Research On Visual Question Answering Based On Deep Stacked Attention Network

Zhu Xiaoqing¹ and Han Junjun²

¹ Shanghai University, Shangda Road 200444, China
² Shanghai University, Shangda Road 200444, China
E-mail: laoyao@shu.edu.cn

Abstract. Aiming at the problem that the existing visual question answering model has a language bias in high-level logical reasoning, the model describes images or answers questions with low accuracy. A visual question answering model based on deep stacked attention networks is proposed. At the same time, the attention mechanism is introduced to help the model fully mine the image information; secondly, the problem attention mechanism is introduced to pay attention to the image and the question information at the same time when the problem feature is extracted; finally, the features of the image and the question are integrated and the answer is selected in the classifier. Taking the Visual Genome and VQA2.0 data sets as examples for empirical analysis, the results show that the accuracy of predicting answers is 1.13% higher than the existing best model, which proves the effectiveness and applicability of the model.

1. Introduction
With the rapid development of computer technology and deep learning, network information has become more and more comprehensive. Faced with massive amounts of data, how to filter useful information has become an important task for the development of the Internet. With the success of deep learning in the field of computer vision and natural language processing, visual understanding and text understanding are no longer the core of limiting model performance, so multimodal learning tasks that combine visual information and text information have become a direction worth exploring. Visual Question Answering[1-2] is a new task that has appeared in recent years. The system can answer questions raised by users with reference to the input picture content, which makes inferences based on visual information and text content to make predictions. In the visual question answering system, computer vision technology is used to understand images, and natural language processing technology is used to understand questions. The two must be combined effectively to answer the questions in the image context [3].

Different from the traditional text-based question answering system, the visual question answering model needs to combine the picture information as the context environment, according to the question asked by the questioner, through the fusion of the graphic cross-modal data, and then use the classifier to select the candidate answer set. At present, the research of visual question answering has certain challenges. The reasons are: (1) Although the continuous development and progress of computer vision and natural language processing technology has promoted the continuous improvement of the accuracy of visual question answering, the accuracy of prediction is relatively low and the distance is high. There is still a long way to go for an intelligent question answering system; (2) The semantic...
features of the question cannot be perfectly integrated with the image features, resulting in the emergence of high-level logical reasoning, the model often fails to give correct predictions; (3) Faced with user open-ended logical reasoning problems, the current methods are too singular.

In response to the above-mentioned problems, this paper proposes a visual question answering model based on a deep stacking attention network. The model uses question information multiple times to apply an attention mechanism to image information, enhances the representation of image content, and fully explores the relationship between image and question information. To obtain multiple image information regions related to the problem information, and set different weight values, input the fused feature information into the classifier to filter, and predict the model in this article in the existing public data set.

2. Related work
The joint embedding model is the most basic one in the research of visual question answering algorithms, which jointly represents two different modalities of vision and text in the feature space. Malinowski [4] and others proposed for the first time a joint embedding model Neural-Image-QA applied to real scenes, using convolutional neural network CNN to extract image features and then transfer them to the long and short-term memory LSTM together with the question text to generate the word sequence of the answer. In terms of text characterization, Zhou et al. [5] chose the word bag model Bow, which is simpler than long and short-term memory LSTM, when processing problem texts, and proposed the iBOWING model, and migrated the pre-trained GoogleNet to extract image features.

The goal of the knowledge base query class is to create knowledge base query sentences based on images and text to obtain answers. Wang et al. [6] introduced the DBpedia knowledge base and proposed the Ahab model, which uses the pre-trained Faster R-CNN and two different VGG-nets. Wu et al. [7] proposed a method of introducing the external knowledge base into the joint embedding model. First, the semantic attributes in the extracted image are associated with the knowledge in the external knowledge base, and then a fixed-size knowledge vector is formed using Doc2Vec and sent to the LSTM model. Finally, it is fused with the question feature and then filtered by the classifier to get the answer.

The attention mechanism was first used in machine translation, and the soft and hard attention mechanisms proposed by Xu et al. [8] have become the mainstream method of VQA. Yang et al. [9] proposed a cascading attention model, which focused on the saliency area of the image many times, and finally learned the image area related to the question and inferred the answer. Lu et al. [10] proposed a joint attention mechanism, which paid attention to image areas and problems at the same time, and learned their attention weights to obtain the interaction of these two modalities. In addition, the TransformerNguyen architecture [11], which aggregates information in each mode by using the key-query attention mechanism, has also become popular.

