Bilinear Semi-Tensor Product Attention (BSTPA) model for visual question answering

Zongwen Bai [1,2,3], Ying Li [1,2], Meili Zhou [2,3], Di Li[1], Dong Wang[1], Dawid Połap[4], Marcin Woźniak[4]
[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, CHINA
[2] Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, Yan'an 716000, CHINA
[3] School of Physics and Electronic Information , Yan'an University, Yan'an 716000, CHINA
[4] Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44-100 Gliwice, POLAND
E-mails: ydbzw@yau.edu.cn, lybyp@nwpu.edu.cn, zml@yau.edu.cn,
lidi1101@163.com, dongwang@mail.nwpu.edu.cn, dawid.polap@polsl.pl, marcin.wozniak@polsl.pl

***Abstract*—We propose a semi-tensor product attention network model as a visual question answering tool for complex interaction over image features. Proposed model performs matrix multiplication of two arbitrary dimensions, which is used to overcome possible dimensional limitations and improve recognition flexibility. In used block-wise operation we preserve spatial and temporal information but reduce the number of parameters by using low-rank pooling scheme. Applied BERT pre-train model is tuned to recognize question features. The proposed model is evaluated on the VQA2.0 dataset. Research results show that our model has good accuracy and easy reconfiguration for future research.**

***Keywords*—visual question answer, bidirectional encoder representation from transformers, semi-tensor product attention, multimodal feature fusion**

I. Introduction

Visual Question Answering (VQA) becomes one of emerging topics in the field of computational intelligence. It combines methods from computer vision and natural language processing. Complete VQA system not only requires sophisticated understanding of both image and natural language, but it also depends on the remarkable diversity of knowledge and experience. In general information processing in VQA system starts with extracting discriminative features from image and formulation of questions, than we have attention mechanism to catch the content for final stage of intelligent reasoning.

Unfortunately not all tasks can be easily solved by standard VQA models. With the development of deep learning the new chances for computer vision opened, pre-trained models like Visual Geometry Group (VGG) [1] or Resnet [2] based on ImageNet with more than ten million images are reported to gain more and more capabilities. Classic VQA relies on extracting information both from images and questions. In joint embedding method, image representation is obtained by pre-trained Convolutional Neural Networks (CNNs), which usually adopt pre-trained VGG model. Question representations are obtained by Recurrent Neural Network which is pre-trained on large text corpora. Then image and question features are obtained. Information extraction from the question is still limited to utilize the Long Short-Term Memory (LSTM) or Recurrent Neural Network (RNN) if the data set is not large enough. Therefore pre-trained models on large-scale datasets are highly demanded to improve reasoning abilities of VQA solutions. Additionally operation fusion for image and question features can boost the reasoning stage. The key is to make all the matrix operations interpretable in lowest possible dimension space. Therefore, finding a flexible matrix operation is a key solution to real-time interaction between image and question features. One of pre-trained deep bidirectional transformers for language understanding is Bidirectional Encoder Representation from Transformers (BERT) proposed by Google [3] which was trained on 330 million Wikipedia records. Semi-Tensor Product (STP) was proposed in [4], which is able to generalize conventional matrix multiplication what leads to dimension match restriction of conventional matrix multiplication.

A. Related works

Attention mechanisms consider local or global correlation between images, questions and answers to simulate computation models of the human vision. In practice, some questions are closely related to local images, other are related to global images while some other questions even go beyond what the image contains. Attention mechanisms search for the most relevant region to the question and assign different weight coefficients to features from different regions. Zhu et al. [5] proposed spatial attention to the standard LSTM model. Chen et al. [6] introduced the Question-Guided Attention Map (QAM). Yang et al. [7] proposed stacked attention networks. Lu et al. [8] and Wang et al. [10] presented hierarchical co-attention model.

Due to the limitations of VQA, not only the content understanding of images is involved, but also prior non-visual information can be required. Therefore, it is imperative to introduce knowledge-based visual question answering approaches. Wu et al. [9] proposed Ask Me Anything (AMA). It combines image features with the external knowledge, constructs a textual representation of an image and then merge this representation with the textual knowledge. Finally, they fed merged information to LSTM to produce an answer. However, the AMA only extracts discrete pieces of the text and ignores the structured representation. Furthermore, it cannot provide explanation for how it reaches to the answer. Wang et al. [6] proposed model which performs reasoning about the content of images and provides explanations of the reasoning behind the answers. Then it processes NLP question query and runs the query over the combined image and information. Andeas et al. [11] employed compositional models and proposed Neural Module

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Networks (NMN). Kumar et al. [12] designed Dynamic Memory Networks (DMN) devoted to logical reasoning. However, studies showed that the DMN scheme had inherent limitations such as the bottleneck during parsing of the question.

