A DNN Model for Speaker Identification: Voice to Face Matching

Unveiling the Relationship Between Voice and Facial Features

Objective:

- Binary classification - Given 1 face embedding and 1 voice embedding of a main entity and 1 face embedding of a different entity should return the ordinal of the main entity (0/1).

Main Algorithm flow:

- Create the training dataset - create the input triplets (v(i),f(i),f(j)) from the voice and image embeddings pickle files.
- Design a simple MLP model.
- Train the model.
- Evaluate the model.

Dataset Preparation

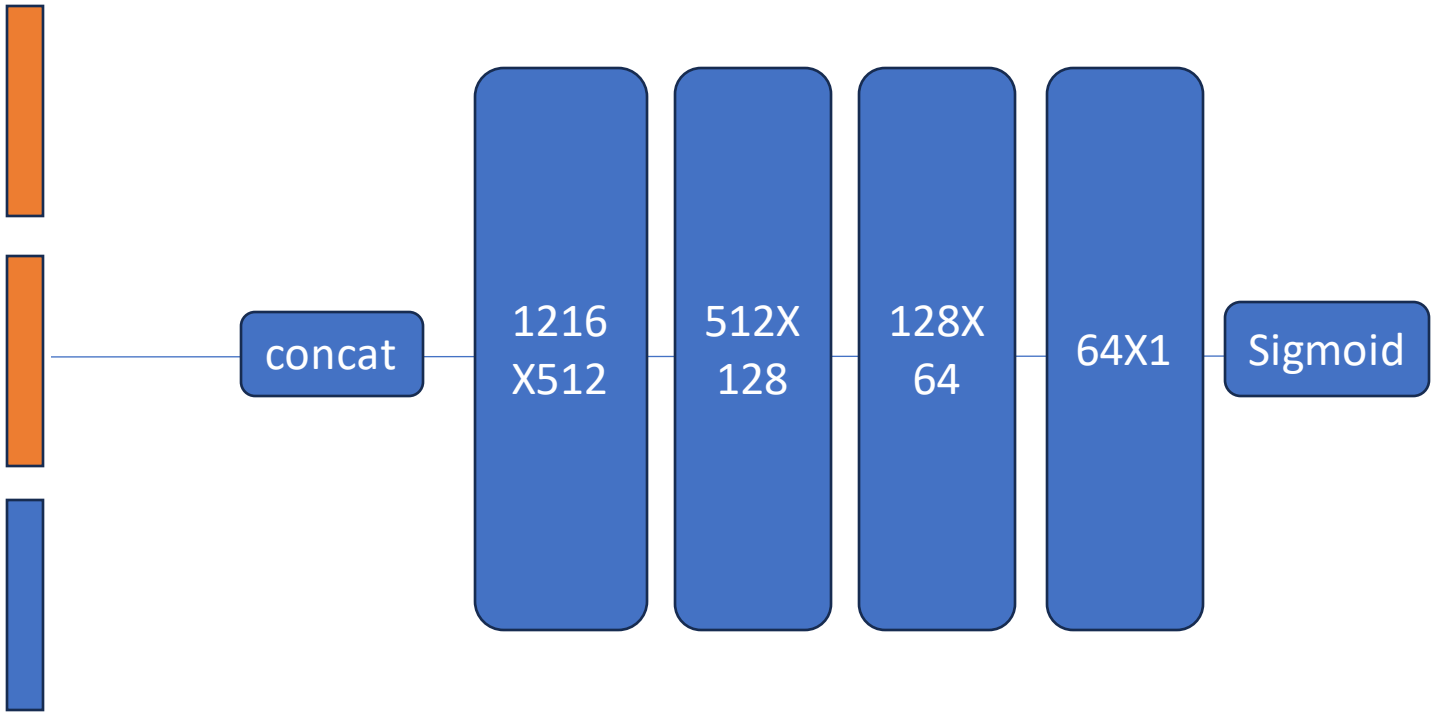
- As some of the embedding keys were not in ascii format, the matching between audio and image embeddings was not working properly and the dataset size was too small. I have added a conversion of speaker names to Unicode using the '**unidecode**' library which solved the problem.
- The dataset contains triplets of the form: {voice_i, face_i, face_j} (as in 'Seeing Voices and Hearing Faces: Cross-modal biometric matching' paper)
- The file 'dataset_construction.py' contains the functionality for creating the triplets- for each speaker and for each of his audio and image embeddings a new triplet is form where the negative sample is randomly sampled from all other speakers embeddings.

Feature Level Augmentation

- Feature-level augmentation for both voice and face embeddings before passing them into the deep neural network can enhance the model's robustness and generalization by simulating variability in the data.
- A random noise, scaling, and shifting were applied. These operations simulate variations in the voice / image data that might occur naturally.

