MSAM: Deep Semantic Interaction Network for Visual Question Answering

Fan Wang, Bin Wang, Fuyong Xu, Jiaxin Li, and Peiyu Liu(✉)

Shandong Normal University, Jinan 250358, China
liupy@sdnu.edu.cn

Abstract. In Visual Question Answering (VQA) task, extracting semantic information from multimodalities and effectively utilizing this information for interaction is crucial. Existing VQA methods mostly focus on attention mechanism to reason about answers, but do not fully utilize the semantic information of modalities. Furthermore, the question and the image relation description through attention mechanism may cover some conflicting information, which weakens multi-modal semantic information relevance. Based on the above issues, this paper proposes a Multi-layer Semantics Awareness Model (MSAM) to fill the lack of multi-modal semantic understanding. We design a Bi-affine space projection method to construct multi-modal semantic space to effectively understand modal features at the semantic level. Then, we propose to utilize contrastive learning to achieve semantic alignment, which effectively brings modalities with the same semantics closer together and improves multi-modal information relevance. We conduct extensive experiments on the VQA2.0 dataset, and our model boosts the metrics even further compared to the baseline, improving the performance of the VQA task.

Keywords: Visual Question and Answering · Contrastive Learning · Semantic Alignment · Semantic Information

1 Introduction

The goal of the Visual Question Answering (VQA) task is to predict answers based on given questions and relevant images. There are two main variants of VQA, Free Form Open End (FFOE) and Multiple Choice (MC). In FFOE, the answers are free response to given image-question input pairs, whereas in MC, the answers are chosen from a predefined list of ground truths. In both cases, extracting meaningful features from both images and questions plays a key role. In addition, semantic features mapped from images and questions can strongly influence the results [1]. Most of the existing VQA task solutions rely on visual relations [2,3], attention mechanism [4,5] and external knowledge [6].

H. Gao et al. (Eds.): CollaborateCom 2023, LNICST 562, pp. 39–56, 2024.
https://doi.org/10.1007/978-3-031-54528-3_3

VQA modeling process often involves feature extraction, cross-modal interaction, cross-modal feature fusion, and answer prediction. Feature extraction involves extracting visual and textual features from raw data. Cross-modal interaction mechanisms (such as attention mechanism) enable the model to reason about the relations between visual and textual features. Cross-modal feature fusion aims to integrate the complementary information from questions and images, enhancing the overall representation of the input. Finally, after obtaining the fused features from questions and images, the VQA task predicts techniques such as classification or generation to predict the answers.

Most of the current VQA methods extract rich features from questions and images through attention mechanism. Yang et al. [7] propose a method to obtain features in key regions of images by stacking attention networks, which aims to solve the problem of too extensive feature regions. Based on this, Nguyen et al. [8] design a co-attention learning method that alternates image attention and question attention to obtain fine-grained key feature representations. Yu et al. [9] introduce the Self-Attention and the Guided-Attention, which result in more detailed representations of the question keywords and the image key regions, facilitating multi-modal interaction. Thus attention mechanism plays a crucial role in VQA task by allowing the model to focus on relevant parts of questions and images. By attending to specific regions, the model can gather more informative features. There are also some VQA approaches that focus on the importance of multi-modal semantic information [10] fusion. Nguyen et al. [11] design fusion for multi-modal at different semantic levels. Chen et al. [12] emphasize the importance of contextual semantic information in the fusion process. Tu et al. [13] propose internal contextual information to enhance memory in an attentional model.

Although there are many approaches to VQA, VQA is still challenging: 1). While attention mechanism has been effective in improving the performance of VQA task, they have limitations in terms of fully understanding the semantic multi-modal information. For example, the question is "What is this person doing?", and the image shows a person standing in a kitchen. Although attention mechanism can focus on the image region of the person, it cannot understand the specific activity of the person in the kitchen, such as whether he is cooking, washing dishes, or cleaning. 2). Existing semantic fusion methods usually focus on combining visual and textual features without explicitly resolving conflicts. This may result in the fusion process ignoring conflicting information and failing to effectively capture the relevance of multi-modal information.

To address the above issues, we propose the MSAM model, in which the self-private semantic space and the co-public semantic space are constructed to further understand multi-modal from the semantic level. Then, we utilize contrastive learning to deeply mine the multi-modal semantic space. By constructing positive pairs (aligned question-image pairs from co-public semantic space) and negative pairs (question-image pairs from self-private semantic space) of multi-modal samples, the method can stretch positive pairs in the semantic space closer together and negative pairs farther apart. In the multi-modal fusion stage,

Self-Attention(SA) and Guided-Attention(GA) [9] are used to achieve multi-modal interaction.

Our contributions are summarized as follows:

- To bridge the difficulty of attention mechanism in understanding multi-modal features at the semantic level, we design the MSAM model, which constructs the modal semantic space by Bi-affine projection to understand modal features at the semantic level.
- To alleviate the problem of conflicting information affecting the relevance of multi-modal semantic information, we employ contrastive learning to achieve semantic alignment. The distance between multi-modal features with identical semantics is further narrowed, thus enhancing the multi-modal semantic relevance.
- Our method is validated on the VQA2.0 dataset and the results confirm that our method outperforms previous methods.

The rest of the paper is organized as follows: Sect. 2 presents related work; Sect. 3 describes the paper methodology; Sect. 4 presents the experiments of our method on the VQA2.0 dataset; Sect. 5 presents conclusion.

