Hindawi
Scientific Programming
Volume 2022, Article ID 4378553, 11 pages
https://doi.org/10.1155/2022/4378553

*Research Article*

# Multiple Context Learning Networks for Visual Question Answering

**Pufen Zhang, Hong Lan, and Muhammad Asim Khan**

*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China*

Correspondence should be addressed to Hong Lan; lanhong@jxust.edu.cn

Received 2 September 2021; Revised 16 December 2021; Accepted 22 December 2021; Published 9 February 2022

Academic Editor: Roberto Natella

A novel Multiple Context Learning Network (MCLN) is proposed to model multiple contexts for visual question answering (VQA), aiming to learn comprehensive contexts. Three kinds of contexts are discussed and the corresponding three context learning modules are proposed based on a uniform context learning strategy. Specifically, the proposed context learning modules are visual context learning module (VCL), textual context learning module (TCL), and visual-textual context learning module (VTCL). The VCL and TCL, respectively, learn the context of objects in an image and the context of words in a question, allowing object and word features to own intra-modal context information. The VTCL is performed on the concatenated visual-textual features that endows the output features with synergic visual-textual context information. These modules work together to form a multiple context learning layer (MCL) and MCL can be stacked in depth for deep context learning. Furthermore, a contextualized text encoder based on the pretrained BERT is introduced and fine-tuned, which enhances the textual context learning at the feature extraction stage of text. The approach is evaluated by using two benchmark datasets: VQA v2.0 dataset and GQA dataset. The MCLN achieves 71.05% and 71.48% overall accuracy on the test-dev and test-std sets of VQA v2.0, respectively. And an accuracy of 57.0% is gained by the MCLN on the test-standard split of GQA dataset. The MCLN outperforms the previous state-of-the-art models and the extensive ablation studies examine the effectiveness of the proposed method.

## 1. Introduction

An artificial intelligence agent must be able to understand not only the semantics of text but also the content of images. Most multimodal tasks involving image and text modalities require this ability; these tasks include grounding referring expressions [1], image captioning [2], image-text matching [3], and visual question answering (VQA) [4, 5]. Compared to other multimodal tasks, VQA is more complex that requires associating visual content in the image with the semantic meaning in the question, together with visual reasoning to make the correct answer. In addition, VQA has a wide range of applications in practice, such as assisting the blind and early childhood education [6]. Considering the challenges and significance of VQA, visual question answering has attracted much attention in computer vision and natural language processing communities.

In the early stages, the VQA methods adopted the convolutional neural networks (CNN) [7] and recurrent neural networks (RNN) [8] to extract the global image and text features [9, 10], respectively. However, the extracted global features are not fine-grained. With the development of deep learning, some fine-grained VQA approaches proposed employing attention mechanisms to locate the question keywords and the image objects related to the answers [11–13]. For example, question-guided visual attention on image regions was first proposed in [11]. Following that, a large variety of attention-based variations including co-attention and compositional attention have been proposed for VQA [14, 15]. Although aforementioned attention-based approaches have dramatically ameliorated the performance of VQA, the context information on image and question are not considered by these methods.

As context reflects the high-order interactive information between entities (image objects or question words) and helps to distinguish the targets from other entities, several attention-based VQA models have begun to emphasize multimodal context learning. For example, to capture the visual context information, two shallow context learning models, ReGAT [16] and v-AGCN [17], are proposed. They both structured the object relation graph on the image and used the graph attention network [18] to generate the visual context-aware image representations. Furthermore, LGCN learned the deep visual context representation through iterative message passing the conditioned textual input on the fully connected image object graph [19]. However, these models only consider the context for image modality. More recently, some deep models have attempted to learn the extra textual context. DFAF [20] used the intra-modality attention flow to capture the intra-modal contexts within each modality. Two deep co-attention models, namely, MCAN [21] and MEDAN [22], have been introduced to VQA task to capture the visual and textual contexts based on the encoder-decoder framework.

Despite the effectiveness of these methods, we still find that more comprehensive context learning is rarely explored. On one hand, a special context, named visual-textual context and expressed as the synergic context-dependence under two modal information, is almost ignored by the above shallow and deep context learning models. On the other hand, how to explore an accordant context learning way integrating multicontext learning into a unified framework and modelling deep multiple contexts, there is still an open question. Besides, in terms of text representations, most existing methods use RNN architecture to extract the text features. However, due to long-term dependence between words, the RNN cannot sufficiently capture the textual context at the textual features extraction stage. To tackle these issues, we propose a novel Multiple Context Learning Network (MCLN) to learn comprehensive contexts for VQA. In this paper, three contexts containing an ignored visual-textual context and two intra-modal contexts (i.e., the visual context in the image modality and textual context in the text modality) are explored. The key-query attention mechanism [23] is able to model the context-dependence of all entities (objects or words) in an entity set. Inspired by this, we adopt key-query attention to uniformly model the multiple context-dependence and propose the corresponding context learning modules. Figure 1 shows the proposed uniform multiple contexts learning strategy. The proposed visual context learning module (VCL), textual context learning module (TCL), and visual-textual context learning module (VTCL) are constructed based on the key-query attention. Firstly, two intra-modal contexts are learned through the VCL and TCL, respectively. Then, the image and question features with intra-modal context information are concatenated to form the visual-textual features, which will be fed into VTCL to extract the visual-textual context information. After that, by modular composition of the three modules, the multiple contexts learning layer (MCL) is structured. Furthermore, we also tend to capture the deep context information. By cascading MCL in depth, the deeper level and more complex context learning can be reached handily. Additionally, the pretrained Bidirectional Encoder Representations from Transformers (BERT) [24] models the textual context by stacking the bidirectional encoder. To further enhance the textual context learning ability of MCLN at the textual features extraction stage and provide better contextualized representation for textual features, a BERT-based text encoder is introduced and fine-tuned to learn contextualized text representations, resulting in better VQA performance. Finally, the Multiple Context Learning Network that consists of cascaded MCL layers is proposed. The preprint is shown in [25].

The main contributions of this study can be summarized as follows:

(1) We explore multiple contexts including visual context, textual context, and visual-textual context ignored by previous methods for VQA task. Particularly, we propose corresponding context learning module based on a uniform context learning strategy and compose the context learning modules to structure a multiple context learning layer (MCL), which can model and learn multiple contexts.

(2) We attempt to consider the deep contexts and the proposed MCLN is a deep model stacking several MCL layers, meaning that it can capture deep-level context information.

(3) To further consider more comprehensive context learning, we adopt the pretrained BERT as a contextualized text encoder and fine-tune its learning rate during the training of MCLN, enhancing the textual context at the textual features extraction stage.

(4) Evaluation results on two benchmark VQA datasets demonstrate that MCLN achieves state-of-the-art performance. The proposed MCLN achieves 71.05% and 71.48% overall accuracies for test-dev and test-std sets on the VQA v2.0 dataset [4], respectively. On the GQA [5], the overall performance reaches 56.8% for test-dev set and 57.0% for test set. In addition, adequate ablation studies are conducted to quantitatively and qualitatively prove the significance of different modules in the proposed model, verifying the effectiveness of the proposed MCLN architecture.

