Voice Classification

Project by Jurgen Arias

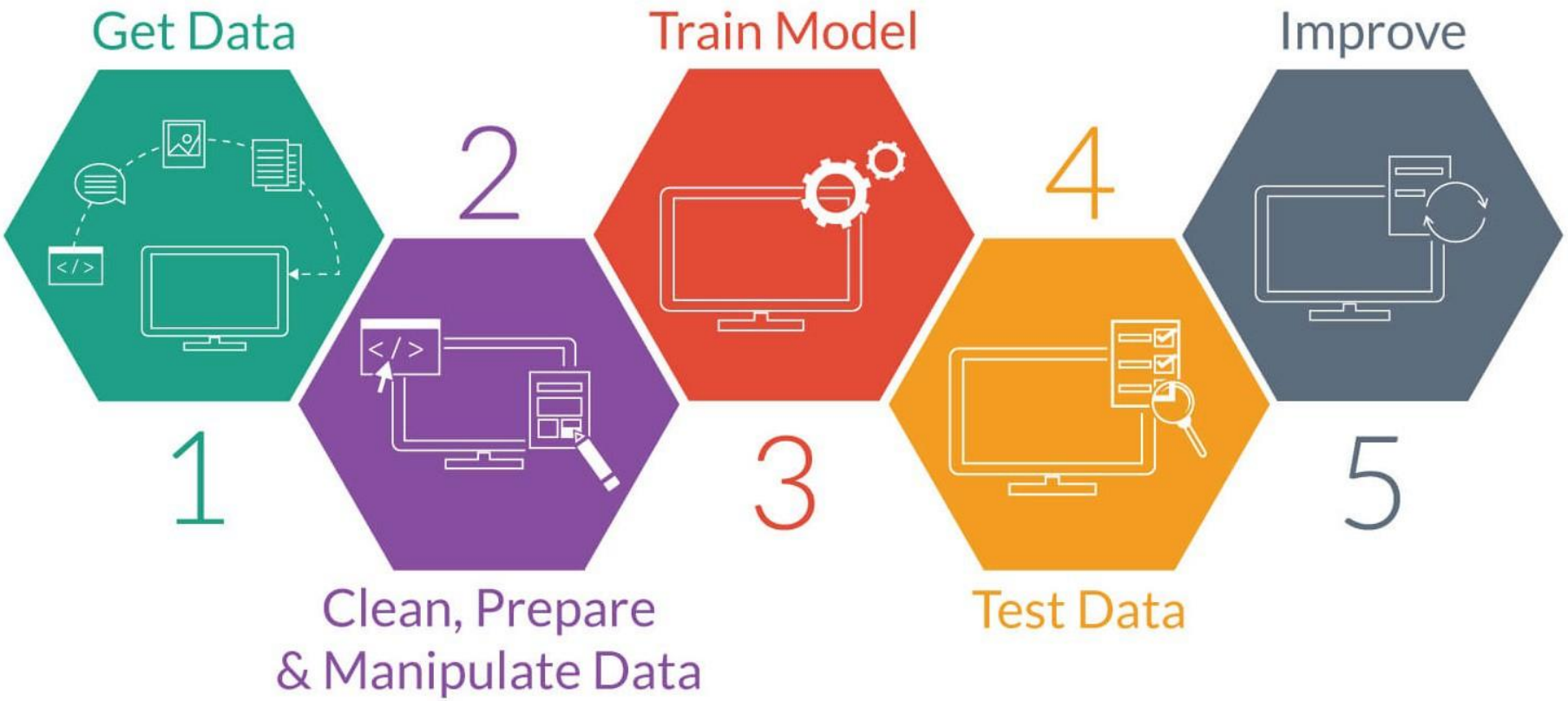
Who said what?

