Compact Trilinear Interaction for Visual Question Answering

Tuong Do†, Thanh-Toan Do†, Huy Tran, Erman Tjiputra, Quang D. Tran

1 AIOZ Pte Ltd, Singapore 2 University of Liverpool

Abstract

In Visual Question Answering (VQA), answers have a great correlation with question meaning and visual contents. Thus, to selectively utilize image, question and answer information, we propose a novel trilinear interaction model which simultaneously learns high level associations between these three inputs. In addition, to overcome the interaction complexity, we introduce a multimodal tensor-based PARALIND decomposition which efficiently parameterizes trilinear interaction between the three inputs. Moreover, knowledge distillation is first time applied in Free-form Open-ended VQA. It is not only for reducing the computational cost and required memory but also for transferring knowledge from trilinear interaction model to bilinear interaction model. The extensive experiments on benchmarking datasets TDIUC, VQA-2.0, and Visual7W show that the proposed compact trilinear interaction model achieves state-of-the-art results when using a single model on all three datasets. The source code is available at https://github.com/aioz-ai/ICCV19_VQA-CTI.

1. Introduction

The aim of VQA is to find out a correct answer for a given question which is consistent with visual content of a given image [25, 3, 10]. There are two main variants of VQA which are Free-Form Open-ended (FFOE) VQA and Multiple Choice (MC) VQA. In FFOE VQA, an answer is a free-form response to a given image-question pair input, while in MC VQA, an answer is chosen from an answer list for a given image-question pair input.

Traditional approaches to both VQA tasks mainly aim to learn joint representations between images and questions, while the answers are treated in a “passive” form, i.e., the answers are only considered as classification targets. However, an answer is expected to have high correlation with its corresponding question-image input, hence a jointly and explicitly information extraction from these three inputs will give a highly meaningful joint representation. In this paper, we propose a novel trilinear interaction model which simultaneously learns high level associations between all three inputs, i.e., image, question, and answer.

The main difficulty in trilinear interaction is the dimensionality issue which causes expensive computational cost and huge memory requirement. To tackle this challenge, we propose to use PARALIND decomposition [6] which factorizes a large tensor into smaller tensors which reduces the computational cost and the usage memory.

The proposed trilinear interaction takes images, questions and answers as inputs. However, answer information in FFOE VQA [1, 40, 26, 59] is only available in the training phase but not in the testing phase. To apply the trilinear interaction for FFOE VQA, we propose to use knowledge distillation to transfer knowledge from trilinear model to bilinear model. The distilled bilinear model only requires pairs of image and question as inputs, hence it can be used for the testing phase. For MC VQA [47, 19, 27, 15, 30, 44], the answer information can be easily extracted, thanks to the given answer list that contains few candidate answers for each image-question pair and is available in both training and testing phases. Thus, the proposed trilinear interaction can be directly applied to MC VQA.

To evaluate the effectiveness of the proposed model, the extensive experiments are conducted on the benchmarking datasets TDIUC, VQA-2.0, and Visual7W. The results show that the proposed model achieves state-of-the-art results on all three datasets.

The main contributions of the paper are as follows. (i) We propose a novel trilinear interaction model which simultaneously learns high level joint presentation between image, question, and answer information in VQA task. (ii) We utilize PARALIND decomposition to deal with the dimensionality issue in trilinear interaction. (iii) To make the proposed trilinear interaction applicable for FFOE VQA, we propose to use knowledge distillation for transferring knowledge from trilinear interaction model to bilinear interaction model. The remaining of this paper is organized as follows. Section 2 presents the related work. Section 3 presents the proposed compact trilinear interaction (CTI). Section 4 presents the proposed models when applying CTI to FFOE VQA and MC VQA. Section 5 presents ablation studies, experimental results and analysis.

