Image Captioning based on Recurrent Neural Network Model

Heng Qiao
Department of Electrical and Computer Engineering
University of Florida
Email:

Tong Shao
Department of Electrical and Computer Engineering
University of Florida
Email: stlm1991@ufl.edu

Yichen Liang
Department of Electrical and Computer Engineering
University of Florida
Email:

***Abstract*—Image captioning is the task to automatically describe the content of an image, which is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this course project, we plan to develop an image captioning system based on the recurrent neural network(RNN) Model. As the most widely used scheme, two neural networks are introduced. The first one is a pre-trained convolutional neural network (CNN) that converts the image into feature vectors, such as the VggNet, ResNet and Inception. Serving as the core of the scheme, the second one adopts the recurrent neural network (RNN) model. It takes the image feature as the input and generate the word vectors of a sentence (caption). The model is trained to maximize the likelihood of the target description sentence given the training image. Based on this, we aim at implementing some assistant techniques to further improve the performance, such as the semantic structures and more optimized neural network structures. This scheme will be implemented in Tensorflow. Experiments will be conducted on public dataset MS COCO and related evaluation scores such as the BLEU-4 score will be provided. Also, the final report, slides and other related material will be prepared as well.**

I. Introduction

Deep learning technology has been widely used in many aspects [1], such as face recognition, natural language processing and etc. It has shown remarkable learning ability in many areas. Among them, the most common form of machine learning, deep or not, is supervised learning [2]. Supervised learning requires labeled training data and typically outputs a label when given a new input [3]. To deal with inputs with correlations in time such as an sentence, recurrent neural network (RNN) [4], typically implemented with long short-term memory (LSTM) units, is employed.

Image captioning is the procedure to automatically describe the content of an image, as shown in Fig. 1. And there has been a recent surge of interest in developing models that can generate captions for images or videos. Most of these approaches learn a probabilistic model of the caption, conditioned on an image or a video. Most recent work on visual captioning first extracts features from an image, obtaining a fixed-length vector representation of a given image or video. Then a language, usually a recurrent neural network (RNN), typically implemented with long short-term memory (LSTM) units [5], is employed to generate a caption while the image's vector representation is the input [6].

Fig. 1. Examples of image captioning.

Though good results have been achieved, most methods are still far away from human performance and the models usually don't consider the grammar rules. Recent work shows that adding explicit high-level semantic concepts of the input image/video can further improve visual captioning. For instance, in [7], a model of semantic attention is proposed which selectively attends to semantic concepts through a soft attention mechanism.

goals, deliverables, overview architecture, use cases, modules/subsystems, flowcharts, algorithms, potential results (performance metrics, success evaluation criteria, mock-up result figures x,y,z curves etc.).

II. Project Description

We aim at designing a typical image captioning system while adding some techniques to further enhance the performance. The detailed project description is as following.

A. Goals

Our main goal is to implement an image captioning system which could automatically describe the content of an image as described before. The scheme will be implemented in

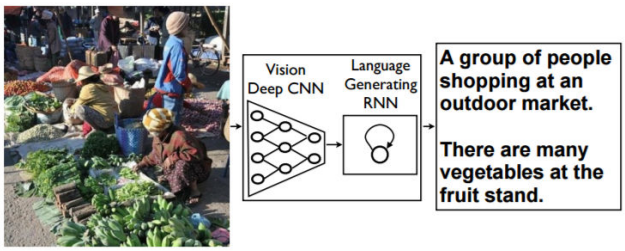

Fig. 2. Overview Architecture of our image captioning scheme.

Tensorflow and experiments will be conducted on MS COCO dataset.

The syetem should be able to produce reasonable captions for input images. And the scores (based on popular metrics) on the MS COCO datasets should be close to, i.e. of the same level as the state-of-the-art methods.

B. Deliverables

We will implement the scheme in Tensflow. Thus we will provide the codes, the pre-trained model as well as the final report, slides.

C. Overview Architecture

As shown in Fig. 2, the basic architecture has two parts. The first one is a pre-trained convolutional neural network (CNN) that converts the image into feature vectors, such as the VggNet, ResNet and Inception. And the second one adopts the recurrent neural network (RNN) model. It takes the image feature as the input and generate the word vectors of a sentence (caption).

D. potential results

As described before, this scheme should produce reasonable captions for input images. We will present some results of images with typical scenes such as sports activities, natural scenes, indoor scenes and etc.

Meanwhile, we will test the scheme on the widely used MS COCO dataset. And the scores based on several popular metrics will be provided including the BLEU, METEOR and CIDER.

We expect our scheme will produce pretty good results on the MS COCO dataset. It should be close to, i.e. of the same level as the state-of-the-art methods such as the Show and Tell [6].

E. Weekly Progress Plan

The schedule of each task is as following:
1) Scheme design
March 4-March 11
2) Code implementing
March 12-March 31
3) Final Report Writing
April 1-April 15
4) Slides and Presentation
April 16-April 23

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[5] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[7] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.