PAPER *Special Section on Deep Learning Technologies: Architecture, Optimization, Techniques, and Applications*

A Visual Question Answering Network Merging High- and Low-Level Semantic Information

Huimin LI[†a)], **Dezhi HAN**[†b)], **Chongqing CHEN**[†c)], ***Nonmembers***, **Chin-Chen CHANG**[††d)], ***Member***,
Kuan-Ching LI[†††e)], ***and*** **Dun LI**[†f)], ***Nonmembers***

SUMMARY Visual Question Answering *(VQA)* usually uses deep attention mechanisms to learn fine-grained visual content of images and textual content of questions. However, the deep attention mechanism can only learn high-level semantic information while ignoring the impact of the low-level semantic information on answer prediction. For such, we design a High- and Low-Level Semantic Information Network *(HLSIN)*, which employs two strategies to achieve the fusion of high-level semantic information and low-level semantic information. Adaptive weight learning is taken as the first strategy to allow different levels of semantic information to learn weights separately. The gate-sum mechanism is used as the second to suppress invalid information in various levels of information and fuse valid information. On the benchmark *VQA-v2* dataset, we quantitatively and qualitatively evaluate *HLSIN* and conduct extensive ablation studies to explore the reasons behind *HLSIN's* effectiveness. Experimental results demonstrate that *HLSIN* significantly outperforms the previous state-of-the-art, with an overall accuracy of 70.93% on test-dev.

key words: *Visual Question Answering (VQA), deep attention mechanisms, adaptive weight learning,gate-sum mechanism*

1. Introduction

The development of artificial intelligence has accelerated the advancement of technologies related to computer vision and natural language processing. With the two mature, invloving visual and language came into being multimodal task. In contrast to single-modal tasks, multimodal tasks require extracting and understanding information from a single modality, which combining information from two different modalities for reasoning. Although this is challenging, current researchers have achieved many multimodal tasks, for instance, image-text matching [1], [2], image captioning [3], [4], and *VQA* [5], [6], [26]. As a typical representative of multimodal tasks, *VQA* requires understanding visual information and image information; what's more, combining the two to reason about the answer. Thus, *VQA* requires that the model must combine another modal information when it understands single-modal information.

Inspired by this concept, many deep neural networks based on co-attention have been applied to *VQA*. The dual attention networks *(DANS)* proposed in [1] collect the necessary information from the two feature vectors by focusing on specific question words and corresponding image regions. Kim et al. [2] presented the Bilinear Attention Network *(BAN)*, which generates the attention graph by calculating the bilinear interaction between each pair of images; Moreover, from the fusion features of the attention map, the final joint representation of the question feature is obtained. Although shallow co-attention can highlight important visual features and textual features, it lacks deep fine-grained multimodal interaction.

To solve this problem, some deep attention mechanisms [7], [8] have been proposed and widely used. The deep attention mechanism usually pays more attention to feature interaction within and between modalities in multimodal tasks. However, in the process of multimodal interaction, the deep attention mechanism often pays attention to the impact of high-level semantic information on the output result, thereby ignoring the impact of low-level semantic information on the output result.

We adopt two strategies to achieve the fusion of high- and low-level semantic information while maintaining the advantage of the deep attention mechanism. Will it be more beneficial to reasoning about answers? Based on this idea, we present a High- and Low-Level Semantic Information Network *(HLSIN)*, which employs two strategies to achieve the fusion of high-level semantic information and low-level semantic information.

In summary, our contributions in this article are as follows:

- We design a High- and Low-Level Semantic Information Network *(HLSIN)*, which employs two strategies to achieve the fusion of high-level semantic information and low-level semantic information—using adaptive weight learning as the first strategy to allow different levels of semantic information to learn weights separately; The gate-sum mechanism is used as the second to suppress invalid information in various levels of information and fuse valid information.
- The experimental results on the *VQA-v2.0* dataset

Manuscript received September 13, 2021.
Manuscript revised November 5, 2021.
Manuscript publicized January 6, 2022.

[†]The authors are with the Department of Computer, Shanghai Maritime University, Shanghai 201306, China.

[††]The author is with Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan.

[†††]The author is with the Dept. of Computer Science and Information Engineering, Providence University, Taiwan.

a) E-mail: lygsmartcar@163.com
b) E-mail: dzhan@shmtu.edu.cn
c) E-mail: chenchongqing0305@163.com
d) E-mail: webadmin@fcu.edu.tw
e) E-mail: kuancli@pu.edu.tw
f) E-mail: lidumshmtu@outlook

DOI: 10.1587/transinf.2022DLP0002

prove the effectiveness of *HLSIN* under the two fusion strategies. The accuracy of *HLSIN* on test-dev is 70.93%, and the accuracy on test-std is 71.33%.

The remaining of the article is organized as follows. Section 2 gives an overview of the related work in relative areas. In Sect. 3, the presented methods are discussed in detail, include Traditional Transformer Structure, The core Structure of *MCAN* and *HLSIN*, The Overall Structure of *HLSIN*. Section 4 shows the results of our experiments using several comparison methods. Finally, Sect. 5 concludes the article.

2. Related Work

In this section, we will give a exhaustively introduction about the related works in attention mechanism (Sect. 2.1), deep attention mechanism based on high-level semantic information (Sect. 2.2), gate mechanism (Sect. 2.3), and method of modal feature fusion (Sect. 2.4).

