VQA Model Based on Image Descriptive Paragraph and Deep Integration of BERT

Zhaochang Wu, Jianing Zhang, 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 Answering (VQA) is a fast developing field involving multiple disciplines, and it is constantly challenging more complex tasks. The classic combination of CNN+LSTM can effectively extract images and language representation to complete the VQA task, but there are still many problems, such as excessively long sequence processing, etc. In recent years, BERT model has expanded rapidly from the field of natural language processing to a broader multi-modal field with its strong learning ability. In this paper, we propose a novel way to apply BERT model in the VQA field. We use the descriptive paragraph generation technology to transform the picture into a text paragraph description, and integrate question information and image information on BERT model. Our model achieves an excellent performance on the VQA2.0 dataset with an overall accuracy 5% higher than previous models.

Keywords: VQA; BERT; CNN; LSTM.

1. Introduction
Along with the progress of natural language processing and image understanding, more complex and more demanding tasks can already be done. In the Stanford reading comprehension Dataset (SquAD2.0), machines have outperformed human performance. The most basic VQA is the joint application of CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory). At present, in the field of natural language processing, transformer's BERT (Bidirectional Encoder Representation from Transformers) pretraining model has better performance than LSTM.

The BERT (Bidirectional Encoder Representation from Transformers) model derived from the Transformer model is a general pre-trained language representation model first proposed by Google in 2017 [3]. The full name of BERT is. BERT greatly improved the benchmark performance of each NLP task by virtue of the strong feature learning ability of Transformer and Two-Encoding realized by the mask language model. The performance of BERT model on 11 NLP tasks broke the record [4], including Question Answering (SQuAD v1.1), reasoning Natural Language Inference (MNLI), etc. In addition to handling problem text information, the VQA also needs to handle information about the image. The Image caption technique first appeared in [1][2] to extract Image features by using some operators of Image processing. The emergence of dense captioning technology [14] further makes it more likely to develop the Image caption.

We propose a method to integrate BERT model into the visual response field of VQA. Pre-trained language model BERT is added by fine-tuning to enhance the processing of language. With image processing, the visual question-and-answer task could be better completed and higher accuracy could be achieved. In our model, we innovatively combine the descriptive text generation technology with BERT model. Instead of inputting the image feature extraction into BERT for the model to learn, we
applied the phased description hierarchy method of the generated image into a reading comprehension problem. The contribution of the article is two-fold: 
1. The hierarchical method of phased description of images replaces the input of images to obtain more comprehensive and fine-grained image information. Text-text fusion is faster and more consistent than image-text fusion in the multiple models fusion of VQA. 
2. The application of BERT model in VQA achieves faster speed and higher accuracy than LSTM + CNN model.

2. Related Work

2.1. Visual Question Answering
Visual Question Answering (VQA) is a learning task involving computer vision and natural language processing. Its main goal is to make the computer output a reasonable answer that conforms to the rules of natural language according to the input pictures and questions. Great progress has been made in building VQA models using LSTMs [15] and convolutional networks [16].

2.2. Preprocessing for Natural Language Preprocessing
What is interesting about NLP field is that its pre-training technology lags far behind that of computer vision. There is some work beginning to improve word embedding, but it is only a low-level representation of the feature. Most importantly, different NLP tasks design different architectures. In the landmark work of Transformer [3], it is recommended to use the attention module in Transformer as a general module of NLP. Then many methods mainly used for training general representation were proposed such as GPT [7], BERT [4] and XLM [8], etc.

2.3. Image Caption and Intensive Captioning Techniques
The early [1,2] approaches of image caption relied heavily on image feature extraction and sentence generation rules. But the effect was not very satisfactory. Inspired by RNN in machine translation, Neural Talk [9] and Show and Tell [10] emerged that used similar techniques to extract image features to implement text generation and description. In recent years, a hierarchical method of generating descriptive paragraphs of images [11], which overcomes the feature that dense captioning technology cannot produce coherent stories, can generate paragraph-based picture descriptions.

2.4. Application of BERT in the Field of VQA
The current application of BERT to VQA is divided into single-stream and dual-stream models. VisualBERT [5] and VLBERT [6] are all single-stream models. In single-stream models, text information and visual information are fused at the beginning. And ViLBERT [12] is one of the most representative of the dual-stream model. However, it is unclear whether single-stream is better or double-stream is better. It seems that the result is related to specific tasks and more rigorous comparison experiments are needed for further verification.

