Data Science and Engineering (2022) 7:339–353
https://doi.org/10.1007/s41019-022-00200-9

RESEARCH PAPER

A Multi-level Mesh Mutual Attention Model for Visual Question Answering

Zhi Lei[1] · Guixian Zhang[1] · Lijuan Wu[1] · Kui Zhang[1] · Rongjiao Liang[1]

Received: 4 August 2022 / Revised: 4 September 2022 / Accepted: 16 October 2022 / Published online: 30 October 2022

Abstract
Visual question answering is a complex multimodal task involving images and text, with broad application prospects in human–computer interaction and medical assistance. Therefore, how to deal with the feature interaction and multimodal feature fusion between the critical regions in the image and the keywords in the question is an important issue. To this end, we propose a neural network based on the encoder–decoder structure of the transformer architecture. Specifically, in the encoder, we use multi-head self-attention to mine word–word connections within question features and stack multiple layers of attention to obtain multi-level question features. We propose a mutual attention module to perform information exchange between modalities for better question features and image features representation on the decoder side. Besides, we connect the encoder and decoder in a meshed manner, perform mutual attention operations with multi-level question features, and aggregate information in an adaptive way. We propose a multi-scale fusion module in the fusion stage, which utilizes feature information at different scales to complete modal fusion. We test and validate the model effectiveness on VQA v1 and VQA v2 datasets. Our model achieves better results than state-of-the-art methods.

Keywords Visual question answering · Multi-level · Mutual attention · Multi-head

1 Introduction

Visual question answering [1] combines the fields of natural language processing and computer vision. One of the most challenging tasks in machine learning is visual question answering. This technology has broad application prospects in human–computer interaction [2] and medical assistance [3]. Visual question answering requires a simultaneous understanding of visual and linguistic information. So achieving information interaction and fusion across modalities is a major challenge.

In early work, some scientists [4, 5] add or concatenate extracted image features and question features to obtain fused features. However, this processing does not tap into the interactions between modalities, which is important for visual question answering. Lu et al. [6, 7] considered the interaction between two modalities but ignored the dense interaction within a single modality.

One must first comprehend the question and image meanings, as shown in Fig. 1, before grasping keywords and image regions. Objects can be represented by multiple modalities. For example, the word "table" in Fig. 1 corresponds to the table area in the image. So we need to map keywords and image regions together. This method can only obtain a rough inter-modal relationship between the two [6, 7]. According to the human brain's thinking process, we must first focus on the image and then understand the question. We can get the correct answer by paying close attention to each other many times. We miss the implicit link between image and question without the mutual attention stage and thus miss the most relevant features between the two.

In addition, in the fusion stage, simply adding the two modal features will bring much noise, because the image or question respectively contains a lot of noise information irrelevant to answering the question [8]. For example, The "table" in Fig. 1 on the right picture contains many electronic products that are not related to mobile phones. Furthermore, image features at different scales may represent the same information [9]-different models of mobile

✉ Zhi Lei
leiz@stu.gxnu.edu.cn

[1] Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, GuiLin 541000, GuangXi, China

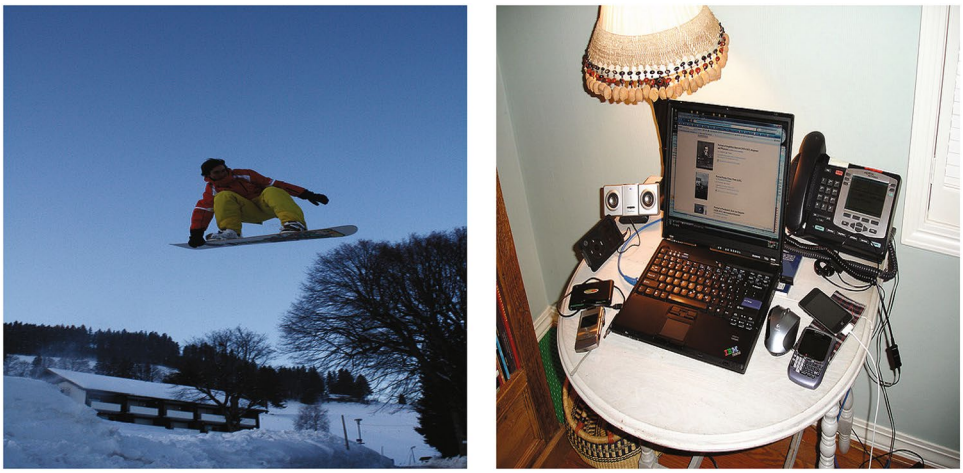


Fig. 1 Two examples from VQA v2 dataset. Simply stitching or fusing two modal features can make the model misunderstand the fused features and fail to answer the question accurately. Therefore, we need to strengthen the information interaction between modalities. In addition, critical information within the modality also needs to be mined

phones in different areas as shown in the picture on the right in Fig. 1. Therefore, establishing mutual attention between modalities, mining hidden relationships between modalities, and exploring abstract information between different modalities at multiple scales in the fusion stage are urgent problems to be solved in visual question answering.

Aiming to solve the above problems, we design a novel multi-level mesh mutual attention model. Different from [4, 10–12], this approach can be well used to explore the relationship between features [13–16]. Unlike previous works [7, 17–19], we achieve a more concise and effective way of information interaction between modes, making full use of multi-level question features to refine the abstract connections between modes.

Considering that for the same image, different questions focus on different objects. The area objects involved in different images are also different for the same question. The interaction can not be expressed in the feature fusion stage by simply splicing or adding two features. Therefore, we also design a multi-scale adaptive fusion module. The module multiples mini-batch transformations of different dimensions to solve the above problem, and the fusion information of all scales is aggregated adaptively.

Briefly speaking, our contributions are summarized as follows:

1. We build a multi-level mesh mutual attention model with an encoder–decoder architecture. The multi-level mesh decoder performs mutual attention operations on multi-level question features and image features, aggregating information from all levels in an adaptive manner. We explore and verify that using both low-dimensional and high-dimensional multi-level question features is beneficial for visual question answering.
2. We design an adaptive pyramid-shaped multi-scale fusion module in the fusion stage. Pyramid linear transformation is performed on the fusion features in multi-layer mini-batches, and multi-scale fusion is adaptively completed.
3. Numerous experiments on VQA v1 and VQA v2 datasets demonstrate that our model achieves state-of-the-art results on the comparison algorithms. In ablation experiments, we build our baseline model, incrementally adding modules, and verify the effect of each proposed module.

The article is organized as follows: Sect. 2 reviews previous research on visual question answering. Section 3 presents the overall framework of our proposed method and details each module in the framework diagram. Section 4 is our experimental part, which includes comparative experiments, ablation experiments, and quantitative analysis. Section 5 concludes the article.

2 Related Work

Visual question answering Since the concept of visual question answering was proposed in [20], many large datasets such as VQA v1 [1], VQA v2 [21] and other datasets have been released to the public, attracting a large number of scholars to conduct research. Antol et al. [1] extended visual question answering to free-form and open-ended and proposed a model combining convolutional neural networks (CNN) and long short-term memory networks (LSTM) to solve visual question answering problems. Different from [5], Gao et al. [22, 23] adopted later fusion in the feature fusion strategy, obtained question features and image features separately, and then performed the feature fusion operation. The above works all use convolutional neural networks, resulting in incomplete object information extracted in image feature extraction. Anderson et al. [10] used the object detection network Faster-RCNN [24] to extract features of the objects in the image and used a threshold to

select some detected objects as visual input. After that, this method became the mainstream image feature processing operation in visual question answering tasks. Teney et al. [6] improved the model on this basis and introduced several techniques to improve model performance. Lu et al. [25] combined the features obtained by CNN with the features obtained by the target detection network.

