ISIA at the ImageCLEF 2017 Image Caption Task

Sisi Liang, Xiangyang Li, Yongqing Zhu, Xue Li, 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
{sisi.liang, xiangyang.li, yongqing.zhu, xue.li, shuqiang.jiang}@vipl.ict.ac.cn

Abstract. This paper describes the details of our methods for participation in the caption prediction task of ImageCLEF 2017. The dataset we use is all provided by the organizers and doesn't include any external resources. The key components of our framework include a deep model part, an SVM part and a caption retrieval part. In deep model part, we use an end to end architecture with Convolutional neural network (CNN) and a Long Short-Term Memory (LSTM) to encode and decode images and captions. According to the statistics of training dataset, we train different models with different lengths of captions. Then in SVM part, we use Support Vector Machine (SVM) to determine which model to use when generating the description for a test image. In this way, we can combine these models from the previous deep model part. In caption retrieval part, we use the image feature extracted from CNN and apply Nearest Neighbor method to retrieve the most similar image with caption in the training dataset. The final description is the aggregation of the generated sentence and the caption retrieved from the training dataset. The best performance of our 10 submitted runs ranks the 3rd in group which doesn't use external resources.

Keywords: Convolutional neural network, Long Short-Term Memory, Support Vector Machine, Nearest Neighbor, Image caption.

1 Introduction

Over the past few years, there has been a huge interest in the task of automatically generating captions for images. It is interesting to see how a machine can solve this problem which is very easy to a person. Many progress [2, 4, 6, 11, 12] has been achieved after so many years endeavor and research.

There are three main approaches to generate image caption: one is using templates [2, 6], these methods rely on detectors and map the output to linguistic structures. Another approach is using language models, like many widely used deep models [11, 12]. This method may yield more expressive captions because it can overcome the limitation of templates. Many deep learning architectures use Long Short-Term Memory (LSTM) based Recurrent Neural Networks (RNN) as language models. The third approach is caption retrieval and recombination [2, 4]. Instead of generating new captions, these methods retrieve captions based on training data.

The purpose of ImageCLEF 2017 caption task [5, 14] is interpreting and summarizing the insights gained from medical images. So the dataset is very different from previous datasets like MSCOCO [7] or Flick8K [4] and has its own characteristics. One is that almost half of the captions are more than 20 words and the longest caption reaches 606 words. Therefore it is a tough task to fully use the semantic information of data. Another is that some of images consist of several small images, like CT images from different perspectives, photos before and after treatment. It is hard to detect the internal relation of items in image and to reflect the change of small images.

Our method can be separated into three parts. The first part is deep model part. This part bases on deep model proposed by Vinyals [11]. The model is an end to end architecture using Convolutional Neural Network (CNN) for image encoding and Long Short-Term Memory (LSTM) based Recurrent Neural Network for sentence decoding. We divide the training dataset into three parts according to the length of captions and train three different models with different lengths. The second part is SVM part. In this part a three class SVM classifier is trained to determine which model to use in predicting a caption. The third part is caption retrieval part. We use the caption retrieval approach and apply Nearest Neighbor method to retrieve a most similar image. Then the caption of this image will be used as a supplement in the final generated caption.

2 Methods

The main structure of our method can be seen in the Fig. 1. The structure is composed of three parts: deep model, SVM and caption retrieval part. Each part will be introduced in detail below.

