Tomayto, Tomahto

Speech recognition using Convolutional Neural Network
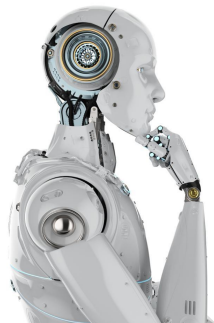
———

Tom Leung

Problem

The word "tomato" helps distinguish between American and British

There is no "tomayto, tomahto" equivalent in Mandarin Chinese, but there are phonetic differences depending on the speaker's origin (Taiwan vs China)

Objectives

1. Train a model to predict if a speaker is from Taiwan or China
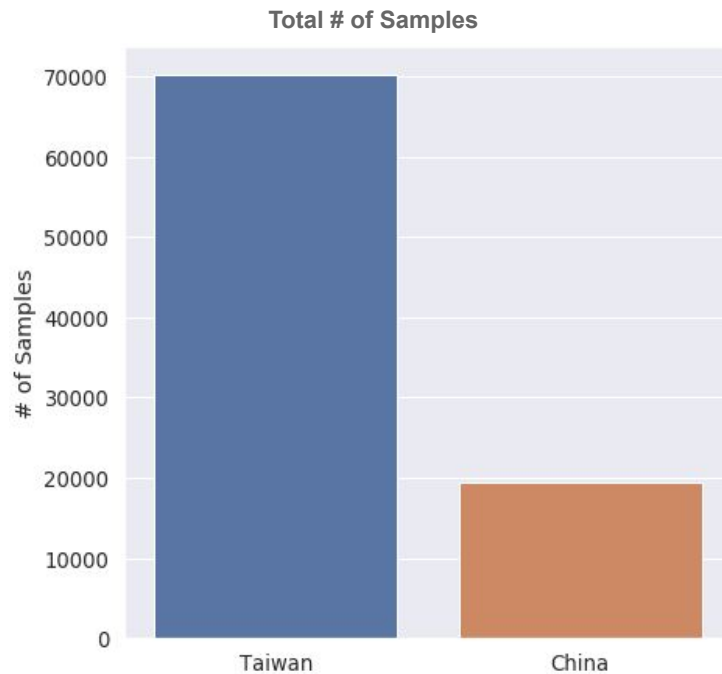2. Create a prototype serving the model

Data

- Obtained 90 hours of Chinese audio recordings from the Mozilla Common Voice project
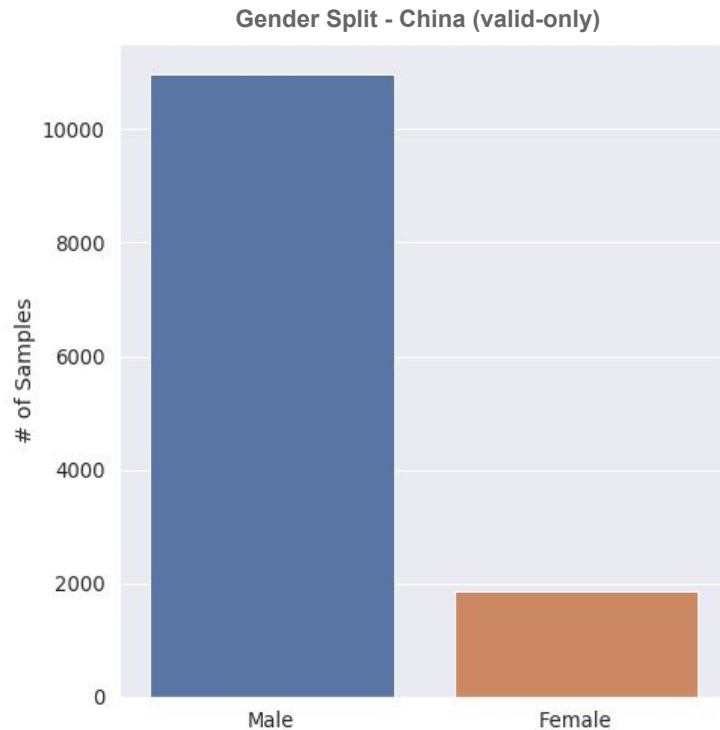- Recordings of volunteers speaking short sentences (3-4 seconds)

