Deep Attention Neural Tensor Network for Visual Question Answering

Yalong Bai[1,2], Jianlong Fu[3], Tiejun Zhao[1], and Tao Mei[2]

[1] Harbin Institute of Technology, Harbin, China
[2] JD AI Research, Beijing, China
[3] Microsoft Research Asia, Beijing, China
{baiyalong,tmei}@jd.com, jianf@microsoft.com, tjzhao@hit.edu.cn

Abstract. Visual question answering (VQA) has drawn great attention in cross-modal learning problems, which enables a machine to answer a natural language question given a reference image. Significant progress has been made by learning rich embedding features from images and questions by bilinear models, while neglects the key role from answers. In this paper, we propose a novel deep attention neural tensor network (DA-NTN) for visual question answering, which can discover the joint correlations over images, questions and answers with tensor-based representations. First, we model one of the pairwise interaction (e.g., image and question) by bilinear features, which is further encoded with the third dimension (e.g., answer) to be a triplet by bilinear tensor product. Second, we decompose the correlation of different triplets by different answer and question types, and further propose a slice-wise attention module on tensor to select the most discriminative reasoning process for inference. Third, we optimize the proposed DA-NTN by learning a label regression with KL-divergence losses. Such a design enables scalable training and fast convergence over a large number of answer set. We integrate the proposed DA-NTN structure into the state-of-the-art VQA models (e.g., MLB and MUTAN). Extensive experiments demonstrate the superior accuracy than the original MLB and MUTAN models, with 1.98%, 1.70% relative increases on VQA-2.0 dataset, respectively.

Keywords: Visual question answering· Neural Tensor Network· Open-ended VQA

1 Introduction

After deep learning techniques have achieved great success in solving natural language processing and computer vision tasks, automatically understanding the semantics of images and text and eliminating the gap between their representations has received intensive research attention. It has stimulated many new research topic like image captioning [8], text to image synthesis [23] and visual question answering [4, 10].

The Visual Question Answering (VQA) is a task about answering questions which posed in natural language about images. The answers can either be se-

lected from multiple pre-specified choices or generated by a model. A natural solution for VQA is to combine the visual based image understanding with the question based on natural language understanding and reasoning. Recently, many studies have explored the multi-modal feature fusion of image representation learned from deep convolutional neural network and question representation learned from time sequential model. Nearly all of these previous works train a classifier based on the fusion of image and question feature to predict an answer, and the relationship of image-question-answer triplets is ignored. While in theory, "these approaches have the potential simple reasoning, it is still not clear whether they do actually reason, or they do so in a human-comprehensible way" as Allan *et al.* [12] mentioned. To model the relational information in triplet, there are some other related works try to use pretrained answer representations to help reasoning, by simply concatenating the features of image, question and answer [12], or projecting the image-question feature into answer feature space [27], but the relational information of image-question-answer triplet is too complex to be modeled by using simply concatenating feature vectors or applying element-wise sum or product. Moreover, the answer representations learned from natural language corpus, which is supervised by the syntactic and semantic information in the corpus, still has a certain gap to describe visual information.

Inspired by the success of neural tensor network for explicitly modeling multiple interactions of relational data [26, 22], we proposed a neural tensor network based framework to model the relational information of image-question-answer triplets and learn the VQA task-specific answer representations from scratch. As we know, typically different triplets in VQA correspond to different kinds of relationship and different reasoning process. In most cases, these relationship is associated with the type of question well. Moreover, the responses of candidate answers are also helpful to predict the question's type. Thus we introduce a novel question and answers' responses guided attention mechanism into our proposed deep neural tensor VQA framework by adaptively reasoning for different triplets according to their implicit relation types. After that, we use a regression-based method to approximate the distributions of image-question-candidate answers instead of the traditional classification-based method. We optimize our proposed model by learning a label regression with KL-divergence losses. Such a design enables scalable training and fast convergence over a large number of answer set.

Different from the previous works, we introduce the answer embedding learning in our method for three purposes. First, we want to model the relationship among image-question-answer triplet to help to reason. Second, the answer embedding may correct the question misunderstanding especially for questions with complex syntactic structures. Third, the answer embedding can help to determinate the type of question and to decide using which kind of reasoning process.

We evaluate the impact of our proposed framework on VQA-1.0 and VQA-2.0 datasets. Since our proposed framework is designed to be applicable to most of the previous image-question multimodal feature learning based models, we selected two of the most powerful bilinear pooling based VQA models to equip our proposed framework, and prove that our proposed method can achieve more

reasonable answer representations and further result in significant improvement on the VQA performance.

In the next section, we provide more details on related works and highlight our contributions. Our proposed method is presented in section 3, and the successful experiments are reported in section 4. We analyzed the result of experiments in section 5, and conclude with a discussion in Section 6.