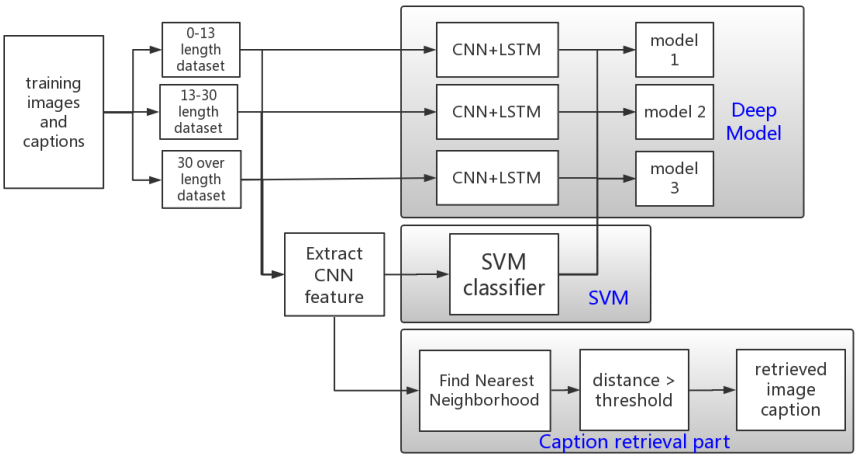

Fig. 1. Structure of our method

Preprocessing. In the preprocessing of data, we notice that training dataset is different from the other well-known Image Caption dataset [4, 7]. The training data is medical images and one caption for one image. Some captions in the training data are very long. The statistics of the sentence length in training dataset are shown in the Table 1. From Table 1 we can see that length of captions under 20 only accounts for 23.73%. We implement experiments about the influence of sentence length. The result in Table 4 shows that combine different models of different sentence lengths can achieve better result. So we divide the dataset to three parts according to the length of captions. The sentence length of each subset are 0-13, 13-30 and 30 over. For each subset of dataset, we train a different deep model based on different lengths. Sentences longer than the max length will be clipped to keep the sentence max length, the max length we use can be seen in the Table 2. The reason we choose length 0-13, 13-30 and 30 over as the subsets is that each subset accounts for around 1/3 of the total sentences.

Table 1. The statistics of the sentence length in training dataset

Length of sentences	Percentage of the total sentences
0-10	23.73%
10-20	26.51%
20-30	15.98%
30-40	10.16%
40-50	6.88%
50-60	4.68%
60-70	3.31%
70-80	2.41%
80-90	1.75%
90-100	1.28%
100 over	3.31%

Table 2. Max sentence length in models

Data sentence length	Max length used
1-13	13
13-30	30
30-606	100

Training deep model. Deep model contain the following parts: Convolutional Neural Network for image encoding, Long-Short Term Memory based Recurrent Neural Network (LSTM-RNN) for sentence encoding and decoding. We use a pre-trained VGGNet [9] for image feature extraction and each image will be transformed into a 4096-dimensional vector. Then we train a LSTM-RNN for encoding and decoding sentences. The LSTM-RNN implementation is based on the NeuralTalk2[1] project. As we divide the training data to three subsets in the preprocessing, we train three differ-

[1] https://github.com/karpathy/neuraltalk2

ent deep models in training stage. For each deep model, the input training data is a subset of dataset after preprocessing and the max length we set are shown in Table 2. Other initial parameters are the same in all the three deep models.

Training SVM classifier. We try to use SVM classifier combine the three models together to generate captions of images. So in SVM part, we attempt to train a SVM classifier which can predict the three kinds of sentence length 0-13, 13-30 and 30 over. We use all the images from training dataset and extract image features from fc7 layer in VGGNet to train a three class SVM classifier. This SVM classifier will be used to determine which deep model to use in the prediction stage. The accuracy of this SVM classifier in predicting the validation data is shown in Table 3.

Table 3. The accuracy of three class SVM classifier

Length	Precision	Recall	F1-score	Support
1-13	0.62	0.61	0.62	3343
13-30	0.44	0.47	0.45	3047
30-606	0.71	0.68	0.70	3610
Avg/Total	0.60	0.59	0.60	10000

Caption retrieval: using Nearest Neighbor method. The performance of the model which only use deep model and SVM classifier is not optimal. Therefore, we attempt to use Nearest Neighbor method to retrieve the most similar image in caption retrieval part. If the Euclidean distance between the image CNN feature of predicted image and the image CNN feature of retrieved image is larger than a threshold, we will use the caption of retrieved image as a supplement in the final caption. The performance of model get an optimal when the threshold are 300 in CNN feature and 25 in normalized CNN features after many trials. In table 5, we can see that there is an improvement in the performance after applying the NN method.

