Accent classification

Huanhan Liu

Ishaan Singhal

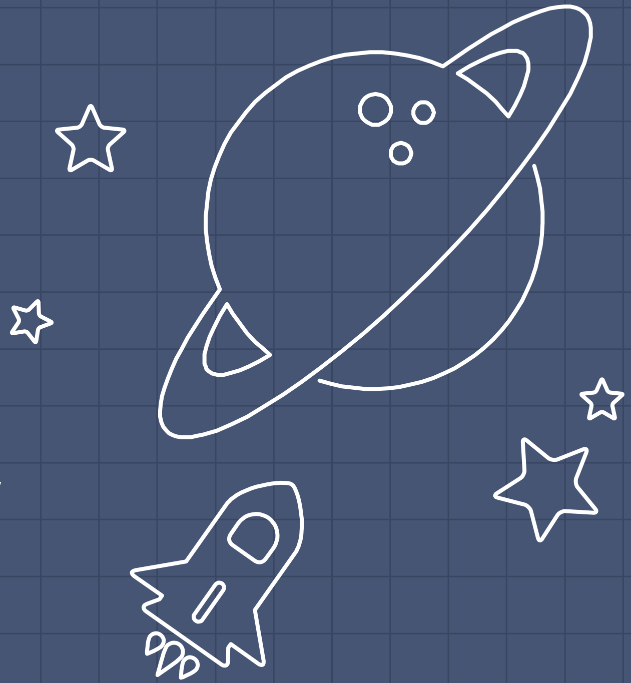
Kavin Chandrasekaran

Big Concept

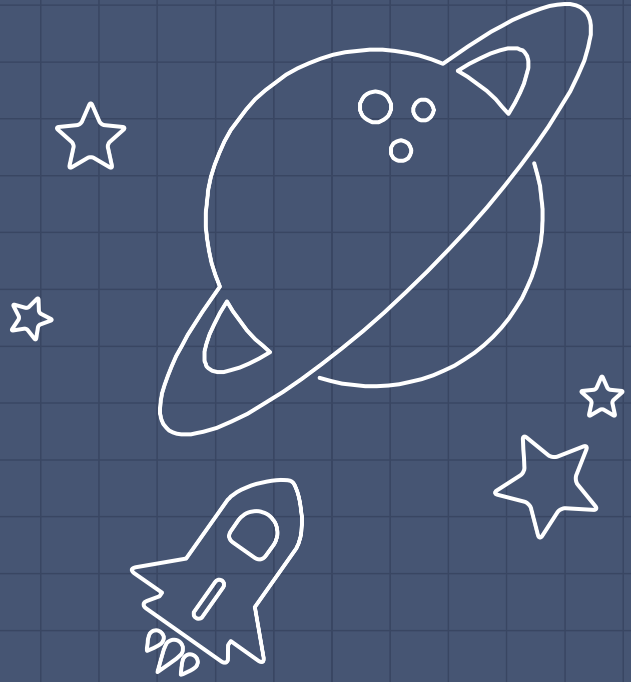
"Finding the country you are from based on your accent"

Motivation

1. If you want to become a super-spy
2. At airport and border immigration checkpoints



Data



Initial Dataset

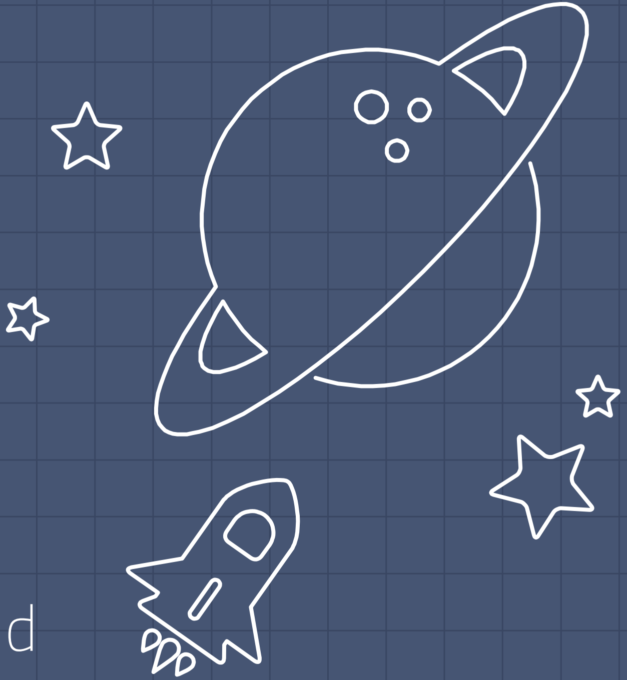
- 2172 audio recordings
- 176 unique countries
- 214 native languages