2 Related Work

2.1 Multi-modal Feature Representation for VQA

The basis of the VQA task is to extract visual features of images and textual features of questions accurately and effectively. In early VQA methods, the Visual Geometry Group (VGG) [15] network is commonly applied to extract visual features. With the ResNet (Residual Network) network proposed by Kaiming et al. [16], the researchers gradually shift to the ResNet network that performs better than VGG in visual feature extraction. Recently, the bottom-up attention network [17] derives from Faster R-CNN [18] outperforms better than ResNet. By using the bottom-up attention network, the VQA model can focus on relevant regions in images, allowing it to better understand the visual context of questions. For the extraction of question features, the Long Short-Term Memory network (LSTM) [19] and the Gated Recurrent Unit (GRU) network [20] are commonly applied. Additionally, pre-training methods on Global Vectors for Word Representation (GLoVe) [21] or Bidirectional Encoder Representations from Transformers (Bert) [22] are also applied to obtain better features. Usually, VQA methods combine GLoVe and LSTM for the question features extraction. Multi-modal feature representation is crucial for improving the performance of VQA task. VQA task involves understanding both the visual content of images and the textual information conveyed by questions. By effectively representing multi-modal features, VQA models can better capture the relations between images and questions, leading to accurate answers.

2.2 Multi-modal Semantic Information for VQA

VQA aims to answer questions about images. To do so effectively, it needs to understand the semantics of both questions and images. In this regard, researchers present a number of approaches. The first one is the attention based approach, Peng et al. [23] propose a new Cross Fusion Network (CF-Net) for fast and efficient extraction of multi-scale semantic information. Tian et al. [24] introduce a Multi-level Semantic Context Information (MSCI) network with an overall symmetrical structure. The interconnections between three different semantic layers are utilized to extract contextual information between multi-modalities. Secondly, based on graph structure methods, Li et al. [25] design a graph embedding method that introduces pointwise mutual information to compute the semantic similarity between nodes. The correlation between the question and the image is measured by calculating the similarity. Adhikari et al. [26] present a new semantic fusion network that fuses the semantic information of questions and images by constructing a semantic graph. The above works focus on the importance of semantic information in the multi-modal fusion process, but ignore the fact that unimodal plays a complementary role in multi-modal semantic fusion. The individual modals themselves carry specific and important semantic information that can contribute to the overall understanding and answering of questions. Therefore, we propose to construct semantic space by Bi-affine space projection, constructing self-private semantic space for each modality separately and co-public semantic space for multi-modality at the same time. By constructing self-private semantic space for each modality, we can capture the unique semantic information of each modality. By building co-public semantic space, we can capture the relation between multi-modal semantic.

2.3 Multi-modal Semantic Alignment for VQA

Semantic alignment refers to the process of aligning the semantic information from multi-modal, to facilitate understanding and interaction between them. In recent years, there has been significant research progress in addressing the challenge of multi-modal semantic alignment in the VQA task. Li et al. [27] introduce a dynamic interaction network that combines Self-Attention and Cross-modal Attention. The ability to dynamically explore the various semantic depths of these different modal representations and fuse them together at a matching semantic level. Bao et al. [28] propose a method is designed for multi-task learning, i.e., learning shared semantic representations of different modals and introducing constraints to narrow the semantic distance between different modals. Based on previous works, we design a method to deeply mine the semantic space utilizing contrastive learning [14]. This approach enables semantic alignment, bridges the semantic distance that exists between modalities, and improves multi-modal semantic information relevance. Semantic Information relevance refers to the degree of semantic matching due to representational differences between heterogeneous modal features. We compute the similarity of image-question pairs and select sample pairs with high similarity as inputs for

subsequent modal fusion, so that semantically similar multi-modal feature representations are more compact, while semantically different ones are mutually exclusive, thus improves the information relevance of the multi-modal.

2.4 Multi-modal Feature Fusion for VQA

Multi-modal feature fusion is an indispensable process in VQA task that combines information from different modalities to generate better representations for answering questions. Early stages utilize simple joint modal representations [10] for multi-modal feature fusion to capture high-level interactions between images and questions. Fukui et al. [29] apply MCB, a bilinear ensemble-based feature fusion method for VQA. This method improves the accuracy of VQA task by bilinear interaction of questions and answers. Kim et al. [30] propose the MLB method, which has a performance equal to MCB but with fewer weighting parameters. Nowadays, most of the multi-modal fusion in VQA task is done in an attention-based manner. Yu et al. [31] propose the MHF method which is superior to the above methods. MFH utilizes the correlation between multi-modal features to achieve more effective multi-modal feature fusion. Yang et al. [7] propose a stacked attention network based on simple fusion to learn attention to image regions through multiple iterations. Chowdhury et al. [8] propose a co-attentive learning method that alternates image attention and question attention based on attention to unimodal. Yu et al. [9] reduce co-attention to two steps, one for self-attentive learning of the question and the other for question-guided attention learning of the image. Based on the above work, we chose to introduce the fusion method of Modular Co-Attention (MCA) [9]. Before fusion, we narrow the semantic distance between multi-modal features through modal semantic alignment, which improves the information relevance between modal features. This approach addresses a limitation of the fusion model mentioned in MCA [9], which is the lack of ability to infer correlations between the question words and the image regions.