The rest of the paper is organized as follows. In Section 2, the proposed Multiple Context Learning Networks (MCLN) are presented. In Section 3, the experimental results and analysis are provided. Finally, the paper is concluded in Section 4.

## 2. Multiple Context Learning Networks for VQA

As common practice, identifying a correct answer related to the given image-question pair from a set of candidate answers is a typical formulation of VQA:

$$\widehat{a} = \arg\max_{a \in A} p_\theta (a | I, Q) \tag{1}$$

where the model predicts the correct answer $\widehat{a}$ from the candidate answers $A$ for a given image $I$ and question $Q$ pair,

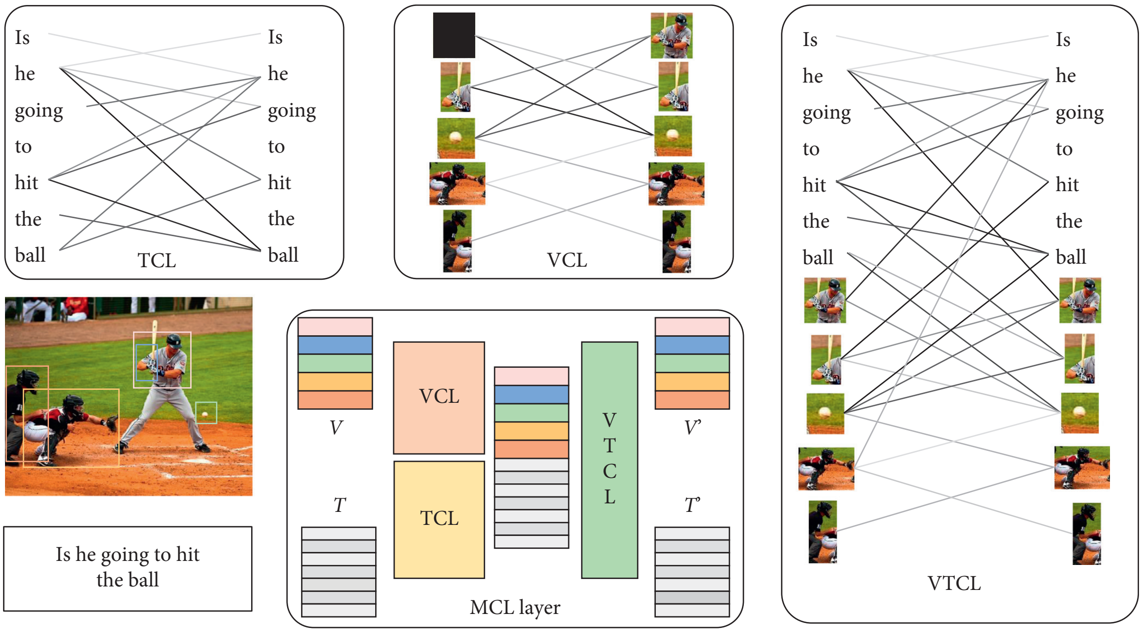

Figure 1: A diagram of the proposed context learning method, which simultaneously learns multiple contexts by using a uniform context learning framework. VCL and TCL model the intra-modal contexts in the MCL layer, while VTCL models the visual-textual context. For image features (V) and question feature (T), the handled features with context information are (V)′ and (T)′, respectively.

and $\theta$ is the learned parameters of model. Without loss of generality, the proposed method also follows this convention.

Figure 2 illustrates the overall pipeline of the proposed MCLN, which consists of a series of subnetworks: (a) image and question representation, which includes image encoder and text encoder. The image encoder detects objects in an image and extracts the object features, while text encoder embeds the question words and encodes the word features; (b) multiple contexts learning that performs multicontext learning and it is composed of multiple learning layers by the cascaded way; and (c) multimodal fusion and answer prediction that fuses image representation and question representation to predict a correct answer.

### 2.1. Image and Question Representation

*2.1.1. Image Encoder.* Inspired by humans using bottom-up attention to process visual information [13], we extract the object-level features for a given image. Our image encoder encapsulates three main steps: (i) a forward pass of a Faster R-CNN [26] object detector pretrained on the Visual Genome dataset [27] for extracting the $K$ most salient objects $O = \{o_i\}_{i=1}^{K}$ within a given image $I$, thereby achieving the bottom-up attention; (ii) a ResNet-101 network [28] that extracts the feature map of detected objects, followed by a mean pooling layer that generates the 2048-dimensional vector representation $r_i$ for the original object $o_i$; and (iii) a linear transformation layer to project the object feature $r_i$ into a 768-dimensional vector.

$$\begin{aligned} r_i &= \text{Faster} - \text{RCNN}(I), \quad i \in [1, K], \\ v_i &= W_v r_i + b_v. \end{aligned} \tag{2}$$

$W_v \in R^{2048\times 768}$ and $b_v \in R^{768}$ are the parameter of the linear layer. Finally, image $I$ is represented as visual object features set $V = \{v_i \in R^{768}\}_{i=1}^{K}$.

*2.1.2. Text Encoder.* RNN-based text encoder first tokenizes question $Q$ into words and uses a maximum of $L$ words to trim the question. Then, the $l$-th word in the question is further embedded into a 300-dimensional vector representation $e_l \in R^{300}$ by using the GloVe word embeddings [29]. The questions shorter than $L$ words are padded at the end with zero vectors. Thus, $Q$ is initialized as a word embeddings sequence $E = \{e_l \in R^{300}\}_{l=1}^{L}$. Here, a one-layer and 768-dimensional long short-term memory network (LSTM) [10] is utilized, which scans the word embeddings sequence $E$ from $e_1$ to $e_L$ and picks up current $e_l$ into its unit to obtain the word feature $t_l$.

$$(t_1, \ldots, t_l, \ldots t_L) = \text{LSTM}(e_1, \ldots, e_l, \ldots e_L). \tag{3}$$

To enhance the textual context at the textual features extraction stage, we also introduce a contextual text encoder based on the pretrained BERT model [23] and fine-tune it to extract the contextualized word features. The BERT-based text encoder is represented as

$$(t_1, \ldots, t_l, \ldots t_L) = \text{BERT}(Q). \tag{4}$$

Finally, the question $Q$ is represented as word features set $T = \{t_l \in R^{768}\}_{l=1}^{L}$.

