Show And Tell: Image Caption Generator Based On CNN And RNN

Yue
yzk4v@mail.umkc.edu
University of Missouri-Kansas city
School of Computing & Engineering
Kansas City, MO 64110, USA

Abstract

In recent years, information on the Internet has been increasing a lot, especially the image data. However the image is unstructured data and most of the are without captions, which make it hard for searching. Generating image caption automatically by artificial intelligence and deep learning method is a good way to make the image be recognized. In recent years, the development of computer power and the big data application make the machine learning and deep learning grow fast. Convolutional neural network is usually used for visual recognition, recurrent neural network is used for human language analysis some times. The show and tell model combine the two neural network for generating captions for images.

In this paper, we will introduce the show and tell model, which is combined with the CNN and RNN. In NLP(nature language processing), RNN is use to translate, first it encode the source language and then decode it into target language. For the case of image captioning, CNN is used to encode the image into represent vector, on the other hand RNN is used to decode the vector to caption.

We will also show the result of the model and evaluate it in metrics score: BLEU((bilingual evaluation understudy), CIDER(Consensus-based Image Description Evaluation), METEOR(Metric for Evaluation of Translation with Explicit ORdering) and ROGUE(Recall-Oriented Understudy for Gisting Evaluation). These metrics are mostly used in nature language processing to evaluating automatic summarization and machine translation. Here we use these metrics to evaluate the caption of the image comparing with the human being caption.

Introduction

Since the theme of project is animals, we don't need to train the whole dataset for the model, so we select several images and the corresponding captions that match the
theme to train the model.
First we prepare the dataset for the model, including image and corresponding captions, then train the model with the pre-trained checkpoint, finally generate caption for the test data.
Then we will show the result that produced by these model and approaches. First the proposed work of the network respectively, then implement them by combining these model, finally we evaluate the result and compare with other related work.

Related work

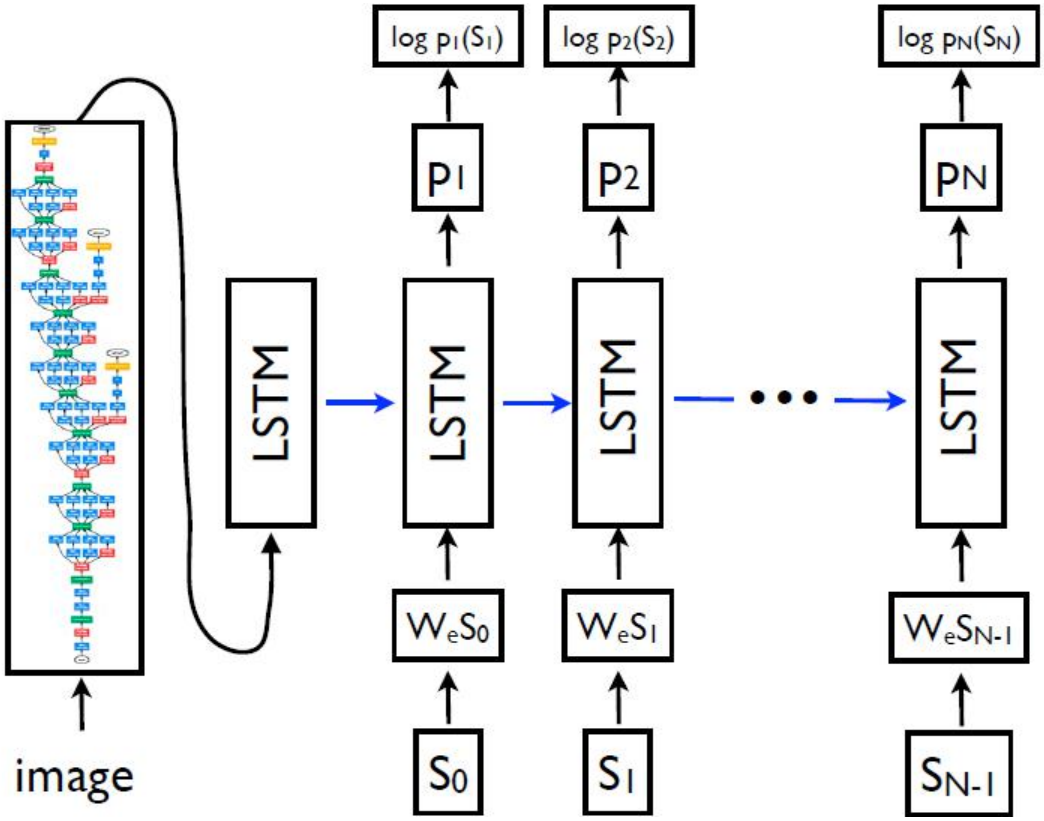

Figure 1 show and tell model architecture

There are many approaches and model in deep learning that affect the accuracy of the result. Some neural network such as CNN[1], RNN[2], RCNN[3]. And there are many usage of these deep learning network like NLP[4], show and tell[5]. There are also some improvement like SIFT[6] for CNN, and LSTM[7] for RNN.

