Project Proposal For 11-785

Yilang Liu
Mechanical Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
yilangl@andrew.cmu.edu

Yao Lu
Mechanical Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
yaolu2@andrew.cmu.edu

Wuzhou Zu
Mechanical Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
wuzhouz@andrew.cmu.edu

Xupeng Shi
Electrical and Computer Engineering Department
Carnegie Mellon University
Pittsburgh, PA 15213
xupengs@andrew.cmu.edu

Abstract

This is the project proposal for 11-785 course. Our project is to use RNN specifically Long short-term memory (LSTM) to transform audio, video input into transcripts and determine who are saying. The data sets are captured from Japanese anime "君の名は" Development will be facilitated by open source solutions such as the NLP From Scratch tutorial by PyTorch, where our primary contribution will be either labeling the speaker or segmenting the transcript to create visually-aware subtitles. We will choose a dataset that is reasonable for our hardware requirements. We expect a model with 20000 parameters to take 30 minutes long to perform (training / validation) on one batch of data. Therefore, for the entire dataset, we expect (training / validation) to take 180 minutes. This is where the 1.2 $ estimate comes from, since (3h x 0.4 $/hour = 1.2 $). Our main resource package is PyTorch. We will try to explore more open source methods in NLP with PyTorch to train our model. The TA mentor for our project is Joseph Konan.

1 Project Overview

1.1 Problem Address

In order to enable content creators and media providers to create better NLP applications for animated film, we selected this project to create and train a LSTM model focused on NLP and Computer Vision in anime Kimi no Na Wa.[1] [2] Our main job in the project is to generate a labeled transcript based on the given anime video as the dataset. In the transcript, we first need to translate the dialogue between characters. There is lots of similar research and application on it, but the main problem of us is to improve the efficiency and performance of the translation with multimodal approaches. We also need to distinguish the speakers of the dialogue. During the project, we will show our process of establishing and training our neural networks. In the end, we will report our results for reproducible research. In the process of our project, we have to overcome many engineering problems with putting forward the end-to-end deeping learning solutions.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

1.2 Datasets Source: 君の名は (Kimi no Na wa)

We have two datasets. The first dataset is Audio recording with dialogue. There are Japanese and English versions. The second dataset is the transcripts of the dialogue showing the content of it and who is the speaker.

2 Literature Review

In recent decades, speech recognition has been drawing much attention in applications of deep learning. As of today, the accuracy and reliability of automatic speech recognition have improved a lot. Recurrent neural networks (RNN) has been playing an important role in this remarkable progress.

RNN is a kind of neural network that contains a hidden state which can link the input and previous state together. Based on Long short term Memory Networks (LSTM), a variant category of RNN, many automatic speech recognition models have been developed and proven to be accurate. Graves et al. introduced a hybrid model, HMM-Deep Bidirectional LSTM (DBLSTM) that improves frame-level accuracy.[3] DBLSTM makes use of previous and future contexts by setting two separate hidden layers and processing the data in both directions. After proper training, this network shows promising results on TIMIT experiments.

A widely applied example of automatic speech recognition is the automatic captions feature on Youtube. In general, such system works by first recognizing the speech in audio and matching it with an existed vocabulary as much as possible. Then, the generated text is modified according to the context to increase the closeness with original meanings.[4]

The project will utilize Long short term Memory Networks (LSTM), a variant category of RNN, as the core model. The reason why we use the LSTM model to generate transcripts is that LSTM works better than standard RNNs for speech recognition and image captioning. Due to the property of this dataset which is continuous and lengthy, the LSTM(RNN) fits the task perfectly since the structure of RNN is continuous and RNNs effectively have an internal memory that allows the previous inputs to affect the subsequent predictions. It's much easier to predict the next word in a sentence with more accuracy, if the previous words are known.

3 Project Goal

The primary function of our LSTM model is to achieve dialogue recognition and generate a transcript. Additionally, the LSTM model we designed should also determine who is speaking in the audio. Since the data is in mp3 format, we have to preprocess the data and transform it into the waveform. Then, the dataset will be divided into three parts: 60 % training set, 20% testing set, and 20% validation set. The final goal of the project should take an arbitrary video, audio segment from Kimi no Na wa as input and output transcript with respective characters.

The accuracy of the transcript is calculated using Cross Entropy Loss. We will experiment with different optimizers such as SGD, Adam, AdamW to lower the loss. If overfitting happens during the training process, we may need to research appropriate regularization techniques for Variational Autoencoders. Since the training process will require huge computing power, we will be using a Cascade Lake CPU and a T4 GPU for computation. We expect to use 4 cores for development and 8 cores for training, though this may change.

References

[1] Ralf C. Staudemeyer and Eric Rothstein Morris. Understanding lstm – a tutorial into long short-term memory recurrent neural networks, 2019.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[3] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278, 2013.

[4] Anthony Romero. How does automated closed captioning work?, Nov 2018.