2018 Fall - IE 534 Final Project Proposal

1. Team member

Chunlei Liu(chunlei2)

Hanwen Hu (hanwenh3)

Renjie Wei(renjiew2)

Xinyan Yang(xinyany2)

2. Project name

Scene/image description as in “Show and Tell: A Neural Image Caption Generator”

3. Brief introduction

In this paper, they present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. The model is often quite accurate, which they verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, their approach yields 59, to be compared to human performance around 69. They also show BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, they achieve a BLEU-4 of 27.7, which is the current state-of-the-art.

In conclusion, they present NIC(Neural Image Caption Generator) --- an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. NIC is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence.

4. Project plan

4.1. Datasets

COCO (http://cocodataset.org/#captions-2015) is a large-scale object detection, segmentation, and captioning dataset. This captioning challenge is part of the Large-scale Scene Understanding (LSUN) CVPR 2015 workshop organized by Princeton University. We choose the COCO train2014 data(Including 2014 Train images [83K/13GB] , 2014 Val images [41K/6GB] and 2014 Train/Val annotations [241MB]) for the training set and COCO test2014 data(including 2014 Test images [41K/6GB] and 2014 Testing Image info [1MB]) for the testset. API will be used to manipulate the data after downloading.

4.2. Plan for code

The coding task is divided into the following parts:

1) Dataset Preparation (Image+Sentence, Corpora) - ***(Together)***
    - Overview the training and testing dataset
    - Unify input image sizes and choose initializing corpus
    - Initialize parameters in all models
2) Image Representation - Deep CNN (Dropout, ensembling) - ***Hanwen Hu***
    - Build a Deep CNN framework to map image tensors to vectors
    - Initialize on a pre-trained model
    - Involve techniques such as data augmentation, batch normalization, dropout, ensembling models
3) Word Representation - Word embedding model - ***Chunlei Liu***
    - Tokenize each sentence with NLTK in python.
    - Use word2vec and glove to embed those tokens
4) Sentence Generator - LSTM network - ***Xinyan Yang, Renjie Wei***
    - Combined with Deep CNN image and word embeddings
    - Train the LSTM network to predict each word of the sentence after it has seen the image as well as all preceding words
5) Model Evaluation. - ***TBD***
    - Perform a series of experiments to assess the effectiveness of the model using several metrics (e.g. BLEU score, METEOR, Cider etc.) compared with prior art as well as human evaluation
    - Analyze generation result and generation diversity as the paper does

Weekly meeting is held on every Thursday in the following weeks. Group members will report their progress, and set their goals for next week.

4.3. Computations and computational hours

This project focuses on generating sentences on test set that has the accuracy no less than the original paper by Vinyals et.al. on 2015. If possible, we will try to make improvement by introducing methods from other papers on Neural Image Captioning.
For now, we plan to implement Deep CNN, Word Embedding and LSTM networks in this project.

It's estimated that 1000~1200 GPU hours will be needed for the whole project.