Improve Visual Question Answering Based On Text Feature Extraction

Jing Wang1*, Yang Dong1

1 School of Computer Science and Engineering, Northeastern University, Shenyang, 110000, China
*Corresponding author’s e-mail: 1971852@stu.neu.edu.cn

Abstract. VQA requires the machine to be able to answer questions related to the image based on the image. VQA is mainly divided into four modules: text extraction module, image extraction module, text-image feature fusion module and answer prediction module. In this paper, the most mainstream VQA2.0[1] data set is selected for experiment, and the text extraction module is improved on the basis of a baseline model. The Glove vector is used to preprocess the problem text data, and the GRU network replaces the traditional LSTM network encoding and processing text vector. Experimental results show that the improvement of the text extraction module improves by 2.42% on the basis of the original model, and the GRU network can also accelerate the training speed.

1. Introduction

Visual Question Answering (VQA)[2] takes an image and related questions as input and requires the machine to give a correct answer. VQA has been substantially improved in recent years, but the computer's understanding of text is far from the level of human beings. When we look at problems, sometimes we need to understand the semantic information in the problems, so the VQA task has some limitations in the processing of text features. In the VQA task, the input is the image and natural language questions, and the output is the natural language answers. Because most of the questions have explicit answers, VQA is relatively easy to evaluate, which is one of the important reasons why it gets a lot of attention.

After extracting the question text feature vector and image feature vector, it is necessary to fuse them. After injecting the image information into the question, the machine can filter the information according to the image, so as to improve the performance of question answering. We proposed an improved method in the text extraction module, using the Global Vectors for Word Representation (Glove) [3] word vector to process text data, and using the Gated Recurrent Unit (GRU)[4] network structure instead of the Long Short-Term Memory network (LSTM)[5] network to code the problem text.

2. Related Work

In traditional Natural Language Processing (NLP)[6], we often use one-hot coding. Alone the heat code can simply and accurately said a word, but a word requires a one dimensional vector, its take up the space is too big, and the sole hot code has a fundamental flaw is that can't really express the semantic coding way, because the inner product of any two word similarity to 0, it cannot represent the relevance of words. Word2Vec[7] is very effective, it caused the research upsurge of words in terms of NLP embedded model, however, because of its locality, it ignores the word contact outside
their window word, also ignore the word order information, Stanford university Glove model put forward by the use of word co-occurrence matrix at the same time, taking into account the global and local word information, its essence is a matrix, each row represents a word, each column represents about the words of a certain characteristic value, can more accurately reflect the actual relationship between word and word. In this experiment, the Glove word vector will be used to process text data.

After the text is encoded as the feature vector, it only gets the vectorized representation of the word, and the sentence information needs to be further encoded to understand the sentence. Loop neural network Recurrent Neural Network (RNN) has a certain memory, can in theory capture all historical information. However, since every time step in the RNN structure needs to be activated by Sigmoid, and this function is in the interval from 0 to 1, it is easy to cause the problem of gradient disappearance or explosion. Therefore, an improved RNN was invented: LSTM, which can effectively alleviate the gradient disappearance problem and solve the long-distance dependence problem to some extent. Another classic variant of RNN is the GRU network, where the forgetting gate and the input gate in the LSTM are combined into an Update Gate, as well as the hidden layer state and memory state of the cell. But in fact, the reason why LSTM and GRU are effective lies in the addition of information containment mechanism, which uses Gate to filter valid information. GRU model is less complex than LSTM model, but it often has the same effect as LSTM, so it is often more favored. In this experiment, GRU will also be used as the baseline model.

3. Model

3.1. Text feature extraction module

Text feature extraction is an important step, because now that is a question and answer, you must let the machine can understand the problem, but the text stored in the form of characters. A simple text logo is difficult to express the semantic, so we need word embedding technology to convert text to have the value of semantic information vector, the following will detail the problems in laboratory model text processing.

We encode the problem text into individual terms. The one-hot representation of the text is the count-Based Representation, whose vector values range from all non-negative integers, with each value representing the number of times the word in the corresponding position appears in the sentence. This method solves the problem that classifiers cannot handle attribute data well, and requires each category to be independent of each other. If there is some continuous relation between each category, this method has some defects and may encounter the disaster of too many features. This article uses the Glove word vectors, in this way the characteristics of the said every word number is not increased with the increase of words set, can be limited to a certain size, at the same time can be judged by calculating the vector relationship, the relationship between the word and the word in understanding the text semantic get some meaningful results[8].

First of all, we symbolized the input of the problem. The input received by our model is English question , \( Q = \{ w_1, ..., w_N \} \) where \( w_i \) is each word and \( N \) is the length of the question. After Glove word vectors of training, the question of encoded text to \( Q^o = \{ w_1^o, ..., w_N^o \} \in \mathbb{R}^{300 \times N} \), and then it is transformed into the word vector space through the word embedding layer, as shown in Formula 1.

\[
Q^e = W_e Q^o
\]  
(1)

The i-th column of \( w_i \) is the i-th word vector in the Glove word table. In this paper, the 300-dimensional Glove word vector is adopted and the word vector of the filler word is set as all 0 vector. The pre-trained vector can achieve better effect than the randomly initialized one.