Pre-training is to design a network structure to do the recognition task by using a large number of endless natural language texts to train the network. The starting point of the training is to apply the neural network to build a language model and realize the word prediction task, where the by-product of the model after the optimization process is the word vector. Pre-training tasks extract a large amount of linguistic knowledge and encode them into the network structure. When the data of the task with annotated information is limited, these initial linguistic features provide extensive common sense. Bengio et al. [13] proposed the first Neural Networks Language Model (NNLM), which employs a neural network to calculate probability and optimize the model according to the objective function from the language model. Mikolov et al. [14] presented Continuous Bag-of-Words (CBOW) model and Skip-Gram Model, which are typical word-embedding schemes. However, they cannot distinguish different semantics of polysemy. It means that embedding matrix does not change with the variation of the context scene. In order to overcome the issues Embedding from Language Models (ELMO) were proposed. Initially they learn an embedding of the word with a language model but the polysemy is not distinguished at this stage. Then it utilizes a fine-tuning in accordance with the context to adapt semantic alteration so that the polysemy can be distinguished. However,there is still a pronounced defect, which takes RNN to extract the feature. Recently, the transformer has been widely used as powerful feature extractor. Generative Pre-Training (GPT) was proposed [15]. Similar to ELMO, the GPT takes the pre-training and fine-tuning policy, while their differences originate from replacing RNN with a transformer. The transformer is a superimposed "self-attention" deep network, which is the most reliable feature extractor in the NLP field.

Computer Vision (CV) research has successfully resulted in powerful pre-trained models, including VGG16, VGG19, RESNET101 and other. Although the obtained answers are impressive, the capability of these systems is still far from satisfactory. Most of CNN have given the same attention to each pixel of entire input image. However in human brain vision system, the visual attention mechanism gives different attention to different regions of the input image. The brain is modular and different functions correspond to different brain regions. In other words, when a specific task is carried out, only the corresponding part of brain is activated. Inspired by the environmental perception system of the biological vision, researchers have explored multiple ways of information fusion. Jin-Hwa et al. [16] proposed a bilinear attention to utilize given vision-language information seamlessly. Lu et al. [8] introduced a hierarchical question-image co-attention which jointly reason about the visual and question attentions. Peng Wang et al. [10] extended this pattern and improved the performance of the VQA. Caglar Gulcehre et al. [17] provided a hyperbolic attention network with capacity to match various data structure.

In recent time some similar systems have been presented. In [30] was discussed model of cross-modal version. While some other solutions are sourced in BERT derivatives, like: [31] as some baseline, [33] as task-agnostic model and [32] as generic version.

In this work we propose a novel semi-tensor product, which combines the aforementioned tools for faster solving existing VQA problems. Novel use of combined LSTM and RNN to extract features from text sequences. We use the BERT pre-trained model on the Wikipedia corpus which acquires question features. Proposed novel combination has capability of better reasoning and multilayer feature output. Proposed semi-tensor attention conducts product between two features with any dimension. Our method, in contrast to previous VQA feature fusion methods, is more flexible and powerful fusion. BERT generates different scale question features at each level, which are utilized with different levels of image features as semi-tensor product, regardless of the dimensional differences. Semi-Tensor Product Attention (STPA) network adopts the low-rank bilinear pooling and semi-tensor product. Our low-rank bilinear pooling remarkably reduce the number of parameters. Moreover, the multi-layer and multi-scale attention model makes information fusion possible at any level. This promotes improved performance of VQA system.

II. Bilinear Semi-Tensor Product Attention model architecture

Most visual question answering models are constructed in image feature extractor and question feature extractor. The image feature extractor is CNN which maps an image to a multi-layer feature vector. The question feature extractor is LSTM or GRU, which converts text into vector. Fusion of these vectors is done at the end due to constraint of matrix operations.

In this work, a more generalized model is proposed. Fig.1 shows the architecture of the proposed model. It contains three components: image feature extractor, question feature extractor and bilinear semi-tensor product attention mechanism. The image feature extractor is CNN model, which is shown in the blue dotted box. The question feature extractor is the transformer model, which is shown in the yellow dotted box. Proposed bilinear semi-tensor product attention module is located between both of them (presented in green square brackets). Detailed schematic block diagram of the multi-head attention module and bilinear attention network module is given in the last row of Fig. 1. The model gets features of image and question at the same time, then it conducts bilinear semi-tensor product to fuse the information in. In this figure, only one transformer decoder layer is presented however in practice we use N identical decoder layers.