*2.2. Multiple Context Learning Layers.* The key-query attention mechanism [23] can model context-dependence of all entities (objects or words) across whole entities set. Inspired by this, we adopt key-query attention to uniformly

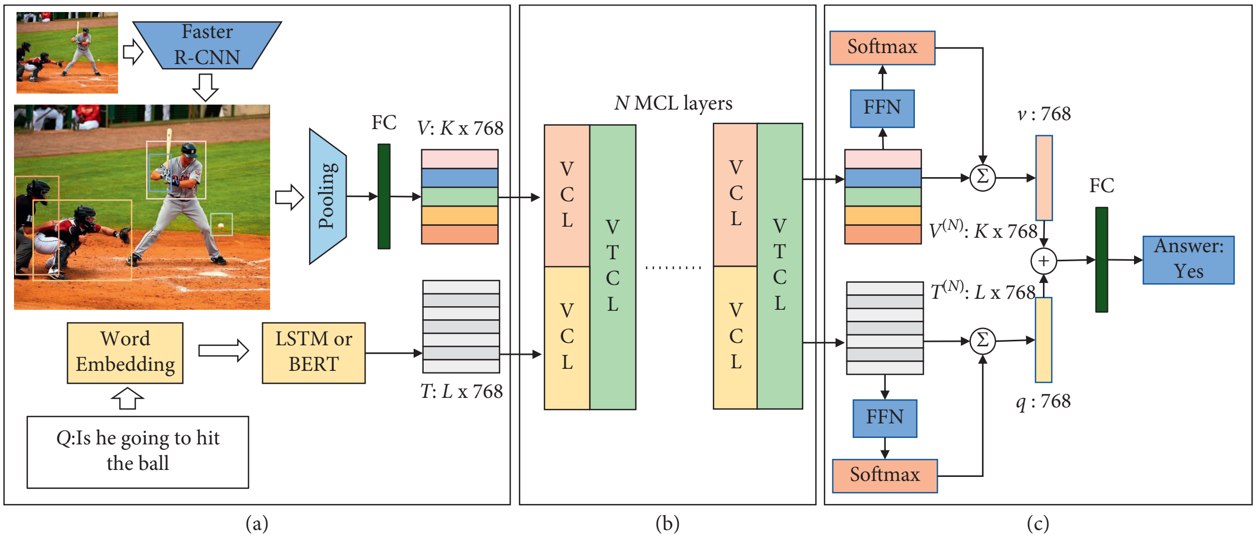

Figure 2: Overall flowchart of the MCLN that consists of three subnetworks. (a) Image and question representation; (b) multiple context learning; (c) multimodal fusion and answer prediction.

model the multiple context-dependence and propose the corresponding context learning modules. Using visual context learning as an example, Figure 3 shows the design of the visual context learning module (VCL), and the implementation will be discussed in detail below.

To capture the visual context of all objects, we adopted the key-value attention to model the context-dependence of objects. Obtained image object features set $V$ can be denoted as a features matrix $V \in R^{K\times 768}$ and its $i$-th row represents the feature of $i$-th object. The matrix $V$ would be converted to key matrix $K^V$, query matrix $Q^V$, and value matrix $V^V$ by three linear transformations.

$$\begin{aligned} K^V &= VW^K, \\ Q^V &= VW^Q, \\ V^V &= VW^V, \end{aligned} \tag{5}$$

where $W^K$, $W^Q$, and $W^V$ are the linear transformation parameters that calculate the key, query, and value matrices. Then, calculate the dot products between query and key matrices to get the attention weights matrix $A^V$.

$$A^V = \mathrm{softmax}\left(\frac{Q^V\left(K^V\right)^{\mathrm{T}}}{\sqrt{d}}\right), \tag{6}$$

where $\sqrt{d}$ is a normalization constant. In this way of dot product, the contextual path dependency distance between all objects is set to 1. So $A^V \in R^{K\times K}$ reflects the context-dependence of all objects and stores the attention weights between all objects. The row attention weights $A^V_{i:}$ represent the context-dependence of $i$-th object in the image. Therefore, the visual context feature for $i$-th object can be obtained by calculating a weighted combination of the $A^V_{i:}$ and all object features, whereas the form of matrices for the visual context feature of all objects can be stated as follows:

$$\mathrm{ATT}(V) = \mathrm{softmax}\left(\frac{Q^V\left(K^V\right)^{\mathrm{T}}}{\sqrt{d}}\right)V^V. \tag{7}$$

In addition, the multihead attention based on key-value attention is performed on $V$.

$$\begin{aligned} \mathrm{MultiHead}(V) &= \left(\mathrm{head}_1 \| \cdots \| \mathrm{head}_H\right)W^O, \\ \mathrm{head}_h &= \mathrm{ATT}_h(V) = \mathrm{softmax} \\ &\cdot \left(\frac{\left(VW_h^V\right)\left(VW_h^K\right)^T}{\sqrt{d_h}}\right)\left(VW_h^V\right). \end{aligned} \tag{8}$$

$\mathrm{head}_h$ is $h$-th head key-query attention. $W_h^K \in R^{768\times d_h}$, $W_h^Q \in R^{768\times d_h}$, and $W_h^V \in R^{768\times d_h}$ are the parameter matrices of $\mathit{head}_h$. $W^O$ is the projection matrix for all heads and $\|$ represents concatenation of all heads. $d_h$ is the dimensionality of the output features from each head and usually making $d_h = 768/H$. The multihead attention allows the model to jointly attend to context information from different representation subspaces, improving the representation capacity of features.

After acquiring the visual context features of all objects using multihead attention, residual connection followed by layer normalization is applied to integrate the visual context into the object features.

$$\widehat{V} = \mathrm{LN}(V + \mathrm{MultiHead}(V)), \tag{9}$$

where LN $(\cdot)$ represents layer normalization. Aiming to further adjust the object representations, the position wise feed-forward network (FFN) transforms object features $\widehat{V}$ with two fully connected layers. And it can be described as

$$\mathrm{FFN}(\widehat{V}) = \left(\mathrm{ReLu}\left(\widehat{V}W_1 + b_1\right)\right)W_2 + b_2, \tag{10}$$

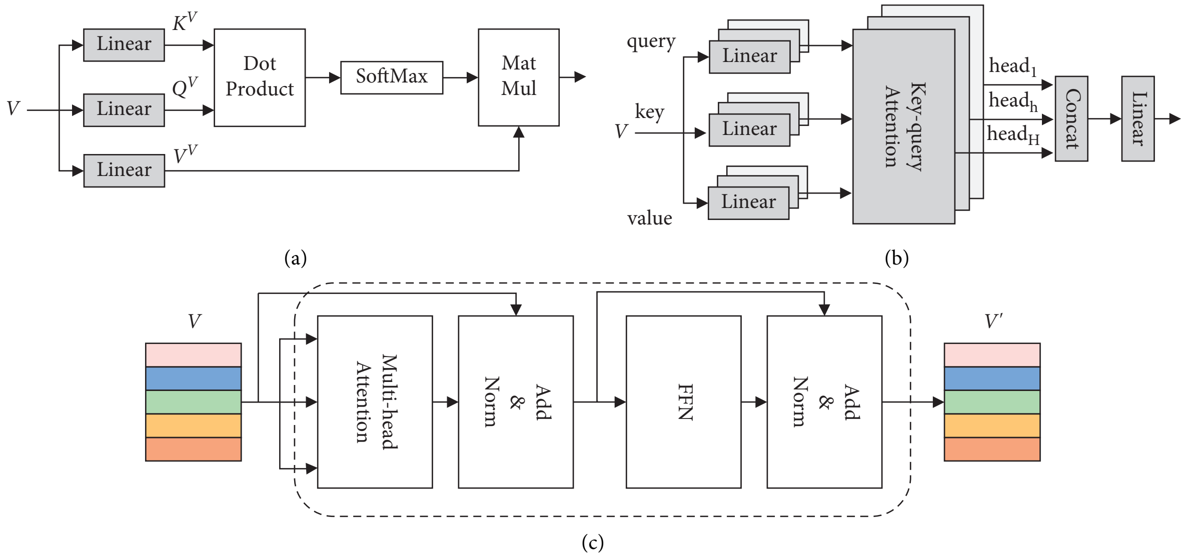

Figure 3: The detailed design of proposed visual context learning module. (a) Key-query attention, (b) multihead attention, (c) visual context learning module.

