Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering

Zhou Yu, Jun Yu, *Member, IEEE*, Chenchao Xiang, Jianping Fan, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Visual question answering (VQA) is challenging, because it requires a simultaneous understanding of both visual content of images and textual content of questions. To support the VQA task, we need to find good solutions for the following three issues: 1) fine-grained feature representations for both the image and the question; 2) multimodal feature fusion that is able to capture the complex interactions between multimodal features; and 3) automatic answer prediction that is able to consider the complex correlations between multiple diverse answers for the same question. For fine-grained image and question representations, a "coattention" mechanism is developed using a deep neural network (DNN) architecture to jointly learn the attentions for both the image and the question, which can allow us to reduce the irrelevant features effectively and obtain more discriminative features for image and question representations. For multimodal feature fusion, a generalized multimodal factorized high-order pooling approach (MFH) is developed to achieve more effective fusion of multimodal features by exploiting their correlations sufficiently, which can further result in superior VQA performance as compared with the state-of-the-art approaches. For answer prediction, the Kullback–Leibler divergence is used as the loss function to achieve precise characterization of the complex correlations between multiple diverse answers with the same or similar meaning, which can allow us to achieve faster convergence rate and obtain slightly better accuracy on answer prediction. A DNN architecture is designed to integrate all these aforementioned modules into a unified model for achieving superior VQA performance. With an ensemble of our MFH models, we achieve the state-of-the-art performance on the large-scale VQA data sets and win the runner-up in VQA Challenge 2017.**

Manuscript received July 28, 2017; revised December 12, 2017 and March 1, 2018; accepted March 14, 2018. Date of publication April 9, 2018; date of current version November 16, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61702143, Grant 61622205, Grant 61472110, and Grant 61772161, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LR15F020002, and in part by the Australian Research Council Projects under Grant FL-170100117, Grant LP-150100671, and Grant DP-180103424. *(Corresponding author: Jun Yu.)*

Z. Yu, J. Yu, and C. Xiang are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China (e-mail: yuz@hdu.edu.cn; yujun@hdu.edu.cn; hdu_xcc@163.com).

J. Fan is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: jfan@uncc.edu).

D. Tao is with the UBTECH Sydney Artificial Intelligence Centre, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia, and also with the School of Information Technologies, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia (email: dacheng.tao@sydney.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2817340

***Index Terms*— Coattention learning, deep learning, multimodal feature fusion, visual question answering (VQA).**

I. Introduction

THANKS to recent advances of deep neural networks (DNNs) 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 also continue to redefine and drive research in image–text retrieval [1]–[3], image captioning [4], [5], and visual question answering (VQA) [6], [7].

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, 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 accurately. 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 performance that is close to that of human beings, is an important step toward the general artificial intelligence.

To support the VQA task, we need to address the following three issues effectively (see Fig. 1): 1) extracting discriminative features for image and question representations; 2) combining the visual features from the image and the textual features from the question to generate the fused image–question features; and 3) using the fused image–question features to learn a multiclass classifier for predicting the best matching answer correctly. DNNs are very effective and flexible, and most of the existing VQA approaches tackle these three issues in one single DNNs model and train the model in an end-to-end fashion through back-propagation.

For feature-based image representation, directly using the global features extracted from the whole image may introduce noisy information (i.e., irrelevant features) that are irrelevant to the given question, e.g., the given question may strongly relate to only a small part of the image (i.e., image attention region) rather than the whole image. Therefore, it is intuitive to introduce *visual attention* mechanism [5] into the VQA task to adaptively learn the most relevant image regions for a given question. Modeling visual attention may significantly improve performance [8]. On the other hand, the questions interpreted

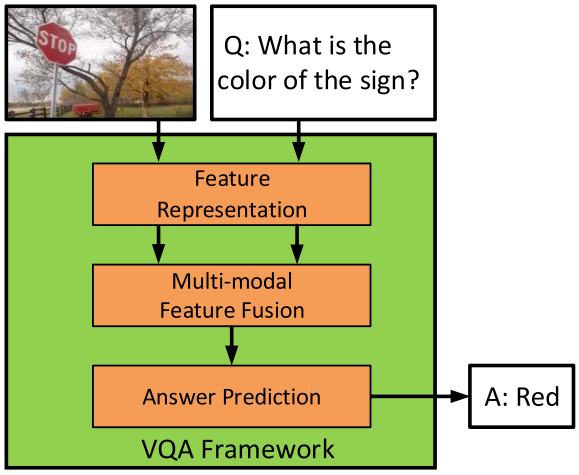


Fig. 1. General framework for the VQA task. Given an arbitrary image and an open vocabulary question (Q) as the inputs, the VQA model outputs the answer (A) in natural language.

in natural languages may also contain colloquialisms that can be treated as noise, thus it is very important to model the question attention simultaneously. Unfortunately, most existing approaches only model the image attention without considering the question attention. Motivated by these observations, we design a deep network architecture for the VQA task using a *coattention learning* module to jointly learn the attentions for both the image and the question, which may allow us to extract more discriminative features for image and question representations.

For multimodal feature fusion, most existing approaches simply use linear models (e.g., concatenation or elementwise addition) to integrate the visual feature from the image with the textual feature from the question even their distributions may vary dramatically [9], [10]. Such linear models may not be able to generate expressive image–question features that are able to fully capture the complex correlations between multimodal features. In contrast to linear pooling, bilinear pooling [11] has recently been used to integrate different CNN features for fine-grained image recognition [12]. Unfortunately, such bilinear pooling approach may output high-dimensional features for image–question representation, and the underlying deep networks for feature extraction may contain huge number of model parameters, which may seriously limit its applicability for VQA. To tackle these problems effectively, Multimodal compact bilinear (MCB) pooling [8] and Multimodal low-rank bilinear (MLB) pooling [13] have been developed to reduce the computational complexity of the original bilinear pooling model and make it practicable for VQA. However, MCB needs very high-dimensional feature to guarantee good performance, and MLB needs a great many training iterations to converge to a satisfactory solution. To tackle these problems, we propose a multimodal factorized bilinear pooling approach (MFB) that enjoys the dual benefits of compact output features of MLB and robust expressive capacity of MCB. Moreover, we extend the bilinear MFB model to a generalized high-order setting and proposed a multimodal factorized high-order pooling (MFH) method to achieve more effective fusion of multimodal features by exploiting their complex correlations sufficiently. By introducing more complex high-order interactions between multimodal features, our MFH method can achieve more discriminative image–question representation and further result in significant improvement on the VQA performance.

For answer prediction, some data sets such as VQA [6] provides multiple answers for each image–question pair and such diverse answers are typically annotated by different users. As the answers are represented in natural languages, for a given question, different users may provide diverse answers or expressions that have same or similar meaning, thus such diverse answers may have strong correlations and they are not independent at all. For example, both *a little dog* and *a puppy* could be the correct answers for the same question. Motivated by these observations, it is important to design an appropriate mechanism to model the complex correlations between multiple diverse answers for the same question. In MCB [8], an *answer sampling* strategy was proposed to randomly pick an answer from a set of candidates during the training course. In this way, the complex correlations between multiple diverse answers could be eventually learned by the model with sufficient training iterations. In this paper, we formulate the problem of answer prediction as a *label distribution learning* (LDL) problem. The answers for an image–question pair in the training data set are converted to a probability distribution over all possible answers. We use the Kullback–Leibler divergence (KLD) as the loss function to achieve more accurate characterization of the consistency between the probability distribution of the predicted answers and the probability distribution of the ground truth answers given by the annotators. Compared with the answer sampling method in MCB [8], using the KLD loss can achieve faster convergence rate and obtain slightly better accuracy on answer prediction.

In summary, we have made the following contributions in this paper.

