Multi-modality Latent Interaction Network for Visual Question Answering

Peng Gao[1], Haoxuan You[3], Zhanpeng Zhang[2],
Xiaogang Wang[1], Hongsheng Li [1]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research [3]Tsinghua University
{*1155102382@link, xgwang@ee, hsli@ee*}*.cuhk.edu.hk*

Abstract

Exploiting relationships between visual regions and question words have achieved great success in learning multi-modality features for Visual Question Answering (VQA). However, we argue that existing methods [29] mostly model relations between individual visual regions and words, which are not enough to correctly answer the question. From humans' perspective, answering a visual question requires understanding the summarizations of visual and language information. In this paper, we proposed the Multi-modality Latent Interaction module (MLI) to tackle this problem. The proposed module learns the cross-modality relationships between latent visual and language summarizations, which summarize visual regions and question into a small number of latent representations to avoid modeling uninformative individual region-word relations. The cross-modality information between the latent summarizations are propagated to fuse valuable information from both modalities and are used to update the visual and word features. Such MLI modules can be stacked for several stages to model complex and latent relations between the two modalities and achieves highly competitive performance on public VQA benchmarks, VQA v2.0 [12] and TDIUC [20]. In addition, we show that the performance of our methods could be significantly improved by combining with pre-trained language model BERT[6].

1. Introduction

Visual Question Answering [2, 53, 12] has received increasing attention from the research community. Previous approaches solve the Visual Question Answering (VQA) by designing better features [25, 44, 13, 17, 1], better bilinear fusion approaches [10, 7, 22, 3, 52] or better attention mechanisms [48, 29, 49, 45, 36]. Recently, relational reasoning has been explored for solving VQA and significantly improved performance and interpretability of VQA systems.

Despite relationships has been extensively adopted in different tasks, such as object detection [14], language modelling [6], image captioning [51] and VQA [36, 11]. Relational approaches for VQA were only proposed for modelling relationship between words and visual regions. Thus, relational reasoning requires large GPU memories because it needs to model relations between every pair. For VQA, modeling relationships between individual words and visual regions is not enough to correctly answer the question.

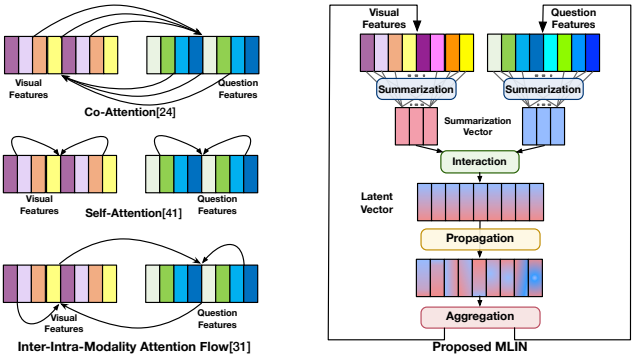

Figure 1: Illustration of the information flow in our proposed MLI compared with previous approaches, namely, co-attention [29], self-attention [45] and intra-inter modality attention(DFAF) [36]. Left side of each image represent visual feature while right side stands for question features.

To model more complex cross-modality relations, we propose a novel Multi-modality Latent Interaction Network (MLIN) with MLI modules. Different from existing relational VQA methods, the MLI module first encodes question and image features into a small number of latent visual and question summarizaiton vectors. Each summarization vector can be formulated as the weighted pooling over visual or word features, which summarizes certain aspect of each modality from a global perspective and therefore encodes richer information compared with individual word and region features. After acquiring summarizations for

each modality, we establish visual-language associations between the multi-modal summarization vectors and propose to propagate information between summarization vectors to model the complex relations between language and vision. Each original visual region and word feature would finally aggregate information from the updated latent summarizations using attention mechanisms and residual connections to predict the correct answers.

Our proposed MLIN achieves competitive performance on VQA benchmarks, including VQA v2.0 [12] and TDIUC [20]. In addition, we experiment how to combine pre-trained language model BERT [6] to improve VQA models. After integrating with BERT [6], MLIN achieves better performance compared with state-of-the-art models.

Our proposed MLIN is related to the attention-based approaches. An illustration between previous approaches can be seen from Figure 1. Previous attention approaches that aggregate information can be classified into the following categories: (1) The co-attention mechanism [29] aggregates information from the other modality. (2) Transformer [45] aggregates information inside each modality using key-query attention mechanism. (3) The intra- & inter-modal attention(DFAF) [36] propagate and aggregate information within and across multiple modalities. For intra-modality feature aggregation, attention is dynamically modulated by the other modality using the pooled features. Compared with previous approaches, MLIN does not aggregate features just from the large number of individual visual-word pairs but from the small number of multi-modal latent summarization vectors, which can capture high-level visual-language interactions with much smaller modal capacity.

