Pre-proposal discussion for Chinese ASR

6 messages

Twofyw Yang <ywywywx@gmail.com> Thu, Apr 4, 2019 at 12:07 PM
To: redhenlab@gmail.com

Hi, I'm writing to discuss my pre-proposal for project 1: Chinese ASR. Until now, my communication with Red Hen Lab happened on Slack. Below is a copy of previous communication for the record.

> Me:
> Hi! I'm an undergraduate research CS student from China.
>
> I discovered Red Hen Lab only recently but I've skimmed through the website and I'm truly interested in contributing to one of the projects. Sorry for my rudeness, but since it's already kind of late and I want to make the most out of the remaining days before deadline to craft my proposal and maximize my chance, I'd like to ask directly that do you have any project that lacks competitive proposals or would like to see more fresh ideas?
>
> My experience is mainly with computer vision, with a publication to CVPR workshop on image segmentation. I'm familiar with both image and video data processing, but I'm currently new to OCR (have played around with tesseract). My experience also includes NLP and I can prototype a deep learning system using PyTorch fairly quickly. I'm proficient in Python and C++.
>
> Thanks in advance for any pointer about choosing a project to look into deeply.
>
> MBT (Mark Turner)
> Please write to redhenlab@gmail.com. We remain interested in strong proposals for the Chinese pipeline. Where in China are you located? We have an ASR pipeline that needs substantial improvement. We also want good Chinese NLP, which will be needed once the ASR is better.

In response to the last message, I'm located in Shanghai, China, and I've looked into the current Chinese ASR pipeline. According to the project requirement, I have studied the code, blog posts listed on the wiki page. In this email, I'd like to discuss what I'll address in my proposal before submitting more detailed specifications.

Firstly, the current system uses outdated packages, such as Python 2, due to the restriction of the PaddlePaddle implementation of Deep Speech 2 used. The difficulty of upgrading the existing system is mainly raised by two reasons: one is the lack of better implementation and another is the lack of corpora. The project description states: "the system should be upgraded to Deep Speech 3", however, I'm having a problem finding an open source implementation of Deep Speech 3. I'll need more help and discussion about this. Alternatives to upgrading the system to Deep Speech 3 would be 1.to improve the current system by further data processing and fine-tuning, or 2.to train a new Chinese model using Facebook's wav2letter++, which is newer than Deep Speech 2, and migrate to that. Fine tuning will suffer from the lack of labeled data in our dataset (has the situation improved?), but we can try to fine tune on more open Chinese datasets and see if it works. If we want to migrate to wav2letter++, the performance is not guaranteed to improve since the model used by the current Deep Speech 2 system is trained with Baidu Internal Corpus, which we can't obtain and we can only train it with open datasets. Discussions about training wav2letter++ with Chinese can be found.

The input to the current system is audio cut at fixed intervals, generated with a simple shell script (split.sh), which the participant from the last year's GSoC explains to be due to the limitation of time. There are a variety of existing audio segmentation methods, such as pyAudioAnalysis. We can experiment with both supervised (fine tuned) and

unsupervised methods to see what best suit our need. The data can be further cleaned by applying music removal, open source projects for which can be easily found.

Extending from the wish of the last year's participants, the following improvements can be made:

- Currently the code in real production is not efficient. I have to implement monitor-new-data function for it, and mutli-thread techniques will help me save a lot of time.
  - Write paralleled code to process the data.
- Use the data to do some easy nlp tasks like word segmentation, word counts and so on. It would be easily done in 1 or 2 days.
  - I'd love to do so.
- Alignments would be a nice task to try.
  - Open source implementations and papers can be found.

From the user's perspective, the documentation of Chinese ASR hasn't been merged or linked from the Chinese Pipeline wiki page, which should be improved. The automation level is relatively low, so it may take some efforts to reproduce the result, and a mix of python and shell scripts is required for the pipeline to work. Repackaging and refactoring into a python app using Google Fire would make for a consistent user interface.

I appreciate further feedback about aspects I can improve my proposal or points I'm missing. I'm also ok to draft a Chinese NLP proposal as suggested by Mr. Turner if you already have strong proposals for the ASR project. Look forward to your reply so I can draft a timeline based on your suggestion :)

Francis Steen <distributedlittleredhen@gmail.com> Thu, Apr 4, 2019 at 5:18 PM
To: Twofyw Yang <ywywywx@gmail.com>, redhenlab@gmail.com

Dear Twofyw Yang,

We appreciate your thoughtful analysis of the opportunities for going forward with the Chinese ASR project. Prof Turner is tracking the Chinese ASR project more closely than me, but let me provide some comments.

1. Could you comment on the status or utility of this code?

   https://github.com/mozilla/DeepSpeech

2. My recollection is that we downloaded a 70GB training set from Baidu. If this can be improved on, along the lines you suggest, this could represent one strategy for improving our results with the current implementation.

3. Improving the audio processing for the input is definitely worth doing.

4. Adding a new implementation using Facebook's wav2letter++ would be great.

Would it be feasible to include both of these strategies in your proposal? That is to say, attempt to improve the recall of the current implementation, and also add a new implementation? At the same time, improve the input. If this is feasible, we encourage you to detail the steps in your proposal.