Where the $W_1, b_1, W_2$, and $b_2$ are the parameters of FFN, and its hidden and output dimensions are $768 \times 4$ and 768, respectively. ReLu is the rectified linear unit. In addition, the residual connection and layer normalization are also applied after FFN to facilitate optimization.

By performing the above procedure, the visual contexts that represent high-order interaction and context-dependence among all objects can be captured. Such procedure is simplified by the visual context learning module (VCL) as

$$V' = \text{VCL}(V). \tag{11}$$

The textual context learning module can work in the same manner on the word features $T$, but with a different set of parameters to be learned; thus details are omitted and it is also simplified as

$$T' = \text{TCL}(T). \tag{12}$$

Since the contextual path dependency distance between all entities is 1, the visual-textual context can be easily modelled by feeding the connected visual-textual feature matrix $V' \| T' \in R^{(K+M)\times 768}$ into the visual-textual context learning module (VTCL) with the same context learning mechanism.

$$V_{\text{out}}, T_{\text{out}} = \text{VTCL}\left(V' \| T'\right). \tag{13}$$

Then, we combine three context learning modules to form a multiple context learning layer (MCL). Finally, such MCL layer is stacked in depth to learn the deeper and more high-level context information, and the output of the last layer is the input of the next layer, expressed as $V^N, T^N = M\,CL_N\,(V^{N-1}, T^{N-1})$.

*2.3. Multimodal Fusion and Answer Prediction.* After $N$ MCL layers, the output image object features $V^N$ and question word features $T^N$ not only contain the intra-modal contexts but also capture the inter-modal context dependencies. To distinguish important entities from context information and fuse multimodal features, an attention model with an FFN (its hidden and output dimensions are 768 and 1) is designed for $V^N$ and $T^N$ to obtain attended image feature $v$ and question feature $q$. Taking image object features $V^N$ as an example, the attended image features $v$ are calculated as follows:

$$v = \sum_{i=1}^{K} a_i v_i^N, \quad \text{where } \alpha_i = \text{softmax}\left(\text{FFN}\left(v_i^N\right)\right). \tag{14}$$

$\alpha_i$ is the learned object attention weight for $i$-th object. The learned word attention weight $\beta_l$ and the attended question feature $q$ can be obtained using an independent attention model by analogy.

After obtaining image feature $v$ and question feature $q$, two linear transformations are implemented on $v$ and $q$, and using the addition to get joint representation $z$. Finally, $z$ is fed into a fully connected layer followed by a sigmoid function to generate the answer vector $p \in R^{|A|}$.

$$\begin{aligned} z &= \text{LN}\left(W_v^T v + W_q^T q\right), \\ p &= \text{sigmaid}\left(W_z z + b_z\right). \end{aligned} \tag{15}$$

$W_v^T \in R^{1024\times 768}$, $W_q^T \in R^{1024\times 768}$, and $W_z \in R^{|A|\times 1024}$ are three linear projection matrixes, $b_z$ is the bias parameter, sigmoid $(\cdot)$ is used for classification, and $|A|$ is the number of candidate answers.

## 3. Experiment

*3.1. Datasets and Evaluation Metric.* VQA v2.0 dataset is the most commonly used large-scale VQA dataset [4], which contains 1.1 M questions asked by humans and 10 answers are collected for each image-question pair from human annotators. The answer with the highest number of occurrences will be regarded as the correct answer. The dataset is divided into three parts: a training set containing 80 K images and 444 K questions, a validation set containing 40 K images and 214 K questions, and a test set containing 80 K images and 448 K questions. Additionally, based on the answer category, all questions are divided into three types: yes/no, number, and other; we use the following accuracy as the evaluation metric for answering quality.

$$\text{accuracy}\,(\text{answer}) = \min\left(\frac{\text{provided}\,(\text{annotators})}{3}, 1\right). \tag{16}$$

To demonstrate the generalization of approach, we further evaluate the model on the GQA dataset [4]. GQA dataset is the latest large-scale VQA dataset containing more than 110 K images and 22 M questions [5]. The dataset is divided randomly into proportions of 87%, 12%, and 1% for train, validation, and test sets, respectively.

### 3.2. *Experimental Setup*

*3.2.1. Universal Setup.* The proposed MCLN is implemented by PyTorch and all the experiments are conducted on a workstation; the experimental environment is shown in Table 1. The universal hyperparameters of MCLN model that are used in the experiments are listed as follows. For the pretrained Faster R-CNN, we follow the strategies in [13] to set its parameters and obtain a dynamic number of objects $K \in [10, 100]$. For textual features, we use either a LSTM with a 300-dimensional GloVe word embedding size or a pretrained BERT model with the 768-dimensional embedding size. In all context learning modules, we set the number of heads $H$ as 12, so the latent dimensionality for each head is $d_h = d/12 = 64$. Adam Optimizer [31] is used to optimize the model with 64 batch sizes.

*3.2.2. Setup for VQA v2.0.* The answers with an occurrence rate less than 8 times in the training and validation sets are discarded, which resulted in $|A| = 3129$ candidate answers. The maximum length of tokenized words is $L = 14$. Moreover, the number of MCL layers is $N \in \{1, 2, 3, 4\}$. For the choice of learning rate $lr$, the warm-up strategy is employed. Specifically, the initial learning rate is $2.5 \times 10^{-5}$ and grows by $2.5 \times 10^{-5}$ at each epoch till it reaches $1 \times 10^{-4}$ at epoch 4. After 10 epochs, the learning rate is decreased by 1/5 for every epoch up to 12 epochs. Since there exist multiple correct answers for a question in the VQA v2.0 dataset, the binary cross-entropy loss (BCE) is applied to optimize MCLN:

$$\text{BCE loss} = -\sum_{i=1}^{|A|} \left(y_i \log\left(p_i\right) + \left(1 - y_i\right)\left(\log 1 - p_i\right)\right), \tag{17}$$

where $y_i$ is the given score by the datasets and $p_i$ is the predicted score by the MCLN model and corresponds to the $i$-th element in the answer vector $p$.

