

Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering

Zhou Yu[†], Jun Yu^{†*}, Jianping Fan[‡], Dacheng Tao[§]

[†] Key Laboratory of Complex Systems Modeling and Simulation,
School of Computer Science and Technology, Hangzhou Dianzi University, P. R. China

[‡] Department of Computer Science, University of North Carolina at Charlotte, USA

[§] UBTECH Sydney AI Centre, School of IT, FEIT, The University of Sydney, Australia

yuz@hdu.edu.cn, yujun@hdu.edu.cn, jfan@uncc.edu, dacheng.tao@sydney.edu.au

Abstract

Visual question answering (VQA) is challenging because it requires a simultaneous understanding of both the visual content of images and the textual content of questions. The approaches used to represent the images and questions in a fine-grained manner and questions and to fuse these multi-modal features play key roles in performance. Bilinear pooling based models have been shown to outperform traditional linear models for VQA, but their high-dimensional representations and high computational complexity may seriously limit their applicability in practice. For multi-modal feature fusion, here we develop a Multi-modal Factorized Bilinear (MFB) pooling approach to efficiently and effectively combine multi-modal features, which results in superior performance for VQA compared with other bilinear pooling approaches. For fine-grained image and question representation, we develop a ‘co-attention’ mechanism using an end-to-end deep network architecture to jointly learn both the image and question attentions. Combining the proposed MFB approach with co-attention learning in a new network architecture provides a unified model for VQA. Our experimental results demonstrate that the single MFB with co-attention model achieves new state-of-the-art performance on the real-world VQA dataset. Code available at <https://github.com/yuzcccc/mfb>.

1. Introduction

Thanks to recent advances in computer vision and natural language processing, computers are expected to be able to automatically understand the semantics of images and natural languages in the near future. Such advances have

also stimulated new research topics like image-text retrieval [35, 37], image captioning [5, 34], and visual question answering [3, 19].

Compared with image-text retrieval and image captioning (which just require the underlying algorithms to search or generate a free-form text description for a given image), *visual question answering* (VQA) is a more challenging task that requires fine-grained understanding of the semantics of both the images and the questions as well as supports complex reasoning to predict the best-matching answer correctly. In some aspects, the VQA task can be treated as a generalization of image captioning and image-text retrieval. Thus building effective VQA algorithms, which can achieve close performance like human beings, is an important step towards enabling artificial intelligence in general.

Existing VQA approaches usually have three stages: (1) representing the images as visual features and questions as textual features; (2) combining these multi-modal features to obtain fused image-question features; (3) using the integrated image-question features to learn a multi-class classifier and to predict the best-matching answer. Deep neural networks (DNNs) are effective and flexible, many existing approaches model the three stages in one DNN model and train the model in an end-to-end fashion through back-propagation. In the three stages, feature representation and multi-modal feature fusion particular affect VQA performance.

With respect to multi-modal feature fusion, most existing approaches simply use linear models for multi-modal feature fusion (e.g., concatenation or element-wise addition) to integrate the image’s visual feature with the question’s textual feature [39, 18]. Since multi-modal feature distributions may vary dramatically, the integrated image-question representations obtained by such linear models may not be sufficiently expressive to fully capture complex associations between the visual features from images and the textual

*Jun Yu is the corresponding author

features from questions. In contrast to linear pooling, bilinear pooling [28] has recently been used to integrate different CNN features for fine-grained image recognition [17]. However, the high dimensionality of the output features and the huge number of model parameters may seriously limit the applicability of bilinear pooling. Fukui *et al.* proposed the Multi-modal Compact Bilinear (MCB) pooling model to effectively and simultaneously reduce the number of parameters and computation time using the Tensor Sketch algorithm [6]. Using the MCB model, the group proposed a network architecture for the VQA task and won the VQA challenge 2016. Nevertheless, the MCB model lies on a high-dimensional output feature to guarantee robust performance, which may limit its applicability due to huge memory usage. To overcome this problem, Kim *et al.* proposed the Multi-modal Low-rank Bilinear (MLB) pooling model based on the Hadamard product of two feature vectors [12]. Since MLB generate output features with lower dimensions and models with fewer parameters, it is highly competitive with MCB. However, MLB has a slow convergence rate and is sensitive to the learned hyper-parameters. To address these issues, here we develop the Multi-modal Factorized Bilinear pooling (MFB) method, which enjoys the dual benefits of compact output features of MLB and robust expressive capacity of MCB.

With respect to feature representation, directly using global features for image representation may introduce noisy information that is irrelevant to the given question. Therefore, it is intuitive to introduce *visual attention* mechanism [34] into the VQA task to adaptively learn the most relevant image regions for a given question. Modeling visual attention may significantly improve performance [6]. However, most existing approaches only model image attention without considering question attention, even though question attention is also very important since the questions interpreted in natural languages may also contain colloquialisms that can be regarded as noise. Therefore, based on our MFB approach, we design a deep network architecture for the VQA task using a *co-attention learning* module to jointly learn both image and question attentions.