If the results are good, it would certainly be worth doing forced alignment and some basic NLP, but the first task is to improve the output. In your proposal, you could include these as stretch goals.

Could you propose a Chinese researcher who may be interested in mentoring this project for Red Hen Lab? Could you give provide us with some more information about your university, your expected graduation, and your future plans and interestes?

Best wishes,
Prof Steen
Red Hen Lab

[Quoted text hidden]

Francis Steen <distributedlittleredhen@gmail.com> Thu, Apr 4, 2019 at 6:15 PM
To: Twofyw Yang <ywywywx@gmail.com>, redhenlab@gmail.com

P.S. You make good points about documentation; please include improving the documentation and documenting your own work in the proposal.

Red Hen aims to generate code that is as simple to maintain as possible. It's not entirely clear that Python Fire will be needed. Please give careful thought to how to design your system so that it will be easy to maintain, modular in design where possible, and avoiding any layers of code that aren't strictly necessary.

Best wishes,
Prof Steen

[Quoted text hidden]

Twofyw Yang <ywywywx@gmail.com> Sat, Apr 6, 2019 at 5:24 PM
To: Francis Steen <distributedlittleredhen@gmail.com>

Respected Prof. Steen,

Following is my response to your comments from the last email.

1. Could you comment on the status or utility of this code? https://github.com/mozilla/DeepSpeech

    This system from Mozilla is an open-source implementation based on Baidu's Deep Speech 1 back from 2014. The project is stable, under active development and used by many other open source projects. However, the model architecture used is older than Deep Speech 2 and wav2letter++. Moreover, they don't have trained weights for Chinese like the weights our system already has been using that is trained with Baidu Internal Corpus. There are attempts to training a Chinese model using this project, but I can't find benchmarks or comparison with other implementations online, which means I don't know the optimal hyper-parameters suitable for this model given the difference in vocabulary size between Chinese and English.

    On the other hand, because this project is a complete implementation of an ASR pipeline, the preprocessing steps of the input audio has been well developed, such VAD (voice activity detection, used to cut audio between sentences). We could also learn from other open source projects based on this project, for example, I learned that there is a good python interface to the WebRTC VAD because I found it's used by a front-end of DeepSpeech.

2. My recollection is that we downloaded a 70GB training set from Baidu. If this can be improved on, along the lines you suggest, this could represent one strategy for improving our results with the current implementation.

    I believe the 70GB data downloaded last year was the weights of the large language model used by the model pre-trained with Baidu Internal Corpus they provided. We could try to fine-tune this model on more Chinese datasets other than Baidu Internal Corpus and see if it improves performance. Fine tuning works better on training data similar to test data (Chinese TV shows in our case), but only if we can obtain labeled TV shows to fine-tune the model, otherwise we can only try to fine tune on other open Chinese datasets.

3. Improving the audio processing for the input is definitely worth doing.

The audio processing can be improved in two stages: first use VAD to properly cut audio, after that try other advanced audio processing techniques, such as Music Removal and Speech Enhancement.

4. Adding a new implementation using Facebook's wav2letter++ would be great. Would it be feasible to include both of these strategies in your proposal? That is to say, attempt to improve the recall of the current implementation, and also add a new implementation? At the same time, improve the input.

Both the input and the model are equally important to the performance of the pipeline, and yes, I will include these strategies in my proposal. The proposal would consist of two parts: enhancement to the input data and the model. These two parts can branch out and process independently, taking advantage of the existing modular pipeline, to avoid stall when training the models.

5. Documentation and code quality

I'll include improving the documentation in my proposal and keep the code quality and design in mind.

About me, I'm a third year CS student at Tongji University, Shanghai, supervised by Prof. Yin Wang. My university isn't particularly strong in NLP and to my knowledge no lab is doing relevant research. I've talked to my friends at two top universities in NLP in Shanghai, Sudan University and Jiao Tong University, asking for potential mentors for this project, but unfortunately, they don't know labs or professors doing ASR research either. Can I ask what would be your requirement for a student to mentor this project? Concerning my plan, I'm going to take a gap year next academic year because of personal reasons, and I plan to apply for a master program in Computer Science in the US. My current plan for my gap year is to be doing research at my supervisor's lab during the first semester and apply for exchange in the US for the second (spring) semester.

I'll send you a copy of my proposal as soon as I finish writing it.

Yours,
Wenxiang Yang

Twofyw Yang <ywywywx@gmail.com> Sat, Apr 6, 2019 at 5:28 PM
To: Francis Steen <distributedlittleredhen@gmail.com>

There is a typo in the last paragraph. It should be **Fudan** University and Jiao Tong University.

Francis Steen <distributedlittleredhen@gmail.com> Sat, Apr 6, 2019 at 7:00 PM
To: Twofyw Yang <ywywywx@gmail.com>, Red Hen Lab <redhenlab@gmail.com>

Dear Wenxiang,

Thank you for discussing your educational situation. I think we can manage this without a Chinese mentor, but let's keep our eyes open in case an opportunity arises.

The python interface to the WebRTC VAD is a great discovery. Please ensure you complete and submit the proposal by the April 9 deadline without waiting for further feedback.

Best wishes,
Prof Steen