Joint Learning of CNN and LSTM for Image Captioning

Yongqing Zhu, Xiangyang Li, Xue Li, Jian Sun,
Xinhang Song, and Shuqiang Jiang

Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology Chinese Academy of Sciences,
No.6 Kexueyuan South Road Zhongguancun, Haidian District, 100190 Beijing, China
{yongqing.zhu,xiangyang.li,xue.li,sun.jian,
xinhang.song,shuqiang.jiang}@vipl.ict.ac.cn

Abstract. In this paper, we describe the details of our methods for the participation in the subtask of the ImageCLEF 2016 Scalable Image Annotation task: Natural Language Caption Generation. The model we used is the combination of a procedure of encoding and a procedure of decoding, which includes a Convolutional neural network(CNN) and a Long Short-Term Memory(LSTM) based Recurrent Neural Network. We first train a model on the MSCOCO dataset and then fine tune the model on different target datasets collected by us to get a more suitable model for the natural language caption generation task. Both of the parameters of CNN and LSTM are learned together.

Keywords: Convolutional neural network, Long Short-Term Memory, Image caption, Joint learning

1 Introduction

With the rapid development of Internet technologies and extensive access to digital cameras, we are surrounded by a huge number of images, accompanied with a lot of related text. However, the relationship between the surrounding text and images varies greatly, how to close the loop between vision and language is a challenging problem for the task of scalable image annotation [1, 2].

It is easy for our human beings to describe a picture after a glance of it. However, it is not easy for a computer to do the same work. Though great progress has been achieved in visual recognition, it is still far away from generating descriptions that a human can compose. The approaches automatically generating sentence descriptions can be divided into three categories. The first method is template-based [3, 4]. These approaches often rely heavily on sentence templates, so the generated sentences lack variety. The second method is retrieval-based [5, 6]. The advantage of these methods is that the captions are more human-like. However, it is not flexible to add or remove words based on the content of the target image. Recently, many researchers have used the combination of CNN and LSTM to translate an image into a linguistic sentence [7, 8].

Our method is based on deep models proposed by Vinyals [7] which takes advantage of Convolutional Neural Network (CNN) for image encoding and Long-Short Term Memory based Recurrent Neural Network (LSTM) for sentence decoding. We firstly train a model on the MSCOCO [9] dataset. We then fine tune the model on different datasets to make the model more suitable for the target task. In training and finetuning, the parameters of both CNN and LSTM are learned together.

Next, we introduce our methods in Section 2, followed by our experimental results in Section 3. At last, the section 4 concludes the paper.

2 Method

The model we use contains two types of neural networks, as illustrated in Figure 1. The first stage is CNN for image encoding and the second stage is Long-Short Term Memory(LSTM) based Recurrent Neural Network for sentence encoding [7, 8]. For CNN, we use the pre-trained VGGNet [10] for feature extraction. Using the VGGNet , we transform the pixels inside an image to a 4096-dimensional vector. After getting the visual features, we train an LSTM to obtain linguistic captions. In the LSTM training procedure, we change the parameters of both CNN and LSTM together. At last we fine tune the pre-trained model to get a more suitable model for the natural language caption generation task.

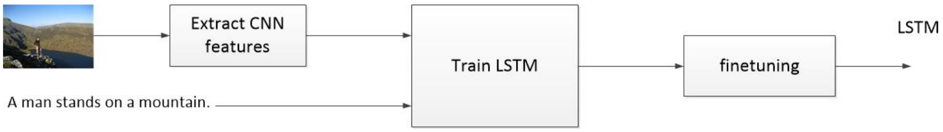


Fig. 1. An illustration of training stage.

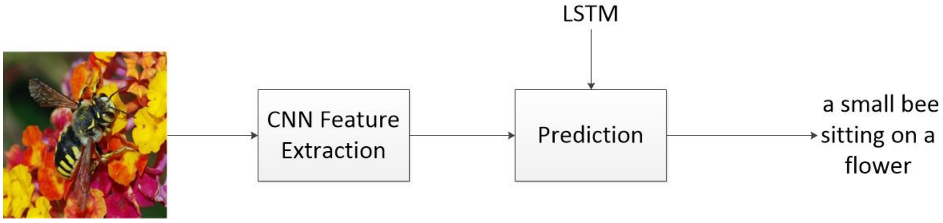


Fig. 2. An illustration of predicting stage.

Training stage

As shown in Figure 1, we use the pre-trained VGGNet [10] for CNN feature extraction. We first train the LSTM on corpora with paired image and sentence captions, such as MSCOCO [9] and Flickr30k [11]. In the training procedure of the LSTM, we change not only the parameters of the LSTM model, but also the parameters of the CNN model, which is a joint learning of CNN and LSTM. We then fine tune our model on different datasets as described in Section 3. At last, We use the trained models to predict linguistic sentence of a given image.

