Exploring deep learning for multimodal understanding

Lao, M.

Chapter 3

Multi-Stage Hybrid Embedding Fusion Network for Visual Question Answering

In this chapter, we focus on the first research question (**RQ 1**) on the model architecture sides for VQA systems. To establish an efficient multimodal integration scheme to enhance the vision-language understanding, we attempt to establish a multi-stage and multi-space fusion approach to achieve efficient feature interactions.

Multimodal fusion is a crucial component of Visual Question Answering (VQA), which involves joint understanding and semantic integration between visual and textual information. In this chapter, we intend to achieve multiple and fine-grained multimodal interactions for enhancing fusion performance. To this end, we propose a Multi-stage Hybrid Embedding Fusion (MHEF) network to fulfill our improvements in two folds: First, we introduce a Dual Embedding Fusion (DEF) approach that transforms one modal input into the reciprocal embedding space before integration, and the DEF is further incorporated with the LEF to form a novel Hybrid Embedding Fusion (HEF). Second, we design a Multi-stage Fusion Structure (MFS) for the HEF to form the MHEF network, so as to obtain diverse and better fusion features for answer prediction. By jointly training the multi-stage framework, we can not only improve the performance in each single stage, but also obtain additional accuracy improvements by integrating all prediction results from each stage. Extensive experiments verify both our proposed HEF and MFS are beneficial to multi-modal fusion. The full MHEF model outperforms the baseline LEF model with 2% accuracy improvements, and achieves promising performance on the VQA-v1 and VQA-v2 datasets.

This chapter is based on the published journal paper:

- **Lao, M.**, Guo, Y., Pu, N., Chen, W., Liu, Y. and Lew, M. S. "Multi-Stage Hybrid Embedding Fusion Network for Visual Question Answering." Neurocomputing, 2021.

3.1 Introduction

Recently, Visual Question Answering (VQA) [30] has received extensive attention in both the academia and industry. It requires a high level understanding of visual/textual information, and aims to accurately answer natural-language questions about given images. VQA can significantly benefit a variety of applications, such as robot tutors, smart home management systems and private virtual assistant.

For achieving simultaneous understanding based on given images and questions, multimodal fusion is an indispensable part in VQA. Its purpose is to incorporate image and question features and to generate integrated visual-textual features for answer prediction. Multimodal fusion has been widely studied in VQA. The prevailing fusion schemes utilized in current VQA learning frameworks include MLB [81], MFB [82], MUTAN [83] and MFH [84]. There are two characteristics for these mainstream multimodal fusion schemes: First, most methods rely on latent embedding that designs two-branch networks in which the visual and textual features are embedded into a common latent space, and then using some operations like multiplication or summation to fuse them as shown in Fig. 3.1(a). Second, these methods are inclined to perform a single interaction between visual and textual inputs, and predict correct answers based on a single fusion feature. For improving such two characteristics, we propose a Multi-stage Hybrid Embedding Fusion (MHEF) network to achieve multi-space and multi-stage interactions in a unified framework.

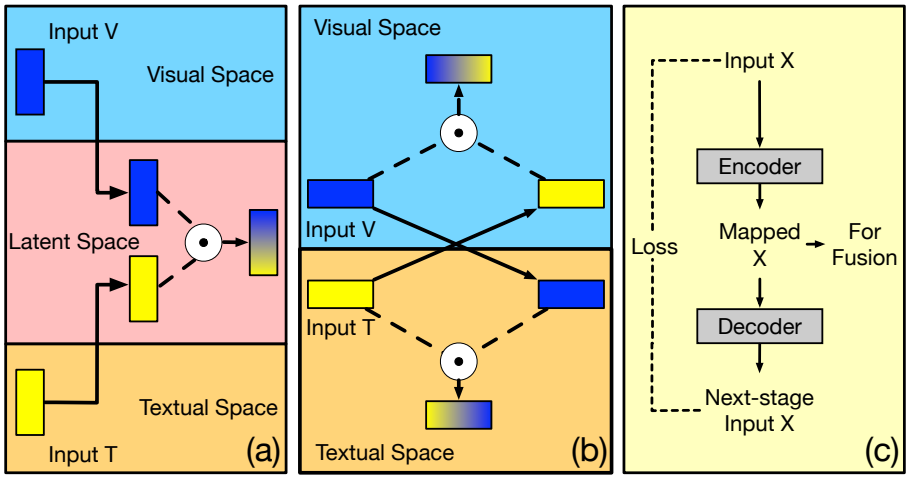

Figure 3.1: (a) Latent Embedding Fusion, $\odot$ is the element-wise multiplication; (b) Dual Embedding Fusion; (c) The core idea of MFS.

For the first characteristic, Latent Embedding Fusion (LEF) is an efficient approach for combining two modal features. To achieve better fusion feature than LEF, motivated by dual mapping [244] in the researches of text-image matching, we intend to explore multimodal relationships in diverse feature spaces for feature fusion, and propose a novel Dual Embedding Fusion (DEF) method. As depicted in Fig. 3.1(b),

DEF establishes two-branch layers to project visual/textual inputs into their reciprocal space (textual/visual space) and then fuse with the original inputs from the other modality in each space. In addition, we combine our DEF and the mainstream LEF in parallel to form a Hybrid Embedding Fusion (HEF) network. In the HEF, after generating fusion features in visual, textual and latent embedding spaces, we unify multi-space fusion features into a final fusion feature for answer prediction. We assume that performing multiple interactions in diverse spaces can gain effective and productive multimodal correlations for fusion features.

For the second characteristic, in order to reduce the potential semantic loss and implement more interactions for improving multimodal fusion performance, we attempt to perform a multi-stage fusion method. To this end, we propose a Multi-stage Fusion Structure (MFS). Specifically, the embedded feature is not only utilized for fusion, but also for reconstructing the original input. In this way, we can ensure the embedded features bring less information loss. Additionally, the reconstructed features can be proceeded for the next round prediction as well. Fig. 3.1(c) illustrates the core idea in our proposed MFS.