2 Related Works

The task of VQA has gathered increasing interest in the past few years. Most of the previous works pose the visual question answering as a classification problem and solved with a deep neural network that implements the joint representation of the image and question. Only a few related works introduce answer representation for reasoning. Meanwhile, the question-guided visual regions attention is also very important for VQA. In this section, we briefly review these related works.

Attention mechanisms have been a real breakthrough in VQA task. Chen *et al.* [7] proposed a question-guided attention mechanism to adaptively learn the most relevant image regions for a given question. Zichao *et al.* [31] proposed to stack multiple question-guided attention mechanisms to learn the attention in an iterative way. Fukui *et al.* [9] and Hedi *et al.* [6] used bilinear pooling to integrate the visual features from the image spatial grids with question features to predict attention. Considering the questions in natural languages may also contain some noise or useless information or words, some co-attention based frameworks designed for jointly learn the attention for both the image and question are also proposed [17, 32]. In this paper, we apply the attention mechanisms used in [9, 6] to learn the attention on the relevant visual regions and discard the useless information regarding the question.

Classification based Methods. The answers in the current VQA datasets only span a small set of words and phrases. Thus most of the related works posed the VQA task as a classification over a set of candidate answers. As a result, the image and question feature fusion strategies become the key factor for improving the performance of VQA. Early works modeled interactions between image and question with first order interactions like concatenation [24] or element-wise product [32, 13, 16]. Second order pooling is a more powerful way to model interactions between two feature spaces. It has shown great success in the fine-grained visual recognition task. Fukui *et al.* [9] first introduced the bilinear pooling on VQA task. They proposed the Multimodal Compact Bilinear pooling (MCB), which use the outer product of image and question feature vectors in different modalities to produce a very high-dimensional feature for quadratic expansion. However, MCB usually needs high-dimensional features to guarantee robust performance, which may seriously limit its applicability for VQA due to the limitations in GPU memory. To overcome this problem, Multimodal Low-rank Bilinear pooling (MLB) [14] are proposed, in which the bilinear interactions between image and question feature spaces are parametrized by a tensor and the tensor is

constrained to be a low rank. After that, Hedi *et al.*proposed Multimodal Tucker Fusion (MUTAN) [6] which is also a multimodal fusion scheme based on bilinear interactions between modalities but relying on a low-rank Tucker tensor-based decomposition to explicitly constrain the interaction rank.

Image-question-answer Triplet based Reasoning. Different from the classification based methods, there are some other related works try to introduce the answer representations into the reasoning for visual question answers. Shih *et al.* [25] combined the question and answer as input for the model to determine whether a question-answer pair is a good match given evidence from the image. Allan *et al.* [12] concatenate image feature vector, question feature vector and answer embedding as input variables and predict whether or not an image-question-answer triplet is correct. The work in [27] try to project the image-question jointly representation into the answer embedding space learned from a text corpus. Both the work of Allan *et al.* [12] and Teney *et al.* [27] used the answer embedding learned from text corpus which has been proved as having limited ability to represent visual information [5]. Moreover, reasoning about the relations among image-question-answer triplets should be very complex, it should be hard to be model by using simple concatenating feature vectors or element-wise product.

In this work, we introduce DA-NTN, a deep attention based neural tensor network for reasoning the complex relations between image-question-answer triplet. The answer embedding used in this work is learned from scratch by supervision of VQA task. DA-NTN can be applied to traditional classification based VQA models easily and significantly boost the performance of these methods.

3 Approach

Figure 1 provides an overview of the architecture of our open-ended visual question answering framework. The goal of the VQA task is to provide an answer given an image $I \in \mathcal{I}$ and a corresponding question $q \in \mathcal{Q}$. Most of the previous works regard the open-ended VQA as a classification task:

$$\arg\max_{a_i \in \mathcal{A}} p_\theta(a_i|q, I) \tag{1}$$

where θ means the whole set of parameters of the model, and $\mathcal{A}$ is the set of candidate answers. However, in our proposed model, we treat the open-ended VQA as a regression task, that our proposed method target at measure the relevance score $s_\theta(q, I, a_i)$ among image I, question a, and answer a_i, and then predicts whether or not an image-question-answer triplet is correct.

The inputs to our model contain a question and a corresponding image and candidate answers. A convolutional neural network and a GRU recurrent network are used for extracting feature vectors for image and question respectively. Then the representations of image and question are integrated as multi-modal features by using bilinear pooling module such as MLB [14], MUTAN [6]. At last, a DA-NTN module is applied to measure the correlation degree between the integrated feature vector v_{qI} of question-image pair and representation of input answer.