However, the above work selects all image features in feature extraction, which contain noise information that is irrelevant to the question, and feeding these into the classifier will affect the prediction of the answer. Before feature fusion, the above methods only map the features of the two modalities to the same space in a linear projection manner, without considering the information interaction between modalities [8].

Attention mechanism Thanks to the research progress of the attention mechanism in machine translation and image description, many works have introduced the attention mechanism into the field of visual question answering, which has improved the correct rate of answering questions. Shih et al. [12] simply multiplied image features and question features to obtain attention weights. The attention weights are then used to guide the model to focus on image features that are most relevant to the question. Yang et al. [26] built a stacked attention model on its basis, which made the model pay more attention to question-related regions in the image through multiple iterative attention operations. Xiong et al. [27] proposed a gated recurrent unit with attention to facilitate answer prediction. The above work filters out image features that are unrelated to the question by introducing an attention mechanism. Lu et al. [6] proposed a hierarchical attention network to construct joint attention at the word level, phrase level, and sentence level, respectively, and provided two different attention construction methods. Nam et al. [28] introduced a memory vector based on it to obtain more detailed information about images and questions. The above methods mainly focus on using the attention mechanism to select question features and image features. In terms of feature fusion, Fukui et al. [8] built a multi-modal compact bilinear pooling model, which simultaneously uses the outer product and the Kronecker product to complete the multi-modal fusion operation. Kim et al. [7] modified it and built a multi-modal low-rank bilinear pooling model, which uses Hadamard product to fuse the two features. Through this way, the number of parameters of the model is reduced. Nguyen et al. [17] proposed a new attention mechanism that enables dense bidirectional interaction between two modalities, which improves the accuracy of answering questions. Patro et al. [18] argued that previous image attention focuses on regions inconsistent with humans and proposed a differentiated attention mechanism. Yang et al. [19] constructed a mutual attention network and considered different question categories when fused. The previous work considers the information interaction between modalities, but the method is more complicated and ignores the interaction of key information within the modalities. Our work considers the information interaction between modalities and mines the implicit relationships between keywords or key regions within the modalities. The use of the Transformer method makes the information interaction between modalities more convenient. In addition, in the fusion stage, we also consider the information aggregation of fusion features from different scales.

3 Methodology

This section will introduce the proposed multi-level mesh mutual attention model in detail. As shown in Fig. 2, this model is mainly composed of five parts: feature extraction, encoder, multi-level mesh decoder, adaptive multi-level feature fusion, and answer classifier.

3.1 Feature Extraction

In Sect. 2, we have mentioned that using basic CNN to extract image features results in incomplete object information extracted in the image features extraction step. Anderson et al. [10] adopted Faster-RCNN to extract image features and won the VQA 2017 challenge. Furthermore, they also analyze in detail the performance impact of each choice on the challenge-winning model in [29]. After that, this image feature extraction method was widely adopted and became the mainstream image processing method in visual question answering. Following Anderson et al. [10, 24, 29], we use the Faster-RCNN model[1] pretrained on Visual Genome (VG) [30][2] as the image feature extraction network. After extracting image features from Faster-RCNN, we change the threshold for object detection [31] to obtain a dynamic object detection candidate region K, where $K \in [10, 100]$. The input image will be denoted as $I = [v_1, v_2, \ldots v_k]^T$, where $v_i \in R^{2048}$ is the convolutional features obtained after average pooling of the image in the target detection candidate box. Considering different images, the number of detected candidate regions is different. In order to facilitate processing, for different numbers of target detection candidate regions obtained by using Faster-RCNN in different pictures, we uniformly use zero vectors padding to fill the candidate regions to the maximum scale K. Ultimately, the image feature we get is a feature matrix $I \in R^{K \times 2048}$.

For question features, we first perform word segmentation on the input question text. In the VQA dataset, only 0.25% of the questions are longer than 14 words [29]. This part of

[1] https://github.com/peteanderson80/bottom-up-attention.

[2] http://visualgenome.org/.

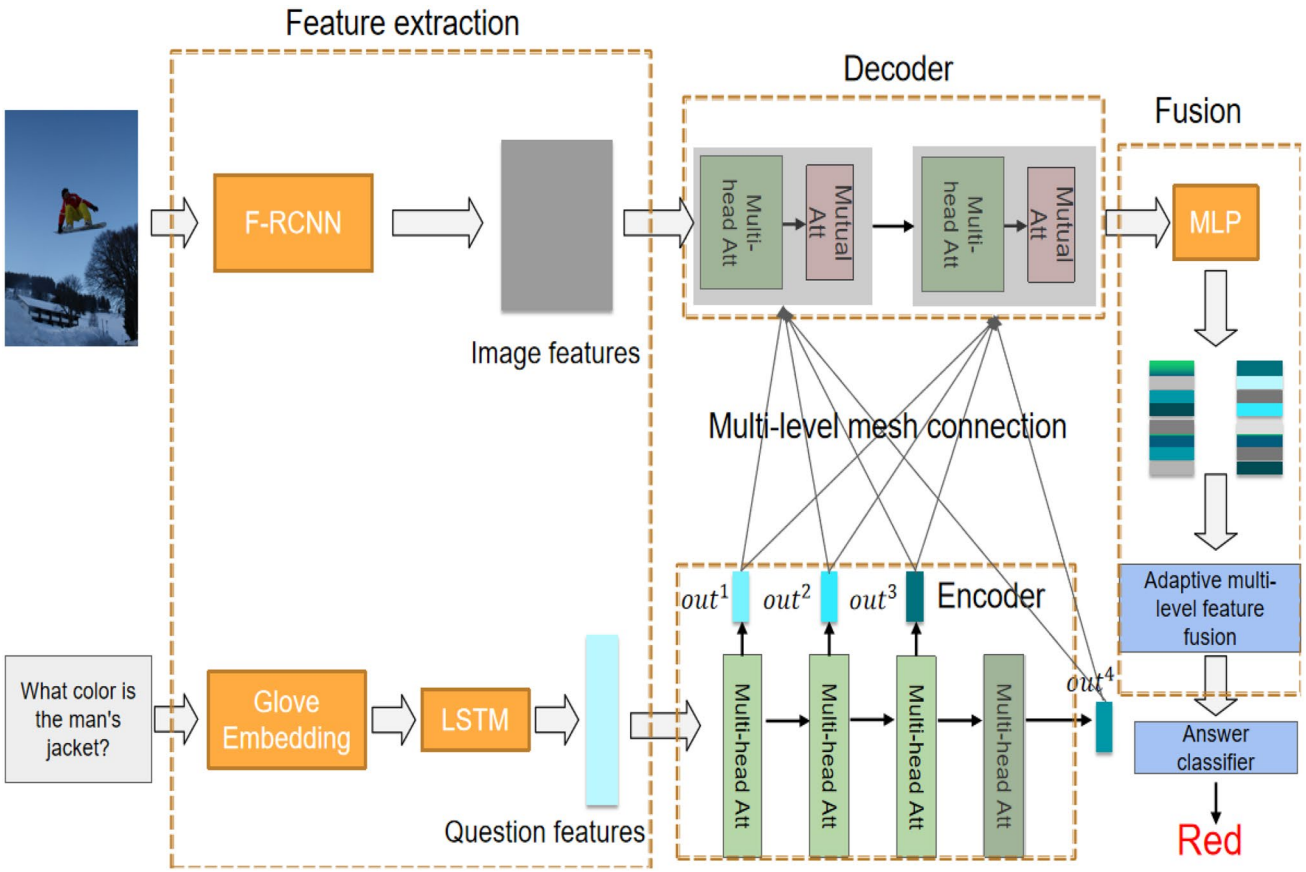


Fig. 2 The overall structure of the proposed multi-level mesh mutual attention model under the example question–answer pair: "What color is the man's jacket? Red." The model consists of five parts: feature extraction, encoder, multi-level mesh decoder, adaptive multi-level fusion module, answer classifier. Figure 3 shows the details of mutual attention module. Multi-level mesh decoder consists of mutual attention module and multi-level mesh connection as shown in Fig. 4. More details of the adaptive multi-level fusion module are in Fig. 5

the question data has little impact on model performance. Based on this, we compress each question into 14 tokens, and the tokens after the 14th will be discarded. For questions whose length is less than 14 tokens, we use zero vectors for padding. After this, we use 300-D GloVe model to perform word embeddings for each token in the question, converting each token into a 300-D word vector. This step obtains a word vector sequence of length $n \times 300$, where $n \in [1, 14]$. n represents the number of tokens in each question. We then feed the sequence of GloVe encoded word vectors into a d_Q dimensional single-layer LSTM network, where $d_Q \in R^{512}$. Finally, the question feature is obtained as a vector matrix $Q \in R^{n \times d_Q}$.