Model Architecture

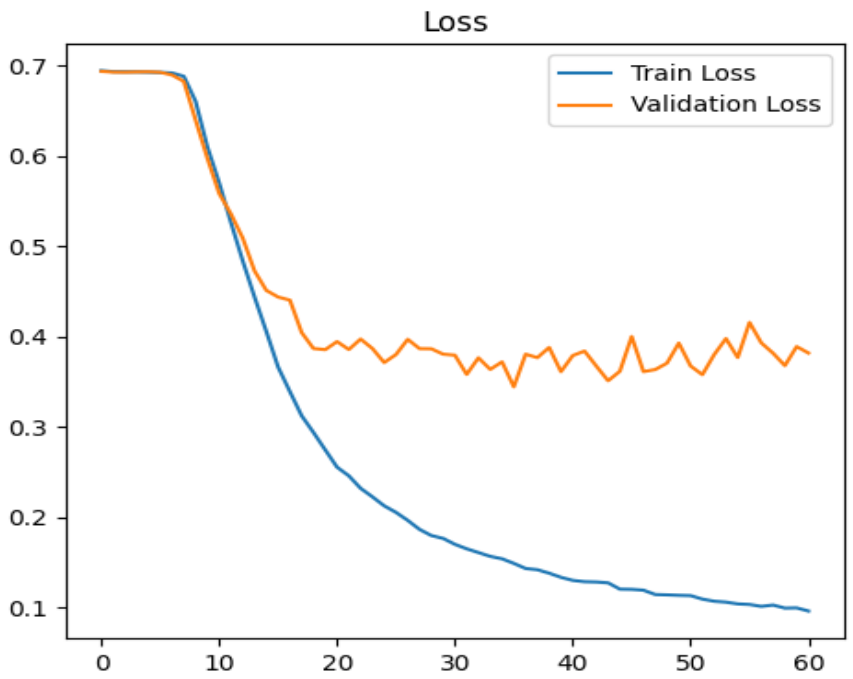
- I used a simple model with 3 fully connected hidden layers and a sigmoid output layer (512,128,64,1)

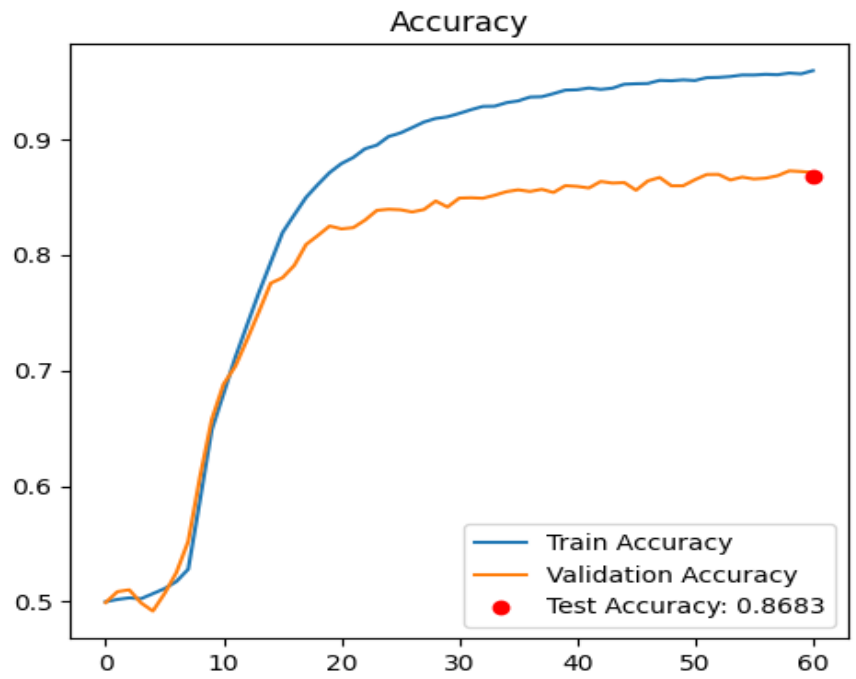


Model training

- BCE loss
- 50 epochs
- Learning rate of .001
- Adam optimizer

Model Performance





Identification Accuracy on a small test set: .86

Overfitting starts after ~20 epochs

Conclusion

- In future work I might plan to deal with the overfitting by adding regularization and/or more complexity to the model.
- The relatively high accuracy (almost .4 above base probability of .5) is quite remarkable and shows the potential of V2F mapping and suggest that there is a strong correlation between a person voice and face.

That's it

- Thanks!