Sentiment Analysis

SCPD presentation link: https://youtu.be/ldo05tf9iDk

fjur@stanford.edu, fjur@google.com
timkim@stanford.edu, timkim@google.com
sbanga@stanford.edu, shiprab@google.com

Stanford Computer Science

Introduction

While standard approaches to sentiment analysis use natural language processing, our approach analyzes features in the frequency domain, specifically, Mel-frequency cepstral coefficients (MFCCs).

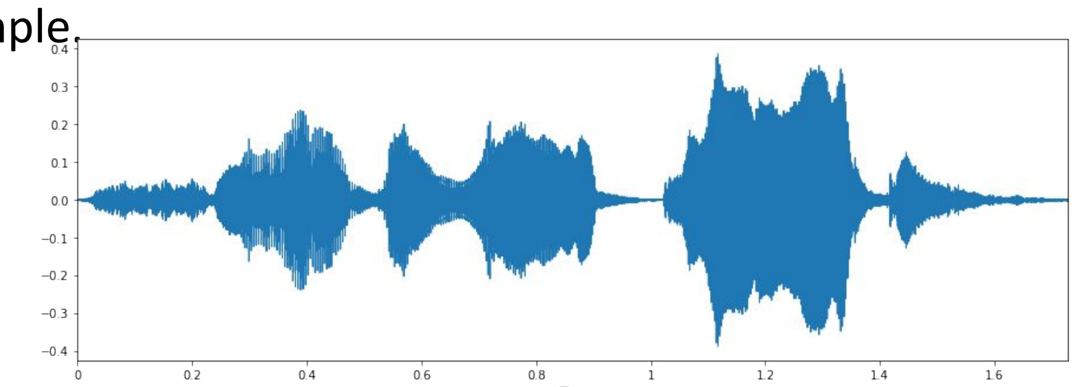
Social Impact

Sentiment analysis has become crucial as the advent of automated intelligent agents in various industries such as finance, automotive industry, and customer support require agents to understand the complex emotions expressed by customers. There are various socio-economic applications to this project, as many industry sectors may apply this to their intelligent agents to better understand customers. While humans are able to understand relatively complex emotions and sentiments, transferring this skill to artificial intelligence has unique challenges.