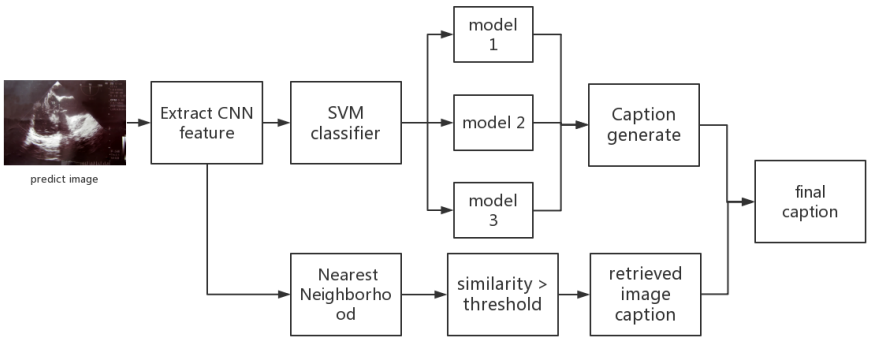

Fig. 2. An illustration of prediction

Prediction. As shown in Fig. 2, the input is an image and output is the generated sentences in the prediction stage. First, we extract the CNN feature from input image, then use the SVM classifier to determine which model to use. Next we use the trained deep model to generate a caption of this input image. Finally we apply the Nearest Neighbor method to retrieve a closest caption. If the Euclidean distance between image CNN feature of input image and image CNN feature of retrieved image is larger than a threshold, this caption will be used as a supplement of the caption.

3 Experiments and Submitted Runs

The dataset we use in our method is all from the provided dataset. None of external datasets is used in our experiments. We first divide the ImageCLEF 2017 caption training data to three parts according the sentence length of training data captions. Then we use the divided subset to train three different deep models based on the CNN and LSTM model. After that, a three classes SVM classifier is trained, using 4096 dimensions vector extracted from VGGNet fc7 (fully connected) layer as feature and using label 0, 1, 2 represent the three kinds of sentence length. Finally, we use NN (Nearest Neighbor) method to find the most similar image in training dataset. The feature is the same feature used in training the SVM classifier and it is normalize from 0 to 1. The distance function is the Euclidean distance. If the similarity is larger than a threshold, we will use this retrieved image caption as a supplement in final caption.

Divide dataset with different sentence length. We conduct experiments to find out that whether the sentence length affects the final performance. Different sentence lengths and their performance in the validation data are shown in Table 4. We use the BLEU [8], METEOR [1], ROUGE [3] and CIDEr [10] scores based on the coco-caption code [13][2] when measure the performance of different models. All the training dataset are divided into different subsets with different sentence length in the preprocessing. When sentence length are 20, 45 and 60, we only use the deep model with CNN+LSTM to generate captions. SVM_two and SVM_three are models which use both deep model and SVM classifier. The results demonstrate that the length of sentence has a significant impact on performance. Training three different sentence length models and using SVM classifier can result in better performance.

Table 4. Performance of different sentence length models

Sentence length	BLEU -1	BLEU -2	BLEU -3	BLEU -4	METEOR	ROUGE_L	CIDEr
20	0.042	0.018	0.007	0.003	0.028	0.088	0.032
45	0.047	0.022	0.009	0.004	0.033	0.105	0.034
60	0.058	0.028	0.013	0.006	0.036	0.110	0.055
SVM_two(length =20,45)	0.098	0.045	0.020	0.009	0.040	0.107	0.044
SVM_three(lengt	0.134	0.061	0.026	0.012	0.043	0.113	0.053

[2] https://github.com/tylin/coco-caption

h =13,30,100)

Caption retrieval: using Nearest Neighbor method. We notice that although using SVM has some improvement in performance, the deep model cannot achieve an optimal result. So we conduct another experiment to explore whether using Nearest Neighbor method to retrieve image caption can help improve the performance.

