Project Proposal

Name: Gurtej Khanooja and Pruthvij Thakar

Date: 10/03/2017

Topic: Speech recognition with DNN-LAS:

Background:

Speech recognition is an interesting topic in deep learning because of its difficulty and its great applications in life: Amazon Echo, Apple Siri, speech typer, etc. The most common architecture used in speech recognition is recurrent neural network (RNN) and its variants. We want to implemented the Listen, Attend, and Spell (LAS) model [1] with an additional Deep Neural Network (DNN) feature extractor. The body of our model consists of an attention-based two multilayer Long Sort Term Memory (LSTMs) networks. This model is very efficient for end-to-end speech recognition tasks, and produces reasonably good results even without any language model rescoring or hidden Markov model (HMM).

Related Work:

There are many good models for doing end-to-end speech recognition tasks. We are particularly interested in the LAS model which was first introduced in Google's paper [1] in 2015, compared to other models such as Connectionist Temporal Classification (CTC) [2] and standard Deep RNN [3]. The paper shows that this model is decently good, and that it feeds the audio features directly into the pyramidal bidirectional LSTM network, and that it used one-hot format for samples from the previous time step. In our project, we want to improve upon these two points.

Data source:

We will be combining the TIMIT speech corpus [4] with VoxForge speech database [5] as our data set. The total number of samples is over 90,000. Audio files that are longer than 8 sec will be excluded, leaving us with 80,000 samples that totaled about 100 hours of audio.

Algorithm:

1) RNN (https://en.wikipedia.org/wiki/Recurrent_neural_network)
2) LAS (http://ieeexplore.ieee.org/document/7472621/)
3) MLP (https://en.wikipedia.org/wiki/Multilayer_perceptron)
4) CNN (https://en.wikipedia.org/wiki/Convolutional_neural_network)

References:

[1] Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. (2015). *Listen,attend and spell. arXiv:1508.01211 [cs, stat].*

[2] Alex Graves, et al. (2006) *Connectionist Temporal Classification: Labelling UnsegmentedSe- quence Data with Recurrent Neural Networks. ICML*

[3] Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton. (2013). *Speech Recognition with Deep Recurrent Neural Networks. arXiv:1303.5778 [cs.stat]*

[4] Garofolo, John, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[5] Voxforge.org. Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org.

http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/