Proposed bilinear semi-tensor product attention overcomes the dimension constraint. Furthermore, in comparison to conventional VQA methods, the proposed method can conduct fusion operation at any layer. Considering the different layer features of image and question, the semi-tensor product can be performed at a homologous layer concurrently. Then, the combination is obtained from image and question and the bilinear semi-tensor product attention is utilized to achieve that.

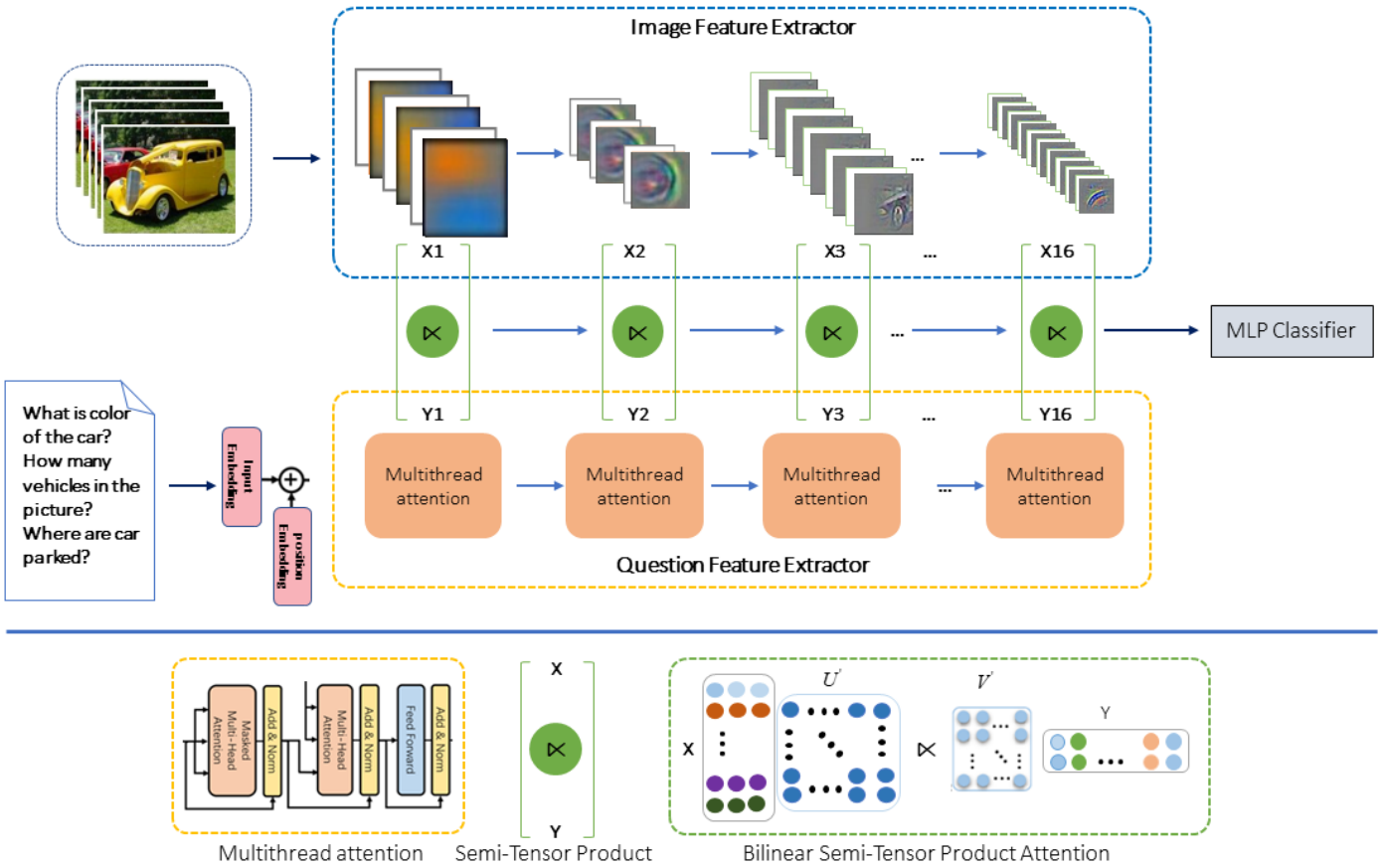

Fig. 1: Architecture of the proposed question answering model, in which image and question features extraction results are combined in one to improve bilinear semi-tensor product (presented in last row) for faster data processing.