3.2 Encoder

In order to mine the connection between words in question features and obtain multi-level question features, encoder is composed of a stacked multi-head self-attention mechanism, as shown in Fig. 2. The self-attention mechanism [13] is defined as follows:

$$Self\_att(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q, K, V are vector matrices of the same dimension, d is the scaling factor, $Softmax(\cdot)$ stands for softmax activation function.

The input of each multi-head self-attention layer is the output of previous layer. Among them, the input of the first layer is the question features after passing through the LSTM network. Through this operation, the upper layers utilize the known information of the previous layer and further mine it, we can obtain multi-level question features Q_{multi} in different semantic dimensions.

$$Q_{att} = Self\_att(W_q Q, W_k Q, W_v Q), \tag{2}$$

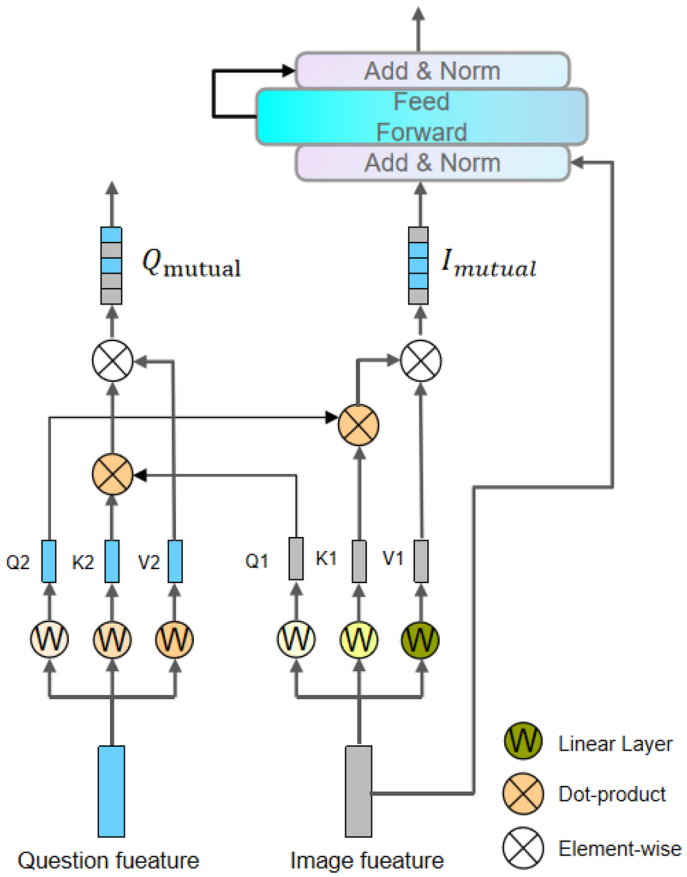


Fig. 3 The process of mutual attention module, which is obtained by changing the self-attention structure. We use Q_2 from the question features to refine the image features and Q_1 from the image features to refine the question features, respectively. Thereby, the interaction information between the two features is obtained. Image features are nested with *Add&Norm*$(\cdot)$ and a pointwise feed-forward layer in the final stage

$$head = Add\&Norm(Q_{att}), \tag{3}$$

$$Out = concat(head_1, head_2, \ldots, head_i), \tag{4}$$

$$Q_{multi} = (Out_1, Out_2, \ldots, Out_m), \tag{5}$$

where Q is the question features obtained after passing through the LSTM network, and W_q, W_k, W_v represents three learnable matrix, *Add&Norm*$(\cdot)$ is composed of residual connection [32] and layer normalization [33], *concat*$(\cdot)$ stands for concatenation operation, m is the number of stacked multi-head attention layers.

3.3 Multi-level Meshed Decoder

We take the multi-level question features Q_{multi} from the encoder and the preprocessed image features I as inputs.

Mutual attention module For the self-attention layers in the Transformer model, Q, K, V all come from the same modality. In this way, only the information inside the modal can be captured, and the information interaction between the modal can not be captured. We implement mutual attention based on the self-attention module to realize the information interaction between different modalities fully.

As shown in Fig. 3, we first pass the image features I through the self-attention module to obtain image features I_s. Then, together with the question features Out, it is sent to the mutual attention module for information exchange between modalities. Image features I_s and question features Out are transformed into the same dimensions Q_1, K_1, V_1 and Q_2, K_2, V_2 through three linear layers with different weight parameters, respectively.

We perform the dot-product operation on Q_2 and K_1 to get the dot-product similarity weight between Q_2 and K_1. Use this weight to refine V_1 to get the question features Q_{mutual} after interacting with the image features I_s.

$$Q_{mutual} = mutual\_att(Out, I_s), \tag{6}$$

$$mutual\_att(Out, I_s) = Self\_att(W_{q_2}Out, W_{k_1}I_s, W_{v_1}I_s), \tag{7}$$

where *mutual_att*$(\cdot)$ represents our mutual attention module, modified on the basis of the self-attention mechanism, details can be seen in Fig. 3, W_{q_2}, W_{k_1}, W_{v_1} represents three different matrices of learnable parameters.

Similarly, Q_1 and K_2 perform the dot-product operation and then use their dot-multiply similarity to refine V_2 to

