Hindawi
Mobile Information Systems
Volume 2021, Article ID 2662064, 9 pages
https://doi.org/10.1155/2021/2662064

Research Article

RDMMFET: Representation of Dense Multimodality Fusion Encoder Based on Transformer

Xu Zhang[1], **DeZhi Han**[1] **and Chin-Chen Chang**[2]

[1]*College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*
[2]*Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan*

Correspondence should be addressed to DeZhi Han; dzhan@shmtu.edu.cn

Received 17 June 2021; Accepted 7 September 2021; Published 18 October 2021

Academic Editor: Chin-Ling Chen

Visual question answering (VQA) is the natural language question-answering of visual images. The model of VQA needs to make corresponding answers according to specific questions based on understanding images, the most important of which is to understand the relationship between images and language. Therefore, this paper proposes a new model, Representation of Dense Multimodality Fusion Encoder Based on Transformer, for short, RDMMFET, which can learn the related knowledge between vision and language. The RDMMFET model consists of three parts: dense language encoder, image encoder, and multimodality fusion encoder. In addition, we designed three types of pretraining tasks: masked language model, masked image model, and multimodality fusion task. These pretraining tasks can help to understand the fine-grained alignment between text and image regions. Simulation results on the VQA v2.0 data set show that the RDMMFET model can work better than the previous model. Finally, we conducted detailed ablation studies on the RDMMFET model and provided the results of attention visualization, which proves that the RDMMFET model can significantly improve the effect of VQA.

1. Introduction

Visual question answering (VQA) [1] is a task that combines CV [2] and NLP [3]. The VQA system takes an image and a question as input and generates an answer as output.

So far, the introduction of the attention mechanism [4] has made VQA a major advancement. It is essentially similar to how humans observe the world, and the purpose is to select what you want to know from a wide range of information. VQA was first proposed by Vaswani et al. [5]. The VQA task is based on the two modalities of image and vision; some single-modal models do not work well on the VQA task. Thus, a key goal of studying VQA models is to effectively aggregate visual and language modal information. For example, to find the correct answer to the question in the VQA task, the model should have the ability to integrate language information and visual information and align language features with visual features to analyze the answer to the question. Models such as MCB [6], BAN [7], DFAF [8], and MCAN [9] all adopt advanced multimodality fusion methods.

Despite various multimodality fusion methods exist in abundance, there is no mature and unified architecture. Hence the pretraining task [10] is proposed to solve the gap for generality and then applied to other tasks. Jacob Devlin et al. propose a large-scale pretraining model "Bidirectional Encoder Representations from Transformers" (BERT) [11]. Owing to its strong learning ability, VideoBert [12] is the first to apply BERT to a multimodality model. Since then, BERT has gradually been used in a variety of multimodality fields. The performance effectively proves that pretraining can significantly improve the effectiveness of the model.

Although the current attention model performs well, the understanding of the problem modal is still insufficient. When a dense language encoder is used to extract problem features, the model can better learn the complex relationship between words, which helps the subsequent multimodality fusion encoder to better understand the relationship between language and image. Therefore, we propose a new model RDMMFET, which uses the transformer as the core to construct a deeply cascaded dense language encoder, image

encoder, and multimodality fusion encoder. The core of each encoder is the self-attention layer and the feed-forward layer. Then the RDMMFET model is pretrained from three aspects of language image and multimodality fusion. We test the performance of the RDMMFET model on the large-scale visual question and answer data set VQA v2.0. The accuracy of the model on the Test-dev set is 72.59%, and the accuracy on the Test-std set is 72.67%, both of which are higher than the previous model.

The main contributions of this paper are as follows:

(1) The RDMMFET model based on dense language encoder and multimodality fusion encoder is proposed.

(2) The RDMMFET model is pretrained from three aspects: masked language model, masked image model, and multimodality fusion task. These pretraining tasks can make the model better understand the association between images and language and achieve cross-modal alignment of image features and language features.

(3) We test the RDMMFET model on the large-scale data set VQA v2.0. The accuracy of the model on Test-dev is 72.59% and on Test-std is 72.67%, which are both higher than the previous model.

(4) Extensive ablation studies are performed on the model, and the results of attention visualization are provided to explain the effectiveness of the RDMMFET model.

The remainder of the paper is organized as follows: Section 2 introduces the latest research on the multimodality fusion model and pretraining task. Then, the research and design of the overall framework of the RDMMFET model are presented in Section 3. Next, Section 4 proposes three types of pretraining tasks for the RDMMFET model. In Section 5, ablation research and attention visualization of the RDMMFET model are presented. Finally, in Section 6, this paper makes a summary.