To summarize, the main contributions of this study are as follows: First, we develop a simple but effective Multi-modal Factorized Bilinear pooling (MFB) approach to fuse the visual features from images with the textual features from questions. MFB significantly outperforms existing multi-modal bilinear pooling approaches such as MCB [6] and MLB [12]. Second, based on the MFB module, a *co-attention learning* architecture is designed to jointly learn both image and question attention. Our MFB approach with co-attention model achieves the state-of-the-art performance on the VQA dataset. We also conduct detailed and extensive experiments to show why our MFB approach is effective. Our experimental results demonstrate

that normalization techniques are extremely important in bilinear models.

2. Related Work

In this section, we briefly review the most relevant research on VQA, especially those studies that use multi-modal bilinear models.

2.1. Visual Question Answering (VQA)

Malinowski *et al.* [19] made an early attempt at solving the VQA task. Since then, solving the VQA task has received increasing attention from the computer vision and natural language processing communities. VQA approaches can be classified into the following methodological categories: the coarse joint-embedding models [39, 3, 11, 26], the fine-grained joint-embedding models with attention [1, 18, 6, 9, 21, 38, 40] and the external knowledge based models [29, 30, 31].

The coarse joint-embedding models are the most straightforward VQA solutions. Image and question are first represented as global features and then integrated to predict the answer. Zhou *et al.* proposed a baseline approach for the VQA task by using the concatenation of the image CNN features and the question BoW (bag-of-words) features, with a linear classifier learned to predict the answer [39]. Some approaches introduce more complex deep models, e.g., LSTM networks [3] or residual networks [11], to tackle the VQA task in an end-to-end fashion.

One limitation of coarse joint-embedding models is that their global features may contain noisy information, making it hard to correctly answer fine-grained problems (e.g., “what color are the cat’s eyes?”). Therefore, recent VQA approaches introduce the *visual attention* mechanism [34] into the VQA task by adaptively learning the local fine-grained image features for a given question. Chen *et al.* proposed a “question-guided attention map” that projects the question embeddings to the visual space and formulates a configurable convolutional kernel to search the image attention region [4]. Yang *et al.* proposed a stacked attention network to learn the attention iteratively [36]. Some approaches introduce off-the-shelf object detectors [9] or object proposals [27] as the attention region candidates and then use the question to identify related ones. Fukui *et al.* proposed multi-modal compact bilinear pooling to integrate image features from spatial grids with textual features from the questions to predict the attention [6]. In addition, some approaches apply attention learning to both the images and questions. Lu *et al.* proposed a co-attention learning framework to alternately learn the image attention and the question attention [18]. Nam *et al.* proposed a multi-stage co-attention learning framework to refine the attentions based on memory of previous attentions [21].

Despite joint embedding models for VQA delivering impressive performance, they are not good enough for answering problems that require complex reasoning or common sense knowledge. Therefore, introducing external knowledge is beneficial for VQA. However, existing approaches have either only been applied to specific datasets [29, 30], or have been ineffective on benchmark datasets [31]. There is room for further exploration and development.

2.2. Multi-modal Bilinear Models for VQA

Multi-modal feature fusion plays an important and fundamental role in VQA. After the image and question features are obtained, concatenation or element-wise summations are most frequently used for multi-modal feature fusion. Since the distributions of two feature sets in different modalities (i.e., the visual features from images and the textual features from questions) may vary significantly, the representation capacity of the fused features may be insufficient, limiting the final prediction performance.

Fukui *et al.* first introduced the bilinear model to solve the problem of multi-modal feature fusion in VQA. In contrast to the aforementioned approaches, they proposed the Multi-modal Compact Bilinear pooling (MCB), which uses the outer product of two feature vectors to produce a very high-dimensional feature for quadratic expansion [6]. To reduce the computational cost, they used a sampling-based approximation approach that exploits the property that the projection of two vectors can be represented as their convolution. The MCB model outperformed the simple fusion approaches and demonstrated superior performance on the VQA dataset [3]. Nevertheless, MCB usually needs high-dimensional features (e.g., 16,000-D) to guarantee robust performance, which may seriously limit its applicability due to limitations in GPU memory.

To overcome this problem, Kim *et al.* proposed the Multi-modal Low-rank Bilinear Pooling (MLB) approach based on the Hadamard product of two feature vectors (i.e., the image feature $x \in \mathbb{R}^m$ and the question feature $y \in \mathbb{R}^n$) in the common space with two low-rank projection matrices: [12]:

$$z = \text{MLB}(x, y) = (U^T x) \circ (V^T y) \quad (1)$$

where $U \in \mathbb{R}^{m \times o}$ and $V \in \mathbb{R}^{n \times o}$ are the projection matrices, o is the dimensionality of the output feature, and \circ denotes the Hadamard product or the element-wise multiplication of two vectors. To further increase model capacity, nonlinear activation like \tanh is added after z . Since the MLB approach can generate feature vectors with low dimensions and deep models with fewer parameters, it has achieved comparable performance to MCB. In [12], the experimental results indicated that MLB may lead to a slow

convergence rate (the MLB with attention model takes 250k iterations, which is about 140 epochs, to converge [12]).