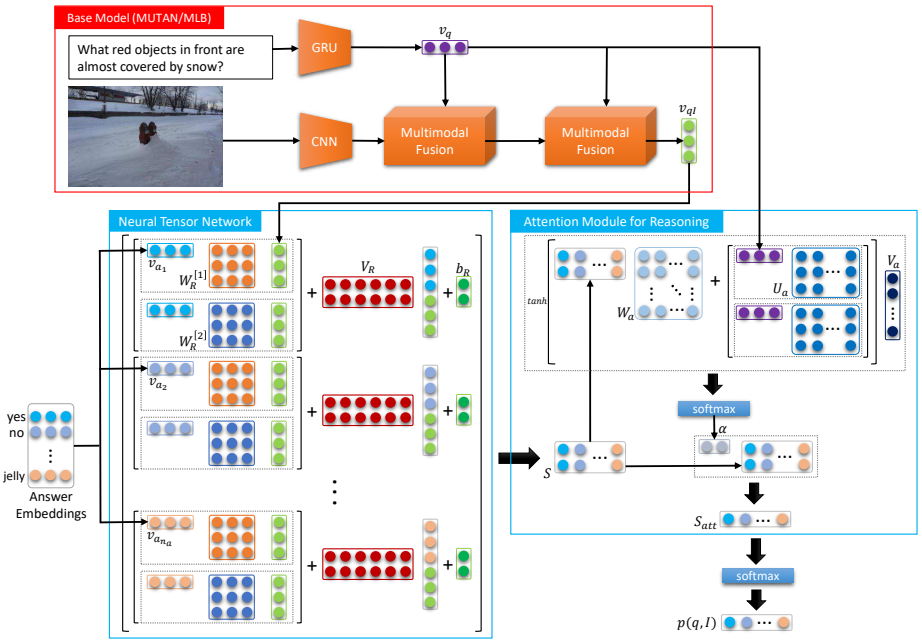

Fig. 1. Overview of our proposed framework for visual question answering. Image, question and all candidate answers are jointly fed into this framework. The structure in the red box is the base model used to generate question representation v_q and the fusion of image and question feature vector v_{qI}. The structure in two blue boxes is our proposed Deep Attention Neural Tensor Network. The blue box named neural tensor network is applied to measure the relevance among image-question-answer triplet, the tensor can represent the implicit relations among triplets. The blue box named attention module for reasoning is designed to adaptively reason for different triplets according to their implicit relation types. (Best viewed in color)

3.1 Neural Tensor Networks for VQA

As we show in Figure 1, DA-NTN module target to measure the relevance score of the image-question-answer triplet. For the VQA task, the image-question pairs are predefined. Thus the relevance of image-question-answer triplet can be rewritten as the relevance between image-question pair and answer.

Following the previous works, we first get the image-question pair's representation v_{qI}. To model the interactions between image-question representation v_{qI} and candidate answer representation v_{a_i}, we need to utilize some metrics to measure their relevance. Given these two feature vector, the traditional ways are to calculate their distance directly or simply concatenate the vectors then feed into a regressor or classifier. However, these two ways cannot sufficiently take into account the complicated interactions between image-question pair and answer.

In this paper, we model the relevant degree of image-question pair and answer in a non-linear way. Considering tensor is a geometric object that describes relations between vectors, and also been able to explicitly model multiple interactions in data [26, 22], we proposed a Neural Tensor Network (NTN) based module to relate the image-question feature vector and answer feature vectors. As a result, the relevance degree between image-question pair $\langle q, I \rangle$ and answer a_i can be measured as shown in Equation 2.

$$s(q, I, a_i) = v_{qI} W_R^{[1:k]} v_{a_i} + V_R \begin{bmatrix} v_{qI} \\ v_{a_i} \end{bmatrix} + b_R \tag{2}$$

where v_{a_i} is the feature vector of answer a_i. R means the implicit relationships between image-question pair and answer. $W_R^{[1:k]} \in \mathbb{R}^{d_{qI} \times d_a \times k}$ is a tensor and the bilinear tensor product $v_{qI} W_R^{[1:k]} v_{a_i}$ results in a k-d vector $h \in \mathbb{R}^k$, where each $\langle q, I, a_i \rangle$ with a special relationship type $rel_r \in R$ can be computed by a corresponding slice $r = 1, ..., k$ of the tensor: $h_i = v_{qI} W_R^{[i]} v_{a_i}$. The other parameters for implicit relationships R are the standard form of a neural network: $V_R \in \mathbb{R}^{k \times (d_{qI} + d_a)}$ and $b_R \in \mathbb{R}^k$. As a result, we can get a k-d vector $s(q, I, R, a_i)$ to measure the relevance degree between image-question pair and answer, and each element in the vector represent the response of image-question-answer triplet with a specific implicit relationship.

Following the settings of previous works, both of the visual representation v_I and question representation v_q are initialized from a pre-trained model, then fine-tuned during the training procedure of VQA task. But for answer a_i, its representation v_{a_i} should be provided with visual information for reasoning. Traditional word embeddings learned from natural language corpus are not suitable for modeling visual information. For example, the nearest words of "dog" in word representations space learned from the natural language corpus are some other words describing animals like "pet", "cat", etc. The word embeddings learned from natural language corpus can distinguish the semantic and syntax differences between answers but it is hard to be used for visual question answering task which requires the ability to describe visual information [5]. Thus, We try to learn the answer representation v_{a_i} for VQA task from scratch instead of directly using the word representations learned from natural language corpus, which is different with previous related works.