2. Related Work

We survey the related work in two parts. First, we present the related research about the multimodality fusion model in Section 2.1, and then we give an overview of the pretraining task in Section 2.2.

2.1. Multimodality Fusion Model. Multimodality fusion refers to the combination of two or more modalities in various forms. The concept of multimodality fusion first came from the image caption generation task [13, 14]. Multimodality fusion methods include linear fusion and bilinear pooling. Linear fusion methods include feature connection, element multiplication, and bilinear merging and then use these methods to calculate the outer product [15].

To reduce the dimensionality of the features without reducing the performance, Fukui et al. propose a multimodality bilinear method to map image and text features to a higher two-dimensional space. Kim et al. present a multimodality low-rank bilinear (MLB) algorithm, which maps the features of the two patterns to the same one-dimensional space, uses Hadamard product for fusion in this space, and finally maps the fused features to another dimension [16]. Yu et al. point out that MLB has the problem of slow convergence speed and proposed the multimodality factorized linearity (MFB) [17], which uses a matrix factorization strategy to calculate the fusion features. Yu et al. propose a multimodality factorization high-order pool (MFH) method [18]. The model has achieved remarkable performance on the VQA data set. Kim et al. present a vision-and-language transformer model (ViLT) [19]. ViLT uses pretrained ViT to initialize the interactive transformer so that you can directly use the interactive layer to process visual features without adding an additional visual encoder. Nagrani et al. introduced a new transformer-based architecture that uses "fusion bottlenecks" for modal fusion at multiple layers. The architecture has achieved state-of-the-art results on multiple audio-visual classification benchmarks [20].

2.2. Pretraining Task. The pretraining task uses self-supervised learning to obtain pretraining models that are not related to specific tasks from large-scale data. Training methods for pretraining tasks can use self-supervised learning techniques (such as autoregressive language models and auto-encoding techniques), which can train single-language, multilanguage, and multimodality models. In 2016, the semisupervised sequence learning proposed by Dai and Le [21] uses language modeling and sequence self-encoding to improve the sequence learning of cyclic neural networks [22], which can be considered as the beginning of the pretraining task. It systematically explains the epoch-making idea that the upstream pretraining language models can be used for the downstream specific tasks. With the development of computing power, the deep model has also been continuously improved, and the architecture of the pretraining task has been advanced from the shallower to the deeper. In 2018, ELMO [23] proposes a context-sensitive text representation method, which performed amazingly well on many typical tasks and could effectively deal with the problem of ambiguity. Finally, many pretrained language models are proposed, such as GPT [24] and BERT.

Compared with models such as ELMO and GPT, the first innovation of BERT is to use Masked LM (MLM) to achieve the purpose of deep two-way joint training. MLM pretraining obtains the probability distribution of the position in the output layer by randomly covering parts of the words in the input text sequence and then maximizes the likelihood probability to adjust the model parameters. Owing to the strong learning ability of BERT, it began to be gradually used in the multimodality field in 2019. Vilbert [25] and LXMERT [26] introduce the dual-stream structure. On the contrary, many models with single-stream structure also are proposed, such as B2T2 [27], Unicode VL [28], VisualBERT [29], and VL-BERT [30].

3. Model Architecture

Our model takes the transformer as the core and establishes a dense language encoder, image encoder, and multimodality fusion encoder. The overall framework is shown in Figure 1. We use an image and an image-related problem as the input of the RDMMFET model. Then we convert the problem and the image into image and problem features. Next, the self-attention language and image features are obtained from the encoder. Finally, the language and image features are introduced into the multimodality fusion encoder at the same time to get the answer to the problem. Next, we will describe our model in detail from three aspects: problem and image representation in Section 3.1, encoder in Section 3.2, and output representation in Section 3.3.

3.1. Problem and Image Representation. The input of the RDMMFET model is a question and an image, and the model transforms these two inputs into the corresponding word vector and object vector. Then these vectors are passed into the encoder for further processing.

3.1.1. Problem Representation. First, the input question q is transformed into a word vector W: $\{\text{CLS}, w_1, \ldots, w_i, \text{EOS}, \ldots, w_n\}$ by WordPiece tokenizer, where n is the length of the transformed word vector and i is the number of words in the specific question. EOS is the end of the word vector, and CLS is a special marker, which can be regarded as the answer to the VQA problem. Then, the transformed word vector W and its corresponding position index vector PI are transformed into fixed-length vectors $\widetilde{W}$ and $\widetilde{\text{PI}}$, respectively, by word embedding and Idx embedding. Finally, $\widetilde{W}$ and $\widetilde{\text{PI}}$ are added and transferred into the LayerNorm layer to learn the problem feature Q. The specific process is as follows:

$$\begin{aligned} W &= \text{WordPiece tokenizer}\ (q), \\ \widetilde{W} &= \text{Word embedding}\ (W), \\ \widetilde{\text{PI}} &= \text{Position embedding}\ (\text{PI}), \\ Q &= \text{LayerNorm}\ (\widetilde{W} + \widetilde{\text{PI}}). \end{aligned} \tag{1}$$