Unlike other models that focus on enriching images and questions information. In our work, we consider interacting with multi-modal from semantic perspective. Narrowing the semantic distance between multi-modal can lead to a more compact relations between multi-modal. Our model provides better results compared to other methods while maintaining a reasonable computational cost for the network.

3 Methodology

3.1 Overview

Figure 1 illustrates our MSAM model. Our model first maps the image features obtained after Faster-RCNN (Sect. 3.2) and the question features obtained after GLoVe+LSTM (Sect. 3.2) by means of a Bi-affine space projection method. The

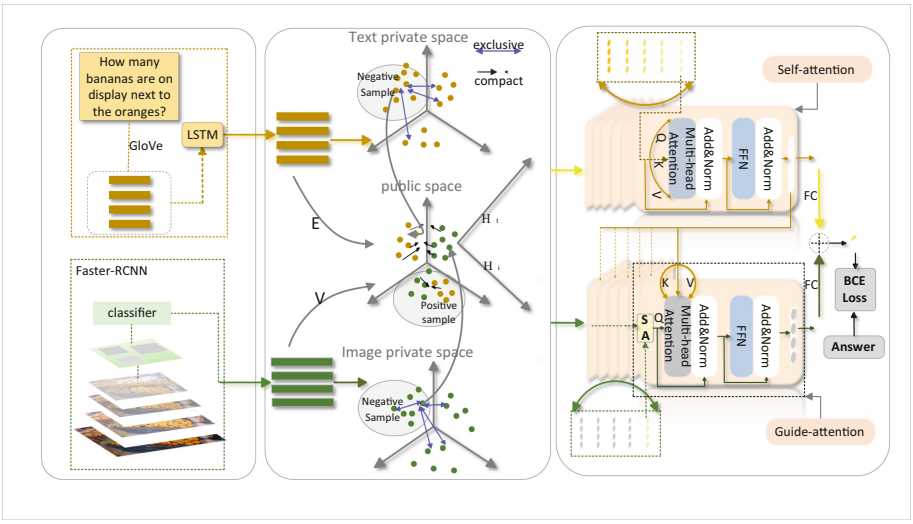

Fig. 1. MSAM model. On the left side of the MSAM model is multi-modal feature extraction. In the middle is semantic space building and multi-modal semantic alignment through contrastive learning. The right side is multi-modal feature fusion.

projections are made to self-private semantic spaces and simultaneously to co-public semantic space (Sect. 3.3). After semantic level alignment (Sect. 3.3), question features and image features with similar semantics are obtained. Finally, the obtained question features and image features with similar semantics are fused (Sect. 3.4) and the fused features are fed to the multi-label classifier to predict the correct answer (Sect. 3.4).

3.2 Images and Questions Representation

Image Representation. Following [32], the input image is represented as a set of image region features. These features are extracted from a Faster R-CNN [33] model pre-trained on the Visual Genome dataset [34]. Give an image I, the image features of I are represented as $V \in \mathbb{R}^{k \times 2048}$. where $k \in [10, 100]$ is the number of object regions. For the i-th word, it is represented as a feature $v_i \in \mathbb{R}^{2048}$.

Question Representation. Firstly, we give a question Q that is tokenized into words. Then, each word of the question is transformed into a vector, and pre-trained on a large-scare corpus by using the 300-dimensional GLoVe word embeddings [35] to get the final size of words $u \times 300$, where $u \in [1, 14]$ is the number of words in question. Finally, the word embeddings are passed through a one-layer and 512-dimensional LSTM network to obtain the question features $E \in \mathbb{R}^{u \times 512}$. This LSTM layer takes as input a sequence of embedded vectors

and processes the sequence step-by-step through time steps. For the j-th word, it is represented as a feature $e_j \in \mathbb{R}^{512}$.

In practice, to handle the different number of object regions k and the variable number of words u, following [9], both of V and E are filled to their maximum sizes (i.e., k = 100, u = 14) by zero-padding. Especially, we conduct a linear transformation of V to make its dimension consistent with the question features.

3.3 Contrastive Learning Deep Mining of Multi-modal Semantic Space

Bi-affine Space Projection. The Bi-affine space projection method can be divided into two steps, first the affine space followed by the affine transformation also known as projection. Affine transformations differ from linear transformations in that they not only maintain the linear nature of vector spaces, but also preserve translation invariance. Based on comprehensive consideration, we choose to obtain the semantic space by affine transformation.

We first store question features E and image features V extracted in Sect. 3.2 into the Bi-affine space. Due to the different modal features of the inputs, they are mapped to different semantic spaces after Bi-affine transformation, which are called question private semantic space and image private semantic space, respectively. Since this space is built on a semantic level, a representation of the distribution of the different modal features in their respective private semantic spaces is obtained at this point. Complementary evidence for subsequent multi-modal semantic fusion is provided. In the subsequent operations we will seize contrastive learning to achieve multi-modal alignment at the semantic level, and therefore use modal features in the private semantic space as negative samples. The unimodal *Hidden representation* obtained in the private semantic space is denoted as H. The output of the text private semantic space is denoted by H_t and the output of the image private semantic space is represented by H_i.