Build a classification model that can predict who is speaking

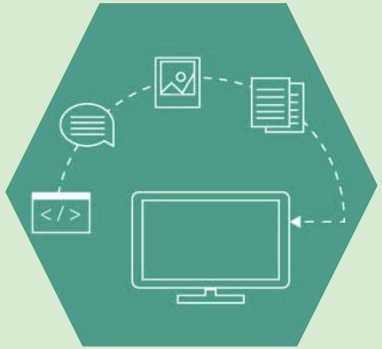
Gender of the speaker?

Build a classification model that can predict the gender of the speaker

Workflow


Get Data
1
2
Clean, Prepare
& Manipulate Data
Train Model
3
4
Test Data
Improve
5

Speaker Classification

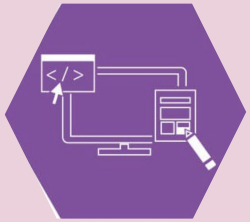


Get Data

Found an awesome dataset from Open SLR

83 datasets (mostly speech)

1000+ hours of spoken English



Clean, Prepare and Manipulate Data

Used 13k voice clips from 115 speakers. The voice clips ranged mostly from 12 seconds to 18 seconds. Average of 100 voice clips from each speaker

Extracted audio features using librosa. Mel-frequency Ceptral Coefficients (MFCCs), Chromagram, Mel-Scaled Spectrogram, Spectral Contrast and Tonal Centroid Features (tonnetz)

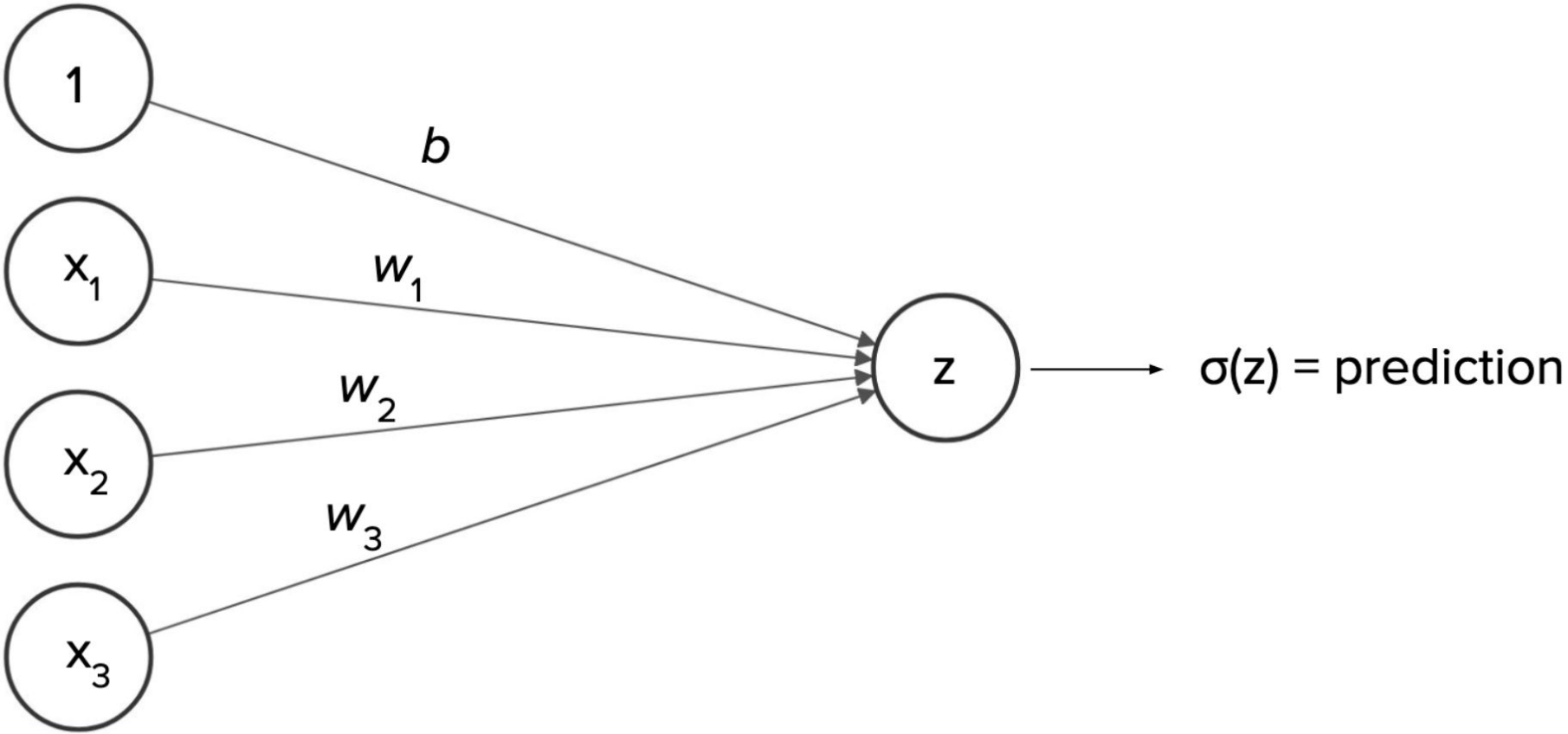
Splitted data into training, validation and test, 10k for training, 2k for validation and 1k for testing

