ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 195-200
doi: 10.18178/wcse.2019.06.029

Feature Fusion Based on Neural Image Captioning with Spatial Attention

Qingqing Lu [1,2], Xiaomei Zhang [1], Xin Kang [2] and Fuji Ren [2 +]

[1] School of Information Science and Technology, Nantong University, No.9 Seyuan Road, Chongchuan District, Nantong, Jiangsu, China

[2] Faculty of Engineering, Tokushima University 2-1 Minami Josanjima, Tokushima, 770-8506, Japan

Abstract. Generating a natural language description of an image is a challenging but meaningful task. The task combines two significant artificial intelligent fields: computer vision and natural language processing. This task is valuable for many applications, such as searching images and assisting the people who have visually impaired to view the world, etc. Most approaches adopt an encoder-decoder framework, and some of the future methods are improved on the basis of this framework. In these methods, image features are extracted by VGG net or other networks, but the feature map will lose important information during the processing. In this paper, we fusing different kinds of image features extracted by the two networks: VGG19 and Resnet50, and put it into the neural network to train. We also add an attention into the a basic neural encoder-decoder model for generating natural sentence descriptions, at each time step, our model will attend to the image feature and pick up the most meaningful parts to generate captions. We test our model on the benchmark dataset called IAPR TC-12, comparing with other methods, we validate our model have state-of-the-art performance.

Keywords: Image captioning, feature fusion, encoder-decoder framework, attention

1. Introduction

With the development of deep learning recently, people have made new discoveries and progress in the research on automatically generating captions from images. Image captioning can be considered as a dynamic target detection, and descriptions are generated from global information. Given a picture with rich content, people can elaborate on it from many angles, but for the machine, it is a complex problem. Because it involves not only a fin-grained understanding of the global and the local entities in an image, but also the relationships and attributes. In addition, it needs to be able to capture the semantic information of the image and generate sentences which conform to human grammar habits.

The generic neural encoder-decoder framework is an effective method for image captioning, it generally employs convolutional neural networks to encode the visual information and utilize recurrent neural networks to decode that information to coherent sentences, but it has two potential disadvantages. Firstly, the image features we used have been extracted by networks on the ImageNet which is a large visualization dataset for visual object recognition research, thus we just need to download the weights of the networks we need instead of re-extracting the image features. The networks can be VGG16, VGG19 [1], Inception v3 [2], Resnet50 [3], etc. However the weights we download may not fully adapt to the application of our model. Secondly, there is a great possibility for an encoder-decoder model to lose its mind on important information which is essential for the system to generate richer and descriptive captions.

+ Corresponding author. Tel.: + 088 - 656 - 9684; fax: + 81886566575.
E-mail address: ren@is.tokushima-u.ac.jp.

In this paper, we propose a method of feature fusion to combine the image features of two networks which will be the input in the next step. Thus the number of image features obtained will more than normal. The process of putting the image features in the neural network to train makes the image features more adaptable to the application of our model. A powerful attention mechanism is added into our model to extract significant information. At last, we compare four methods on the IAPR TC-12 dataset with the metric BLEU-1, 2, 3, the score of the value proves our model's superiority.

2. Related Work

To deal with the problem of generating a natural language description from an image, many approaches have been proposed. But based on the existing models and methods, we can make some changes to improve the performance of the original model.

Neural networks are used widely in machine translation, Kiros et al. [4] proposed firstly to use neural network to generate caption of image, this paper used a multimodal log-bilinear model which constructed a joint multimodal embedding space by combining a powerful computer vision model and an LSTM. And then the encoder-decoder framework [5, 6] was brought into the task of image captioning, method as this framework usually encoded an image as feature vector, and transfer them into recurrent neural network to decoder to generate a sentence corresponding to the image. Mao et al. [7] replaced the feed forward neural network with a recurrent neural network creatively. Vinyals et al. [8] made the image features entered only once which prevented the recurrent of the noise in the picture, and used an LSTM instead of a vanilla RNN as the decoder.