EDA

- Imbalance data
  - More recordings from Taiwan

EDA

- Imbalance data
  - More recordings from Taiwan
  - Gender (male > female)
- Curated my own origin-, and gender-balanced dataset
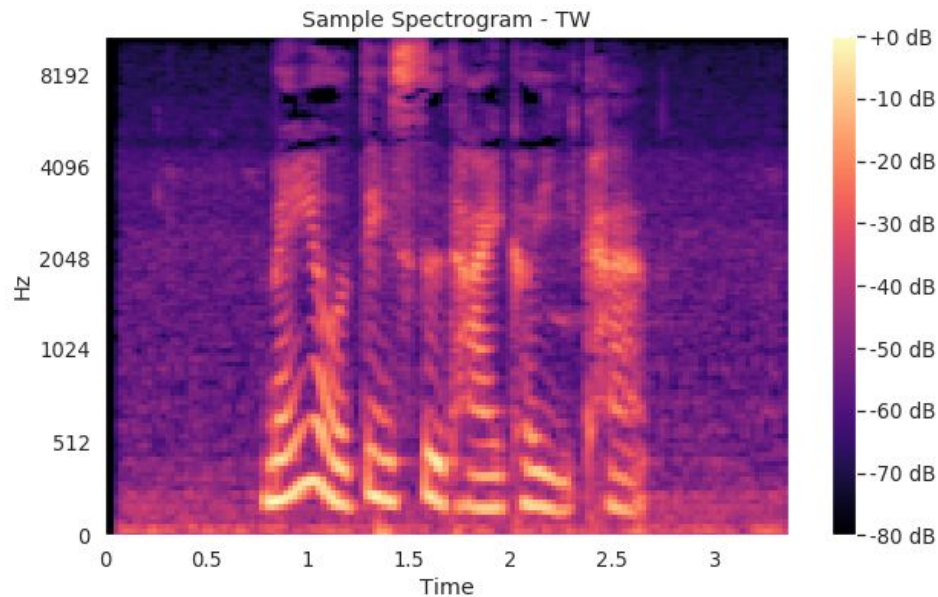 - 7.4k clips in total

Gender Split - China (valid-only)

Building the neural network

Pipeline

- Pre-processing
  - Convert audio into visual representation of sound
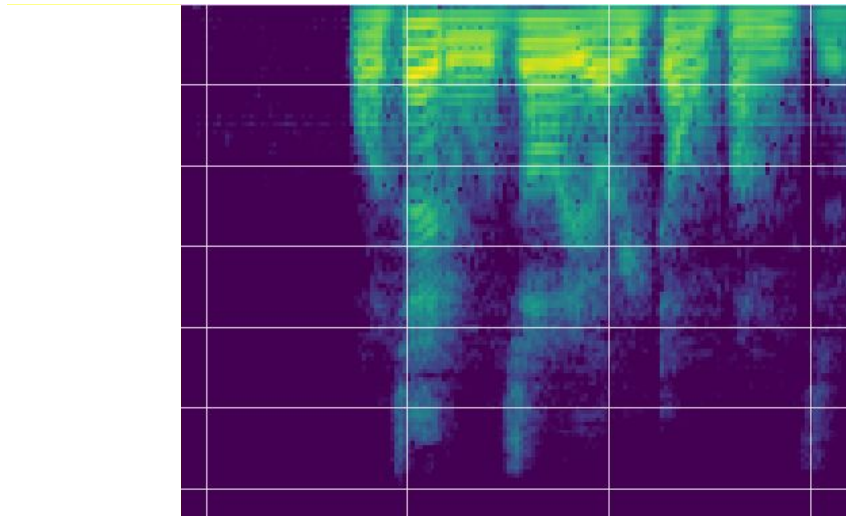 - Mel-scaled spectrograms
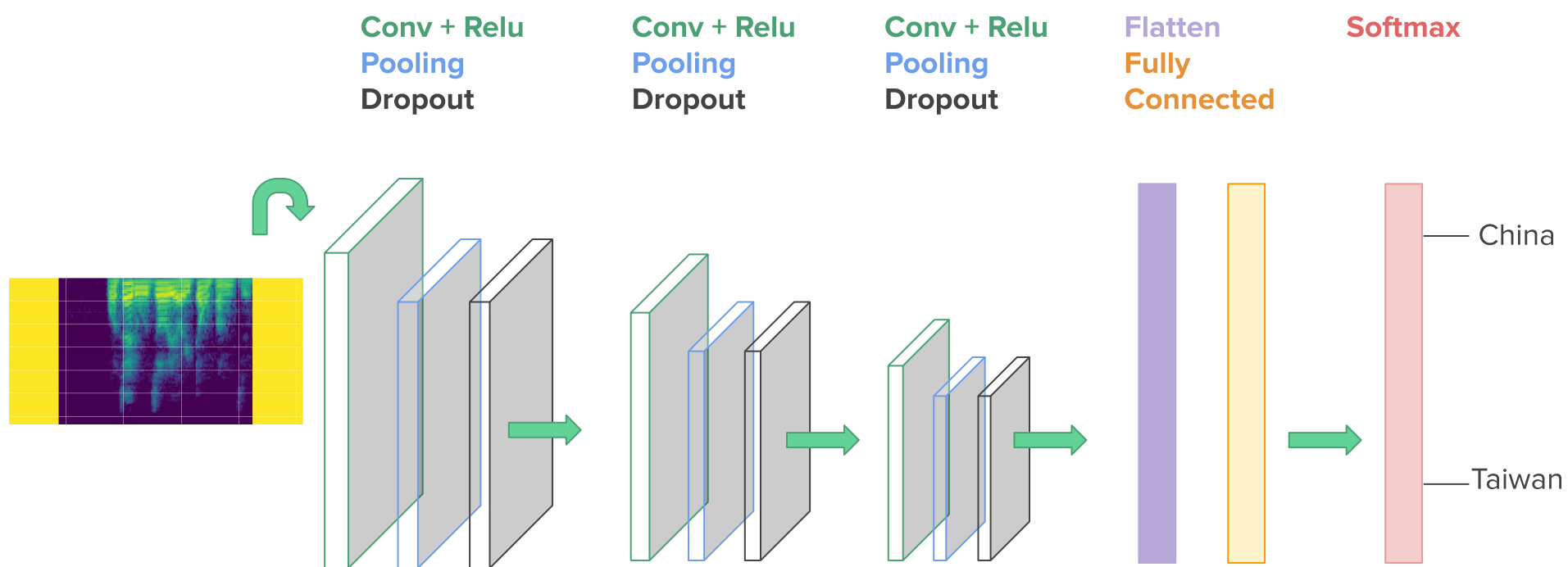