3.2 Attention Module for Reasoning

Since each element in the vector $s(q, I, a_i)$ is designed to correspond to one particular relationship and reasoning process of $\langle q, I, a_i \rangle$, we propose an attention mechanism to combine them by dynamically adjusting the weight of each element in the vector. For VQA task, the relationship of $\langle q, I, a_i \rangle$ triplet usually be decided by the type of question q. For example, the relationships of triplets can be split as object recognition, object location, object counting, object attributes, etc. All of these relationship classes can be recognized according to the meaning of the question. Moreover, the responses of all candidate answers can

also provide more detail information about the question type. For example, if one question is answering about colors, the responses of candidate answers about color should have larger responses than other candidate answers.

Specifically, we use the attention mechanism to obtain the weighted average of each element in the relevant vector $s(q, i, a_i)$ as the output of the finally score about whether or not $\langle q, I, a_i \rangle$ is correct, which is denoted as

$$s_{att}(q, I, a_i) = \sum_{j=1}^{k} s_{i,j} \alpha_j \tag{3}$$

where $s_{i,j}$ is the j-th element in relevance vector $s(q, I, a_i)$, α_j is the attention weight for the j-th element. The attention score α_j is calculated by

$$\alpha_j = \frac{\exp(c_j)}{\sum_{e=1}^{k} \exp(c_e)} \tag{4}$$

and c_j is defined as

$$c_j = V_a \cdot \tanh(W_a S_j + U_a v_q) \tag{5}$$

where $S_j = \{s_{1,j}, s_{2,j}, ..., s_{n_a,j}\}$ is a vector to represent the responses of all candidate answers given image I, question q and one special implicit relationship type. $W_a \in \mathbb{R}^{n_a \times n_a}$, $U_a \in \mathbb{R}^{n_a \times |v_q|}$, $V_a \in \mathbb{R}^{n_a \times 1}$ are weight matrices of the attention module. The combination weights are determined by the response of all candidate answers and question representations. In this way, multiply image-question-answer implicit relationships are taken into consideration and different reasoning results are integrated according to the responses of candidate answers and the contextual information in question.

3.3 Label Distribution Learning with Regression

In practice, an image-question pair is associated with one or several similar answers. In dataset like VQA [4] and VQA-2.0 [10], each image-question pair is annotated with multiple answers by different people. The answers for each sample can be represented as a distribution vector of all the possible answers $y \in \mathbb{R}^{n_a}$, where $y_i \in [0, 1]$ indicates the occurrence probability of the i-the answer in $\mathcal{A}$ across human labeled answers for this image-question pair.

Since our proposed model output as regression of answer scores, a typical training strategy is to use margin-based loss function to maximize the distance between correct answers and any incorrect answers. However, for open-ended VQA task, there are lots of candidate answers need to be considered. The increasing of negative samples lead to much more positive-negative pairs to train and more complex training procedure. As a result, it is very complex to model the structure of VQA reasoning space by using margin-based loss function with limited negative samples and may also introduce uncertainty to the learned model and take much more iterations to converge. To overcome this problem, we transform the margin based learning problem with negative sampling to label distribution learning (LDL) problem with all answers distributions y.

For each image-question pair, we compute the regression score $s_{att}(q, I, a_i)$ for each answer a_i in overall answer candidate set $\mathcal{A}$. Then use a softmax regression to approach the answers distributions:

$$p_i(q, I) = \frac{\exp(s_{att}(q, I, a_i))}{\sum_{j=1}^{n_a} \exp(s_{att}(q, I, a_j))} \tag{6}$$

The KL-divergence loss function is applied to penalize the prediction $p_i \in \mathbb{R}^{n_a}$, our model is trained by minimizing

$$l = \frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{n_a} y_i \log \frac{y_i}{p_i(q_j, I_j)} \tag{7}$$

where N is the amount of image-question pairs for training.

During inference procedure, we just feed the embeddings of all candidate answers into DA-NTN, and then select the answer a_i with biggest triplet relevance score $s_{att}(q, I, a_i)$ as the final answer.

4 Experiments

In this section, we evaluate the performance of our proposed DA-NTN model on VQA task. We also analyze the implicit relationship used for guiding reasoning and the answer representations learned from VQA task.

Existing VQA approaches usually have three stages: (1) getting the representation vectors of image and question respectively; (2) combining these multimodal features to obtain fused image-question representation; (3) using the integrated image-question features to learn a multi-class classifier and to predict the best-matching answer. Bilinear pooling based methods have been widely used for fusing image-question feature in step 2. We build our model based on below two widely used VQA models by applying the attention-based neural tensor network after step 2 to measure the relevance scores of image-question-answer triplets.

MLB [14]. Using low-rank bilinear pooling in step 2 to approximate full bilinear pooling between image representations and question representations.

