Visual Question Answering Based on Question Attention Model

Jianing Zhang, Zhaochang Wu, Huajie Zhang and Yunfang Chen*

School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China

*Corresponding author email: chenyf@njupt.edu.cn

Abstract. Visual Question Answer (VQA), the natural language question of Visual images, has become popular in the field of artificial intelligence. At present, most of the VQA models extract the whole image features, which consume a large amount of computation and have a complex structure. In this paper, we propose a VQA method based on question attention model. Firstly, the Convolutional Neural Networks (CNN) is used to extract image features from the input images, and the question text is processed by the Long Short-Term Memory (LSTM). Then, we design a question attention module to let the learning algorithm focus on the most relevant features of the input text. According to question features, our method utilizes the attention module to add the corresponding weights to the image features and extract the meaningful information for the generation of answer sequence words. Our method performed significantly better than the LSTMQ+I model on the MS COCO visual question answer (VQA) dataset with an accuracy improvement of 2%.

Keywords: VQA; Attention model; CNN; LSTM.

1. Introduction

Look at the pictures and questions in Fig. 1. What is the dog doing? Where is the dog? As human beings, we can answer easily, but machines often fail to give accurate answers. With the advent of deep learning (DL), the visual question and answer system (VQA) performs this task to some extent, which can be intuitively understood as an AI system that takes pictures and questions as input, combines the two kinds of information, and produces a human language answer as output.

Figure 1. Through visual perception of the scene, human beings can identify the key points in the picture and make logical reasoning based on the questions to get the answers.

The neural image caption (NIC) model, as an early VQA model, is based on deep learning [1]. Its inspiration comes from machine translation [2], whose task is to translate the sentence written in the source language into sentences written in the target language. The NIC model is inspired by encoder-decoder model and the deep CNN is used to replace the encoder RNN. After the image classification is pre-trained, the last hidden layer is used as the input of the RNN decoder of the
generated statement. The NIC model combines the visual model CNN with the language model RNN, and forms the original VQA system.

As human beings, we visually perceive the scene, accurately identify the key points in the image, and answer the question through logical reasoning. In the example of figure 1, we can focus on the dog and answer the questions accurately. The visual attention mechanism enables humans not to answer questions based solely on the features of the image, but to choose to understand the image on the basis of semantics and find the key points.

In this paper, we add a human-like attention mechanism to the system. Firstly, we replace RNN with LSTM, then add attention module to the basic model. In the attention module, important information in the output of our semantic model will be fused into the output of image features, giving relative weights and changing attention according to the change of the question. On the network state and output node, the attention module fuses the state of LSTM with the image features, obtains the weighted image features, and finally obtains the final output through the multi-layer perceptron (MLP).

The main contribution of our work is twofold:

• We design a question attention module for VQA. The attention module can learn to focus and identify the special image area related to the question, so as to provide effective analysis for the VQA tasks.

• We propose an overall VQA framework based on question attention, including question attention module, word frequency preprocessing, and MLP fusion, which improves the accuracy of the model.

2. Related Work

In recent years, attention model has been widely used in image processing, natural language processing and other deep learning fields. The human visual mechanism has a special brain signal processing mechanism called visual attention mechanism [3,4]. This mechanism can quickly scan the image to get the global focus and pay more attention to the focus area to get the key information. This is also the goal of the attention mechanism in deep learning. The encoder-decoder framework is a general thinking of the attention model [5], but it mainly deals with text and focuses on translation. The methods of VQA fall into two categories: monolithic and modular. The monolithic approach means using a single model applied to different tasks [7, 8], mostly on VQA datasets of real-world images. The modular approach, on the other hand, is better at answering comprehensive questions with different modules [9,10] for different subtasks. [11] proposed a network tuning method based on people's perception of importance, which uses human supervision to improve visual grounding. By optimizing the alignment of human attention, the model focuses on the correct regional map. Language prior problem for the VQA means that the model did not learn to answer the question according to the image, but simply relied on the proportion of the answer. This problem has attracted a lot of research work [14,15], most of which balance biased data sets or force models to address this problem to understand the image and question better. In this paper, we propose a kind of question attention module, and let the question and the image guide each other to solve the language prior problem.