† indicates equal contribution.
2. Related Work

Joint embedding in Visual Question Answering. There are different approaches have been proposed for VQA \cite{18, 5, 8, 45, 20, 22, 24, 2, 28, 23, 38, 46, 29, 40}. Most of the successful methods focus on learning joint representation between the input question and image \cite{8, 5, 18, 45}. In the state-of-the-art VQA, the features of the input image and question are usually represented under matrix forms. E.g., each image is described by a number of interested regions, and each region is represented by a feature vector. Similar idea is applied for question, e.g., an question contains a number of words and each word is represented by a feature vector. A fully expressive interaction between an image region and a word should be the outer product between their two corresponding vectors \cite{8}. The outer product allows a multiplicative interaction between all elements of both vectors. However, a fully bilinear interaction using outer product between every possible pairs of regions and words will dramatically increase the output space. Hence instead of directly computing the fully bilinear with outer product, most of works try to compress or decompose the fully bilinear interaction.

In \cite{8}, the authors proposed the Multimodal Compact Bilinear pooling which is an efficient method to compress the bilinear interaction. The method works by projecting the visual and linguistic features to a higher dimensional space and then convolving both vectors efficiently by using element-wise product in Fast Fourier Transform space. In \cite{5}, the authors proposed Multimodal Tucker Fusion which is a tensor-based Tucker decomposition to efficiently parameterize bilinear interaction between visual and linguistic representations. In \cite{45}, the author proposed Factorized Bilinear Pooling that uses two low rank matrices to approximate the fully bilinear interaction. Recently, in \cite{18} the authors proposed Bilinear Attention Networks (BAN) that finds bilinear attention distributions to utilize given visual-linguistic information seamlessly. BAN also uses low rank approximation to approximate the bilinear interaction for each pair of vectors from image and question.

There are other works that consider answer information, besides image and question information, to improve VQA performance \cite{16, 9, 36, 14, 41, 54}. Typically, in \cite{14}, the authors learned two embedding functions to transform an image-question pair and an answer into a joint embedding space. The distance between the joint embedded image-question and the embedded answer is then measured to determine the output answer. In \cite{41}, the authors computed joint representations between image and question, and between image and answer. They then learned a joint embedding between the two computed representations.

In \cite{34}, the authors computed “ternary potentials” which capture the dependencies between three inputs, i.e., image, question, and answer. For every triplet of vectors, each from each different input, to compute the interaction between three vectors, instead of calculating the outer products, the author computed the sum of element-wise product of the three vectors. This greatly reduces the computational cost but it might not be expressive enough to fully capture the complex associations between the three vectors.

Different from previous works that mainly aim to learn the joint representations from pairs of modalities \cite{8, 5, 18, 45, 14, 41} or greatly simplify the interaction between the three modalities by using the element-wise operator \cite{34}, in this paper, we propose a principle and direct approach – a trilinear interaction model, which simultaneously learns a joint representation between three modalities. In particular, we firstly derive a fully trilinear interaction between three modalities. We then rely on a decomposition approach to develop a compact model for the interaction.

Knowledge Distillation. Knowledge Distillation is a general approach for transferring knowledge from a cumbersome model (teacher model) to a lighter model (student model) \cite{13, 11, 33, 7, 4}. In FFOE VQA, the trilinear interaction model, which takes image, question, and answer as inputs, can only be applied for training phase but not for testing phase due to the omission of answer in testing. To overcome this challenge and also to reduce computational cost, inspired from the Hinton’s seminar work \cite{13}, we propose to use knowledge distillation to transfer knowledge from trilinear model to bilinear model.

3. Compact Trilinear Interaction (CTI)

3.1. Fully parameterized trilinear interaction

Let $M = \{M_1, M_2, M_3\}$ be the representations of three inputs. $M_t \in \mathbb{R}^{n_t \times d_t}$, where $n_t$ is the number of channels of the input $M_t$ and $d_t$ is the dimension of each channel. For example, if $M_1$ is the region-based representation for an image, then $n_1$ is the number of regions and $d_1$ is the dimension of the feature representation for each region. Let $m_{t_{th}} \in \mathbb{R}^{1 \times d_t}$ be the $e^{th}$ row of $M_t$, i.e., the feature representation of $e^{th}$ channel in $M_t$, where $t \in \{1, 2, 3\}$.