[8] incorporate two model to equip classification of low resolution pictures by combining convolutional high resolution and convolutional grained analysis.

[9] fix the model to resolve the problem of domain shift and learn the regression prototype by a zero-shot method.

[10] transfer unseen category to implanted space of seen category by fake labels without data loss.

Model

The show and tell model is showed in figure 1. A convolutional neural network can create an embedding which is a dense feature vector. And this vector is described as a feature, input to other network and algorithm.

For this image caption generator model, the vector is a image representation and input to the RNN LSTM.

In show and tell model, we use LSTM for the RNN, it's also usually used in transitory dependence problem. It can grasp information from former states to update the present prediction.

The show and tell model can recognize the object in the image, and show the relationship between them, then describe it using nature language. It is an

Figure 2 result of CNN model

encoder-decoder NN model. First it encodes the image to a representation, then it decodes the representation to a caption.

In the encoding step, it uses CNN, CNN can embed the image to fixed-length vector, this vector will became the input of the decoding step.
In the decoding step, it uses RNN with LSTM to generate the representation to natural language caption.
Machine translation is that input a sentence, make the probability that the translation is correct maximal. So we can use the same approach, input a image, use CNN to generate the object, and use RNN to "translate" it into description by maximizing the probability of the correction of description.

Implement

In this paper, we choose animal to be the object and use the data set of Flickr8k for generating the caption.Flickr8k has 8000 pictures and every picture has 5 different captions from human being. The caption clearly describe the important object and interaction between them. All the images in this data set are selected from six separate Flickr image album, it doesn't include famous people or places, and chosen by human to show many different stuff and case. Since 5 captions of each image describe the same picture, the sentence are similar and has many words in common. We choose the picture of animals as our own dataset and separate as training and test dataset.

First, we show the achievement of each model such as CNN, RNN and NLP, and evaluate the result respectively. Then we combine the work to the show and tell model and evaluate the final result accuracy.

Convolutional neural network

In order to recognize the object in the image, we build the CNN model by the pre-trained model and animal training data set. And use the model on the test data set to generate the result by the model. As showed in figure 2, the result of CNN is not as exact as we expect, the main reason of it is the lack of the training data. We will improve it by using more data of animal, which also means more time and computational space at the same time, to train the network.

Recurrent Neural Network

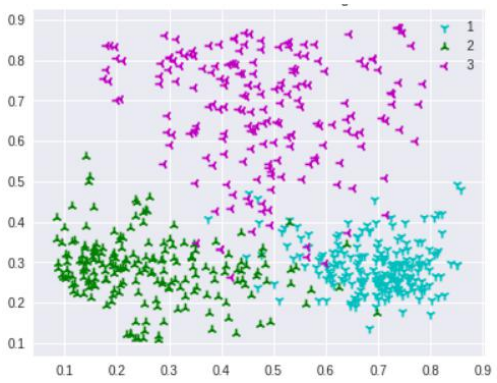

Figure 3

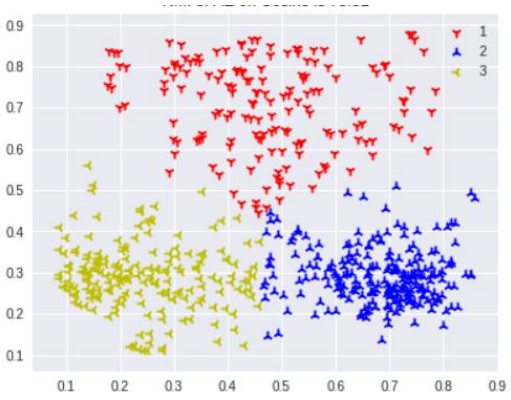

Figure 4

We train the RNN model by the human caption of the image, and use the key word to test the result. For example, the training data including many captions about the animal, such as “Two dogs are wrestling in the grass .” and “a cat sits alone in dry grass .”, when we test the result, we input some word such as “dog”, “cat”, “grass”, the model will output the whole sentence.