3. Multi-modal Factorized Bilinear Pooling

Given two feature vectors in different modalities, e.g., the visual features $x \in \mathbb{R}^m$ for an image and the textual features $y \in \mathbb{R}^n$ for a question, the simplest multi-modal bilinear model is defined as follows:

$$z_i = x^T W_i y \quad (2)$$

where $W_i \in \mathbb{R}^{m \times n}$ is a projection matrix, $z_i \in \mathbb{R}$ is the output of the bilinear model. The bias term is omitted here since it is implicit in W . To obtain a o -dimensional output z , we need to learn $W = [W_1, \dots, W_o] \in \mathbb{R}^{m \times n \times o}$. Although bilinear pooling can effectively capture the pairwise interactions between the feature dimensions, it also introduces huge number of parameters that may lead to high computational cost and a risk of over-fitting.

Inspired by the matrix factorization tricks for uni-modal data [15, 25], the projection matrix W_i in Eq.(2) can be factorized as two low-rank matrices:

$$\begin{aligned} z_i &= x^T U_i V_i^T y = \sum_{d=1}^k x^T u_d v_d^T y \\ &= \mathbf{1}^T (U_i^T x \circ V_i^T y) \end{aligned} \quad (3)$$

where k is the factor or the latent dimensionality of the factorized matrices $U_i = [u_1, \dots, u_k] \in \mathbb{R}^{m \times k}$ and $V_i = [v_1, \dots, v_k] \in \mathbb{R}^{n \times k}$, \circ is the Hadamard product or the element-wise multiplication of two vectors, $\mathbf{1} \in \mathbb{R}^k$ is an all-one vector.

To obtain the output feature $z \in \mathbb{R}^o$ by Eq.(3), the weights to be learned are two three-order tensors $U = [U_1, \dots, U_o] \in \mathbb{R}^{m \times k \times o}$ and $V = [V_1, \dots, V_o] \in \mathbb{R}^{n \times k \times o}$ accordingly. Without loss of generality, we can reformulate U and V as 2-D matrices $\tilde{U} \in \mathbb{R}^{m \times ko}$ and $\tilde{V} \in \mathbb{R}^{n \times ko}$ respectively with simple reshape operations. Accordingly, Eq.(3) can be rewritten as follows:

$$z = \text{SumPooling}(\tilde{U}^T x \circ \tilde{V}^T y, k) \quad (4)$$

where the function $\text{SumPooling}(x, k)$ means using a one-dimensional non-overlapped window with the size k to perform sum pooling over x . We name this model Multi-modal Factorized Bilinear pooling (MFB).

The detailed procedures of MFB are illustrated in Fig. 1(a). The approach can be easily implemented by combining some commonly-used layers such as fully-connected, element-wise multiplication and pooling layers. Furthermore, to prevent over-fitting, a dropout layer is added after the element-wise multiplication layer. Since element-wise multiplication is introduced, the magnitude of the output neurons may vary dramatically, and the model might

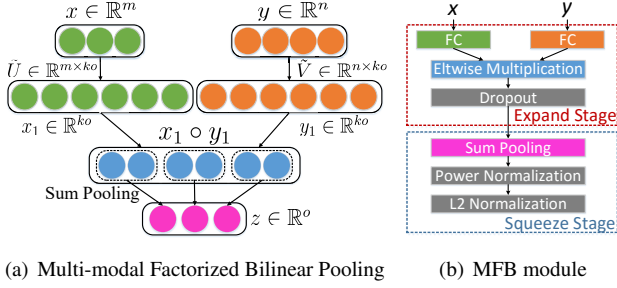


Figure 1. The flowchart of Multi-modal Factorized Bilinear Pooling and completed design of the MFB module.

converge to an unsatisfactory local minimum. Therefore, similar to [6], the power normalization ($z \leftarrow \text{sign}(z)|z|^{0.5}$) and ℓ_2 normalization ($z \leftarrow z/\|z\|$) layers are appended after MFB output. The flowchart of the entire MFB module is illustrated in Fig. 1(b).

Relationship to MLB. Eq.(4) shows that the MLB in Eq.(1) is a special case of the proposed MFB with $k = 1$, which corresponds to the rank-1 factorization. Figuratively speaking, MFB can be decomposed into two stages (see in Fig. 1(b)): first, the features from different modalities are *expanded* to a high-dimensional space and then integrated with element-wise multiplication. After that, sum pooling followed by the normalization layers are performed to *squeeze* the high-dimensional feature into the compact output feature, while MLB directly projects the features to the low-dimensional output space and performs element-wise multiplication. Therefore, with the same dimensionality for the output features, the representation capacity of MFB is more powerful than MLB.