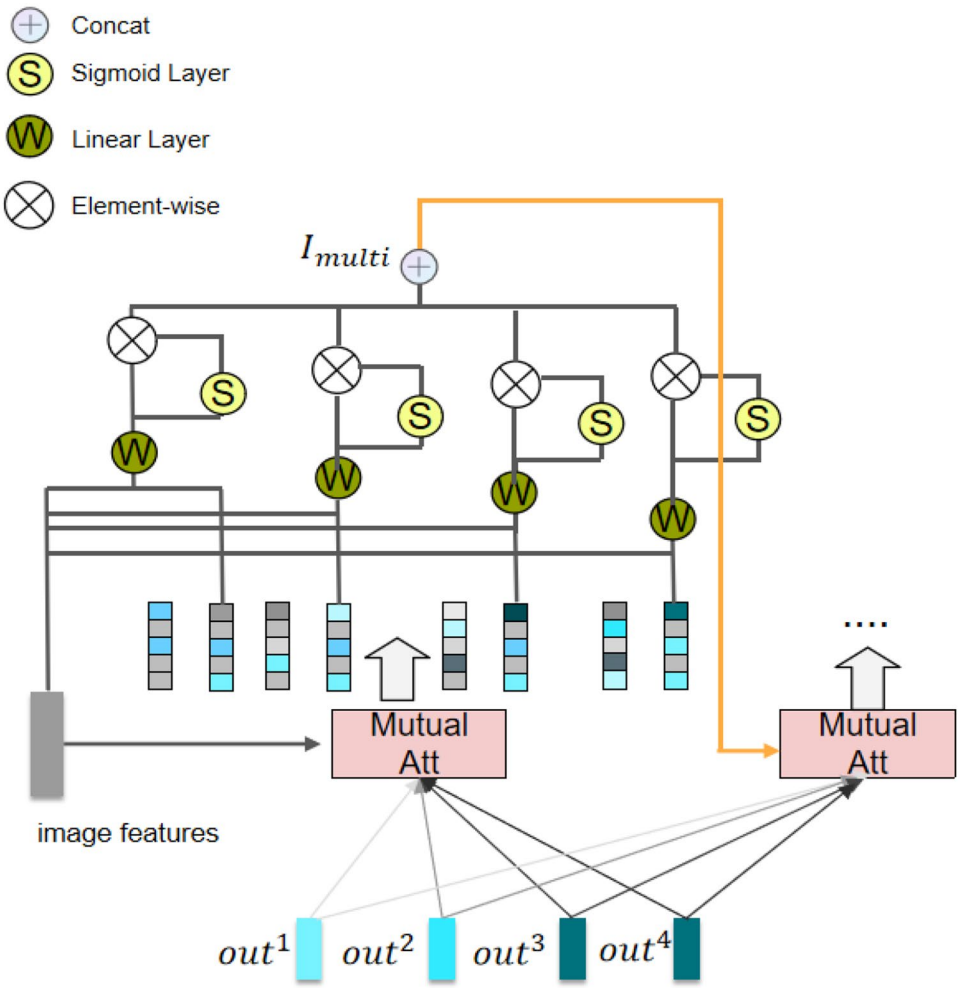


Fig. 4 The processing flow of the multi-level mesh decoder. First, the multi-level question features Q_{multi} and image features I_s are fed into the mutual attention module. The image feature sets I_{mutual} after mutual attention are aggregated according to formula (10) to obtain multi-level image features I_{multi}. The multi-level image features I_{multi} and the multi-level question features Q_{multi} from the encoder will be fed into the next decoding layer as inputs

obtain the image features I_{mutual} after interacting with the question features Out. $Add\&Norm(\cdot)$ and a pointwise feed-forward layer are nested again.

$$I_{mutual} = mutual\_att(I_s, Out), \tag{8}$$

$$mutual\_att(I_s, Out) = Self\_att(W_{q_1} I_s, W_{k_2} Out, W_{v_2} Out), \tag{9}$$

where $W_{q_1}, W_{k_2}, W_{v_2}$ represents three different matrices of learnable parameters.

Multi-level mesh decoding layer In the field of machine translation, the traditional transformer only considers the information of the last layer of the encoder at the decoder side. And each time it passes through a layer of self-attention layer, the important information in the question features will change. We design a multi-level mesh decoding layer that utilizes multi-level question features at the encoder side to interact with image features and aggregates information in a mesh-connected fashion, as shown in Fig. 4.

The inputs to the multi-level mesh decoder are the multi-level image features I_{multi} output from the previous decoding layer and the multi-level question features Q_{multi} from the encoder. For the first layer of the decoder, the inputs are image features I_s and multi-level question features Q_{multi}. The decoder is defined as follows:

$$I_{multi} = \sum_{i=1}^{n} (\alpha_i * concat(mutual\_att(I_s, Q_{multi}), I_s)), \tag{10}$$

where $*$ denotes element multiplication operation, α_i is a weight matrix that can measure the contribution of different levels of question features in the interaction and the similarity between the interaction results, α_i is defined as follows:

$$\alpha_i = Sigmoid(W_i(concat(mutual\_att(I_s, Q_{multi}), I_s)) + b_i), \tag{11}$$

where $Sigmoid(\cdot)$ represents the sigmoid activation function, W_i is a learnable parameter matrix, b_i is a learnable bias vector.

In the decoding layer, image features interact with multi-level question features from different levels in the encoder respectively, resulting in a feature set:

$$I_{mutual} = [I^1_{mutual}, I^2_{mutual}, ..., I^i_{mutual}], \tag{12}$$

$$Q_{mutual} = [Q^1_{mutual}, Q^2_{mutual}, ..., Q^i_{mutual}], \tag{13}$$

According to Fig. 4, we spliced I_{mutual} and I_s, and then multiplied by α_i after a linear layer dimensionality reduction, and finally summed to get I_{multi}, I_{multi} will continue to be sent to the next decoding layer for mesh interaction with multi-level question features. We take the question features Q^i_{mutual} and I_{multi} obtained after the last layer of mesh interactions in the decoder as outputs.

3.4 Adaptive Multi-level Feature Fusion

He et al. [32] demonstrates that multi-layer learning with a small number of hidden units for higher layers can learn more abstract information in the features. Inspired by He et al. [32], to fully and effectively integrate feature representations at different levels of abstraction, we design an adaptive multi-level feature fusion module shown in Fig. 5 to fuse the feature representations of all layers in different dimensions to obtain the final fused features.

Before fusion, we apply a multi-layer perceptron [31] to reduce the dimensionality of features output by the decoder. The adaptive multi-level feature fusion module takes the question features Q^i_{mutual}, image features I_{multi} from the decoder as inputs and finally outputs the multi-level fusion features f in different dimensions.

The adaptive multi-level feature fusion module is defined as follows:

$$f = Tan(concat(\alpha^1_{1024} * f^1_{1024}, \alpha^2_{512} * f^2_{512}, \alpha^3_{512} * f^3_{512})), \tag{14}$$

where $Tan(\cdot)$ stands for tan activation function, f^1_{1024},f^2_{512},f^3_{512} are the fusion features of different dimensions, they are obtained by the following operation:

$$f^1_{1024} = Tan((concat(Q^i_{mutual}, I_{multi})W_1 + b_{1024})), \tag{15}$$

$$f^2_{512} = Tan((concat(Q^i_{mutual}, I_{multi})W_2 + b_{512})), \tag{16}$$

$$f^3_{512} = Tan((concat(Q^i_{mutual}, I_{multi})W_3 + b_{512})), \tag{17}$$

where W_1, W_2, W_3 are three different matrices of learnable parameters respectively, b_{1024}, b_{512} are two different learnable bias vectors. This transformation can learn more abstract information in the fused features [32]. At the same time, multi-layer fusion features can provide more helpful information for the classifier. Furthermore, we introduce α following the inspiration of Multi-level Meshed Decoder:

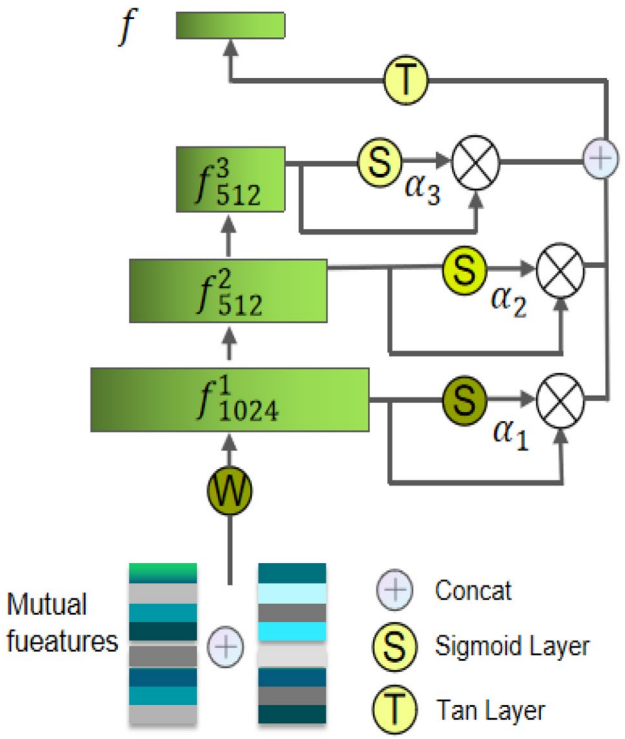


Fig. 5 The process of multi-level fusion modules. We take the multi-level image features and question features from the multi-level mesh decoder as inputs. The fusion features in different dimensions are obtained in the form of pyramids, and the fusion features from different dimensions are adaptively aggregated in the form of formula (14)

$$\alpha^1_{1024} = \text{Sigmoid}(f^1_{1024}), \tag{18}$$

$$\alpha^2_{512} = \text{Sigmoid}(f^2_{512}), \tag{19}$$

$$\alpha^3_{512} = Sigmoid(f^3_{512}) \tag{20}$$

3.5 Answer Classifier

The answer classifier consists of a layer normalization layer, a linear layer, and a sigmoid nonlinear layer. The answer classifier projects the previously obtained fused features as probabilities:

$$logits = Sigmoid(Norm(fW_{1024\times3129} + b_{3129})), \tag{21}$$

where $Norm(\cdot)$ stands for layer normalization operation, $W_{1024\times3129}$ represents a learnable parameter matrix with a dimension of 1024×3129, b_{3129} is a learnable bias vector.