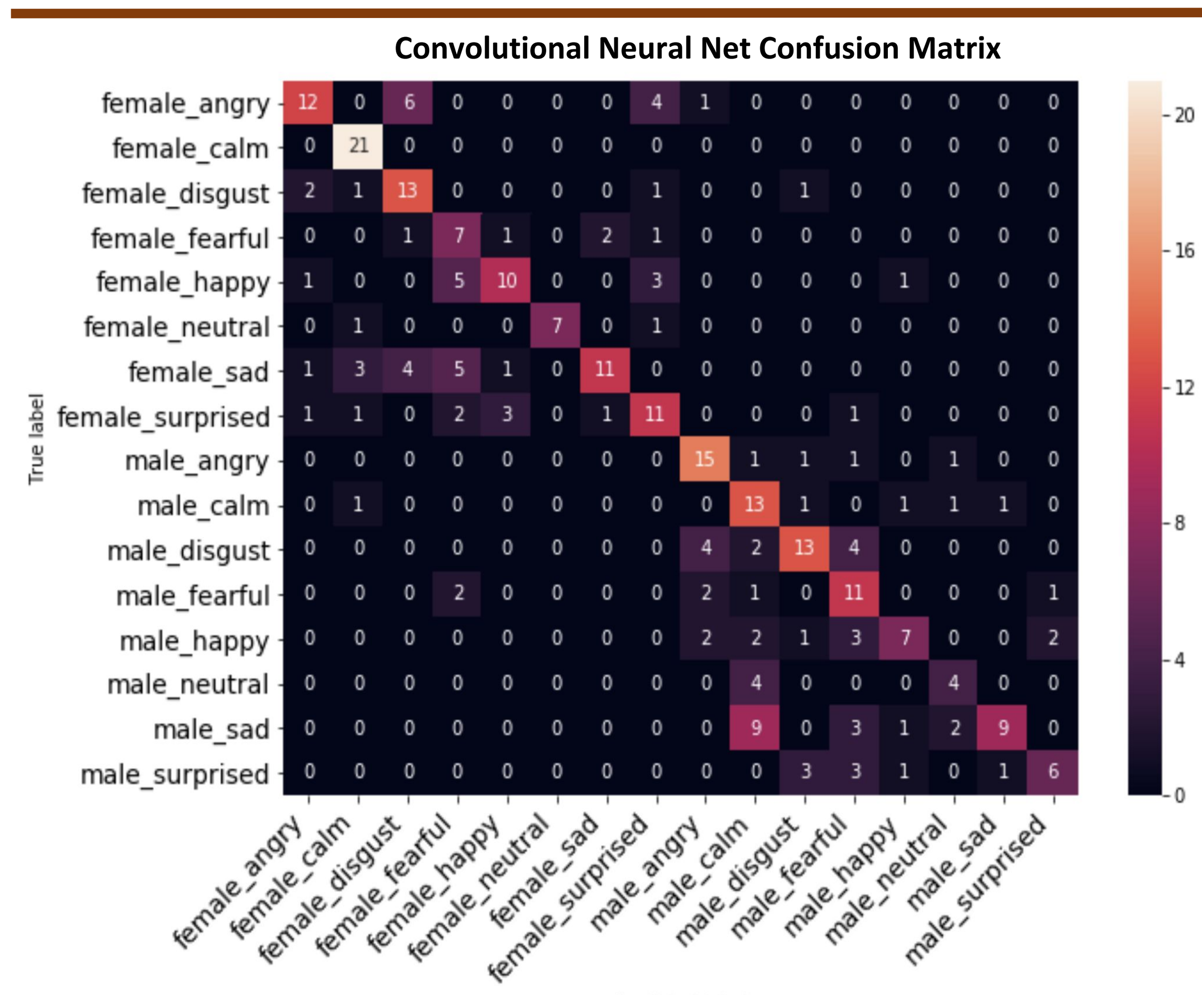
Stanford University

Data and Data Processing

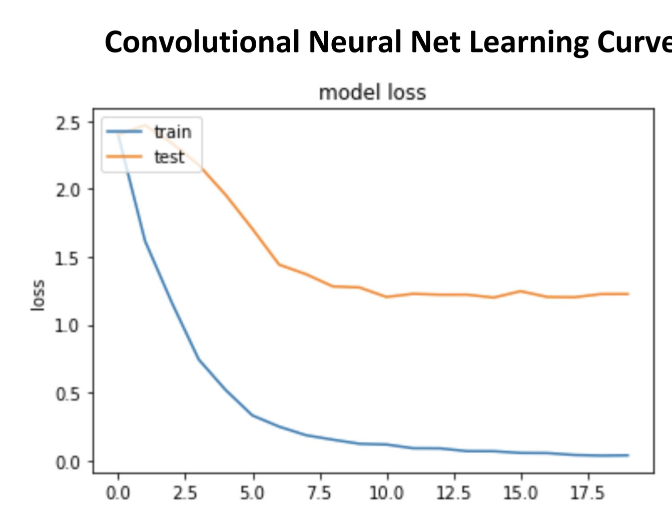
- RAVDESS dataset: 1440 clips of speech data from various male/female voice actors with labeled emotions.
- Challenge: There is a lack of high-quality audio datasets that are labeled by emotion.
  - Labels generated from human classification on audio clips isn't as reliable as recording human speech when the speaker is prompted for a given emotion label.
- Labels: Gender, emotion
  - Gender: male, female
  - Emotion: neutral, surprise, anger, sad, disgust, happy, fear, and calm
- The waveform below shows intensity across time for the sentence "Say the word dog." This is classified as a fearful female example.