3. Methods

Our method aims to integrate text and image more accurately, and the framework is shown in Fig. 2. Firstly, the deep neural network algorithm is used to obtain the image feature representation through the pre-trained CNN. At the same time, TF-idf algorithm is used to pre-process the input question, filter out the ordinary words and special words, retain the core words, and obtain a fixed length vector representation of the question through LSTM. Then, we introduce the question attention module. The image features and the output and hidden states of LSTM are fed into the attention module to obtain the context vector. Finally, the two forms are embedded into the same space and input into the MLP to output the final result. The answer prediction is based on the participating images and question characteristics.
4. VQA Model Based on Seq2Seq

A VQA model is proposed based on Seq2Seq which is a variable length prediction model based on CNN and LSTM [6]. In this model, the visual question and its answer are regarded as a sequence to sequence task with image information as an auxiliary. Firstly, the image features are extracted by a pre-trained deep CNN model, and then the image feature vectors and the question words vectors are sent into LSTM network together. The answer is then generated using the same LSTM network until the terminator ($) is generated. The training process of the model includes the training of LSTM network combining image features and the training of word vector generator.

Given image $x$ and question $q$:

$$
a = \arg \max_{a \in A} p(a|x, q; \theta).$$  (1)

Here, $\theta$ refers to the parameters of the model. $a$ is an answer, and $A$ is a set of candidate answers. $a$ is the best answer to the question $q$ with image $x$.

In the task scenario, the question produces more than one-word answer, so the question is represented as a set of predicted answers: $a_{q,x} = \{a_1, a_2, ..., a_{N(q,x)}\}$, $a_t$ from a limited vocabulary $V'$, $V' = V' \cup \{\$\}$, where $\$ $ represents the end of the sentence. $N(q,x)$ represents the number of words in the answer to a given question $q$ and image $x$.

The recursive prediction process is:

$$\hat{a}_t = \arg \max_{a \in A} p(a|x, q, \hat{A}_{t-1}; \theta).$$  (2)

$\hat{a}_t$ represents the current word to be generated. $\hat{A}_{t-1} = \{\hat{a}_1, ..., \hat{a}_{t-1}\}$ represents a collection of previously generated words. The recursive neural network and the softmax prediction layer are used for parameter distribution modeling $p(\cdot|x, q; \theta)$.

5. The Attention Model Based on NMT

Neural machine translation (NMT) model [12] is encoder-decoder architecture with attention mechanism, as Fig. 3(a) shown. Unlike general encoder-decoder methods, the probability of each target word $y_t$ here is related with the context $c_t$, and $c_t$ depends on the annotation $(h_1, ..., h_{T_x})$. $h_t$ contains information about the entire input sequence and $T_x$ is the number of hidden states. So $c_t$ is represented as the weighted sum of $h_t$:

$$c_t = \sum_{j=1}^{T_x} a_{ij} h_j.$$  (3)

Weight $a_{ij}$ is expressed as:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}, e_{ij} = d(s_{i-1}, h_j).$$  (4)
$e_{ij}$ is used to measure the degree of matching between the $j$th input word and the $i$th output word. The degree is calculated based on the decoder's hidden state $s_{i-1}$ and the $j$th annotation $h_j$ of the input sentence. We changed the attention model to make it applicable to the VQA model, as Fig. 3 (b) shown.

In NMI model, the bidirectional LSTM model is used. Because in the translation task, each word annotation prediction requires not only the preceding word, but also the following word. However, bidirectional method does not be adopted in our attention module since image features in VQA are different from text features in translation tasks. So the input is the image features processed by CNN combined with the vector representation of LSTM processed text. The details will be introduced in the next section.

Figure 3. (a) The NMI model, (b) The variant NMI applied in our VQA model.

6. VQA Model Based on Question Attention

We expect that the extracted image features can rely on the language model. The adaptive question attention mechanism expands the spatial attention. Our model with attention module is shown in Fig. 4.

Figure 4. VQA model based on question attention

$x_t$ is the question word with time $t$ as input, and the output vector $k_t$ is obtained through LSTM:

$$k_t = \sigma(W_x x_t + W_h h_{t-1}).$$

Where $W_x$ and $W_h$ are weight parameters, which need to be obtained thought deep learning. $h_t$ and $h_{t-1}$ are respectively the hidden states of LSTM at time $t$ and time $t-1$. The attention module without adding question features only depends on $V$ and $h_t$, as shown in Fig. 5(a). $V = \{v_1, v_2, ..., v_T\}$ represents image features which obtained by CNN. Then the single-layer neural network and softmax function are used to generate the distribution of attention:
(a) The attention module without adding the question feature, (b) The attention module with adding the question feature:

$$m_t = \omega_t^T \tanh (W_t^h V + (W_t^h h_t)^T), \quad \alpha_t = \text{softmax}(m_t).$$  

(6)

I is a vector with all elements set to 1. $W_v$ and $W_g$ are weight parameters which need to be learned. $\alpha_t$ is the weight of the attention of image $V$. At this time, the context vector $c_t$ without adding question features can be expressed as:

$$c_t = \sum_{i=1}^{k} \alpha_i v_i.$$  

(7)

In order to introduce the question feature to guide the image feature, as shown in Fig. 5(b), we choose a hybrid model to calculate the new context vector $\hat{c}_t$. The hybrid model is defined as:

$$\hat{c}_t = \gamma_t k_t + (1 - \gamma_t) c_t.$$  

(8)

$\gamma_t$ is the parameter at time $t$ to trade off the importance of the image and the question. The range of $\gamma_t$ is [0,1], and 0 means that the question does not provide any critical information while 1 means that the answer can be completely based on the question. Combined with hidden state $h_t$ and new context vector $\hat{c}_t$, the answer can be predicted:

$$f_t = \text{softmax}(W_f(\hat{c}_t + h_t)).$$  

(9)

$W_f$ is the weight parameter which needs to be learned.