The question features E and image features V extracted in Sect. 3.2 are stored into the affine space. After affine projected into the same semantic space, which is called the co-public semantic space, we obtain representations of different modalities in the same semantic space. Since the spatial construction is de-constructed based on semantic information, there is a partial aggregation of different modal features. The aggregation of different modalities realizes a multi-modal fusion representation at the semantic level, which reduces some noise. However, despite the similarity of certain problem features and image features at some semantic levels, due to the semantic gaps between heterogeneous modalities, these modal features are close in the semantic space but still at a certain distance. Therefore it is difficult to provide a good representation of the semantic information correlation of multi-modal. In the previous section, the modal features in the private space are used as negative samples, and we use the modal features in the co-public semantic space as positive samples at this point. The modal *Hidden representation* obtained in the co-public semantic space is denoted as H^P. The question output features in the public semantic space are denoted by H_T^P, and the image output features are represented by H_I^P.

Contrastive Learning for Semantic Alignment. Contrastive learning is called self-supervised learning [36] in some papers and unsupervised learning [37] in others, and self-supervised learning is a form of unsupervised learning. Through contrastive learning, we compute the similarity of the question-image pairs in the batch and obtain sample pairs with the highest possible similarity. Multi-modal semantic alignment is achieved on this basis, shrinking the modal distance that exists. Therefore, multi-modal features with similar semantics lead to more compact modalities, while ones with different semantics are mutually exclusive. For a batch with N training samples, the unsupervised comparison loss for all samples in the batch can be expressed as follows:

$$L = \sum_{m \in M} -\frac{1}{|P(m)|} \sum_{p \in P(m)} \log \frac{\exp(H_m \cdot H_p/\tau)}{\sum_{c \in A(m)} \exp(H_m \cdot H_c/\tau)} \tag{1}$$

where $m \in M = \{1, 2, ..., \mathrm{N}\}$ indicates the index of the samples, $\tau \in \mathbb{R}^+$ denotes the temperature coefficients introduced to control the distance between instances, $P(m) = I_{j=m} - \{m\}$ represents samples with the same category as m while excluding itself, $A(m) = M - \{m, \mathrm{N} + m\}$ indicates samples in a batch except itself.

3.4 Cross-Modal Fusion and Answer Prediction

Cross-Modal Deep Interaction Networks. We are inspired by the previous work of [9] to introduce Self-Attention (SA), Guide-Attention (GA), and their combined Modular Co-Attention (MCA). The SA module consists of a multi-head attention and a feedforward layer, where the multi-head attention is introduced to further improve the representation of the participating features. A set of input features $H_T^P = \left[H_{T1}^P; H_{T2}^P; ...; H_{Tm}^P\right]$ as samples, the multi-head attention learns the pairwise relations between the paired samples $\left(H_{T_i}^P, H_{T_j}^P\right)$ in the samples H_T^P, and outputs the participating output features by weighting the sum of all instances in the H_T^P. Multi-headed attention consists of h parallel "heads", where each head corresponds to an independently scaled dot-product attention function, and the attended output features f is given by:

$$f = \text{softmax}\left(\frac{QK}{\sqrt{d}}\right) V \tag{2}$$

$$Mf = [\text{head}_1, \text{head}_2, ..., \text{head}_h]\, \mathbf{W}^o \tag{3}$$

$$\text{head}_j = f\left(Q\mathbf{W}_j^Q, K\mathbf{W}_j^K, V\mathbf{W}_j^V\right) \tag{4}$$

where $\mathbf{W}_j^Q$, $\mathbf{W}_j^K$, $\mathbf{W}_j^V \in \mathbb{R}^{d \times d_h}$ are the projection matrices for the head, and $\mathbf{W}^o \in \mathbb{R}^{h \times d_h \times d}$. d_h is the dimensionality of the output features from each head.

The input of scaled dot-product attention consists of Queries and Keys with dimension d_{key}, and Values with dimension d_{value}. For simplicity, d_{key} and d_{value} are usually set to the same number d. We are given a Query $Q \in \mathbb{R}^{1\times d}$, n Key-Value pairs(packed into a key matrix $K \in \mathbb{R}^{n\times d}$ and a value matrix $V \in \mathbb{R}^{n\times d}$), $f \in \mathbb{R}^{1\times d}$.

The feedforward layer applies the output characteristics of multi-attention, and further transforms them through two fully-connected layers with ReLU activation and dropout (FC(·)-ReLU-Dropout(0.1)-FC(·)). Moreover, residual connection followed by layer normalization [38] is applied to the outputs of the two layers to facilitate optimization.

The GA module applies two sets of input features H_T^P and $H_I^P = \left[H_{I1}^P; H_{I2}^P; ...; H_{Im}^P\right]$ as samples, where H_T^P guides the attention learning for H_I^P. Note that the shapes of H_T^P and H_I^P are flexible, so they can be applied to represent the features for different modalities. The GA units simulates the pairwise relations between the paired samples $\left(H_{T_i}^P, H_{I_j}^P\right)$ from H_T^P and H_I^P, respectively.

Based on the two attention modules mentioned above, they are combined to obtain MCA layer, which can be utilized to process multi-modal features. Any MCA layer is deeply cascaded [39], and the output of the previous MCA layer is utilized as the input to the next MCA layer. This means that the number of input characteristics is equal to the number of output characteristics without reducing any instances.

$$\left[H_T^{\text{out}(L)}, H_I^{\text{out}(L)}\right] = \text{MCA}^{(L-1)}\left[H_T^{\text{out}(L-1)}, H_I^{\text{out}(L-1)}\right] \tag{5}$$

We take the image features H_I^P and question features H_T^P obtained by contrastive learning as input, and pass the input features through a deep co-attention model composed of L-MCA layer to perform deep concentrative learning (denoted by MCA(1), MCA(2)... MCA(L)). Denoting the input features for MCA(L) as $H_T^{\text{out}(L-1)}$ and $H_I^{\text{out}(L-1)}$ respectively, their output features are denoted by $H_T^{\text{out}(L)}$ and $H_I^{\text{out}(L)}$, which are further fed to the MCA(L+1) as its inputs in a recursive manner.