3.1.2. Problem Representation. First, the faster R-CNN [31] is used to find m objects $\{o_1, \ldots, o_m\}$ in the input image. Each object O is represented by its boundary box coordinates PF and its region of interest feature RF [32]. To balance PF and RF, we add two fully connected LayerNorm layers to learn the new position feature vector $\widetilde{\text{PF}}$ and interest feature vector $\widetilde{\text{RF}}$. Finally, the new position feature $\widetilde{\text{PF}}$ and interest feature $\widetilde{\text{RF}}$ are added and divided by 2 to get the image feature V. The specific process is as follows:

$$\begin{aligned} \widetilde{\text{PF}} &= \text{LayerNorm}\ (\text{PF}), \\ \widetilde{\text{RF}} &= \text{LayerNorm}\ (\text{RF}), \\ V &= \frac{\widetilde{\text{PF}} + \widetilde{\text{RF}}}{2}. \end{aligned} \tag{2}$$

3.2. Encoder. We have built a dense language encoder, an image encoder, and a multimodality fusion encoder based on transformers. Next, we will introduce attention mechanisms, dense language encoders, image encoders, and multimodality fusion encoders.

3.2.1. Attention Layer. The input is the query Q, key K, and value V [33], where the dimension of Q and K is d_k, and the dimension of V is d_v. By calculating the correlation between Q and K, the weight coefficient of V corresponding to each K is obtained, and it is normalized by softmax function. Then, the weight and the corresponding V are weighted to get the attention value. The specific process is as follows:

$$\begin{aligned} S &= \text{Socre}\,(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right), \\ \text{Attention}\,(Q, K, V) &= SV, \end{aligned} \tag{3}$$

where $\text{softmax}\,(\cdot)e$ is the normalization of softmax function and S is the weight matrix.

In this paper, we use the multiattention [34]. The input Q, K, and V are linearly transformed, and then they are input to the scaled dot product attention. This process requires h times. Each time Q, K, and V are queried, the parameter W of the linear transformation is different. Then, H times of attention are sutured as the result of multiattention. The specific process is as follows:

$$\begin{aligned} \text{HeadAtt}_i &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \\ &\text{MHAttention}\ (Q, K, V) \\ &= \text{Concat}\left(\text{HeadAtt}_1, \text{HeadAtt}_2, \ldots, \text{HeadAtt}_h\right)W^o, \end{aligned} \tag{4}$$

where W_i^Q, W_i^K, and W_i^V are the projection matrix corresponding to the Q, K, and V in the ith header; $\text{Concat}\,(\cdot)$ represents the concatenation of h headers; and W^o is the projection matrix of all headers.

3.2.2. Dense Language Encoder. After transforming the input question into question feature Q, we pass it into the dense language encoder. The dense language encoder only pays attention to the single mode of language. A single language encoder consists of a self-attention layer and a feed-forward layer, each of which contains two fully connected layers. We add residual join and layer standardization after each sublayer. Unlike other models, we choose to parallel the two language encoders and overlay the language encoders with N_L layers. We input the problem feature Q_0 into two parallel language encoders, then splice the output of the two language encoders as the input of the next dense language encoder, and splice the output Q_{N_L} of the last dense language encoder as the input of the multimodality fusion encoder. The specific process is as follows:

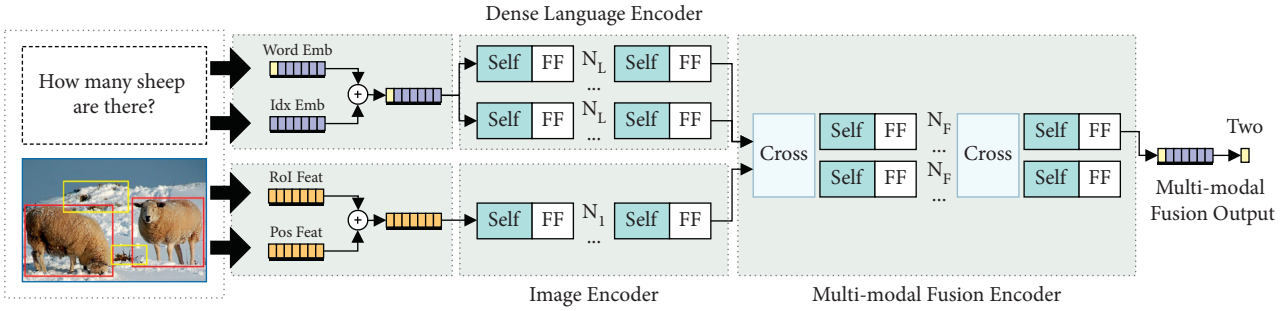


Figure 1: The overall structure of the RDMMFET model for learning visual and linguistic multimodality representations. The RDMMFET model consists of three parts: problem and image representation (a), encoder (b), and output representation (c).