A. Semi-Tensor Product

Consider two matrices $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{p\times q}$. For $X \cdot Y$ basic condition is that n = p what is an insurmountable obstacle for many applications. Chen et al. [4] proposed a semi-tensor product, which can conduct matrix multiply in any dimension. Let $\mathcal{A} \in \mathbb{R}^{m\times n}$ and $\mathcal{B} \in \mathbb{R}^{p\times q}$ then semi-tensor product can be defined as:

$$\mathcal{A} \ltimes \mathcal{B} := \left(A \otimes I_{t/n}\right)\left(B \otimes I_{t/p}\right) \tag{1}$$

where $\otimes$, I and t denote the Kronecker product, identity matrix and the least common multiple of n and p, respectively. For these operations we can assume

1) when $n = p$, $\mathcal{A} \ltimes \mathcal{B}$ is equivalent to $\mathcal{A} \cdot \mathcal{B}$
2) when $n = tp$ or $p = tn$ we have multiple relations $\mathcal{A} \succ_t \mathcal{B}$ and $\mathcal{A} \prec_t \mathcal{B}$ respectively
3) other cases are arbitrary dimension semi-tensor product.

In our model all the operations are case 2). Let $A \in \mathbb{R}^{1\times np}$ be a row vector, and $A \in \mathbb{R}^{p\times 1}$ be a column vector. Then A can be split into p equal-size blocks A^1, A^2 .... A^p which are $1 * n$ rows. Left Semi-Tensor Product (LSTP) denoted by $\ltimes$ is

$$\begin{cases} \mathcal{A} \ltimes \mathcal{B} = \sum_{i=1}^{p} A^i B_i & \in \mathbb{R}^n \\ \mathcal{A}^T \ltimes \mathcal{B}^T = \sum_{i=1}^{p} B_i \left(A^i\right)^T & \in \mathbb{R}^n \end{cases} \tag{2}$$

Let us consider more general case $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, then STP can be rewritten as

$$\begin{pmatrix} Row_1(A)Col_1(B) & \dots & Row_1(A)Col_q(B) \\ \dots & \dots & \dots \\ Row_m(A)Col_1(B) & \dots & Row_m(A)Col_q(B) \end{pmatrix} \tag{3}$$

Suppose $A \succ_t B$ or $A \prec_t B$, when $r \le m, s \le n$ and $r \le p, t \le q$ we can divide $\mathcal{A}$ and $\mathcal{B}$ into blocks as

$$A = \begin{bmatrix} A^{11} & \dots & A^{1s} \\ & \dots & \\ A^{r1} & \dots & A^{rs} \end{bmatrix}; B = \begin{bmatrix} B^{11} & \dots & B^{1t} \\ & \dots & \\ B^{r1} & \dots & B^{rt} \end{bmatrix} \tag{4}$$

If $A^{ik} \succ_t B^{kj}, \forall i, j, k$, then

$$\mathcal{A} \ltimes \mathcal{B} = \begin{pmatrix} C^{11} & \dots & C^{1t} \\ \dots & \dots & \dots \\ C^{r1} & \dots & C^{rt} \end{pmatrix} \tag{5}$$

where

$$C^{ij} = \sum_{k=1}^{s} A^{kj} \ltimes B^{kj} \tag{6}$$

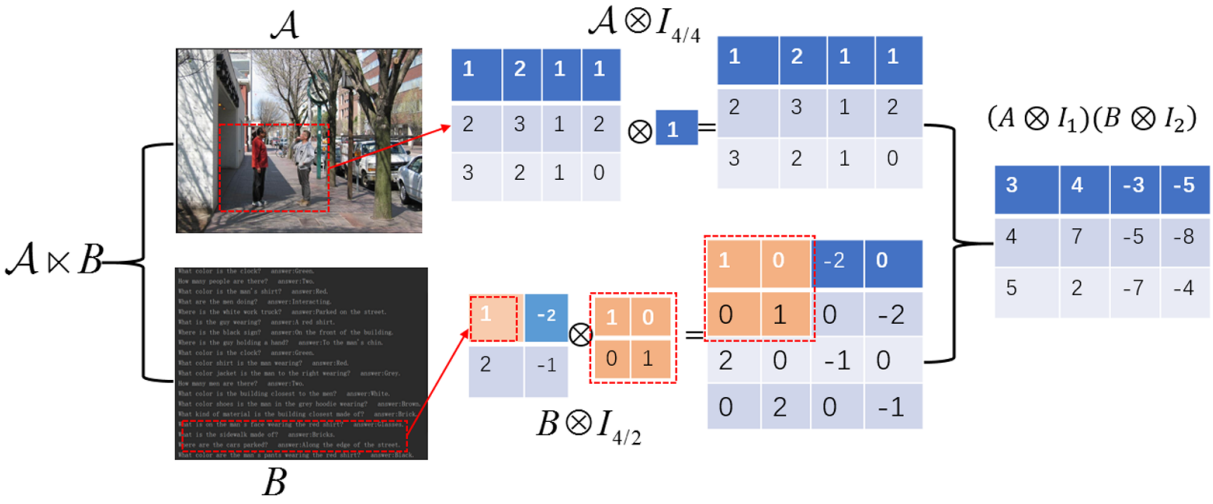

Fig. 2: Semi-Tensor Product in a multi-modal fusion way used in our model.

Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$ and $m/n = p/q = \mu$, the Left Semi-Tensor Addition (STA) [18] of A and B denoted by $\ltimes\!\!\!+$ is

$$A \ltimes\!\!\!+ B := \left(A \otimes I_{t/m}\right) + \left(B \otimes I_{t/p}\right) \tag{7}$$

where $\otimes$, I and t denote the Kronecker product, identity matrix and the least common multiple of n and p, respectively. This rule extends the classical matrix addition to a class of two matrices with different dimensions.

As a generalization of the conventional matrix product, STP is applicable to two matrices of arbitrary dimensions. This generalization keeps all fundamental properties of the conventional matrix product. Moreover, it is a block-wised operation without dimension constraint to be implemented in multi-modal data fusion.

Fig. 2 shows a classical multi-modal information fusion by STP operation. A is an image local region or feature of the image, B is a vector of text or feature of the text, their dimensions are usually variable. The process of STP is divided into two stages. Initially, the Kronecker product operation is conducted. B in Fig. 2 shows that, the Kronecker product extracts the local region information from B and keeps the topology relation, while a signal modulation $B \otimes I_{4/2}$ shows that the Kronecker product extracts the local region information from B and keeps the topology relation. Signal modulation $I_{4/2}$ serves as the carrier signal and B is the modulation signal. Furthermore from Fig. 2 we see that the Kronecker product extends the matrix into a certain form which overcomes the matrix multiplication constraint. The second stage is an ordinary matrix multiplication. Moreover, eq. (5) indicates that STP is a block-wised operation. For example, in application of visual question answering or image caption, some blocks represent local feature of the image, and some blocks represent semantic information from the text.

B. Feature padding

In eq. (1), the condition of STP is that n is a factor of p, it can be expressed as $nk = p$ or $n = pk$. Unfortunately, it can not meet the requirement every time. In our research, feature padding model is exploited to tackle this problem. Just like padding in CNN, when padding is needed, zero paddings are adopted, and padding numbers can be calculated by the following equation

$$P = p - p\%n \tag{8}$$

where % denotes modulo operation.

C. Low-rank bilinear model

At each level, the scale difference between image features dimension and question features is very large. In most cases, the image features scale is larger than the question features. According to STP theory, this model works well. However, the efficiency has to be improved. Here, we introduce a low-rank bilinear model to split one high dimension matrix into two or more low dimension matrices.

Let W_i be weight matrix. It can be substituted with the multiplication of two smaller matrices $U_i V_i^T$ by matrix factorization proposed in [19], where $U_i \in R^{N \times d}$ and $V_i \in R^{M \times d}$. The rank of w_i is at most $d \leq \min(N, M)$ for the scalar output

$$f_i = X^T W_i Y \approx X^T U_i V_i^T Y = \mathbf{1}^T \left(U_i^T X \ltimes V_i^T Y\right) \tag{9}$$

where $\mathbf{1}^T$ denotes the unitary matrix and $\ltimes$ is the STP. X and Y denote input feature vectors, respectively. In order to reduce the number of parameters significantly, we introduced low-rank bilinear pooling

$$f = P^T \left(U_i^T X \ltimes V_i^T Y\right) \tag{10}$$

D. Bilinear attention network of Semi-Tensor Product

The attention mechanism provides an efficient way to explore the distribution of multiple inputs. We introduce it in accordance with the low-rank bilinear pooling and definition of the Semi-Tensor Product Attention defined as

$$\alpha := \text{soft}\max\left(P^T \left(U_i^T X \ltimes V_i^T Y\right)\right) \tag{11}$$

where $\alpha \in \mathbb{R}^{G \ltimes \phi}$, $\phi = \left|\{y_j\}\right|$, $\mathrm{P} \in \mathbb{R}^{d \times G}$, $\mathrm{U} \in \mathbb{R}^{N \times d}$, $\rho = \{X_i\}$, $\mathbf{1} \in \mathbb{R}^{\phi}$, $\mathrm{V} \in \mathbb{R}^{M \times d}$ and $\mathrm{Y} \in \mathbb{R}^{M \times \phi}$. X and Y are features from images and questions, respectively. Furthermore, α is a selective combination of input X and Y using the low-rank bilinear pooling for a downstream application. If $U_i^T X$ and $V_i^T Y$ do not satisfy matching condition of multiple dimension, we take padding solution according to eq. (8).