Final Dataset

- 5 unique countries (USA, China, India, UK, Canada)

Two Approaches:

1. Extract features from audio files and use a Fully-Connected Neural Network.
2. Create spectrograms of the audio and use a Convolutional Neural Network.



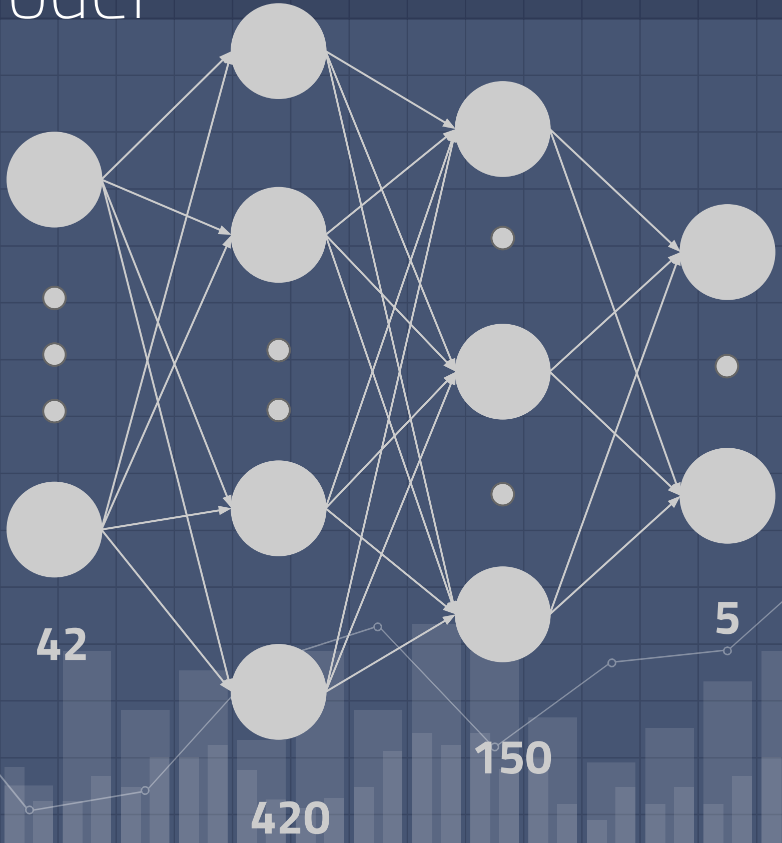
Fully Connected approach

1

Feature extraction

- Used Yaafe & pydub
- MFCC, Energy, Spectral Rolloff, Spectral Flux, Loudness and Flatness.
- Total of 42 features
- Random oversampling to address class imbalance