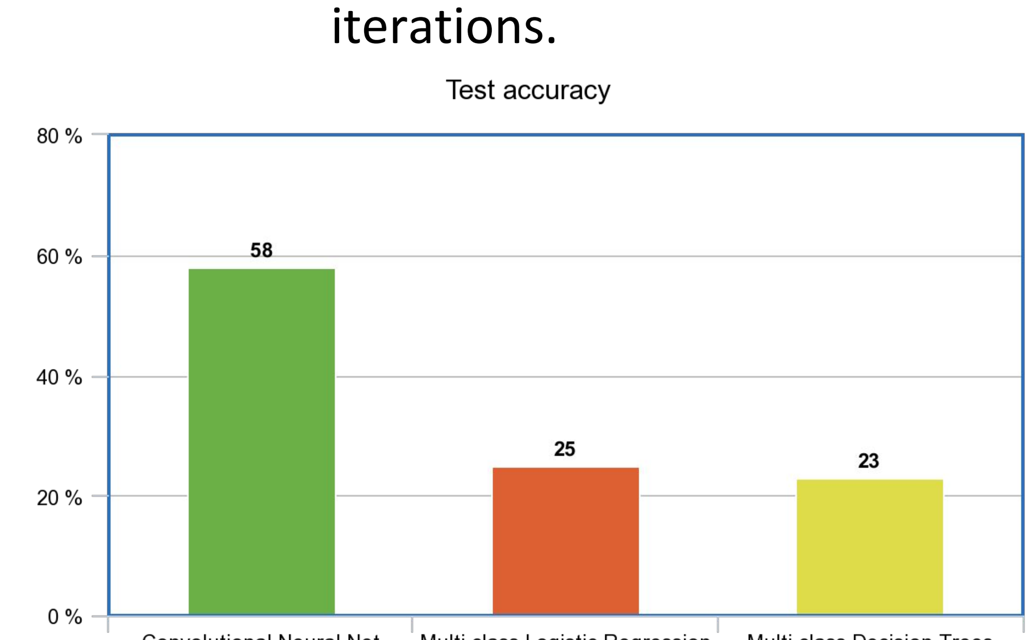


Results



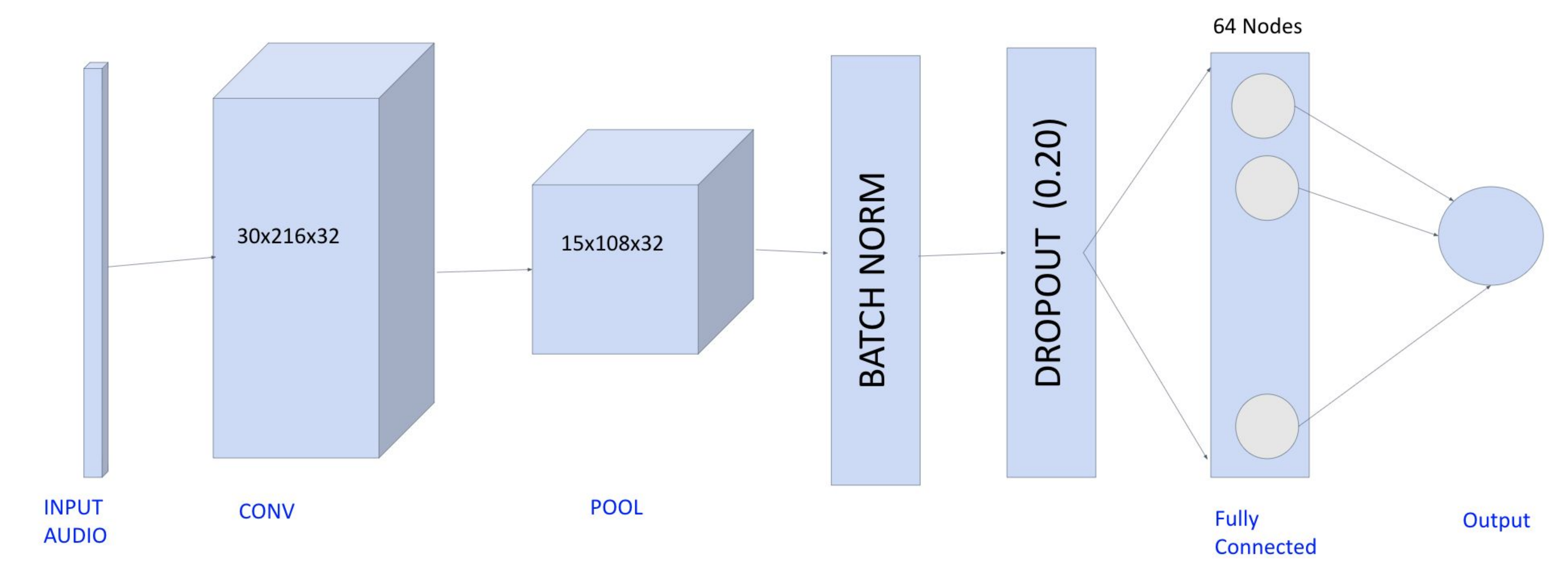
- CNN (58% accuracy) significantly outperforms Linear Classifiers (25% accuracy) and Decision Trees (23% accuracy). The CNN performs similarly to human classification (52% accuracy).
- Test loss for the CNN decreases per epoch, eventually converging around epoch 10
  - Choosing a shallow topology with a single conv-pool layer produces reasonable results with minimal training iterations.