The joint representation resulted from a fully parameterized trilinear interaction over the three inputs is presented by $z \in \mathbb{R}^{d_z}$ which is computed as follows

$$z^T = \left(\langle T \times_1 \text{vec}(M_1) \rangle \times_2 \text{vec}(M_2) \rangle \times_3 \text{vec}(M_3)\right)$$

(1)

where $T \in \mathbb{R}^{d_{M_1} \times d_{M_2} \times d_{M_3}}$ is a learning tensor; $d_{M_t} = n_t \times d_t$; vec($M_t$) is a vectorization of $M_t$ which outputs a row vector; operator $\times_i$ denotes the $i$-mode tensor product.

The tensor $T$ helps to learn the interaction between the three input through $i$-mode product. However, learning such a large tensor $T$ is infeasible when the dimension $d_{M_t}$ of each input modality is high, which is the usual case in VQA. Thus, it is necessary to reduce the size $T$ to make the learning feasible.

Inspired by \cite{43}, we rely on the idea of unitary attention mechanism. Specifically, let $z_p \in \mathbb{R}^{d_z}$ be the joint
where $M_T^z$ can be rewritten as $M_T$ in (1), we now only need to learn two smaller tensors $M_{1r}, M_{2r}, M_{3r}$ of $p^{th}$ triplet is computed as

$$z_p^T = \left( (T_{sc} \times 1 m_{1r}) \times 2 m_{2r} \right) \times 3 m_{3r}$$

(2)

where $T_{sc} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the learning tensor between channels in the triplet.

Follow the idea of unitary attention [4], the joint representation $z$ is approximated by using joint representations of all triplets described in (2) instead of using fully parameterized interaction over three inputs as in (1). Hence, we compute

$$z = \sum_p M_p z_p$$

(3)

Note that in (3), we compute a weighted sum over all possible triplets. The $p^{th}$ triplet is associated with a scalar weight $M_p$. The set of $M_p$ is called as the attention map $M$, where $M \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

The attention map $M$ resulted from a reduced parameterized trilinear interaction over three inputs $M_1, M_2$ and $M_3$ is computed as follows

$$M = \left( (T_M \times 1 M_1) \times 2 M_2 \right) \times 3 M_3$$

(4)

where $T_M \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the learning tensor of attention map $M$. Note that the learning tensor $T_M$ in (4) has a reduced size compared to the learning tensor $T$ in (1).

By integrating (2) into (3), the joint representation $z$ in (3) can be rewritten as

$$z^T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} M_{ijk} \left( (T_{sc} \times 1 m_{1r}) \times 2 m_{2r} \right) \times 3 m_{3r}$$

(5)

where $M_{ijk}$ in (5) is actually a scalar attention weight $M_p$ of the attention map $M$ in (3).

It is also worth noting from (5) that to compute $z$, instead of learning the large tensor $T \in \mathbb{R}^{d_{1r} \times d_{2r} \times d_{3r}}$ in (1), we now only need to learn two smaller tensors $T_{sc} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ in (2) and $T_M \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ in (4).

3.2. Parameter factorization

Although the large tensor $T$ of trilinear interaction model is replaced by two smaller tensors $T_M$ and $T_{sc}$, the dimension of these two tensors still large which makes the learning difficult. To further reduce the computational complexity, the PARALIND decomposition [6] is applied for $T_M$ and $T_{sc}$. The PARALIND decomposition for the learning tensor $T_M \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ can be calculated as

$$T_M \approx \sum_{r=1}^{R} \left[ (G_r \times 1 W_{1r}) \times 2 W_{2r} \right] \times 3 W_{3r}$$

(6)

where $R$ is a slicing parameter, establishing a trade-off between the decomposition rate (which is directly related to the usage memory and the computational cost) and the performance. Each $G_r \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a smaller learnable tensor called Tucker tensor. The number of these Tucker tensors equals to $R$. The maximum value for $R$ is usually set to the greatest common divisor of $d_1, d_2$ and $d_3$. In our experiments, we found that $R = 32$ gives a good trade-off between the decomposition rate and the performance.