4. Network Architectures for VQA

The goal of the VQA task is to answer a question about an image. The inputs to the model contain an image and a corresponding question about the image. Our model extracts both the image and the question representations, integrates the multi-modal features using the MFB module in Figure 1(b), treats each individual answer as one class and performs multi-class classification to predict the correct answer. In this section, two network architectures are introduced. The first is the MFB baseline with one MFB module, which is used to perform ablation analysis with different hyper-parameters for comparison with other baseline approaches. The second network introduces co-attention learning which jointly learns the image and question attentions, to better capture fine-grained correlations between the image and the question, which may lead to a model with better representation capability.

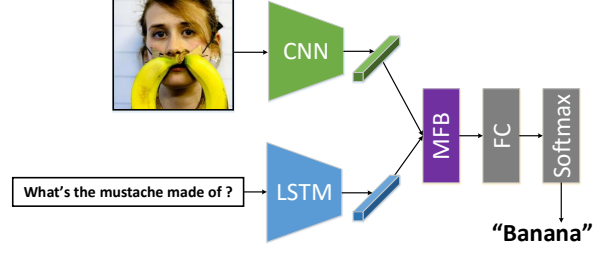


Figure 2. MFB baseline network architecture for VQA.

4.1. MFB Baseline

Similar to [6], we extract the image features using 152-layer ResNet model [7] pre-trained on the ImageNet dataset. Images are resized to 448×448 , and 2048-D *pool5* features (with ℓ_2 normalization) are used for image representation. Questions are first tokenized into words, and then further transformed to one-hot feature vectors with max length T . Then, the one-hot vectors are passed through an embedding layer and fed into a two-layer LSTM networks with 1024 hidden units [8]. Each LSTM layer outputs a 1024-D feature for each word. Similar to [6], we extract the output feature of the last word from each LSTM network, and concatenate the obtained features of two LSTM networks to form a 2048-D feature vector for question representation. For predicting the answers, we simply use the top- N most frequent answers as N classes since they follow the long-tail distribution.

The extracted image and question features are fed to the MFB module to generate the fused feature z . Finally, z is fed to a N -way classifier with the KL-divergence loss. Therefore, all the weights except the ones for the ResNet (due to the limitation of GPU memory) are optimized jointly in an end-to-end manner. The whole network architecture is illustrated in Figure 2.

4.2. MFB with Co-Attention

For a given image, different questions could result in entirely different answers. Therefore, an *image attention* model, which can predict the relevance of each spatial grid to the question, is beneficial for predicting the accurate answer. In [6], 14×14 (196) image spatial grids (*res5c* feature maps in ResNet) are used to represent the input image. After that, the question feature is merged with each of the 196 image features using MCB, followed by some feature transformations (e.g., 1×1 convolution and ReLU activation) and softmax normalization to predict the attention weight for each grid location. Based on the attention map, the attentional image features are obtained by the weighted sum of the spatial grid vectors. Multiple attention maps are generated to enhance the learned attention map, and these attention maps are concatenated to output the

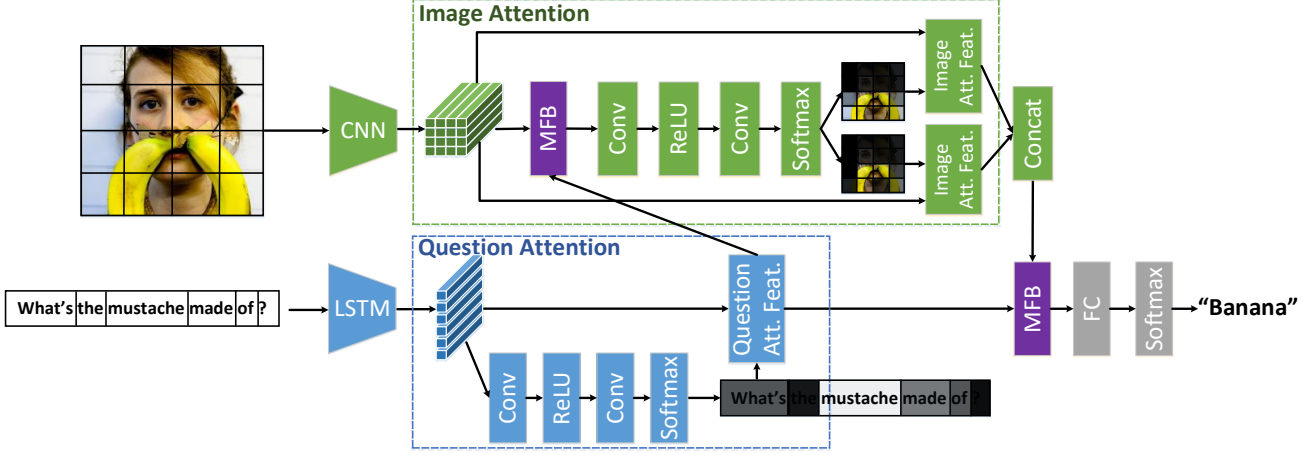


Figure 3. MFB with Co-Attention network architecture for VQA. Different from the network of MFB baseline, the images and questions are firstly represented as the fine-grained features respectively. Then, *Question Attention* and *Image Attention* modules are jointly modeled in the framework to provide more accurate answer predictions.

attentional image features. Finally, the attentional image features are merged with the question features using MCB to determine the final answer prediction.