1) A *coattention* learning architecture is designed to jointly learn the attentions for both the image and the question, which can allow us to reduce the irrelevant features (i.e., noisy information) effectively and obtain more discriminative features for image and question representations.
2) A MFB approach is developed to achieve more effective fusion of the visual features from the image and the textual features from the question. By supporting more effective exploitation of the complex correlations between multimodal features, our MFB approach can significantly outperform the existing bilinear pooling approaches.
3) A generalized MFH approach is developed by cascading multiple MFB blocks. Compared with MFB, MFH captures more complex correlations of multimodal feature to achieve more discriminative image–question representation and further result in significant improvement on the VQA performance.
4) The KLD is used as the loss function to achieve more accurate characterization of the consistency between the

predicted answers and the annotated answers, which can allow us to achieve faster convergence rate and obtain slightly better accuracy on answer prediction.

5) Extensive experiments over multiple VQA data sets are conducted to explain the reason why our approaches are effective. Our experimental results demonstrate that a) our proposed approaches can achieve the state-of-the-art performance on the real-world VQA data sets; b) the normalization techniques are extremely important in bilinear pooling models.

The rest of the paper is organized as follows. In Section II, we review the related work of VQA approaches, especially the ones introducing the bilinear pooling. In Section III, we revisit the bilinear model and its factorized extension. Then, we propose the bilinear MFB model and reveal the fact that MFB is a generalization form of MLB. Based on MFB, we further propose its generalized high-order extension MFH. In Section IV, we propose the coattention learning network architecture for VQA based on MFB or MFH. In Section V, we analyze the importance of modeling answer correlation in VQA and propose a solution with KLD loss. In Section VI, we introduce our extensive experimental results for algorithm evaluation, and multiple real-word VQA data sets are used to evaluate our proposed approaches. Finally, we conclude this paper in Section VII.

II. Related Work

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

A. Visual Question Answering

Malinowski and Fritz [7] made an early attempt at solving the VQA task. Since then, solving the VQA task has received increasing attention from the communities of computer vision and natural language processing. Most existing VQA approaches can be classified into the following three categories: 1) the coarse joint embedding models [6], [9], [14]; 2) the fine-grained joint embedding models with attention [8], [10], [15]–[17]; and 3) the external knowledge-based models [18], [19].

The coarse joint embedding models are the most straightforward solution for VQA. Image and question are first represented as global features, and then integrated to predict the answer. Zhou *et al.* [9] proposed a baseline approach for the VQA task using the concatenation of the image CNN features and the question bag-of-words features, and a linear classifier is learned to predict the answer. Wang and Ji [20] perform a detailed analysis on the modeling of questions using CNN to obtain better question representations for VQA. Some approaches introduce more complex deep models, e.g., long-short term memory (LSTM) networks [6] or residual networks [14], 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 (i.e., irrelevant features), and such noisy global features may not be able to answer the fine-grained problems correctly (e.g., "what color are the cat's eyes?"). Therefore, recent VQA approaches introduce the *visual attention* mechanism [5] into the VQA task by adaptively learning the local fine-grained image features for a given question. Chen *et al.* [21] 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. Yang *et al.* [22] proposed a stacked attention network to learn the attention iteratively. Some approaches introduce off-the-shelf object detectors [16] or object proposals [23] as the candidates of the attention regions and then use the question to identify the relevant ones. Fukui *et al.* [8] proposed MCB pooling to integrate the visual features from the image spatial grids with the textual features from the questions to predict the attention. As the VQA task need to fully understand the semantic of the question in natural language, it is necessary to learn the *textual attention* for question simultaneously. Inspired by the works from the Natural Language Processing community [24], [25], some approaches perform attention learning on both the images and the questions. Lu *et al.* [10] proposed a coattention learning framework to alternately learn the image attention and the question attention. Nam *et al.* [17] proposed a multistage coattention learning model to refine the attentions based on memory of previous attentions.

Despite the joint embedding models can deliver impressive VQA performance, they are not good enough for answering the questions that require complex reasoning or knowledge of common senses. Therefore, introducing external knowledge is beneficial for VQA. However, existing approaches have either only been applied to specific data sets [18] or have been ineffective on benchmark data sets [19]. Thus, they still have rooms for further exploration and development.

B. Multimodal Bilinear Models for VQA

Multimodal feature fusion plays a critical and fundamental role in VQA. Once the image and the question representations are obtained, concatenation or elementwise summations are most frequently used for multimodal 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 simply fused features may be insufficient, limiting the final prediction performance.

Fukui *et al.* [8] first introduced the bilinear model to solve the problem of multimodal feature fusion in VQA. In contrast to the aforementioned approaches, they proposed the MCB pooling, which uses the outer product of two feature vectors in different modalities to produce a very high-dimensional feature for quadratic expansion [8]. 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 data set [6]. Nevertheless, MCB usually needs high-dimensional features (e.g., 16 000-D) to guarantee robust performance, which may seriously limit its applicability for VQA due to the limitations in GPU memory.

To overcome this problem, Kim *et al.* [13] proposed the MLB pooling 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

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

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 elementwise multiplication of two vectors. To further increase the model capacity, nonlinear activation such as *tanh* is added after z. Since the MLB approach can generates feature vectors with low dimensions and deep networks with fewer model parameters, it has achieved very comparable performance to MCB. In MLB [13], the experimental results indicated that MLB may lead to a slow convergence rate (the MLB with attention model takes 250 000 iterations with the batch size 200, which is about 140 epochs, to converge [13]).

III. Generalized Multimodal Factorized High-Order Pooling

In this section, we first revisit the multimodal bilinear models and then introduce the MFB pooling model. We give detailed explanation on the implementation of our MFB model and further analyze its relationship with the existing MLB approach [13]. By treating our MFB model as the basic building block, we extend the idea of bilinear pooling to a generalized high-order pooling, and we further propose a multimodal high-order pooling (MFH) model by simply cascading multiple MFB blocks to capture more complex high-order interactions between multimodal features.

A. Multimodal 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 multimodal bilinear model is defined as follows:

$$z_i = x^T W_i y \tag{2}$$

where $W_i \in \mathbb{R}^{m \times n}$ is a projection matrix and $z_i \in \mathbb{R}$ 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 $W = [W_i, \ldots, W_o] \in \mathbb{R}^{m \times n \times o}$. 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 overfitting.

Inspired by the matrix factorization tricks for unimodal data [26]–[30], the projection matrix W_i in (2) can be factorized as two low-rank matrices

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

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

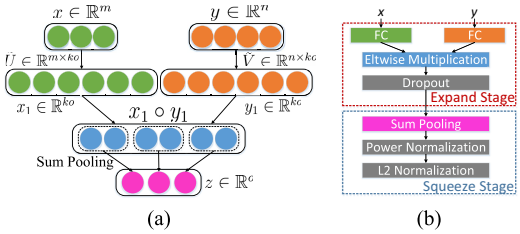


Fig. 2. Flowchart of MFB pooling and complete MFB module. (a) MFB pooling. (b) MFB module.

To obtain the output feature $z \in \mathbb{R}^o$ by (3), the weights to be learned are two three-order tensors $U = [U_1, \ldots, U_o] \in \mathbb{R}^{m \times k \times o}$ and $V = [V_1, \ldots, V_o] \in \mathbb{R}^{n \times k \times o}$ 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, (3) is rewritten as follows:

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

where the function $\text{SumPool}(x, k)$ means using a 1-D nonoverlapped window with the size k to perform sum pooling over x. We name this model as MFB pooling.

The detailed procedures of MFB are illustrated in Fig. 2(a). The approach can be easily implemented by combining some commonly used layers such as fully connected, elementwise multiplication, and pooling layers. Furthermore, to prevent overfitting, a dropout layer [31], [32] is added after the elementwise multiplication layer. Since elementwise multiplication is introduced, the magnitude of the output neurons may vary dramatically, and the model might converge to an unsatisfactory local minimum. Therefore, similar to [8], 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. 2(b).