Here, we have dimension $d_1 = d_1/R, d_2 = d_2/R$ and $d_3 = d_3/R$. $W_{1r} \in \mathbb{R}^{d_1 \times d_1}, W_{2r} \in \mathbb{R}^{d_2 \times d_2}$ and $W_{3r} \in \mathbb{R}^{d_3 \times d_3}$ are learnable factor matrices. Figure 1 shows the illustration of PARALIND decomposition for a tensor $T_M$.

The shorten form of $T_M$ in (6) can be rewritten as

$$T_M \approx \sum_{r=1}^{R} \left[ (G_r; W_{1r}, W_{2r}, W_{3r}) \right]$$

(7)

Integrating the learning tensor $T_M$ from (7) into (4), the attention map $M$ can be rewritten as

$$M = \sum_{r=1}^{R} \left[ (G_r; M_1 W_{1r}, M_2 W_{2r}, M_3 W_{3r}) \right]$$

(8)

Similar to $T_M$, PARALIND decomposition is also applied to the tensor $T_{sc}$ in (5) to reduce the complexity. It is worth noting that the size of $T_{sc}$ directly affects to the dimension of the joint representation $z \in \mathbb{R}^{d_z}$. Hence, to minimize the loss of information, we set the slicing parameter $R = 1$ and the projection dimension of factor matrices at $d_z$, i.e., the same dimension of the joint representation $z$.

Therefore, $T_{sc} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ in (5) can be calculated as

$$T_{sc} \approx \left( (G_{sc} \times 1 W_{z1}) \times 2 W_{z2} \right) \times 3 W_{z3}$$

(9)

where $W_{z1} \in \mathbb{R}^{d_1 \times d_z}, W_{z2} \in \mathbb{R}^{d_2 \times d_z}$, $W_{z3} \in \mathbb{R}^{d_3 \times d_z}$ are learnable factor matrices and $G_{sc} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a smaller tensor (compared to $T_{sc}$).

Up to now, we already have $M$ by (8) and $T_{sc}$ by (9), hence, we can compute $z$ using (5). $z$ from (5) can be rewrit-
ten as

\[ z^T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} M_{ijk} \left( (G_{sc} \times m_1, W_{z_1}) \times m_2, W_{z_2} \right) \times m_3, W_{z_3} \]  \tag{10}

Here, it is interesting to note that \( G_{sc} \in \mathbb{R}^{d_s \times d_s \times d_s} \) in (10) has rank 1. Thus, the result got from \( i \)-mode tensor products in (10) can be approximated by the Hadamard products without the presence of rank-1 tensor \( G_{sc} \). In particular, \( z \) in (10) can be computed without using \( G_{sc} \) as

\[ z^T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} M_{ijk} \left( m_1, W_{z_1} \circ m_2, W_{z_2} \circ m_3, W_{z_3} \right) \]  \tag{11}

Note that \( d_z \), which is the joint embedding dimension, is a user-defined parameter which makes a trade-off between the capability of the representation and the computational cost. In our experiments, we found that \( d_z = 1,024 \) gives a good trade-off.

4. Compact Trilinear Interaction for VQA

The input for training VQA is set of \( (V, Q, A) \) in which \( V \) is an image representation; \( V \in \mathbb{R}^{v \times d_v} \) where \( v \) is the number of interested regions (or bounding boxes) in the image and \( d_v \) is the dimension of the representation for a region; \( Q \) is a question representation; \( Q \in \mathbb{R}^{q \times d_q} \) where \( q \) is the number of hidden states and \( d_q \) is the dimension for each hidden state. \( A \) is an answer representation; \( A \in \mathbb{R}^{a \times d_a} \) where \( a \) is the number of hidden states and \( d_a \) is the dimension for each hidden state.