2.1 Attention Mechanism

Attention mechanism has become an ordinary operation in multimodal systems. The use of attention mechanisms improves the performance of multimodal tasks. It can use the different variants of attention mechanisms to adaptively select the essential features and enhance the accuracy of *VQA*. Xu et al. [9] presented the soft and hard attention mechanism as the mainstream method for *VQA*. Then, Yang et al. [10] presented a stacked attention network, which generates multiple attention maps on the image and is gradually aborbed in the most critical visual regions. Lu et al. [11] proposed a co-attention mechanism that focuses on image regions and questions and learns their attention weights to interact between the two modes. On this basis, Nguyen et al. [12] presented a closely connected *VQA* co-attention mechanism that focuses on image regions and question feature through multiple steps. At present, the most popular framework is the transformer. The *Bert* [13] model is the first transformer model to reach a human-level framework through the self-attention mechanism, and the relationship between words in question modes is modeled to learn the most advanced word embedding [14], [16].

2.2 Deep Attention Mechanism Based on High-Level Semantic Information

In the process of multimodal interaction, the deep attention mechanism based on high-level semantic information is mainly used to study the common fusion intra- and intermodality. Gao et al. [17] believe that each model complements the other and proposed a *DFAF* model that includes the co-attention of different modes as well as within the self-attention. For image modalities, each image region should obtain information from the lexicality of the question word and the corresponding image regions. Subsequently, Gao et al. [33] proposed the *MLIN* (Multimodality Later Interaction Network) model structure. Compared with previous models, the *MLIN* model can extract features from many individual visual word pairs and multimodal potential summary vectors, thus capturing high-level visual-linguistic interactions with a smaller modal capacity. A deep Modular Co-attention NetWork *(MCAN)* was presented by Yu et al. [11]. Based on the previous deep collaborative attention models, a dense self-attention model is constructed in each mode to understand the relationship between regions and words in the image, which further enriches the feature representation of the image and the problem.

2.3 Gate Mechanism

In the process of multimodal task interaction, there may be irrelevant or noisy features that hurt the interaction process, resulting in the inaccuracy of output results. Therefore, to effectively solve such problems, the gate mechanism is proposed in the process of multimodal interaction. In the *MUAN* model [19], a gating model based on the low-rank bilinear pool is designed to reweight the Query Matrix Q and Key Matrix K features before matching the point product. Its immediate implementation is to carry out the element product between Q and K. In addition to multi-head self-attention in the *NMT* [20] model, there is a block that performs multi-head attention on the output of the contextual encoder stack. Although the traditional *VQA* model using the gate mechanism in multimodal interaction mechanism, most of the models in the process of interaction only paid attention to the high-level semantic information integration and ignored the low-level semantic information to predict the answers, therefore, in the process of multimodal task fusion, gate-sum attention mechanism is adopted in this article to fuse the features of high-level and low-level semantic information.

2.4 Method of Modal Feature Fusion

Feature fusion refers to the fusion of visual features and textual features at the feature level. The core of feature fusion is to fuse cross-modal information. With computer vision and natural language processing rapidly developing in deep learning, monomode representation has also made significant progress. Monomode [21] has different meanings in different semantic spaces. The *VQA* task is essentially a multimodal reasoning task, so it is of great significance to effectively integrate information for multimodal reasoning [22], [24]. A bilinear fusion method has been adopted in recent research to improve the ability of cross-modal fusion. The bilinear fusion method considers the relationship between visual feature elements and textual feature elements. However, the direct modelling of these two correlations will produce a square scale of parameters, so the performance of bilinear fusion is often limited by computational machine resources. Later, Fukui et al. [25] presented a Multimodal Compact Bilinear *(MCB)* Pooling method, but the *MCB* Pooling method still requires higher dimensions to ensure robustness. To solve this problem, Kim et al. [27] proposed a Multimodal Low-rank Bilinear *(MLB)* Pooling method, which is based on the matrix *Hadamard* product [28] to calculate the two eigenvectors. Still, the Multimodal Low-rank Bilinear *(MLB)* Pooling has low dimensionality and fewer parameters, and it is susceptible to parameters and converges slowly. So at the later stage, the Multimodal Factorization Bilinear *(MFB)* [29] set and Multimodal sum Factorization High-order *(MFH)* [22] set were proposed, which achieved better results.

3. Our Method

In this section, we present the overall structure of our proposed method in Fig. 1. To better introduce the structure of *HLSIN*, we will specifically introduce *HLSIN* from the four-module.

3.1 The Structure of Traditional Transformer

Figure 2 displays the basic structure of the Transformer, which mainly consists of two parts: *Encoder* and *Decoder*. Furthermore, the *Encoder* and *Decoder* are composed of N stacked layers. And each stacked layer includes a *Multi-Head Attention (MHA)* unit and a *Feed Forward (FF)* unit. The difference is that the *Decoder* also consists of a *Masked MHA* unit for mask marking. Moreover, each unit is followed by a residual connection [30] and layer normalization [31] for optimization. In this section, following [35], we give a exhaustive introduction to the relevant contents of the Transformer.

3.1.1 Multi-Head Attention

The *Multi-Head* attention mechanism is designed to strengthen the characterization ability of features. First, it projects the Query Matrix (Q), Key Matrix (K) and Value Matrix (V) into H sub-query matrices, sub-key matrices and sub-value matrices of the same dimension, respectively. Then, the attention operation is performed separately within each head. Finally, the output within each head is spliced to produce the final feature. We can use the formlula (1) to present the calculation process:

$$\begin{aligned} F &= MH(Q, K, V) \\ &= \mathrm{Concat}\left(\mathrm{head}_1, \ldots\ldots, \mathrm{head}_h\right) W^0, \qquad (1) \end{aligned}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \qquad (2)$$

$$= \mathrm{SoftMax}\left(\frac{QW_i^Q (KW_i^K)^T}{\sqrt{d}}\right) VW_i^V. \qquad (3)$$

Where $W^0 \in R^{d \times hd_h}$ is the projection matrix for all heads, $Concat(\cdot)$ represents concatenation of all heads, $W_i^Q \in R^{d \times d_h}$, $W_i^K \in R^{d \times d_h}$, $W_i^V \in R^{d \times d_h}$ are the projection matrixes for i-th head. $F \in R^{n \times d}$ is the output features.