We use normalized features and non-normalized features for NN method to compare the performance. And the results are shown in Table 5.The results demonstrate that adding the retrieved caption can lead to a better result and the result will be further improved when using normalized features.

Table 5. The performance with and without Nearest Neighbor method

Sentence length	BLEU -1	BLEU -2	BLEU -3	BLEU -4	METEOR	ROUGE_L	CIDEr
SVM_three	0.134	0.061	0.026	0.012	0.043	0.113	0.053
Nn+SVM_three(thresh <3000)	0.144	0.068	0.035	0.020	0.058	0.127	0.049
Nn_normalized+SVM _three (thresh<25)	0.171	0.099	0.065	0.047	0.077	0.144	0.095

Table 6. The performance of submitted runs

Submitted runs	Mean BLEU score
test_13_svm_3_nn_dist_25_normal_noUNK	0.2600
test_5_svm_nn_dist_3000_nounk_modified_2	0.2507
test_12_svm_3_nn_dist_25_normal	0.2454
test_11_svm_2_nn_dist_25_normal_noUNK	0.2386
test_10_svm_2_nn_dist_25_normal	0.2315
test_9_svm_three_nn_3000_noUNK	0.2240
test_6_svm_three_parts	0.2193
test_2_svm_two	0.1953
test_1_wc5sl70	0.1912
test_8_svm_two_remove_UNK	0.1684

test_13_svm_3_nn_dist_25_normal_noUNK: use three classes SVM classifier and NN method. In NN method, we use normalized features and threshold is 25. Besides, we remove the UNK which is used to represent those word account didn't achieve five times in training dataset.

test_5_svm_nn_dist_3000_nounk_modified_2: use two classes SVM classifier and NN method. In NN method, we didn't use normalized features and threshold is 3000. As mention before, this run also removes UNK.

test_12_svm_3_nn_dist_25_normal: use three classes SVM classifier and NN method. We use normalized features and threshold is 25 when applied NN method. The difference is that this run didn't remove UNK.

test_11_svm_2_nn_dist_25_normal_noUNK: use two classes SVM classifier and NN method. Normalized the image feature and threshold is 25. UNK is remain in this run.

test_10_svm_2_nn_dist_25_normal: use two classes SVM classifier and NN method. Use normalized features and threshold is 25. UNK is removed.

test_9_svm_three_nn_3000_noUNK: use three classes SVM classifier and NN method. In NN method, we didn't use normalized features and threshold is 3000. UNK is removed.

test_6_svm_three_parts: only use three lengths classes SVM classifier and deep model.

test_2_svm_two: only use two lengths classes SVM classifier and deep model.

test_1_wc5sl70: only use CNN+LSTM deep model.

test_8_svm_two_remove_UNK: only use two lengths classes SVM classifier and UNK is removed.

We have submitted ten runs in the caption prediction subtask and the performances are shown in the Table 6. The performance of the best run is 0.2600 in Mean BLEU score. Compared to other runs which doesn't including SVM classifier or caption retrieval, the performance has been greatly improved.