Table 1: Experimental environment.

| Configuration | Details |
|---|---|
| Operating system | Ubuntu 18.04 |
| RAM | 64 G |
| Graphic processing unit | NVIDIA GeForce GTX 2080 Ti |
| Programming language | Python 3.6 |
| Deep learning framework | PyTorch 1.0.1 |
| Architecture platform | CUDA10.0 and CUDNN7.4 |

*3.2.3. Setup for GQA.* The GQA dataset provided $|A| = 1878$ candidate answers. The maximum length of tokenized words is $L = 29$. We also use the same learning rate strategy on the GQA dataset, while after 8 epochs the learning rate is decreased by 1/5 for each up to 10 epochs. For the loss function, we choose the cross-entropy loss (CE) to optimize our model on GQA dataset.

$$\text{CE loss} = -\sum_{i=1}^{|A|} y_i \log\left(p_i\right). \tag{18}$$

*3.3. Ablation Studies.* On the validation sets of VQA v2.0 and GQA, the proposed MCLN will execute extensive ablation studies from three different aspects: (1) the effectiveness of the three context learning modules in the proposed MCL network architecture; (2) the impact of different layers on the performance of MCLN; (3) the effect of enhancing the textual context.

*3.3.1. Effect of the Context Learning Module.* As mentioned, three modules are designed to perform corresponding context learning. To demonstrate the effectiveness of the proposed context learning modules, we employ the MCLN model with LSTM to quantitatively ablate three modules. Table 2 shows the results of various ablated versions of the model on the VQA v2.0 and GQA validation sets. It has been discovered that induction and elimination of any module have an impact on the performance. The optimum results are only achieved by integrating all three components. Results indicate that every context learning module contributes significantly to VQA performance and demonstrating the effectiveness of context learning modules. The result on VTCL module is noteworthy. For the models on VQA v2.0 dataset, model 4 only with the VTCL module obtains 62.07% overall accuracy, which outperforms model 1 more than 7.47% and higher performance is gained compared to model 2 and model 3. Furthermore, when the VTCL module is disabled in model 7 compared to model 8, a rapid decrease of 9.92% is observed. It indicates that the visual-textual context ignored by prior approaches is crucial to VQA.

*3.3.2. Effect of MCL Layers.* Next, we stack the full-module MCL layer in depth to evaluate the effect of MCL layers. Figure 4 shows the performance of the MCLN models with the different numbers of MCL layers $N \in \{1, 2, 3, 4\}$ on the VQA v2.0 validation set. Regarding the performance, we observed two phenomena on the overall accuracy: (1) as increasing N, the performances of MCLN models steadily improve and finally saturate at a certain number $N = 3$; (2)

Table 2: The results of ablating the context learning modules on VQA v2.0 and GQA validation sets.

| Model | Module | GQA | VQA v2.0 | | | |
|---|---|---|---|---|---|---|
| | | All | All | Y/N | Num | Other |
| 1 | Without all | 53.08 | 54.60 | 69.79 | 36.02 | 47.50 |
| 2 | Only VCL | 53.45 | 55.13 | 69.82 | 36.09 | 47.99 |
| 3 | Only TCL | 53.50 | 55.53 | 69.82 | 36.44 | 49.72 |
| 4 | Only VTCL | 58.63 | 62.07 | 79.79 | 42.67 | 53.73 |
| 5 | TCL + VTCL | 63.88 | 65.17 | 82.88 | 44.68 | 57.15 |
| 6 | VCL + VTCL | 59.04 | 62.72 | 79.33 | 43.29 | 55.23 |
| 7 | VCL + TCL | 53.72 | 55.76 | 71.01 | 36.37 | 49.29 |
| 8 | Full modules | 64.48 | 65.68 | 83.40 | 45.57 | 57.53 |

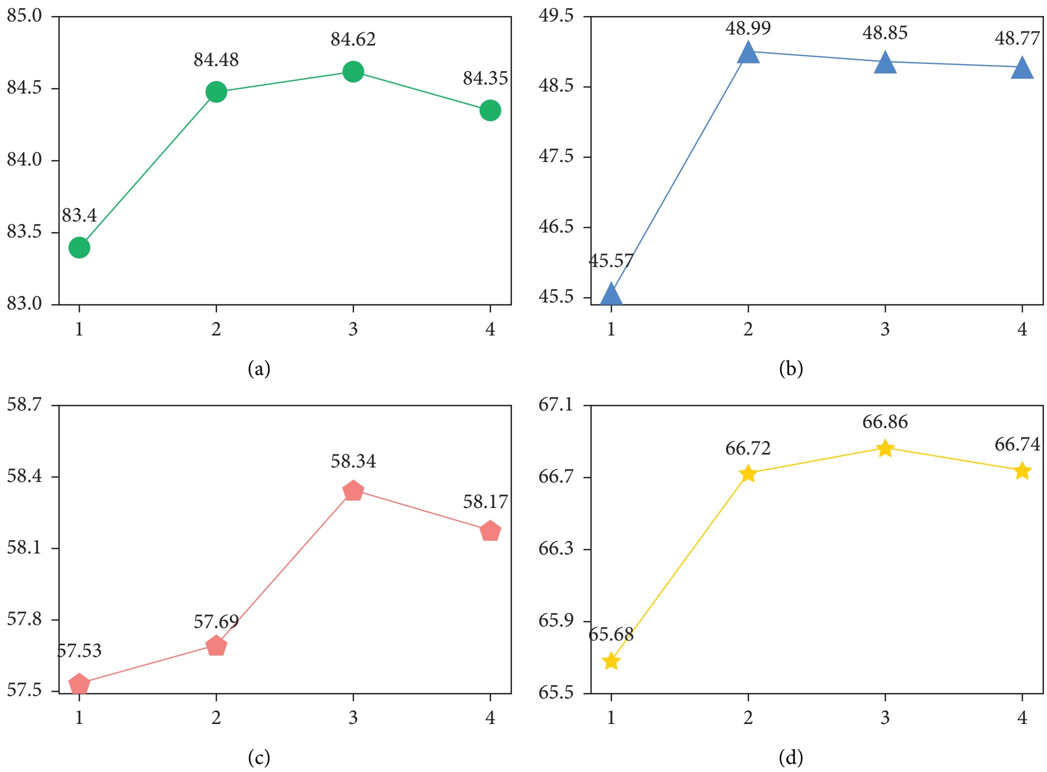

Figure 4: The overall and per-type accuracies of the MCLN with different MCL layers on the VQA v2.0 validation set. (a) Yes/no, (b) number, (c) others, and (d) overall.

the one-layer model does not perform as well as the deeper models, but there is a quite improvement over $N = 1$ while a subtle decrease over $N = 3$ when $N = 4$. In addition, the performance of the MCLN with the different number of MCL layers on the GQA validation set is shown in Figure 5. Similar experimental phenomena are observed and the best performance appears at $N = 2$. From the experimental results, it can be seen that deeper MCL layers usually obtain higher overall accuracy. We attribute it to deep context learning, which captures more complex context information and achieves a better contextual understanding of the image and question contents. However, comparing the shallow model and deep model, the number of layers increases and the parameters of the model rise as well. As a result, the model with deeper MCL layers is difficult to be optimized and suffers from a larger risk of overfitting the training set due to the higher number of parameters. Therefore, the overall performance decreases at $N = 4$ and $N = 3$ on the VQA v2.0 and GQA, respectively.