Model



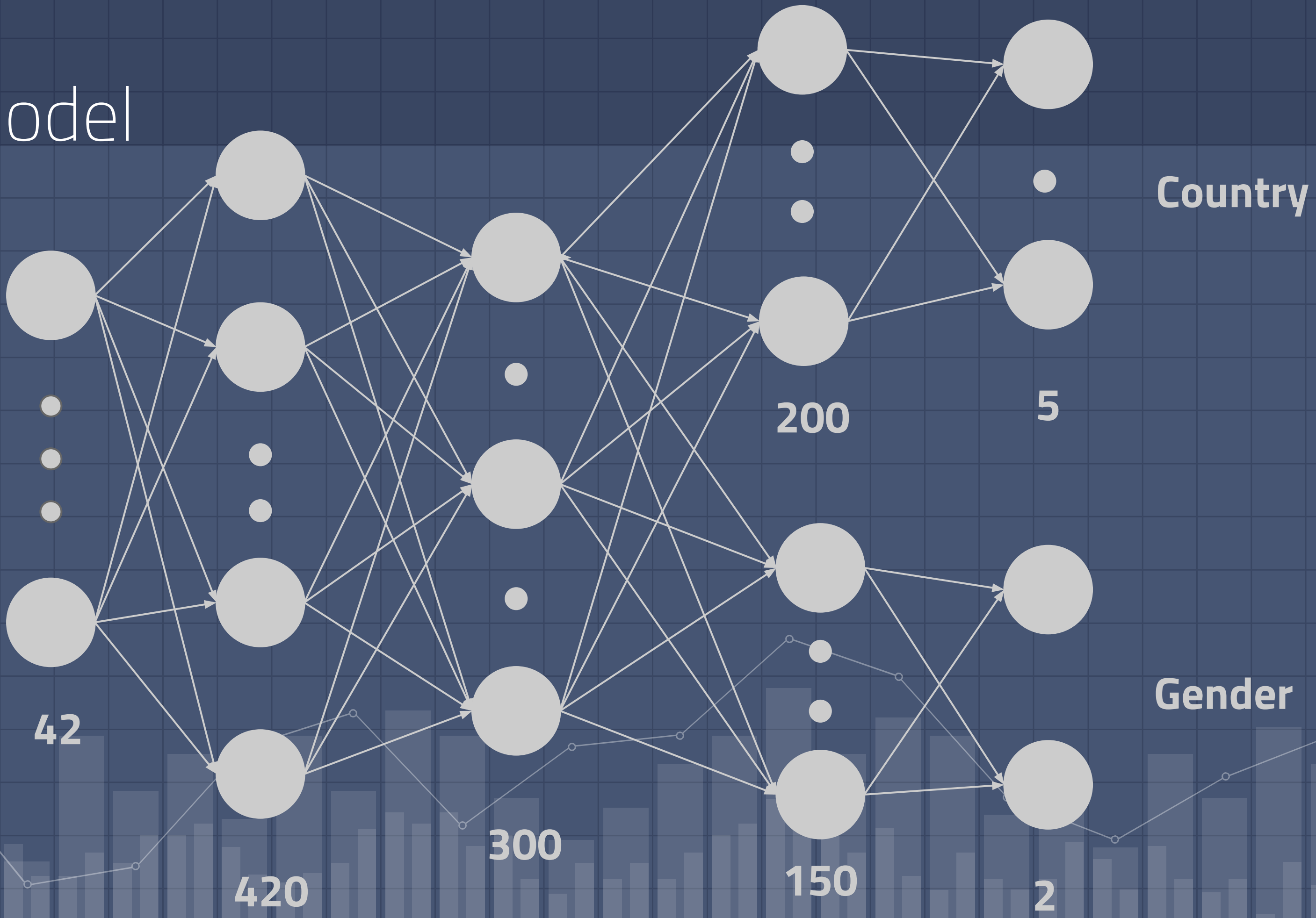
We got performance only slightly better than random. ~30%

The major issue was overfitting.

Multi-task Learning

- A regularization technique.
- While learning one task, learn another task as well, so that the shared parts of the network are regularized.
- We chose to use the gender classification to regularize the country classification.

