Proposal for GSOC 2018

Project Name:

RED HEN LAB --Chinese Pipeline

Abstract

This project is roughly divided into three parts: OCR Recognition, which uses existing tools to extract captions from videos to text; Speech Recognition, which uses deep learning tools(BaiduSpeech) to translate audios to text; NLP tasks, including segmentation, part-of-speech tagging, named entity recognition, dependency parsing, semantic role labeling and so on. The most important part is Speech Recognition. Since there are few guidences about how to use DeepSpeech model to train Chinese, I will pay more attention to this part and train a model as soon as possible.

About Me

- **Name** : Xu Zhaoqing
- **Location**: Beijing,China
- **Time Zone**: GMT +8
- **Education**:Junior, Electronic Information Engineering, Beihang University
- **Email**:15071124@buaa.edu.cn

Project Proposal

Here is a brief introduction to how I would contribute to this project. Specifically, this summer I'd like to concentrate on these aspects:

- For the data collection, I will be very willing to help if you get any problem.
- For OCR Recognition, forced alignment and Multi-program transport stream splitting, since I'm announced that I don't need to do these parts, I'm very willing to be in colleration with other partners and I will try my best to help them improve the performance if needed.
- For Automated Speech Recognition, I will use DeepSpeech from Baidu because it's so powerful in many languages, of course in Mandarin.[1] It's really time-consuming, so I'm trying to get an Azure account and train on the cloud. What's more, I find several great training datasets of Chinese speeches which are contributed by Tsinghua University.[2] So we could use them to train our own model and applied them to convert speech to text. And another very deep learning tool -- Kur, would help us save a lot of time.[3]

- For word segmentation, I would use the most famous Chinese library: jieba. It's confirmed the quickest and most powerful tool in segmentation by Tsinghua University[4].
- For other basic NLP tasks, I'd like to use pyltp by HIT as a main library for pos-tagging, named entity recognition and other tasks. Why do I choose it? Because it covers nearly all of the main tasks in NLP and it has complete documentation and examples[5], and it's totally free without any licenses (nlpir requires it!!).
- For other high-level tasks, I think there are still some interesting ideas for us to explore. I will be very glad to implement ideas from you, and now I have two ideas: the first is to use word2vec method to calculate the word similarity, from which we could know some potential connections among different entities. The other is we could count the frequency of words in order to insight some maybe economical or political or fashion style tendency during these years or in one year. These are both easy work for NLP and could be done in one week. For some complicated ideas, I'm very glad to implement them after summer.

Schedules of Deliverables

- **March 28 - April 10(two weeks):**use DeepSpeech to train a simple Chinese model and run the test
- **April 23 - May 14 (three weeks)**: familiar with the community, mentors and coding styles; continue to be familiar with the tools, help with the mentors to get all the data we require done.
- **May 15 - May 29 (two weeks)**: some pre-processing tasks about OCR extraction,data classification, cleaning, format conversion and forced alignment.
- **May 30 - June 12 (two weeks)**: Use the trained model to do ASR, improving the performance of it while training

Milestone : Get all the pre-work done

- **June 13 - June 20 (one week)**: finish ASR training and got the result.
- **June 21 - June 28 (one week)**: finish all the documentation above.
- **June 29 - July 6 (one week)**: I have to focus on my final exam, so during this time I'm going to revise the bugs in the tasks above.

Milestone : Finish all the work and its documentation except NLP

- **July 7 - July 14 (one week)**: finish all the basic NLP works including segmentation, POS-tagging, named entity recognition, etc.
- **July 15 - July 22 (one week)**: I'm going to caclulate the word simliarity and to find the connections given a word.
- **July 23 - July 30(one week)**: I'm going to analyze the tendancy of commercial and politics using some key words and visualize them using matplotlib, seaborn and other python tools
- **August 1 - August 7(one week)**: I will clean up all the codes I have written down, clearly commenting them, and write complete documentation for them.
- **August 8 - August 14(one week)**: Buffer Time. It's very hard to accurately predict what I could do 3 months before, so I will set apart one week for buffer time.

Milestine : Finish all the tasks

Preparations I have done

- find the best tools for the task after comparsion, which I have mentioned above
- getting familiar with these tools
- read two papers about Deep Speech:
  https://arxiv.org/abs/1412.5567
  https://arxiv.org/abs/1512.02595
  Now I have known the ideas behind the codes
- Continue to train a good Chinese model with DeepSpeech, now I have found a great audio dataset for me to train, which is THCHS-30.And I have trained a great English model using the method of this article.But for the Chinese training, it's very hard to deal with the hundreds of characters. I'm stuck now at this point and still looking for solutions. I will finish the training before 4/10.

More about me

Skills

- basic knowledge about data science, machine learning, deep learning, nlp , cv
- C/C++, python3, javascript, verilog(the language for FPGA platform)
- Fluent in common Python packages like numpy, pandas, scikit-learn, nltk, jieba,etc.
- Tensorflow, Pytorch
- Code in Linux
- Solid background in Mathmatics,for our project: Calculus, Probability, Stochastic Process, Signal&System, Linear Algebra

Open Source

I'm a member and contributor of an open-source organziation: Apachecn, which is aimed to teach more and more Chinese people to be familiar with machine learning and implement the algorithms in code. I contributed the implementation of Perceptron,Linear Unit in deep learning in python3. and co-organized two activities about teaching Kmeans algorithm and Introduction to Deep Learning.

Moocs:

I really like nlp and machine learning, and study all the things online, mostly on Coursera. Here are some courses I have learned or is learning which are related to our project.

- Introduction to Data Science in Python *by UMich*
- Machine Learning Foundations: A Case Study Approach *by UW*
- Applied Text Mining in python *by UMich*
- Machine Learning *by Stanford*
- Natural Language Processing *by NRUHSE*
- Introduction to Deep Learning *by NRUHSE*
- How to win a Data Science Competition: Learn from Top Kagglers *by NRUHSE*

Why RED HEN LAB

The most exciting part for you is that you're dealing with Chinese data, text and speeches. It's undeniable for me because I have spent too much time learning nlp methods about English. I'm really keen on getting deep to find some interesting insights for Chinese political, commercial and cultural use.

What's more, since we're supposed to deal with ASR and model performance's improvement, it's better for a candidate to be familiar with both deep learning methods and the mathmatics behind the signals&systems. Being a EE student, I'm very familiar with FFT, DTW, HMM and other math methods which could be used in traditional ASR and help me read the papers related to it.

References

1. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin
2. Chinese Speech Dataset
3. How to train Baidu's deep speech model with Kur
4. Common Segmentation Tools Comparsion
5. LTP Tutorial