Our contributions can be summarized into two-fold. (1) We propose the MLIN for modelling multi-modality interactions via a small number of multi-modal summarizations, which helps encode the relationships across modalities from global perspectives and avoids capturing too much uninformative region-word relations. (2) We carried out extensive ablation studies over each components of MLIN and achieve competitive performance on VQA v2.0 [12] and TDIUC [20] benchmarks. Besides, we provide visualisation of our LMIN and have a better understanding about the interactions between multi-modal summarizations. We also explore how to effectively integrate the pre-trained language model [6] into the proposed framework for further improving the VQA accuracy.

2. Related Work

2.1. Representation Learning

Learning good representations have been the foundations for advancing vision and Natural Language Processing (NLP) research. For computer vision, AlexNet [25], VGGNet [44], ResNet [13] and DenseNet [17] features achieved great success on image recognition [5]. For NLP, word2vec [30], GloVe [37], Skipthough [24], ELMo [39], GPT [40], VilBERT [28] and BERT [6] achieved great success at language modelling. The successful representation learning in vision and language has much benefitted multi-modality feature learning. Furthermore, bottom-up & top-down features [1] for VQA and image captioning greatly boosted the performance of multi-modality learning based on the additional visual region (object detection [41]) information.

2.2. Relational Reasoning

Our work is mostly related to the relational reasoning approaches. Relational reasoning approaches try to solve VQA by learning the relationships between individual visual regions and words. Co-attention based [29] approaches can be seen as modelling the relationship between each word and visual region pairs using the attention mechanism. Transformer [45] proposed to use the key-query-value attention mechanism to model the relationship inside each modality. Simple relational networks [42, 15] reason over all region pairs in the image by concatenating region features. Besides VQA, relational reasoning has improved performance in other research areas. Relational reasoning has been applied to object detection [14] and show that modelling relationships could help object classification and non-maximum suppression. Relational reasoning has also been explored in image captioning [51] using graph neural networks. Non-local network [46] shows that modelling relationship across video frames can significantly boost video classification accuracy.

2.3. Attention-based Approaches for VQA

Attention-based approaches have been extensively studied for VQA. Many relational reasoning approaches using attention mechanisms to aggregate contextual information. Soft and hard attention [48] has been first proposed by Xu *et al*., which has become the main-stream in VQA systems. Yang *et al*. [49] proposed to stack several layers of attention to gradually focus on the most important regions. Lu *et al*. [29] proposed co-attention-based methods, which can aggregate information from the other modality. Vaswani *et al*. [45] aggregated information inside each modality for solving machine translation. Nguyen *et al*. [31] proposed a densely connected co-attention mechanism for VQA. Bilinear Attention Network [21] generated attention weights by capturing the interactions between each feature channel. Structured attention [55] added a Markov Random Field (MRF) model over the spatial attention map for modelling spatial importance. Besides VQA, Chen *et al*. [4] proposed spatial-wise and channel-wise attention mechanisms, which can modulate information flow spatial-wise and channel-wise for image captioning. In referring expression, Xihui

et al. [27] propose attention guided feature erasing.

2.4. Dynamic Parameter Prediction

Dynamic parameter prediction (DPP) propose another direction for multi-modality feature fusion. Noh *et al*. [33] firstly proposed a DPP-based multi-modality fusion approach by predicting the weights of fully connected layer using question features. Perez *et al*. [38] achieved competitive VQA performance compared with complex reasoning approaches on the CLEVR [19] dataset by predicting the normalisation parameter of visual features. Furthermore, Gao *et al*. [9] proposed to modulate visual features by predicting convolution kernels from the input question. Hybrid convolution was proposed to reduce the number of parameters without hindering the overall performance. Beyond VQA, DPP-based approaches have been adopted for transfer learning between classification and segmentation [16].

3. Multi-modality Latent Interaction Network

Figure 2 illustrate the overall pipeline of our proposed Multi-modality Latent Interaction Network (MLIN). The proposed MLIN consists of a series of stacking Multi-modality Latent (MLI) modules, which aims to summarize input visual-region and question-word information into a small number of latent summarization vectors for each modality. The key idea is to propagate visual and language information among the latent summarization vectors to model the complex cross-modality interactions from global perspectives. After information propagation among the latent interaction summarization vectors, visual-region and word features would aggregate information from the cross-domain summarizations to update their features. The inputs and outputs of the MLI module has the same dimensions and the overall network stacks the MLI module for multiple stages to gradually refine the visual and language features. In the last stage, we conduct elementwise multiplication between the average features of visual regions and question words to predict the final answer.

3.1. Question and Visual Feature Encoding

Given an input image I and a question Q, the task of VQA requires joint reasoning over the multi-modal information to estimate an answer. Following previous approaches [1, 21, 36], we extract visual-region features from I using the Faster RCNN object detector [41, 18] and the word features from Q using a bidirectional Transformer model [45]. The feature extraction stage is shown in the upper part of Figure 2. Each image will be encoded as a series of M visual-region features, denoted as $R \in \mathbb{R}^{M\times 512}$, while sentence will be padded to a maximum length of 14 and be encoded by bidirectional Transformer with random initialization, denoted as $E \in \mathbb{R}^{N\times 512}$. The multi-modal feature encoding can be formulated as

$$\begin{aligned} R &= \text{RCNN}(I;\theta_{\text{RCNN}}), \\ E &= \text{Transformer}(Q;\theta_{\text{Transformer}}), \end{aligned} \tag{1}$$

where θ_{RCNN} and $\theta_{\text{Transformer}}$ denote the network parameters for visual and language feature encoding.