Relationship to MLB: Equation (4) shows that the MLB in (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 Fig. 2(b)]: first, the features from different modalities are *expanded* to a high-dimensional space and then integrated with elementwise 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 elementwise multiplication. Therefore, with the same dimensionality for the output features, we can conjecture that MLB may suffer from insufficient representation.

B. From Bilinear Pooling to Generalized High-Order Pooling

From the previous work such as [8] and [14], we have witnessed that the bilinear pooling models have superior

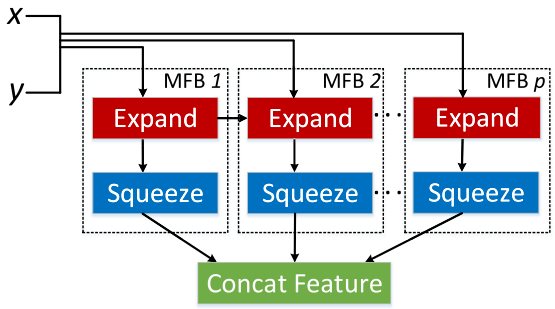


Fig. 3. Flowchart of the MFH^p module based on the cascading of p MFB blocks. Note that MFB is a special case of MFH^p with $p = 1$.

representation capacity than the traditional linear pooling models. This inspires us that exploiting the complex interactions among the feature dimensions is beneficial for capturing the common semantics of multimodal features [33], [34]. Therefore, a natural idea is to extend the second-order bilinear pooling to the generalized high-order pooling to further enhance the representation capacity of fused features. In this section, we introduce a generalized MFH model by cascading multiple MFB blocks.

As shown in Fig. 2(b), the MFB module can be separated into the expand stage and the squeeze stage as follows:

$$z_{\text{exp}} = \text{MFB}_{\text{exp}}(x, y) = \text{Dropout}(\tilde{U}^T x \circ \tilde{V}^T y) \in \mathbb{R}^{ko} \quad (5)$$

$$z = \text{MFB}_{sqz}(z_{\text{exp}}) = \text{Norm}(\text{SumPool}(z_{\text{exp}})) \in \mathbb{R}^{o} \quad (6)$$

where Drop$(\cdot)$, SumPool$(\cdot)$, and Norm$(\cdot)$ refer to the dropout, sum pooling, and normalization layers, respectively. z_{exp} and z are the internal and the output features of the MFB module, respectively.

To make p MFB blocks cascadable, we slightly modify the original MFB$_{\text{exp}}$ stage in (5) as follows:

$$z^i_{\text{exp}} = \text{MFB}^i_{\text{exp}}(x, y) = z^{i-1}_{\text{exp}} \circ (\text{Dropout}(\tilde{U^i}^T x \circ \tilde{V^i}^T y)) \quad (7)$$

where $i \in \{1, 2, \ldots, p\}$ is the index for the MFB blocks. $\tilde{U^i}$, $\tilde{V^i}$, and z^i_{exp} are the weight matrices and the internal feature for ith MFB block, respectively. z^{i-1}_{exp} is the internal feature of $i-1$th MFB block and $z^0_{\text{exp}} \in \mathbb{1}^{ko}$ is an all-one vector.

Once the internal feature z^i_{exp} is obtained for ith MFB block, the output feature z^i for ith MFB block can be computed by (6). The final output feature z of the high-order MFH^p model is obtained by concatenating the output feature of p MFB blocks as follows:

$$z = \text{MFH}^p = [z^1, z^2, \ldots, z^p] \in \mathbb{R}^{op}. \quad (8)$$

The overall flowchart of the MFH approach is illustrated in Fig. 3. With the increase of p, the model size and the dimensionality of the output feature for MFH grow linearly. In order to control the model complexity and the training time that we can afford, we use $p < 4$ in our experiments. It is worth noting that the propoed MFB model in Section III-A is a special case of our MFH^p model with $p = 1$.

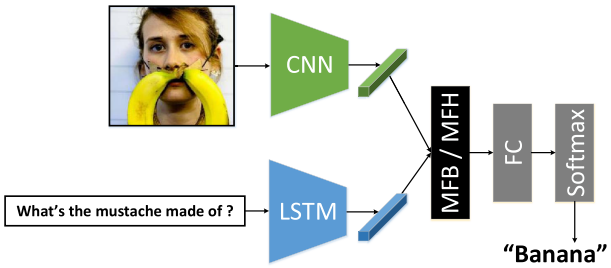


Fig. 4. Baseline network architecture with MFB or MFH and without the attention mechanism for VQA.

IV. 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 the corresponding question about the image. Our model extracts the representations for both the image and the question, integrates multimodal features using the MFB or MFH modules in Fig. 2(b), treats each individual answer as one class and performs multiclass classification to predict the correct answer. In this section, two network architectures are introduced. The first one is the baseline with one MFB or MFH module, which is used to perform ablation analysis with different hyper-parameters for comparison with other baseline approaches. The second one introduces coattention learning to achieve more effective characterization of the fine-grained correlations between multimodal features, which may result in a model with better representation capability.

A. Baseline Model

Similar to MCB [8], we extract the image features using the 152-layer ResNet model [35], which is pretrained on the ImageNet data set. 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 LSTM networks with 1024 hidden units [36]. Similar to MCB [8], we extract the output feature of the last word from the LSTM network to form a 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 multimodal features (that are extracted from the image and the question) are fed to the MFB or MFH module to generate the fused image–question feature z. Finally, z is fed to an N-way classifier to predict the best-matching answer. 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 Fig. 4.

B. Coattention Model

For a given image, different questions could result into an entire different set of answers. Therefore, an *image attention* model, which can predict the relevance between each spatial

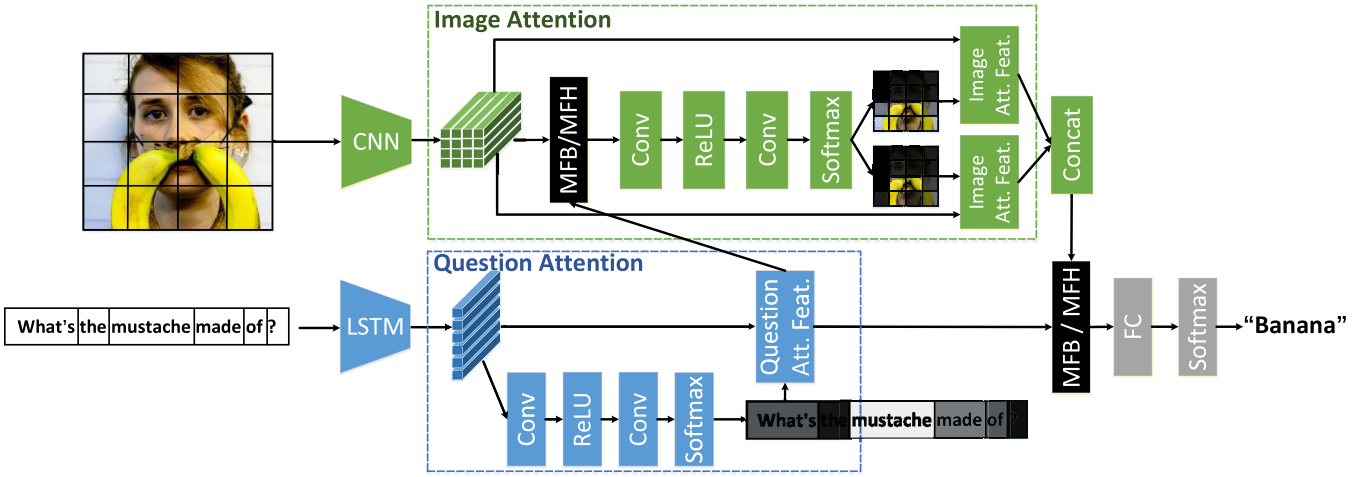


Fig. 5. Coattention network architecture with MFB or MFH for VQA. Different from the network of MFB baseline, the images and questions are first represented as the fine-grained features, respectively. Then, *Question Attention* and *Image Attention* modules are jointly modeled in the framework to provide more accurate answer predictions. For both the image and question attention modules, multiple attention maps (see the example in the image attention module) can be adapted to further improve the representation capacity of the fine-grained features.

grid of the image with the question, is beneficial for predicting the best-matching answer accurately. From the results reported in MCB [8], one can see that incorporating such image attention mechanism allows the model to effectively learn which image region is important for the question, clearly contributing to better performance than the models without using attention. However, their attention model only focuses on learning the image attention, while completely ignoring the question attention. Since the questions are interpreted in natural languages, the contribution of each word is definitely different. Therefore, we develop a coattention learning approach named MFB + CoAtt or MFH + CoAtt (see Fig. 5) to jointly learn the attentions for both the question and the image.

