[15] L. Wang, Y. Li, and L. Svetlana, "Learning deep structure-preserving image-text embeddings," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5005–5013.

[16] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. Eur. Conf. Comput. Vision*, 2017, pp. 1908–1917.

[17] J. Dong, X. Li, and D. Xu, "Cross-media similarity evaluatio for web image retrieval in the wild," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2371–2384, Sep. 2018.

[18] Z. Zhang, Z. Lin, Z. Zhao, and Z. Xiao, “Cross-modal interaction networks for query-based moment retrieval in videos,” in *Proc. Special Interest Group Inf. Retrieval*, 2019, pp. 655–664.

[19] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in *Proc. Europ. Conf. Comput. Vision*, 2018, pp. 711–727.

[20] D. Nguyen and T. Okatani, “Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6087–6096.

[21] H. Nam, J. W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2156–2164.

[22] R. D. Santoro, A. D. Barrett, and M. Malinowski, “A simple neural network module for relational reasoning,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4967–4976.

[23] J. Johnson, B. Hariharan, L. Maaten, and F.-F. Li, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1988–1997.

[24] P. Wang, Q. Wu, and C. Shen, “The VQA-machine: Learning how to use existing vision algorithms to answer new questions,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3909–3918.

[25] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[26] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, “Structured attention networks,” in *Proc. Int. Conf. Learn. Representations*, 2017.

[27] A. Vaswani, N. Shazeer, and N. Parmar, “Attention is all you need,” in *Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. 37th Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[29] A. Fukui, D. H. Park, D. Yang, A. Rohrbach T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 457–468.

[30] J. Kim, K. On, W. Lim, and B. Zhang, “Hadamard product for low-rank bilinear pooling,” in *Proc. Int. Conf. Learn. Representations*, 2017.

[31] N. Hyeonwoo, S. Paul, and B. han, “Image question answering using convolutional neural network with dynamic parameter prediction,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 30–38.

[32] E. Perez, F. Strub, H. Vries, V. Dumoulin, and A. Courville, “FILM: Visual reasoning with a general conditioning layer,” in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 3942–3951.

[33] R. Nikhil *et al.*, “A new approach to cross-modal multimedia retrieval,” in *Proc. ACMMM.*, 2010, pp. 251–260.

[34] R. Viresh, R. Nikhil, and C. V. Jawahar, “Multi-label cross-modal retrieval,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4094–4102.

[35] Z. Ma, Y. Lu, and F. Dean, “Finding linear structure in large datasets with scalable canonical correlation analysis,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 169–178.

[36] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 212–228.

[37] H. Pirsiavash, D. Ramanan, and C. Fowlkes, “Bilinear classifiers for visual recognition,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1482–1490.

[38] Z. Yu, J. Yu, J. Fan, and D. Tao, “Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5947–5959, Dec. 2018.

[39] C. Wu, G. Li, N. Duan, D. Tang, and X. Wang, “Deep reason: A strong baseline for real-world visual reasoning,” in 2019, *arXiv:1905.10226*.

[40] S. Bai, J. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018, *arXiv:1803.01271*.

[41] Y. Fisher and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *Proc. Int. Conf. Learn. Representations*, 2014.

[42] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[43] K. Vijaya, C. Gustavo, and R. Ian, “Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5385–5394.

[44] S. Antol, A. Agrawal, J. Lu, and M. Mitchell, “VQA: Visual question answering,” in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 2425–2433.

[45] T. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[46] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015.

[48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[49] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6325–6334.

[50] Z. Yu, J. Yu, J. Fan, and D. Tao, “Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,” in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 1839–1848.

[51] H. Fan and J. Zhou, “Stacked latent attention for multimodal reasoning,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1072–1080.

[52] Q. Cao, X. Liang, B. Li, G. Li, and L. Lin, “Visual question reasoning on general dependency tree,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7249–7257.

[53] O. Ahmed and W. Samek, “DRAU: Dual recurrent attention units for visual question answering,” *Comput. Vision Image Understanding*, vol. 185, pp. 24–30, 2019.

[54] J. Johnson, B. Hariharan, L. Maaten, and F.-F. Li, “Inferring and executing programs for visual reasoning,” in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 3008–3017.

[55] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, “Learning to reason: End-to-end module networks for visual question answering,” in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 804–813.