Based on the MFS and HEF, we build a Multi-stage Hybrid Embedding Fusion (MHEF) network by constructing the Multi-stage Latent Embedding Fusion and the Multi-stage Dual Embedding into a parallel combination. To be specific, for each fusion stage in the MHEF, we can get fusion features in three spaces (visual, textual and latent spaces), and the same fusion feature processing in HEF are employed to produce a final fusion feature for each stage. Then, to further enhance VQA performance, multi-stage final fusion features are separately exploited to predict answer candidates, and the final result is the average of the predictions from all stages.

In a nutshell, our contributions can be summarized as follows: (1) We propose a novel DEF scheme, and further incorporate the LEF with our DEF in parallel to establish a HEF approach, thereby capturing rich multimodal semantic correlations in multimodal fusion. (2) In order to reduce information loss and further improve fusion performance, we present a novel MFS, and apply it into the HEF model for generating a MHEF network. (3) We carry out extensive ablation studies over each component in the MHEF, and validate the benefits of the HEF and MFS for multimodal fusion. Furthermore, our proposed MHEF remarkably outperforms the dominant multimodal fusion approaches, and yields promising performance on the VQA-v1 and VQA-v2 datasets.

3.2 Related Work

3.2.1 Visual Question Answering

Multimodal learning inspired considerable researches at the boundary of computer vision and natural language processing, among which visual question answering is an important direction. It brings promising prospects in various applications, such as medical assistance and early education. Multimodal feature extraction and multimodal fusion are two crucial parts for VQA learning frameworks. In this paper, we mainly focus on the Multimodal fusion, whose related works will be described in the next subsection.

As the downstream task of image and natural language understanding, one crucial operation in multimodal feature extraction is to use pre-trained models to obtain multimodal features. For visual representation, previous networks exploit VGG-net [58] or Res-net [245] to extract visual spatial features. Recently, most of current VQA approaches tend to employ Faster-RCNN [246] to obtain visual objects features. [191] verifies that the features visual objects can have better representation for image and achieve better VQA performance. For textual features, VQA approaches use the word embeddings [247] as the inputs of the LSTM [248] or GRU [249] networks to encoding question word features.

Multimodal features extracted from pre-trained models is not sufficient for VQA. Most of VQA approaches employ attention mechanism to focus on the most relevant image regions and question words for better visual and textual representation. Yang et al. [66] present a stack attention network to update the attended image regions iteratively. Lu et al. [67] propose a co-attention model to simultaneously focus on important visual and textual information. Kim et al. Li et al. [68] propose a bilinear attention network to achieve bilinear interactions between multimodal inputs. Li et al. [69] present a relation-aware graph attention network to encode images into graphs, and build multi-type inter-object relations via graph attention mechanism.

3.2.2 Multimodal Fusion

As the foundation for VQA, multimodal fusion has been extensive studied. Early methods used simple linear fusion schemes to merge image and question inputs, such as concatenation, summation and multiplication. For achieving high-order interactions between multimodal features, bilinear pooling [250] has been considered as an effective way to fuse information from two sources, since it can take all pairwise interactions among given features into consideration. However, due to the fact that standard bilinear pooling employs outer-product operation and consumes lots of parameters, some advanced bilinear pooling methods (MLB, MFB, MUTAN and MFH) are presented to decrease the employment of computational resources, and further improve the VQA performance.

One crucial rectification based on traditional bilinear pooling in these bilinear approaches is to exploit a low-rank factorization [81] for weight tensors in bilinear fusion. Practically, in neural network operation, this process can be operated as follows: two multimodal vectors are separately translated by two linear weight matrices into a common latent space, and computed by element-wise multiplication. In this chapter, we name this operation as the LEF. The MUTAN, MFB and MFH approaches can be considered as the improved LEF that bring the conception of Tucker decomposition [251], high-dimensional expanding and high-order interaction with the low-rank factorization, which means that LEF is still a crucial component in these state-of-the-art bilinear fusion schemes.

3.2.3 Dual Embedding

In this chapter, we assume that the aforementioned LEF-based approaches may not sufficiently catch visual-textual relationships when fusing multimodal features. Therefore, the dual embedding learning, the widely adopted and effective resolution for multimodal embedding, is taken consideration for improving multimodal fusion in our paper.

Recently, dual embedding learning approach is successfully utilized in the image-text matching [252, 253], which demonstrates its superiority to capture visual/textual relations. The crucial operation in the dual embedding learning is projecting visual features into the textual feature space and vice versa, so as to capture rich multimodal relationships in both visual and textual space. Moreover, Huang et al. [254] also shows that latent embedding can be additionally used in the dual embedding models to enhance cross-modal relations. In this chapter, we attempt to acquire more semantic correlations between image and question inputs. We propose a dual embedding fusion (DEF) approach, and extend the DEF with the dominant LEF scheme to form a novel Hybrid Embedding Fusion (HEF) for visual-textual feature fusion.

3.2.4 Multi-stage Learning

Multi-stage or multi-prediction approaches are also the solutions to boost performance for multimodal tasks. Guo et al. [255] propose a dual prediction network to rectify one-stage captioning generation into a two-stage process (forward and backward prediction), and achieves remarkable improvement in image captioning. Mingrui et al. [256] propose a cross-modal multistep fusion network to fuse visual and textual features recursively, and the learning parameters would not increase linearly. Shen et al. [257] propose a multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection, and achieve promising results on two public datasets. Li et al. [258] introduce a multi-stage object detection network to fully exploit the learned segmentation features, and achieve remarkable performance. In this chapter, we attempt to propose a multi-stage fusion for achieving

two-fold enhancements: 1) Motivated by circle-consistent learning [252, 259], we propose to reconstruct multimodal inputs when implementing multi-modal feature fusion. In this way, we can reduce the potential semantic loss in feature mapping, and keep the semantic consistency for both visual and textual inputs. 2) We exploit the reconstructed visual and textual features for fulfilling the second-stage multi-modal fusion. Extensive experiments verify that joint training for multi-stage fusion can bring mutual benefits for fusion feature representations. Furthermore, we can acquire a more comprehensive answer prediction by taking all stage predictions into account.

3.3 Proposed Model

We first elaborate two important components in our MHEF model: Hybrid Embedding Fusion (HEF) in Section 3.3.1, and Multi-stage Fusion Structure (MFS) in Section 3.3.2. Then, we introduce the full structure of our Multi-stage Hybrid Embedding Fusion (MHEF) network, followed by a VQA framework with the MHEF approach in Section 3.3.3.