Train Model

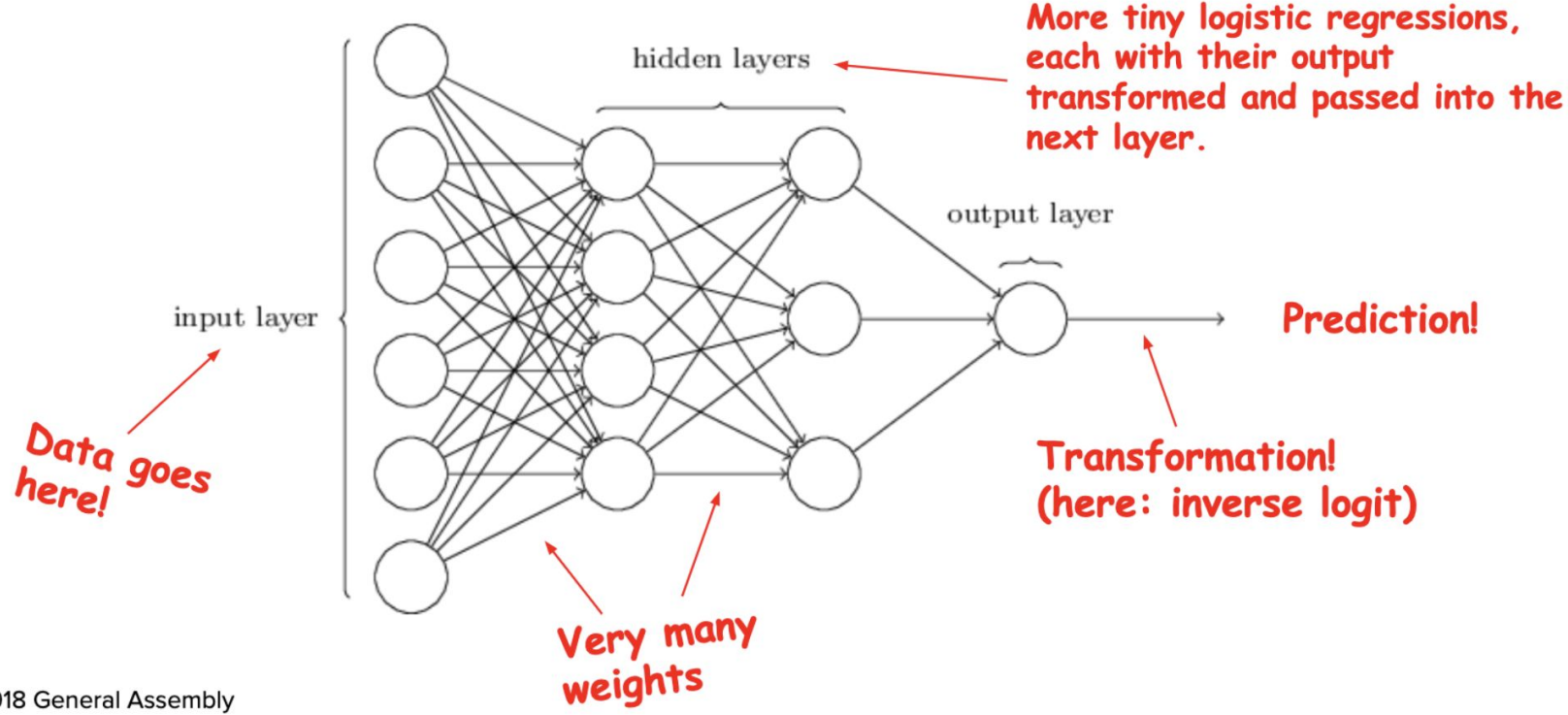
Used a Dense Neural Network with two hidden layers