3. Methods
Our basic idea is to use text alignment instead of image-text alignment, and the overview of our method is shown in Figure 1. The detailed method is based on the hierarchical method of image descriptive paragraphs, which converts images into descriptive text paragraphs of images, and then puts them into BERT’s text reading model as paragraphs, designing fine-tuning to handle reading comprehension and question answering.
We introduce the hierarchical method of generating image descriptive paragraphs in 3.1, describe the BERT model in 3.2, and propose the fine-tuning end of the fusion of the BERT model based on the hierarchical method of image descriptive paragraphs in 3.3.

3.1. Generating Descriptive Image Paragraphs—Dense Captioning

The paper [11] proposed a hierarchical method model for generating descriptive paragraphs of images. Taking images as input, natural language paragraphs describing it are generated, and the structure is designed to make use of hierarchical text of images and paragraphs.

This method can effectively transfer the region knowledge of the picture to the paragraph describing. It also has high interpretability. Descriptive paragraphs can be generated by only using a subset of the image. Because it can restore the original image information well, our model uses this method.  

3.2. BERT Model

3.2.1. BERT's Working Principle:

BERT uses the encoder part of Transformer [3]. Transformer is a model that completely relies on the attention mechanism and can learn the contextual relationship between words in the text and discards the RNN cycle architecture. Transformer's prototype includes two independent mechanisms, an encoder is responsible for receiving text as input, and a decoder is responsible for predicting the results of the task. The goal of BERT is to generate language models, so only the encoder mechanism is needed.

Figure 3 shows the overall structure of BERT. The input is a token sequence. It is first embedding it as a vector, and then input to the neural network. The output is a vector sequence of size H.
When training a language model, it is difficult to define a prediction target. To overcome this challenge, BERT uses two training strategies, Masked LM and Next Sentence Prediction.

3.2.2. Masked LM (MLM) & Next Sentence Prediction (NSP)
In order to achieve deep bidirectional representation, MLM strategy is adopted. We randomly block some input tokens and then predict them. In order not to damage the language comprehension of the model, only 15% of the words in the sequence are replaced.

NSP considers the problem from the perspective of sentences, and pre-trains a binarization next sentence prediction task, which can be easily generated from any monolingual corpus. In the training process, 50% of the input is contextual in the original document. While in the other 50% of the input, it is randomly composed from the corpus and is disconnected from the first sentence. This tactics solves the problem of how to obtain dependencies on sentences required in many NLP tasks. The mission achieved accuracy of 97-98%.

3.2.3. Model Training
Two models of different sizes are proposed in [4]:
1. BERT-Base: L = 12, H = 768, A = 12, Total parameters = 110M
2. BERT-Large: L = 24, H = 1024, A = 16, Total parameters = 340M
Where L represents the number of Transformer layers, H represents the internal dimensions of Transformer, and A represents the number of Heads.

3.3. Modification of Fine-tuning End (Fusion of Hierarchical Method of Descriptive Paragraph of Image and BERT Model)
As shown in Figure 4, the input sequence of the model is the embedding corresponding to the sentence pairs. Sentence pairs consist of questions and text containing answers, and are separated by a special separator "[SEP]". As with other downstream tasks, the first token of the input sequence is a special classification embedded in "[CLS]", and the input sequence is the sum of token embedding, segmentation embedding, and position embedding.
Figure 4. Fine-tuning principle structure diagram of BERT. The output of BERT is the encoding vector corresponding to each token. Because the answer consists of consecutive tokens in the text, the process of predicting the answer is essentially the process of determining the location of the beginning and ending tokens of the answer.

Figure 5. Trains two separate auxiliary vectors and takes the inner product of each vector output to determine the positions of start and end markers according to the degree of correlation. While entering the descriptive paragraph of the image as the paragraph end, VQA system needs to receive a question about the text sequence, and mark the answer in the sequence according to the position of the Start and end tags to obtain the final answer.

4. Experiment

4.1. Dataset
In a VQA task, given a natural image, the question is asked at the perception level and the correct answer is generated or selected according to the algorithm. We have done the experiments with the most widely used VQA2.0 [13] dataset. The VQA v2.0 dataset is split into train (82,783 images and 443,757 questions), validation (40,504 images and 214,354 questions), and test (81,434 images and 447,793 questions) sets. In VQA2.0, each image has an average of 5 questions and each question has 10 pre-selected answers. There is a one-to-one correspondence between images, questions and answers.