3.3.1 Hybrid Embedding Fusion

In this section, we first describe the DEF approach, and then introduce the HEF model by incorporating DEF with the LEF.

(1) Dual Embedding Fusion (DEF) Fig 3.1(b). Input visual feature $v \in \mathbb{R}^{d_v \times 1}$ and textual feature $t \in \mathbb{R}^{d_t \times 1}$ first pass through fully connected layers into the textual/visual space and obtain $v^{pt} \in \mathbb{R}^{d_t \times 1}$ and $t^{pv} \in \mathbb{R}^{d_v \times 1}$ respectively. d_v and d_t are the dimensions of v and t. Next, the projected features t^{pv} and v^{pt} are integrated with the original features v and t, using the element-wise multiplications for fusion features in visual and textual space (f^v and f^t) as follows:

$$f^v = (W_{t-v})^T t \circ v, \tag{3.1}$$

$$f^t = (W_{v-t})^T v \circ t, \tag{3.2}$$

where $W_{t-v} \in \mathbb{R}^{d_t \times d_v}$ and $W_{v-t} \in \mathbb{R}^{d_v \times d_t}$ are the parameters,$(W_{t-v})^T t$ and $(W_{v-t})^T v$ can represent the t^{pv} and v^{pt}. The $\circ$ denotes the element-wise multiplication operation. Typically, the DEF is independently utilized for predicting answer without further improvement, f^v and f^t are fed into two linear layer and mapped into a common latent space. Then, concatenation is used to combine two mapped features for unified-space representation ff,

$$ff = Concat((W^u_{v-l})^T f^v, (W^u_{t-l})^T f^t), \tag{3.3}$$

where the $(W^u_{v-l})^T \in \mathbb{R}^{d_v \times d_l}$ and $(W^u_{t-l})^T \in \mathbb{R}^{d_t \times d_l}$ are learning matrices to embed f^v and f^t into the common latent space.

(2) LEF+DEF The HEF network is built by incorporating the Latent Embedding Fusion (LEF) and DEF in a parallel way. After obtaining f^v and f^t from the DEF, the same original inputs v and t are also adopted to implement the LEF and the fusion feature in latent space f^l can be computed as follows:

$$f^l = (W_{v-l})^T v \circ (W_{t-l})^T t, \tag{3.4}$$

where $W_{v-l} \in \mathbb{R}^{d_v \times d_l}$ and $W_{t-l} \in \mathbb{R}^{d_t \times d_l}$ are the parameters for latent embeddings. d_l is the dimension of the common latent embedding space . By jointly achieving dual and latent embedding fusion, HEF network can implement multiple interactions in three different embedding spaces (visual, textual and latent spaces).

Then, to acquire final fusion feature for answer prediction, we embed f^v and f^t computed from the DEF into the common latent space, and integrate them with the multiplication. After that, the fused feature is concatenated with the f^l or the final fusion feature ff.

$$ff = Concat((W^u_{v-l})^T f^v \circ (W^u_{t-l})^T f^t, f^l), \tag{3.5}$$

3.3.2 Multi-stage Fusion Structure (MFS)

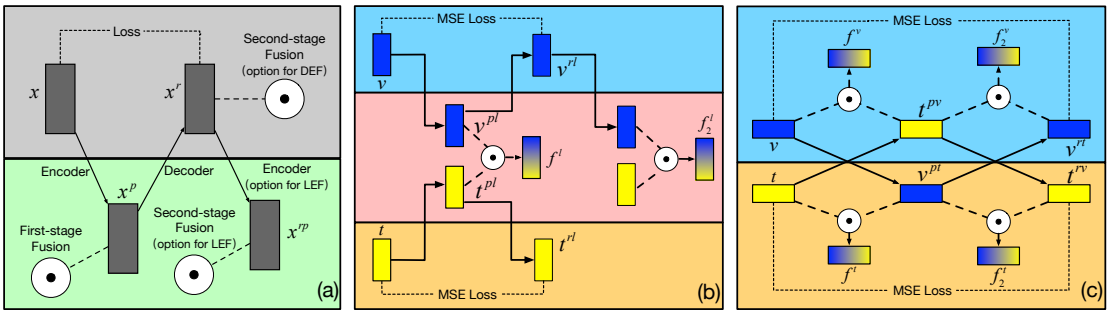

Figure 3.2: (a) Multi-stage Fusion Structure; (b) Multi-stage Latent Embedding Fusion; (c) Multi-stage Dual Embedding Fusion.

In this section, we explore to rectify current single-stage fusion schemes into a multi-stage fusion approaches. Specifically, visual and textual inputs are not only embedded into different spaces for integration, but also employed to reconstruct original features. In addition, the reconstructed features would act as the inputs for the next fusion stage.

Fig. 3.2(a) depicts the conception of our proposed Multi-stage Fusion Structure. As for an input feature x in its original visual/textual space (the gray area), it

first passes through a mapping (encoder) to produce mapped features x^p in another feature space (the green area). Apart from integrating with feature from the other modal in this space (the first-stage fusion), x^p is also fed into an inverse mapping (decoder) to project it back to its input space (x^r). By making x^r and input x similar through supervision, x^r can be seen as the reconstructed feature of x, and can act as the visual/textual input in the next fusion step. Specifically, if the x is for the DEF, x^r can be directly fused with the mapped features from the other modality in the input space (the gray area) to obtain the second-stage fusion feature. On the contrary, for LEF, the green area represents the latent space. x^r needs to be fed into a mapping again, and the projected reconstructed feature x^{rp} should be integrated with the projected reconstructed feature from the other modal in this latent embedding space.

In practice, for a multi-stage fusion scheme, fusion features from each step can be jointly used for answer prediction to further improve the performance in VQA.

(1) Multi-stage Latent Embedding Fusion (MLEF)

In this subsection, we detail how our MFS is incorporated with the LEF. The neural network architecture of the MLEF is described in Fig. 3.2(b).