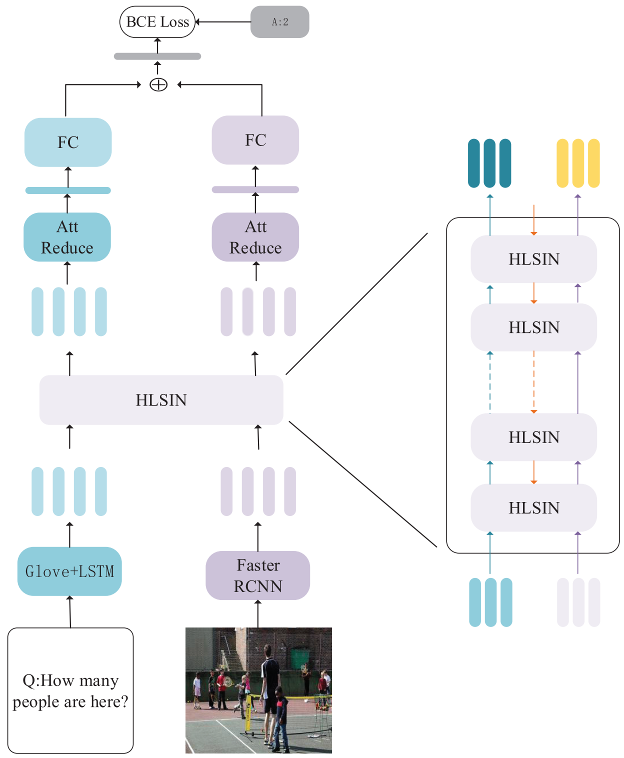


Fig. 1 The structure of HLSIN

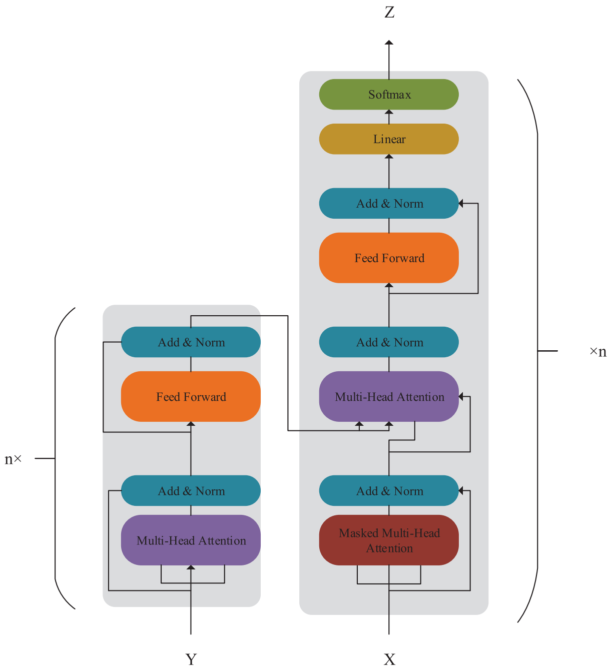
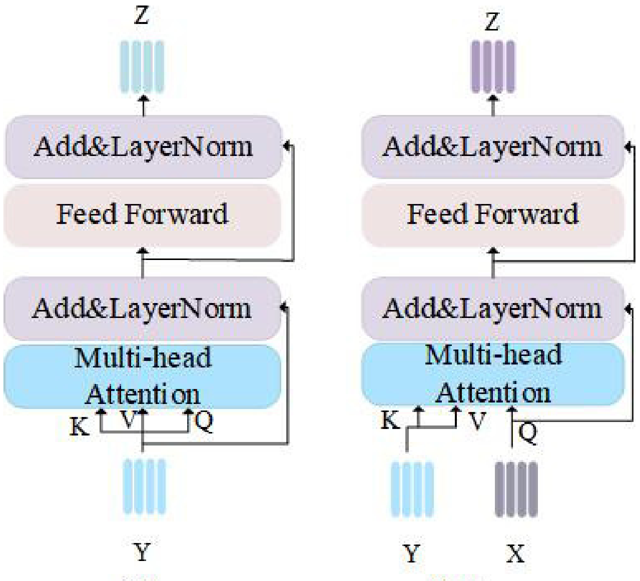


Fig. 2 The structure of Transformer

3.1.2 Feed Forward Network

Apart from *MHA* in Transformer's *Encoder-Decoder* module, each layer is also composed of a fully connected *feed forward* network. The network is composed of two linear



a)SA *b)GA*

Fig. 3 The different attention units with multi-head attention as the different types of inputs.

transformations and a *Relu* [31] activation function. Assuming that the input feature of the *feed forward* network can be represented by $X = [x_1, x_2, \ldots\ldots, x_n] \in R^{n\times d}$, thus, we can use the formula (4) to present the output feature:

$$FFN(X) = max(0, XW_1 + b_1)W_2 + b_2. \tag{4}$$

3.2 The Core Structure of MCAN

Inspired by Transformer, the deep attention model *MCAN* due on high-level semantic information was presented and won the 2019 *VQA* Challenge championship in one fell swoop. Essentially, *MCAN* is applying the standard Transformer architecture in *VQA* tasks. The deep co-attention part used for multimodal information interaction uses *Encoder-Decoder* as the core architecture. In this section, we present the core content of *MCAN* exhaustively.

