Image Caption Combined with GAN Training Method

Zeqin Huang[1,2(✉)] and Zhongzhi Shi[1]

[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100190, China
{huangzeqinl7g, shizz}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. In today's world where the number of images is huge and people cannot quickly retrieve the information they need, we urgently need a simpler and more human-friendly way of understanding images, and image captions have emerged. Image caption, as its name suggests, is to analyze and understand image information to generate natural language descriptions of specific images. In recent years, it has been widely used in image-text crossover studies, early infant education, and assisted by disadvantaged groups. And the favor of industry, has produced many excellent research results. At present, the evaluation of image caption is basically based on objective evaluation indicators such as BLUE and CIDEr. It is easy to prevent the generated caption from approaching human language expression. The introduction of GAN idea allows us to use a new method of adversarial training. To evaluate the generated caption, the evaluation module is more natural and comprehensive. Considering the requirements for image fidelity, this topic proposes a GAN-based image description. The Attention mechanism is introduced to improve image fidelity, which makes the generated caption more accurate and more close to human language expression.

Keywords: GAN · Deep learning · Attention mechanism · Image caption · LSTM

1 Introduction

Image caption, as its name suggests, generates natural language descriptions of specific images. Due to its extensive use in image-text cross-research, early infant education, and assistance from disadvantaged groups, it has become more and more popular in academia and industry in recent years. There has produced a lot of excellent researches and results related to it [1, 2].

With the rapid development of the Internet and computer technology, we have formed a world constructed with images. Using a large number of images to automatically generate easy-to-understand knowledge has become a topic that attracts wide attention. On the one hand, the number of images is increasing, on the other hand, people cannot retrieve and find the required information from such a large number of

Published by Springer Nature Switzerland AG 2020
Z. Shi et al. (Eds.): IIP 2020, IFIP AICT 581, pp. 310–316, 2020.
https://doi.org/10.1007/978-3-030-46931-3_29

images. Therefore, we urgently seek to be able to automatically analyze and understand image information, a simpler and more human-friendly way of image understanding, and image caption was born to meet this need, it can automatically build images consistent with human cognition Semantic information. At the same time, in solving such a problem of interaction between images and NLP, the traditional deep learning technology is further improved and integrated to adapt to such a difficult task.

Computer vision is an important task in the computer field, and image perception and image texting are the main problems to be solved. Studying such an image understanding technology (image caption) has very important progress significance. This is a classic artificial intelligence, brain-computer collaboration and other framework for the classic problem of image information understanding and perception, that is, how to use the computer's brain-computer collaboration, neural Understanding to simulate people's analysis, cognition, recognition and memory functions of images. What's more, this technology will be a sign that traditional artificial intelligence is moving towards true artificial intelligence. In addition, the latest research results in the field of machine learning and artificial intelligence will also be further used for this task to improve performance and theoretical supplementation. In the process, they complement each other and promote each other.

2 Related Work

In general, the current method for image caption tasks in the field of deep learning is mainly the Encoder-decoder model. That is, the basic extension of the model based on CNN + RNN. In addition, after the introduction of the attention mechanism, the performance of the universal Encoder-decoder model has been significantly improved. In 2014, Baidu's Mao Junhua and others creatively combined CNN and RNN to deal with problems such as image annotation and image sentence retrieval. At the same time, they pioneered the application of deep learning to the image caption task and achieved good results. Although the model m-RNN proposed in this paper has some disadvantages, it has achieved very good results, so far, many domain papers still use this model as a baseline [3]. Later in 2014, GOOGLE proposed the NIC model to promote the m-RNN. They replaced RNN with LSTM and AlexNet with GooLeNet, in the end the model was a great success [4].

In recent years, there have been a lot of related work and many gratifying breakthroughs in the image caption task. This has benefited from convolutional neural networks and recurrent neural networks [5], but almost all solutions have not departed from the Encoder-Decoder framework. Lu, Xiong, Parikh and Socher introduce Attention mechanism in to the CNN Part and greatly improved the efficiency of the model [6]. Anderson, He and Buehler et al. added attention-mechanism to feature extraction and caption generation to improve performance [7]. Chen, Mu et al. Tried to introduce GAN's adversarial training ideas into the image caption task to improve performance even more [8]. However, they still leave issues of feature distortion and semantic relevance.