3.2. Modality Summarizations in MLI Module

Summarization module can be seen from the Summarization part of Figure 2. After acquiring visual and question features, we add a lightweight neural network to generate k sets of latent visual or language summarization vectors for each modality. The k sets of linear combination weights are first generated via

$$L_R = \text{softmax}_{\leftrightarrow}(W_R R^T + b_R), \tag{2}$$

$$L_E = \text{softmax}_{\leftrightarrow}(W_E E^T + b_E), \tag{3}$$

where $W_R, W_E \in \mathbb{R}^{k\times 512}$ and $b_R, b_E \in \mathbb{R}^k$ are the k sets of learnable linear transformation weights for each of the modality, and "softmax$_{\leftrightarrow}$" denotes the softmax operation along the horizontal dimension. The individual visual and word features, R and E, can then be converted into k latent summarization vectors, $\overline{R} \in \mathbb{R}^{k\times 512}$ and $\overline{E} \in \mathbb{R}^{k\times 512}$, for the visual and language modalities,

$$\overline{R} = L_R \cdot R, \tag{4}$$

$$\overline{E} = L_E \cdot E. \tag{5}$$

Each of the k latent visual or language summarization vectors (*i.e*., each row of $\overline{R}$ or $\overline{E}$) is a linear combination of the input individual features, which is able to better capture high-level information compared with individual region-level or word-level features. The k summarization vectors in each modality can capture k different aspects of the input features from global perspectives.

3.3. Relational Learning on Multi-modality Latent Summarizations

Relational Latent Summarizations. Relational latent summarization is in correspondence with the Interaction part of Figure 2 .The obtained latent summarization vectors encode high-level information from one of the modalties. To reason the correct answer corresponding to the input image and question, it is important to understand the complex cross-domain relations between the inputs. We therefore propose to utilize a relation learning network to establish the associations across domains. Motivated by the simple relation network [42], we create $k\times k$ latent visual-question feature pairs from the above introduced k latent summarization vectors, $\overline{R}$ and $\overline{E}$, in the two modalities. Such $k \times k$ pairs can be represented as a 3D relation tensor $A \in \mathbb{R}^{k\times k\times 512}$:

$$A(i,j,:) = W_A[\overline{R}(i,:) \odot \overline{E}(j,:)] + b_A \tag{6}$$

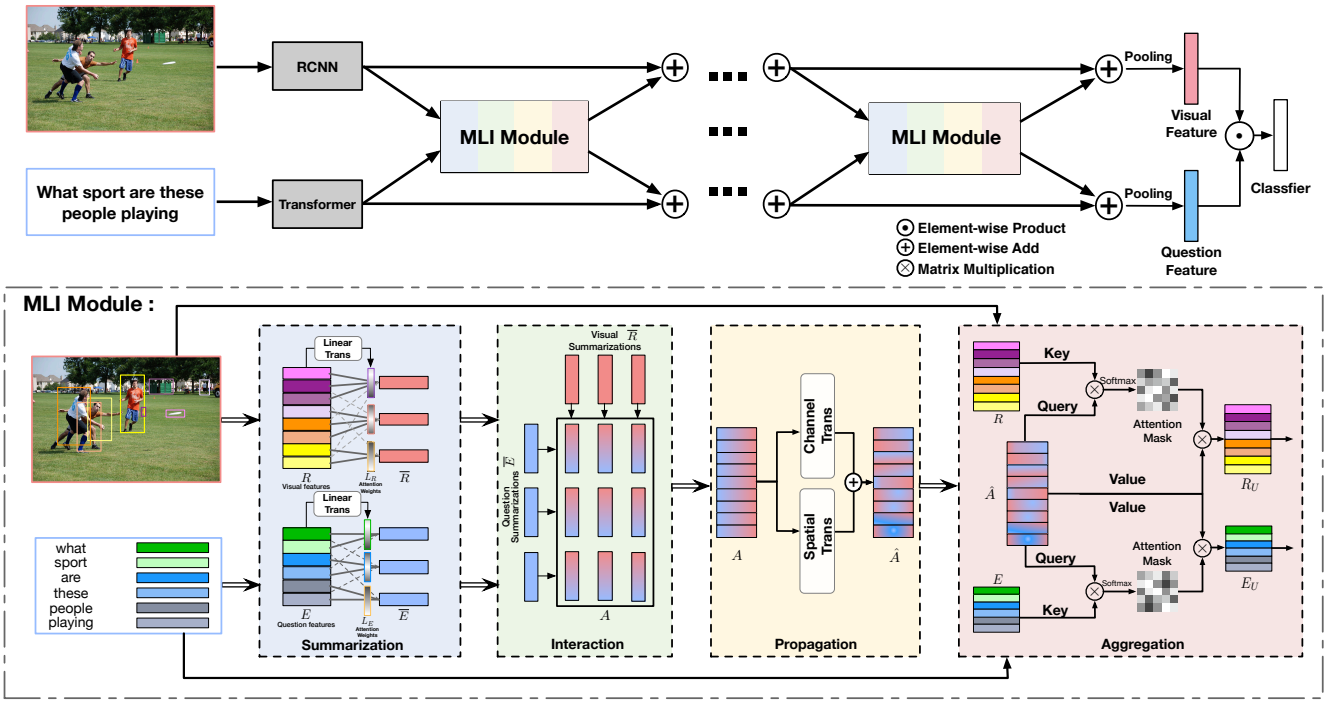

Figure 2: An overview of our proposed stack Multi-modality Latent Interaction Network. Multi-modality reasoning is accomplished inside our proposed MLI modules. After MLI module, residual connection is used for stacking multiple MLI modules. Inside MLI, visual and question features will be summarised into a few summarization vectors, which are fused to create question and visual summarization pairs. After acquiring latent interaction features, we propagate information between latent summarization pairs. After feature propagation, each question and visual feature will gather information from latent summarization vectors using key-query attention mechanism.

where "$\odot$" denotes elementwise multiplication, $W_A \in \mathbb{R}^{512\times512}, b_A \in \mathbb{R}^{512}$ are the linear transformation parameters that further transforms the cross-domain features.

Relational Modeling and Propagation. It is important to propagate information across the two modalities to learn complex relations for answer prediction. Based on our cross-modality relation tensor A, we introduce two operations that passes and aggregate information between the paired features. Before information propagation, the tensor $A \in \mathbb{R}^{k\times k\times512}$ is reshaped to $\tilde{A} \in \mathbb{R}^{k^2\times512}$. The first cross-modal message passing operation performs an additional linear transformation on each paired feature,

$$\tilde{A}_c = \tilde{A} \cdot W_c + b_c \tag{7}$$

where $W_c \in \mathbb{R}^{512\times512}$ and $b_c \in \mathbb{R}^{512}$ are the relation linear transformation parameters that transforms each paired feature $A(i,j,:)$ into a new 512-dimensional feature. The second cross-modal information propagation operation performs information passing between different paired features. The $k \times k = 36$ paired cross-modal features pass messages to each other, which can be considered as "second-order" information for learning even higher non-linear cross-modal relations,

$$\tilde{A}_p = W_p \cdot \tilde{A} + b_p \tag{8}$$

where $W_p \in \mathbb{R}^{36\times36}$ and $b_p \in \mathbb{R}^{36}$ are the linear transformation parameters that propagates information across paired features. The results of the two cross-modal transformations focus on different aspects of the cross-modal paired features to model the complex relations between the input image and question. The first operation focuses on modeling the relation between each individual visual-question latent pair, while the second operation tries to propagate higher-order information between all visual-question pairs to model more complex relations. The summation of the results of the two above operations $\hat{A} \in \mathbb{R}^{k^2\times512}$,

$$\hat{A} = \tilde{A}_c + \tilde{A}_p \tag{9}$$

can be considered as a latent representation that deeply encodes the cross-domain relations between the latent summarization vectors in the two modalities.

Feature Aggregation. The latent multi-modality representation $\hat{A} \in \mathbb{R}^{k^2\times512}$ contains fused question and region

features. Each original visual feature $R(i,:)$ and word feature $E(i,:)$ can aggregate information from the latent representations $\hat{A}$ for improving their feature discriminativeness, which has paramount impact on final VQA accuracy. The feature aggregation process can be modeled by the key-query attention mechanism from Transformer [45]. Each of the region and word features, *i.e.*, $R, E \in \mathbb{R}^{M \text{or} N \times 512}$, would be converted to 128-d query features, $Q_R, Q_E \in \mathbb{R}^{M \text{or} N \times 128}$, as

$$Q_R = R \cdot W_{qr} + b_{qr}, \quad E_Q = E \cdot W_{qe} + b_{qe} \tag{10}$$

where $W_{qr}, W_{qe} \in \mathbb{R}^{512\times128}$, $b_{qr}, b_{qe} \in \mathbb{R}^{512\times128}$ are the linear transformation parameters for calculating the query features. Each feature of the latent representations, *i.e.*, $\hat{A} \in \mathbb{R}^{k^2\times512}$, would be converted to 128-d key and value features $K, V \in \mathbb{R}^{k^2\times128}$,

$$K = \hat{A} \cdot W_k + b_k, \quad V = \hat{A} \cdot W_v + b_v, \tag{11}$$

where $W_k, W_v \in \mathbb{R}^{512\times128}, b_k, b_v \in \mathbb{R}^{128}$ are the linear transformation parameters that calculate the key and value features from latent representations $\hat{A}$. The query features of the region and word features, Q_R, Q_E, would be used to weight different entries from latent representations with their key features K,

$$U_R = \text{softmax}_{\updownarrow}\left(\frac{Q_R \cdot K^T}{\sqrt{\text{dim.}}}\right), \tag{12}$$