From the results reported in [6], one can see that incorporating an attention mechanism allows the model to effectively learn which region is important for the question, clearly contributing to better performance than the model without attention. However, the attention model in [6] only focuses on learning image attention while completely ignoring question attention. Since the questions are interpreted as natural language, the contribution of each word is significantly different. Therefore, here we develop a co-attention learning approach (see Figure 3) to jointly learn both the question and image attentions.

The difference between the network architecture of our co-attention model and the attention model in [6] is that we additionally place a *question attention* module after the LSTM networks to learn the attention weights of every word in the question. Different to other co-attention models for VQA [18, 21], in our model, the image and question modules are loosely coupled such that we do not exploit the image features when learning the question attention module. This is because we assume that the network can directly infer the question attention (i.e., the key words of the question) without seeing the image, as humans do. We name this network MFB with Co-Attention (MFB+CoAtt).

5. Experiments

In this section, we conduct several experiments to evaluate the performance of our MFB models on the VQA task using the VQA dataset [3] to verify our approach. We first perform ablation analysis on the MFB baseline model to verify the efficiency of the proposed approach over existing state-of-the-art methods such as MCB [6] and MLB

[12]. We then provide detailed analyses of the reasons why our MFB model outperforms its counterparts. Finally, we choose the optimal hyper-parameters for the MFB module and train the model with co-attention (MFB+CoAtt) for fair comparison with other state-of-the-art approaches on the VQA dataset [3].

5.1. Datasets

The VQA dataset [3] consists of approximately 200,000 images from the MS-COCO dataset [16], with 3 questions per image and 10 answers per question. The data set is split into three: *train* (80k images and 248k questions), *val* (40k images and 122k questions), and *test* (80k images and 244k questions). Additionally, there is a 25% test split subset named *test-dev*. Two tasks are provided to evaluate performance: Open-Ended (OE) and Multiple-Choices (MC). We use the tools provided by Antol *et al.* [3] to evaluate the performance on the two tasks.

5.2. Experimental Setup

For the VQA dataset, we use the Adam solver with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The base learning rate is set to 0.0007 and decays every 40,000 iterations using an exponential rate of 0.5. We terminate training at 100,000 iterations (200,000 iterations if the training set is augmented with the large-scale Visual Genome dataset [14]). Dropouts are used after each LSTM layer (dropout ratio $p = 0.3$) and MFB module ($p = 0.1$) like [6]. The number of answers $N = 3000$. For all experiments (except for the ones shown in Table 2, which use the train and val splits together as the training set like the comparative approaches), we train on the train split, validate on the val split, and report the results on the test

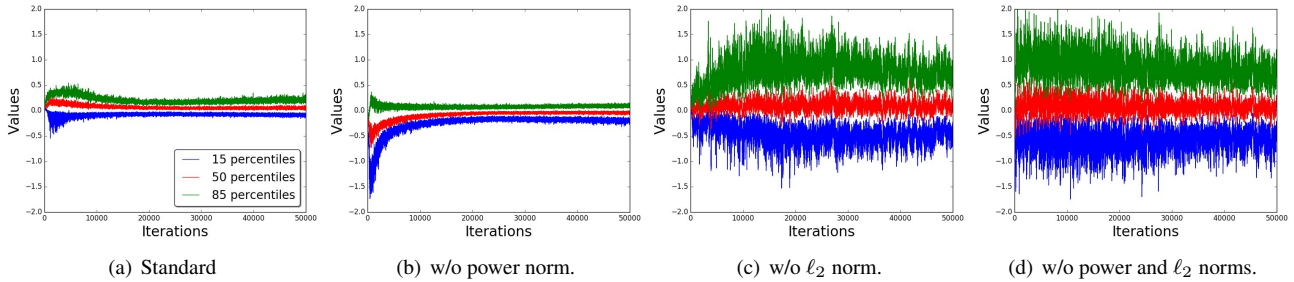


Figure 4. The evolution of the output distribution of one typical neuron with different normalization settings, shown as {15,50,85}th percentiles. Both normalization techniques, especially the ℓ_2 normalization make the neuron values restricted within a narrow range, thus leading to a more stable model. Best viewed in color.

Table 1. Overall accuracies and model sizes of approaches and on the test-dev set of the Open-Ended task. The reported accuracy is the overall accuracy of all question types. The model size includes the parameters for the LSTM networks.