Predicting stage

The process of predicting an image is shown in Figure 2. To generate a sentence caption for an image, we get the CNN features of an image b_v, set the first hidden state h_0=0, x_0 to the START vector and compute the hidden state h_1 and predict the first word y_1. Then we use the word y_1 predicted by our model and set its embedding vector as x_1, the previous hidden state h_1, and then compute the hidden state h_2 and use it to predict the next word y_2. The process is repeated until the END token is generated.

3 Experiments and Submitted Runs

We first use the LSTM implementation from the NeuralTalk project [7]. We train models separately on the MSCOCO [9] dataset, the Flickr8k [12] dataset and the Flickr30k [11] dataset. Then we use the model to predict the images in the ImageCLEF 2016 validation set. The results are shown in Table 1. We use Meteor [13] to evaluate sentences generated by a model. Validation set we use here is the 2000 images and their corresponding sentences provided by the organizers. Because the performance of the model on the MSCOCO dataset is better than the other dataset, So we use the model trained on the MSCOCO as our pre-trained model.

Table 1. The performance of training a LSTM on different datasets.

Training data	Test data	Accuracy
MSCOCO	validation set	0.1326719463
Flickr8k	validation set	0.1168440478
Flickr30k	validation set	0.1231898764

We then do experiments to decide whether jointly learn the parameters of CNN and LSTM together or fixed the CNN and just learn the parameters of the LSTM. Firstly, we train a model which only learns the parameters in LSTM, then we use the model to predict the images in the validation set. The training set we use is the MSCOCO dataset, and the test set we use is the provided validation set. For comparison, we train a new model which not only learns the parameters of the LSTM but also fine tunes the CNN model. We then use the second model to predict the images in the validation set. And the results are shown in Table 2.

Table 2. The performance of joint learning CNN and LSTM or not.

Change parameters in CNN	Training data	Test data	Accuracy
NO	MSCOCO	validation set	0.1326719463
YES	MSCOCO	validation set	0.1646048292

The results demonstrate that the joint learning of CNN and LSTM has a significant improvement in performance. To make full use of the MSCOCO dataset, we jointly train a model using all of the examples in MSCOCO dataset, not just using the *train split*. The results are shown in Table 3. It is demonstrated that more data can result in better performance.

At last, we fine tune the jointly learned model on different datasets to get a more suitable model for the natural language caption generation task. We use the model trained on all of the examples on MSCOCO as the *baseline*, and fine tune the model on different datasets. We firstly fine tune our model on a very small dataset. In this experiment, we use 1500 images and their sentences in the validation set as a training set and use the remaining 500 images to evaluate the performance of the fine tuned model. The results are shown in Table 3. We also fine tune the baseline model on a big dataset, which is the combination of Flickr30K and Flickr8K. We can see that fine tuning on a big dataset can get a better performance. We use the model obtained in the previous step to generate image captions on all the 510123 target images. This time, we manually select 1000 satisfactory pairs of image and generated sentence from all the generated captions and add them to the fine-tuning dataset. As shown in Table 3, the results show that this pipeline has the best performance. We use the model fine tuned in the combination of Flickr8K, Flickr30K and the selected 1000 examples as our final model.

Figure 3 is the illustration of the generated image captions by different models. The results qualitatively demonstrate that our final model can generate more satisfactory captions that reveal the content of the corresponding images.

Table 3. The performance of fine tuning the model on different datasets

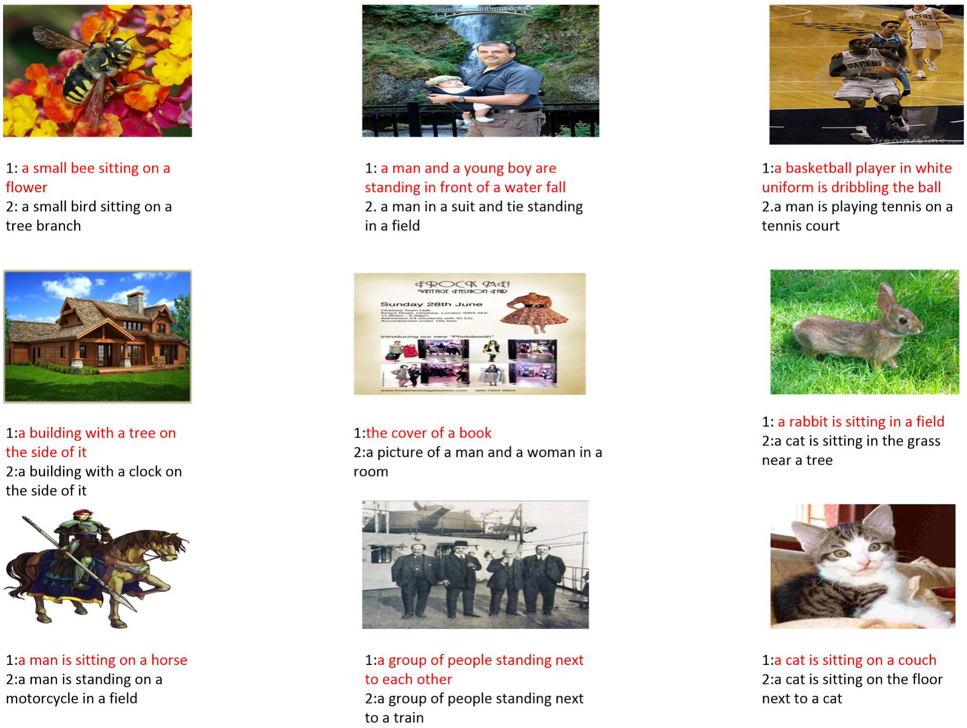
Train	Test	Accuracy
Only the training split of MSCOCO	validation set	0.1368988204
MSCOCO(baseline)	validation set	0.1646048292
FT(3/4 validation set)	1/4 validation set	0.1159403729
FT(Flickr8k+Flickr30k)	validation set	0.1738749916
FT(Flickr8k+Flickr30K+1000_selected)	validation set	0.2042696662