$$\begin{aligned}
\overline{Q}_{L1}^{k} &= \text{SelfAtt}\left(q_1^{k-1}, \ldots, q_n^{k-1}\right),\\
\widehat{Q}_{L1}^{k} &= \text{FeedForward}\left(\overline{Q}_{L1}^{k}\right),\\
\overline{Q}_{L2}^{k} &= \text{SelfAtt}\left(q_1^{k-1}, \ldots, q_n^{k-1}\right),\\
\widehat{Q}_{L2}^{k} &= \text{FeedForward}\left(\overline{Q}_{L2}^{k}\right),\\
\widehat{Q}_{L1}^{k} &= \widehat{Q}_{L1}^{k} + \widehat{Q}_{L2}^{k}.
\end{aligned} \tag{5}$$

3.2.3. Image Encoder. We transform the input image into an image feature V, and then transfer it to the image encoder. Unlike the encoder in the BERT model, which is only used for text coding, we apply the encoder to image coding. The structure of a single image encoder is the same as that of a single language encoder. We also add residual join and layer standardization after each sublayer. Our image encoder has N_I layers. The specific process is as follows:

$$\begin{aligned}
\overline{V}_I^k &= \text{SelfAtt}\left(v_1^{k-1}, \ldots, v_m^{k-1}\right),\\
\widehat{V}_I^k &= \text{FeedForward}\left(\overline{V}_I^k\right).
\end{aligned} \tag{6}$$

3.2.4. Multimodality Fusion Encoder. Each multimodality fusion layer of the multimodality fusion encoder is composed of a bidirectional fusion attention layer, two self-attention layers, and two feed-forward layers. Our multimodality fusion layer has N_F layers in total. In the kth multimodality fusion layer, we first transfer the language feature $Q_{N_{k-1}}$ and image feature $V_{N_{k-1}}$ from the $(k-1)$th multimodality fusion layer to a bidirectional fusion attention layer to generate Q_{N_k} and V_{N_k}. The sublayer contains two unidirectional fusion attention layers, which facilitate the fusion of image and language. The specific process is as follows:

$$\begin{aligned}
\overline{Q}_F^k &= \text{FusionAtt}_{L\longrightarrow I}\left(q_i^{k-1}, \left\{v_1^{k-1}, \ldots, v_m^{k-1}\right\}\right),\\
\overline{V}_F^k &= \text{FusionAtt}_{I\longrightarrow L}\left(v_i^{k-1}, \left\{q_1^{k-1}, \ldots, q_n^{k-1}\right\}\right).
\end{aligned} \tag{7}$$

The multimodality fusion layer enables both the interaction of problem and image information and the alignment of entities between the two modalities for better learning of multimodality representations. To better establish the internal connection of the mode, we transfer the output of the bidirectional fusion attention layer to the two self-attention layers. The specific process is as follows:

$$\begin{aligned}
\widetilde{Q}_F^k &= \text{SelfAtt}_{L\longrightarrow I}\left(q_i^k, \left\{q_1^k, \ldots, q_n^k\right\}\right),\\
\widetilde{V}_F^k &= \text{SelfAtt}_{I\longrightarrow L}\left(v_i^k, \left\{v_1^k, \ldots, v_m^k\right\}\right).
\end{aligned} \tag{8}$$

Finally, we transfer the output to the feed-forward layer to generate the output of the kth multimodality fusion layer. Like language and image encoder, we add residual join and layer standardization after each sublayer. The specific process is as follows:

$$\begin{aligned}
\widehat{Q}_F^k &= \text{FeedForward}\left(\widetilde{Q}_F^k\right),\\
\widehat{V}_F^k &= \text{FeedForward}\left(\widetilde{V}_F^k\right).
\end{aligned} \tag{9}$$

3.3. Output Representation. As shown in Figure 1, the output of our model is the output of the modal fusion encoder. In the problem representation, we add CLS at the beginning of each question, which is shown as the top yellow block of the problem feature in Figure 1. Finally, we use the feature vector corresponding to the special tag CLS in the language feature sequence as the answer to the VQA task.