By applying the Compact Trilinear Interaction (CTI) to each \( (V, Q, A) \), we achieve the joint representation \( z \in \mathbb{R}^{d_z} \). Specifically, we firstly compute the attention map \( M \) by (8) as follows

\[ M = \sum_{r=1}^{R} \left[ G_r; VW_{v_r}, QW_{q_r}, AW_{a_r} \right] \]  \tag{12}

Then the joint representation \( z \) is computed by (11) as follows

\[ z^T = \sum_{i=1}^{v} \sum_{j=1}^{q} \sum_{k=1}^{a} M_{ijk} \left( V_i W_{v_i} \circ Q_j W_{q_j} \circ A_k W_{a_k} \right) \]  \tag{13}

where \( W_{v_1}, W_{q_1}, W_{a_1} \) in (12) and \( W_{v_2}, W_{q_2}, W_{a_2} \) in (13) are learnable factor matrices; each \( G_r \) in (12) is a learnable Tucker tensor.

4.1. Multiple Choice Visual Question Answering

To make a fair comparison to the state of the art in MC VQA \([14, 41]\), we follow the representations used in those works. Specifically, each input question and each answer are trimmed to a maximum of 12 words which will then be zero-padded if shorter than 12 words. Each word is then represented by a 300-D GloVe word embedding \([32]\). Each image is represented by a \( 14 \times 14 \times 2048 \) grid feature (i.e., \( 196 \) cells; each cell is with a \( 2,048 \)-D feature), extracted from the second last layer of ResNet-152 which is pre-trained on ImageNet \([12]\).

Follow \([41]\), input samples are divided into positive samples and negative samples. A positive sample, which is labelled as 1 in binary classification, contains image, question and the right answer. A negative sample, which is labelled as 0 in binary classification, contains image, question, and the wrong answer. These samples are then passed through our proposed CTI to get the joint representation \( z \). The joint representation is passed through a binary classifier to get the prediction. The Binary Cross Entropy loss is used for training the proposed model. Figure 2 visualizes the proposed model when applying CTI to MC VQA.

4.2. Free-Form Opened-Ended Visual Question Answering

Unlike MC VQA, FFOE VQA treats the answering as a classification problem over the set of predefined answers. Hence the set possible answers for each question-image pair is much more than the case of MC VQA. Therefore the model design proposed in Section 4.1 i.e. for each question-image input, the model takes every possible answers from its answer list to computed the joint representation, causes high computational cost. In addition, the proposed CTI requires all three \( V, Q, A \) inputs to compute the joint representation. However, during the testing, there are no available answer information in FFOE VQA. To overcome these challenges, we propose to use Knowledge Distillation \([13]\) to transfer the learned knowledge from a teacher model to a student model. Figure 3 visualizes the proposed design for FFOE VQA.

Our teacher model takes triplets of \( \text{image-question-right answer} \) as inputs. Each triplet is passed through the proposed CTI to get the joint representation \( z \). The joint representation \( z \) is then passed through a multiclass classifier (over the set of predefined answers) to get the prediction which is similar to \([37]\). The Cross Entropy loss is used for training the teacher model. Regarding the student models, any state-of-the-art VQA can be used. In our experiments, we use BAN2 \([18]\) or SAN \([43]\) as student models. The student models take pairs of image-question as inputs and treat the prediction as a multiclass classification problem. The loss function for the student model is defined as

\[ L_{KD} = \alpha \gamma^2 L_{CE}(Q_S, Q_T^\gamma) + (1 - \alpha)L_{CE}(Q_S, y_{true}) \]  \tag{14}

where \( L_{CE} \) stands for Cross Entropy loss; \( Q_S \) is the standard softmax output of the student; \( y_{true} \) is the ground-truth answer labels; \( \alpha \) is a hyper-parameter for controlling the im-
important of each loss component; \( Q_S^T, Q_T^T \) are the softened outputs of the student and the teacher using the same temperature parameter \( T \) [13], which are computed as follows

\[
Q_i^T = \frac{\exp(l_i/T)}{\sum_i \exp(l_i/T)}
\]

where for both teacher and the student models, the logit \( l \) is the predictions outputted by the corresponding classifiers.