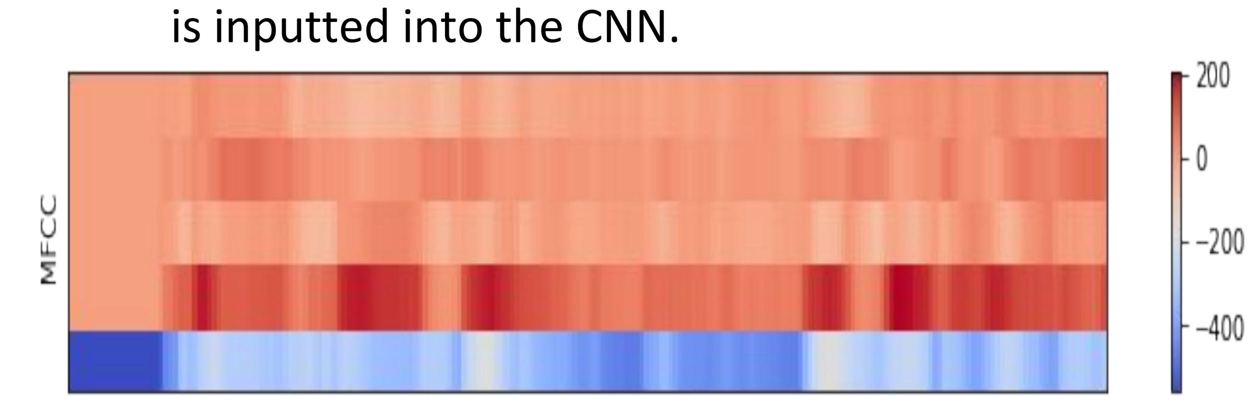
Models

- Using 2-D Convolutional Neural networks (CNN), since the problem is similar to Image Classification.
- Convolution layer of 32 size 3×3 kernels
- Max pool layer after convolutions layer
- Dropout layer for regularization to avoid overfitting on training data
- Batch normalization to prevent vanishing and exploding gradient
- After Conv-Pool layer, dense layer of 64 hidden nodes
- Selected one conv-pool layer, because the dataset contains only 1440 data samples



Features

- Primary features are the Mel-frequency cepstral coefficients (MFCCs).
  - The Mel transform converts audio clips in the time domain to the mel frequency domain.
  - This frequency is similar to the scale at which humans perceive sound.
- Each audio clip is converted to a 30x216 (MFCC band x time period) pixel image, which is inputted into the CNN.



Future Work

- We want to generalize to unseen phrases:
  - Approach 1: Gather more training data from other datasets with a variety of sentences and use OneShot learning.
  - Approach 2: Use Transfer learning with a pretrained CNN to generalize our model.
- Generating audio samples with a given sentiment

Discussion

- Using a shallow modern CNN architecture performs well on classifying MFCCs as images, especially given human agents perform inaccurately on classifying emotions due to the inherent subjectivity of classifying emotions.
  - Using deep learning techniques to classify emotions removes subjectivity of emotion classification and improves results.
- Due to small training set size, the model overfits, this can be verified by the training loss curve which diverges from the test loss.
- Error analysis based on CNN Confusion Matrix shows difficulty in differentiating:
  - male sad and calm, possibly due to similar low audio intensity
  - female angry and disgust, possibly due similar high audio intensity

References

[1]Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):
A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.
https://doi.org/10.1371/journal.pone.0196391.

[2] Lok, Eu Jin. "Audio Emotion Recognition." Kaggle, Kaggle, 12 Sept. 2019, https://www.kaggle.com/ejlok1/audio-emotion-part-6-2d-cnn-66-accuracy.

[3] Peng, Zeshan. "Acoustic feature-based sentiment analysis of call center data." (2017).

[4] Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.