4. Pretraining Strategy

We first introduce several pretraining tasks used in the model in Section 4.1 and then describe the pretraining data in Section 4.2. The framework of pretraining is shown in Figure 2.

4.1. Pretraining Method

4.1.1. Masked Language Model. As aforementioned in the part of the branches, we first mask the input problem with a 15% probability using a special marker MSAK and then train the model to predict the masked word cases through other word cases. Different from previous masked language models, our masked language model can not only predict masked words from nonmasked words but also predict

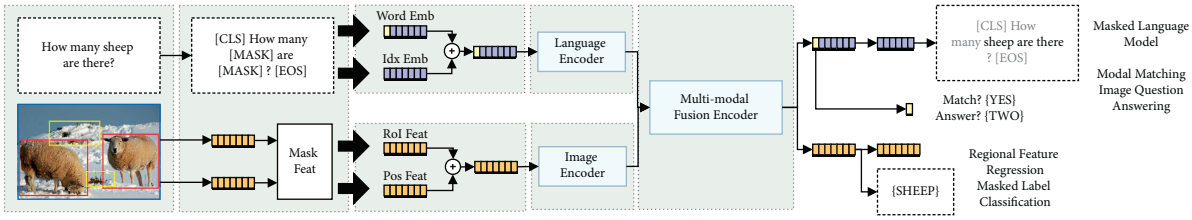


Figure 2: Pretraining strategy of the RDMMFET model. The strategy consists of three parts: masked language model (a), masked image model (b), and multimodality fusion task (c).

masked words from image features using a multimodality fusion model. We masked the words "sheep" and "there" in the question in Figure 2. It is hard to predict "sheep" from other word examples, but it is easy to use image features. Therefore, our masked language model can make the model better understand the relationship between language and language, language and image.

4.1.2. Masked Image Model. We also mask the image features with a probability of 15%. Our model can use unmasked image features to predict the masked image features or use the word examples in the problem to predict the masked image features. We propose two different pretraining tasks: regional feature regression and masked label classification. Regional feature regression: for each region, there is a high-dimensional vector. We want the output vector of the model to be as close as possible to the feature vector of the masked area; therefore, we use L2 loss to make the distance between the two vectors as small as possible so that the two features are more similar. The L2 loss function is also called least square error (LSE), which is to minimize the sum of squares of the difference between the target value and the estimated value. Masked label classification: after each region gets the feature vector, R-CNN will predict a label to classify the masked region. We use the model to predict the masked areas so that the model can learn the classification of each masked area.

4.1.3. Multimodality Fusion Task. For the pretraining task of multimodality fusion, we also propose two different tasks: modal matching and image question answering. Modal matching: we replace the input problem with a 50% probability, that is, replace the original problem with a problem that does not match the image and then use the model to predict whether the current problem matches the image. This task is used to learn the global information correspondence between image and text. Image question answering: we increase the amount of pretraining data by adding other image question answer data sets, and classify the image data pairs using the answers of questions as labels. Then, we use modal matching to pretrain the model. When the model predicts that the current question matches the image, the model needs to predict again whether the answer to the question is consistent with the label. Multimodality fusion task can better learn the in-depth relationship between images and problems, making the connection between images and problems closer.

4.2. Pretraining Data Sets. To better establish the relationship between language and vision, improve the model reasoning ability, we collect five different data sets as our pretraining data sets. These data sets are MS COCO [35], visual genome (VG) [36], VQA V2.0, GQA [37], and VG-QA [38].

To get a data set without any test set, we combine the original training and validation segmentation of MS COCO and VG. To avoid the overlap of the images in the training set and the images in the validation set and test set, we delete duplicate images. Then preprocess the large-scale data set to create alignment problems and image pairs. This provides us with 180,000 different images on the 9.18 m problem and image pairs of large alignment problems and image data sets. The number of images and questions for each data set is shown in Table 1. As the MS COCO verification set is too large, we sample a group of 5K images from the MS COCO verification set as the minimum verification set. The rest images in the MS COCO training and verification set as well as the training set and verification set in VG are used as the training set of the pretraining task.

5. Experiments

Our experiment is based on the VQA v2.0 data set. In this section, we evaluate the RDMMFET model through experiments to prove its effectiveness:

In Section 5.1, the benchmark data set VQA v2.0 used in the experiment is introduced.

In Section 5.2, the pretraining task and VQA downstream task are introduced in detail.

In Section 5.3, we have conducted extensive ablation studies on the parameters used in the experimental details.

In Section 5.4, we compare and analyze the RDMMFET model and some of the latest VQA models. The accuracy of the RDMMFET model is higher than other models, which proves the effectiveness of the RDMMFET model.

In Section 5.5, we increase our understanding of the RDMMFET model by visualizing the RDMMFET model.