To be explicit, two inputs v and t re fed into two linear layers to transform them into the latent space, and two transformed feature vectors v^{pl} and t^{pl} are integrated with multiplication which is computed as Equ. (3.4). Meanwhile, v^{pl} and t^{pl} are also utilized to yield reconstructed visual and textual features (v^{rl} and t^{rl}) by inverse projection, and we apply Mean Squared Error (MSE) loss to narrow the difference between reconstructed and original features. For the visual feature, these operations can be formulated as follows,

$$v^{rl} = (W_{l-v})^T \sigma(v^{pl}), \tag{3.6}$$

$$l(v, v^{rl}) = \left\| v - \sigma(v^{rl}) \right\|_2, \tag{3.7}$$

where $l(v, v^{rl})$ is the MSE loss to supervise v and v^{rl} to be similar. $W_{l-v} \in \mathbb{R}^{d_l \times d_v}$ is the parameter matrix to project v^{pl} back to visual space, and σ represents the ReLU function. The reconstructed textual feature t^{rl} and $l(v, v^{rl})$ for the textual feature can be computed by analogy.

Furthermore, two reconstructed features v^{rl} and t^{rl} will be adopted as the visual and textual inputs and used to generate a new fusion f_2^l in the second fusion stage by LEF:

$$f_2^l = (W_{v-l})^T v^{rl} \circ (W_{t-l})^T t^{rl}. \tag{3.8}$$

Specifically, the parameters W_{v-l} and W_{t-l} are the same as the parameters in the Equ. (3.4).

(2) Multi-stage Dual Embedding Fusion (MDEF)

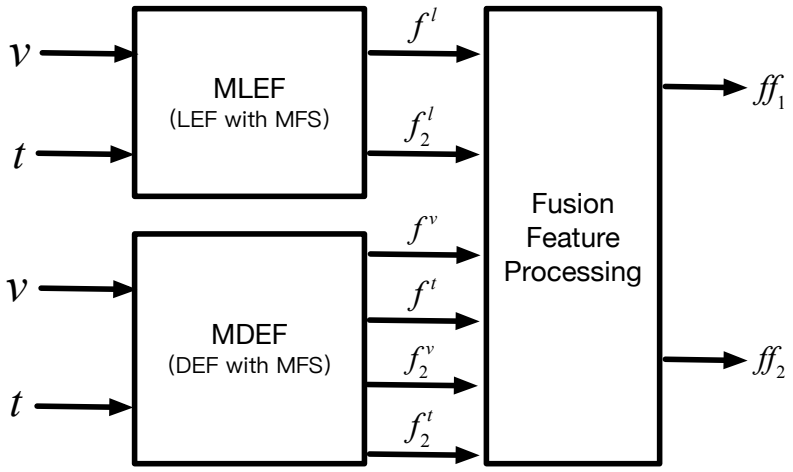

Figure 3.3: The structure of the MHEF network

In Multi-stage Dual Embedding Fusion (MDEF) network (Fig. 3.2(c)), apart from fulfilling DEF in the first stage, projected features v^{pl} and v^{pl} are also fed into two inverse mappings to reconstruct their original inputs supervised by MSE loss. Specifically, reconstructed multimodal features v^{rt}, t^{rv} and their corresponding loss function $l(v, v^{rt})$, $l(t, t^{rv})$ can be computed as Equ. (3.6) and (3.7).

Unlike MLEF, to perform the second stage fusion, reconstructed multimodal features v^{rt} and t^{rv} are not necessary to be mapped into another space for fusion. Reconstructed visual feature v^{rt} are directed fused with mapped textual space t^{pv} for the second stage fusion feature in visual space f_2^v,

$$f_2^v = v^{rt} \circ t^{pv}. \tag{3.9}$$

Meanwhile, the textual space output f_2^t in the second stage is calculated as follows,

$$f_2^t = t^{rv} \circ v^{pt}. \tag{3.10}$$

3.3.3 Multi-stage Hybrid Embedding Fusion Network (MHEF)

Fig. 3.3 depicts our proposed MHEF for multimodal fusion. The front end of MHEF is a module combining MLEF and MDEF in parallel. By feeding visual and textual inputs into this module, as described in section 3.2, we can get six fusion features, among which two fusion features (f^l and f_2^l) from MLEF and four features (f^v, f^t, f_2^v and f_2^t) from MDEF.

Then, we divide the six fusion features into two groups. The first group are the fusion features from HEF in the first fusion stage (f^l, f^v and f^t). The remaining three features f_2^l, f_2^v and f_2^t (fusion features in the second stage) are in the other group. Next, we need to generate the final joint representations for two groups. Specifically, the feature fusion processing module for three features in each group is the same as the formula (5), which is exploited to unify multi-space fusion features into the unified-space representations (ff_1 and ff_2) for two-stage answer prediction,

$$ff_1 = Concat((W_{v-l}^{u1})^T f^v \circ (W_{t-l}^{u1})^T f^t, f^l), \tag{3.11}$$

$$ff_2 = Concat((W_{v-l}^{u2})^T f_2^v \circ (W_{t-l}^{u2})^T f_2^t, f_2^l), \tag{3.12}$$

where $W_{v-l}^{u1}/W_{v-l}^{u2} \in \mathbb{R}^{d_v \times d_l}$ and $W_{t-l}^{u1}/W_{t-l}^{u2} \in \mathbb{R}^{d_t \times d_l}$ are parameters to unify f^v/f_2^v and f^t/f_2^t into common latent spaces.

3.3.4 VQA framework with MHEF

In this subsection, we describe a VQA framework with our proposed MHEF network for multimodal fusion. The architecture of the VQA framework is described in Fig. 3.4.

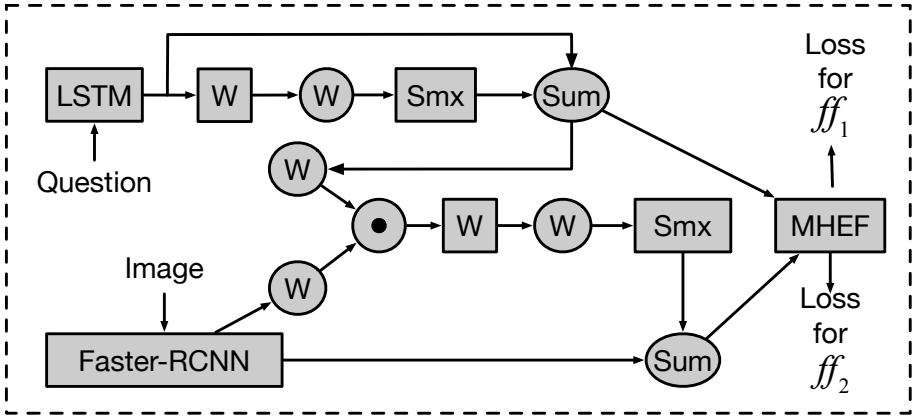

Figure 3.4: VQA framework with MHEF. The rounds and squares with W label represent the linear and non-linear layers. The Smx denotes the softmax function, and Sum is the weighted summation for word-level question features and object level image features.