3.2.1 Self-Attention (SA) Unit and Guide-Attention (GA) Unit

Figure 3 shows two core units designed in *MCAN*, each of which has the same composition as a layer of *Encoder* in Transformer. According to the different inputs, the *SA* unit and the *GA* unit are respectively defined to simulate the intra- and inter-modality attention relationship.

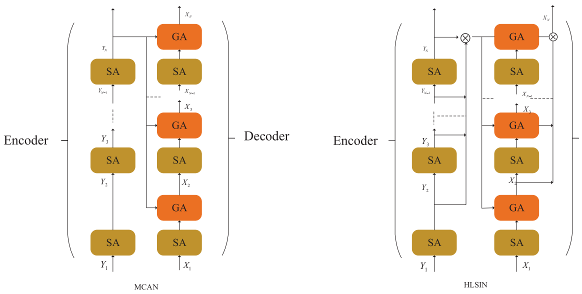


Fig. 4 The encoder-decoder frame of MCAN and HLSIN.

3.2.2 Encoder in MCAN

In fact, the *Encoder* in *MCAN* is the same as that of the *Encoder* in the Transformer. As shown in the left part of Fig. 4 (a), the *Encoder* of *MCAN* is composed of a profoundly stacking of N layers *SA* units, which is used to simulate the intra-modality information interaction of problem words. Specifically, take the question feature $Y = Y_0 = [y_1, y_2, \ldots\ldots, y_n] \in R^{n\times 512}$ as the initial input. After each *SA* unit, an *MHA* operation and an *FF* operation are performed. After N layers, the interactive question feature $Y_N = \left[y_1^N, y_2^N, \ldots\ldots, y_n^N\right] \in R^{n\times 512}$ is obtained. The derivation process is represented by formula (5):

$$Y_i = \mathrm{SA}^{[i]}(Y_0). \tag{5}$$

Where $i \in [1, N]$ represents the stacking amount of *SA* units. $Y_i \in R^{n\times 512}$ stands for the question feature of the output of the i-th layer.

3.2.3 Decoder in MCAN

Similar to the *Decoder* structure in Transformer, the *Decoder* in *MCAN* comprises a combination of *SA* unit and *GA* unit (defined as *SGA* unit in *MCAN*) through *N* deep stacking. The *SA* unit is used to simulate the intra-modality information interaction of the image region, and the *GA* unit affects the inter-modality information interaction between the question word and the image region. As expressed in the right part of Fig. 4 (a), the initial image feature $X = X_0 = [x_1, x_2, \ldots\ldots, x_n] \in R^{n\times 512}$ and the question feature obtained in the above *Encoder* is used as two inputs, and the final image feature $X_N = \left[x_1^N, x_2^N, \ldots\ldots, x_n^N\right] \in R^{n\times 512}$ is obtained after N-layers of *SGA* units. We can use the formula (6) to present the process:

$$X_i = \mathrm{SGA}^{[i]}(X_0, Y_i). \tag{6}$$

Where $i \in [1, N]$ represents the stacking number of *SGA* units. $X_i \in R^{n\times 512}$ is the question feature of the output of the i-th layer.

Obviously, in *MCAN*, deep co-attention based on the *Encoder-Decoder* is used to simulate the intra- and inter-modality interaction of problem-image. In addition, the model will obtain the high-level semantic information of the question and image for the subsequent fusion and prediction classification. The results in the *VQA* task show that the attention model based on *Encoder-Decoder* architecture is effective. However, from the overall framework of Fig. 4 (a), it is easy to understand the features used for classification prediction are only the last high-level semantic information. Objectively, the intermediate process is a black-box model. Although features can be made more fine-grained by stacking layers, they also bring more noise. It is assumed that low-level semantic information can also play a positive role in classification prediction. Therefore, to maintain the advantages of *Encoder-Decoder* architecture, it is urgent to find a model that can effectively utilize the high- and low-level semantic information.

3.3 The Core Structure of HLSIN

On the premise of maintaining the advantages of *MCAN*, we designed a network model—*HLSIN*, which integrates high- and low-level semantic information and is based on the *Encoder-Decoder* framework of *MCAN*. Its main structure is shown in Fig. 4 (b). To realize the fusion of high- and low-level semantic information, *HLSIN* usually adopts two different methods to fuse information. We adopt adaptive weight learning as the first strategy to allow different levels of semantic information to learn weights separately. The gate-sum mechanism is used as the second to suppress invalid information in various levels of information and fuse valid information. The following will exactly introduce the control methods of these two attention mechanisms.

(1) The adaptive weight control mode can be expressed by formulas (7) and (8):

$$\delta_i = Sigmoid(W_i[\overline{T_i} + T_i] + b_i), \tag{7}$$

$$T = \sum_{i=1}^{N}(\delta_i \bigotimes \overline{T_i} + (1 - \delta_i) \bigotimes T_i). \tag{8}$$

Where $\bigotimes$ represents element multiplication. We use $i \in [1, N]$ to express the stacking amount of *SA (SGA)* units in the *Encoder (Decoder)*. The question feature (image feature) of the previous N layer in the *Encoder (Decoder)* is represented by $\overline{T_i}$. T_i represents the question feature (image feature) of the N-th layer.δ_i represents the weight of the correlation coefficient, the weight of the correlation coefficient is obtained by Sigmoid() function. T represents the output result of the final question feature (image feature) obtained by summing the question feature (image feature) of the former N-1 layer and the question feature (image feature) of the N-th layer through the accumulation of adaptive weights.