7. Experiment

7.1. Datasets

We evaluated the proposed model on the MS COCO visual question answer dataset [13]. The dataset contains manually annotated questions and answers, including 248,349 training questions, 121,512 validation questions, and 244,302 test questions, for a total of 6,141,630 question-answer pairs. The feature of MS COCO dataset is object segmentation, which can be recognized in the context. There are three types of answers: yes/no, Numbers, and others. We used the evaluation criteria in the experiment.

8. Results and Comparison

We compared our method with the question+image-based model [13] in VQA, and the results are in Table 1 and Table 2. For open-ended questions our model achieved a 3% improvement. For the multiple-choice question, it improved by 2%. Because of the prominence of the question attention mechanism, our model is significantly better at "other" types of questions.
Table 1. Test results for open-ended questions answered on the VQA dataset

| model         | Y/N | Num  | Others |
|---------------|-----|------|--------|
| LSTMQ+I[13]  | 79.5| 39.6 | 43.0   |
| Ours          | 78.6| 40.2 | 48.3   |

Table 2. Test results for multiple-Choice questions on VQA dataset

| model         | Y/N | Num  | Others |
|---------------|-----|------|--------|
| LSTMQ+I[13]  | 79.5| 38.3 | 53.0   |
| Ours          | 78.6| 39.4 | 56.3   |

Our method gives more weight to the person or animal in the picture and focuses visual attention on the action. So, our method can answer such questions more accurately, but not to the Y/N and Num questions.

9. Summary

In this paper, we take question attention module from the traditional machine translation model and apply it to the VQA. The question attention module focuses on the appropriate areas of the image according to the words in the question. The model establishes a natural connection between the image and the question, and it has potential application capability in various areas involving computer vision and natural language processing.

Acknowledgement

This research was financially supported by Science and Technology Innovation Training Program (STITP) of Nanjing University of Posts and Telecommunications 2019.

References

[1] O. Vinyals, A. Toshev, S. Bengio, Show and tell: a neural image caption generator. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015, pp. 3156-3164.
[2] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 3rd International Conference on Learning Representations (ICLR) 2015.
[3] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, In Matters of intelligence, 10.1007/978-94-009-3833-5 (Chapter 5):115-141.
[4] M. W. Spratling, M. H. Johnson, A feedback model of visual attention, Journal of cognitive neuroscience, 16 (2004) 219-237.
[5] K. Xu, J. Ba, R. Kiros, Show, attend and tell: neural image caption generation with visual attention, 32nd International Conference on Machine Learning (ICML). 2015, pp. 2048-2057.
[6] M. Malinowski, Ask your neurons: a neural-based approach to answering questions about images, 2015 International Conference on Computer Vision (ICCV). 2015, pp. 1-9.
[7] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2018, pp. 6077-6086.
[8] D. A. Hudson, C. D. Manning, Compositional attention networks for machine reasoning, 6th International Conference on Learning Representations (ICLR). 2018.
[9] R. Hu, J. Andreas, T. Darrell, K. Saenko, Explainable neural computation via stack neural module networks, In Proceedings of the European Conference on Computer Vision (ECCV). 2018, pp. 53-69.
[10] J. Shi, H. Zhang, J. Li, Explainable and explicit visual reasoning over scene graphs, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019, pp. 8368-8376.
[11] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, D. Batra, D. Parikh, Taking a hint: Leveraging explanations to make vision and language models more grounded, IEEE International Conference on Computer Vision (ICCV), 2019. arXiv:1902.03751v2
[12] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and
translate, International Conference on Learning Representations (ICLR). 2015

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: visual question answering, International Conference on Computer Vision (ICCV). 2015, pp. 2425-2433.

[14] S. Ramakrishnan, A. Agrawal, S. Lee, Overcoming language priors in visual question answering with adversarial regularization, Advances in Neural Information Processing Systems (NeurIPS). 2018, pp. 1541-1551.

[15] A. Agrawal, D. Batra, D. Parikh, A. Kembhavi, Don't just assume; look and answer: overcoming priors for visual question answering, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 4971-4980.