MUTAN [6]. A multimodal tensor-based tucker decomposition to efficiently parametrize bilinear interactions between image and question representations.

In order to get convincing comparison between baseline methods and our method, we directly apply the best hyper-parameters on MLB and MUTAN to DA-NTN based MLB and MUTAN respectively. We also reference other previous works for comparison with our DA-NTN based MUTAN and MLB models.

4.1 Dataset

In this paper, we use the VQA-1.0 dataset [4] and VQA-2.0 dataset [10] for evaluating our proposed method.

The VQA-1.0 dataset consists of approximately 200,000 images from MS-COCO dataset with nearly 3 questions per images and each question is answered

by 10 annotators. There are 248k question-answer pairs for the training set, 121k pairs for validation and 244k pairs for testing. Additionally, there is a 25% test split subset named *test-dev*.

VQA-2.0 is another dataset for VQA task. Compared to the VQA-v1.0 dataset, it contains more training samples (440k question-answer pairs for training and 214k pairs for validation) and is more balanced to weaken the potential that an overfitted model may achieve good results. Thus we use VQA-2.0 dataset for experimental results analysis.

In this paper, we focus on the open-ended VQA task, where the ground truth answers are given in free natural language phrases. And we evaluate the VQA accuracy by using the tools provided by Antol *et al.* [4], where the accuracy of a predicted answer a_i is given by:

$$\min\left(1, \frac{\#\text{ annotators the provided } a_i}{3}\right) \tag{8}$$

It means that if the predicted answer a_i appears greater than or equal to three times in human labeled answer list, the accuracy is calculated as 1.

4.2 Experimental Settings

To be fair, we use the same image representations and question representations models for all of the experiments in this paper. We used image features with bottom-up attention [1] from Faster R-CNN as the visual features which produce feature maps of size $K \times 2048$, since the features can be interpreted as ResNet features centered on the top-K objects in the image, where $K < 100$. A GRU initialized with the parameters of a pre-trained Skip-thoughts model [15] is used for learning question representations.

We use the Adam solver as the optimizer for training. The hyper-parameters such as initial learning rate, dropout ratio, the dimension of the image-question feature are set as same with the best settings in the original publications about MLB and MUTAN respectively. Both of them are equipped with visual regions attention module.

DA-NTN Setup. For our proposed attention-based neural tensor network module, we set the dimension of answer representation as 360 for all of the experiments in this paper. The candidate answer set $\mathcal{A}$ is fixed to the top-2000 most frequent answers since the answers in VQA-2.0 dataset follow the long-tail distribution. For the inference procedure, only image and question are required as inputs, then embedding of all candidate answers will be fed into the model, and the answer with biggest triplet relevance score s_{att} will be selected as predicted answer of DA-NTN. To avoid over-fitting, we apply L2-regularization for embeddings of all candidate answers. By default, we set $k = 6$ by considering the trade-off between training complex and performance on the validation set.

4.3 Experimental Results

In Table 1, we compare the performance of our proposed method with the base models. The models are trained on the train set and evaluated on the valida-

Table 1. Comparison between different models for open-ended VQA on the validation split of VQA-2.0 dataset. The model size indicates the number of all learnable parameters, including the parameters of GRU for question representation learning. NTN means neural tensor network without attention module for reasoning. For NTN we use sum-pooling instead of our proposed attention module for reasoning. All: overall accuracy in percentage, Yes/No: accuracy on yes-no questions, Numb: accuracy on questions that can be answered by numbers or digits, Other: accuracy on other types of questions.

Model	Model Size	VQA-2.0 val set			
		Yes/no	Numb.	Other	All
MUTAN	38.0M	81.09	42.25	54.41	62.84
MUTAN + NTN ($k = 3$)	39.3M	81.69	43.88	55.35	63.74
MUTAN + NTN ($k = 6$)	39.9M	81.96	43.63	55.39	63.83
MUTAN + NTN ($k = 10$)	40.6M	82.23	43.34	55.33	63.86
MUTAN + DA-NTN ($k = 3$)	48.1M	81.96	44.59	55.63	64.07
MUTAN + DA-NTN ($k = 6$)	48.7M	81.98	44.85	55.72	64.16
MUTAN + DA-NTN ($k = 10$)	49.4M	82.24	44.55	55.43	64.07
MLB	67.2M	81.89	42.97	53.89	62.98
MLB + DA-NTN ($k = 6$)	87.5M	83.09	44.88	55.71	**64.58**

tion set. Furthermore, we explore different hyper-parameters for our proposed attention-based neural tensor network. It worth to note that the average accuracies of our implemented baseline MLB and MUTAN on VQA-2.0 dataset are already 5.7% and 4.9% higher than the performance reported in previous works [6] respectively.