5.1. Data Sets. In the experiment, we select the VQA v2.0 data set for training and verification. The VQA v2.0 data set is based on the MS COCO data set. VQA v2.0 contains 204,721 pictures in the COCO data set and 1,105,904 questions about these pictures. Among them, each picture corresponds to a different number of questions, with an average of 5.4 questions per picture. We carefully segment

the data set above to ensure that all test data did not involve any pretraining or fine-tuning. The data set is divided into training, validation, and test sets, with the proportions of 40%, 20%, and 40% respectively. The types of all questions in the VQA v2.0 data set are yes/no, quantity, and others. The VQA v2.0 is balanced enough to deal with the possibility of accuracy improvement due to overfitting. In addition, the VQA v2.0 collects more samples than VQA v1.0.

5.2. Experiment Setup

5.2.1. Pretraining Program. The input questions are transformed into word vectors by WordPiece tokenizer in BERT, and the input images were detected by the faster R-CNN with 36 objects. In the encoder layer, we set the number of layers of the dense language encoder, the image encoder and the multimodality fusion encoder $N_L N_I$ and N_F to 9, 4, and 5, respectively. The size of the hidden layer in each encoder is 768, and the potential dimension d of multihead attention in each encoder to 512. The multiattention contains 8 heads, and the potential dimension of each head is up to 64.

We set all training parameters in each encoder, and randomly initialize or set the model parameters to zero. RDMMFET pretrains through multiple pretraining tasks, so it involves multiple losses. We add these losses with the same weight to get the final loss. We take Adam [39] as the optimizer and adopt the linear decay learning rate plan. The peak learning rate is $1e^{-4}$. We set the training epoch to 20 and the batch size to 256. Since the image question answering pretraining task converges fast, we set the task epoch to 10.

5.2.2. Fine-Tuning. We fine-tune the model to adapt to the VQA task. In the encoder layer, we set the depth DL of the text encoder to 2. In the multimodality fusion layer, the final output is the top eigenvector of the output of the RDMMFET model. In the RDMMFET model, the basic learning rate is set to $5e^{-5}$, and the dropout is set to 0.1 to prevent overfitting. The RDMMFET model has four epochs, where the batch size is 32.

5.3. Ablation Experiment. In the VQA v2.0 data set, we conducted some extensive ablation experiments on the RDMMFET model. In all ablation experiments, all features and other parameters are the same except for the studied parameters.

5.3.1. Iterations. To explore the influence of iteration number on the accuracy of the RDMMFET model, we set the iteration number epoch $\in \{3, 4, 5, 6\}$. As shown in Table 2, as the number of iterations increases, the overall accuracy of VQA improves. When the number of stacked layers is 5, the performance of VQA will start to decline rapidly. Therefore, when the number of iterations is 4, the accuracy of the model is the highest, and our optimal model sets the number of iterations to 4.

5.3.2. Layers of the Encoder. To explore the influence of the number of encoder layers on the accuracy of the RDMMFET model, we compare the text encoder, image encoder, and multimodality fusion encoder with different layers, respectively. In the experiment, we find that VQA tasks benefit from a larger number of layers. We set the number of layers $N_L \in \{8, 9, 10\}$ for the dense language encoder, the number of layers $N_I \in \{3, 4, 5\}$ for the image encoder and the number of layers $N_F \in \{4, 5, 6\}$ for the multimodality fusion encoder, respectively. As shown in Table 2, when the number of layers of the dense language encoder, image encoder, and multimodality fusion encoder $N_L N_I$ and N_F are 9, 4, and 5, respectively, the accuracy of the RDMMFET model is the highest. Therefore, we set the number of layers of dense language encoder, image encoder, and multimodality fusion encoder $N_L N_I$ and N_F to 9, 4, and 5.

5.4. Comparison with the Latest Models. We compare the RDMMFET model with the latest models in the same experimental setup. In Table 3, DFAF, MCAN, and MUAN [40] are the best models of deep common attention without pretraining. Based on cross-modal self-attention and cross-modal co-attention, DFAF proposes a multimodality feature fusion method. MCAN proposes a deep modular common attention network, which is composed of deeply cascaded modular common attention (MCA) layers. MUAN proposes a general "unified attention" model, which simultaneously captures the intramode and intermode interactions of multimodal features and outputs its corresponding participation representations.