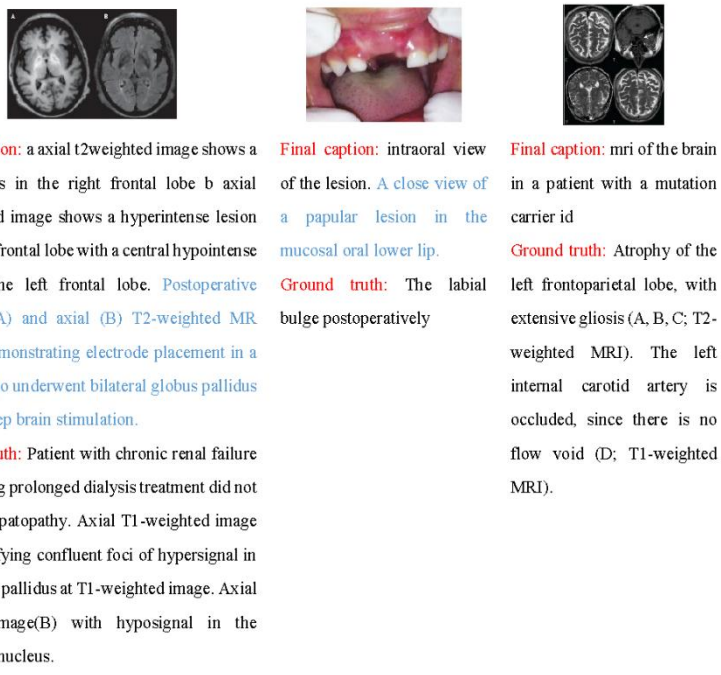

Fig. 3. Some examples of generated caption and ground truth from validation dataset. The blue sentence is the caption using Nearest Neighbor method.

As shown in the Fig.3, the model can generate different lengths of captions according to the picture. The first example shows that the caption retrieved by NN method is similar to the caption generated by deep model and is also relate to the ground truth. Besides, the image of the first example contains images A and B. We are delighted to see the model can learn the pattern and generate a caption include alphabetic number and information in the two images. The third example shows that the distance between input image and the similar image retrieved using Nearest Neighborhood method is far from threshold, so the retrieved caption will not be used in the final caption.

4 Conclusions

In this paper, we describe our method in ImageCLEF 2017 caption prediction subtask. We use statistics of training dataset and divide the training dataset into three parts. Then in the training stage, we train three deep learning models and use CNN and LSTM to generate natural language sentences. A three classes SVM classifier is trained at the same time to determine which deep model to use when predicting image caption. Besides, Nearest Neighbor method is also applied to retrieve a similar image and its caption in the training data as a supplement in final caption. After performing the experiments above, we get the following conclusions. Firstly, the sentence length parameter in training can affect the performance. By training separately models and using SVM classifier, we achieve a better result compared to the model only use CNN+LSTM. Secondly, the similar image can provide useful information in caption. After applying the Nearest Neighbor method to retrieve a similar image and caption, the performance of the model can be greatly improved. However, we limit the sentence length and remove the words that only appear one or two times during the training. The removed words in the training dataset cannot be made full use of. This makes some generated captions lack of readability. In addition, the generated language model is too simple to generate a complex and fully descriptive caption. Compared with other participants in this task, the best performance of our 10 submitted runs is 0.2600 in Mean BLEU score, ranks the 3rd in group which doesn't use external resources.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and 61322212, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

References

1. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: The Workshop on Statistical Machine Translation. pp. 376–380 (2014)
2. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: European Conference on Computer Vision. pp. 15–29 (2010)
3. Flick, C.: Rouge: A package for automatic evaluation of summaries. In: The Workshop on Text Summarization Branches Out. p. 10 (2004)
4. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. AI Access Foundation (2013)
5. Ionescu, B., M¨uller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.T., Dicente Cid, Y., Eickhoff, C., Garcia Seco de Herrera, A., Gurrin, C., Islam, B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: Information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017. Lecture Notes in Computer Science, vol. 10456. Springer, Dublin, Ireland (September 11-14 2017)
6. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: understanding and generating simple image descriptions. In: Computer Vision and Pattern Recognition. pp. 1601–1608 (2011)
7. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.L.: Microsoft coco: Common objects in context 8693, 740–755 (2014)
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Meeting on Association for Computational Linguistics. pp. 311–318 (2002)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
10. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. Computer Science pp. 4566–4575 (2014)
11. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator pp. 3156–3164 (2014)
12. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. Computer Science pp. 2048–2057 (2015)
13. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.:Microsoft coco captions: Data collection and evaluation server. Computer Science (2015)
14. Eickhoff, C., Schwall, I., Garc´ıa Seco de Herrera, A., M¨uller, H.: Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomed-ical images. In: CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings,CEUR-WS.org<http://ceur-ws.org>, Dublin, Ireland (September 11-14 2017)