Sample Spectrogram - TW

Pipeline

- Pre-processing
  - Uneven clips
  - Trimmed or padded to 3 seconds
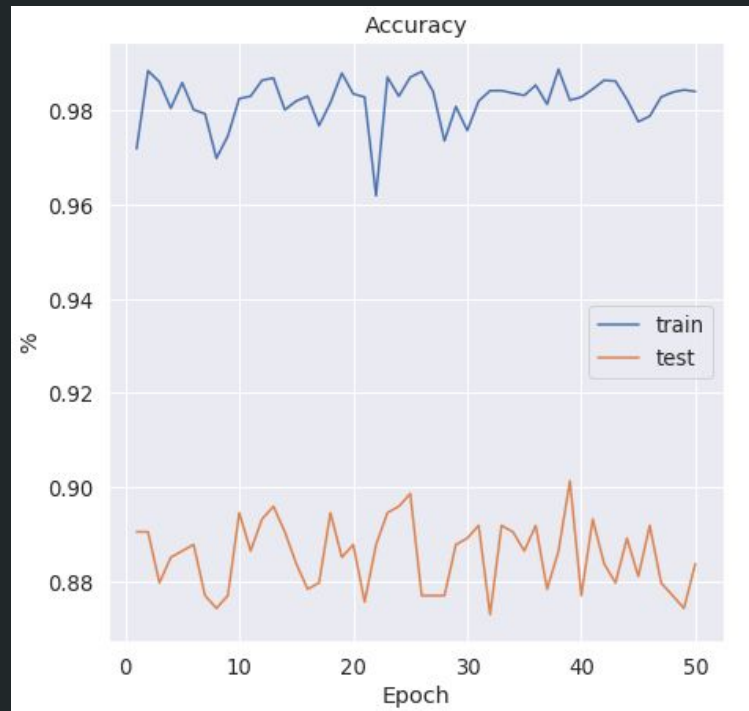- Identically shaped images ready as input

128

256

CNN Structure

Results

90% val accuracy (test)
90% val accuracy (hold-out)

Serving the model

Web demo which allows users to:

1. Record a short clip and upload to the server

2. It makes the prediction using the tensorflow model in the back-end

Conclusion

Modeling

- Neural Networks are difficult to set-up but powerful
- Less is more with epoch

Demo
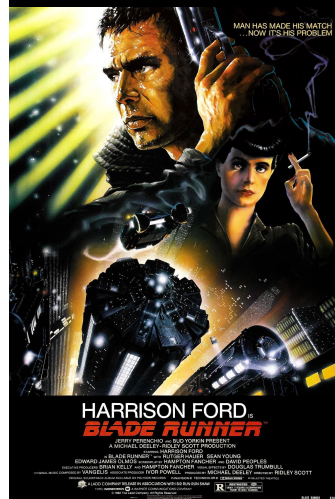
- Field-tested with friends with good results
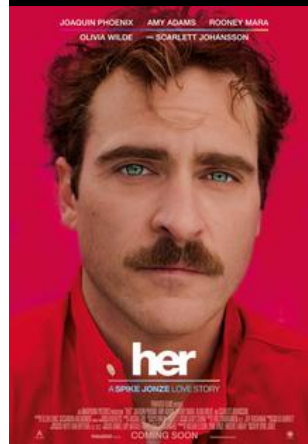
Overall

- Satisfying results given the data

What's next?

- Include other languages, dialects, accents
- Personalized search content
- Auto tag uploaded content

What's next?

- Build talking robots

Demo available now @

https://tomayto-tomahto.herokuapp.com/

Github: tchleung/tomayto_tomahto
Email: tom.ch.leung@gmail.com
Linkedin: linkedin.com/in/tom-ch-leung/

Thank you!