The multi-level fusion features first pass through a layer normalization layer. Then a linear layer is used to convert

them to 3129 dimensions, which performs pre-classification operations on the features. After that, the sigmoid activation function projects the 3129-dimensional fused features into answer probabilities. Following Teney et al. [29], we train the classifier with a binary cross-entropy loss function on fused features *f*.

4 Experiments

In this section, we validate the performance of the proposed model on the VQA v1 [1] and VQA v2 [21] dataset.[3] First, we will introduce the dataset and detailed parameter settings in the experiments. Then, we compare and discuss the experimental results with different methods. Finally, considering the influence of different structures and parameters on the model performance, we conduct ablation experiments on the method proposed in this paper. We also conduct a quantitative analysis to explore the effect of combining different levels of multi-level features and different numbers of decoding layers on the model efficiency.

4.1 Dataset

There are two primary large-scale datasets about visual question answering: VQA v1 [1] and VQA v2 [21]. Both are built on the Microsoft Common Objects in Context(MSCOCO) [34] image dataset and divided into training, validation, and test set.

VQA v1 VQA v1 [1] is the first large dataset in the visual question answering domain, consisting of 204,721 images from the MSCOCO [34] image dataset, with at least three questions per image and ten manually annotated answers per question. There are 614,163 questions and 3,698,610 answers. The questions were divided into three subsets: the training set (248,349 questions), the validation set (121,512 questions), and the test set (244,302 questions). All questions can be divided into three categories: yes/no, number, and other. Answers are divided into training set answers (2,483,490 answers) and validation set answers (1,215,120 answers). Furthermore, the test set is divided into a test development set and a test standard set. The results of these two test subsets can only be obtained online on the VQA Challenge.[4]

VQA v2 VQA v2 [21] dataset, updated from VQA v1 [1] dataset, also contains 204,721 images and is currently the most commonly used large-scale public dataset in the visual question answering field. Unlike the VQA v1 [1], the VQA v2 [21] dataset has a larger sample of questions, solving the data imbalance problem in the VQA v1 [1] dataset and making the dataset smoother in terms of linguistic bias. Specifically, VQA v2 [21] has 1,105,904 questions divided into three subsets: the training set (443,757 questions), the validation set (214,354 questions) and the test set (447,793 questions). The data set had a total of 6,581,110 answers and was divided into training set answers (4,437,570 answers) and validation set answers (2,143,540 answers).

Data Augmented We have used VG [30] dataset as a means of data enhancement in the experimental part for a fair comparison. Similar to existing strategies [10, 11, 35], we first select images in the VG [30] dataset that appear in both the MSCOCO [34] train and val datasets, and obtain question–answer pairs associated with these images. Next, we select the question–answer pairs whose answers appear in the candidate answer set(described in 4.2) as the final data-augmented question–answer pairs. We will briefly introduce the VG [30] dataset later.

Visual Genome The VG [30] dataset consists of 108,249 images from YFCC100M [36] and MSCOCO [34]. The dataset has 1.7 million question–answer pairs, with an average of 17 question–answer pairs per image. The VG [30] dataset does not have the yes/no binary of the VQA dataset, in order to encourage the use of more complex questions.

VQA v1 [1] and VQA v2 [21] are large-scale mainstream general datasets in visual question answering. Therefore, we will conduct experiments on these two datasets. We will validate the model's overall performance and conduct ablation experiments to verify individual module performance.

4.2 Evaluation Metric

Open visual question answering is defined as a multi-category classification problem. Simple accuracy can also be used in it. But in this case, the answer predicted by the algorithm must be exactly the same as the ground truth. This evaluation standard is too strict, which will lead to ambiguity problems [37]. The creator of the VQA dataset, Antol et al. [1], proposed a new VQA evaluation metric widely accepted as a consensus in the VQA field. The VQA evaluation metric is defined as follows:

$$\text{Accuracy}(a) = \min\left\{\frac{n}{3}, 1\right\}, \tag{22}$$

where n is the same number of predicted answers as correct answers, a represents an answer, $\min(\cdot)$ is the minimum value operation.

Under the VQA evaluation metric, as long as the answer predicted by the algorithm can be consistent with three or

[3] https://visualqa.org/download.html.

[4] https://eval.ai/web/challenges/challenge-page/830/submission.

Table 1 The performance comparison of our method and other methods on the VQA v1 [1] dataset

Methods	VQA v1 test-dev				VQA v1 test-std			
	Overall	Yes/No	Numbers	Other	Overall	Yes/No	Numbers	Other
LSTM Q+I [1]	53.74	78.94	35.24	36.42	54.10	79.00	35.60	36.80
DPPnet [23]	57.22	80.71	37.24	45.77	57.36	80.28	36.92	42.24
MLB [38]	65.08	84.14	38.21	54.87	65.07	84.02	37.90	54.77
DCN [18]	65.4	83.8	39.1	55.2	–	–	–	–
DCA [17]	66.89	84.61	42.35	57.31	67.02	85.04	42.34	56.98
CAQT [19]	66.37	82.63	42.02	57.98	66.53	82.88	41.15	58.05
ATCG [39]	69.47	**86.72**	43.05	60.70	69.64	**86.87**	42.63	60.77
ALMA [40]	68.94	85.49	42.09	59.97	68.76	84.11	42.59	58.06
UFSCAN [35]	69.06	–	–	–	69.34	–	–	–
Ours	**69.74**	84.63	**45.25**	**61.51**	**69.86**	86.67	**45.48**	**60.95**

Results in Table 1 are the performance of a single model of the compared methods on the same training set. "–" indicates the result is not available. "Overall" represents the final overall accuracy. "Yes/No", "Num", and "Other" respectively indicate the accuracy of three different question types under the subdivision

Bold value is the best among all the methods

more manually annotated answers, it is judged as the correct answer. The category label of visual question answering is a predefined set of candidate answers, and we select the answers that are more than nine times in the VQA dataset as the candidate answer set [29]. There are 2,185 candidate answers in the candidate answer set for the VQA v1 [1] dataset and 3,129 candidate answers in the candidate answer set for the VQA v2 [21] dataset.

4.3 Implementation Details

All experiments we conducted are based on the PyTorch deep learning framework and use a TITAN XP to train the model. Our models are trained in an end-to-end manner. Below we will introduce the parameters used by the model in the experiments. For image features, the image features extracted from Faster R-CNN are 2048-dimensional. For question features, the length of each question is set to 14, and the dimension of the question features processed by a single layer LSTM network is 512-dimensional. Following the setting in [13], for the parameters of the multi-head self-attention and mutual attention layers, we set the dimension of hidden features to 512. Number of attention heads is 8, and the feature dimension of the attention head is 64. We set the initial learning rate as min(2.5te?5, 1e?4), where t is the current step and the starting value is 1. When t is greater than 10, the learning rate decreases by 1/5 every two steps. The parameters of the dropout layer used in the model are set to 0.1. For the linear layers in the model, we initialize the parameters with a uniform distribution. The bias of the linear layer is initialized to 0. We set the batch size to 64 and train the model for a total of 15 epochs. We choose the best epoch parameter to test the model and generate a json file for online submission. For the VQA v2 [21] validation set results, we use only the training set to train the model. For online test results on test-dev set and test-std set, we use the training set and validation set for model training. In addition, VG [30] dataset is also used as a means of data augmentation.

