Intelligent Voice Identification with Neural Networks

Lloyd Maza (lmaza) - Michael Thompson (mt1) - David Troner (dtroner)
CS 230 Final Project, Winter 2018

Stanford University

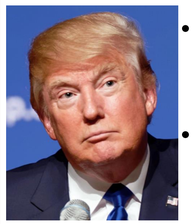
Motivation

Automated speech recognition is commonplace in dozens of technology domains. There are many potential applications for automated speaker recognition:

- Automated air traffic control
- Over-the-phone identity verification
- Speaker diarization for speech-to-text

Dataset

- **Audio clips of 17 people's voices**
- About 10 minutes of audio per person
- Audio collected from YouTube



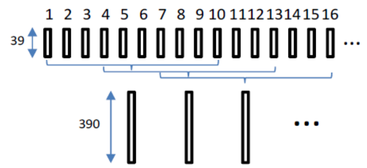
- Dataset breakdown: **95% train, 2.5% dev, 2.5% test**
- Per person: 9.5 min of training audio, 15s for dev, 15s for test

- Varying degrees of background noise
- Only one person speaking at a time
- Examples: Trump, Obama, characters from The Office

Pre-Processing

Mel-Frequency Cepstral Coefficients

- Need to represent data in a tractable way for a simple NN
- MFCCs are the standard in speech recognition tasks
- Derived from audio frequency content
- Modeled after human auditory system
- Extracted in sliding windows



Frame: 10 ms
Window: 10 frames
Stride: 3 frames

Uniformly scale audio amplitude

↓

Remove unvoiced audio with voice activity detection (VAD)

↓

Extract MFCCs and their derivatives for each window

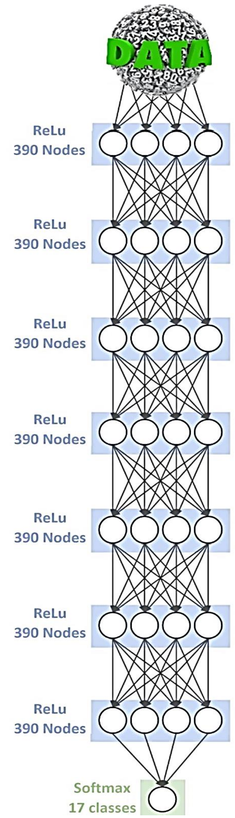
↓

Normalize speaker feature sets

Architecture

Neural Network Parameters:

- 7 FC layers with ReLu activations
- Softmax output with 17 classes
- Run for 25 epochs
- Mini-Batch Size of 64
- Early Stopping with $\delta_{min} = 0,\ \ patience = 10$
- L2 Regularization with $\lambda = 0.7$
- Adam Optimization with $\alpha = 0.001,\ \beta_1 = 0.9,\ \ \beta_2 = 0.999,$ $learning\ decay = 0.00001$
- Batch Normalization with $Momentum\ = \ 0.99$
- Tuned hyperparams in space $N = [2,4,7,10],\ \ \lambda = [0,.1,.3,.5,.7,1],$ $L = [10,100,390,500,1000],$ $Batchsize = [16,32,64,256]$

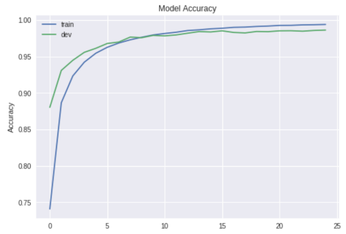


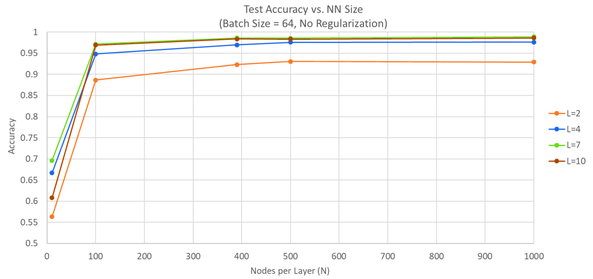
Results

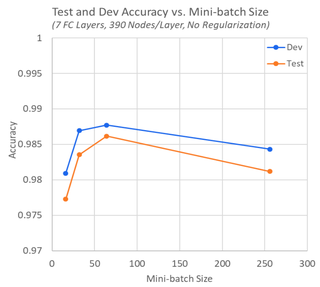
Accuracy on Dataset:

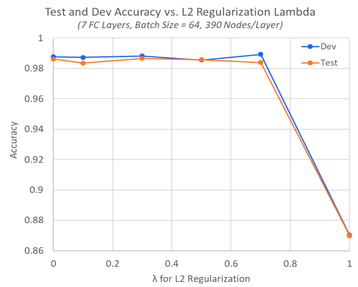
$Dev\ Accuracy\ = 98.64\%$

$Test\ Accuracy = 98.69\%$









Conclusions

- High accuracy over 17 speakers was achieved with an all FC layer architecture
- Small improvements in test accuracy were made through tuning of $\lambda_{L2}$and mini-batch size
- Asymptotic test accuracy behavior above ~400 nodes/layer and above ~7 hidden layers

Future Work

- Expand dataset to include more people
- Introduce "unknown" class
- Real-time speaker prediction
- Deeper hyperparameter search
- Automated sub-captioning

References

[1]: Z. Ge, A. Iyer, S. Cheluvaraja, R. Sundaram, A. Ganapathiraju, "Neural Network
Based Speaker Classification and Verification Systems with Enhanced Features," in Intelligent Systems Conference, London, UK, September 7-8, 2017.