In the area of object recognition, feature fusion is adopted to achieve a higher classification rate. Guan et al. [9] introduced a novel framework that combines interactive image segmentation with multifeature fusion to achieve improved Mobile Landmark Recognition (MLR) with high accuracy. Shi et al. [10] proposed the multiple feature fusion image retrieval algorithm based on the texture feature and rough set theory, they fused the different features with operation of normalization and the rough set theory will assist them to enhance the robustness of retrieval system when facing with incomplete data warehouse. Hoashi et al. [11] put forward an automatic food image recognition system for 85 food categories by fusing various kinds of image features including bag-of-features (BoF), color histogram, Gabor features and gradient histogram with Multiple Kernel Learning (MKL).

Recently, attention mechanisms [12, 13, 14, 15] have been applied into the encoder-decoder neural framework in the aspect of image captioning. Xu et al. [16] added an attention mechanism to learn a latent alignment from scratch when generating corresponding words. You et al. [15] injected semantic concepts and attributes into neural image captioning. Lu et al. [17] proposed a special attention mechanism that can decide to ignore the non-visual words such as "the", "of", and "to".

3. Model

3.1. Framework

Shown in the Fig1, given an image and its corresponding caption, we calculate the maximum value of the probability of the correct description, $S = \{S_1, S_2, \dots, S_N\}$ represents the true sentence describing of the image. We denote S_0 as a special start word and S_N as a special stop word, which designates the start and end of the sentence. Both the image and words are mapped to the attention layer, the image by fusing VGG19 network and Resnet50 network, the words by using word embedding W_e. Our loss is the sum of the negative log likelihood of the correct word at each step.

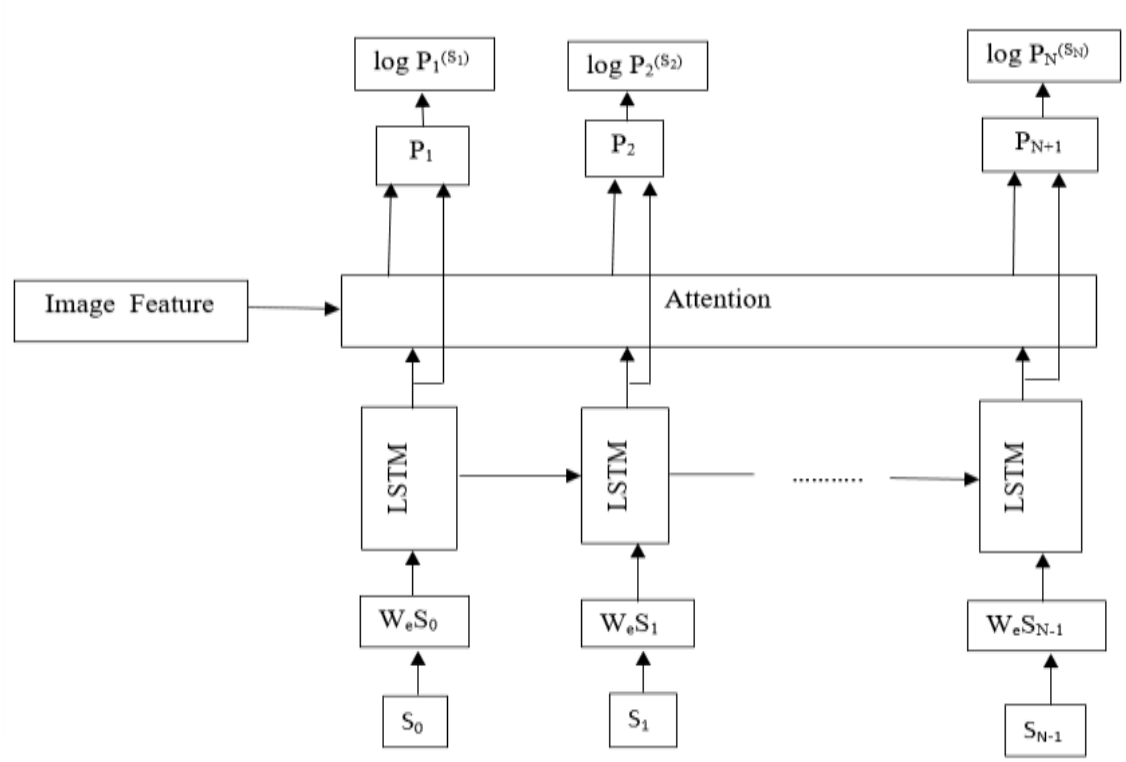


Fig. 1: Image feature fusion for an encoder-decoder model with attention mechanism.