Specifically, 14×14 (196) spatial grids of the image (*res5c* feature maps in ResNet) are used to represent the input image and T output features from the LSTM networks are used to represent each word in the input question. After that, the T question features are fed into a *question attention* module and output an attentive question representation. This attentive question representation is fed into an *image attention* module (with 196 image features), and MFB or MFH is used to generate a fused image–question representation. Such fused image–question representation is further used to learn a multiclass classifier for answer prediction. In our excrements, we find that using MFH rather than MFB in the image attention module does not improve the prediction accuracy significantly while inducing much higher computational cost. Therefore, in most of our experiments (unless in the final model ensemble experiment), the MFH module is only used in the feature fusion stage for integrating the attentive features extracted from the image and the question.

Both the image attention module and question attention module consist of sequential 1×1 convolutional layers and ReLU layers followed by the softmax normalization layers to predict the attention weight for each input feature. The attentive feature is obtained by the weighted sum of the input features. To further improve the representation capacity of the attentive feature, multiple attention maps are generated to enhance the learned attention map, and these attention maps are concatenated to output the attentive image features.

It is worth noting that the question attention in our network architecture is learned in a *self-attentive* manner using the question feature itself. This is different from the image attention module which is learned using both the image features and question features. The reason is that we assume that the question attention (i.e., the key words of the question) can be inferred without seeing the image, as humans do.

V. Answer Correlation Modeling

In most existing VQA approaches, the answering stage is formulated as a multiclass classification problem and each answer refers to an individual class. In practice, this assumption may not hold for the VQA task, because the answers with the same or similar meaning can be expressed diversely by different annotators. For example, both the answers "*a little dog*" and "*a puppy*" could be correct for a given image–question pair. Therefore, it is crucial to model the answer correlations in the VQA task so that the learned model could be more robust.

In some data sets such as VQA [6], each question is annotated with multiple answers by different people. To exploit the answer correlations, an *answer sampling* strategy was used in MCB [8]. Specifically, for each image–question pair in the training set, the multiple answers for each sample are represented as a distribution vector of all the possible answers $y \in \mathbb{R}^N$, where N is the total number of answers for the whole training set. $y_i \in [0, 1]$ indicates the occurrence probability of the ith answer with that $\sum_i y_i = 1$. In each epoch the sample is accessed, a single answer is obtained by sampling from probability distribution y as the label for this sample in this epoch. In this way, the problem become the traditional multiclass classification problem with single label and traditional softmax loss function could be used to train

the model. With sufficient number of iterations, the model can learn the answer correlation eventually.

In practice, using answer sampling strategy may introduce uncertainty to the learned model and may take more iterations to converge. To overcome the problem, we transform the single-label multiclass classification problem with sampled answers to the LDL problem [37] with the answer distribution y. Accordingly, we use the KLD loss function to penalize the prediction $z \in \mathbb{R}^N$ after the softmax activation of the last fully connected layer

$$\ell(y, z)_{\mathrm{KL}} = \sum_i y_i \log\left(\frac{y_i}{z_i}\right). \tag{9}$$

Note that KLD loss contains an additional constant term compared to the multilabel cross-entropy loss. They are equivalent during optimization.

VI. Experiments

We have conducted several experiments to evaluate the performance of our MFB models for the VQA task using the VQA data sets [6], [38] to verify our approach. We first perform ablation analysis on the MFB and MFH baseline models to verify the superior performance of the proposed approaches over the existing state-of-the-art methods such as MCB [8] and MLB [13]. We then provide detailed analysis of the reasons why our models outperform their counterparts. Finally, we choose the optimal hyperparameters for the MFB or MFH module and train the models with coattention for fair comparison with the state-of-the-art approaches on the real-world VQA data sets. The corresponding source codes and pretrained models are released online.[1]

A. Data Sets and Evaluation Criteria

We have evaluated the performances of our proposed approaches over multiple VQA data sets. In addition, we have compared our proposed approaches with the state-of-the-art algorithms.

1) VQA-1.0: The VQA-1.0 data set [6] consists of approximately 200 000 images from the MS-COCO data set [39], with three questions per image and 10 answers per question. The data set is split into three: *train* (80 000 images and 240 000 question–answer pairs), *val* (40 000 images and 120 000 question–answer pairs), and *test* (80 000 images and 240 000 question–answer pairs). Additionally, there is a 25% test 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.* [6] to evaluate the accuracy on the two tasks. Specifically, the accuracy of a predicted answer a is calculated as follows:

$$\mathrm{Accuracy}(a) = \min\left\{\frac{\mathrm{Count}(a)}{3}, 1\right\} \tag{10}$$

where $\mathrm{Count}(a)$ is the count of the answer a voted by different annotators.

[1]https://github.com/yuzcccc/vqa-mfb

2) VQA-2.0: The VQA-2.0 data set [38] is the updated version of the VQA data set. Compared with the VQA data set, it contains more training samples (440 k question-answer pairs for training and 214 000 pairs for validation), and is more balanced to weaken the potential that an overfitted model may achieve good results. Specifically, for every question there are two images in the data set that result in two different answers to the question. At this point only the train and validation sets are available. Therefore, we report the results of the OE task on validation set with the model trained on train set. The evaluation criterion on this data set is same as the one used in the VQA-1.0 data set.

B. Experimental Setup

For the VQA and VQA 2.0 data sets, we use the Adam solver with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The base learning rate is set to 0.0007 and decays every 40 000 iterations using an exponential rate of 0.5 for MFB and 0.25 for MFH. All the models are trained up to 100 000 iterations. Dropouts are used after each LSTM layer (dropout ratio $p = 0.3$) and MFB and MFH modules ($p = 0.1$). The number of answers $N = 3000$. For all experiments (except for the ones shown in Table I, which use the train and val sets together as the training set like the comparative approaches), we train on the train set, validate on the val set, and report the results on the test-dev and test-standard sets.[2] 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 [40] and performed on the workstations with NVIDIA TitanX GPUs.

C. Ablation Study on the VQA-1.0 Data Set

We design the following ablation experiments to verify the efficacy of our MFB and MFH modules, as well as the advantage of the KLD loss in modeling answer correlations.

1) Design of the MFB and MFH Module: In Table II, we compare the performance of MFB and MFH with other baseline multimodal fusion models (i.e., feature concatenation, elementwise summation, elementwise product, and their variants with one additional fully connected layer followed by ReLU activation). Besides, the state-of-the-art multimodal bilinear pooling models, namely, MCB [8] and MLB [13], are fairly compared. The models are trained on the train set and evaluated on the test-dev set. For fair comparison, all the compared bilinear pooling approaches use power + ℓ_2 normalizations. None of these approaches introduces the attention mechanism. We explore different hyperparameters and normalizations introduced in MFB to explore why MFB outperform the compared bilinear models. Finally, we evaluate MFH^p with different p to explore the effect of high-order feature pooling.