*3.3.3. Effect of Enhancing the Textual Context.* Since the pretrained BERT [23] was trained on large text corpus and models the textual context by stacking the bidirectional encoder, to enhance the textual context at the textual features extraction stage, a contextualized text encoder based on BERT is introduced and fine-tuned. The results of fine-tuning the weight of pretrained BERT with different learning rates are shown in Tables 3 and 4. On the VQA v2.0, the best performance is achieved at $lr \times 0.1$ under one-layer MCL. With this learning rate, by increasing the MCL layers, the performance grows and also saturates at three MCL layers.

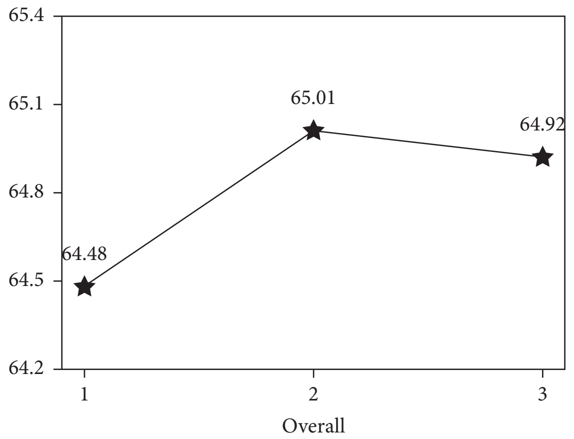

Figure 5: The overall accuracies of the MCLN with different MCL layers on the GQA validation set.

Table 3: The results of fine-tuning BERT on VQA v2.0 validation set.

| Model | $lr\times$ | All | Y/N | Num | Other |
|---|---|---|---|---|---|
| $N=1$, BERT | 0.001 | 65.21 | 82.61 | 45.74 | 57.13 |
| $N=1$, BERT | 0.01 | 66.12 | 84.08 | 45.97 | 57.80 |
| $N=1$, BERT | 0.1 | 66.61 | 85.09 | 46.14 | 57.98 |
| $N=2$, BERT | 0.1 | 67.52 | 85.18 | 49.09 | 58.97 |
| $N=3$, BERT | 0.1 | 67.86 | 85.35 | 49.75 | 59.27 |
| $N=4$, BERT | 0.1 | 67.80 | 85.73 | 49.94 | 58.94 |
| $N=3$, LSTM | - | 66.86 | 84.62 | 48.85 | 58.34 |

Table 4: The results of fine-tuning BERT on GQA validation set.

| Model | $lr\times$ | All |
|---|---|---|
| $N=1$, BERT | 0.001 | 64.25 |
| $N=1$, BERT | 0.01 | 64.63 |
| $N=1$, BERT | 0.1 | 64.51 |
| $N=2$, BERT | 0.01 | 65.22 |
| $N=3$, BERT | 0.01 | 65.11 |
| $N=2$, LSTM | - | 65.01 |

On the GQA dataset, due to the differences in datasets and experimental settings, the best performance is achieved by fine-tuning the learning rate of BERT to $lr\times 0.01$ and setting the MCL layer to 2. Compared with the MCLN-LSTM using three-layer MCL, 1% overall accuracy is improved by the MCLN-BERT using three-layer MCL and $lr\times 0.1$ learning rate on the VQA v2.0 validation set. On the GQA validation set, when the MCL layer is set to two and fine-tuning the learning rate of BERT to $lr\times 0.01$, the gain is 0.11% compared with MCLN-BERT and MCLN-LSTM. From these experimental and compared results, it can be found that BERT is compatible with MCLN and conducive to boosting the performance by fine-tuning it to enhance textual context at the textual features extraction stage.

*3.4. Visualization Analysis.* Assuming that, after processing the object and word features through multiple MCL layers, the keywords in a question and the relevant image objects related to the answer can be well distinguished by the multimodal fusion and answer prediction networks according to the multiple context information. Thus, to intuitively illustrate the effectiveness of MCL layers, we selected three models to visualize the learned attention weights by equation (14): the MCLN-w/o without the MCL layer, the MCLN-1 with one-layer MCL, and the MCLN-3 with three MCL layers. The top-2 object attention weights and all word attention weights are visualized in Figure 6. As can be seen from the visualization of attention weights, MCL contributes to focusing on relevant objects and keywords. For the correctly predicted example by three models, although the MCLN-w/o can locate the elephants, the MCLN-1 and MCLN-3 can focus on the target objects more fine-grained and the target objects with the higher attention weight value. The incorrectly predicted examples by the MCLN-w/o model are more complex and require a comprehensive understanding of the contexts of image and question. Due to the lack of MCL, MCLN-w/o is not able to accurately locate keywords and relevant objects according to the context information, resulting in wrong answers. For example, MCLN-w/o mainly pays attention to the "people" but ignores other keywords (e.g., "many" and "board"), which means that the context-dependency of "people on board" cannot be captured. In addition, compared with MCLN-1, more reasonable attention weights are usually learned by the MCLN-3. As shown in Figure 6, MCLN-2 mainly attend the keywords like "people," "on," and "board" in the second question. In the third question, the mainly attended keywords by the MCLN-2 are "person," "white," and "holding." In other words, the deep-level textual contexts like "people on board" and "white person holding" are captured by MCLN-3.

*3.5. Comparison with State-of-the-Art Methods.* In Tables 5 and 6, we compare our MCLN model with the current state-of-the-art (SOTA) models on the GQA and VQA v2.0, respectively. And the text encoder of compared models is based on RNN architecture. The proposed MCLN-LSTM is based on RNN architecture, while the text encoder of proposed MCLN-BERT is based on BERT and it enhances the textual context learning at the textual features extraction stage. Furthermore, two proposed MCLN are deep context learning models; they not only consider two intra-modal contexts but also learn the ignored visual-textual context.