Question and image feature extraction: For language feature extraction, 300-D GLoVe [247] word embedding is applied to pass through the LSTM [248] for acquiring word-level question features. In addition, we adopt pre-trained Faster R-CNN [246] model to extract visual objects features.

Textual and visual attention mechanism: Attention mechanism is a significant part to boost performance, and has widely used in current VQA researches. The frame-

work in Fig. 3.4 also implements textual and visual attention mechanism to generate attentive image and question features for acquiring better visual and textual representations. For textual attention, word-level question features pass through a non-linear layer and a linear layer followed by a softmax function to distribute weights for each word, and the attentive question feature is computed by the weighted summation of question input features.

To obtain visual attentive feature, object-level image features and attentive question feature are mapped into a common space and fused together. Similar to textual attention, the combination is fed into a non-linear followed by a linear layer, and the attention weights are computed by softmax function. Finally, we use weighted summation for object-level visual features to acquire attentive visual feature.

Answer prediction and training loss: After getting two-stage final fusion features ff_1 and ff_2 from our proposed MHEF model, we propose to exploit two features to separately predict the probability distribution for answer prediction. Specifically, for predicting answers, ff_1 and ff_2 are mapped mapped into vectors $pred_1, pred_2 \in \mathbb{R}^N$, where N denotes the most frequent answers in the training dataset. Then, the final decision making is the average prediction results based on two-stage fusion features.

It is worth noting that there are two groups of training loss in our VQA framework. The first group is two KL-divergence (KLD) losses l_1 and l_2 to train two k-way classifiers generated by ff_1 and ff_2,

$$l_1 = D_{KL}(y_i \,||\, z_i^1) = \sum_i y_i log(\frac{y_i}{z_i^1})), \tag{3.13}$$

$$l_2 = D_{KL}(y_i \,||\, z_i^2) = \sum_i y_i log(\frac{y_i}{z_i^2})), \tag{3.14}$$

where $y \in \mathbb{R}^N$ is the ground-truth distribution, and N is the total number of answer candidates. $y_i^1 \in \mathbb{R}^N$ and $y_i^2 \in \mathbb{R}^N$ are the answer prediction generated by ff_1 and ff_2. The rest are four MSE losses ($l(v, v^{rl})$, $l(v, v^{rt})$, $l(t, t^{rl})$, $l(t, t^{rv})$), utilized to make reconstructed visual (v^{rl}, v^{rt}) and textual features (t^{rl}, t^{rv}) similar to their original inputs (v and t). The total loss l_r for these four losses can be written as follows:

$$l_r = \alpha(l(v, v^{rl}) + l(t, t^{rl}) + l(v, v^{rt}) + l(t, t^{rv})), \tag{3.15}$$

where α is applied to adjust the weights of the loss terms. As a result, the final loss function l_f can be formulated as the Equ. (3.16),

$$l_f = l_1 + l_2 + l_r \tag{3.16}$$

3.4 Experiments

3.4.1 Datasets and Evaluation Metric

We conduct extensive experiments to evaluate the performance of MHEF on two commonly used datasets: VQA-v1 [30] and VQA-v2 [48]. VQA-v1 dataset consists of 614,163 samples, and 204,721 images from MSCOCO. VQA v1.0 utilizes the 204K images and related captions from the MSCOCO dataset for annotating nearly three questions per image and ten optioned answers per question. The dataset provides 123K images for training and validation and 81K images for testing. VQA v2.0 is a new version of VQA v1.0, which is a more balanced version by reducing data bias. It provides the question-answers pairs for newly collected complementary images to diminish the language bias. It enlarges the dataset as it contains 195K images for train set, 93K images for val set, and 191K images for test set. It is noteworthy that, we do not exploit any augmented or supplemental dataset like Visual Genome to facilitate training or further improve VQA performance.

To estimate the performance of a VQA framework , an accuracy-based evaluation metric is determined as follows, which is robust to inter-human variability in phrasing the answer:

$$ACC(answer) = min\{\frac{count(answer)}{3}, 1\}, \tag{3.17}$$

Where count(answer) is a function that count the answer voted by varied annotators. It illustrates that the predicting result of answer candidate is 100%, if at least 3 annotators provided the answer.

3.4.2 Implementation Details

Network Details: The configurations of our VQA framework are as follows: We set the max length of words in question is 15, and the size of answer vocabulary is 3129, which is consistent to answers appearing more than eight times in the train set. For visual input, we set a threshold for Faster R-CNN to obtain 10-100 visual object features. The dimensions of image and question inputs are 2048 and 1024 respectively. In addition, the attention glimpses [80] in both visual and textual attention are equal to 2. The dimension of the latent space is 1024, and the hyper-parameter α in total training loss is 1/8. Dropout [260] ($p = 0.1$) and L2 normalization are added after each multiplication operation.

Training Details: We carried out our proposed approach based on Pytorch library. The initial learning rate is 0.002, and it is warmed-up for 4 epochs to 0.008. After 9 epochs, the learning rate is decayed by 1/4 for every epoch up to 13 epochs and stop training. All models are optimized employing the Adam solver [261], and the mini-batch is set to 64.

3.4.3 Ablation Study

In this section, we design some ablation studies to ablate our proposed modules from its complete form. VQA-v2 dataset is applied for these experiments, where we train our model on the train dataset and test on the val dataset. From Tab. 3.1, our proposed full model MHEF remarkably outperforms the baseline model LEF by more than 2%. Next, we will explicitly verify the improvements and benefits for two-fold contributions HEF and MFS in our MHEF model respectively.