In the next step, Bilinear Semi-Tensor Product Attention (BSTPA) is introduced

$$f'_k = \left(X^T U\right)_k^T A \left(Y^T V'\right)_k \tag{12}$$

where $\mathrm{U}' \in \mathbb{R}^{N \times K}$, $\mathrm{V}' \in \mathbb{R}^{M \times K}$, $\left(X^T U'\right)_k \in \mathbb{R}^{\rho}$, $\left(Y^T V'\right)_k \in \mathbb{R}^{\phi}$, $A \in \mathbb{R}^{\rho \times \phi}$. BSTPA reduces the two channels simultaneously. Moreover, f'_k denotes the $k-th$ element of intermediate representation, and k is the index of the column.

We can rewrite eq. (7) as an explicit form of computation

$$\begin{aligned} f'_k = & \sum_{i=1}^{\rho} \sum_{j=1}^{\phi} A_{ij} \left(X_i^T U'_k\right) \left(V_k^T Y_j\right) \\ = & \sum_{i=1}^{\rho} \sum_{j=1}^{\phi} A_{ij} X_i^T \left(U'_k V'_k\right) Y_j \end{aligned} \tag{13}$$

where X_i^T and Y_j are the $i-th$ channels of input X and $j-th$ channel of input Y, respectively. U'_k and V'_k denote the $k-th$ column of U' and V' matrices, respectively. Moreover A_{ij} denotes an element in $i-th$ and $j-th$ column of A. According to eq. (6) by using STP and bilinear transformation the attention map A can be defined

$$A := \mathrm{soft\,max}\left(\left(\left(1 \cdot P^T\right) \times X^T U\right) V^T Y\right)) \tag{14}$$

Then, each logit A_{ij} of the softmax is the output of low-rank bilinear pooling

$$A_{ij} := P^T \left(\left(U^T X_i\right) \times \left(V^T Y_j\right)\right) \tag{15}$$

For brevity, we define Semi-Tensor Product Attention network as a multi-modal function with different dimensions parameterized by the bilinear map

$$f = STPA(X, Y; A) \tag{16}$$

Compare to other attention mechanism using point-wise operation, Semi-Tensor Product Attention is a block vs. block form operation. It keeps temporal and spatial information of image and question, and the attention has clear practical significance.

E. Stacked learning of attention

According to Semi-Tensor Addition definition and inspired by Multi-modal Residual Networks (MRN) [20], a novel stacked learning attention is proposed in the present study to integrate the attention form different levels in Fig. 1. The $i+1-th$ output is

$$f_{i+1} = STPA(X, Y; A) + f_i \tag{17}$$

where f_0 is the first layer output of the STPA and f_i is current layer attention. It accumulates all the attention form first layer to last layer, therefore we called that stacked learning of attention.

III. Research results and discussion

In the research we have used VQA 2.0 data and Visual Genome QA data to evaluate our proposed model.

In the first part of our experiments, to extract features from images we have used VGG tool for images from ImageNet organized in accordance with WordNet hierarchy. We have chosen this way since concepts in WordNet have many words of description and ImageNet illustrations are of high quality. This composition of extraction tools is widely used in various tasks of computer vision. Applied BERT model was utilized in the research to extract question vector, and a total of 3.3 billion words were used in BERT training. Of these, 2.5 billion words came from Wikipedia, while the remaining 800 million words came from Books Corpus. Then, the proposed STPA visual question answering model was evaluated on VQA 2.0. Research results show that proposed model has significant superiorities, including more balancing and reduced language biases over VQA 1.0. Moreover, the size of VQA 2.0 is almost twice of that for VQA 1.0. Furthermore proposed STPA approach gives fine-grained visual understanding and a multilevel semantic understanding.

In the second part of our experiments, the sizes of question embedding and image features for our model are set to 2048 and 1024, respectively. Furthermore, the size of the joint representation is the same as the rank in the low-rank bilinear pooling. Additional model parameters used in training are as suggested in literature, dropout is 0.2 [22], weight normalization as [21] and Adamax optimizer as [23]. The learning rate changes from 16 to 22, and the batch size is 512. We have used Visual Genome QA data set to verify our model. In general it contains 1445322 questions on 108077 images. Since the official split of this dataset is not released, we have applied our split constructed randomly for this study. In our split we have 443757 training questions, 214354 validation questions and 447793 testing questions/answers. The following accuracy metric was considered to evaluate results

$$ACC = \min\{\frac{\text{humans that provided answer}}{3}, 1\} \tag{18}$$

In Fig. 3 we can the training process statistics in measures of loss and score functions in relation to iterations. We can see that our model absorbs training well and with each new iteration score measure is growing, while loss function has significant changes in first iterations wile further the adjustment is smooth.