(2) The gate-sum mechanism is used to suppress invalid information, which is expressed in detail by formulas (9), (10), (11), (12), and (13):

$$T_{ia} = \mathrm{W}_i^a(T_i^a + \overline{T_i^a}) + b_i^a, \tag{9}$$

$$T_{ic} = \mathrm{W}_i^c(T_i^c + \overline{T_i^c}) + b_i^c, \tag{10}$$

$$T_{\overline{ia}} = Sigmoid(T_{ia}), \tag{11}$$

$$T_{\overline{ic}} = Sigmoid(T_{ic}), \tag{12}$$

$$T = \sum_{i=1}^{N}(T_{\overline{ia}} * T_{ia} + T_{ic} * T_{\overline{ic}}). \tag{13}$$

Where we use $i \in [1, N]$ to represent the stacking number of *SA (SGA)* units in the *Encoder (Decoder)*. $\overline{T_i^a}$ is the question feature of the previous N-1 layer in the *Encoder*, $\overline{T_i^c}$ represents the image feature of previous N-1 layer in the *Decoder*. T_i^a is the question feature of the N-layer in the *Encoder*, T_i^c is the image feature of the N-layer in the *Decoder*. T_{ia}, T_{ic}represents the question of the i-th layer or the relevant features of the image after linear operation; $T_{\overline{ia}}, T_{\overline{ic}}$ stands for the Sigmoid() function. W_i^a, W_i^c denote the weight coefficients when performing linear operations and b_i^a, b_i^c represent the deviation vector. T represents the final output result of the gate-sum suppression of non-keywords.

3.4 The Overall Structure of HLSIN

3.4.1 Multimodal Feature Extraction

The High- and Low-Level Semantic Information Network *(HLSIN)* mainly extracts multimodal tasks features from two aspects: images and questions. The structure of the *HLSIN* is primarily shown in Fig. 1. We usually use the bottom-up attention module [3] to pick up image features with input size $X \in R^{K\times2048}$, where $K \in [10, 100]$ indicates all amount of the target detection features. K is generally set at 100 to achieve better results during the experiment.

The input question words are defined as a maximum of 14 to improve feature extraction efficiency. With pre-trained *Glove* [36] for word embedding, and each word is initialized to a 300-dimensional feature vector. Finally, input the initialized word features into a 512-dimensional single-layer *LSTM* [37] network to obtain the question feature $Y \in R^{14\times d_y}$, where d_y represents the feature dimension of each problem.

3.4.2 High- and Low-Level Semantic Information Learn

HLSIN is based on the *Encoder-Decoder* framework of *MCAN*. To solve the problem, *MCAN* only focuses on the high-level semantic information in the feature extraction process and ignores low-level semantic information on answer prediction. Therefore, *HLSIN* is based on two different fusion strategies to achieve the fusion of features. One is to let different levels of semantic information learn the weights separately by the adaptive weight learning method; the other is to design a gate-sum mechanism to suppress the invalid information in different levels of information and fuse the valid information instead.

We are taking the encoder of *HLSIN* as an example to introduce *HLSIN* in detail. The encoder usually saves the output results of the previous N-1 layers of each $SA(\cdot)$ unit in $\overline{Y} = [Y_1, Y_2, \ldots\ldots, Y_{N-1}] \in R^{(n-1)\cdot512}$ and then converts $\overline{Y} \in R^{(n-1)\cdot512}$ to $\overline{Y} \in R^{1\times512}$ through linear operations. Finally, enter the final question feature Y_N into each $SGA(\cdot)$ unit of the *decoder* to realise the problem feature's guidance to the image feature.

The *Encoder-Decoder* module of *HLSIN* usually fuses the image-question features of the previous N-1 layer with the image-question features of the N-th layer. *HLSIN* usually uses two different fusion methods to fuse these features, and the results obtained by various fusion methods are also different. The specific experimental results and related parameters will be described in detail in the fourth part.

3.4.3 Multimodal Fusion

After extracting N-layer *HLSI* features, question feature *Y* and image feature *X* contain rich semantic information. Therefore, a two-layer *MLP (FC(4d)-ReLU-Drop(0.1)-FC(d))* is designed to reduce relevant information in the multimodal feature fusion mechanism. As the following example of question feature *Y*, the question feature obtained after the fusion of multimodal features is expressed by formulas (14) and (15):

$$\epsilon = SoftMax(MLP(Y_i)), \tag{14}$$

$$F_u = \sum_{j=1}^{N}(\epsilon_j Y_j^i). \tag{15}$$

Where $\epsilon = [\epsilon_1, \epsilon_2, \ldots\ldots, \epsilon_n] \in R^n$ is the weight learned through the $SoftMax(\cdot)$ function,and i is the number of layers stacked by HLSI layers,namely i = 7. By analogy, we can further use layer normalization to reduce the problem features obtained by the model, which can be expressed by formulas (16), (17) and (18):

$$Z = Linear(W_x^N F_u + W_y^N F_u), \tag{16}$$

$$C = Linear(Z), \tag{17}$$

$$A = Sigmoid(C). \tag{18}$$

Where $W_x^n, W_y^n \in R^{(d\times d_z)}$ is the *Linear* projection matrix of image features and question features, d_z is the common dimension of fusion features. $Linear(\cdot)$ is used to optimize training, and $Sigmoid(\cdot)$ is used for classification. According to [33], we use *BCE* as the *Loss* function to train the variety of answers.

4. Experiments and Results

This section first describes the datasets for evaluation in (Sect. 4.1) and hyperparameter settings in (Sect. 4.2). Then, We present ablation results in (Sect. 4.3). Finally, the feasibility analysis of HLSIN is carried out in (Sect. 4.4).