Table 3.1: Experimental results for backbone model LEF and our full MHEF on VQA-v2 val dataset

Model	**Accuracy(%)**
LEF (backbone model)	64.33
MHEF (full model)	**66.40**

(1) Improvements of Hybrid Embedding Fusion (HEF)

The results of ablation studies for the HEF is in Tab. 3.2:

Table 3.2: Experimental results in ablation studies for HEF on VQA-v2 val dataset

Model	**Accuracy(%)**
LEF	64.33
DEF	64.66
HEF	**65.15**

DEF vs LEF: The dominant fusion scheme LEF is seen as the backbone model in our experiments. Compared with the LEF, our DEF surpasses it by around 0.3% on accuracy, which shows that DEF can produce better fusion features. The reason is that the DEF achieve multimodal feature interactions in both visual and textual spaces, and capture more correlations between question and image features.

HEF vs LEF/DEF: Furthermore, our HEF fusion scheme outperforms both LEF and DEF methods, and provides a gain of around 0.8% over the backbone model. It verifies that the HEF can provide more effective fusion features by incorporating DEF with LEF, so as to boost VQA performance by achieving multiple and multi-space (visual, textual and latent spaces) fusions.

(2) Improvements of Multi-stage Fusion Structure (MFS)

We explicitly investigate the MFS with different components in Tab. 3.3.

Multi-stage fusion vs Single-stage fusion: From the accuracies of MHEF and HEF model (E), multi-stage fusion outperforms the sing-stage fusion by more than 1.2%. Specifically, we conclude three benefits for Multi-stage Fusion Structure (MFS): multi predictions, joint training for each fusion stage and multimodal feature reconstruction.

Table 3.3: Experimental results in ablation study for MFS on VQA-v2 val dataset. The MHEF denotes the MHEF full model. The pred and recons are prediction (in one fusion stage) and feature reconstruction operation respectively. The I-st and II-st represent the first and the second fusion stage.

	MHEF	**(A)**	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**
II-st pred	✓		✓		✓		✓
I-st pred	✓	✓		✓		✓	
II-st loss	✓	✓	✓		✓		✓
I-st loss	✓	✓	✓	✓		✓	
Recons	✓	✓	✓	✓	✓		
Accuracy(%)	**66.40**	65.77	65.68	65.41	65.10	65.15	64.90

Effectiveness of multi predictions: One crucial improvement in the MHEF model is that we simultaneously exploit two-stage predictions for selecting answer candidates. From the results of MHEF, (A) and (B), jointly considering answer prediction from all stage features (MHEF) achieves improvement by at least 0.6% over the model (A and B) merely making decision with a single fusion feature. It reveals that taking multi-stage predicted results into consideration is beneficial to fulfill more comprehensive and precise answer prediction.

Effectiveness of joint training for each stage: Through the comparative analysis, model (A) and (B) achieve higher results than those of model (C) and (D) respectively. It demonstrates that training the first-stage and second-stage classifiers can simultaneously bring reciprocal enhancements for final fusion feature from all stages, which shows that achieving multi-stage fusions is superior to the single-stage fusion schemes even under the one-classifier prediction.

Effectiveness of multimodal feature reconstruction: In this subsection, we make comparisons between single-stage fusion models with multimodal feature reconstruction (C and D) and without reconstruction (E and F). We can notice that ablating inverse mapping (the decoder in the autoencoder structure) and supervision loss impairs the model performance and get lower accuracy. This tends to indicate that recovering the mapped features can decrease the semantic loss in embedding process and generate better representations for mapped and reconstructed features, which is consistent to the motivation of the circle-consistency.

(3) The Effect of Fusion Stage Number

In Tab. 3.4, we show the results of LEF, DEF and HEF models with the numbers of fusion stage $N \in \{1, 2, 3, 4\}$.Compared with the single-stage fusion model, all multi-stage models obtain remarkable better performance, increasing the accuracy by at least 1%. This provides support that multistage fusion can achieve more efficient multimodal interactions in latent (LEF), original (DEF), or hybrid (HEF) embedding spaces. On the other hand, more fusion stages do not necessarily mean better performance, which may be due to the potential overfitting in the approach.

Table 3.4: Results of the model LEF, DEF and HEF with different fusion stage numbers on VQA-v2 val dataset.

N	LEF	DEF	HEF
1	64.33	64.66	65.15
2	65.68	65.88	**66.40**
3	65.75	65.85	66.38
4	65.60	65.75	66.32

In addition, the excessive fusion stages also lead to higher computational costs. Therefore, the optimal setting we utilize is $N = 2$.

(4) The Effect of Embedding Dimension in Latent Space

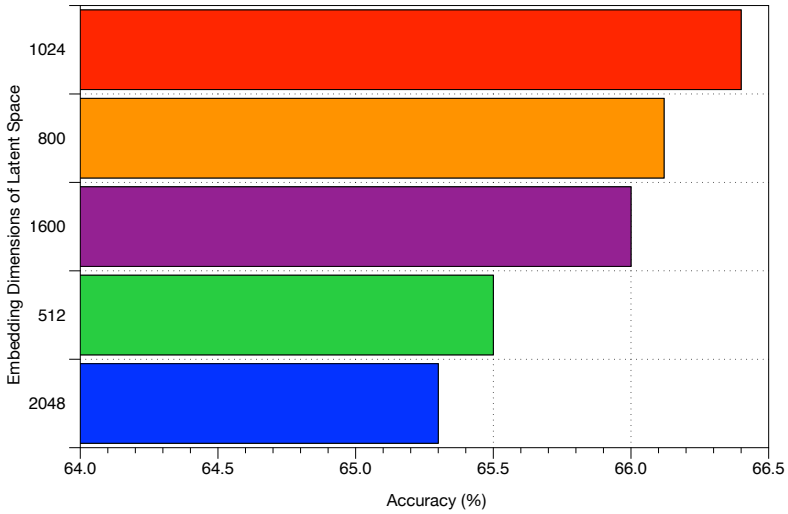

Figure 3.5: The accuracies of MHEF models with different embedding dimensions of latent spaces.