Model

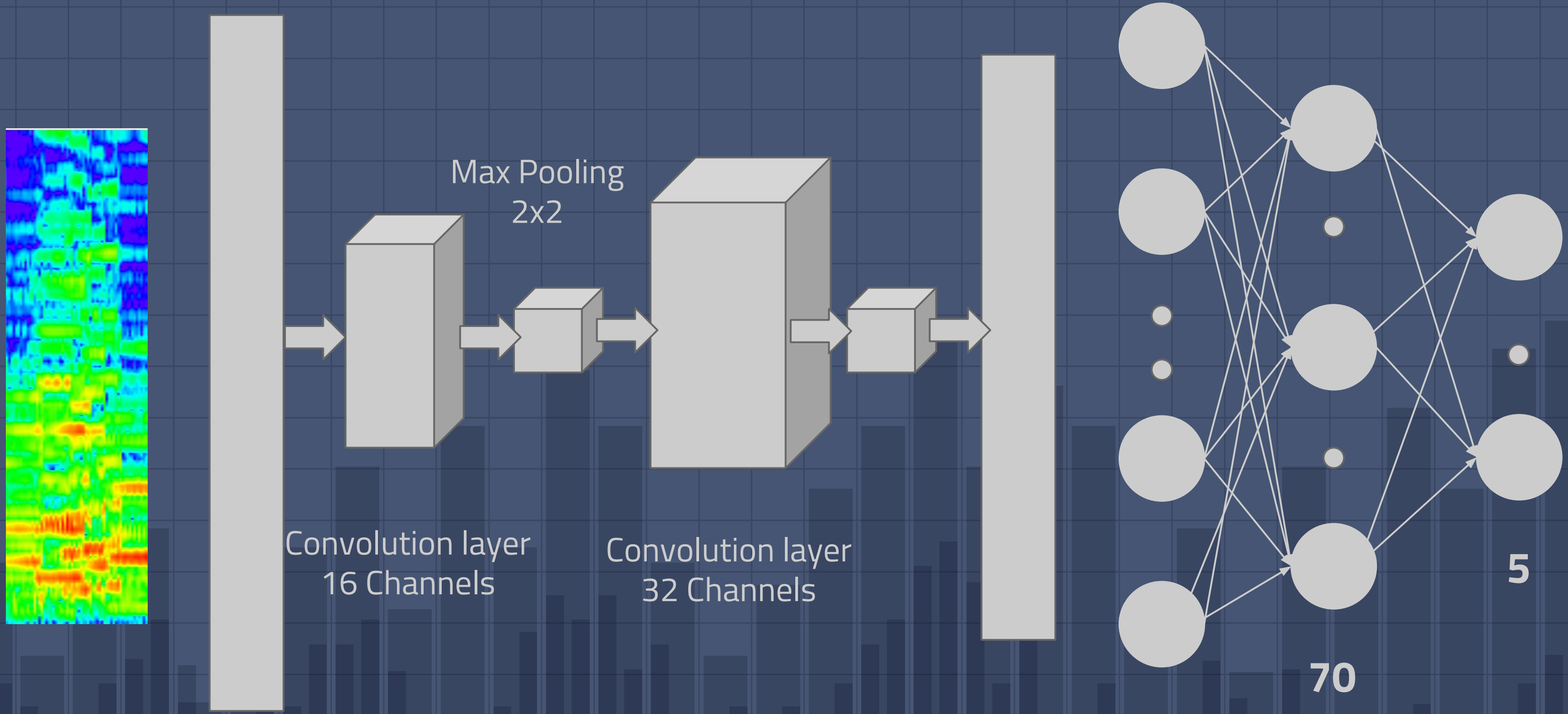


Multi-task Learning

- Since the inputs are based on audio signals, the variance of the input values was much higher, ranging from 1e-6 to 1e8.
- We used batch normalization address the covariate shift.
- We were able to minimize overfitting by using Multi-task learning.