[56] D. Husdon and C. Manning, “Compositional attention networks for machine reasoning,” in *Proc. Int. Conf. Learn. Representations*, 2018.

[57] C. Lluis, A. Yusuf, V. Carl, P. Hamed, and T. Antonio, “Learning aligned cross-modal representations from weakly aligned data,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2940–2949.

[58] T. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[59] A. Karpathy and F.-F. Li, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3128–3237.

[60] F. Faghri, D. Fleet, R. Kiros, and S. Fidler, “VSE++: Improved visual-semantic embeddings with hard negatives,” in *Proc. Brit. Mach. Vision Conf.*, 2018.

[61] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, “Order-embeddings of images and language,” in *Proc. Int. Conf. Learn. Representations*, 2016.

[62] L. Ma, Z. Lu, L. Shang, and H. Li, “Multimodal convolutional neural networks for matching image and sentence,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2623–2631.

[63] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, “Hierarchical multimodal lstm for dense visual-visual embedding,” in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 1899–1907.

[64] L. Wang, Y. Li, and S. Lazebnik, “Learning deep structure-preserving image-text embeddings,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5005–5013.

[65] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7254–7262.

[66] L. Wang, Y. Li, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.

[67] B. Klein, G. Lev, G. Sadeh, and L. Wolf, “Associating neural word embeddings with deep image representations using fisher vectors,” in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4437–4446.

[68] X. Lin and D. Parikh, “Leveraging visual question answering for image-caption ranking,” in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 261–277.

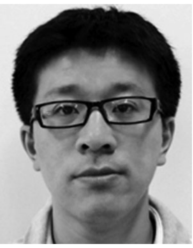
[69] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 707–723.

[70] Y. Huang and L. Wang, “ACMM: Aligned cross-modal memory for few-shot image and sentence matching,” in *Proc. Int. Conf. Comput. Vision*, 2019, pp. 5774–5783.

Jing Yu received the B.S. degree in automation science from Minzu University, Beijing, China, in 2011, the M.S. degree in pattern recognition from Beihang University, Beijing, China, in 2014, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2019. She is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. She works on vision and language problems, including visual question answering, visual dialogue, and cross-modal information retrieval, etc.

Weifeng Zhang received the B.S. degree in electronic information engineering from the Beijing University of Technology, Beijing, China, in 2009, the M.S. degree in pattern recognition from Beihang University, Beijing, China, in 2012, and the Ph.D. degree in computer science from Hangzhou Dianzi University, Hangzhou, China, in 2019. He is currently an Associate Professor with Jiaxing University, Jiaxing, China. His research interests include machine learning and multimedia modeling.

Yuhang Lu received the B.S. degree in automation science from Minzu University, Beijing, China, in 2016, and the M.S. degree from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2018. He is an Algorithm Engineer with Alibaba. His research interests include natural language processing and information retrieval.



Zengchang Qin received the B.S. degree from Heilongjiang University, Harbin, China, in 2001, and the M.S. and Ph.D. degrees from the University of Bristol, Bristol, U.K., in 2002 and 2005, respectively. He is currently an Associate Professor with Intelligent Computing and Machine Learning Lab, the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. His research interests mainly include machine learning, data mining, multimedia learning, and natural language processing.

Yue Hu received the Ph.D. degree in computer science from the University of Science and Technology, Beijing, China, in 2000. She is currently a Researcher with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. She is a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include the area of natural language processing and social network analysis.

Jianlong Tan received the B.S. and M.S. degrees from Xiangtan University, Xiangtan, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include the area of multimedia analysis and hardware algorithm design.

Qi Wu received the M.Sc. and Ph.D. degrees in computer science from the University of Bath, Bath, U.K., in 2011 and 2015, respectively. He is an Assistant Professor with The University of Adelaide, Adelaide, SA, Australia and also an Associate Investigator with the Australia Centre for Robotic Vision (ACRV), Brisbane City, QLD, Australia. He received ARC Discovery Early Career Researcher Award (DECRA) Fellow between 2019 and 2021. He works on the vision and language problems, including image captioning, visual question answering, visual dialog, etc. His work has been published in prestigious journals and conferences, such as TPAMI, CVPR, ICCV, AAAI, and ECCV.