Answering Prediction. After passing through the six layers of SA module and MCA module, the question features $H_T^{\text{fina-out}(L)}$ and image features $H_I^{\text{fina-out}(L)}$ already contain rich and accurate informations, so it is also important to fuse the features to predict the correct answer. In this process, a two-layer MLP(FC(·)-ReLU-Dropout(0.1)-FC(·)) structure is introduced to obtain the participation features, and the formula is as follows (seizing the image features as an example):

$$a = \text{softmax}\left(\text{MLP}\left(H_I^{\text{out}(L)}\right)\right) \tag{6}$$

$$H_I^{\text{fina-out}(L)} = \sum_{i=1}^{m} a_i H_I^{\text{out}(L)} \tag{7}$$

where $a = [a_1, a_2, ..., a_m] \in \mathbb{R}^m$ are the learned attention weights.

Using the computed $H_T^{\text{fina-out}}$ and $H_I^{\text{fina-out}}$, we apply the linear multi-modal fusion function as follows:

$$\mathrm{Z} = \text{LayerNorm}\left(\mathbf{W}_t^T H_T^{\text{fina-out}(L)} + \mathbf{W}_i^T H_I^{\text{fina-out}(L)}\right) \tag{8}$$

where $\mathbf{W}_t$,$\mathbf{W}_i \in \mathbb{R}^{d \times d_z}$ are two linear projection matrices, d_z is the common dimensionality of the fused features. LayerNorm is introduced here to stabilize training. Z represents the fusion features.

The fused features Z is projected into vector $\mathbf{s} \in \mathbb{R}^{\mathrm{D}}$ followed by a sigmoid function, where D is the number of the most frequent answers in the training set. Following [40], we utilize binary cross-entropy (BCE) as the loss function to train an D-way classifier on top of the fused features Z.

4 Experiments

In this section, we conduct experiments to evaluate the performance of our model on the largest VQA benchmark dataset. After introducing contrastive learning in our model, we learn that the temperature coefficients in contrastive learning can have an impact on the experimental results. Therefore, we conduct a number of quantitative and qualitative abatement experiments to explore the conditions under which our model MSAM performs best. Finally, we introduce optimal hyperparameters to compare MSAM with current methods.

4.1 Datasets

VQA2.0 is the most commonly used VQA benchmark dataset [41]. It contains human-annotated question-answer pairs relating to the images from the MS-COCO dataset [42], with 3 questions per image and 10 answers per question. The dataset is divided into three parts: train split (80k images and 444k QA pairs), val split (40k images and 214k QA pairs) and test split (80k images and 448k QA pairs). Additionally, there are two test subsets called test-dev and test-standard to evaluate model performance online. The results consist of three per-type accuracies (Yes/No, Number, and Other) and an overall accuracy.

4.2 Implementation Details

The model hyperparameters used in the experiments are shown below. The dimensions of input image features, input question features, public semantic space output image features, public semantic space output question features and fused multi-modal features are 2048, 512, 512, 512 and 1024, respectively.

The potential dimension of multihead attention is 512 and the number of heads h is set to 8. The potential dimension of each header is $d_h = d/h = 64$. The size of the answer vocabulary is set to 3129. The number of layers of MCA is set to $\mathrm{L} = 6$. To train this model, we still apply the same Adam solver [43] $\beta_1 = 0.9$ and $\beta_2 = 0.98$ introduced in the MCAN model. The basic learning rate in the model

is set to a minimum value ($2.5e^{-5}$, $1e^{-4}$) and the current epoch experienced starting from 1. The learning rate decreases according to the epoch experienced. All models are trained in the same batch of 15 epochs with a batch size of 64.

4.3 Baselines

In order to verify the effectiveness and generalization of our proposed approach to the VQA task, we compare it with the following models.

- **UpDn** [44]. This model is able to combine bottom-up and top-down attention mechanism and can compute attention at the level of objects and other salient image regions.
- **BAN** [21]. This model is able to compute the similarity between the image and the question using bilinear ensemble operations, and weight the features of the image and the question using the attention mechanism.
- **MUTAN** [45]. This model can be efficiently parameterized in a visual and textual bilinear interaction models (bilinear models) based on multi-modal tensor Tucker decomposition.
- **TRN+UpDn** [46]. This model maximizes the likelihood of estimating the joint distribution between observed questions and predicted answers.
- **DFAF** [1]. This model is used to dynamically fuse multi-modal features so that information is transmitted alternately between the image and the question modality.
- **DOG** [47]. It is able to learn an end-to-end VQA model from pixels directly to answers, and demonstrates good performance without the use of any region annotations in pre-training.
- **MLIN** [48]. The model is capable of utilizing multi-modal information, fusing sensitive information from each modality, and updating visual and textual features using multi-modal information.
- **HAN** [49]. This model is enabled to add semantic and structural information (hypergraph subgraph matching) to construct common concern graphs.
- **SUPER** [50]. The model is customized by five powerful specialized modules and dynamic routers to build a compact routing space. A variety of routing customizables can be used through explicitly calibrated visual semantic representations.
- **Co-VQA** [51]. This model is able to break down a complex question into a series of simple sub-questions and ultimately arrive at an answer to the question.
- **MRA-Net** [52]. The model explores both textual and visual relationships to improve performance and interpretation.
- **CAM** [53]. The Cascading Answer Model (CAM) is proposed, which extends the traditional single-stage VQA model to a two-stage model.
- **COB** [54]. A new regularization method is proposed for the VQA model, which exploits the theory of baroque (COB) to improve the information content of the joint space by minimizing redundancy.