From Table II, we can see that, first, MFB significantly outperforms all the baseline multimodal fusion models. MFB is at

[2]The submission attempts for the test set are strictly limited. Therefore, we report most of our results on the test-dev set and the best results on the test-standard set.

TABLE I

OE AND MC RESULTS ON VQA-1.0 DATA SET COMPARED WITH THE STATE-OF-THE-ART APPROACHES IN TERMS OF ACCURACY IN PERCENTAGE. ATT. INDICATES WHETHER THE APPROACH INTRODUCE THE ATTENTION MECHANISM EXPLICITLY, W.E. INDICATES WHETHER THE APPROACH USES EXTERNAL WORD EMBEDDING MODELS, AND E.D. INDICATES WHETHER THE APPROACH USES EXTERNAL DATA SETS. 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.	E.D.	Accuracy									
				Test-dev					Test-Standard				
				OE				MC	OE				MC
				All	Y/N	Num	Other	All	All	Y/N	Num	Other	All
iBOWIMG [9]				55.7	76.5	35.0	42.6	-	55.9	78.7	36.0	43.4	62.0
DPPnet [42]				57.2	80.7	37.2	41.7	-	57.4	80.3	36.9	42.2	-
VQA team [6]				57.8	80.5	36.8	43.1	62.7	58.2	80.6	36.5	43.7	63.1
AYN [43]				58.4	78.4	36.4	46.3	-	58.4	78.2	36.3	46.3	-
AMA [19]				59.2	81.0	38.4	45.2	-	59.4	81.1	37.1	45.8	-
MCB [8]				61.1	81.7	36.9	49.0	-	61.1	81.7	36.9	49.0	-
MRN [14]				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	-	-	-	-	-
MFH (Ours)				**62.9**	**83.1**	36.8	**51.5**	**67.9**	-	-	-	-	
SMem [44]	✓			58.0	80.9	37.3	43.1	-	58.2	80.9	37.3	43.1	-
NMN [45]	✓			58.6	81.2	38.0	44.0	-	58.7	81.2	37.7	44.0	-
SAN [22]	✓			58.7	79.3	36.6	46.1	-	58.9	-	-	-	-
FDA [16]	✓			59.2	81.1	36.2	45.8	-	59.5	-	-	-	-
DNMN [15]	✓			59.4	81.1	38.6	45.4	-	59.4	-	-	-	-
HieCoAtt [10]	✓			61.8	79.7	38.7	51.7	65.8	62.1	-	-	-	-
RAU [46]	✓			63.3	81.9	39.0	53.0	67.7	63.2	81.7	38.2	52.8	67.3
MCB+Att [8]	✓			64.2	82.2	37.7	54.8	-	-	-	-	-	-
DAN [17]	✓			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	-	-	-	-	-
MFH+CoAtt (Ours)	✓			**65.8**	**84.1**	38.1	**56.5**	**70.6**	-	-	-	-	-
MCB+Att+GloVe [8]	✓	✓		64.7	82.5	37.6	55.6	-	-	-	-	-	-
MLB+Att+StV [13]	✓	✓		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
MFH+CoAtt+GloVe (Ours)	✓	✓		**66.8**	**85.0**	39.7	**57.4**	**71.4**	66.9	85.0	**39.5**	57.4	71.5
MCB+Att+GloVe+VG [8]	✓	✓	✓	65.4	82.3	37.2	57.4	-	-	-	-	-	-
MLB+Att+StV+VG [13]	✓	✓	✓	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
MFH+CoAtt+GloVe+VG (Ours)	✓	✓	✓	**67.7**	**84.9**	**40.2**	**59.2**	**72.3**	**67.5**	**84.9**	39.3	**58.7**	**72.1**

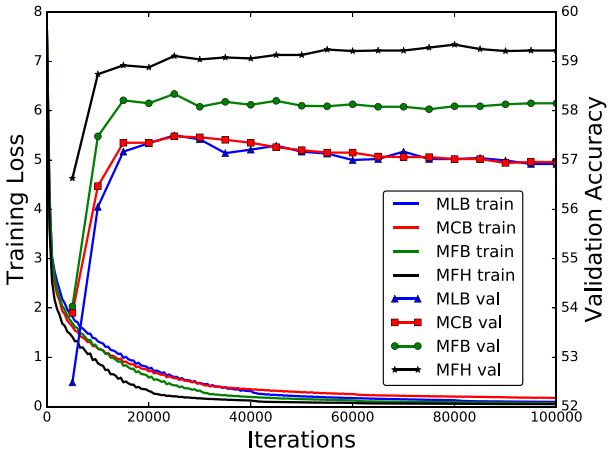


Fig. 6. Training loss and validation accuracy versus iterations of MCB, MLB, MFB, and MFH^2 ($k = 5, o = 1000$). KLD loss is used for all the methods. Best viewed in color.

least two points higher than the compared baseline models of similar sizes: the MFB ($k = 5$, $o = 200$) model outperforms the EltwiseProd model by 2.1 points, and the MFB ($k = 5$, $o = 1000$) model outperforms the EltwiseProd + fully-connected + ReLU model by 2.2 points. These results demonstrate the advantage of the second-order bilinear pooling models over the first-order pooling models on learning discriminative multimodal feature representations.

Second, MFB outperforms other multimodal bilinear pooling approaches. With 5/6 parameters, MFB ($k = 5$, $o = 1000$) achieves an improvement of about 1.0 points compared with MCB and MLB. Moreover, with only 1/3 parameters and 2/3 GPU memory usage, MFB ($k = 5$, $o = 200$) obtains similar results to MCB. These characteristics allow us to train our model on a memory limited GPU with larger batch size. In Fig. 6, we show the courses of validation, from which it can be seen that MFB significantly outperforms the two other methods in terms of accuracy on the validation set. Furthermore, it can be seen from the accuracy curve of MCB that its performance gradually falls after 25 000 iterations, indicating that it suffers from overfitting with the high-dimensional output features. In comparison, the performance of MFB is relatively robust.

Third, 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 points 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 lose some information. When k is fixed, increasing o does not produce any further improvement. This suggests that high-dimensional output features may be easier to overfit. Similar results can be seen in MCB [8]. In summary, $k = 5$ and $o = 1000$ may be

TABLE II
OVERALL ACCURACIES AND MODEL SIZES OF APPROACHES AND ON THE VQA-1.0 TEST-DEV SET OF THE OE TASK. ALL THE COMPARED APPROACHES USE THE SAME INPUT FEATURES AND DOES NOT INTRODUCE THE EXTERNAL DATA SETS OR THE ATTENTION MECHANISM. THE MODEL SIZE INCLUDES THE PARAMETERS FOR THE LSTM NETWORKS

Model	Acc.	Size
Concat	57.1	29M
Concat+FC(4096)+ReLU	58.4	45M
EltwiseSum	56.4	23M
EltwiseSum+FC(4096)+ReLU	58.3	37M
EltwiseProd	57.8	23M
EltwiseProd+FC(4096)+ReLU	58.7	37M
MCB [8] ($d = 16000$)	59.8	63M
MLB [13] ($d = 1000$)	59.7	25M
MFB($k = 1, o = 5000$)	60.4	51M
MFB($k = 5, o = 1000$)	60.9	46M
MFB($k = 10, o = 500$)	60.5	38M
MFB($k = 5, o = 200$)	59.8	22M
MFB($k = 5, o = 500$)	60.4	28M
MFB($k = 5, o = 2000$)	60.6	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	-
MFH^2($k = 5, o = 1000$)	**61.6**	62M
MFH^3($k = 5, o = 1000$)	61.5	79M

a suitable combination for our MFB model on the VQA data set, so we use these settings in our follow-up experiments.