Preprocessing of text data encodes text into feature vectors, and further encoding of sentence information is needed to understand the semantics of sentences. When dealing with the current node vocabulary, the information of previous node vocabulary is expected to be received, and the sentence vocabulary sequence relation is captured. LSTM can effectively alleviate the problem of gradient disappearance and solve the problem of long distance dependence to some extent. GRU network is improved on the basis of LSTM network. The model is less complex than LSTM model and the effect
is equivalent to LSTM network. Experimental data show that LSTM and GRU have similar performance in VQA tasks, and the latter can complete modeling with less memory and faster speed. Therefore, GRU is used as the baseline model in this experiment to extract problem features and obtain the final feature vector $V_q$ of the problem text, as shown in Formula 2.

$$V_q = \text{GRU}(Q)$$  \hspace{1cm} (2)

3.2. Image feature extraction

Image data is stored directly in pixels, each pixel has three channels directly represent the image semantics, and spatial semantics can also be directly reflected by the pixel matrix. Image feature extraction module takes pixel image $I \in R^{W \times H \times 3}$ as input and feature vector $V_t$ containing high-level semantic information as output. $V_t$ is the feature vector obtained by image $I$ through ResNet forward propagation coding, which can standard the information of the original image. Since Softmax is ultimately required to obtain the probability density of each answer in the candidate set, the dimensions will be mapped to the dimensions of the candidate set by the Fully Connected Layer after ResNet[9]. As shown in Formula 3.

$$V_t = \text{ResNet}_{\text{conv}}(I)$$  \hspace{1cm} (3)

3.3. Feature fusion and answer prediction

After the completion of image and question text feature encoding, it is necessary to fuse the two, build a bridge between image and text semantics, and enable the machine to find the answer to the question based on the image information[10]. The image and text are spliced and fused, $F_c$ is the full connection layer of the feature vector which is used to integrate the splicing[11]. The multi-modal characteristic process of fusion is shown in Formula 4.

$$\alpha_{c,t} \propto F_c(V_q, V_t)$$  \hspace{1cm} (4)

$$x_c = \sum_{t} \alpha_{c,t} V_t$$  \hspace{1cm} (5)

Attention weighting $\alpha_{c,t}$ in the fusion image and text function each weight, $c=1,2,\ldots,C$. Image feature $X_t$ is all spatial positions of the image $t = \{1,\ldots,T\}$, in fact $F = [F_1,F_2,\ldots,F_C]$ is modeled with two-layer convolution. Thus, $F_t$ shares parameters in the first layer. We simply rely on different initializations to produce different distributions of attention.

Image $I$ is a real world image in pixel form, while Question $Q$ is text in natural language form. The task of VQA is to get answer $A$ as correct as possible from the machine. We use the neural network algorithm to the conditional probability $P (A \mid I, Q; \theta)$, and obtain the optimal parameter through training, the correct answer when $\theta$ is maximum, as shown in Formula 6.

$$\Theta^* = \arg \max_{\theta} \sum_{l,q,a} P(A \mid I, Q; \theta)$$  \hspace{1cm} (6)

where there are three addresses, you should insert numbered superscripts 1, 2 and 3 to link surnames to addresses and then insert footnotes 4 and 5. Note that the first footnote in the main text will now be number 6.

3.4. Model and related parameters

The model designed in this paper is shown in Figure 1. In the text data module, data preprocessing is required. We unified the sentence length of the question as 14. Then, the Glove word vector of 300 dimensions is used for marking, and the generated sentence feature vector is embedded into the GRU network, and the GRU state size is set to 1024.

In the image extraction module, the preprocessed image information is passed through the ResNet network. At this time, the dimension of the feature tensor $V$ of an image is $2048*14*14$. Where, $14*14$ corresponds to the Spatial dimension of the image, and each 2048 dimension vector corresponds to the local features in the receptive field of the region, and the image features are fully connected. Attention in order to calculate weight of the image characteristics, we will deal with the text of the GRU helped network characteristic vector and ResNet processing of image feature vectors, and by a depth of 512
layer 1 x 1 d convolution, followed by ReLU[12] nonlinear layer, output characteristics and through the depth of another 1 of 2 x 1 convolution.

Concentration distribution by softmax function calculation, we use the distribution by calculating a weighted average of the image characteristics to calculate a glimpse of two images, image feature weighting matrix. The weight matrix is connected to the text feature vector through a fully connected layer of size 1024 with ReLU nonlinearity. The output is fed to the linear layer of size M=3000, followed by softmax, to produce the probability of the most frequent category.