Fig. 3.5 illustrates the effect of latent space dimensions for our proposed MHEF model. We select the embedding dimensions of 512, 800, 1024, 1600,2048 for experiments. We can notice that MHEF with the dimension of 1024 achieve the best result. When the dimension is 512, the performance of MHEF reduces by around 1%. It shows that low embedding dimension is not sufficient to capture multimodal relationships. However, from the experimental result of the dimension 2048, selecting a high embedding dimension would also bring the overfitting problem.

3.4.4 Fusion Scheme Comparison

In this subsection, we implement six prevailing multimodal fusion schemes to conduct comparative experiments with our proposed MHEF. For fair comparison, all fusion schemes are fulfilled within the same VQA framework detailed in section 3.4. All of the models are trained on the VQA-v2 train+val dataset, and examined on

the test-dev dataset. We follow the optimal parameter settings in the original works of MLB, MUTAN, MFB and MFH: The dimension of common latent space in MLB is 1200. The dimension of latent space for both MFB and MFH is set to 1000, and the non-overlapped window in MFB/MFH is 5. For the MUTAN approach, the dimensions of all mappings are equal to 360, and the parameter of the constant is 10.

Table 3.5: Results of MHEF and other state-of-the-art fusion approaches on VQA-v2 test-dev dataset. Y/N, Num and Other are the question types of yes/no, number and other.

Model	Y/N	Num	Other	All
Concatenation	81.50	43.57	57.02	65.58
Summation	82.83	45.88	57.99	66.85
MLB	83.50	48.25	58.86	67.81
MUTAN	83.52	49.11	58.34	67.66
MFB	83.84	48.39	58.89	67.98
MFH	84.44	48.92	58.85	68.27
MHEF	**85.91**	**50.72**	**60.02**	**69.63**

Tab. 3.5 shows the accuracies of the fusion approaches, we can summarize that the traditional linear fusion schemes like Concatenation and Summation are obviously inferior to bilinear-based approaches (MLB, MUTAN, MFB and MFH), which demonstrates bilinear interaction can achieve more effective integration for multimodal features. Among these bilinear fusion approaches, MFB reaches slightly higher accuracy than MUTAN and MLB, as projecting multimodal feature into high-dimensional latent space can obtaining richer information before fusion. As the extension method based on the MFB, MFH implements extra high-order interactions between multimodal inputs and achieves better results than MFB. As for our proposed MHEF model, it achieves overall the best result which outperforms the competitive MFH model by more than 1.3%, which demonstrates the superiority of our MHEF model. It further proves that the two characteristics of the MHEF model, HEF and MFS, are beneficial for multimodal fusion.

3.4.5 Qualitative analysis

In this section, we present some examples to qualitatively describe the superiorities of our proposed MHEF model in Fig. 3.6. These examples cover a broad range of the answer types, including counting objects, color identification, verification and location. From these examples, given a simple image-question pair, both LEF and our MHEF can select the correct answer. However, if the question is relatively challenging to solve based on the given image, the LEF model fails to predict the right answer and tends to choose the answer based on the superficial correlation between image-question pairs. For instance, for the first examples related to the counting

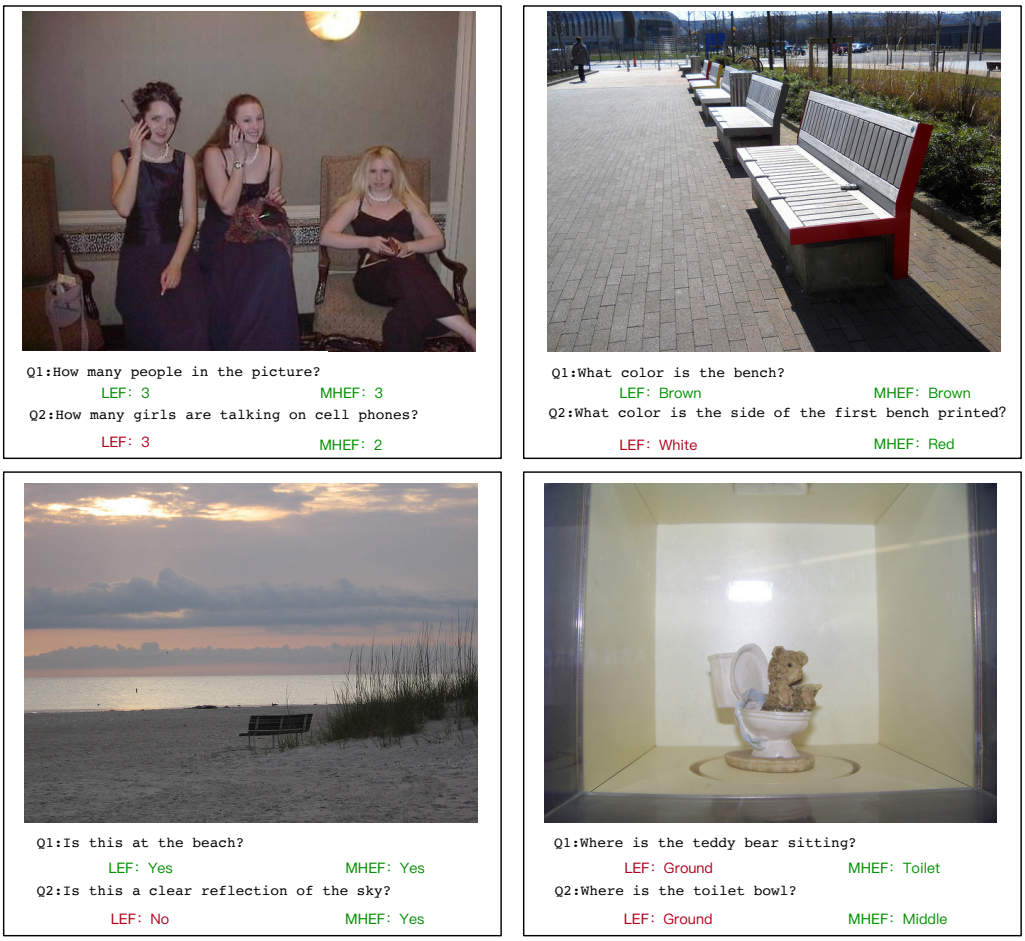

Figure 3.6: Qualitative examples from the VQA-v2 val dataset, with predictions of the backbone LEF model and our proposed MHEF model. For each image, we provide two questions from the same question type, and the predictions of LEF and MHEF models are listed below. The predicting answers in green are the correct answers, while results in red are the wrong answers to the corresponding questions.

objects, both two models can predict the right answer "3" for the straightforward question "How many people in the picture?" On the contrary, given a more detailed question like "How many girls are talking on cell phones?", the prediction of the MHEF model is correct, while the LEF model still chooses the answer "2" as the right answer. It demonstrates that, compared with the backbone LEF model, our proposed MHEF model can better capture the visual and textual semantic correlations for fusion features, and fulfill more precise and comprehensive decision making for answer prediction.