$$U_E = \text{softmax}_{\updownarrow}\left(\frac{Q_E \cdot K^T}{\sqrt{\text{dim.}}}\right), \tag{13}$$

where $\text{softmax}_{\updownarrow}$ denotes conducting softmax operation along the vertical dimension and "dim." $= 128$ is a normalization constant. $U_R, U_E \in \mathbb{R}^{M \text{or} N \times k^2}$ stores each region or word feature's weights to aggregate the k^2 latent representations. The original region and word features can therefore be updated as

$$R_U = R + U_R \cdot \hat{A} \tag{14}$$

$$E_U = E + U_E \cdot \hat{A} \tag{15}$$

where $U_R \cdot \hat{A}$ and $U_E \cdot \hat{A}$ aggregate the informamtion from the latent representations to obtain the updated region and word features R_U and E_U. The feature aggregation process has been illustrated in the Aggregation module in Figure 2.

The input features R, E and output features R_U, E_U of the above introduced MLI module shares the same dimension. Motivated by previous approaches [21, 36], we stack MLI modules for multiple stages to recursively refine the visual and language features. After several stages of MLI modules, we average pool the visual and word features separately and elementwisely multiplicate the deeply refined region and word features for multi-modal feature fusion. A final linear classifier (W_{cls}, b_{cls} as parameters) with softmax non-linearity function is adopted for answer prediction,

$$R_{\text{pool}} = \frac{1}{M}\sum_{i=1}^{M} R_U(i,:), \tag{16}$$

$$E_{\text{pool}} = \frac{1}{N}\sum_{i=1}^{N} E_U(i,:), \tag{17}$$

$$\text{Answer} = \text{Classifier}\left[R_{\text{pool}} \odot E_{\text{pool}}\right] \tag{18}$$

Accordingly, the overall system is trained in an end-to-end manner with cross-entropy loss function.

3.4. Comparison of Message Passing Complexity

In this section, we compared the message passing complexity between co-attention [29], self-attention [45] and intra-inter attention [36]. The information flow pattern has been illustrated in Figure 1. For co-attention, the number of message passings is $\mathcal{O}(2 \times M \times N)$ because each word would calculate an attention matrix from each visual region and vice versa. For self-attention, the number of message passings is $\mathcal{O}(M \times M + N \times N)$. The number of message passings for intra- and inter-modality attention is the summation of those of self-attention and co-attention, $\mathcal{O}((M+N)\times(M+N))$. Generally, in bottom-up & top-down attention [1], 100 region proposals would be used for multi-modal feature fusion. The quadratic number of message passings in self attention [45] and intra- and inter-modality attention flow [8] would requires large GPU memories and hinders the relational learning as well. For our proposed MLIN framework, the MLI module generates k latent summarization vectors for each modality. After relational reasoning, $k \times k$ features are generated. In the final feature redistribution stage, $\mathcal{O}(k \times k \times N)$ message passings are performed for question feature update, and $\mathcal{O}(k \times k \times M)$ message passings are required for updating region features. The total number of message passings for our proposed MLIN in each stage is therefore $\mathcal{O}(k \times k \times (M+N))$. Our proposed multi-modality latent representations could better capture multi-modality interactions with much fewer message passings and achieved competitive performance compared with DFAF. A performance comparison has been conducted in the experiments session.

4. Experiments

4.1. Dataset

We conduct experiments on VQA v2.0 [2] and TDIUC [20] datasets. Both VQA v2.0 and TDIUC contain question-image pairs collected from Microsoft COCO [26] dataset and annotated questions. VQA v2.0 is an updated version of VQA v1.0 by reducing data bias. VQA v2.0 contains train, validation and test-standards and 25% of test-

standards serve as the test-dev set. Performance evaluation on VQA v2.0 includes evaluating accuracies of different types of questions: YES/NO, Number, Others and overall accuracy. Train, validation and test sets contain 82,743, 40,504 and 81,434 images, with 443,757, 214,354 and 447,793 questions, respectively. We carry out extensive ablation studies on the validation set of VQA v2.0 trained on train split. Also, we report final performance on VQA v2.0 test set trained on the combination of train and validation set, which is a common practice of most previous approaches listed in Table 2. Although VQA v2.0 has been commonly adopted as the most important benchmark on VQA. However, Kafke *et al*. [20] found that the performance of VQA v2.0 is dominated by simple questions, which make it difficult to compare different approaches. To solve the bias problem existing in VQA v2.0, TDIUC collect 1.6 million questions divided into 12 categories.