3. Method for VQA
Most attention mechanisms used for visual question answering are top-down, that is, focusing on the parts of the image that are closely related to the problem according to the problem in the current task. These attention mechanisms are all based on image features extracted by CNN. The input area corresponds to a unified grid composed of neural receptive areas of the same size and shape. It has nothing to do with image content and cannot better characterize image information. In order to be more in line with human visual attention and improve the ability to express image content, the model in this paper pays more attention to objects and salient areas in the image, and enhances the representation of image content. The overall model flowchart is shown in Figure 1.
Firstly, the pre-trained ResNet model [12] is used to extract image features with spatial information and location information. Secondly, LSTM is used to extract high-level semantic features related to image problems, and then an attention mechanism is applied to the image and the problem. The force is further concentrated on the area related to the question, and finally the image and text information of interest are spliced and fused and sent to the softmax classifier to predict the answer. The training process of the visual question answering system implemented in this paper can be divided into three stages: feature extraction module, stack attention module, and answer generation module.

### 3.1. Feature extraction module

#### 3.1.1. Image feature extraction

Since the performance improvement of AlexNet [13] due to its simple initial structure, subsequent improvements to the method of visual feature recognition have been improved on this basis. In order to better understand the relationship between image regions, this paper uses the ResNet model pre-trained on the data set to train a deeper model. ResNet is easy to optimize and can improve the accuracy of classification by increasing the depth of the network. In this paper, a 152-layer residual neural network is used to process image information. ResNet-152 contains 4 residual learning blocks. The number of units in each block is 3, 8, 36, and 3, respectively. In the experiment, the image size in the data set is modified to 224*224 as input, and the fully connected layer of ResNet-152 is selected as output, and finally the image feature representation is obtained:

\[
V = (V_1, V_2, \ldots, V_N)
\]

Where \( V_N \) is the feature vector at the spatial location \( N \).

#### 3.1.2. Text feature extraction

Text representation is to convert the preprocessed text into a computer-recognizable text vector form. One-hot encoding is the simplest and easy to understand representation method. Its disadvantage is that as the dimension increases, it will cause a lack of connection between languages and cannot be represented. A higher-level feature semantics [14]. In this paper, a long and short-term memory neural network LSTM is used to avoid repetitive operations on a single matrix. The training sample is composed of no more than 20 words. The pre-trained GloVe word embedding method (Global Vectors for Word Representation) [14] is used for word vector Initialize, encode the input question word \( Q_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(l)}) \). Then use LSTM to extract features and use \( Q_i \) as input node data. The phrase input to the network at time \( t \) is updated according to the LSTM cell state \( c_{-1} \) and hidden layer state \( H_{t-1} \) at time \( t-1 \) to obtain \( H_t \). And input the obtained \( H \) as a feature vector of a
phrase into the next layer of the network, and output the calculation result at the last moment as the expression of the problem data:

\[ H = LSTM(x_t^{(i)}) \]  

(2)

3.2. Stack attention module

The attention mechanism is mainly to simulate the human attention distribution when watching a picture or a paragraph of text [15]. For a given input and target, the attention mechanism will generate a weight distribution vector, which will integrate the input information to get the result of attention.

Since the previous attention mechanism did not make good use of the problem features, the key point of the deep stack attention mechanism proposed in this paper is to pay attention to the image and the problem information, and to focus on the relationship between the image spatial location information and the problem text. The specific methods are as follows, through the residual neural network and the long and short-term memory neural network, we obtain the characteristic representation of the image and the problem, and then obtain the attention matrix through the tanh activation function:

\[ C = \tanh(Q^T W_b V) \]  

(3)

Where \( Q \) represents the problem feature, \( V \) represents the image feature, \( W_b \) represents the weight parameter.

When traditional algorithms segment image regions at the image level, there will be problems such as attention errors at the image level. Through the image attention mechanism, multiple target areas in the image can be detected, objects in the image can be initially located, and background information can be weakened. However, the model cannot determine the association with the problem. By introducing the problem attention mechanism, the attention to the image is more focused on the area related to the problem.

After obtaining the attention matrix \( C \), combine the attention matrix with image features and problem features to generate image information features with text attention and text information features with image attention:

\[ H'^q = \tanh(W_q Q + C) \]  

(4)

\[ H'^v = \tanh(W_v V + C) \]  

(5)

\[ \alpha'^v = \text{soft max}(w_v^T H_v) \]  

(6)

\[ \alpha'^q = \text{soft max}(w_q^T H_q) \]  

(7)

Where \( \alpha'^v \) and \( \alpha'^q \) are the attention probability of each image area \( v \) and each question \( q \) respectively.