From Table 1, we can find that: (1) MUTAN + NTN gives better results than MUTAN, even with a small number of implicit triplet relationship, like $k = 3$. This shows that the neural tensor network is able to learn powerful correlations among image-question-answer triplets. (2) The attention module for reasoning benefit the performance of NTN, we can see that the DA-NTN achieves better performance than NTN. This phenomenon proved, that different types of image-question-answer triplets should correspond to different reasoning process, and the attention module for associating the triplet with its relevant reasoning process is important for VQA.(3) Even using the same DA-NTN hyper-parameter settings of MUTAN ($v_{qI} \in \mathbb{R}^{512}$) for MLB ($v_{qI} \in \mathbb{R}^{4800}$), our proposed DA-NTN still can significantly boost the accuracy of MLB.

Table 2 reports the experimental results on *test-dev* and *test-stand* set of VQA-2.0 dataset. All of the models in Table 2 are trained on the combination of train set and validation set, without any data augmentation. From the results, We can find that the models with DA-NTN have stable improvements than base

Table 2. The performance of different single model for open-ended VQA on the test-dev and test-stand set of VQA-2.0 dataset.

Model	VQA-2.0 Test-dev set				VQA-2.0 Test-standard set			
	Y/N	No.	Other	All	Y/N	No.	Other	All
Prior [10]	-	-	-	-	61.20	0.36	1.17	25.98
LSTM (blind) [10]	-	-	-	-	67.01	31.55	27.37	44.26
MCB [10]	-	-	-	-	78.82	38.28	53.36	62.27
MUTAN	82.88	44.54	56.50	66.01	83.06	44.28	56.91	66.38
MLB	83.58	44.92	56.34	66.27	83.96	44.77	56.52	66.62
MUTAN + DA-NTN	83.58	46.78	57.77	67.15	83.92	46.64	58.0	67.51
MLB + DA-NTN	84.29	47.14	57.92	**67.56**	84.60	47.13	58.20	**67.94**

models, and our DA-NTN based models archived the best accuracy on all of the three different types of questions.

Considering that most of the previous works compare their performance on VQA-1.0 dataset, we also provide the experimental results on the VQA-1.0 dataset in Table 3. Similar with the experimental results on VQA-2.0 dataset, our proposed DA-NTN can provide steady improvement.

5 Analysis

To get the deep insight of our proposed method, in this section, we conduct the studies to investigate how the reasoning attention module helps to improve the performance of the base model, and we also analyze the answer embedding learned from VQA task.

5.1 Attention Module Analysis

As we mentioned in Section 3.2, the relationship among image-question-answer triplet and its relevant reasoning process should be decided by the type of question. To further analyze, how the reasoning attention module works, we counted the average attention scores corresponding to different implicit relationships for each type of question. Figure 2 presents the distributions of attention score computed by MUTAN + DA-NTN on different types of questions in the validation set of VQA-2.0. Since we set $k = 6$ in this experiment, each question type has six attention scores corresponding to six different kinds of implicit relationship and reasoning process.

From Figure 2, we can observe that each implicit relationship pay attention to at least one specific question type. For example, the attention score α_1 for implicit relationship rel_1 is significantly bigger than others on questions about color. α_2 is bigger than other attention scores on questions about the number of objects. The combination of rel_3 and rel_4 is focussing on questions that have

Table 3. Comparison between different single models for open-ended VQA on the test-dev and test-stand set of VQA-1.0 dataset. † : use GloVe [21] as pretrained word embedding model for question representation. ‡ : use Skip-thought [15] as pretrained word embedding model for question representation. * : use image features with bottom-up attention [1].

Model	VQA-1.0 Test-dev set				VQA-1.0 Test-standard set			
	Y/N	No.	Other	All	Y/N	No.	Other	All
iBOWIMG [33]	76.5	35.0	42.6	55.7	76.8	35.0	42.6	55.9
DPPnet [20]	80.7	37.2	41.7	57.2	80.3	36.9	42.2	57.4
VQA team [4]	80.5	36.8	43.1	57.8	80.6	36.5	43.7	58.2
SMem [30]	80.9	37.3	43.1	58.0	80.9	37.5	43.5	58.2
AYN [18]	78.4	36.4	46.3	58.4	78.2	36.3	46.3	58.4
NMN [3]	81.2	38.0	44.0	58.6	81.2	37.7	44.0	58.7
SAN [31]	79.3	36.6	46.1	58.7	-	-	-	58.9
AMA [28]	81.0	38.4	45.2	59.2	81.1	37.1	45.8	59.4
D-NMN [2]	81.1	38.6	45.5	59.4	-	-	-	59.4
FDA [11]	81.1	36.2	45.8	59.2	-	-	-	59.5
DMN+ [29]	80.5	36.8	48.3	60.3	-	-	-	60.4
MRN [13]	82.3	38.9	49.3	61.7	82.4	38.2	49.4	61.8
HieCoAtt [17]	79.7	38.7	51.7	61.8	-	-	-	62.1
RAU [19]	81.9	39.0	53.0	63.3	81.7	38.2	52.8	63.2
MCB† [9]	82.5	37.6	55.6	64.7	-	-	-	-
MLB‡ [14]	84.1	38.2	54.9	65.1	84.0	37.9	54.8	65.1
MFB† [32]	84.0	39.8	56.2	65.9	83.8	38.9	56.3	65.8
MUTAN‡*	83.3	39.7	56.6	65.7	83.2	40.3	56.4	65.8
MLB‡*	85.1	39.9	55.4	65.9	84.7	39.5	55.5	65.9
MUTAN‡* + DA-NTN	84.5	41.8	57.8	67.1	84.3	41.9	58.0	67.1
MLB‡* + DA-NTN	85.8	41.9	58.6	**67.9**	85.8	42.5	58.5	**68.1**