Model	Acc.	Model Size
MCB[6] ($d = 16000$)	59.8	63M
MLB[12] ($d = 1000$)	59.7	25M
MFB($k = 1, o = 5000$)	60.4	51M
MFB($k = 5, o = 1000$)	60.9	46M
MFB($k = 10, o = 500$)	60.6	38M
MFB($k = 5, o = 200$)	59.8	22M
MFB($k = 5, o = 500$)	60.4	28M
MFB($k = 5, o = 2000$)	60.7	62M
MFB($k = 5, o = 4000$)	60.4	107M
MFB($k = 5, o = 1000$)	-	-
-w/o power norm.	60.4	-
-w/o ℓ_2 norm.	57.7	-
-w/o power and ℓ_2 norms.	57.3	-

split¹. The batch size is set to 200 for the models without the attention mechanism, and set to 64 for the models with attention (due to GPU memory limitation). All experiments are implemented with the *Caffe* toolbox [10] and performed on a workstation with GTX 1080 GPUs.

5.3. Ablation Analysis

In Table 1, we compare MFB’s performance with other state-of-the-art bilinear pooling models, namely MCB [6] and MLB (for fair comparison, we replace the tanh function in MLB with the proposed power+ ℓ_2 normalizations) [12], under the same experimental settings. None of these methods introduce the attention mechanism. Furthermore, we explore different hyper-parameters and normalizations introduced in MFB to explore why MFB outperform the compared bilinear models.

From Table 1, we can see that:

¹the submission attempts for the test-standard split are strictly limited. Therefore, we evaluate most of our settings on the test-dev split and only report the best results on the test-standard split.

First, MFB significantly outperforms MCB and MLB. With 5/6 parameters, MFB($k = 5, o = 1000$) achieves about a 1% accuracy improvement compared with MCB. Moreover, with only 1/3 parameters, MFB($k = 5, o = 200$) obtains similar results to MCB. These characteristics allows us to train our model on a memory limited GPU with larger batch-size. Furthermore, the validation accuracy of MCB suffers from overfitting with the high-dimensional output features. In comparison, the performance of our MFB model is relatively robust.

Second, when ko is fixed to a constant, e.g., 5000, the number of factors k affects the performance. Increasing k from 1 to 5, produces a 0.5% performance gain. When $k = 10$, the performance has approached saturation. This phenomenon can be explained by the fact that a large k corresponds to using a large window to sum pool the features, which can be treated as a compressed representation and may loss some information. When k is fixed, increasing o does not produce further improvements. This suggests that high-dimensional output features may be easier to overfit. Similar results can be seen in [6]. In summary, $k = 5$ and $o = 1000$ may be a suitable combination for our MFB model on the VQA dataset, so we use these settings in our follow-up experiments.

Finally, both the power and ℓ_2 normalization benefit MFB performance. Power normalization results in about 0.5% improvement and ℓ_2 normalization, perhaps surprisingly, results in about 3% improvement. Results without ℓ_2 and power normalizations were also reported in [3] and are similar to those reported here. To explain why normalization are so important, we randomly choose one typical neuron from the MFB output feature before normalization to illustrate how its distribution evolves over time in Figure 4. It can be seen that the standard MFB model (with both normalizations) leads to the most stable neuron distribution and without the power normalization, about 10,000 iterations are needed to achieve stabilization. Without the ℓ_2 normalization, the distribution varies seri-

Table 2. Open-Ended (OE) and Multiple-Choice (MC) results on VQA dataset compared with the state-of-the-art approaches in terms of accuracy in %. Att. indicates whether the approach introduce the attention mechanism, W.E. indicates whether the approach uses external word embedding models. VG indicates the model is trained with the Visual Genome dataset additionally. All the reported results are obtained with a *single model*. For the test-dev set, the best results in each split are bolded. For the test-standard set, the best results overall all the splits are bolded.