Following by the current state of the art in FFOE VQA [18], for image representation, we use object detection-based features with FPN detector (ResNet152 backbone) [22], in which the number of maximum detected bounding boxes is set to 50. For question and answer representations, we trim question and answer to a maximum of 12 words which will then be zero-padded if shorter than 12 words. Each word is then represented by a 600-D vector that is a concatenation of the 300-D GloVe word embedding [32] and the augmenting embedding from training data as [18]. In the other words, a question is with a representation with size \( 12 \times 600 \). It is similar for answer.

5. Experiments

5.1. Dataset and evaluation protocol

**Dataset.** We conduct the experiments on three benchmarking VQA datasets that are Visual7W [47] for the MC VQA, VQA-2.0 [10] and TDIUC [17] for the FFOE VQA. We use training set to train and validation set to evaluate in all mentioned datasets when conducting ablation study.

**Implementation details.** Our CTI is implemented using PyTorch [31]. The experiments are conducted on a NVIDIA Titan V GPUs with 12GB RAM. In all experiments, the learning rate is set to \( 10^{-3} \). Batch size is set to 128 for training MC VQA and 256 for training FFOE VQA. When training both MC VQA model (Section 4.1) and FFOE VQA model (Section 4.2), except the image representation extraction, other components are trained end-to-end. The temperature parameter \( T \) in (15) is set to 3. The dimension of the joint representation \( z \) is set at 1,024 for both MC VQA and FFOE VQA.

**Evaluation Metrics.** We follow the literature [3, 17, 47] in which the evaluation metrics for each VQA task are different. For FFOE VQA, the single accuracy, which is a standard VQA accuracy \( \text{Acc} \) [3], is applied for both TDIUC and VQA-2.0 datasets. In addition, due to the imbalance in the question types of TDIUC dataset, follow [17], we also report four other metrics that compensate for the skewed question-type distribution. They are Arithmetic MPT \( \text{Ari} \), Arithmetic Norm-MPT \( \text{Ari-N} \), Harmonic MPT \( \text{Har} \), and Harmonic Norm-MPT \( \text{Har-N} \). For MC VQA, we follow the evaluation metric \( \text{Acc-MC} \) proposed by [47] in which the performance is measured by the portion of correct answers selected by the VQA model from the candidate answer set.

5.2. Ablation study

The effectiveness of CTI on FFOE VQA. We compare our distilled BAN2 (BAN2-CTI) and distilled SAN (SAN-CTI) student models to the state-of-the-art baselines BAN2 [18] and SAN [43]. Table 1 presents a comprehensive evaluation on five different metrics on TDIUC. Among all metrics, on overall, our BAN2-CTI and SAN-CTI outperform corresponding baselines by a noticeable margin.
Figure 3. The proposed model when CTI is applied on FFOE VQA. The details are described in Section 4.2. Best view in color.

Table 1. Overall performance of the proposal and the baselines BAN2, SAN in different evaluation metrics on TDIUC validation set. The performance is shown with and without considering Absurd question category. BAN2-CTI and SAN-CTI are student models trained under our proposed CTI teacher model.

These results confirm the effectiveness of our proposed CTI for learning the joint representation. In addition, the proposed teacher model (Figure 3) is also effective. It successfully transfers useful learned knowledge to the student models. Note that in Table 1, the “Absurd” question category indicates the cases in which input questions are irrelevant to the image contents. Thus, the answers are always “does not apply”, i.e., “no answer”. Using these meaningless answers when training the teacher causes negative effect when learning the joint representation, hence, reducing the model capacity. If the “Absurd” category is not taken into account, the proposed model achieves more improvements over baselines.

Table 2. Performance (Acc) of the proposal and the baselines BAN2, SAN for each question-type on TDIUC validation set. BAN2-CTI and SAN-CTI are student models trained under our compact trilinear interaction teacher model.