4.2. Experimental Setup

We use common feature extraction, preprocessing and loss function as most previous approaches listed in Table 2. For visual features, we extract the first 100 region proposals with dimension of 2048 for VQA v2.0. While on TDIUC, we extract the first 36 region features. Region features are generated by Faster RCNN [41]. For the question encoder, we pad all questions with 0 to a maximum length of 14 and extract $\mathbb{R}^{14\times786}$ question features using a single layer Bidirectional Transformer [45] with random initialization. After acquiring visual and word features, we transform them into 512 dimension using linear transform. For all layers, we use a dropout rate 0.1 and clip the gradients to 0.25. Default batch size is 512 with Adamax [23] optimiser with a learning rate of 0.005. We gradually increase the learning rate to 0,005 in the first 1000 iterations because our Bidirectional Transformer Encoder is initialised randomly while previous approaches use pretrained Glove [37] and Skipthought [24] embedding. We also augment our MLIN with a Masked Word Prediction for transformer regularisation. We trained the model for 7 epochs and decay the learning rate 0.0005 and fix it for the following epochs. All layers are initialised randomly with Pytorch's [35] random initialisation. For pretrained language models, we adopt a base BERT [6] model which is trained by randomly masking words.

4.3. Ablation Study on VQA2 Validation

We carried out extensive ablation studies on evaluating the effectiveness of each module in our proposed MLIN in Table 1. The default setting is one stage MLIN where all features are transformed into dimension of 512. We create 6 summarizations for each modality. For the feature aggregation key-query attention module, we adopted a 12 head multi-head attention with each head calculating 128-dimensional features. In ablation study, we check the influence of the number of MLIN stacks, number of latent summarisation vectors, latent interaction, latent propagation, feature aggregation and final feature fusion operator.

Component	Setting	Accuracy
Bottom-up [1]	Bottom-up	63.37
Bilinear Attention [21]	BAN-1	65.36
	BAN-4	65.81
	BAN-12	66.04
DFAF [36]	DFAF-1	66.21
	DFAF-8	66.66
	DFAF-8 + BERT	67.23
Default	MLI-1	66.04
	MLI-8 + BERT	**67.83**
# of stacked blocks	MLI-5	66.32
	MLI-8	**66.53**
# of Question and Visual Summary Heads	3 by 3	65.63
	6 by 6	66.04
	6 by 12	66.15
	12 by 12	**66.21**
Latent Interaction Operator	Concat	65.99
	Product	66.04
	Addition	65.69
	MUTAN	**66.20**
Embedding dimension	512	66.04
	1024	**66.18**
Latent Propagation Operator	Linear	**66.04**
	Self Attention	65.84
	Dual Attention	66.01
Feature Gathering Operator	Key-query	**66.04**
	Transpose	65.78
# of Parallel Heads in Feature Gathering Operator	8 heads	65.84
	12 heads	**66.04**
	16 heads	66.19
BERT Finetuning	Freezing	65.51
	lr 1/10 fintuning	**67.83**
	lr 1/100 finetuning	66.99
	lr 1/1000 finetuning	66.74

Table 1: Ablation studies of our proposed MLIN on VQA v2.0 validation dataset. Default setting is represented by underline while best performance will be highlighted. Our proposed MLIN takes both simplicity and performance into consideration.

Similarly with BAN [21] and DFAF [36], we stack the proposed MLI module for 5 and 8 times denoted as MLIN-5 and MLIN-8 for multiple stage reasoning. We observe that deeper layers will improve the performance and can be optimized by SGD thanks to the residual connections [13].

Then we study the influence of the number of question and visual summarization vectors. Too few summarization vectors will be unable to capture different aspects of the input which deteriorates the overall performance. Too many

summarization vectors will require too much GPU memory and computations with marginal improvement. We choose 6 question summarization and 6 visual summarization vectors as a trade-off between performance and computation.

For the interaction operator to create paired summarization vectors, we compare between element-wise product, element-wise addition and bilinear fusion (MUTAN) [3] for multi-modality summarization fusion. Bilinear fusion [3] gives the best performance. However, we choose elementwise product in our final model considering the overall simplicity and efficiency of the network design. Different from our approaches, Simple Relational Reasoning Network [42] choose concatenation by default.

For the simplicity of hyper-parameter selection, we set all layers have the same dimension. Extracted visual and question features are transformed into the same dimension by linear transform. 1024 leads to better performance than 512. However, stacking multiple MLI modules can lead to more performance improvement than being wide. Our final model chooses 512 dimensions by default.

Among the latent paired summarization vectors, there exist several ways for propagating information between them. Self-attention [45] uses key-query attention to aggregate information from the other latent summarizations. while dual attention aggregate information inside and outside each feature vector simultaneously using self attention. In our experiment, our proposed relational propagation operations (e.g. Equation 7,8,9) could achieve better performance than the complicated dual attention.

After acquiring latent interaction features, the original question and visual features will gather information from the latent vectors to complete multi-modality relational learning. We tested two approaches for feature gathering from latent vectors. We use the key of visual and word feature to gather information from the query of latent vectors and perform weighted pooling of latent summarization vectors. Motivated by the dynamic attention weight prediction network [47], we use the the transpose of attention weight in the summarization stage to gather information from latent summarization vectors. Key-query attention approach outperform dynamic attention weight prediction.

Another hyper-parameter in feature gathering stage is the number of attention heads and head dimension in the feature aggregation stage, we keep the dimension of each heads as 128 and test the number of parallel attention head with number of 8, 12 and 16. The obtained features of different heads are concatenated to obtain the final features.