Unlike other single-stream BERT models, ViLBERT creates a multimodality dual-stream model. ViLBERT first inputs image and question information into two identical single-stream models, and then these streams interact through the transformer layer in the common attention. Then, ViLBERT performs pretraining on a large data set to learn the basics of vision. VisualBERT proposes a transformer layer that can automatically align the elements of input text with the regions in the associated input image implicitly. Then VisualBERT further puts forward two basic visual language models for image subtitle data pretraining. VL-BERT uses a simple and powerful transformer model as the backbone network and extends its input to a multimodality form including both visual and linguistic inputs. VL-BERT is pretrained on large-scale concept subtitle data sets and pure text corpus, which is suitable for most visual language downstream tasks. LXMERT proposed a model with cross-modality encoder as the core and five different pretraining tasks.

We tested our model on Test-dev and Test-std of the VQA v2.0 data set. It can be seen from Table 3 that most of the pretraining models have higher accuracy than the nonpretraining models in Test-dev and Test-std, which proves the necessity of pretraining models. The RDMMFET model is 1.68%, 1.60%, 0.38%, and 0.06% higher than ViLBERT, VisualBert, VL-BERT(large), and LXMERT in Test-std, respectively. Using the dense language encoder, the pretraining model RDMMFET based on modal fusion

Table 1: Statistics of data sets used for pretraining.

Image (K)	Questions					
	MS COCO (K) [33]	VG (M) [34]	VQA v2.0 (K)	GQA (M) [35]	VG-QA (M) [36]	All (M)
180	617	5.39	658	1.07	1.44	9.18

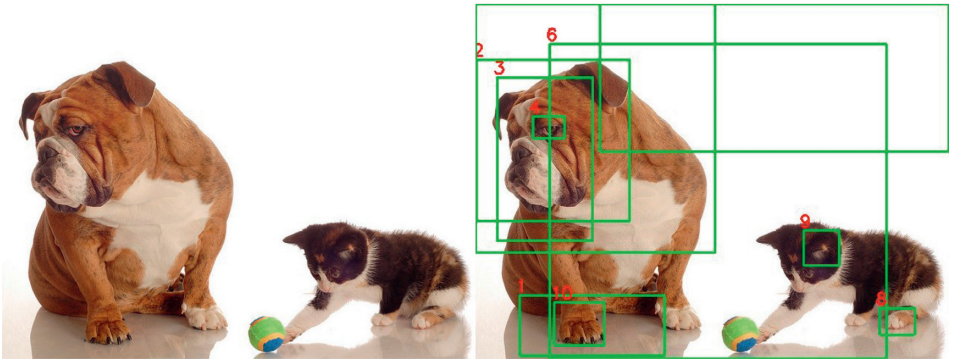
Table 2: Ablation studies on VQA v2.0 test-dev with iterations and layers of each encoder.

Module	Setting	Accuracy
Number of iterations	Epoch = 3	72.46
	Epoch = 4	**72.59**
	Epoch = 5	72.46
	Epoch = 6	72.22
Number of encoder layers	$N_L, N_I, N_F = 8, 4, 5$	72.37
	$N_L, N_I, N_F = 10, 4, 5$	72.35
	$N_L, N_I, N_F = 9, 3, 5$	72.47
	$N_L, N_I, N_F = 9, 5, 5$	72.44
	$N_L, N_I, N_F = 9, 4, 4$	72.25
	$N_L, N_I, N_F = 9, 4, 6$	72.37
	$\mathbf{N_L}, \mathbf{N_I}, \mathbf{N_F} = 9, 4, 5$	**72.59**

Table 3: Comparison with the latest models on the VQA v2.0 data set.

Label	Method	Test-dev	Test-std
No pretraining	DFAF [8]	70.22	70.34
	MCAN [9]	70.63	70.90
	MUAN [38]	70.82	71.10
Pretraining	ViLBERT [23]	70.55	70.92
	VisualBert [27]	70.80	71.00
	VL-BERT(base) [28]	71.16	-
	VL-BERT(large) [28]	71.79	72.22
	LXMERT [24]	72.42	72.54
	RDMMFET (ours)	**72.59**	**72.67**

Q: How many buses are there? Q: How many buses are there?



Q: What is the dog doing? Q: What is the dog doing?

Figure 3: Visualization of the last layer of RDMMFET's multimodal fusion encoder.

achieves the highest accuracy. The RDMMFET model can achieve a good result, which proves that the RDMMFET model has good performance.