It took 20 seconds to fit the model

94k parameters


1
b
x1
w1
z
σ(z) = prediction
w2
x2
w3
x3

Output of one gets transformed and then passed as input to another:



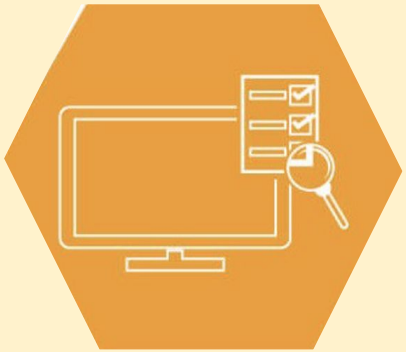
Train Model

Used a Dense Neural Network with two hidden layers

It took 20 seconds to fit the model

94k parameters

Test Data

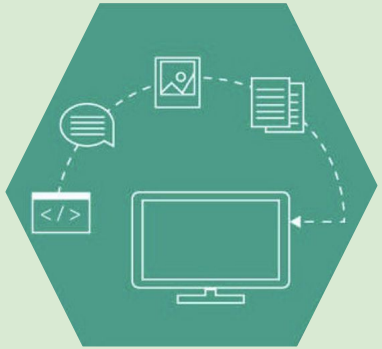


Generated predictions on test data

Remember we had 115 different speakers

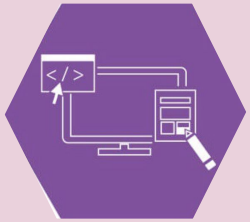
We got **99.8%** accuracy when predicting the speaker

Speaker’s Gender Classification using Convolutional Neural Networks



Get Data

Used the same data as before

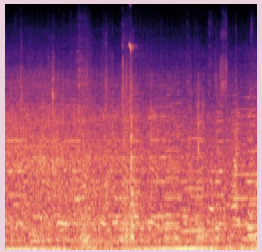


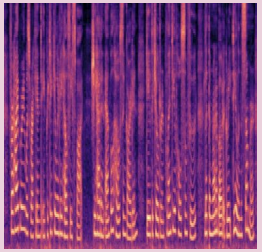
Clean, Prepare and Manipulate Data

Labeled Data

Converted the voice clip to an image using librosa

Sample images:





Splitted data into training, validation and test, 10k for training, 2k for validation and 1k for testing

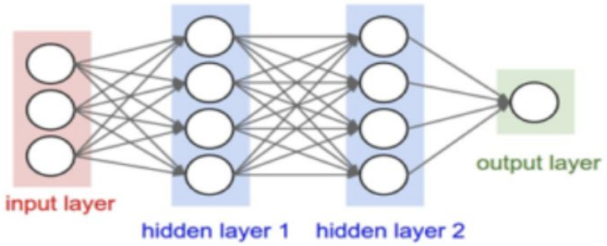
Train Model

Used a Convolutional Neural Network with five hidden layers

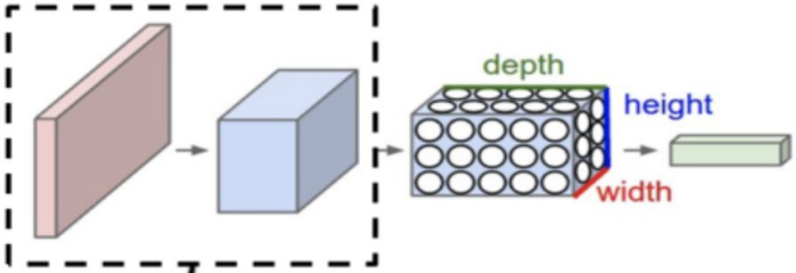
It took about two hours to fit the model