3.2. Caption Generation

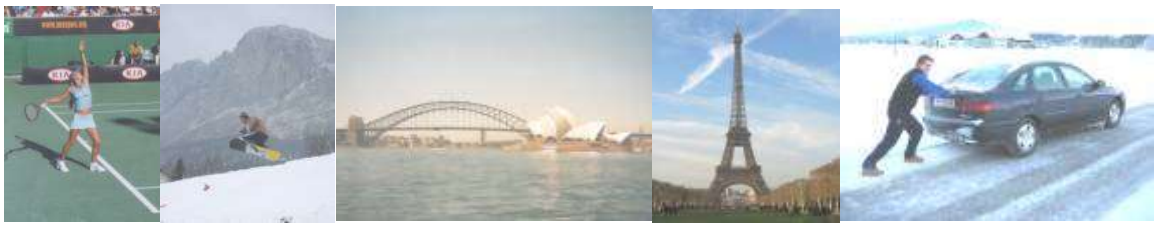
Our model is an encoder-decoder framework, it has two inputs: image and reference caption. The encoder uses a CNN to get the representation of images. Different from [17] which uses the feature outputs of the last convolutional layer of Resnet, we fuse the image features of VGG19 and Resnet50, which makes it possible for the model to identify more objects in the image. In the decoder part, we concatenate the word embedding vector and global image feature vector as the input vector of LSTM at time step t. The hidden state of the LSTM at time t is influenced by the hidden state of the LSTM at time t-1, memory cell vector at time t-1 and the input vector at time t. The spatial image features and the hidden state of LSTM at time t are computed with attention function to get the context vector, which provides visual evidence for caption generation. The probability over a vocabulary of possible words at time t is effected by the hidden state of LSTM at time t and the context vector, and the probability is calculated with softmax function.

4. Experiment

We next depict the dataset used for testing, followed by an evaluation of experimental result on sentence generation.

4.1. Dataset

In the paper, we use the IAPR TC-12 Benchmark [18], which consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. Example images can be seen in Fig2.



Sports City pictures Action shots

Landscape shots Animal pictures People shots

Fig. 2: Example images.

4.2. Training

We train our model with a batch of 64, giving roughly 500 epochs with early stopping if the value of validation loss had not improved. We use the Adam optimizer. Due to the high value of epoch, overfitting is existing in the experiment. We tried dropout model to drop some parameters, and the BLEU points were improved a few. In order to deal with another problem of incomplete description, we attempt to fuse the image features extracted by different networks in the part of image processing in the basis of the encoder-decoder neural model with spatial attention. Increased number of image features makes our result better, the score of BLEU is higher and the value of loss reduces.

4.3. Result

We next describe our caption generation results, beginning with a short discussion of evaluation metric. We adopt BLEU scores (i.e. B-1, B-2, B-3) [19] as the evaluation metric in this paper. BLEU was originally designed for automatic machine translation, with several reference sentences gave, they rated the quality of the sentences translated. BLEU uses the accuracy of the generated sentence n-grams to calculate the geometric mean to obtain the similarity of the sentences. The accuracy of the so-called n-grams is the proportion of the n-word string in the generated sentence appearing in the reference sentence. BLEU remains the standard evaluation metric for image captioning generation, though it has drawbacks. Even if the sentence generated is better and smoother than the original sentence , the score is low.

In the experiment, we conduct three different methods on the IAPR TC-12 dataset. As we can see in Table 1, our approach works best on the BLEU-1, 2, 3 metrics. Compared to the method of multimodal recurrent neural network (M-RNN) [7], neural image captioning has a certain improvement, this progress owes to the neural image captioning input the image features only once. Neural image captioning with spatial attention is proved superior to the neural image captioning(NIC) [8], this dues to the use of attention, the model can decide to attend to the most striking part of the image. Our model combines image features of two networks based on the the model of neural image captioning with spatial attention [17], and then put the feature fusion into the neural network to train, thereby the image features we get are more suitable for our model, the consequence indicates the advantage of our model.

Table 1: Performance of our method on the IAPR TC-12 dataset, comparing with state-of-the-art methods

Method	B-1	B-2	B-3
M-RNN [7]	0.395	0.183	0.131
NIC [8]	0.445	0.211	0.206
NIC+ Spatial attention [17]	0.546	0.331	0.233
Our model	**0.577**	**0.406**	**0.303**

(1)

NIC	A valley with a large palm tree.
NIC+ Spatial attention [17]	A square with a red and yellow building with green columns.
Our model	People on a square with many standing trees and red flowers and lamps.
Original sentence	A square with a lawn red and purple flowers a palm tree and street lamps.