3 Model Combined with GAN

Next, we will propose an Image Caption Model combining GAN's adversarial training ideas, based on Convolution Neural Network (CNN) and Long Short Term Memory Network Network (LSTM). And introduce with the Attention Mechanism to improve performance. When we describe an image, we need to pay attention to the content of the image as well as the language foundation. When we get the word "cat," we focus on the cat part of the image and ignore the rest. The prediction of a word requires not only the introduction of attention mechanism in the language model, but also in the image.

3.1 CNN for Feature Extraction

Bottom-up Model. In this model, we define spatial regions based on bounding boxes and use Faster R-CNN to achieve bottom-up attention model [9, 10]. Faster R-CNN is an object detection model designed to identify object instances belonging to certain classes and localize them using bounding boxes. The final output of the model includes the softmax distribution on the class labels and the class-specific bounding box optimization proposed by each box. To pre-train the bottom-up attention model, we first initialize Faster R-CNN using pre-trained ResNet-101 for classification on ImageNet. To predict the attributes of region *i*, we embed the average merged convolutional feature vi with the learned ground truth object classes and feed them to define the softmax distribution on each attribute class and the additional output of the "no attribute" class Layer.

3.2 LSTM for Caption Generation

Top-down Model. General RNN cannot save too much information, there is only one state in the hidden layer, if we add another state C to save long-term information, the problem will be solved. LSTM is an improved recurrent neural network, it uses gate to control long-term status C [11]. The gate can be expressed as:

$$g(x) = \sigma(Wx + b) \tag{1}$$

W is the weight vector of the gate and b is the bias term. $\boldsymbol{\sigma}$ is the sigmoid function and the range is (0,1), so the state of the gate is half open and half closed. The final output of the LSTM is jointly controlled by the output gate and cell state:

$$\boldsymbol{h_t} = \boldsymbol{o_t} \cdot \tanh(\boldsymbol{c_t}) \tag{2}$$

Because of the control of oblivion gate, it can save the information of a long time ago, and because of the control of the input door, it can avoid the current inconsequential content from entering the memory.

3.3 Attention Mechanism

The AM model is one of the most important developments in the field of NLP in the past few years, and appears in most current papers with the Encoder-Decoder framework [12]. But the AM model can be used as a general idea. When the general RNN model generates a language sequence, the predicted next word is only related to its first n words.

$$y_i = f(y_{i-1}, y_{i-2}, y_{i-3}, \ldots, C) \tag{3}$$

The above formula C represents semantic encoding. Obviously, the semantic meaning of each word is unreasonable. A word will not only be related to the nearest word, so we give each word a probability distribution and express its relevance to other words. And replace C with C_i.

$$C_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{4}$$

T_x represents the number of other words related to C_i, and α_{ij} represents the probability of attention between two words. h_j is the information of the word itself. The question now is how to calculate the probability distribution. It is usually to calculate the similarity between the current input information H_i and the previous information h_j

After the multi-layer convolution structure, the image information is compressed into a vector I. When predicting each word, you need to associate some information in the vector. The image attention parameter W is a parameter that we need to train to obtain.

$$A_i = I \cdot W_i \tag{5}$$

Finally, the functional relationship of each predicted word is as follows:

$$y_i = f(A_i, C_i, y_{i-1}, y_{i-2}, y_{i-3}, \ldots) \tag{6}$$

3.4 Adversarial Training of GAN

The Generative Adversarial Network consists of a Generative Network and a Discrimination Network [13, 14]. The generating network randomly samples from the latent space as input, and its output needs to mimic real samples in the training set as much as possible. The input of the discriminating network is the real sample or the output of the generating network. The purpose is to distinguish the output of the generating network from the real samples as much as possible. The generation network should try to deceive the identification network as much as possible. The two networks oppose each other and constantly adjust the parameters. The ultimate goal is to make the judgment network unable to judge whether the output of the generated network is true. Introduced the idea of GAN's adversarial training in LSTM to make the generated caption closer to human natural language expression, that is, more authentic. In the

LSTM model discussed above, we add a discriminator $D_{\emptyset}$ and an evaluator E, where $D_{\emptyset}$ is used to determine that the caption is a machine-generated probability d and E is used to evaluate the accuracy of caption. Combine the two scores and feed them back to the network to train the network:

$$r = \alpha * d + (1 - \alpha) * e \tag{7}$$

And the whole model is shown as Fig. 1:

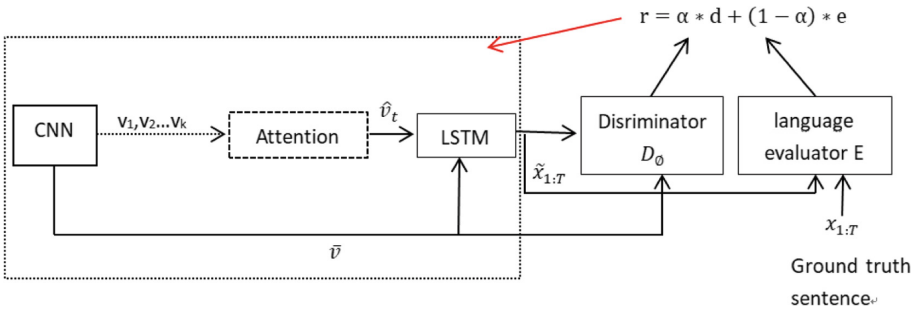

Fig. 1. Our whole model

4 Experiments

In order to compare with the prior art, we conducted a large number of experiments using the BLEU metrics [15] to evaluate the effectiveness of our model. Experiments used the MSCOCO 2014 data set, which includes 123000 images. The dataset contains 82783 images in the training set, 40504 images in the validation set and 40775 images in the test set. we use the whole 82783 training set images for training, and selects 5000 images for validation and 5000 images for testing from the official validation set.

4.1 Evaluating Indicator

A popular automatic evaluation method is the BLEU algorithm proposed by IBM. The BLEU method first calculates the number of matching n-grams in the reference sentence and the generated sentence, and then calculates the ratio of the number of n-grams in the generated sentence. As an evaluation indicator. It focuses on the accuracy of generating words or phrases in sentences. The accuracy of each order N-gram can be calculated by the following formula:

$$P_n = \frac{\sum_i \sum_k \min\left(h_k(c_i), max_{j \in m} h_k\left(s_{ij}\right)\right)}{\sum_i \sum_k \min(h_k(c_i))} \tag{8}$$

The upper limit of N is 4, which means that only the accuracy of 4-gram can be calculated.

4.2 Results and Discussion

This model is improved on the basis of the basic model and improves the performance on the basis of the baseline. By comparing with the two previous models, we get a higher accuracy rate. As we can see in the Table 1, compared with other mainstream algorithms, one and two keywords perform better, but the effect of getting more keywords is not good.

Table 1. Comparison with other methods on the MSCOCO dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GLA	56.8	37.2	23.2	14.6
BRNN	64.2	45.1	30.4	20.3
Google NIC	66.6	46.1	**32.9**	**24.6**
OURS	**67.6**	**47.1**	31.8	22.8

5 Conclusions

Image caption is a complex task, and deep learning-based frameworks have become the current mainstream method. This paper proposes a multi-attention mechanism based on GAN training methods, which requires understanding the syntax of sentence generation and the content in images. Different words have different levels of attention to image content, and contexts have different levels of attention. Experimental results show that under the BLEU evaluation standard, the attention mechanism combined with GAN training methods can achieve better results.

References

1. Shi, Z.: Mind Computation. World Scientific Publishing, Singapore (2017)
2. Vinyals, O., et al.: Show and tell: a neural image caption generator. In: Computer Vision and Pattern Recognition, pp. 3156–3164. IEEE (2015)
3. Mao, J., Xu, W., Yang, Y., et al.: Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090 (2014)
4. Vinyals, O., Toshev, A., Bengio, S., et al.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 652–663 (2016)
5. Hollink, L., Little, S., Hunter, J.: Evaluating the application of semantic inferencing rules to image annotation. In: International Conference on Knowledge Capture, pp. 91–98. ACM (2005)
6. Lu, J., Xiong, C., Parikh, D., et al.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 375–383 (2017)
7. Anderson, P., He, X., Buehler, C., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)

8. Chen, C., Mu, S., Xiao, W., et al.: Improving image captioning with conditional generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8142–8150 (2019)
9. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
10. Girshick, R.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448. IEEE (2015)
11. Jia, X., et al.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415. IEEE (2016)
12. Yan, S., Xie, Y., Wu, F., et al.: Image captioning via hierarchical attention mechanism and policy gradient optimization. Sig. Process. **167**, 107329 (2020)
13. Yu, L., Zhang, W., Wang, J., et al.: SeqGAN: sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
14. Dai, B., Fidler, S., Urtasun R., et al.: Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2970–2979 (2017)
15. Papineni, K., Roukos, S., Ward, T., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)