3.3. Answer generation module

The answer generation module uses the fused features as input to predict the answer. This paper uses the nonlinear layer to obtain the image feature vector \( H^v \) and the problem feature vector \( H^q \), and uses the tanh activation function to normalize the vector, and then fuse the features by element-wise multiplication:

\[ h = f(H'^q) \circ f(H'^v) \]  

(8)

Each question in the VQA data set is associated with one or more answers, and these answers are labeled with a precision between 0 and 1. The multi-label classifier in this paper inputs the \( h \) vector of the joint representation into the non-linear layer \( f_0 \) fusion feature, and uses the Sigmoid function to normalize the score, and then predicts the score in the candidate answer through linear mapping, and selects the output with higher score:

\[ \hat{s} = \sigma(w_h f(h)) \]  

(9)
Where $\sigma$ is the activation function, and $w_0$ is the initial weight matrix. Through this function, the final score is normalized to the (0,1) interval, and the maximum probability is taken as the output:

$$L = -\sum_{i}^{M} \sum_{j}^{N} s_{ij} \log(\hat{s}_{ij}) - (1 - s_{ij}) \log(1 - \hat{s}_{ij})$$  \hspace{1cm} (10)

The indices $i$ and $j$ respectively cover $N$ candidate answers of $M$ training questions, and the true value $s_{ij}$ is the true answer to the question.

4. Experiment and analysis

4.1. Experimental Data

For effective evaluation, this article uses the two largest publicly available visual question and answer data sets: the Visual Genome data set and the VQA2.0 data set. These two data sets contain a large number of question pairs, and they are mainly composed of realistic images and abstract cartoon pairs, which are challenging to the research of visual question and answer.

Visual Genome dataset: contains 108077 images, 5.4 million area descriptions, 1.7 million question answer pairs, 3.8 million object instances, 2.8 million attributes, and 2.3 million relationships. Each image has an average of 35 objects, 26 attributes, and 21 pairs of object relationships.

VQA2.0 data set: On the basis of VQA1.0, it has expanded the scale and reduced the existing language bias problem. It contains 123287 training images and 81434 test images, mainly from the MS-COCO data set. In order to evaluate the quality of answers, the data set uses standard VQA evaluation indicators to predict the accuracy of answers.

4.2. Evaluating Indicator

The text uses the evaluation index report provided by the data set author to evaluate the performance of visual question answering. Among them, only when three or more annotators vote for the answer, the predicted answer is considered correct, that is, for the VQA data set, if the predicted answer is the same as the marked answer in the data set, it can be considered correct answer.

$$\text{Acc}(\text{ans}) = \min \left\{ \frac{\# \text{humans vote for ans}}{3}, 1 \right\}$$  \hspace{1cm} (11)

4.3. Parameter Settings

For the convenience of calculation, when processing data, for the question and answer text pair, this experiment chooses the answer that appears 8 times or more in Visual Genome and VQA2.0 to constitute the candidate answer. When dealing with problem information, filter out some meaningless words, and only keep the words with semantic information, so as to avoid double calculation.

Before training the network, we preprocess the candidate answer set through annotation technology. The network is trained end-to-end through backpropagation, using Adam to optimize the model. In order to obtain a more accurate image description and minimize the loss function value, the experimental parameters in the article are set as follows: the learning rate is set to 0.001, the batch size is 64, the maximum number of iterations is 30, and the learning rate is attenuated at 10 and 20 respectively. The attenuation rate is fixed at 0.1.

4.4. Results and Analysis

In order to verify the effectiveness of the model in this article, we tested the model algorithm proposed in the article on public data sets such as Visual Genome and VQA2.0, and divided the questions into four categories according to the answers: Yes/No, Num, Other, All. The model prediction results on different data sets are shown in Figure 2.
In addition, a comparative experiment was carried out to study the influence of each attention in the model, and the weight value p was set to further verify on the VQA2.0 test set. The initial challenge model M[16] adopted the simplest element addition and fusion method. On this basis, this article uses element-wise multiplication for feature fusion. As shown in Table 1 below, different weight values are given in The comparison result on forecast accuracy.

Through comparative analysis, it can be seen that the prediction accuracy rate of the model in the article is better than that of the M model. However, the results are different when different weight values are selected. On the VQA2.0 verification set, when p=0.2, the result is the best Yes, the overall prediction accuracy rate is 1.59% higher than that of the M model. When the weight values p=0.1 and p=0.2, the results are similar when predicting other types of problems. After comparison, the idea of this article is further verified. The model needs to pay attention to multiple important areas of complex problems. It is not possible to use only one superimposed attention.