4.1 Datasets

The proposed *VQA* model was mainly evaluated and tested on two well-known datasets: *VQA v1.0* and *VQA v2.0*. Both datasets contain many open-ended questions and more than 50% of other types of questions.

VQA v1.0 The image data in the *VQA v1.0* dataset mainly comes from the *MSCOCO* [38] dataset. The data in this dataset is primarily split from the *MSCOCO* dataset, which contains 248349 training questions, 121512 verification questions, and 244302 test questions. In addition, the questions in the dataset are divided into three categories: binary (yes/no), number, and other. For each question, different annotators give different forms of answers.

VQA v2.0 The *VQA v2.0* dataset is an updated version of *VQA v1.0*, which pays more attention to linguistic bias and requires a more fine-grained recognition capability of the *VQA* model. Compared to the *VQA v1.0* dataset, the *VQA v2.0* dataset has a much larger data size and contains over 1.1 million *MSCOCO* image-based questions and 15 answer pairs. For two images with similar semantics, although each pair of images includes the same questions, the corresponding answer is different. To perform a better performance evaluation of *HLSIN*, we carry on experiments on *HLSIN* on the *VQA v2.0* dataset. In order to reduce the possibility of overfitting, the dataset usually uses 443757 pairs (images, questions, answers) for training, 214354 pairs for verification, and 447793 pairs for testing.

4.2 Parameter Setting Experiments

The hyperparameter settings of the model employed in the experiment are as follows. The input image features, the question features, and the fused multimodal features are set to $d_x = 2048$, $d_y = 512$, $d_z = 1024$, respectively. According to the previous work, in multi-head attention,the amount of heads is set at 8. As described in [11], the length of the questions is 14, and questions insufficient than 14 words are filled to 14. The number of *HLSI* layers is $N \in \{6, 7, 8\}$. In the experiments, we train all models with the same batch size for 13 rounds and then select the best training results.

4.3 Ablation Studies

Many ablation experiments were conducted on the *VQA v2.0* dataset to study the reasons for the effectiveness of *HLSIN*.

4.3.1 *HLSI* Variants

HLSIN employs two strategies to achieve the fusion of high- and low-level semantic information. Adaptive weight learning is used as the first strategy to allow different levels of semantic information to learn weights separately. The gatasum mechanism is used as the second to suppress invalid information in various levels of information and fuse valid information. From the results in Table 1, it can prove the information fusion of *HLSI* by these two methods is effective. By observation, we can find that the performance of *SA(Y)-GSGA(X, Y)* is significantly better than other fusion methods. Therefore, *SA(Y)-GSGA(X, Y)* is used as our default *HLSI* in the following experiments unless otherwise specified.What's more,SA (x)-GSGA (X, Y) in HLSI variant means that HLSIN adopts gate-sum mechanism in decoder to restrain invalid information and fuse valid information.

4.3.2 *HLSI* vs Depth

Since *MCAN* has achieved better experimental results when the number of layers is N=6, *HLSIN* is based on the Encoder-Decoder framework of *MCAN*, combines high-level semantic information with low-level semantic infor-

Table 1 Model accuracy of *HLSIN* under different *HLSI*. The first row represents three different variants to fuse information in *HLSI* employing adaptive weight. For example: WSA(X)-SGA(X,Y) indicates that HLSIN uses adaptive weight in decoder section; to suppress the invalid information using the gate-sum mechanism, we can use the next row to denote three different variants of *HLSI*. For example, GSA(x)-SGA(x,y) indicates that HLSIN uses the gate-sum mechanism in the decoder section; *WSA(X)-GSGA(X, Y)*, *GSGA(X)-WSGA(X, Y)* represent that two information fusion methods are used at the same time in *HLSI* to conducts information fusion.

Model	Test-dev				Test-std
	Y/N	Num	Other	All	All
WSA(X)-SGA(X,Y)	86.86	53.64	60.8	70.72	71.18
WSA(X)-SGA(X,Y)	86.96	53.13	60.96	70.78	71.13
WSA(X)-WSGA(X,Y)	86.52	53.22	60.99	70.66	70.96
GSA(X)-SGA(X,Y)	86.7	52.96	61.15	70.75	71.13
SA(X)-GSGA(X,Y)	**87.09**	**53.17**	**60.95**	**70.83**	**71.16**
GSA(X)-GSGA(X,Y)	86.55	52.88	61.04	70.62	71.22
WSA(X)-GSGA(X,Y)	86.73	53.3	60.83	70.64	70.92
GSGA(X)-WSGA(X,Y)	87.0	53.19	60.73	70.7	70.93

Table 2 From the results in Table 1, it can be found the performance of *SA(X)-GSGA(X, Y)* is significantly better than the performance of the other models. Therefore, we do experiments on the number of layers for *HLSIN* based on *SA(X)-GSGA(X, Y)*, where the number of layers $N \in \{6, 7, 8\}$

N	Test-dev				Test-std
	Y/N	Num	Other	All	All
6	87.09	53.17	60.95	70.83	71.16
7	**87.01**	**53.31**	**61.18**	**70.93**	**71.33**
8	86.98	53.43	60.92	70.8	71.23

Table 3 *HLSIN* integrates high-level semantic information with low-level semantic information based on *MCAN*. By comparing *HLSIN* with the most state-of-the-art model at present, we find the effectiveness of high- and low-level semantic information fusion.