4.4 Overall Accuracy

All comparison algorithms are trained on training and validation sets, and tested on test-dev and test-std sets for a fair comparison. We also conduct fair comparisons for some additional methods using the VG [30] dataset. For ease of reference, we will first briefly describe the models used for comparison[5] in Tables 1 and 2.

Methods based on CNN+LSTM structure:

- LSTM Q+I [1] uses VGGNet to extract image features, a two-layer LSTM network encodes question features, and finally fuses them by element multiplication.
- DPPnet [23] proposes a CNN dynamic parameter layer that can be adaptively changed according to the questions.
- VQA Team-LSTM+CNN [21] aims to address the linguistic bias that exists in VQA v1 [1] dataset, the VQA v2 [21] dataset is proposed by collecting complementary images to balance the dataset. The model extracts image and text features using a CNN+LSTM network structure.

Methods based on attention mechanism:

[5] Please note that if one of the following methods does not appear in Tables 1 or 2, it means that the method has no test results on the dataset corresponding to Tables 1 or 2.

Table 2 The performance comparison of our method and other methods on the VQA v2 [21] dataset

Methods	VQA v2 test-dev				VQA v2 test-std			
	Overall	Yes/No	Numbers	Other	Overall	Yes/No	Numbers	Other
VQA Team-Prior [21]	–	–	–	–	25.98	00.36	01.07	61.20
VQA Team-Language Only [21]	–	–	–	–	44.26	31.55	27.37	67.01
VQA Team-LSTM+CNN [21]	–	–	–	–	54.22	35.18	41.83	73.46
MLB [38]	–	–	–	–	62.54	79.85	38.64	52.95
MF-SIG* [11]	64.73	81.29	42.99	55.55	–	–	–	–
Adelaide Model+detector* [10]	65.32	81.82	44.21	57.10	65.67	82.2	56.26	43.9
DCA [17]	65.12	83.18	47.32	56.10	66.08	83.48	56.33	47.12
CAQT [19]	66.37	82.63	42.02	57.98	66.53	82.88	58.05	47.15
SOMA [41]	68.38	84.86	47.59	59.06	68.67	–	–	–
ATCG [39]	69.13	85.80	51.54	59.17	69.57	86.17	**59.27**	51.46
ALMA [40]	68.12	84.62	47.08	58.14	–	–	–	–
UFSCAN* [35]	69.83	85.21	50.98	**60.98**	70.09	85.51	61.22	51.46
Ours	69.54	86.05	50.91	59.67	70.08	86.23	51.23	**60.35**
Ours*	**70.03**	**86.32**	**52.21**	60.16	**70.28**	**86.69**	51.83	60.22

Results in Table 2 are the performance of a single model of the compared methods on the same training set. "*" indicates augmented with VG [30] dataset. "–" indicates the result is not available. "Overall" represents the final overall accuracy. "Yes/No", "Num", and "Other" respectively indicate the accuracy of three different question types under the subdivision

Bold value is the best among all the methods

- MF-SIG [11] proposes a Conditional Random Field(CRF) method to construct structural attention for image regions, which solves the problem of limited perceptual field of CNN.
- MLB [38] proposes a low-rank bilinear pooling method using Hadamard product for multimodal fusion learning.
- Adelaide Model+detector [10] applies the bottom-up attention mechanism with Faster R-CNN model to the field of visual question answering. This strategy can obtain question-relevant region-level objects in the image.
- DCN [18] proposes a differentiated attention mechanism to solve the inconsistency between the attention regions of previous methods and human attention regions.

Methods based on modal interaction:

- DCA [17] proposes a dense co-attention layer. The dense co-attention layer improves the representation of fused features by considering a dense symmetric interaction between the input image features and the problem features.
- CAQT [19] obtains the interaction features between modalities through a common attention mechanism and introduces problem categories in the fusion stage.
- SOMA [41] proposes a second order-enhanced multi-glimpse attention model, which utilizes a second order module to accurately model the interaction between question features and co-embedded features in multi-glimpse outputs.
- ATCG [39] proposes a multi-step attention mechanism, which allows the model to gradually adjust its attention to image regions guided by the question features.
- ALMA [40] uses the siamese similarity learning method to achieve multimodal attention between images and text. Furthermore, an adversarial learning mechanism is introduced so that the learned multimodal features contain answer-related information.
- UFSCAN [35] proposes a feature-wise attention mechanism. This mechanism provides more discriminative features for the representation of image and question features by suppressing irrelevant features and emphasizing informative features.

Results on VQA v1 From Table 1, we can see that our model achieves an overall accuracy of 69.74% and 69.86% on the test-dev and test-std of the VQA v1 dataset, respectively, which are higher than all the comparison algorithms. We can also see that the methods MLB [38] and DCN [18] using the

attention mechanism perform better than the LSTM Q+I [1] and DPPnet [23] based purely on the CNN+LSTM structure but are weaker than the methods based on modal interaction. Our model belongs to the modal interaction-based approach. Our performance is 0.27% and 0.22% higher than the best method among the comparison algorithms, ATCG [39], on test-dev and test-std, respectively. In addition, our performance is also much higher than other methods based on modal interaction in the comparison algorithm, such as 2.85% higher than DCA [17] and 3.37% higher than CAQT [19] on the test-dev dataset.

Results on VQA v2 As can be seen from Table 2, our model outperforms all contrasting algorithms, achieving a new state-of-the-art performance of up to 70.03% on test-dev set and 70.28% on test-std set, respectively. The first three methods in Table 2 are all based on CNN+LSTM structure and use simple addition, multiplication, and splicing operations to complete feature fusion. On the test-dev set and test-std set, the performance of our model is significantly higher than the above three methods. Therefore, the attention mechanism and feature fusion strategy have a great impact on the overall performance of the model. The fourth, fifth, and sixth methods use the attention mechanism but only filter image features. Our model outperforms these methods and outperforms Adelaide Model+detector [10] by 4.22% on the test-dev set and 4.41% on the test-std set without augmentation with the VG dataset. This is because the attention mechanism in this paper considers the attention of the question to the image and considers the attention of the image to the question and realizes the information interaction between modalities. In Table 2, DCA [17], CAQT [19], SOMA [41], ATCG [39], ALMA [40], UFSCAN [35] all use attention mechanism to realize the implicit information exchange between modalities. It can be seen that these methods outperform the previous ones, which validates the importance of inter-modal interactions in visual question answering.

For results on test-dev set, Our model performs better than other methods mentioned above that realize the information interaction between modalities, which are 0.41%, 1.42%, 3.17% and 4.42% higher than those of ATCG [39], ALMA [40], CAQT [19] and DCA [17], respectively. For the UFSCAN [35] method additionally using the VG dataset, our model also outperforms by 0.2% on the test-dev set. Besides, for test-std set, our model outperforms DCA [17] by 4%, CAQT [19] by 3.55%, SOMA [41] by 1.42%, and ATCG [39] by 0.51%, respectively. It is worth mentioning that our model achieves the same effect as the UFSCAN [35] method without additional use of the VG dataset. With the VG dataset, our model outperforms UFSCAN [35] by 0.19% on the test-std set.