Model	Att.	W.E.	Test-dev					Test-standard				
			OE				MC	OE				MC
			All	Y/N	Num	Other	All	All	Y/N	Num	Other	All
iBOWIMG [39]			55.7	76.5	35.0	42.6	-	55.9	78.7	36.0	43.4	62.0
DPPnet [23]			57.2	80.7	37.2	41.7	-	57.4	80.3	36.9	42.2	-
VQA team [3]			57.8	80.5	36.8	43.1	62.7	58.2	80.6	36.5	43.7	63.1
AYN [20]			58.4	78.4	36.4	46.3	-	58.4	78.2	36.3	46.3	-
AMA [31]			59.2	81.0	38.4	45.2	-	59.4	81.1	37.1	45.8	-
DMN+ [32]			60.3	80.5	36.8	60.3	-	60.4	-	-	-	-
MCB [6]			61.1	81.7	36.9	49.0	-	61.1	81.7	36.9	49.0	-
MRN [11]			61.7	82.3	38.9	49.3	-	61.8	82.4	38.2	49.4	66.3
MFB (Ours)			62.2	81.8	36.7	51.2	67.2	-	-	-	-	-
SMem [33]	✓		58.0	80.9	37.3	43.1	-	58.2	80.9	37.3	43.1	-
NMN [2]	✓		58.6	81.2	38.0	44.0	-	58.7	81.2	37.7	44.0	-
SAN [36]	✓		58.7	79.3	36.6	46.1	-	58.9	-	-	-	-
FDA [9]	✓		59.2	81.1	36.2	45.8	-	59.5	-	-	-	-
DNMN [1]	✓		59.4	81.1	38.6	45.4	-	59.4	-	-	-	-
HieCoAtt [18]	✓		61.8	79.7	38.7	51.7	65.8	62.1	-	-	-	-
RAU [22]	✓		63.3	81.9	39.0	53.0	67.7	63.2	81.7	38.2	52.8	67.3
MCB+Att [6]	✓		64.2	82.2	37.7	54.8	-	-	-	-	-	-
DAN [21]	✓		64.3	83.0	39.1	53.9	69.1	64.2	82.8	38.1	54.0	69.0
MFB+Att (Ours)	✓		64.6	82.5	38.3	55.2	69.6	-	-	-	-	-
MFB+CoAtt (Ours)	✓		65.1	83.2	38.8	55.5	70.0	-	-	-	-	-
MCB+Att+GloVe [6]	✓	✓	64.7	82.5	37.6	55.6	-	-	-	-	-	-
MLB+Att+StV [12]	✓	✓	65.1	84.1	38.2	54.9	-	65.1	84.0	37.9	54.8	68.9
MFB+CoAtt+GloVe (Ours)	✓	✓	65.9	84.0	39.8	56.2	70.6	65.8	83.8	38.9	56.3	70.5
MCB+Att+GloVe+VG [6]	✓	✓	65.4	82.3	37.2	57.4	-	-	-	-	-	-
MLB+Att+StV+VG [12]	✓	✓	65.8	83.9	37.9	56.8	-	-	-	-	-	-
MFB+CoAtt+GloVe+VG (Ours)	✓	✓	66.9	84.1	39.1	58.4	71.3	66.6	84.2	38.1	57.8	71.4

ously over the entire training course. This observation is consistent with the results shown in Table 1.

5.4. Comparison with State-of-the-art

Table 2 compares our approaches with the current state-of-the-art. The table is split into four parts over the rows: the first summarizes the methods without introducing the attention mechanism; the second includes the methods with attention; the third illustrates the results of approaches with external pre-trained word embedding models, e.g., GloVe [24] or Skip-thought Vectors (StV) [13]; and the last includes the models trained with the external large-scale Visual Genome dataset [14] additionally. To best utilize model capacity, the training data set is augmented so that both the train and val splits are used as the training set, result in about 1% ~ 2% overall accuracy improvement on the OE task. Also, to better understand the question semantics, pre-trained GloVe word vectors are concatenated with the learned word embedding. The MFB model corresponds to the MFB baseline model. The MFB+Att model indicates the model that replaces the MCB with our MFB in the

MCB+Att model [6]. The MFB+CoAtt model represents the network shown in Figure 3.

From Table 2, we have the following observations:

First, the model with MFB outperforms other comparative approaches significantly. The MFB baseline outperforms all other existing approaches without the attention mechanism for both the OE and MC tasks, and even surpasses some approaches with attention. When attention is introduced, MFB+Att consistently outperforms current next-best model MCB+Att, highlighting the efficacy and robustness of the proposed MFB.

Second, the co-attention model further improve the performance over the attention model with only considering the image attention. By introducing co-attention learning, MFB+CoAtt delivers a 0.5% improvement on the OE task compared with the MFB+Att model in terms of overall accuracy, indicating the additional benefits of the co-attention learning framework.

Finally, with the external pre-trained GloVe model and the Visual Genome dataset, the performance of our models are further improved. The MFB+CoAtt+GloVe+VG model



Figure 5. Typical examples of the learned image and question of the MFB+CoAtt+GloVe model. The top row shows four examples of four correct predictions while the bottom row shows four incorrect predictions. For each example, the query image, question (Q), answer (A) and prediction (P) are presented from top to bottom; the learned image and question attentions are presented next to them. The brightness of images and darkness of words represent their attention weights.

Table 3. Comparison with the state-of-the-art results (with model ensemble) on the test-standard set of the VQA dataset. The best results are bolded.

Model	OE				MC
	All	Y/N	Num	Other	All
HieCoAtt [18]	62.1	80.0	38.2	52.0	66.1
RAU [22]	64.1	83.3	38.0	53.4	68.1
7 MCB models [6]	66.5	83.2	39.5	58.0	70.1
7 MLB models [12]	66.9	84.6	39.1	57.8	70.3
7 MFB models (Ours)	68.4	85.6	40.6	59.8	72.5
Human [3]	83.3	95.8	83.4	72.7	91.5

significantly outperforms the best reported results with a single model on both the OE and MC task.