CNN approach

2



Max Pooling
2x2

Convolution layer
16 Channels

Convolution layer
32 Channels

5

70

Input

Flatten 200

[1] https://www.lunaverus.com/cnn - Spectrogram source

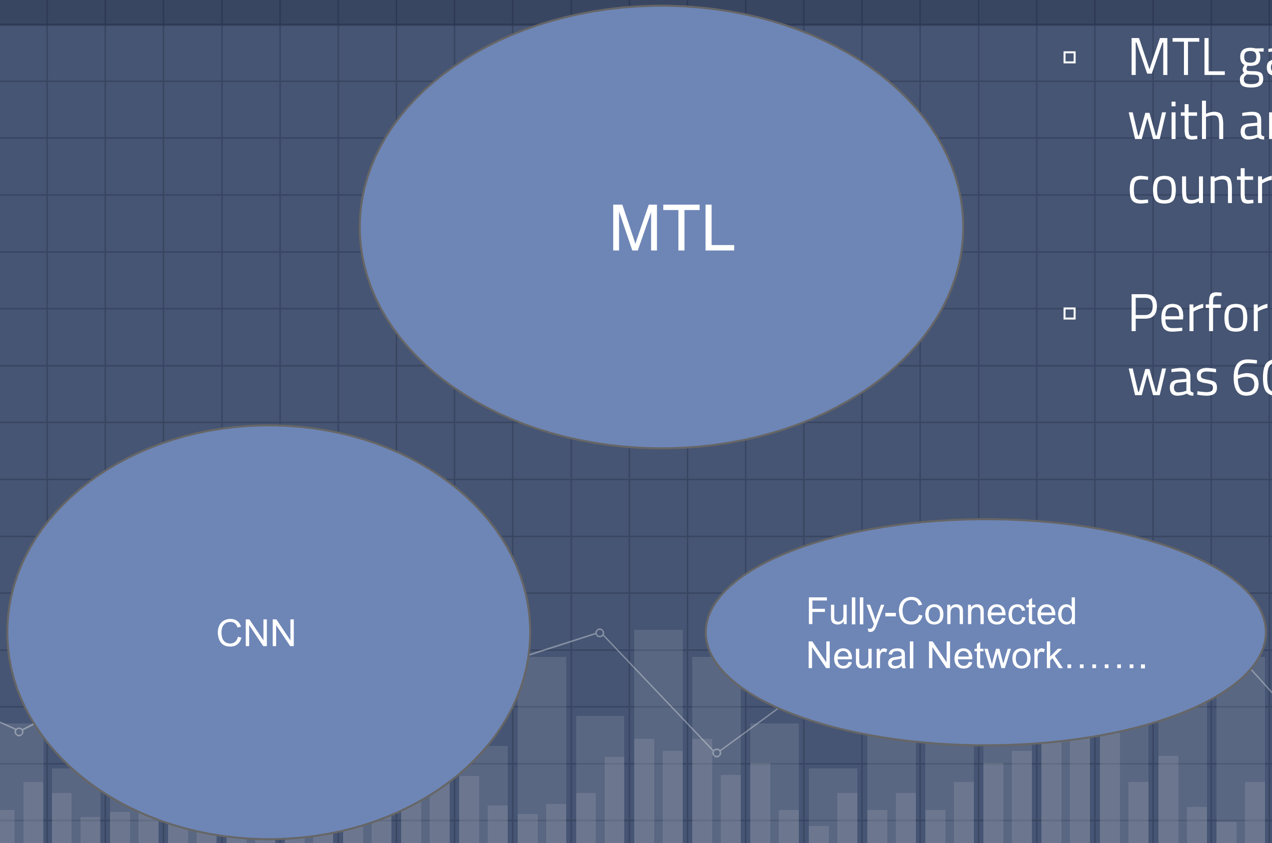
Overfitting vs Underfitting

- Our initial model was overfitting during the training.
- We employed many regularization techniques:
  - Batch normalization
  - Dropouts
  - Input data augmentation
  - Gaussian noise layer
- We regularized to a point the model was underfitting.
- We tuned the parameters to find the middle ground.

Results

3

Performance



MTL

CNN

Fully-Connected
Neural Network…….

- MTL gave us the best results with an accuracy of 88.67% for country prediction
- Performance of the CNN model was 60%

Conclusion & Future

4

MTL>CNN>>Fully-Connected

- Overfitting caused a large gap between training and testing accuracies.

- Hyper-parameter tuning needs to consider minimizing the gap and not just focus on accuracy.

- MFCC has limitations

- The information content of the spectrogram can be as high as for the signal in the time-domain

Possible improvements

- We could make use of the temporal information and train a RNN/LSTM model and use an ensemble method.
- We would like to do more research in the future on the accent-related information contained in the sound data.
- We could also try to develop a hierarchical classifier that initially considers groups of countries based on geography then makes more fine-grained model.

THANKS!

Any questions?