Fourth, both the power and ℓ_2 normalization benefit MFB performance. Power normalization results in an improvement of 0.5 points and ℓ_2 normalization, perhaps surprisingly, results in an improvement of about three points. Results without ℓ_2 and power normalizations were also reported [6] and are similar to those reported here. To explain why normalization is so important, we randomly choose one typical neuron from the MFB output feature before normalization to illustrate how its distribution evolves over time in Fig. 7. It can be seen that the standard MFB model (with both normalizations) leads to the most stable neuron distribution (i.e., small neuron variance) and without the power normalization, about 10 000 iterations are needed to achieve stabilization. Without the ℓ_2 normalization, the distribution varies seriously over the entire training course. This observation is consistent with the results shown in Table II. The effects of power normalization and ℓ_2 normalization are also observed in [41]. Furthermore, although MLB does not use any normalization, it introduces the *tanh* activation after the fused feature, which regularizes the distribution of the output feature in some way.

Finally, MFH^2 and MFH^3 further outperform MFB with an improvement of about 0.7 points on the test-dev set. This observation demonstrates the efficacy of high-order pooling model for VQA. However, the performance of MFH^3 is slightly worse than MFH^2 even with a more complex model. This may be explained that the representation capacity of MFH is saturated with $p = 2$ for the VQA task. Therefore, in our following experiments, $p = 2$ is used for MFH and the superscript p is omitted for simplicity.

2) Answer Correlation Modeling Strategies: In Fig. 8, the validation accuracies of MFB and MFB + CoAtt models with respect to different answer sampling strategies are demonstrated, respectively. *Max Prob* means using the most frequent answer of the sample as the unique label and formulate the optimization for VQA as the traditional multiclass problem with single label. This strategy refer to the baseline approach that does not consider answer correlation. *Answer Sampling* is the strategy used in MCB [8], which random sample an answer from the candidate answer set at each time. KLD is the strategy proposed in Section V of this paper.

From the results, we have the following observations. First, modeling answer correlation bring remarkable improvement on the VQA-1.0 data set. The *Answer Sampling* and *KLD* strategies which model the answer correlation, significantly outperform the *Max Prob* strategy. Second, compared with the *Answer Sampling* strategy, the proposed KLD strategy has the merits of faster convergence rate and slightly better accuracy, especially on the complex MFB + CoAtt model.

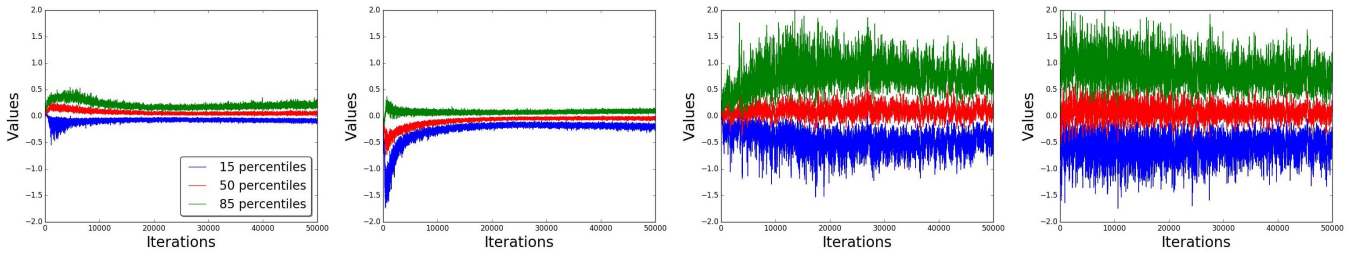
D. Results on the VQA-1.0 Data Set

Table I 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 pretrained word embedding models, e.g., GloVe [47] or Skip-thought Vectors (StV) [48]; and the last includes the models trained with the external large-scale Visual Genome data set [49] additionally. To best utilize model capacity, the training data set is augmented so that both the train and val sets are used as the training set. Also, to better understand the question semantics, pretrained 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 [8]. The MFB + CoAtt model represents the network shown in Fig. 5. The MFB + CoAtt + GloVe model additionally concatenates the learned word embedding with the pretrained GloVe vectors. The MFB + CoAtt + GloVe + Visual Genome dataset (VG) model further introduce the data from the Visual Genome data set [49] into the training set.

From Table I, we have the following observations.

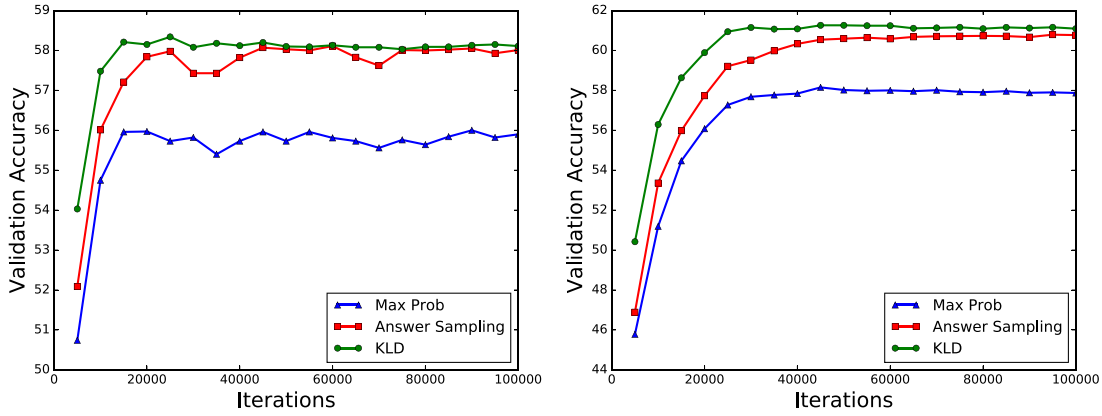
First, the model with MFB outperforms other comparative approaches significantly. The MFB baseline outperforms all other existing approaches without the attention mechanism for both the OE and MC tasks, and even surpasses some approaches with attention. When attention is introduced, MFB + Att consistently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

Second, the coattention model further improve the performance over the attention model with only considering the image attention. By additionally introducing the self-attention module for questions, MFB + CoAtt delivers an improvement of 0.5 points on the OE task compared with the MFB + Att model in terms of overall accuracy. Moreover, for each question type (i.e., Y/N, Num, or Others), the



(a) (b) (c) (d)

Fig. 7. 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. (a) Standard. (b) Without power norm. (c) Without ℓ_2 norm. (d) Without power and ℓ_2 norms.



(a) (b)

Fig. 8. Validation accuracies of MFB and MFB + CoAtt models with respect to different answer correlation modeling strategies. (a) MFB with respect to different strategies. (b) MFB + CoAtt with respect to different strategies.

improvement of MFB + CoAtt over MFB + Att is significant, indicating the effect of the self-attention module in our coattention learning framework.

Third, by replacing MFB with MFH, the performance of all of our models further enjoy an improvement of about 0.7∼1.1 points steadily. The performance of a single MFH + CoAtt + GloVe model has even surpassed the best published results with an ensemble of seven MLB or MFB models shown in Table III on the test-standard set.

Finally, with the external pretrained GloVe model and the Visual Genome data set, the performance of our models are further improved. The MFH + CoAtt + GloVe + VG model significantly outperforms the best reported results with a single model on both the OE and MC task.