On the GQA dataset, the compared results of two MCLN models and the SOTA models are shown in Table 5. Among them are CNN + LSTM [5], BUTD [13], and MAC [15] without any context learning. LCGN [19] is a deep context learning model and OCCAM [31] is a shallow context learning model, but they both only take into account the context of the image modality. The highest accuracy of the model without context learning is 54.1% and the highest accuracy of the model with visual context learning is 56.3% on the test set. For the MCLN-LSTM also using RNN to extract the word features, compared with MAC and OCCAM, the accuracy is 2.5% and 0.3% higher, respectively, while for the MCLN-BERT employing BERT, higher gains,

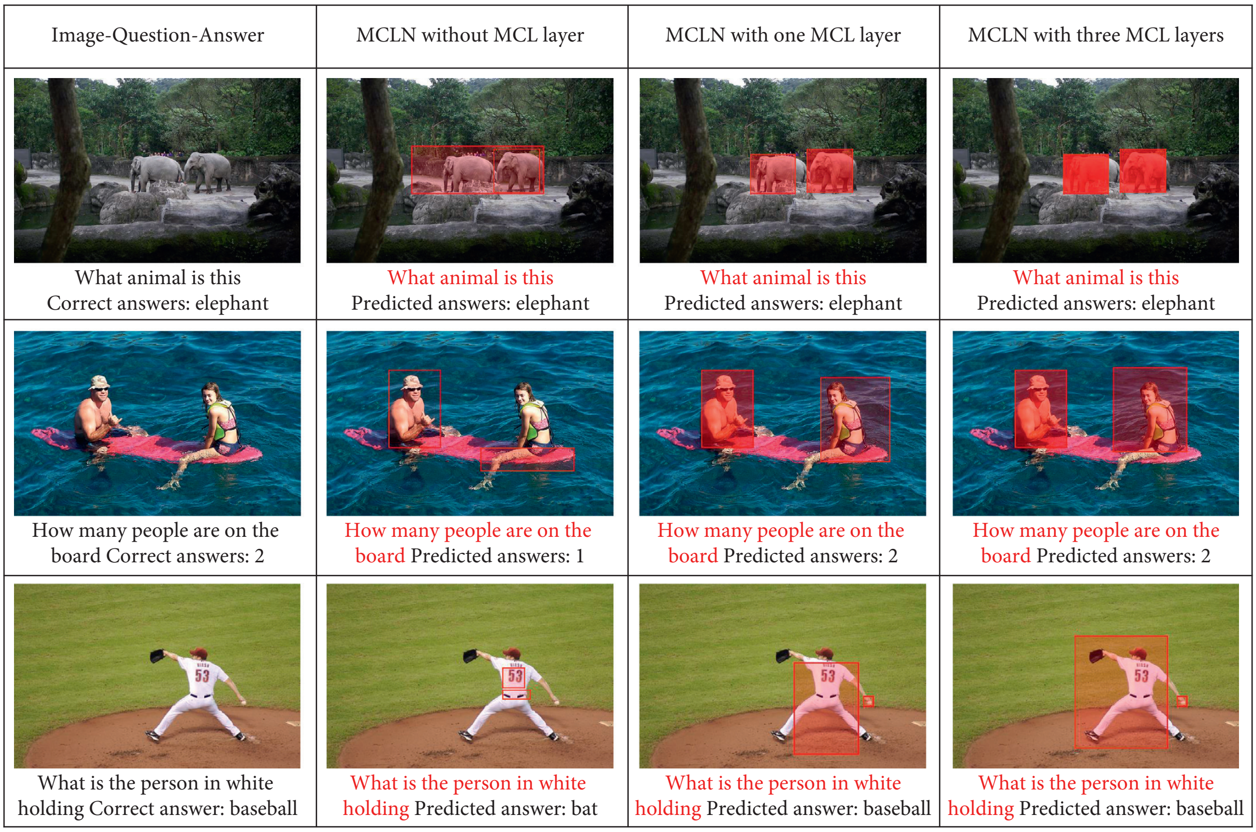
Image-Question-Answer
MCLN without MCL layer
MCLN with one MCL layer
MCLN with three MCL layers
What animal is this Correct answers: elephant
What animal is this Predicted answers: elephant
What animal is this Predicted answers: elephant
What animal is this Predicted answers: elephant
How many people are on the board Correct answers: 2
How many people are on the board Predicted answers: 1
How many people are on the board Predicted answers: 2
How many people are on the board Predicted answers: 2
What is the person in white holding Correct answer: baseball
What is the person in white holding Predicted answer: bat
What is the person in white holding Predicted answer: baseball
What is the person in white holding Predicted answer: baseball

Figure 6: Visualizations of the learned attention weights on the VQA v2.0 dataset. The colors ranging from clear to red on image objects or words denote the attention weights from 0 to 1.

Table 5: Comparison with the current state-of-the-art methods on GQA test datasets.

| Model | Test-dev | Test |
|---|---|---|
| CNN + LSTM [5] | — | 46.6 |
| BUTD [13] | — | 49.7 |
| MAC [15] | — | 54.1 |
| LCGN [19] | 55.8 | 56.1 |
| OCCAM [31] | 56.2 | 56.3 |
| MCLN-LSTM | 56.4 | 56.6 |
| MCLN-BERT | 56.8 | 57.0 |

Table 6: Comparison with previous state-of-the-art methods on VQA v2.0 test dataset.

| Model | Test-dev | | | | Test-std |
|---|---|---|---|---|---|
| | All | Y/N | Num | Other | All |
| BUTD [13] | 65.32 | 81.82 | 44.21 | 56.05 | 65.67 |
| MFH [32] | 68.76 | 85.31 | 49.56 | 59.89 | - |
| Counter [33] | 68.09 | 83.14 | 51.62 | 58.97 | 68.09 |
| v-AGCN [17] | 65.94 | 82.58 | 45.12 | 56.71 | 66.17 |
| ReGAT [16] | 70.27 | 86.08 | 54.42 | 60.33 | 70.58 |
| DFAF [20] | 70.22 | 86.09 | 53.32 | 60.49 | 70.34 |
| MCAN [21] | 70.63 | 86.82 | 53.26 | 60.72 | 70.90 |
| MEDAN [22] | 70.60 | 87.10 | 52.69 | 60.56 | 71.01 |
| MCLN-LSTM | 70.26 | 85.95 | 53.18 | 60.72 | 70.63 |
| MCLN-BERT | 71.05 | 87.43 | 53.28 | 61.08 | 71.48 |

2.9% and 0.7% are achieved compared with MAC and OCCAM. Additionally, MCLN-LSTM and MCLN-BERT both overperform the SOTA models, proving the excellent performance of the proposed model on the GQA.