The simple RNN model can only output the sentence that already exit in the data set, which makes the result inaccurate because it’s not normal for two pictures that have the same caption. For example, there is a sentence “A dog is playing ball on the field”, we input “dog”, “ball”, “field” and expect the sentence “A man and a child on the field is playing ball with a dog. ”, but since it’s not in the training caption dataset, it still output the first sentence.

Image classification

We use unsupervised learning to classify the image of different objects. Since unsupervised learning is classify the similar image without the label, we only use the image to classify without caption. We encode the images and use K-Means[11] for classifying.

Figure 3 shows the representations in 2 dimensions, it’s generated by auto-encoder and the categories of images used is dog, cat and bird. In figure 4 we classify the representation by using K-Means. We can see the result is acceptable for human experience.

Show and tell

In the result of show and tell model(figure 5), we can tell that the caption is brief and lack of details, some of them are wrong and some are the same. The main reason is lack of training data. The image data is not enough so the CNN model can’t recognize the object, the caption data is lack so it can’t generate more detail and more variety.

In table 1 we show the metrics of our model and human caption. BLEU[12] (bilingual evaluation understudy) exam the capacity of the message. The correlation between two text that the closer of the text, the better it is. METEOR[13] (Metric for Evaluation of Translation with Explicit Ordering) is

1. A dog runs down stairs to grass.

2. A bird standing on hand and eating

3. A black dog on the beach

4. A dog running on a field

5. A dog play on the field

6. A dog running on a field

7. Two dogs running on white field

8. A black dog on the field

9. A dog is jumping

Figure 5 Some example of the captions. We can see that 5 is wrong, 4 and 6 are totally the same (although they are both right)

Metrics	BLEU-4	METEOR	CIDER	ROGUE
Show and tell	26.5	22.3	84.6	0.503
Human	22.8	25.4	86.1	0.496

Table 1: the metrics on show and tell model and human being caption

based on accuracy and recall to evaluate the similarity of two text, with stemmer and synonymy matching. ROUGE[14] (Recall-Oriented Understudy for Gisting Evaluation) compares the translation or summary with a reference of translation or summary CIDEr[15] (Consensus-based Image Description Evaluation) is a novel paradigm to exam the caption of image by human consensus. It shows the average metrics scores of the caption generated by model, in order to compare the accuracy with the real human caption, we select other

human caption and calculate the metrics scores.

Conclusion

As the result showed above, it still need to improve. Since the data set isn't enough, the recognizing is not accurate, and the RNN model need to improve so it can generate the new caption out of the training dataset. However, although the result of show and tell model is easy and inaccurate sometimes, it still can work for simple animal image and summarize the picture.

The future work is to implement another CNN such as recurrent-CNN[3] and more data to generate more accurate result. This will require more computational space and time, so using the distributed big data tools is also one of the future work.

Reference

[1]Keiron O'Shea and Ryan Nash, "An Introduction to Convolutional Neural Networks", Dec 2015

[2] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS", Mar 2013

[3] Ming Liang, Xiaolin Hu, "Recurrent Convolutional Neural Network for Object Recognition", 2015

[4] Julia Hirschberg, Christopher D. Manning, "Advances in natural language processing", April 5, 2019

[5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge", Sep 2016

[6] Tony Lindeberg, "Scale Invariant Feature Transform", Stockholm, Sweden, 2015

[7] Haşim Sak, Andrew Senior, Françoise Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition", Feb 2014

[8] D. Cai , K. Chen , Y. Qian , J.-K. Kämäräinen , Convolutional low-resolution fine–grained classification, Pattern Recognit. Lett. 119 (2019) 116–171 .

[9] C. Luo , Z. Li , K. Huang , J. Feng , M. Wang , Zero-shot learning via attribute re- gression and class prototype rectification, IEEE Trans. Image Process. 27 (2) (2018) 637–648 .

[10] Y. Guo , G. Ding , J. Han , Y. Gao , Zero-shot learning with transferred samples, IEEE Trans. Image Process. 26 (7) (2017) 3277–3290 .

[11] Khaled Alsabti, Sanjay Ranka, Vineet Singh, "An efficient k-means clustering algorithm", 1997.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", 2002

[13] Satanjeev Banerjee and Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", 2005

[14] Lin, Chin-Yew, "ROUGE: a Package for Automatic Evaluation of Summaries.", 2004

[15] Ramakrishna Vedantam, C. Lawrence Zitnick and Devi Parikh, "CIDEr: Consensus-based Image Description Evaluation", 2015