- **CCC** [55]. The model divides questions into skills and concepts and combines them in new ways to improve the generalization of the VQA model.
- **DAQC** [56]. The model solves the VQA problem based on the visual problem construction of double attention and question categorization.
- **ALSA** [57]. The model is freeform and detection-based and aims to utilize prior knowledge for attention distribution learning.

4.4 Results

Validation Results on VQA2.0. Table 1 summarizes the results of our comparison with different methods in the VQA task. On the VQA2.0 dataset, Since our model focuses more on deeper understanding of multi-modal information at the semantic level, we have a fuller understanding of questions and images. The more comprehensively understood question and image features are fed into the fusion model to accomplish answer prediction. On the VQA2.0 dataset, our model outperforms the existing new model in terms of overall accuracy and accuracy in other categories. Since most of the existing papers validate the effect of online testing on the test set, we further compare it with the existing new method in Table 4. It shows that our model MSAM works better than the existing models.

Table 1. Results of MSAM compared with the new current models on VQA2.0 val. All Accuracy represents the overall accuracy, Yes/No represents the accuracy of the question type where the answer to the question is yes/no, Number represents the accuracy of the question with numeric answers, and Other represents the accuracy of the question with other question types.

Model	All Accuracy	Other	Yes/No	Number
UpDn [44]	63.15	55.81	80.07	42.87
BAN [21]	66.04	–	–	–
MUTAN [45]	63.61	–	–	–
TRN+UpDn [46]	65.10	57.10	82.61	45.10
DFAF [1]	66.20	–	–	–
DOG [47]	64.29	55.70	82.16	45.45
MLIN [48]	66.53	–	–	–
HAN [49]	65.50	–	–	–
MCAN [9]	66.88	58.04	84.62	49.00
SUPER [50]	66.59	58.07	85.15	48.27
COB [54]	63.80	55.86	81.36	43.30
Ours	**67.27**	**58.53**	**85.00**	**49.73**

Results of Introducing Hard Samples. With the addition of contrastive learning, we introduce hard samples. The role of hard samples is to enhance modal alignment. By enhancing modal alignment, question-image pairs with the same semantics are made better distinguishable from other sample pairs that do not have the same semantics. Table 2 shows the effect of adding hard samples and changing the number of hard samples on the results. Based on the results, it can be seen that as the number of hard sample increases to some extent, the various metrics are improved. However, due to the limitation of the number of batches, the number of hard samples cannot be increased without limit. Figure 2 visualizes the changes produced by various indicators by changing the number of hard samples.

Table 2. The effect of the number of hard samples on MSAM.

Model	All Accuracy	Other	Yes/No	Number
Ours$_{\tau=15}$	67.03	58.31	84.53	49.41
Ours+hard10 sample$_{\tau=15}$	67.25	58.45	84.91	49.45
Ours+hard16 sample$_{\tau=15}$	67.27	58.53	85.00	49.73
Ours+hard24 sample$_{\tau=15}$	67.24	58.43	84.92	49.40

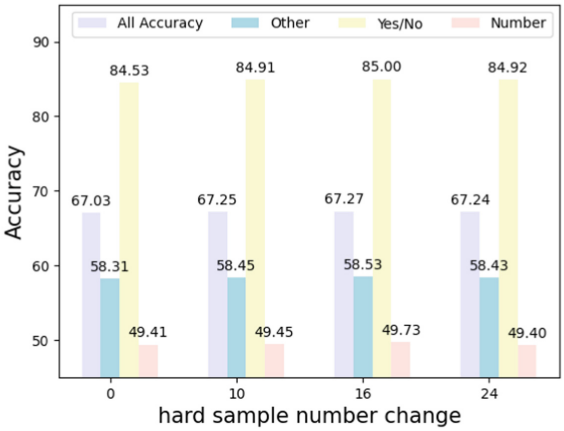

Fig. 2. Purple represents the All Accuracy indicator, blue represents the Other indicator, yellow represents the Yes/No indicator, and pink represents the Number indicator. (Color figure online)

Results of Changing Temperature Coefficients. With the introduction of contrastive learning, we learn that the presence of temperature coefficients in the contrastive learning formulation affects the extent to which comparative loss focuses on hard samples. For this reason, we conduct the following experiments to explore the effect of varying different temperature coefficients on the accuracy of the model. By setting the temperature coefficients to {1, 15, 50, 75, 100}, we obtain that the model is optimal when the temperature coefficient is 15. The results are shown in Table 3. In order to express more clearly the variation of the experimental results due to the temperature coefficients, we plot the following line graph (as shown in the Fig. 3).

Table 3. The effect of temperature coefficient variation on MSAM.