Table 3 Model ablation experiment results on VQA v2 validation set. Model is trained using only the training set and tested on the validation set

Models	Overall	Yes/No	Number	Other
$Model_{baseline}$	54.57	69.62	36.31	47.93
$Model_{baseline+att}$	55.40	69.87	37.08	49.24
$Model_{baseline+mutual}$	62.14	79.07	42.34	54.51
$Model_{baseline+fusion}$	59.08	77.60	39.52	50.16
$Model_{baseline+multi-mesh}$	65.40	83.31	45.31	57.10
$Model_{baseline+transformer}$	64.25	82.30	43.32	56.06
$Model_{full}$	66.31	83.94	48.69	57.57

4.5 Ablation Study

Our proposed multi-level mesh mutual attention model comprises multiple modules. To explore the individual effects of the proposed module, we build a baseline model, incrementally add the proposed modules, and evaluate the effect of each module on the VQA v2 validation set.

For baseline model $Model_{baseline}$, encoder keeps one multi-layer perceptron, and decoder keeps one multi-head self-attention layer and one multi-layer linear perceptron. Question features and image features are simply added and fed into the answer classifier. Based on the baseline model, we gradually add the proposed modules. $Model_{baseline+att}$ represents adding one multi-head self-attention layer only in the encoder. $Model_{baseline+mutual}$ stands for adding a mutual attention module to the decoder. $Model_{baseline+fusion}$ represents the use of multi-level fusion modules in the fusion stage. $Model_{baseline+multi-mesh}$ represents the use of multi-level mesh connection. In this model, encoder consists of two stacked multi-head self-attention layers. A mutual attention layer is included in the decoder. $Model_{baseline+transformer}$ represents using the connection method in the traditional transformer, only using the information of the last layer in the encoder. Other settings of the model remain the same as $Model_{baseline+multi-mesh}$. $Model_{full}$ represents our final population model.

Table 3 demonstrates the results of the ablation experiments. The addition of each proposed module over the baseline model improves the model effect, which confirms the effectiveness of the proposed module. The data in the first row and second row in Table 3 show $Model_{baseline+att}$ is 0.83% higher than $Model_{baseline}$, which means that adding a self-attention module on the encoder side and mining the relationship between words in the question information is conducive to improving the effect. The results in the third and fourth rows in Table 3 show that after adding the mutual attention module and the multi-level fusion module,

the model effect is improved by 7.57% and 4.51%, respectively, over the baseline model. This fully demonstrates the importance of information interaction between modalities and multi-scale modal information fusion in visual question answering. Furthermore, the mutual attention module and multi-level fusion module are also verified. Compared with the connection method of traditional Transformer in the field of machine translation, $Model_{baseline+multi-mesh}$ is 1.15% higher than $Model_{baseline+transformer}$. The multi-level mesh connection method has significantly improved, which verifies our proposed assumption. At the same time, using the low-dimensional and high-dimensional question information in the encoder for modal interaction, we can obtain better feature representation and improve the model effect. In the last row of Table 3, the effect of our overall model is still improved, which indicates that the proposed modules play a positive role in promoting each other.

4.6 Quantitative Analysis

To further explore the effect of the multi-level mesh decoder, we quantitatively analyze the decoding layers and mesh multi-level features in the decoder, respectively.

Table 4 demonstrates the overall effect of different multi-level question features connection methods under other numbers of decoding layers. The rows represent different mesh connection methods. "only 1" means to use only the last layer of question features in the multi-level question features for mesh interaction. "2-to-1" and "4-to-1" respectively indicate using two-layer question features and four-layer question features in the multi-level question features for mesh interaction at the decoder side. "DLayer-1", "DLayer-2", and "DLayer-4" indicate the use of one, two, and four decoding layers, respectively.

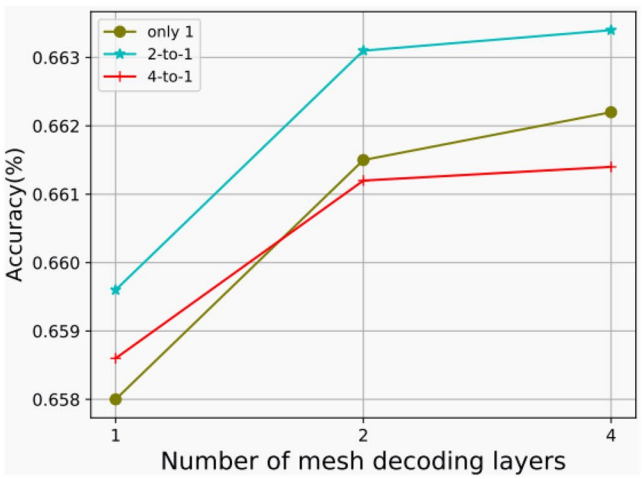
From Table 4, we can see that when the number of decoding layers is fixed, as the number of multi-level question features increases, the overall effect of the model first increases and then decreases. The model achieves the best overall performance in "2-to-1" when using two layers of question features for decoder-side mesh interactions. This means that using multi-level question features containing different levels of question information helps to improve the correct rate of question answering. However, the multi-level question features with too many levels have little improvement even a slight decrease in the model effect. This is because different levels of question features focus on inconsistent objects, and there is partially redundant noise information for the final question answer.

When the feature connection mode of the multi-level question is fixed, the number of decoding layers increases and the model's overall effect is improved. When the number of decoding layers is increased from one layer to two layers, the impact of the model is greatly enhanced. When the number of decoding layers is increased from two layers to four layers, the effect of the model is minimal, and the bottleneck of the model has been reached. This means that the decoding layer can reasonably complete the mesh interaction between modalities, further improving the efficiency of the model.

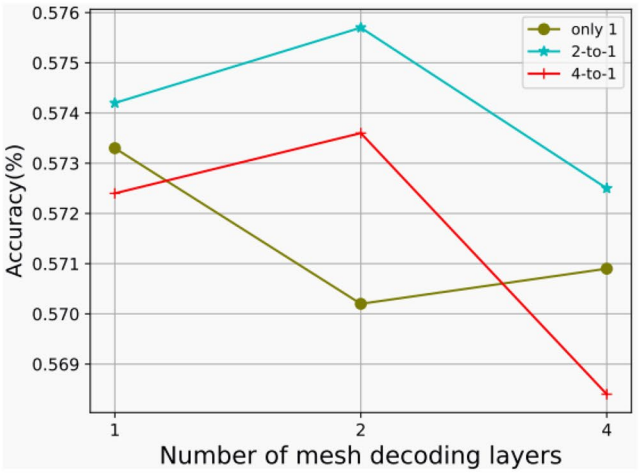
Figure 6 shows the details of the model performance with different levels of multi-level question features and different decoding layers. From Fig. 6c, we can see that the connection method that only uses the last layer of multi-level question features has a higher correct rate on the question type "Yes/No" than the "2-to-1" and "4-to-1" connection methods. The situation is just the opposite in Fig. 6b, d. Among the question types, the "Yes/No" question type is linguistically biased, often does not require reasoning about pictures and questions, and has a 50% chance of being correct for random answers. The correct answer to the "Other" and "Num" type questions needs to find the relevant attributes between the picture and the question and make multiple inferences to get it. Therefore, the above cases also further demonstrate the effectiveness of our proposed multi-level features and mesh decoder for answering questions. An interesting phenomenon is that in Fig. 6a, as the number of decoding layers increases, the overall effect of the "only-1" connection method is greater than that of the "4-to-1" connection method. This also verifies that different levels of question features focus on inconsistent objects, and there is noise redundant information.