3.4.6 Comparison with the State-of-the-Art

The results of current state-of-the-art VQA approaches on the VQA-v1 dataset in Tab. 3.6. Our MHEF remarkably surpasses these three approaches based on bilinear fusion (MCB [80], MLB and MUTAN) by at least 2.5%, which highlights the effectiveness of our MHEF network for multimodal fusion. We also compare

Table 3.6: Results of MHEF and other state-of-the-art fusion approaches on VQA-v1 test-dev dataset. Y/N, Num and Other are the question types of yes/no, number and other.

Approach	Test-dev(%)				Test-std(%)			
	Y/N	Num	Other	All	Y/N	Num	Other	All
MCB[80]	83.40	39.80	58.50	66.70	83.20	39.50	58.00	66.50
MLB[81]	85.57	39.32	57.36	67.03	84.39	38.70	58.20	66.96
MUTAN[83]	85.14	39.81	58.52	67.42	84.91	39.79	58.35	67.36
ODA[262]	85.82	43.03	58.07	67.83	-	-	-	-
CoR-3[263]	85.69	44.06	59.08	68.37	85.83	43.93	59.11	68.54
CRA[264]	86.51	44.60	59.88	69.11	85.21	44.60	59.42	69.28
MHEF	**86.80**	**45.52**	**59.90**	**69.91**	**86.67**	**45.48**	**60.95**	**69.94**

Table 3.7: Results of MHEF and other state-of-the-art fusion approaches on VQA-v2 test-dev dataset. Y/N, Num and Other are the question types of yes/no, number and other.

Approach	Test-dev(%)				Test-std(%)			
	Y/N	Num	Other	All	Y/N	Num	Other	All
Bottom-Up[191]	81.82	44.21	56.05	65.32	82.20	43.90	56.26	65.67
DCN[191]	83.51	46.61	57.26	66.87	83.85	47.19	56.95	66.97
MuRel[191]	84.77	49.84	57.85	68.03	-	-	-	-
Counter[191]	83.14	**51.62**	58.97	68.09	83.56	**51.39**	59.11	68.41
ODA[191]	84.66	48.04	58.68	68.17	-	-	-	-
CRA-Net[191]	84.87	49.46	59.08	68.61	85.21	48.43	59.42	68.92
CoR-3[191]	85.22	47.95	59.15	68.62	85.76	48.40	59.43	69.14
MFH[191]	84.27	49.56	59.89	68.76	-	-	-	-
Ban+Glove[191]	85.46	50.66	**60.50**	**69.66**	-	-	-	-
MHEF	**85.91**	50.72	60.02	69.63	**86.01**	50.17	**60.19**	**69.80**

three VQA approaches with advanced attention mechanisms (ODA [262], CoR-3 [263], CRA-Net [264]). From the results, although these approaches adopt advanced attention approaches that focus on the multimodal relationships, our MHEF can still show its superiority over these methods and achieves state-of-the-art performance on VQA-v1 dataset. It further proves our MHEF is effective for multi-modal fusion, and remarkably enhance the VQA performance.

We compare our MHEF model to state-of-the-art approaches on the VQA-v2 in Table 3.7. Bottom-Up [191] model is the backbone framework for most state-of-the-art VQA models. MHEF obviously outperforms Bottom-Up model by more than 4%. It highlights our MHEF is powerful to implement productive multimodal interactions, and achieves remarkable improvement for VQA performance. Through the comparison between the MHEF model and the MFH method with the MFH-based co-attention framework, our model still achieves superior accuracy. It shows that our MHEF is more effective than the MFH for multimodal fusion, which is consistent with the experimental results in section 3.4.3. Some approaches implemented with state-of-the-art attention network like BAN [68], Counter [265], MuRel [266] and DCN [74] are selected to make some comparative analysis with MHEF, and we find that MHEF can still achieve competitive performance. It is noteworthy that

BAN is a state-of-the-art VQA approach that exploits bilinear attention mechanism. Unlike our MHEF model which fuses overall (attentive) visual and textual feature together, this mechanism can consider every pair (the pairs of question word and image regions) of multimodal inputs. Our model achieves competitive performance compared with the BAN approach, which demonstrates that, apart from capturing pairwise relationships for multimodal fusion, exploring multi-space and multi-stage fusion is also a competitive and potential direction.

Recently, some advanced approaches have been proposed to boost the VQA performance, such as relation-aware graph attention [69], and deep modular attention [74]. Compared with our VQA experimental model depicted in Fig. 3.4, these approaches tend to explore more accurate attention weights for multimodal inputs before multimodal fusion. In our future research, we will try to build relationships between attention mechanism and multistage fusion scheme, and further enhance VQA accuracy.

3.5 Conclusion

In this chapter, we propose a novel Multi-stage Hybrid Embedding Fusion (MHEF) network, which includes Hybrid Embedding Fusion (HEF) and Multi-stage Fusion Structure (MFS) to fulfill two-fold improvements for multimodal fusion. Through extensive experiments, we verify the effectiveness of each component in the MHEF model, and achieve remarkably better performance on the VQA-v1 and VQA-v2 datasets. In the future, we will incorporate MHEF with some recent advanced attention modules to further enhance its performance, and exploit its potentiality for other multimodal deep learning tasks.

Future works: From the practical aspect, we will incorporate MHEF with some recent advanced attention modules to further enhance its performance, and exploit its potentiality for other multimodal deep learning tasks, such as visual commonsense reasoning, video grounding, and multimodal sentiment analysis. From the theoretical aspect, we seek to analyze and quantify the model behaviours in different fusion stage, and further design a concise and efficient multi-stage fusion strategy.