In Table 3, we compare our approach with the state-of-the-art methods with model ensemble. Similar with [6, 12], we train 7 individual MFB+CoAtt+GloVe models and average the prediction scores of them. Four of the seven models additionally introduce the Visual Genome dataset [14] into the training set. For fair comparison, only the published results are demonstrated. From Table 3, the ensemble of MFB models outperforms the next best approach by 1.5% on the OE task and by 2.2% on the MC task respectively. Finally, compared with the results obtained by human, there is still a lot of room for the improvement to approach the human-level.

To better demonstrate the effects of co-attention learning, in Figure 5 we visualize the learned question and image attentions of some examples from the validation set. The examples are randomly picked from different question types. It can be seen that the learned question and image attentions are usually closely focus on the key words and the most relevant image regions. From the incorrect examples, we can also draw conclusions about the

weakness of our approach, which are perhaps common to all VQA approaches: 1) some key words in the question are neglected by the question attention module, which seriously affects the learned image attention and final predictions (e.g., the word *catcher* in the first example and the word *bottom* in the third example); 2) even the intention of the question is well understood, some visual contents are still unrecognized (e.g., the *flags* in the second example) or misclassified (the *meat* in the fourth example), leading to the wrong answer for the counting problem. These observations are useful to guide further improvements for VQA in the future.

6. Conclusions

In this paper, we develop a Multi-modal Factorized Bilinear pooling (MFB) approach to fuse multi-modal features for the VQA task. Compared with existing bilinear pooling methods, the MFB approach achieves significant performance improvement for the VQA task. Based on MFB, we design a network architecture with co-attention learning that achieves new state-of-the-art performance on the real-world VQA dataset. This explorations of multi-modal bilinear pooling and co-attention learning are applicable to a wide range of tasks involving multi-modal data.

7. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 61622205, 61602136 and 61472110, the Zhejiang Provincial Natural Science Foundation of China under Grant LR15F020002, the Australian Research Council under Project FL-170100117, DP-140102164, and LP-150100671.

References

- [1] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016. 2, 7
- [2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016. 7
- [3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1, 2, 3, 5, 6, 7, 8
- [4] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 2015. 2
- [5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. 1
- [6] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2, 3, 4, 5, 6, 7, 8
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4
- [8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4
- [9] I. Ilievski, S. Yan, and J. Feng. A focused dynamic attention model for visual question answering. *arXiv preprint arXiv:1604.01485*, 2016. 2, 7
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia (ACMMM)*, pages 675–678, 2014. 5
- [11] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in neural information processing systems (NIPS)*, pages 361–369, 2016. 2, 7
- [12] J.-H. Kim, K. W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 2, 3, 5, 6, 7, 8
- [13] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *Advances in neural information processing systems (NIPS)*, pages 3294–3302, 2015. 7
- [14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 5, 7
- [15] Y. Li, N. Wang, J. Liu, and X. Hou. Factorized bilinear models for image recognition. *arXiv preprint arXiv:1611.05709*, 2016. 3
- [16] 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. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5
- [17] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015. 2
- [18] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems (NIPS)*, pages 289–297, 2016. 1, 2, 5, 7, 8
- [19] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems (NIPS)*, pages 1682–1690, 2014. 1, 2
- [20] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. 7
- [21] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016. 2, 5, 7
- [22] H. Noh and B. Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv preprint arXiv:1606.03647*, 2016. 7, 8
- [23] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30–38, 2016. 7
- [24] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 7
- [25] S. Rendle. Factorization machines. In *International Conference on Data Mining (ICDM)*, pages 995–1000, 2010. 3
- [26] K. Saito, A. Shin, Y. Ushiku, and T. Harada. Dualnet: Domain-invariant network for visual question answering. *arXiv preprint arXiv:1606.06108*, 2016. 2
- [27] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4613–4621, 2016. 2
- [28] J. B. Tenenbaum and W. T. Freeman. Separating style and content. *Advances in neural information processing systems (NIPS)*, pages 662–668, 1997. 2
- [29] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015. 2, 3
- [30] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick. Fvqa: Fact-based visual question answering. *arXiv preprint arXiv:1606.05433*, 2016. 2, 3
- [31] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4622–4630, 2016. 2, 3, 7

- [32] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning (ICML)*, pages 2397–2406, 2016. 7
- [33] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision (ECCV)*, pages 451–466, 2016. 7
- [34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, volume 14, pages 77–81, 2015. 1, 2
- [35] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Transactions on Multimedia*, 10(3):437–446, 2008. 1
- [36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2016. 2, 7
- [37] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *International Conference on Research on Development in Information Retrieval (SIGIR)*, pages 395–404, 2014. 1
- [38] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang. Video question answering via hierarchical spatio-temporal attention networks. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 2
- [39] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 1, 2, 7
- [40] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann. Uncovering temporal context for video question and answering. *arXiv preprint arXiv:1511.04670*, 2015. 2