On the test sets of VQA v2.0, the proposed models are compared with SOTA models to further verify their advantage. The compared models include three models without context learning (BUTD [13], MFH [32], Counter [33]), two shallow models that only learn the visual context (ReGAT [16], v-AGCN [17]), and three deep models with intra-modal context learning (DFAF [20], MCAN [21], MEDAN [22]). Table 6 shows the compared experiment results on the VQA v2.0 test-dev and test-std sets. For the RNN-based MCLN-LSTM model, compared with three methods that ignore the context learning, on the test-dev set, 4.94%, 1.5%, and 2.17% overall accuracy are improved by the MCLN-LSTM. ReGAT only models the visual context on the image modality, which combines various relation learning models and the Counter model [33]. By comparison, the single MCLN-LSTM not only is higher than ReGAT on test-std, but also has an advantage in the number of models. The DFAF model ignoring the visual-textual context captures intra-modal contexts via two intra-modality attention flows and uses two separate inter-modality attention flows to achieve the inter-modal interactions, but the two separate inter-modality information flows do not reflect the synergic context dependence of image objects and question words under the two-modality information. On the contrary, the MCLN-LSTM not only captures intra-modal contexts, but also achieves the synergic visual-textual context learning as well as the inter-modal interactions through a single VTCL module. As shown in Table 3, MCLN-LSTM is 0.04 and 0.29 points higher than DFAF on test-dev and test-std. By using the encoder-decoder architecture, MCAN and MEDAN achieve optimal performance. Even though the overall accuracy of the proposed MCLN-LSTM model is lower than MCAN and MEDAN, where BERT is introduced to enhance the textual context at the textual features extraction stage, the MCLN-BERT model overperforms these two models on the overall accuracy. The results indicate that MCLN considering comprehensive contexts is superior. Furthermore, compared with other state-of-the-art models, the significant improvements are obtained by MCLN-BERT and it gains 71.05% and 71.48% overall accuracies on the test-dev and test sets, respectively.

From the compared experiment results on GQA and VQA v2.0 dataset, it can be seen that the proposed MCLN gains better results due to the fact that it considers more comprehensive contexts. On one hand, MCLN not only captures two intra-modal contexts but also learns the ignored visual-textual context. On the other hand, MCLN enhances the textual context learning at the textual features extraction stage.

## 4. Conclusions

This article presents a novel framework, the Multiple Context Learning Network (MCLN), to model multiple context learnings for visual question answering. The MCLN exploits three types of contexts, the visual context and textual context in the intra-modalities, and the visual-textual context, to learn a context-aware representation through the corresponding context learning module. The core idea of the context learning module is to establish the contextual dependency of all entities in an entity set by using key-query attention mechanism. Our context learning approach is simple but extremely effective. The MCLN is better able to learn complex contexts and improve VQA performance by composing three context learning modules to construct MCL layer and stacking such layer in depth. Furthermore, BERT-based text encoder is introduced and fine-tuned to facilitate the textual context learning at the textual features extraction stage. Experimental results on two large-scale benchmark datasets show that our proposed method outperforms the previous state-of-the-art methods. And the extensive ablation studies demonstrate the effectiveness of context learning modules, stacking MCL layers, and enhancing the textual context. However, stacking MCL layers will increase the training time of the model and make the model difficult to be optimized in practical experiments; we will optimize our model to achieve a rapid network framework in a future study. In addition, we plan to promote the proposed context learning method to other multimodal tasks such as image captioning and image-text matching.

## Data Availability

The dataset used to support the findings of this study is available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61762046 and the Natural Science Foundation of Jiangxi Province under Grant 20161BAB212048.

## References

[1] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, “Learning to compose and reason with language tree structures for visual grounding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, 2019.

[2] J. Donahue, L. A. Hendricks, M. Rohrbach et al., “Long-term recurrent convolutional networks for visual recognition and description,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[3] F. Huang, X. Zhang, Z. Zhao, and Z. Li, “Bi-directional spatial-semantic attention networks for image-text matching,” *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 2008–2020, 2018.

[4] 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 *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern*

*Recognition (CVPR)*, pp. 6325–6334, Honolulu, HI, USA, July 2017.

[5] D. A. Hudson and C. D. Manning, “GQA: a new dataset for real-world visual reasoning and compositional question answering,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, Long Beach, CA, USA, June 2019.

[6] J.-H. Kim, S.-W. Lee, D.-H. Kwak et al., “Multimodal residual learning for visual QA,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 361–369, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, “Simple baseline for visual question answering,” ‘, https://arxiv.org/abs/1512.02167, 2015.

[10] M. Ren, R. Kiros, and R. Zemel, “Exploring models and data for image question answering,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, vol. 28, pp. 2953–2961, Bali, Indonesia, November 2015.

[11] K. J. Shih, S. Singh, and D. Hoiem, “Where to look: focus regions for visual question answering,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4613–4621, Las Vegas, NV, USA, June 2016.

[12] J. Song, P. Zeng, L. Gao, and H. T. Shen, “From pixels to objects: cubic visual attention for visual question answering,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 906–912, Stockholm, Sweden, July 2018.

[13] P. Anderson, X. He, C. Buehler et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.

[14] D.-K. Nguyen and T. Okatani, “Improved fusion of visual and language representations by dense symmetric Co-attention for visual question answering,” in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6087–6096, Long Beach, CA, USA, June 2018.

[15] D. A. Hudson and C. D. Manning, “Compositional attention networks for machine reasoning,” in *Proceedings of the International Conference on Learning Representation (ICLR)*, Vancouver, Canada, May 2018.

[16] L. Li, Z. Gan, Y. Cheng, and J. Liu, “Relation-aware graph attention network for visual question answering,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10313–10322, Long Beach, CA, USA, June 2019.

[17] J. Yu, W. Zhang, Z. Yang, Z. Qin, and Y. Hu, “Cross-modal learning with prior visual relation knowledge,” *Knowledge-Based Systems*, vol. 203, Article ID 106150, 2020.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proceedings of the International Conference on Learning Representations*, Vancouver Canada, April 2018.

[19] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, “Language-conditioned graph networks for relational reasoning,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10294–10303, Long Beach, CA, USA, June 2019.

[20] P. Gao, Z. Jiang, H. You et al., “Dynamic fusion with intra- and inter-modality attention flow for visual question answering,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6639–6648, Long Beach, CA, USA, June 2019.

[21] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular Co-attention networks for visual question answering,” in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6281–6290, Long Beach, CA, USA, June 2019.

[22] C. Chen, D. Han, and J. Wang, “Multimodal encoder-decoder attention networks for visual question answering,” *IEEE Access*, vol. 8, pp. 35662–35671, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 30, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional Transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Long and Short Papers, Minneapolis, MI, USA, June 2018.

[25] P. Zhang and H. Lan, “Multiple context learning networks for visual question answering,” 2021, https://europepmc.org/article/ppr/ppr407117.

[26] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[27] R. Krishna, Y. Zhu, O. Groth et al., “Visual Genome: connecting language and vision using crowdsourced dense image annotations,” *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[29] J. Pennington, R. Socher, and C. D. Manning, “Glove: global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014.

[30] D. P. Kingma and L. B. Jimmy, “Adam: a method for stochastic optimization,” in *Proceedings of the ICLR 2015: International Conference on Learning Representations 2015*, San Diego, CA, USA, May 2015.

[31] Z. Wang, K. Wang, M. Yu et al., “Interpretable visual reasoning via induced symbolic space,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1878–1887, Seattle, WA, USA, June 2020.

[32] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, “Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[33] Y. Zhang, J. Hare, and P.-B. Adam, “Learning to count objects in natural images for visual question answering,” in *Proceedings of the International Conference on Learning Representations*, Vancouver, Canada, April 2018.