5 Conclusion

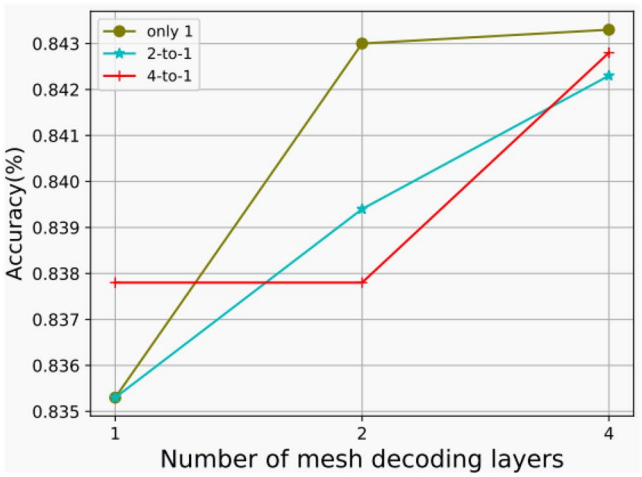
This paper proposes a multi-level mesh mutual attention model for visual question answering. The multi-level mesh mutual attention model utilizes mutual attention to fully explore the information interaction between visual and language modalities and improve the model efficiency. The model cleverly uses a multi-level mesh connection and at the same time utilizes low-dimensional and high-dimensional question information at different levels, providing more feature information for modal interactions. Besides, an adaptive multi-scale feature fusion module is designed to mine abstract information in fused features at different scales in the fusion stage. We perform comparative experiments on VQA v1 and VQA v2 datasets. The results verify the significance of our proposed module and the performance of the overall model, respectively. As for future work, we intend to consider introducing inference mechanisms, such as causal inference, graph neural networks, etc., to establish more complex relationships and reasoning between modalities, so as to improve the accuracy about answering questions.



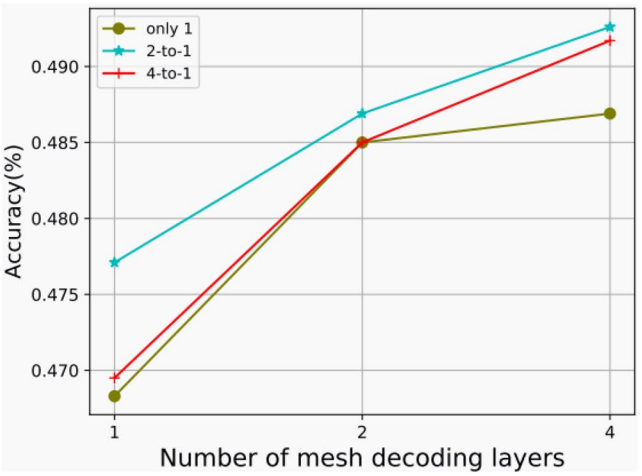
(a) Overall



(b) Other



(c) Yes/no



(d) Num

Fig. 6 Details of the accuracy of the model with different levels of multi-level question features and different numbers of decoding layers. There are three types of questions: binary choice questions "Yes/No", counting reasoning questions "Num" and other types of reasoning questions "Other". "Overall" represents the overall model accuracy. The question features are connected in three ways: "only-1", "2-to-1" and "4-to-1". The number of decoding layers is 1, 2, and 4, respectively

Table 4 Results of models with different levels of multi-level question features and different numbers of decoding layers on the VQA v2 val dataset

Multi-mesh	DLayer-1	DLayer-2	DLayer-4
Only 1	65.80	66.15	66.22
2-to-1	**65.96**	**66.31**	**66.34**
4-to-1	65.86	66.12	66.14

Bold value is the best among all the methods

Acknowledgements This work was supported by Research Fund of Guangxi Key Lab of MultisourceInformation Mining & Security (Grant Number: MIMS20-04, 20-A-01-02)and the Innovation Project of Guangxi Graduate Education (Grant Number:YCSW2022124).

Authors' contributions ZL: Conceptualization, Methodology, Software; GZ: Data Curation; LW: Visualization; KZ: Validation; RL: Reviewing and Editing.

Data availability All data generated or analysed during this study are included in this publishedarticle.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
2. Weiss M, Chamorro S, Girgis R, Luck M, Kahou SE, Cohen JP, Nowrouzezahrai D, Precup D, Golemo F, Pal C (2020) Navigation agents for the visually impaired: a sidewalk simulator and experiments. In: Conference on robot learning. PMLR, pp 1314–1327
3. Bghiel A, Dahdouh Y, Allaouzi I, Ben Ahmed M, Anouar Boudhir A (2019) Visual question answering system for identifying medical images attributes. In: The proceedings of the third international conference on smart city applications. Springer, pp 483–492
4. Malinowski M, Rohrbach M, Fritz M (2015) Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision, pp 1–9
5. Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R (2015) Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167
6. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical co-attention for visual question answering. In: Advances in neural information processing systems (NIPS) 2
7. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. In: Advances in neural information processing systems 31
8. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847
9. Ma Y, Lu T, Wu Y (2021) Multi-scale relational reasoning with regional attention for visual question answering. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 5642–5649
10. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) 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
11. Zhu C, Zhao Y, Huang S, Tu K, Ma Y (2017) Structured attentions for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 1291–1300
12. Shih KJ, Singh S, Hoiem D (2016) Where to look: focus regions for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4613–4621
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30
14. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
16. Yu W, Luo M, Zhou P, Si C, Zhou Y, Wang X, Feng J, Yan S (2021) Metaformer is actually what you need for vision. arXiv preprint arXiv:2111.11418
17. Nguyen D-K, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6087–6096
18. Patro B, Namboodiri VP (2018) Differential attention for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7680–7688
19. Yang C, Jiang M, Jiang B, Zhou W, Li K (2019) Co-attention network with question type for visual question answering. IEEE Access 7:40771–40781
20. Malinowski M, Fritz M (2014) A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in neural information processing systems 27
21. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) 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
22. Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W (2015) Are you talking to a machine? Dataset and methods for multilingual image question. In: Advances in neural information processing systems 28
23. Noh H, Seo P.H, Han B (2016) Image question answering using convolutional neural network with dynamic parameter prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 30–38
24. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems 28
25. Lu P, Li H, Zhang W, Wang J, Wang X (2018) Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
26. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
27. Xiong C, Merity S, Socher R (2016) Dynamic memory networks for visual and textual question answering. In: International conference on machine learning. PMLR, pp 2397–2406
28. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299–307
29. Teney D, Anderson P, He X, Van Den Hengel A (2018) Tips and tricks for visual question answering: learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4223–4232
30. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
31. Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) 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

32. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
33. Lei Ba J, Kiros JR, Hinton GE (2016) Layer normalization. arXiv e-prints, 1607
34. Lin TY, Maire M, Belongie S, Hays J, Zitnick CL (2014) Microsoft coco: common objects in context. Springer, pp 740–755
35. Zhang S, Chen M, Chen J, Zou F, Li Y-F, Lu P (2021) Multimodal feature-wise co-attention method for visual question answering. Inf Fusion 73:1–10
36. Thomee B, Elizalde B, Shamma DA, Ni K, Friedland G, Poland D, Borth D, Li LJ (2016) Yfcc100m: the new data in multimedia research. Commun ACM 59(2):64–73
37. Kafle K, Kanan C (2017) Visual question answering: datasets, algorithms, and future challenges. Comput Vis Image Underst 163:3–20
38. Kim J-H, On KW, Lim W, Kim J, Ha J-W, Zhang B-T (2017) Hadamard Product for Low-rank Bilinear Pooling. In: The 5th international conference on learning representations
39. Li W, Sun J, Liu G, Zhao L, Fang X (2020) Visual question answering with attention transfer and a cross-modal gating mechanism. Pattern Recognit Lett 133:334–340
40. Liu Y, Zhang X, Huang F, Cheng L, Li Z (2020) Adversarial learning with multi-modal attention for visual question answering. IEEE Trans Neural Netw Learn Syst 32(9):3894–3908
41. Sun Q, Xie B, Fu Y (2020) Second order enhanced multi-glimpse attention in visual question answering. In: Proceedings of the Asian conference on computer vision