569k parameters

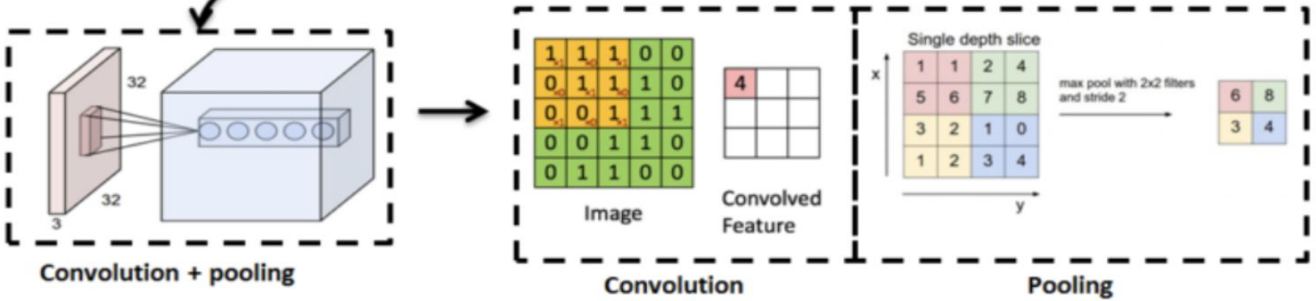
CNN VISUAL



A regular 3-layer fully connected Neural Network



Convolutional Neural Network

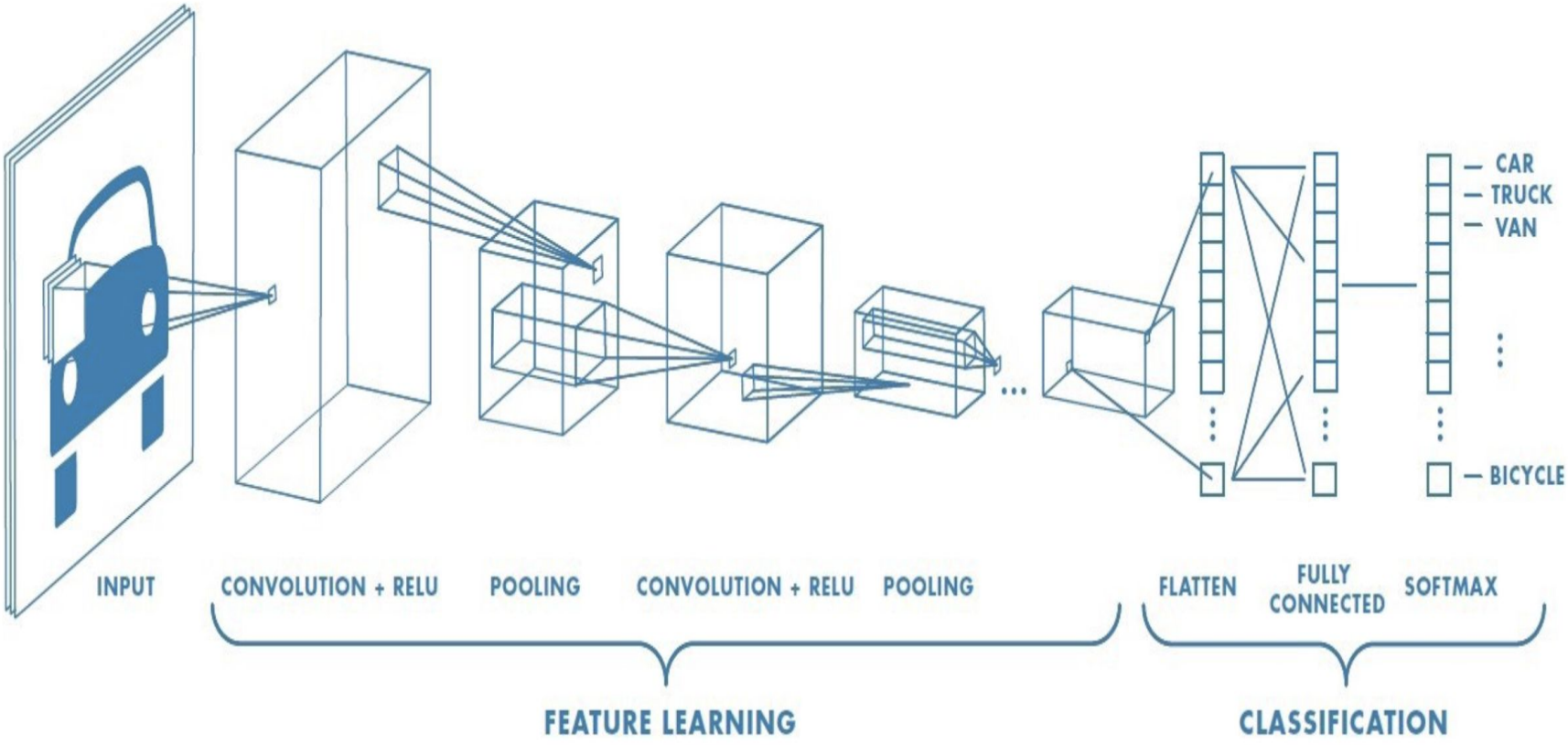


Convolution + pooling

Convolution

Pooling

Source: http://ufldl.stanford.edu/tutorial/images/Convolution_schematic.gif, http://cs231n.github.io/assets/nn1/neural_net2.jpeg, http://cs231n.github.io/assets/cnn/depthcol.jpeg, http://cs231n.github.io/assets/cnn/maxpool.jpeg


CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
FEATURE LEARNING
CLASSIFICATION

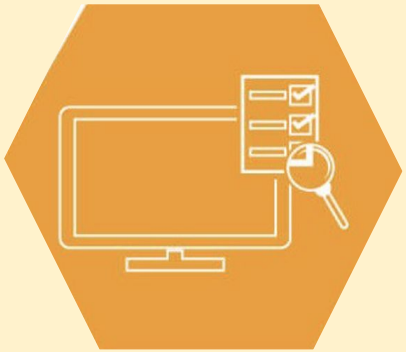
Train Model

Used a Convolutional Neural Network with five hidden layers

It took about two hours to fit the model

569k parameters

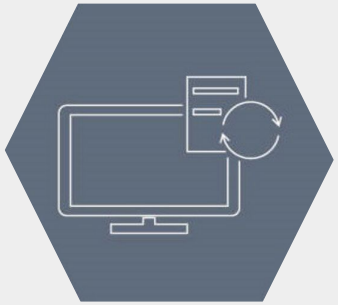
Test Data



Generated predictions on test data

We got 97.7% accuracy when predicting the speaker’s gender from previously known speakers

We got 100 new speakers and got **95%** accuracy when predicting the gender of never before heard speakers



Future Improvements

Use RNN

Gridsearch over CNN with Google Colab

Voting Classifier

Add more data / Take away data

Interactive

Thanks

Sources

https://ehq-production-us-california.imgix.net/3aa184fc578b28f2551285532bb84326503960ad/projects/images/000/002/142/original/running-effective-meeting.jpg?auto=compress%2Cformat&w=1080

https://cdn.ttgtmedia.com/rms/onlineImages/mobile_computing-mobile%20biometrics_05.png

https://machinelearningblogcom.files.wordpress.com/2017/11/1_kzmiuypmxgehhxx7slbp4w.jpeg?w=5400

https://git.generalassemb.ly/DSI-US-9/course-info