Table 2 presents detail performances with Acc metric over each question category of TDIUC when all categories, including “Absurd”, are used for training. The results show that we achieve the best results on all question categories but “Absurd”. We note that in the real applications, the “Absurd” question problem may be mitigated in some cases by using a simple trick, i.e., asking a “presence question” before asking the main question, e.g., we have an image with no human but the main question is “Is the people wearing hat?”, i.e., a “Absurd” question. By asking a “presence
Table 3. Performance of the proposal and baselines BAN2, SAN in VQA-2.0 validation set and test-dev set. BAN2-CTI and SAN-CTI are student models trained under proposed teacher model.

| Ref models | Validation Accuracy | Test-dev Accuracy |
|------------|--------------------|------------------|
| Bottom-up [37] | 63.2              | 65.4             |
| SAN [43]   | 61.7              | 63.0             |
| SAN-CTI    | 62.1              | 63.4             |
| BAN2 [18]  | 65.6              | 66.5             |
| BAN2-CTI   | 66.0              | 67.4             |

Table 4. The performance (Acc-MC) and the number of parameters of the proposed MC VQA model and the baselines BAN2, SAN on Visual7W validation set.

| Ref models | Visual7W validation set | Number of parameters |
|------------|-------------------------|----------------------|
| BAN2 [18]  | 65.7                    | ~86.5M               |
| SAN [43]   | 59.3                    | ~69.7M               |
| CTI        | 67.0                    | ~66.5M               |

Figure 4. The visualization of an attention map (b) computed from Eq. (12) for an image-question-answer input (a). The attention map indicates attention weights over triplets of "detected bounding box - word in question - word in answer". The higher weight of a triplet is, the more contribution it makes to the joint representation. We can see that there are triplets (V=4, Q=“tube”, A=“white”), (V=4, Q=“tube”, A=“red”), (V=4, Q=“tube”, A=“blue”) have high weight values. That means that these triplets give high contribution to the joint representation. The input sample (a) is from Visual7W validation set. Best view in color.

5.3. Comparison with the state of the art

To further evaluate the effectiveness of CTI, we conduct a detailed comparison with the current state of the art. For FFOE VQA, we compare our proposal with the recent state-of-the-art methods on TDIUC and VQA-2.0 datasets, including SAN [43], QTA [35], BAN2 [18], Bottom-up [37], MCB [8], and RAU [29]. For MC VQA, we compare with the state-of-the-art methods on Visual7W dataset, including BAN2 [18], SAN [43], MLP [16], MCB [8], STL [41], and FPMC [14]. It is worth noting that depending on tasks FFOE VQA or MC VQA, we use different representations for images and questions as clearly mentioned in Section 4. This ensures a fair comparison with compared methods.

Regarding FFOE VQA, Tables 3 and 5 show comparative results on VQA-2.0 and TDIUC respectively. Specifically, Table 5 shows that our distilled student BAN2-CTI outperforms all compared methods over all metrics by a large margin, i.e., our model outperforms the current state-of-the-art QTA [35] on TDIUC by 3.4% and 5.4% on Ari and Har metrics, respectively. The results confirm that the proposed trilinear interaction has learned informative repre-
### Evaluation metrics

| Models     | Acc | Art | Har | Art-N | Har-N |
|------------|-----|-----|-----|-------|-------|
| BAN2 [18]  | 85.5| 67.4| 54.9| 37.4  | 15.7  |
| SAN [43]   | 82.3| 65.0| 53.7| 35.4  | 14.7  |
| QTA [35]   | 85.0| 69.1| 60.1| -     | -     |
| MCB [18]   | 79.2| 65.8| 58.0| 39.8  | 24.8  |
| RAU [29]   | 84.3| 67.8| 59.0| 41.0  | 24.0  |
| SAN-CTI    | 84.5| 68.7| 59.9| 41.3  | 23.3  |
| BAN2-CTI   | 87.0| 72.5| 65.5| 45.8  | 28.6  |

Table 5. Performance comparison between different approaches with different evaluation metrics on TDIUC validation set. BAN2-CTI and SAN-CTI are the student models trained under our compact trilinear interaction teacher model.