Model	Test-dev				Test-std
	Y/N	Num	Other	All	All
BUTD [3]	81.82	44.21	56.05	65.32	65.67
MFH [22]	85.31	49.56	59.89	68.76	NULL
BAN [2]	85.42	50.93	60.26	69.52	NULL
DFAF [17]	86.09	53.32	60.49	70.22	70.34
MCAN [11]	86.82	53.26	60.72	70.63	70.9
MUAN [19]	86.77	54.4	60.89	70.82	71.1
MEDAN [39]	87.1	52.69	60.56	70.6	71.01
DCAN [40]	60.88	88.02	53.4	70.89	71.21
HLSIN-6(Ours)	**87.09**	**53.17**	**60.95**	**70.83**	**71.16**
HLSIN-7(Ours)	**87.01**	**53.31**	**61.18**	**70.93**	**71.33**

mation. Therefore, the research of the number of *HLSI* layers starts from N=6. With the amount of layers N increases, the gap of *HLSI* performance has appeared, too. From Table 2, we can find when N=7, *HLSIN* has the best performance.

4.3.3 Comparison with State-of-the-Art *VQA* Model

As shown in Table 3, we compare the best model, *HLSIN*, with the current state-of-the-art. MEDAN is introduced to capture rich and reasonable question features and image features. Compared with *MEDAN (Adam)* [39], the accuracy of *HLSIN* on test-std is improved by 0.32%. Next, to achieve fine-grained interaction between question words and image regions, *MCAN* model is proposed. By comparing the accuracy of *HLSIN* and the *MCAN*, we can easily find the accuracy of *HLSIN* on test-std has improved by 0.33%, and the accuracy on test-dev has improved by 0.3%. Therefore, we conjecture that the attention mechanism that combines high-level semantic information with low-level semantic information based on *MCAN* is instructive for the future development of *VQA*.

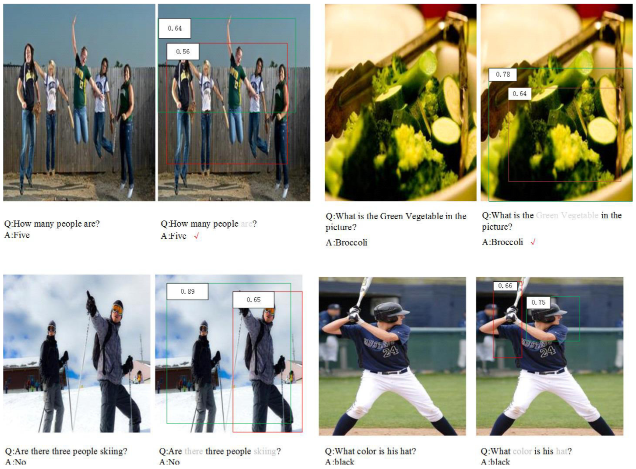


Fig. 5 Feasibility analysis.

4.4 Attention Visualization

As shown in Fig. 5, there are four examples randomly selected from the dataset. The first row shows an example where *HLSIN* and *MCAN* can choose the correct answer in the meanwhile, while the second row shows an example where *HLSIN* can choose the correct answer, but *MCAN* can not select the correct answer. The textual brightness represents a keyword in the question, and the content selected from the picture denotes the answer in the question. The higher the probability value of the selected content in the picture frame, the more likely the range of interest is the correct answer. Making full use of these visualizations will benefit us make further improvements to the model in the future.

5. Conclusion

This article introduces a new *VQA* model, High- and Low-Level Semantic Information Network *(HLSIN)*, which employs two strategies to achieve the fusion of high-level semantic information and low-level semantic information. Adaptive weight learning is used as the first strategy to allow different levels of semantic information to learn weights separately. The gate-sum mechanism is used as the second to suppress invalid information in various levels of information and fuse valid information. The results of the fourth section of the ablation experiment prove that these two information fusion methods are effective, and the *HLSIN* composed of the seven-layer *SA (Y)-GSGA (X, Y)* unit has the best effect.

Although the accuracy of HLSIN model is greatly improved compared with that of MCAN model, the counting of HLSIN model still needs to be greatly improved compared with other models. Therefore, in the process of future research, I will try to combine the model with graph reasoning to improve the counting ability of the model.

In recent years, with the continuous development of *VQA* technology, a variety of new research directions have emerged. Such as: it can help visually impaired people "see" the world better, in particular, for example: when a blind people in the supermarket or other place, want to know what object in front of him or her, he or she can take a photo and enter into a *VQA* system, this can be a better tool, let them get information from outside world. *VQA* can also be used for image retrieval, selecting from a large number of images matching a question, or reasoning through the answer to find a video containing the question, etc. Other applications include medical question answering, intelligent driving and virtual reality avatars.However, there are still some problems in the development of visual VQA, such as: the images in the dataset are not close to the reality, the reasoning ability is not strong enough, and the semantic features of the questions can not be well combined with the image features. Therefore, how to build a VQA system to understand the relationship between different objects and the relationship between questions is the challenge of the future.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under Grant 61873160 and Grant 61672338, Natural Science Foundation of Shanghai under Grant 21ZR1426500.

References

[1] H. Nam, J.W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," CoRR, abs/1611.00471, 2016.

[2] J.H. Kim, J. Jun, and B.T. Zhang, "Bilinear attention networks," CoRR, abs/1805.07932, 2018.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.6077–6086, 2018.

[4] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," CoRR, abs/1411.4389, 2014.

[5] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C.L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," International Journal of Computer Vision, vol.123, pp.4–31, 2015.

[6] Z. Guo, D. Han, and K.-C. Li, "Double-layer affective visual question answering network," Comput. Sci. Inf. Syst., vol.18, pp.155–168, 2021.

[7] D. Han, N. Pan, and K.-C. Li, "A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection," IEEE Transactions on Dependable and Secure Computing, pp.316–327, 2020.

