Machine Learning 2

Final Project Proposal

Alex • Anamay • Tanaya

For our final project, we hope to examine and develop a voice recognition network with voice-to-text capability by adapting easily available, pre-trained voice recognition algorithms. We aim to apply our network to a specific dataset and observe its performance compared to voice-to-text networks developed by technology giants such as Google or Amazon. If successful, this project will show that one can develop better, domain-specific applications using a generalized pre-trained network with similar capabilities than some of the more comprehensive models. We selected this topic because of our interest in using neural networks in a speech-to-text capacity, and felt this would be a great way to both practice new techniques as well as see how expertly-developed networks are constructed and trained. Voice-to-text also has great applications in video transcription and education, especially for those with hearing disabilities.

We hope to use the LibriSpeech corpus, which contains around 1000 hours of English speech in the form of audio books. The clean development corpus contains 337M of annotated text, and the smaller of the training sets contains 6.3G, or 100 hours, of cleaned (human-understandable) English speech. We believe this dataset should be enough to adapt an existing network to a sufficiently robust dataset. Additionally, the comparison datasets produced by Google, Amazon, or other companies will be trained on gigabytes or terabytes of data, which proves more than sufficient for training networks.

For our network architecture, we aim to use a pre-trained RNN which goes by the name Deep Speech 2 (DS2), implemented by Sean Naren. Given that the network is pre-trained on specific corpuses, we will produce our network by altering and tweaking DS2 for our corpus of interest. Additionally, it can be assumed that the comparison networks are also variations of an RNN architecture. Since DS2 is developed using PyTorch, it will be much easier to continue using PyTorch than rebuilding the model under a different framework.

For references, we will be focused on reading on RNN architectures, bi-directional networks and voice-to-text processing through both academic papers and more high-level, summary internet articles. The Deep Speech 2 documentation and codebase should serve as the main reference material for this project.

The performance of our adapted network will be compared to the pre-trained, commercially available networks developed by large technology companies by comparing the correct number of transcribed words in a previously unseen set of audio files. This will be an overall *accuracy*

measure, meaning the number of correctly transcribed words divided by the total number of words in the audio files.

This current week will be focused on understanding the DS2 model, and downloading and pre-processing the audio files. The second week will revolve around attempting to implement the Deep Speech 2 model given the new data, as well as making the specific modifications necessary for more focused training. The final week will be additional model training and development, and evaluation in comparison to the tech-company provided transcription services, as well as preparing for the final presentation.

Reference links:

Paper: https://arxiv.org/pdf/1512.02595v1.pdf

Codebase: https://github.com/SeanNaren/deepspeech.pytorch

Database: http://www.openslr.org/12/