Yes or No answers, meanwhile, the combination of rel_4 and rel_6 usually focus on questions about "what" and "how".

We can also find that some implicit relationships which are hard to distinguish by simple classification based method also can be detected. For example, all questions with answers related to number or digit are treated in the same way by using traditional methods. However, in practice, the questions about "how many objects" should have totally different reasoning process comparing to questions about "what number is ", because the former target at counting objects in images, while the latter target at recognizing the digits in images. By using our proposed DA-NTN, these two types of questions can be classified into two different implicit relationships and associated with two different reasoning process. In Figure 2 we can find that questions asking about "what number is" have biggest attention score on rel_5, while question asking about "how many (people are/ people are in/)" have biggest attention score on rel_2.

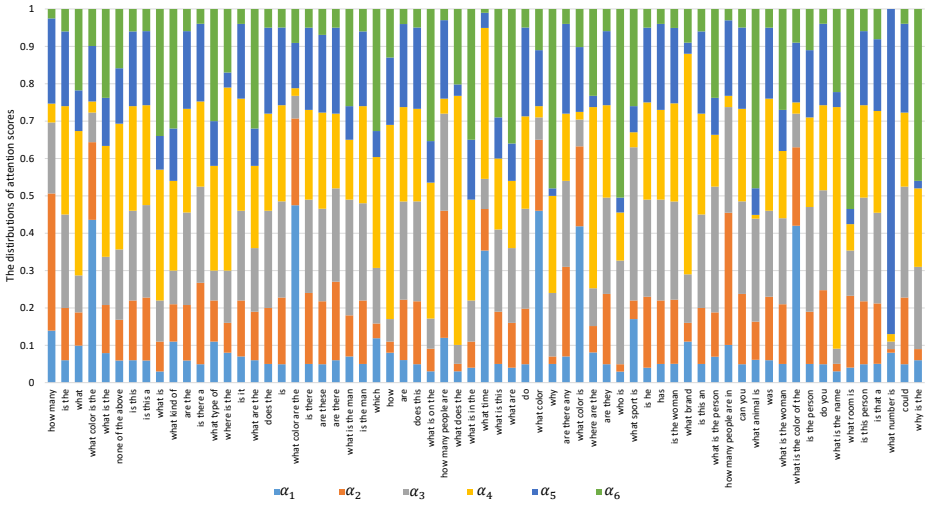

Fig. 2. The distributions of average attention scores across different types of questions. Each attention score α_i is relevant to one specific implicit relationship rel_i. The length of column in different colors indicates the value of attention score of different implicit relationship, higher column means higher attention score. Since we use the softmax function (Equation 4) normalized the distribution of attention score α_i, the sum of average attention scores for each type of question is 1. (Best viewed in color)

With these observations, we can conclude that DA-NTN can effectively model diverse relationships among image-question-answer triplets and benefit the reasoning process of visual question answering.

5.2 Answer Representations Analysis

To gain a deeper understanding of how proposed DA-NTN learn answer representations according to the supervision from VQA, we look at nearest neighbors of several exemplary answers given by word embeddings, where cosine similarity is used as a distance metric. We compare word embeddings learned by DA-NTN with GloVe [21] word embeddings, since GloVe has been used for many previous VQA models [27, 9, 32]. For GloVe, if the answer is a phrase, we averaged the word embedding of each word in the phrase as phrase representation. The experimental results are shown in Table 4.

Obviously, our word representations reflect more about visually similar. For example, it returns “red”, “yellow” and “brown” as the nearest neighbors of word “orange”, since these three colors are very close to red in the standard gradual color bar. Due to lack of the supervision from VQA, the words in GloVe embedding space distribute in a mess, we can find that for each answer, there are many nearest neighbors, and all of these nearest neighbors usually have very small cosine distance with the central word. This makes it more difficult to distinguish candidate answers. Moreover, since the GloVe word vectors are learned

Table 4. For query words, we show their most similar words based on our method and context based word embedding [21]. We also show the cosine similarity scores between query word and its nearest neighbors, only the words whose cosine similarity scores are smaller than -0.3 are shown in this table.