Model	All Accuracy	Other	Yes/No	Number
Ours$_{\tau=1}$	67.20	58.32	84.75	49.42
Ours$_{\tau=15}$	67.27	58.53	85.00	49.73
Ours$_{\tau=50}$	67.25	58.39	84.90	49.63
Ours$_{\tau=75}$	67.22	58.43	84.89	49.60
Ours$_{\tau=100}$	67.21	58.36	84.87	49.55

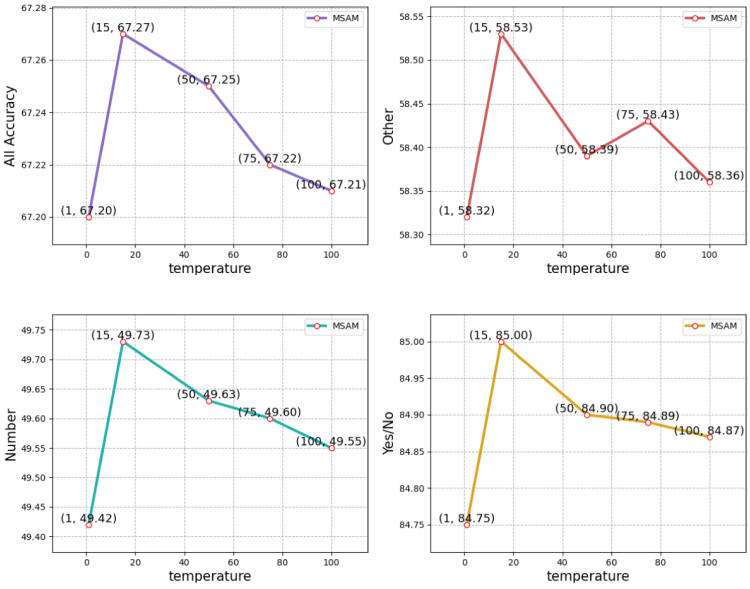

Fig. 3. The four line graphs show the effect on the various metrics as the temperature coefficients change.

Online Test for VQA2.0. Finally, we evaluate our results on the test-dev and test-std sets. The results show that our model MSAM can be used for modal interactions by digging deeper into the multi-modal semantic information from the semantic point of view. The VQA task is accomplished by deeply understanding the modal semantic information and combining the semantic information. As shown in Table 4.

Table 4. Results of MSAM compared with the new current models on test-dev and test-std sets.

Model	test-dev				test-std
	All Accuracy	Other	Yes/No	Number	All Accuracy
TRN+UpDn [46]	67.00	57.44	83.83	45.61	67.21
MRA-Net [52]	69.06	59.62	86.79	43.89	69.22
CCC [55]	69.78	–	–	–	70.09
SUPER [50]	69.23	59.35	85.42	51.16	69.69
Co-VQA [51]	–	–	–	–	70.39
CAM [53]	68.82	59.76	85.18	47.35	68.99
DAQC [56]	64.51	56.39	82.15	43.57	–
ALSA [57]	69.21	59.17	85.73	48.98	–
ours	**70.38**	**60.12**	**87.35**	**51.70**	**70.65**

5 Conclusions

In this paper, we design the MSAM for deep mining multi-modal semantics for VQA task. By constructing semantic space, we understand the information deeply at the semantic level. Multi-modal semantic alignment is achieved through contrastive learning, which reduces the distance between heterogeneous modalities and improves the relevance of semantic information. The attention mechanism, which is widely used in VQA task, has limited ability in deep semantic understanding, and it does not has the ability of multi-modal semantic alignment and reasoning to deeply understand the semantic relationship between different modalities. Our propose method well compensates the deficiencies existing in the attention mechanism. Extensive experiments on the VQA2.0 dataset show that our approach achieves competitive results compared to current methods.

References

1. Gao, P., Jiang, Z., You, H., 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 (2019)
2. Zhang, W., Yu, J., Hu, H., et al.: Multimodal feature fusion by relational reasoning and attention for visual question answering. Inf. Fusion **55**, 116–126 (2020)

3. Chen, T., Yu, W., Chen, R., et al.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6163–6171 (2019)
4. Zhou, H., Du, J., Zhang, Y., et al.: Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2617–2629 (2021)
5. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
6. Gu, J., Zhao, H., Lin, Z., et al.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1969–1978 (2019)
7. Yang, Z., He, X., Gao, J., et al.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
8. Chowdhury, M.I.H., Nguyen, K., Sridharan, S., et al.: Hierarchical relational attention for video question answering. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 599–603. IEEE (2018)
9. Yu, Z., Yu, J., Cui, Y., et al.: 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 (2019)
10. Chang, L., Zhang, C.: Vehicle taillight detection based on semantic information fusion. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) ICONIP 2021. CCIS, vol. 1517, pp. 528–536. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92310-5_61
11. Nguyen, B.X., Do, T., Tran, H., et al.: Coarse-to-fine reasoning for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4558–4566 (2022)
12. Chen, C., Han, D., Chang, C.C.: CAAN: context-aware attention network for visual question answering. Pattern Recogn. **132**, 108980 (2022)
13. Tu, G., Wen, J., Liu, C., et al.: Context-and sentiment-aware networks for emotion recognition in conversation. IEEE Trans. Artif. Intell. **3**(5), 699–708 (2022)
14. Xiong, L., Xiong, C., Li, Y., et al.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808 (2020)
15. Donahue, J., Anne Hendricks, L., Guadarrama, S., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
16. Xu, K., Ba, J., Kiros, R., et al.: Neural image caption generation with visual attention. In: Proceedings of the ICML, pp. 2048–2057 (2015)
17. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307 (2017)
18. Wang, Y., Yasunaga, M., Ren, H., et al.: VQA-GNN: reasoning with multimodal semantic graph for visual question answering. arXiv preprint arXiv:2205.11501 (2022)
19. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
20. Zhao, Z., Zhang, Z., Xiao, S., et al.: Open-ended long-form video question answering via adaptive hierarchical reinforced networks. IJCAI **2**, 8 (2018)

21. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Peng, C., Zhang, K., Ma, Y., et al.: Cross fusion net: a fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2021)
24. Tian, P., Mo, H., Jiang, L.: Image caption generation using multi-level semantic context information. Symmetry **13**(7), 1184 (2021)
25. Li, D., Li, D., Wang, C., et al.: Network embedding method based on semantic information. In: Proceedings of the 3rd International Conference on Advanced Information Science and System, pp. 1–6 (2021)
26. Adhikari, A., Dutta, B., Dutta, A., et al.: Semantic similarity measurement: an intrinsic information content model. Int. J. Metadata Semant. Ontol. **14**(3), 218–233 (2020)
27. Li, B., Lukasiewicz, T.: Learning to model multimodal semantic alignment for story visualization. arXiv preprint arXiv:2211.07289 (2022)
28. Bao, Y., Lattimer, B.M., Chai, J.: Human inspired progressive alignment and comparative learning for grounded word acquisition. arXiv preprint arXiv:2307.02615 (2023)
29. Fukui, A., Park, D.H., Yang, D., et al.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
30. Kim, J.H., On, K.W., Lim, W., et al.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
31. Yu, Z., Yu, J., Xiang, C., et al.: Beyond bilinear: generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. Neural Networks Learn. Syst. **29**(12), 5947–5959 (2018)
32. Chen, C., Han, D., Wang, J.: Multimodal encoder-decoder attention networks for visual question answering. IEEE Access **8**, 35662–35671 (2020)
33. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
34. Krishna, R., Zhu, Y., Groth, O., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**, 32–73 (2017)
35. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
36. Li, L., Liang, Y., Shao, M., et al.: Self-supervised learning-based Multi-Scale feature Fusion Network for survival analysis from whole slide images. Comput. Biol. Med. **153**, 106482 (2023)
37. Zheng, Z., Feng, X., Yu, H., et al.: Unsupervised few-shot image classification via one-vs-all contrastive learning. Appl. Intell. **53**(7), 7833–7847 (2023)
38. Yeo, Y.J., Sagong, M.C., Park, S., et al.: Image generation with self pixel-wise normalization. Appl. Intell. **53**(8), 9409–9423 (2023)
39. Ye, Y., Pan, Y., Liang, Y., et al.: A cascaded spatiotemporal attention network for dynamic facial expression recognition. Appl. Intell. **53**(5), 5402–5415 (2023)
40. Kulkarni, C., Rajesh, M,. Shylaja, S.S.: Dynamic binary cross entropy: an effective and quick method for model convergence. In: 2022 21st IEEE International

Conference on Machine Learning and Applications (ICMLA), pp. 814–818. IEEE (2022)
41. Goyal, Y., Khot, T., Summers-Stay, D., et al.: Making the v in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6904–6913 (2017)
42. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
43. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
44. Anderson, P., He, X., Buehler, C., 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 (2018)
45. Ben-Younes, H., Cadene, R., Cord, M., et al.: MUTAN: multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2612–2620 (2017)
46. Han, X., Wang, S., Su, C., Zhang, W., Huang, Q., Tian, Q.: Interpretable visual reasoning via probabilistic formulation under natural supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part IX. LNCS, vol. 12354, pp. 553–570. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_32
47. Jiang, H., Misra, I., Rohrbach, M., et al.: In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10267–10276 (2020)
48. Gao, P., You, H., Zhang, Z., et al.: Multi-modality latent interaction network for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5825–5835 (2019)
49. Kim, E.S., Kang, W.Y., On, K.W., et al.: Hypergraph attention networks for multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14581–14590 (2020)
50. Han, Y., Yin, J., Wu, J., et al.: Semantic-aware modular capsule routing for visual question answering. arXiv preprint arXiv:2207.10404 (2022)
51. Wang, R., et al.: Co-VQA: answering by interactive sub question sequence. In: Findings of the Association for Computational Linguistics: ACL (2022)
52. Peng, L., Yang, Y., Wang, Z., et al.: MRA-Net: improving VQA via multi-modal relation attention network. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 318–329 (2020)
53. Peng, L., Yang, Y., Zhang, X., et al.: Answer again: improving VQA with Cascaded-Answering model. IEEE Trans. Knowl. Data Eng. **34**(04), 1644–1655 (2022)
54. Jha, A., Patro, B., Van Gool, L., et al.: Barlow constrained optimization for visual question answering. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1084–1093 (2023)
55. Whitehead, S., Wu, H., Ji, H., et al.: Separating skills and concepts for novel visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5632–5641 (2021)
56. Mishra, A., Anand, A., Guha, P.: Dual attention and question categorization-based visual question answering. IEEE Trans. Artif. Intell. **4**(1), 81–91 (2022)
57. Liu, Y., Zhang, X., Zhao, Z., et al.: ALSA: adversarial learning of supervised attentions for visual question answering. IEEE Trans. Cybern. **52**(6), 4520–4533 (2022)