[8] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J.M.F. Moura, D. Parikh, and D. Batra, "Visual dialog.," IEEE Trans. Pattern Anal. Mach. Intell., vol.41, no.5, pp.1242–1256, 2019.

[9] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," Sensors (Basel, Switzerland), vol.20, no.7, 2020.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.21–29, 2016.

[11] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.6274–6283, 2019.

[12] D.-K. Nguyen and T. Okatani, "Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6087–6096, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," NAACL-HLT, 2019.

[14] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," IEEE transactions on neural networks and learning systems, vol.29, no.12, pp.5947–5959, 2018.

[15] H. Li and D. Han, "Edurss: A blockchain-based educational records secure storage and sharing scheme," IEEE Access, vol.7, pp.179273–179289, 2019.

[16] H. Liu, D. Han, and D. Li, "Fabric-iot: A blockchain-based access control system in iot," IEEE Access, vol.8, pp.18207–18218, 2020.

[17] P. Gao, Z. Jiang, H. You, P. Lu, S.C.H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.6632–6641, 2019.

[18] P. Gao, H. Y ou, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp.5824–5834, 2019.

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

[20] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, "Context-aware neural machine translation learns anaphora resolution," arXiv, abs/1805.10163, 2018.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[22] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," arXiv preprint arXiv:1512.02167, 2015.

[23] D. Han, S. Zhou, K.C. Li, and R.F de Mello, "Cross-modality co-attention networks for visual question answering," Soft Comput., vol.25, pp.5411–5421, 2021.

[24] M. Cui, D. Han, and J. Wang, "An efficient and safe road condition monitoring authentication scheme based on fog computing," IEEE Internet of Things Journal, vol.6, no.5, pp.9076–9084, 2019.

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

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," ICML, 2015.

[27] J.B. Delbrouck and S. Dupont, "Multimodal compact bilinear pooling for multimodal neural machine translation," CoRR, abs/1703.08084, 2017.

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

[29] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear

pooling with co-attention learning for visual question answering," Proceedings of the IEEE international conference on computer vision, pp.1821–1830, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.

[31] J. Ba, J. Kiros, and G.E. Hinton, "Layer normalization," arXiv, abs/1607.06450, 2016.

[32] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Densely connected attention flow for visual question answering," IJCAI, 2019.

[33] G. Peng, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp.5824–5834, 2019.

[34] T. Wang, H. Luo, X. Zeng, Z. Yu, A. Liu, and A.K. Sangaiah, "Mobility based trust evaluation for heterogeneous electric vehicles network in smart cities," IEEE Transactions on Intelligent Transportation Systems, vol.22, no.3, pp.1797–1806, 2020.

[35] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv, abs/1706.03762, 2017.

[36] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. Shamma, M.S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol.123, pp.32–73, 2016.

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

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," ECCV, vol.8693, pp.740–755, 2014.

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

[40] S. He and D. Han, "An effective dense co-attention networks for visual question answering," Sensors (Basel, Switzerland), vol.20, no.17, 2020.

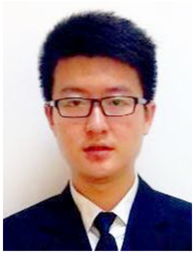
Huimin Li graduated from Xinxiang College with a bachelor's degree. In addition, she is currently studying for a M.S. degree at Shanghai Maritime University. At this stage, she mainly follows her instructor to research visual question answering and machine learning.

Dezhi Han is a Professor, doctoral supervisor, head of the Department of Computing, director of the master's program of the first-level discipline of computer science and technology, and the first outstanding discipline leader of Shanghai Maritime University. The doctoral supervisor of information management and information systems has guided and trained seven postgraduate students; the supervisor of computer science and technology, software engineering, and computer technology majors has instructed and trained more than 50 graduate students.

Chongqing Chen entered Shanghai Maritime University to study for a M.S. degree in 2018. During this period, he has published many papers and won many awards. In addition, he has successfully applied for a Ph.D. and started his Ph.D. study in September 2020.

Chin-Chen Chang received the Ph.D. degree in computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1982, and the B.E. and M.E. degrees in applied mathematics, computer and decision sciences from National Tsinghua University, Hsinchu, Taiwan, in 1977 and 1979, respectively. He was with National Chung Cheng University, Minxiong, Taiwan. Currently, he is a Chair Professor with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan, since 2005. His current research interests include database design, computer cryptography, image compression, and data structures. Prof. Chang was a recipient of many research awards and honorary positions by and in prestigious organizations both nationally and internationally, such as the Outstanding Talent in Information Sciences of Taiwan. He is currently a Fellow of the IEEE, a Fellow of the IEE, U.K. and a Member of the IEICE.

Kuan-Ching Li (SM'07) received the Ph.D. and M.S. degrees in electrical engineering and the Licenciate degree in mathematics from the University of São Paulo, Brazil,in 2001,1996, and 1994, respectively. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering, Providence University, Taiwan. He has been involved actively in many major conferences and workshops as a program/general/steering conference chairman positions and member of the program committee and has organized numerous conferences related to high-performance computing and computational science and engineering. Besides publishing numerous research papers and articles, he is co-author/co-editor of several technical professional books published by CRC Press/Taylor Francis, Springer, McGraw-Hill, and IGI Global. His research interests include parallel and distributed processing, GPU/many-core computing, big data, and cloud.



Dun Li is with undergraduate studies at Macau University of Science and Technology, and his main research field is economics. He taught at the Beijing Institute of Technology for one year. He is studying for a Ph.D. degree at Shanghai Maritime University and will go to France in September 2021 for a one-year study abroad.