In Table III, we compare our model with the state-of-the-art results with model ensemble. Similar with [8] and [13], we train seven individual MFB (or MFH) + CoAtt + GloVe models and average the prediction scores of them. Four of the seven models additionally introduce the Visual Genome data set [49] into the training set. All the reported results are fetched from the leaderboard of the VQA-1.0 data set.[3] For fair comparison, only the published results are demonstrated. From the results, the ensemble of MFB models outperforms the next best result by 1.5 points on the OE task and by 2.2 points on the MC task, respectively. Furthermore,

[3]The *Standard* tab in http://www.visualqa.org/roe.html

TABLE III

COMPARISON WITH THE STATE-OF-THE-ART RESULTS (WITH MODEL ENSEMBLE) ON THE TEST-STANDARD SET OF THE VQA-1.0 DATA SET. ONLY THE PUBLISHED RESULTS ARE DEMONSTRATED. THE BEST RESULTS ARE BOLDED

Model	OE				MC
	All	Y/N	Num	Other	All
HieCoAtt [10]	62.1	80.0	38.2	52.0	66.1
RAU [46]	64.1	83.3	38.0	53.4	68.1
7 MCB models [8]	66.5	83.2	39.5	58.0	70.1
7 MLB models [13]	66.9	84.6	39.1	57.8	70.3
7 MFB models	68.4	85.6	41.0	59.8	72.5
7 MFH models	**69.2**	**86.2**	**41.8**	**60.7**	**73.4**
Human [6]	83.3	95.8	83.4	72.7	91.5

the result of the ensemble of MFH models obtains a further improvement of 0.8 points and achieves the new state-of-the-art approach. Finally, compared with the results obtained by human, there is still a lot of room for improvement to approach the human level.

To demonstrate the effects of coattention learning, we visualize the learned question and image attentions of some image–question pairs from the val set in Fig. 9. The examples are randomly picked from different question types. It can 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

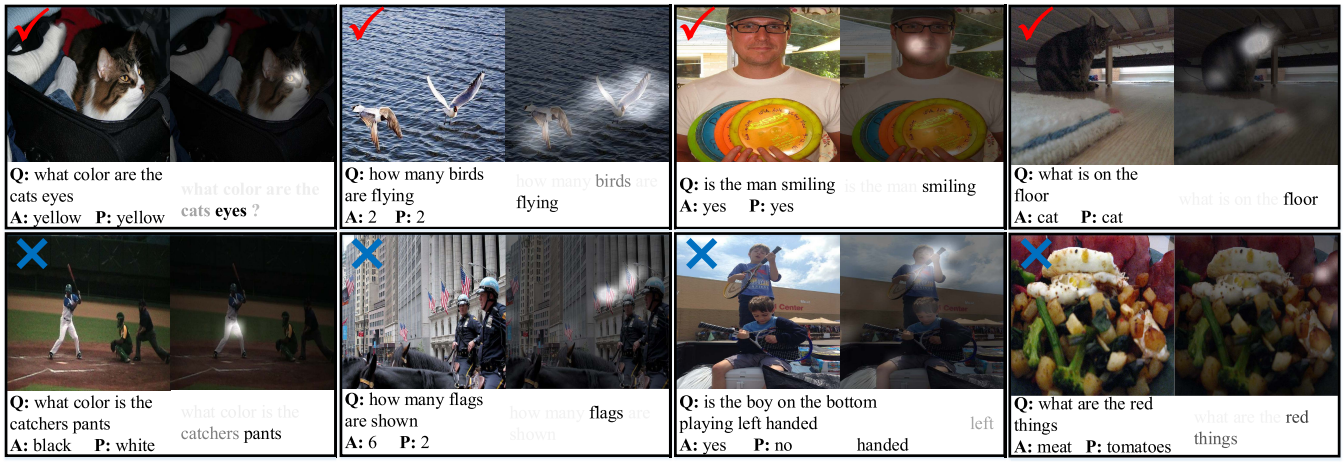


Fig. 9. Typical examples of the learned image and question of the MFB + CoAtt + GloVe model on the VQA-1.0 data set. Top row: four examples of four correct predictions. Bottom row: four incorrect predictions. For each example, the query image, question (Q), answer (A), and prediction (P) are presented 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 IV

OVERALL ACCURACIES ON THE TEST-DEV AND TEST-CHALLENGE SETS OF THE VQA-2.0 DATA SET

Model	Test-Dev	Test-Challenge
vqateam-Prior	-	25.98
vqateam-Language	-	44.34
vqateam-LSTM-CNN	-	54.08
vqateam-MCB	-	62.33
Adelaide-ACRV-MSR [51] (1st place)	-	**69.00**
DLAIT (2nd place)	-	68.07
LV_NUS (4th place)	-	67.62
1 MFB model	64.98	-
1 MFH model	65.80	-
7 MFB models	67.24	-
7 MFH models	67.96	-
9 MFH models (2nd place)	**68.02**	68.16

TABLE V

OVERALL ACCURACY ON THE TEST-DEV SET OF THE VQA-2.0 DATA SET. MFH MODELS WITH DIFFERENT HYPERPARAMETERS ARE REPORTED. VG INDICATES WHETHER THE TRAINING SET IS AUGMENTED WITH VISUAL GENOME. MFH/ MFB(I) MEANS WHETHER THE MFH OR MFB MODULE IS USED IN THE IMAGE ATTENTION MODULE; # Q_{att} AND # I_{att} INDICATE THE NUMBER OF GLIMPSES (I.E., ATTENTION MAPS) FOR THE QUESTION AND IMAGE ATTENTION MODULES, RESPECTIVELY

index	VG	MFH / MFB(I)	# Q_{att}	# I_{att}	Accuracy(%)
1		MFB	1	2	65.70
2		MFB	2	2	65.74
3		MFB	2	3	65.80
4	✓	MFB	1	2	65.95
5	✓	MFB	2	2	66.12
6	✓	MFB	2	3	66.01
7	✓	MFH	1	2	65.93
8	✓	MFH	2	2	66.12
9	✓	MFH	2	3	66.03

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) and 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 improvement for the VQA task in the future.

E. Results on the VQA-2.0 Data Set

Table IV demonstrates our results on the VQA-2.0 data set (also known as, VQA challenge 2017). We compare our models with the results of baseline models (including the MCB model which is the champion of VQA Challenge 2016) and the results of the top-ranked teams on the leaderboard. We use the same training strategies aforementioned for this data set.

From the results, our single MFB and MFH models (with CoAtt + GloVe but without the Visual Genome data argumentation) significantly surpass all the baseline approaches. If we neglect the tiny difference between the results on test-dev and test-standard sets, MFB and MFH are about 2.7 points and 3.5 points higher than the MCB model, respectively. Finally, with an ensemble nine models, we report the accuracy of 68.02% on the test-dev set and 68.16% on the test-challenge set, respectively,[4] which ranks the second place (tied with another team) in VQA Challenge 2017. The details of the nine models are illustrated in Table V.

In the solution of the champion team, they introduce the region-based visual features extracted from the Faster R-CNN model, which is pretrained on the large-scale Visual Genome data set [50]. Using these visual features instead of the convolutional features from the ResNet model brings surprisingly good performance even with a simple VQA model. Using their visual features as the backbone for our models with

[4]http://visualqa.org/roe_2017.html

MFH, we are in the first place on the real-time leaderboard of the VQA-2.0 data set up to now (March 15, 2018). We report the overall accuracy 70.92% on the test-standard set of VQA-2.0 with eight models, while they report the accuracy 70.34% with up to 30 models [51].

VII. Conclusion

In this paper, a network architecture with coattention learning is designed to model both the image attention and the question attention simultaneously, so that we can reduce the irrelevant features effectively and extract more discriminative features for the image and question representations. An MFB approach is developed to achieve more effective fusion of the visual features from the images and the textual features from the questions, and a generalized high-order model called MFH is developed to capture more complex interactions between multimodal features. Compared with the existing bilinear pooling methods, our proposed MFB and MFH approaches can achieve significant improvement on the VQA performance because they can achieve more effective exploitation of the complex correlations between multimodal features. Using the KLD as the loss function, our proposed answer prediction approach can achieve faster convergence rate and obtain better performance as compared with the state-of-the-art strategies. Our experimental results have demonstrated that our approaches have achieved the state-of-the-art or comparable performance on two large-scale real-world VQA data sets.

References

[1] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[2] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. ACM SIGIR*, 2014, pp. 395–404.