Language model has been actively investigated in NLP related tasks. Language models [30, 37, 39, 6] can generate feature that better capture language meanings. BERT [6] is a language model pretrained by randomly masking a word or predicting whether one sentence is next to the other sentence. As can be seen from the table, finetuning the MLIN+BERT model by setting its learning rate to 1/10 of the main learning rate will awaken the full power of BERT.

Model	test-dev				test-std
	Y/N	No.	Other	All	All
Feature Fusion					
BUTP [1]	81.82	44.21	56.05	65.32	65.67
MFH [12]	n/a	n/a	n/a	66.12	n/a
MFH+BUTD [12]	84.27	49.56	59.89	68.76	n/a
BAN+Glove [21]	85.46	50.66	**60.50**	69.66	n/a
Relation Learning					
DCN [31]	83.51	46.61	57.26	66.87	66.97
Relation Prior [50]	82.39	45.93	56.46	65.94	66.17
Graph [34]	82.91	47.13	56.22	n/a	66.18
Counter [54]	83.14	51.62	58.97	68.09	68.41
DFAF [54]	86.09	53.32	60.49	70.22	70.34
DFAF-BERT [54]	86.73	52.92	**61.04**	70.59	70.81
MLIN(ours)	85.96	52.93	60.40	70.18	70.28
MLIN-BERT(ours)	**87.07**	**53.39**	60.49	**71.09**	**71.27**

Table 2: Comparison with previous state-of-the-art methods on VQA 2.0 test dataset.

4.4. Comparison with State of the art methods

In this section, we compare our proposed MLIN with previous state-of-the-art methods on VQA v2.0 and TDIUC datasets in Table 2 and 3. Following previous methods, we compare our methods on VQA v2.0 test dataset trained with train, validation split and visual genome augmentation.

On VQA v2.0, we divide previous approaches into non-relational and relational approaches which are two orthogonal research directions and can assist each other. Bottom-Up-Top-Down(BUTD) [1] approach proposed to use object detection features in a simple attention module for answering the question related to the input image. MFH [52] is the state-of-the-art bilinear fusion approach. By switching from Residual features to Bottom-up-top-down features, better accuracy can be achieved. BAN [21] proposed a bilinear attention mechanism which generates a multi-modality attention using information of each channel and has won the first place in the single model task of VQA competition 2018.

Besides feature fusion, relational reasoning has been paid much attention in solving VQA. DCN [31] proposed a densely connected co-attention module for cross-modality feature learning. $< subject, predicate, object >$ triples are created for VQA reasoning in Relation prior [50]. Conditional Graph [34] built a graph among all region proposals and condition this graph on visual question. Although Conditional Graph is less competitive compared with other approaches. However, the interpretation from conditional graph is quite useful for diagnosing VQA problem. Counter [54] dives into the number question of VQA by utilising the relative position between bounding box for learning efficient Non Maximum Suppression(NMS). DFAF [36] is a multi-layer stacked network by combining intra- and inter- modality information flow for feature fu-

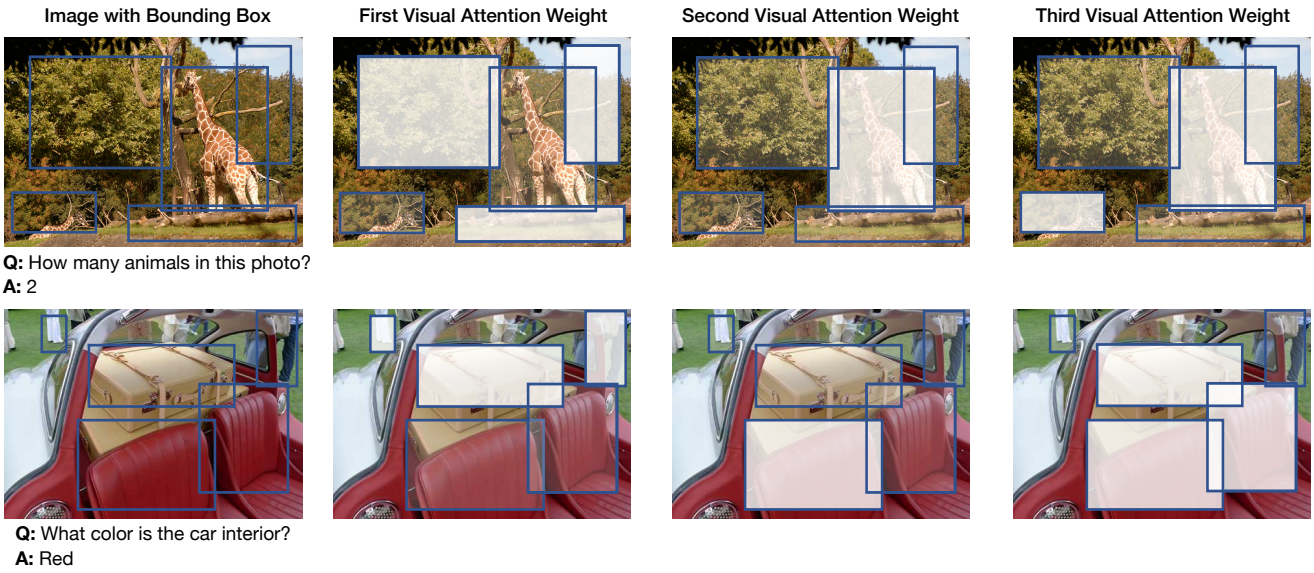

Figure 3: We visualize the first three visual attention weights for creating visual summarization vectors. Bounding boxes generated by Faster RCNN are shown in the first column. For visual summarization, the colors ranging from clear to white in bounding boxes denote the attention weights from 0 to 1. After training, the first attention focuses on the background regions. The second and the third attention weights concentrate on single and multiple foreground objects