Answers	DA-NTN	GloVe
0	1:-0.43, 2:-0.32	1:-0.60, 5:-0.53, 9:-0.51, 6:-0.51, 3:-0.50, 4:-0.50, 8:-0.50, etc.
orange	red:-0.39, yellow:-0.33, brown:-0.32	orange and yellow:-0.90, orange and blue-0.89, orange juice:-0.88, green and orange:-0.87, etc.
table	on table:-0.35, desk:-0.30	on table:-0.84, picnic table:-0.84, chairs:-0.62, dining room:-0.60, etc.
rectangle	square:-0.34	triangle:-0.64, squares:-0.61, circle:-0.60, oval:-0.59, etc.
glove	baseball glove: -0.34, mitt:-0.33	baseball glove:-0.82, gloves:-0.81, knee pads:-0.57, helmet:-0.56, etc.
playing frisbee	catching frisbee:-0.37, throwing frisbee:-0.35	frisbee:-0.81, throwing frisbee:-0.80, playing tennis:-0.80, playing:-0.80, etc.
river	lake:-0.32, pond:-0.32	lake:-0.72, shore:-0.63, railroad crossing:-0.58, bridge:-0.58, water:-0.58, etc.
middle	center:-0.30	end:-0.64, in corner:-0.64, right side:-0.63, left one:-0.63, etc.

from natural language corpus without visual supervision, there are many semantic or syntactic similarity but visual irrelevance and noisy words are introduced for nearest neighbors during using GloVe. For example, all of nearest neighbors of "middle" (like "end", "in corner", "right side") in the GloVe space has no visual relevance with "middle". This kind of noisy words can mislead the reasoning process for visual question answering.

6 Conclusion

In this paper, a reasoning attention based neural tensor network is designed for visual question answering. We applied our proposed method to different VQA models and got substantial gains for all types of questions. Our analysis demonstrates that our proposed method can not only model the diverse implicit relationship among image-question-answer triples to benefit the reasoning of visual question answering, but also learn reasonable answer representations.

One direction for future work is to apply our DA-NTN to more VQA models, the other direction is to model the relationships of triplet by measuring the relevance between question-answer pair and image, image-answer pair and question, or some more complex combinations of image, question, and answer. We are also interested in learning better answer representations for some specialized tasks such as reading.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017)
2. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Learning to compose neural networks for question answering. arXiv preprint arXiv:1601.01705 (2016)
3. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 39–48 (2016)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015)
5. Bai, Y., Yang, K., Yu, W., Xu, C., Ma, W.Y., Zhao, T.: Automatic image dataset construction from click-through logs using deep neural network. In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 441–450 (2015)
6. Ben-younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: The IEEE International Conference on Computer Vision (ICCV). vol. 1, p. 3 (2017)
7. Chen, K., Wang, J., Chen, L.C., Gao, H., Xu, W., Nevatia, R.: Abc-cnn: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960 (2015)
8. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
9. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. vol. 1, p. 9 (2017)
11. Ilievski, I., Yan, S., Feng, J.: A focused dynamic attention model for visual question answering. arXiv preprint arXiv:1604.01485 (2016)
12. Jabri, A., Joulin, A., van der Maaten, L.: Revisiting visual question answering baselines. In: European conference on computer vision. pp. 727–739. Springer (2016)
13. Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T.: Multimodal residual learning for visual qa. In: Advances in Neural Information Processing Systems. pp. 361–369 (2016)
14. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling (2017)
15. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in neural information processing systems. pp. 3294–3302 (2015)
16. Li, R., Jia, J.: Visual question answering with question representation update (qru). In: Advances in Neural Information Processing Systems. pp. 4655–4663 (2016)
17. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems. pp. 289–297 (2016)

18. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A deep learning approach to visual question answering. International Journal of Computer Vision **125**(1-3), 110–135 (2017)
19. Noh, H., Han, B.: Training recurrent answering units with joint loss minimization for vqa. arXiv preprint arXiv:1606.03647 (2016)
20. Noh, H., Hongsuck Seo, P., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 30–38 (2016)
21. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014) **12** (2014)
22. Qiu, X., Huang, X.: Convolutional neural tensor network architecture for community-based question answering. In: IJCAI. pp. 1305–1311 (2015)
23. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)
24. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in neural information processing systems. pp. 2953–2961 (2015)
25. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4613–4621 (2016)
26. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: Advances in neural information processing systems. pp. 926–934 (2013)
27. Teney, D., Anderson, P., He, X., Hengel, A.v.d.: Tips and tricks for visual question answering: Learnings from the 2017 challenge. arXiv preprint arXiv:1708.02711 (2017)
28. Wu, Q., Wang, P., Shen, C., Dick, A., van den Hengel, A.: 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. pp. 4622–4630 (2016)
29. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: International Conference on Machine Learning. pp. 2397–2406 (2016)
30. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision. pp. 451–466. Springer (2016)
31. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–29 (2016)
32. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proc. IEEE Int. Conf. Comp. Vis. vol. 3 (2017)
33. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167 (2015)