4.2. Comparison of Results
In order to verify the effectiveness of the method, several relevant models on the official VQA website [17] were selected and compared with the model in this paper on VQA2.0. The results are shown in table 1:
Table 1. BERT model compared with test results of relevant models on VQA website.

| Model          | Yes/No | Number | Other  | overall |
|----------------|--------|--------|--------|---------|
| Prior          | 61.26  | 0.40   | 1.26   | 26.13   |
| Language-only  | 67.88  | 30.59  | 29.13  | 46.21   |
| d-LSTM + n-I   | 72.23  | 36.20  | 39.68  | 57.27   |
| Our-1          | 79.89  | 38.51  | 56.23  | 61.78   |
| Our-2          | 84.24  | 40.48  | 58.95  | 64.12   |

Prior means to answer questions on the test set with the most common answers on the training set. Language-only which use a single LSTM architecture just use question to answer to forecast. d-LSTM+n-I is a basic visual question and answer model, Our-1 represents the improved model based on BERT\textsubscript{BASE}. Our-2 represents the improved model based on BERT\textsubscript{LARGE}.

In Table. 1, "yes/no", "number" and "other" correspond to the accuracy rate of the model's prediction under three different answers types respectively, and "overall" is the overall performance on the corresponding data set. As can be seen from Table 1, for questions with the answer of yes/no, the accuracy of the model in our-1 is about 80%. In the question with the answer of number type, the accuracy of our-1 is about 40%. In other type of questions, the accuracy of our-1 is about 57%. The model of our-2, which has a larger amount of pre-training, has been improved more significantly. In general, the accuracy of our models is about 5-10% higher than the accuracy of the model mentioned above. Since the hierarchical method is more able to restore the existence and action of objects in the picture, the accuracy of yes/no type and other type is greatly improved, while that of number type is not significantly improved.

It can be seen that the idea of applying BERT to VQA field proposed in this paper is effective, and the improvement on accuracy is significant compared with previous models.

5. Conclusion

In this paper, we successfully applied BERT model which has an excellent performance in NLP field to VQA field based on the hierarchical method of descriptive paragraphs of images. We have demonstrated that the model performs well in VQA tasks. Compared with the traditional VQA model, the fusion model of BERT and VQA obtains faster speed and higher accuracy. It also has considerable applications in fields related to Natural- Language Question Answering and Computer vision. However, due to the limitation that the image description cannot fully restore the image, how to generate a more problem-oriented text description containing the image context is our further research direction.

Acknowledgements

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

References

[1] Kulkarni G and Premraj V, et al 2013 Babytalk: understanding and generating simple image descriptions IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12) pp 2891-2903.
[2] Farhadi A and Hejrati S M M, et al 2010 Every Picture Tells a Story: Generating Sentences from Images Computer Vision – ECCV 2010. Springer Berlin Heidelberg pp 15-29.
[3] Vaswani A and Shazeer N, et al 2017 Attention Is All You Need. Advances in Neural Information Processing Systems 30 (NIPS 2017)
[4] Devlin J and Chang M W, et al 2018 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv:1810.04805
[5] Li L H and Yatskar M, et al 2019 VisualBERT: A Simple and Performant Baseline for Vision and Language arXiv:1908.03557
[6] Su W and Zhu X, et al 2019 VL-BERT: Pre-training of Generic Visual-Linguistic Representations arXiv:1908.08530
[7] Radford A and Narasimhan K, et al 2018 Improving language understanding by generative pre-training URL https://blog.openai.com/language-unsupervised/.

[8] Lample G and Conneau A 2019 Cross-lingual Language Model Pretraining Advances in Neural Information Processing Systems 32 (NIPS 2019)

[9] Karpathy A and Fei-Fei L 2016 Deep Visual-Semantic Alignments for Generating Image Descriptions The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 3128-3137.

[10] Vinyals O and Toshev A, et al 2017 Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4) pp 652-663

[11] Krause J and Fei-Fei L, et al 2017 A Hierarchical Approach for Generating Descriptive Image Paragraphs The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 317-325

[12] Lu J and Batra D, et al 2019 ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks Advances in Neural Information Processing Systems 32 (NIPS 2019)

[13] Goyal Y and Khot T, et al 2017 Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6904-6913

[14] Johnson J and Karpathy A, et al 2016 DenseCap: Fully Convolutional Localization Networks for Dense Captioning The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 4565-4574

[15] Hochreiter S and Schmidhuber J 1997 Long Short-Term Memory Neural computation 9(8) pp 1735–1780

[16] LeCun Y and Bottou L, et al 1998 Gradient-based learning applied to document recognition Proceedings of the IEEE, 86(11) pp 2278–2324

[17] https://visualqa.org/.