(2)

NIC	A man and a woman are standing are standing in front of a brown rock.
NIC+ Spatial attention [17]	A man and a woman are standing in a brown sandy desert.
Our model	Eight tourists are green on a brown and brown hill at the sea.
Original sentence	Eight tourists are posing on a brown hill at a steep coast at the sea.

Fig. 3: Visualization of generated captions on the IAPR TC-12 dataset.

Examples of the sentences generated can been seen in Fig3 , the important features were illustrated in the sentences. We record the captions generated by three models and their corresponding descriptions of the two images to compare.

In Fig3 (1), the sentence generated by neural image captioning is simplest, but the object "palm tree" is described correctly. The sentence generated by neural image captioning with spatial attention recognizes the color "yellow" and the object "square", which are absent in the first method. Our model generate a sentence with the richest content, the sentence has "square", "trees", "flowers" and lamps, which is most similar to the original sentence. In Fig3 (2), the sentences generated by the first two models look the same roughly expect for the description of the environment. Our model has outstanding performance on the description of characters comparing with the other two methods, and the the sentence also has the environments "hill" and "sea".

5. Conclusion

In this paper, we present a model that combines feature fusion and an encoder-decoder framework with spatial attention, which achieves state-of-the-art performance on the benchmark dataset using the BLEU metric compared with the other three approaches: multimodal recurrent neural network, neural image captioning and neural image captioning with spatial attention. Different from previous methods, our method utilizes feature fusion to replace the image features acquired by a single network, and put it into the neural net work to train, this step of processing can not only get abundant image features, but also make the image features more suitable for our model. Spatial attention is added in our model as well to attend on striking semantic information detected from the image to generate captions. For next steps, we plan to expand the training data and then add a mechanism of reinforcement learning into our model to get a better result.

6. Acknowledgment

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

7. References

[1] Simonyan, Karen, and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Computer Science, 2014.

[2] Szegedy. C, Vanhoucke. V, Ioffe. S, Shlens. J, Wojna. Z. Rethinking the Inception Architecture for Computer Vision. In: Computer Vision & Pattern Recognition, 2016.

[3] He. K, X. Zhang, S. Ren, J. Su. Deep Residual Learning for Image Recognition. In: Computer Vision and Pattern Recognition, 2015.

[4] Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unifying visual-semantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.

[5] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In CVPR, 2015.

[6] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars. Guiding the longshort term memory model for image caption generation. In ICCV, 2015.

[7] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In ICLR, 2015.

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, pages 3156–3164, 2015.

[9] T. Guan, Y. Wang , L. Duan , and R. Ji. On-device mobile landmark recognition using binarized descriptor and multifeature fusion. In: ACM Transactions on Intelligent Systems Technology, 2015.

[10] X. Shi, Y. Shao. Research on the Multiple Feature Fusion Image Retrieval Algorithm based on Texture Feature and Rough Set Theory. In Computer Vision and Pattern Recognition, 2016.

[11] Hoashi. H, Joutou. T, Yanai. K. Image Recognition of 85 Food Categories by Feature Fusion. In: ISM, 2010.

[12] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.

[13] P. Jiang, F. Ren and N. Zheng. A new approach to data clustering using a computational visual attention model. In International Journal of Innovative Computing, Information and Control, vol.5, no.12(A), pp.4597-4606, 2009.

[14] X. Wang, M. Peng, L. Pan, M. Hu, C. Jin, F. Ren. Two-level Attention with Two-stage Multi-task Learning for Facial Emotion Recognition. In Computer Vision and Pattern Recognition, 2018.

[15] Q. You, H. Jin, Z.Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In CVPR, 2016.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[17] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In CVPR, 2017.

[18] M. Grubinger, P. Clough. On the Creation of Query Topics for ImageCLEFphoto. Proceedings of the Third Workshop on Image and Video Retrieval Evaluation, pages 50-63, Budapest, Hungary, 2007.

[19] He. K, X. Zhang, S. Ren, J. Su. Deep Residual Learning for Image Recognition. In: Computer Vision and Pattern Recognition, 2015.