In Tab. II we present comparisons to four reference methods. Unitary attention is a method in which question embedding vector is used to calculate attention weights for multiple image features. For research we have selected methods which are similar in the way of features selection and information processing. All selected approaches have neural mechanism of vector comparison and various propositions of self-attention mechanism utilized to combine hierarchical question embedding into the single embedding vector. Results from Tab. II show that proposed BSTPA is significantly better than other attention methods. Other examined approaches have very similar results. Additional advantage of proposed method is that BSTPA is capable of more information processing.

Tab. III shows comparisons of the results to a simple unitary model used as a reference for direct performance

TABLE I: ablation study on the Visual Genome QA dataset.Accuracy for different question type are shown.

Model	Accuracy							WUPS	
	What	Where	When	Who	Why	How	Overall	0.9	0.1
VGG+LSTM [29]	35.12	16.33	52.71	30.03	11.55	42.69	32.46	38.30	58.39
HieCoAtten+VGG [8]	36.88	16.85	52.74	32.30	11.65	44.00	33.88	39.	0.6568
VQA-machine [10]	44.91	18.87	52.33	38.87	12.88	46.08	39.30	44.96	61.21
BAN [16]	58.42	48.28	81.77	57.49	47.90	61.19	58.37	68.26	85.73
BSTPA (ours)	**59.21**	**48.40**	**81.94**	**57.82**	**49.32**	**61.15**	**59.07**	**68.63**	**86.28**

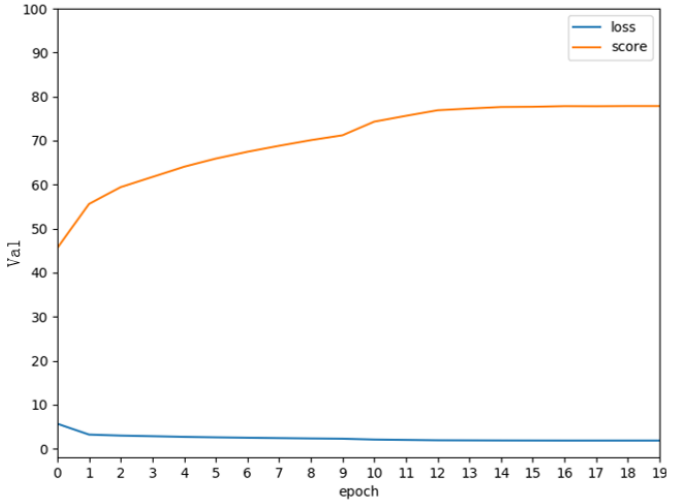

Fig. 3: Training measures visible in loss and score functions calculated for each of iterations.

TABLE II: Validation scores on VQA 2.0 datasets for different attention and integration mechanisms.

Model	nParameter	VQA score
Unitary attention [24]	31.9M	64.59±0.04
Co-attention [25]	32.5M	64.79±0.06
Bilinear-attention	32.2M	65.36±0.14
BAN-4 [16]	44.8M	65.81±0.09
BSPTA (ours)	**91.3M**	**79.90±0.06**

evaluation. Proposed BSTPA model again has superiority over other attention networks. We can see that overall accuracy of this model is 71.02, which outperforms the best of other models. BSPTA network appliance elevates the overall performance steadily. Regarding yes/no and other issues, the accuracy variation tendency is similar to the overall accuracy. For the number issue, the counter model is superior to other models. The reason why the proposed BSTPA surpasses other methods is that introduced BERT learns in more common sense, while the standard block-wise operation of the semi-tensor product can realize information fusion from multiple levels. BERT is a method for pre-training language representations. In this method, it is intended to train a general-purpose "language understanding" model on a large text corpus such as Wikipedia and then use that model for question answering of a special task. We have applied BERT since it outperforms other methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. Moreover, pre-trained representations are classified into context-free or contextual presentations, while contextual representations are

TABLE III: Comparison of test-dev and other test-standard scores on VQA 2.0 dataset.

Model	Overall	Yes/no	Number	Other	Test-std
Bottom-Up [26], [27]	65.32	81.82	44.21	56.05	65.67
Counter [28]	68.09	83.14	51.62	58.97	68.41
MFH+Bottom-Up [25]	68.76	84.17	49.56	59.89	–
BAN [16]	69.52	85.31	50.93	60.26	–
BSTPA (ours)	**71.02**	**85.80**	**54.22**	**60.97**	**69.86**

classified into unidirectional or bidirectional representations. BERT is unsupervised system, what means that only a plain text corpus is utilized to train BERT. It is significant since an enormous amount of plain text data is publicly available on the web in arbitrary languages. Our pre-trained domain is conducted on the VQA dataset. All of questions and answers are initially organized in the dataset to match the required input format of BERT. Then the domain pre-training is performed on the original BERT. Finally, the parameters of the BERT's penultimate layer are extracted as the word embedding of the question.