We submitted four runs in the natural language caption generation task:

id_sentence2.txt is our baseline method. The model is trained only using all the examples in the MSCOCO dataset. The *median score* (provided by the server) of the generated sentences is 0.1676 (The model is used twice to generate both the two runs, so ***id_sentence3.txt*** is the same as ***id_sentence2.txt***).

id_sentence.txt is the results generated by the model which is firstly trained only using the examples in the MSCOCO dataset and then fine tuned on the combination of Flickr8K and Flickr30K. The median score of the generated sentences is 0.1710.

id_sentence4.txt is the results generated by our final model which is firstly trained only using the examples in the MSCOCO dataset and then fine tuned on the combination of Flickr8K, Flickr30k and the manually selected examples. And the median score of the generated sentences is 0.1711. This submitted run is the best of our submitted runs and also the best one for natural language caption generation of ImageCLEF 2016.



1: a small bee sitting on a flower
2: a small bird sitting on a tree branch

1: a man and a young boy are standing in front of a water fall
2. a man in a suit and tie standing in a field

1:a basketball player in white uniform is dribbling the ball
2.a man is playing tennis on a tennis court

1:a building with a tree on the side of it
2:a building with a clock on the side of it

1:the cover of a book
2:a picture of a man and a woman in a room

1: a rabbit is sitting in a field
2:a cat is sitting in the grass near a tree

1:a man is sitting on a horse
2:a man is standing on a motorcycle in a field

1:a group of people standing next to each other
2:a group of people standing next to a train

1:a cat is sitting on a couch
2:a cat is sitting on the floor next to a cat

Fig. 3. An illustration of the generated image captions. Sentence 1 is generated by our final model. Sentence 2 is generated by our baseline model.

4 Conclusions

After performing the experiments above, we get the following conclusions. By learning the parameters in CNN and LSTM, the performance of the model can be greatly improved. When we just change the parameters of LSTM, the accuracy on the test set is 0.133. However, when we change parameters in both neural network, the accuracy is 0.165. Secondly,more data can result in better performance. We train a model only on the training split of the MSCOCO dataset and the score of the generated sentences is 0.137. However, when we use all the data in the MSCOCO dataset, the score is 0.165. Thirdly, when fine-tuning the model only using a small dataset, the result we get is worse. When we fine tuned our model only using 3/4 of the validation set, the result on the remaining 1/4 of the validation set is 0.116 which is worse than the model before fine-tuning. However, the datasets we use to train our model don't include all the concepts of the ImageCLEF 2016, so some sentences predicted by our model might be weird.

5 Acknowledgments

This work was supported in part by the National Basic Research Program of China (973 Program): 2012CB316400, the National Natural Science Foundation of China: 61532018, 61322212, the National Hi-Tech Development Program (863 Program) of China: 2014AA015202. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

References

1. Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, Emmanuel Dellandrea, Robert Gaizauskas, Mauricio Villegas, and Krystian Mikolajczyk. Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In *CLEF2016 Working Notes*, CEUR Workshop Proceedings, Évora, Portugal, September 5-8 2016.
2. Mauricio Villegas, Henning Müller, Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, Andrew Gilbert, Luca Piras, Josiah Wang, Fei Yan, Arnau Ramisa, Emmanuel Dellandrea, Robert Gaizauskas, Krystian Mikolajczyk, Joan Puigcerver, Alejandro H. Toselli, Joan-Andreu Sánchez, and Enrique Vidal. General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science. Springer International Publishing, 2016.
3. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. CVPR, 2011.
4. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. ECCV, 2010.
5. Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1,ACL '12*, pages 359–368, Stroudsburg,PA,USA, 2012. Association for Computational Linguistics.

6. R. Mason and E. Charniak. Nonparametric method for datadriven image captioning. In *ACL*, 2014.
7. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015.
8. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044, 2015.
9. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. DollÂŽar, and C. L. Zitnick. Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312, 2014.
10. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
11. P. Young, M. Hodosh A. Lai, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2014.
12. M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013.
13. M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.