On the test-std test set of VQA2.0, the method of this article is compared with the experimental results of other six network models, including MCB model, MLB model, Language-only model, d-LSTM+nI model, and baseline model Prior And mainstream algorithms such as the Up-down model that combines bottom-up and top-down attention, and compare the algorithm performance from four problem categories. The experimental results of this method and other methods are shown in Table 2.

| Weights | Yes/No | Num | Other | All |
|---------|--------|-----|-------|-----|
| M       | 80.07  | 42.87 | 55.82 | 63.14 |
| P=0     | 82.14  | 42.96 | 55.73 | 64.32 |
| P=0.1   | 82.78  | 44.38 | 57.03 | 64.36 |
| P=0.2   | 84.03  | 43.65 | 57.04 | 64.73 |
| P=1     | 82.26  | 43.71 | 56.89 | 64.64 |

| Models   | Yes/No | Num | Other | All |
|----------|--------|-----|-------|-----|
| MCB      | 79.81  | 38.28 | 53.36 | 62.27 |
| MLB      | 83.96  | 44.77 | 56.52 | 66.62 |
| Language-only | 67.01  | 31.55 | 27.37 | 44.26 |
| d-LSTM+nI | 73.45  | 35.28 | 41.89 | 54.23 |
| Prior    | 61.24  | 0.43  | 1.26  | 25.91 |
| Up-down  | 82.25  | 43.97 | 56.24 | 65.70 |
| Ours     | 82.04  | 46.88 | 58.47 | 67.75 |
According to the table, the deep-stacked attention network model proposed in this paper achieves the best performance in accuracy and classification of the four problems. The overall accuracy of the model is also 1.13% higher than the accuracy of other best models. The worst of them is the Prior model, which has an accuracy rate of 41.84% higher. The method proposed in this article is lower than the Up-down model and the MLB model when answering the "yes/no" question. It is in answer to the "count" and "other" categories. When it comes to related issues, the model in this paper has improved accuracy compared to the six existing models proposed in the paper, and the prediction accuracy of the best model is 2.11% and 1.95% higher than that of the best model, which verifies that the model is in The overall prediction effect of the visual question answering system is the best.

5. Conclusion
This paper proposes a visual question answering model based on deep stacked attention network. Through the deep learning model based on image attention and question attention stacking, the visual and text features are processed to realize the interactive attention between image and text features, and a lot of comparative experiments are carried out Through comparison and analysis with existing models, the effectiveness of the method is verified. Due to the inexplicability of deep learning and the low accuracy of model predictions, the next stage of work is to add reasoning ability to the visual question answering system and apply it to preschool education, so that the model can learn knowledge and reasoning, and optimize human-computer interaction ability to better understand images and natural language.

References
[1] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, Min Sun. (2017) Leveraging Video Descriptions to Learn Video Question Answering, AAAI.
[2] Hyeonwoo Noh, Bohyung Han. (2016) Training Recurrent Answering Units with Joint Loss Minimization for VQA, arXiv:1606.03647.
[3] Le T, Huy N T, Minh N L. (2020) Integrating Transformer into Global and Residual Image Feature Extractor in Visual Question Answering for Blind People[C]/ 2020 12th International Conference on Knowledge and Systems Engineering (KSE).
[4] Malinowski M, Rohrbach M, Fritz M. (2017) Ask Your Neurons: A Deep Learning Approach to Visual Question Answering[J]. International Journal of Computer Vision.
[5] Zhou B, Tian Y, Sukhbaatar S, et al. (2015) Simple Baseline for Visual Question Answering[J]. Computer Science.
[6] Wang P, Wu Q, Shen C, et al. (2017) Explicit Knowledge-based Reasoning for Visual Question Answering[C]/ Twenty-Sixth International Joint Conference on Artificial Intelligence.
[7] Wu Q, Shen C, Wang P, et al. (2017) Image Captioning and Visual Question Answering Based on Attributes and External Knowledge[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1-1.
[8] XU K, BA J, KIROS R, et al. (2015) Show, attend and tell: Neural image caption generation with visual attention [C]/Proceedings of the 32nd international conference on Machine Learning. New York: ACM, 2048-2051.
[9] 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. Piscataway, NJ: IEEE, 21-29.
[10] LU J, YANG J, BATRA D, et al. (2016) Hierarchical question image co-attention for visual question answering[C]/ Proceedings of Advances in Neural Information Processing Systems. Cambridge, MA:MIT Press,289-297.
[11] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention Is All You Need[J]. arXiv.
[12] He K, Zhang X, Ren S, et al. (2016) Deep Residual Learning for Image Recognition[J].
[13] Krizhevsky A, Sutskever I, Hinton G E. (2017) ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 60(6):84-90.

[14] Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap, (2017) A simple neural network module for relational reasoning, arXiv:1706.01427.

[15] Lu J , Yang J , Batra D , et al. (2016) Hierarchical Question-Image Co-Attention for Visual Question Answering[J].

[16] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge[C]. in CVPR.