Figure 1. An overview of our model

4. Experiment

4.1. Experimental data sets
In this experiment, VQA 2.0 data set was used, and the images used were from MS-COCO data set. There were 82,783 images in the training set and 40,504 images in the validation set. In this experiment, I will answer the candidate set to training set answers to a number greater than or equal to nine times, a total of 3192, discarded are not the answer to this question and answer for the sample of a set of the answers.

4.2. Evaluation indicators
Because of the VQA data set, each question is accompanied by 10 human tellers' answers, and because of the tellers' misinterpretations of the question and the answers given, everyone may have different edits to the same question. Although the most frequent answers given by the markers to this question were used as the true labels during the training, the creators of VQA data set designed an evaluation scheme based on voting mechanism when evaluating scores, and the calculation formula of accuracy is shown in Formula 7.

\[
\text{accuracy}(a) = \frac{1}{10} \sum_{k=1}^{10} \min\left(\frac{\sum_{i=1}^{10} \delta(k,a_i)}{3}, 1\right)
\]  \hspace{1cm} (7)

We believe that as long as the answer predicted by the model is more than or equal to the labeled answer of 3, it is considered to be completely correct, otherwise, the corresponding decimal score will be obtained.
4.3. Analysis of experimental results

Table 1 shows the comparison of the results of different experimental models on the VQA 2.0 validation set. Our model is improved in a text feature extraction module. Glove-LSTM model dealing with the text of the text data will be the relationship between the words, in the understanding of computer text semantic information to improve its accuracy, on the validation set won 62.09% of the improved accuracy compared to the original model. In the VQA task, the performance of LSTM is almost the same as that of GRU, and the latter can complete the modeling with less memory and faster speed. GRU is obviously better than LSTM in the time of code training.

| Methods                        | Y/N  | Num  | Other | All   |
|--------------------------------|------|------|-------|-------|
| HieCoAtt [13]                  | 71.80| 36.53| 46.25 | 54.57 |
| MCB [14]                       | 77.37| 36.66| 51.23 | 59.14 |
| Stacked(Embed-LSTM) [15]       | 77.45| 38.46| 51.76 | 59.67 |
| Stacked (Glove-LSTM)           | 79.15| 41.46| 53.37 | 62.09 |
| Stacked (Glove-GRU)            | 79.23| 41.58| 53.50 | 62.17 |

5. Conclusion

In this paper, an improved visual question-answering model based on text modules is proposed. By using Glove vector to encode text information, the accuracy of text information extraction is improved. GRU network and LSTM network have almost the same difference in the experimental results, but it greatly improves the experimental training time and the performance of the model. The improvement of text feature extraction module is also applicable to text encoding in NLP field and has practical application value.

Acknowledgments

I sincerely thank my teachers and classmates for taking time out of their busy study and life to help me with my study. I am very grateful for the conditions and spiritual support given by the laboratory. They gave me help and encouragement when I met with difficulties, so that I could solve the difficulties as soon as possible, thank them very much.

References

[1] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2016) Making the v in vqa matter: Elevating the role of image understanding in visual question answering. CoRR, abs/1612.00837, 1, 2, 3, 7
[2] Antol, S., Agrawal, A., Lu, J., Et Al. (2015) VQA: Visual Question Answering . Proceedings of The IEEE International Conference on Computer Vision. pp: 2425-2433.
[3] Jeffrey, P., Richard, S., and Christopher M. (2014) Glove: Global vectors for word representation. In The Conference on Empirical Methods in Natural Language Processing, pp 1532–1543.
[4] Chung, J., Gulcehre, C., Cho, K., and Bengio Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555.
[5] Hochreiter, S., and Schmidhuber, J. (1997) Long short-term memory. Neural Computation. 9:1735–1780, 1997. 2
[6] Bahdanau, D., Cho, K., and Bengio Y. (2015) Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations, 2015.
[7] Mikolov, T., Chen, K., Corrado, G., Et Al. (2013) Efficient Estimation of Word Representations in Vector Space. ICLR (Workshop Poster) 2013.
[8] Nam H, Ha J W, Kim J. (2017) Dual Attention Networks for Multimodal Reasoning and Matching. Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. 2017: 299-307.
[9] He, K., Zhang, X., Ren, S., Sun. J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778.

[10] Xu, H., Saenko K. (2016) Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering[C]//European Conference on Computer Vision. Springer, Cham, pp: 451-466.

[11] Yang, Z., He, X., Gao, J., Et Al. (2016) Stacked Attention Networks for Image Question Answering[C]//Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. pp: 21-29.

[12] Nair, V., Hinton, G. E. (2010) Rectified linear units improve restricted boltzmann machines. In ICML. pp: 2- 4.

[13] Lu, J., Yang, J., Batra, D., and Parikh, D. (2016) Hierarchical question-image co-attention for visual question answering. In Advances In Neural Information Processing Systems, pp289–297.

[14] Fukui, A., Park, D. H.,Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M. (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In EMNLP. pp: 1-7.

[15] Kazemi, K., Elqursh, A. (2017) Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. CoRR abs/1704.03162. pp:1-6.