Tab. I presents the accuracy of different types of questions using Visual Genome question answering model. We have selected four representative models to be compared to the proposed BSTPA approach. The first baseline is the VGG+LSTM discussed by [29], which simultaneously takes two layers of the LSTM and VGG schemes to process question and image, respectively. The second approach is the HieCoAtt+VGG scheme proposed by [8], which jointly gives reasoning about image and question. The third reference model is VQA-machine introduced by [10], where external off-the-shelf algorithms were in-composed for devoted tasks. The percentage for each type of question is calculated. We can see that for each question type proposed BSTPA gives best results, even if for questions "where", "when", and "how" the results are only slightly better or almost equal to BAN model. Moreover, WUPS at 0.9 and 0.1 for different models are calculated. In both categories proposed approach is the best however the results are only slightly better from other tested approaches.

IV. Conclusions and Future Works

In this work, a new attention mechanism called BSTPA was discussed. In our approach we have replaced semi-tensor product attention module with bilinear attention module with the same parameters. The obtained results indicate that the proposed method has stronger selectivity and efficient integration capability, and that overall performance was improved.

One of advances was introduction of BERT scheme to improve VQA performance. As a pre-trained model using large-scale corpus, BERT plays an important role in advanced

processing. Results show that introduction of BERT improves computer vision capabilities in the same way as VGG. BERT scheme provides features, which are beneficial for various applications with pre-trained domains of particular solution spaces. Moreover, we see that semi-tensor product has a widespread application in the field of information fusion. Proposed BSTPA approach gives more generalized form of the product. It provides a block-wise mechanism and substitutes point-wised operations. Combination of VGG and BERT produce multilayer feature, so that a residual network can be conducted by the semi-tensor product. Therefore as a result we receive a hierarchy and multi-scale attention network.

The future works will be oriented on additional acceleration of feature extraction by the use of another CNN architectures for images and improved language models for sentences. Another important aspect will be more efficient parallelization of processes in our model.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No.: 61761042, 61871460, 61941112, 61702091), Key Research and Development Program of Yanan (Grant No. 2017KG-01, 2017WZZ-04-01), and the Natural Science Foundation of Shaanxi (Grant No. 2020JM-556). This work was also supported by the Key Research and Product Program of Shaanxi Province (No.: CXY201909, YDZ2019-05).

REFERENCES

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

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

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'On Semi-tensor Product of Matrices and its Applications," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Daizhan, Cheng and Lijun ,Zhang. "Semi-tensor compressed sensing for hyperspectral image," in *ACTA, Math, App., Sinica*.2003, pp. 219-228

[5] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[6] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[7] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[8] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[9] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4622–4630.

[10] P. Wang, Q. Wu, C. Shen, and A. van den Hengel, "The vqa-machine: Learning how to use existing vision algorithms to answer new questions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1173–1182.

[11] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 39–48.

[12] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*, 2016, pp. 1378–1387.

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[15] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*, 2018.

[16] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 1564–1574.

[17] C. Gulcehre, M. Denil, M. Malinowski, A. Razavi, R. Pascanu, K. M. Hermann, P. Battaglia, V. Bapst, D. Raposo, A. Santoro *et al.*, "Hyperbolic attention networks," *arXiv preprint arXiv:1805.09786*, 2018.

[18] D. Cheng, "One quivalence of matrices," *arXiv preprint arXiv:1605.09523v4*, 2017.

[19] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Bilinear classifiers for visual recognition," in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009.*, 2009, pp. 361–369.

[20] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in *Advances in neural information processing systems*, 2016, pp. 361–369.

[21] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 901–909.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[25] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–13, 2018.

[26] A. Peter, H. Xiaodong, B. Chris, T. Danien, S. G. Mark ohnson, and L. Zhang, "Bottom-up -up and top-down attention for image caption and visual question answering," *arXiv preprint arXiv:1707.07998*, 2017.

[27] X. H. Danien Tency, Peter Anderson and A. V. den Gengel, "Tips and tricks for visual question answering:learning form the 2017 challenge," *arXiv preprint arXiv:1708.02711*, 2017.

[28] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," *arXiv preprint arXiv:1802.05766*, 2018.

[29] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[30] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," *arXiv preprint arXiv:1908.06066*, 2019.

[31] L. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "Visualbert:

A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[32] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[33] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.