### 5.4. Further analysis

#### Compact Trilinear Interaction as the generalization of BAN [18]

The proposed compact trilinear interaction model can be seen as a generalization of the state-of-the-art joint embedding BAN [18].

In BAN, each input contains an image representation $V \in \mathbb{R}^{d_v \times v}$ and a question representation $Q \in \mathbb{R}^{d_q \times q}$. The trilinear interaction model can be modified to adapt to these two inputs. The joint representation $z \in \mathbb{R}^{d_z}$ in (1) can be adapted for two input as

$$z^T = (T_{vq} \times_1 vec(V)) \times_2 vec(Q)$$

where $T_{vq} \in \mathbb{R}^{d_V \times d_Q \times d_z}$ is a learnable tensor; $vec(V)$ is the vectorization of $V$ and $vec(Q)$ is the vectorization of $Q$ which output row vectors; $d_V = d_v \times v$; $d_Q = d_q \times q$.

By applying “Parameter factorization” described in Section 3.2, $z$ in (16) can be approximated based on (13) as

$$z^T = \sum_{i=1}^{v} \sum_{j=1}^{q} \mathcal{M}_{ij} (V^T W_{zv} \circ Q_j^T W_{zq})$$

where $W_{zv} \in \mathbb{R}^{d_z \times d_v}$ and $W_{zq} \in \mathbb{R}^{d_z \times d_q}$ are learnable factor matrices; $\mathcal{M}_{ij}$ is an attention weight of attention map $\mathcal{M} \in \mathbb{R}^{v \times q}$ which can be computed from (12) as

$$\mathcal{M} = \sum_{r=1}^{R} \left[ G_r; V^T W_{v v^r}, Q^T W_{q q^r} \right]$$

where $W_{v v^r} \in \mathbb{R}^{d_v \times d_{v^r}}$ and $W_{q q^r} \in \mathbb{R}^{d_q \times d_{q^r}}$ are learnable factor matrices; $d_{v^r} = d_v / R$; $d_{q^r} = d_q / R$; each $G_r \in \mathbb{R}^{d_{v^r} \times d_{q^r}}$ is a learnable Tucker tensor.

Interestingly, (17) can be reorganized to have a form of BAN [18] as

$$z_k = \sum_{i=1}^{v} \sum_{j=1}^{q} \mathcal{M}_{ij} \left( V_i^T W_{zv} W_{zv k}^T Q_j \right)$$

where $z_k$ is the $k$th element of the joint representation $z$; $W_{zv k}$ and $W_{zv k}$ are $k$th column in factor matrices $W_{zv}$ and $W_{zq}$. Note that in (19), our attention map $\mathcal{M}$ is resulted from the PARALIND decomposition, while in BAN [18], their attention map is computed by bilinear pooling.

### 6. Conclusion

We propose a novel compact trilinear interaction which simultaneously learns high level associations between image, question, and answer in both MC VQA and FFoE VQA. In addition, knowledge distillation is the first time applied to FFoE VQA to overcome the computational complexity and memory issue of the interaction. The extensive experimental results show that the proposed models achieve the state-of-the-art results on three benchmarking datasets.
[40] Damien Teney and Anton van den Hengel. Visual question answering as a meta learning task. In ECCV, 2018.
[41] Zhe Wang, Xiaoyi Liu, Limin Wang, Yu Qiao, Xiaohui Xie, and Charless Fowlkes. Structured triplet learning with pos-
tag guided attention for visual question answering. In WACV, 2018.
[42] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In ECCV, 2016.
[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In CVPR, 2016.
[44] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-
level attention networks for visual question answering. In CVPR, 2017.
[45] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-
modal factorized bilinear pooling with co-attention learning for visual question answering. In ICCV, 2017.
[46] Bolei Zhou, Yuanlong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167, 2015.
[47] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-
Fei. Visual7W: Grounded question answering in images. In CVPR, 2016.