Model	RAU [32]	MCB [7]	QTA [43]	DFAF [36]	MLI
Accuracy	84.26	81.86	85.03	85.55	87.60

Table 3: Comparison with previous state-of-the-art methods on TDIUC test dataset.

sion. Furthermore, DFAF can dynamically modulate the intra modality information flow using the average pooled features from the other modality. MLI use 100 region proposals for fair comparison.

VQA 2.0 has been mostly adopted as the most important benchmark in VQA. Since VQA 2.0 is dominated by simple samples, which is hard to discriminate between different methods. We also compare with approaches on the TDIUC dataset. QTA [43] is the state-of-the-art methods on TDIUC, which proposed a question type guided attention with both bottom-up-top-down features and residual features. Our proposed MLIN can achieve better performance even with bottom-up-top-down features only. Our method also outperform DFAF on this dataset.

4.5. Visualization

We visualize the attention weight of summarization vector in Figure 3. We discover the following patterns. Different summarization have a specific function. As can be seen from the visualization of attention weight, different summarization vectors focus on different global information. The first attention weight collect information from the background, while the second attention weight focuses on the most important regions for answering the question. While the third attention performs weighted pooling of regions with a strong interaction for answering the question.

5. Conclusion

In this paper, we proposed a novel MLIN for exploring relationship for solving VQA. Inside MLIN, multi-modality reasoning is realized through the process of Summarisation, Interaction, Propagation and Aggregation. MLIN can be stacked several layers for better relationship reasoning. Our method achieved competitive performance on benchmark VQA dataset with much smaller message passing times. Furthermore, we show a good pre-trained language model question encoder is important for VQA performance.

6. Acknowledgements

This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14208417, CUHK14239816, in part by CUHK Direct Grant.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 2, 3, 5, 6, 7

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 5

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2612–2620, 2017. 1, 7

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2, 6, 7

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1, 8

[8] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. 5

[9] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven CH Hoi, and Xiaogang Wang. Question-guided hybrid convolution for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 469–485, 2018. 3

[10] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016. 1

[11] Shijie Geng, Ji Zhang, Hang Zhang, Ahmed Elgammal, and Dimitris N Metaxas. 2nd place solution to the gqa challenge 2019. *arXiv preprint arXiv:1907.06794*, 2019. 1

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 6

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 1, 2

[15] Ping Hu, Ximeng Sun, Kate Saenko, and Stan Sclaroff. Weakly-supervised compositional featureaggregation for few-shot recognition. *arXiv preprint arXiv:1906.04833*, 2019. 2

[16] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. 3

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 2

[18] Zhengkai Jiang, Peng Gao, Chaoxu Guo, Qian Zhang, Shiming Xiang, and Chunhong Pan. Video object detection with locally-weighted deformable neighbors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8529–8536, Jul. 2019. 3

[19] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017. 3

[20] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 1, 2, 5, 6

[21] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1571–1581, 2018. 2, 3, 5, 6, 7

[22] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 2, 6

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence

Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[27] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019. 3

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 1, 2, 5

[30] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 2, 7

[31] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018. 2, 7

[32] Hyeonwoo Noh and Bohyung Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*, 2016. 8

[33] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 30–38, 2016. 3

[34] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *Advances in Neural Information Processing Systems*, pages 8344–8353, 2018. 7

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[36] Gao Peng, Hongsheng Li, Haoxuan You, Zhengkai Jiang, Pan Lu, Steven Hoi, and Xiaogang Wang. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. *arXiv preprint arXiv:1812.05252*, 2018. 1, 2, 3, 5, 6, 7, 8

[37] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2, 6, 7

[38] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3

[39] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, 2018. 2, 7

[40] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 6

[42] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017. 2, 3, 7

[43] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–166, 2018. 8

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1, 2, 3, 5, 6, 7

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[47] Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019. 7

[48] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1, 2

[49] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 1, 2

[50] Zhuoqian Yang, Jing Yu, Chenghao Yang, Zengchang Qin, and Yue Hu. Multi-modal learning with prior visual relation reasoning. *arXiv preprint arXiv:1812.09681*, 2018. 7

[51] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018. 1, 2

[52] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering.

IEEE transactions on neural networks and learning systems, (99):1–13, 2018. 1, 7

[53] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[54] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018. 7

[55] Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. Structured attentions for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1291–1300, 2017. 2