5.5. Visualization. As shown in Figure 3, we visualize the attention of the last layer in RDMMFET's multimodality encoder to reveal the relationship between the problem and the image. For the problem text, we highlight the words with high attention weight and bold the keywords. There are a total of 10 boxes on each image. Each box represents the concentration area of the last layer of an RDMMFET multimodal encoder. The order is determined by the number in the upper left corner of the box. The smaller the number, the more focused the model is on this area. For the image, we select the top 10 objects with the highest attention score and show them in the boxes in Figure 3. The attention ranking is consistent with the label in the upper left corner of the box. We find that the attention of the multimodality fusion encoder is mainly focused on nouns and pronouns (as highlighted in Figure 3) because the number of nouns and pronouns is the largest in visual and linguistic tasks. Most of the attention boxes in Figure 3 focus on the keywords "bus" and "dog". In the upper part of Figure 3, most of the attention of the question text is on the "bus", and in the corresponding picture, most of the attention boxes is also on the "bus". The same is true for the lower part of Figure 3. It can be seen that the keywords in the question are closely related to the box with a high attention score in the image, and there is a good correlation between the keywords and the related image regions. Visualization can help us further understand the model and improve the model in the future.

6. Conclusion

In this paper, we propose a model RDMMFET based on the dense language encoder and the multimodality fusion encoder, which can be used to learn the association between problems and vision. The RDMMFET model consists of the dense language encoder, image encoder, and multimodality fusion encoder. We use different pretraining tasks to pretrain RDMMFET on large-scale data sets and then fine-tune the model for VQA tasks. Finally, we evaluate the RDMMFET model on the VQA v2.0 data set and get good results. Although our model has achieved the latest results, there is still a big gap compared with human reasoning ability. Therefore, in future work, we will design a more perfect multimodality fusion model and pretraining task to improve the understanding of images and texts.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant nos. 61672338 and 61873160).

References

[1] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, Seattle, WA, USA, June 2020.

[2] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[3] T. Wolf, J. Chaumond, L. Debut et al., "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, November 2020.

[4] J. Qiu, B. Wang, and C. Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism," *PLoS One*, vol. 15, no. 1, Article ID e0227222, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[6] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," arXiv preprint arXiv:1606.01847, 2016.

[7] J. H. Kim, J. Jun, and B. T. Zhang, "Bilinear attention networks," arXiv preprint arXiv:1805.07932, 2018.

[8] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6639–6648, Long Beach, CA, USA, June 2019.

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

[10] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, Long Beach, CA, USA, June 2019.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[12] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: a joint model for video and language representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7464–7473, Long Beach, CA, USA, June 2019.

[13] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," arXiv preprint arXiv:1412.6632, 2014.

[14] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 203–212, Las Vegas, NV, USA, June 2016.

[15] D. Han, S. Zhou, K. C. Li, and R. F. de Mello, "Cross-modality co-attention networks for visual question answering," *Soft Computing*, vol. 25, no. 7, pp. 5411–5421, 2021.

[16] J. H. Kim, K. W. On, W. Lim, J. Kim, J. W. Ha, and B. T. Zhang, "Hadamard product for low-rank bilinear pooling," arXiv preprint arXiv:1610.04325, 2016.

[17] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1821–1830, Venice, Italy, October 2017.

[18] 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.

[19] W. Kim, B. Son, and I. Kim, "Vilt: vision-and-language transformer without convolution or region supervision," arXiv preprint arXiv:2102.03334, 2021.

[20] A. Nagrani, S. Yang, and A. Arnab, "Attention bottlenecks for multimodal fusion," arXiv preprint arXiv:2107.00135, 2021.

[21] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," arXiv preprint arXiv:1511.01432, 2015.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[23] M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," 2018. arXiv preprint arXiv: 1802.05365.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[25] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," arXiv preprint arXiv:1908.02265, 2019.

[26] H. Tan and M. Bansal, "Lxmert: learning cross-modality encoder representations from transformers," 2019. arXiv preprint arXiv:1908.07490.

[27] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," 2019. arXiv preprint arXiv:1908.05054.

[28] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11336–11344, 2020.

[29] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, and K. W. Chang, "Visualbert: a simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.

[30] W. Su, X. Zhu, Y. Cao et al., "Vl-bert: pre-training of generic visual-linguistic representations," arXiv preprint arXiv: 1908.08530, 2019.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," arXiv preprint arXiv:1506.01497, 2015.

[32] 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 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.

[33] Z. Tan, M. Wang, and J. Xie, "Deep semantic role labeling with self-attention," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[34] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned," 2019. arXiv preprint arXiv:1905.09418.

[35] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Proceedings of European Conference on Computer Vision*, pp. 740–755, Springer, Zurich, Switzerland, September 2014.

[36] 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.

[37] D. A. Hudson and C. D. Manning, "Gqa: a new dataset for compositional question answering over real-world images," vol. 2, no. 3, p. 11, 2019 arXiv preprint arXiv:1902.09506.

[38] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: grounded question answering in images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4995–5004, Las Vegas, NV, USA, June 2016.

[39] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[40] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, "Multimodal unified attention networks for vision-and-language interactions," arXiv preprint arXiv:1908.04107, 2019.