[3] X. Shen, W. Liu, I. W. Tsang, Q.-S. Sun, and Y.-S. Ong, "Multilabel prediction via cross-view search," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2763967.

[4] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015, pp. 2625–2634.

[5] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, vol. 14. 2015, pp. 2048–2057.

[6] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. ICCV*, 2015, pp. 2425–2433.

[7] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. NIPS*, 2014, pp. 1682–1690.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. (2016). "Multimodal compact bilinear pooling for visual question answering and visual grounding." [Online]. Available: https://arxiv.org/abs/1606.01847

[9] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. (2015). "Simple baseline for visual question answering." [Online]. Available: https://arxiv.org/abs/1512.02167

[10] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. NIPS*, 2016, pp. 289–297.

[11] J. B. Tenenbaum and W. T. Freeman, "Separating style and content," in *Proc. NIPS*, 1997, pp. 662–668.

[12] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. ICCV*, 2015, pp. 1449–1457.

[13] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *Proc. ICLR*, 2017, pp. 1–10.

[14] J.-H. Kim *et al.*, "Multimodal residual learning for visual QA," in *Proc. NIPS*, 2016, pp. 361–369.

[15] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. (2016). "Learning to compose neural networks for question answering." [Online]. Available: https://arxiv.org/abs/1601.01705

[16] I. Ilievski, S. Yan, and J. Feng. (2016). "A focused dynamic attention model for visual question answering." [Online]. Available: https://arxiv.org/abs/1604.01485

[17] H. Nam, J.-W. Ha, and J. Kim. (2016). "Dual attention networks for multimodal reasoning and matching." [Online]. Available: https://arxiv.org/abs/1611.00471

[18] P. Wang, Q. Wu, C. Shen, A. V. D. Hengel, and A. Dick. (2015). "Explicit knowledge-based reasoning for visual question answering." [Online]. Available: https://arxiv.org/abs/1511.02570

[19] 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 *Proc. CVPR*, Jun. 2016, pp. 4622–4630.

[20] Z. Wang and S. Ji. (2017). "Learning convolutional text representations for visual question answering." [Online]. Available: https://arxiv.org/abs/1705.06824

[21] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia. (2015). "ABC-CNN: An attention based convolutional neural network for visual question answering." [Online]. Available: https://arxiv.org/abs/1511.05960

[22] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. CVPR*, 2016, pp. 21–29.

[23] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. CVPR*, 2016, pp. 4613–4621.

[24] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[25] Z. Lin *et al.*. (2017). "A structured self-attentive sentence embedding." [Online]. Available: https://arxiv.org/abs/1703.03130

[26] Y. Li, N. Wang, J. Liu, and X. Hou. (2016). "Factorized bilinear models for image recognition." [Online]. Available: https://arxiv.org/abs/1611.05709

[27] S. Rendle, "Factorization machines," in *Proc. ICDM*, Dec. 2010, pp. 995–1000.

[28] D. Tao, J. Cheng, M. Song, and X. Lin, "Manifold ranking-based matrix factorization for saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1122–1134, Jun. 2016.

[29] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang, "Person re-identification by dual-regularized kiss metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2726–2738, Jun. 2016.

[30] D. Tao, L. Jin, Y. Yuan, and Y. Xue, "Ensemble manifold rank preserving for acceleration-based human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1392–1404, Jun. 2016.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] X. Shen, X. Tian, T. Liu, F. Xu, and D. Tao, "Continuous dropout," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2750679.

[33] C. Shi, Z. Liu, X. Dong, and Y. Chen, "A novel error-compensation control for a class of high-order nonlinear systems with input delay," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2751256.

[34] X. Zhao, N. Wang, Y. Zhang, S. Du, Y. Gao, and J. Sun, "Beyond pairwise matching: Person reidentification via high-order relevance learning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2736640.

[35] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385

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

[37] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[38] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. (2016). "Making the V in VQA matter: Elevating the role of image understanding in visual question answering." [Online]. Available: https://arxiv.org/abs/1612.00837

[39] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[40] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.

[41] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. ECCV*, 2010, pp. 143–156.
[42] H. Noh, P. Hongsuck Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. CVPR*, 2016, pp. 30–38.
[43] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. ICCV*, 2015, pp. 1–9.
[44] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. ECCV*, 2016, pp. 451–466.
[45] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. CVPR*, 2016, pp. 39–48.
[46] H. Noh and B. Han. (2016). "Training recurrent answering units with joint loss minimization for VQA." [Online]. Available: https://arxiv.org/abs/1606.03647
[47] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, vol. 14. Oct. 2014, pp. 1532–1543.
[48] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. NIPS*, 2015, pp. 3294–3302.
[49] R. Krishna *et al.*. (2016). "Visual genome: Connecting language and vision using crowdsourced dense image annotations." [Online]. Available: https://arxiv.org/abs/1602.07332
[50] P. Anderson *et al.*. (2017). "Bottom-up and top-down attention for image captioning and visual question answering." [Online]. Available: https://arxiv.org/abs/1707.07998
[51] D. Teney, P. Anderson, X. He, and A. van den Hengel. (2017). "Tips and tricks for visual question answering: Learnings from the 2017 challenge." [Online]. Available: https://arxiv.org/abs/1708.02711

Zhou Yu received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China, in 2010 and 2015, respectively.

He is currently a Lecturer with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His current research interests include multimodal data analysis, computer vision, machine learning, and deep learning.

Jun Yu (M'13) received the B.Eng. and Ph.D. degrees from Zhejiang University, Zhejiang, China.

He was an Associate Professor with the School of Information Science and Technology, Xiamen University, Xiamen, China. From 2009 to 2011, he was with Nanyang Technological University, Singapore. From 2012 to 2013, he was a Visiting Researcher with Microsoft Research Asia, Beijing, China. He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. He has authored or co-authored more than 60 scientific articles. His research interests included multimedia analysis, machine learning, and image processing.

Dr. Yu is a Professional Member of the Association for Computing Machinery and the China Computer Federation. He was a recipient of the IEEE SPS Best Paper Award. He has co-chaired several special sessions, invited sessions, and workshops. He served as a Program Committee Member or Reviewer for top conferences and prestigious journals.

Chenchao Xiang received the B.Eng. degree from the School of Management, Hangzhou Dianzi University, Hangzhou, China, in 2016, where he is currently pursuing the M.Eng. degree with the School of Computer Science and Technology.

His current research interests include multimodal analysis, computer vision, and machine learning.

Jianping Fan received the M.S. degree in theory physics from Northwestern University, Xian, China, in 1994, and the Ph.D. degree in optical storage and computer science from the Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997.

From 1997 to 1998, he was a Post-Doctoral Researcher with Fudan University, Shanghai. From 1998 to 1999, he was a Researcher with the Japan Society of Promotion of Science, Department of Information System Engineering, Osaka University, Osaka, Japan. From 1999 to 2001, he was a Post-Doctoral Researcher with the Department of Computer Science, Purdue University, West Lafayette, IN, USA. He is currently a Professor with University of North Carolina at Charlotte, Charlotte, NC, USA. His current research interests include image/video privacy protection, automatic image/video understanding, and large-scale deep learning.

Dacheng Tao (F'15) is Professor of Computer Science and ARC Laureate Fellow in the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, at the University of Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 500+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, Advances in Neural Information Processing Systems, International Conference on Machine Learning, IEEE Conference on Computer Vision and Pattern Recognition, IEEE International Conference on Computer Vision, European Conference on Computer Vision, IEEE International Conference on Data Mining, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM07, the best student paper award in IEEE ICDM13, the distinguished student paper award in the 2017 IJCAI, the 2014 ICDM 10-year highest-impact paper award, and the 2017 IEEE Signal Processing Society Best Paper Award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the IEEE, AAAS, OSA, IAPR, and SPIE.