W251 | Door Lock / Audio-Visual Authentication

Carlos Sancini
Madia Taher
Matt Brimmer
Xander Hathaway

April 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Product

- Multimodal, double-security audio-visual edge authentication

- Namesake case: Home Door Lock

End-User Problem / Motivation

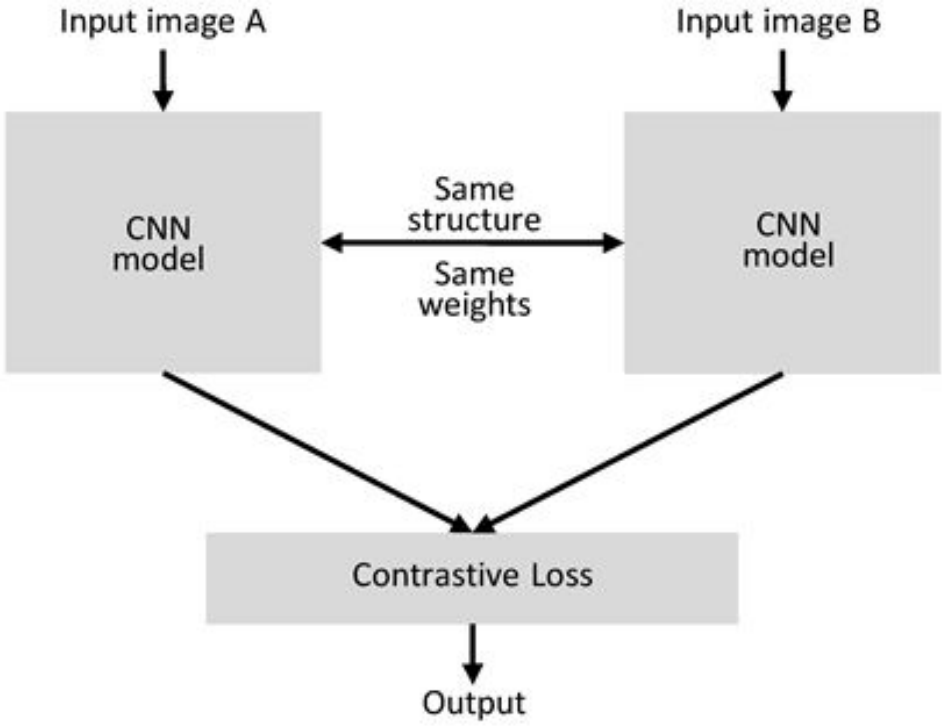
- Privacy - users are wary of sending sensitive biometric data to be stored in cloud

- Security - Many edge devices leverage facial images for authentication. As ubiquity increases, so will attacks against them.

Approach to the problem

- Typical machine learning tasks rely on large datasets
- The door lock task implies just a few observations for a given user, in both audio and video perspectives
- This type of task is usually referred as One-Shot learning

One-Shot learning is solved with a Siamese NN



- The input data is structured in pairs
- Positive pairs contains observations of the same person
- Negative pairs contains mismatched people
- The model learns to output a similarity score
- During inference the model calculates a similarity score for a input pair never seen before

The Siamese model in the context of the door lock system

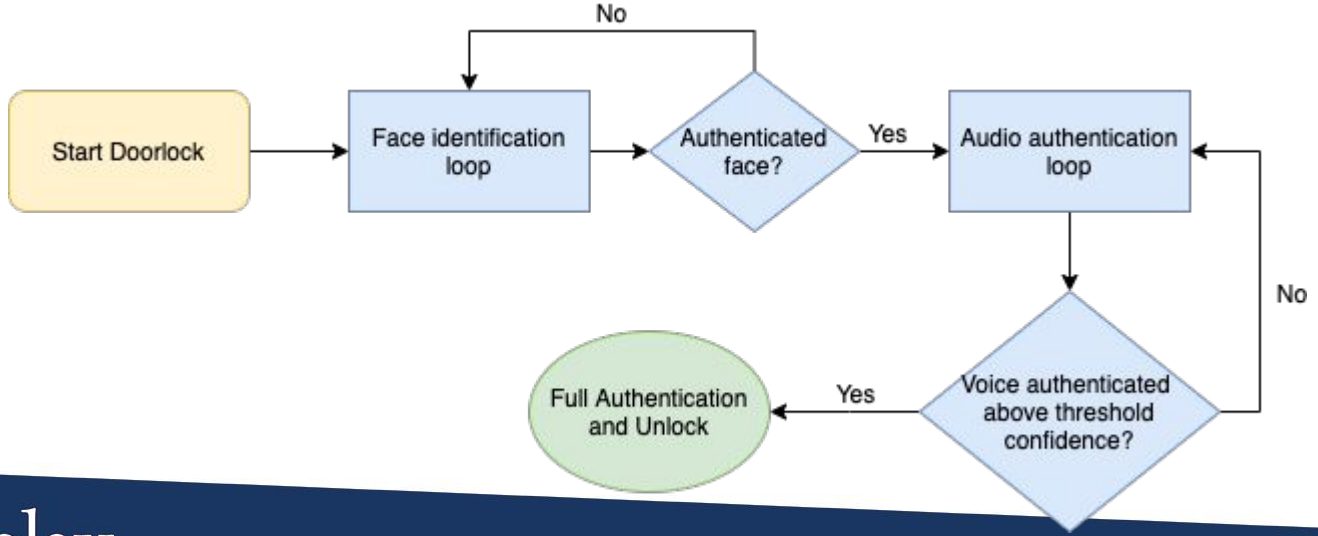
A reference speech and picture of the user is captured during the setup of the door lock system

The edge device captures the user's face and voice and calculates two similarity scores

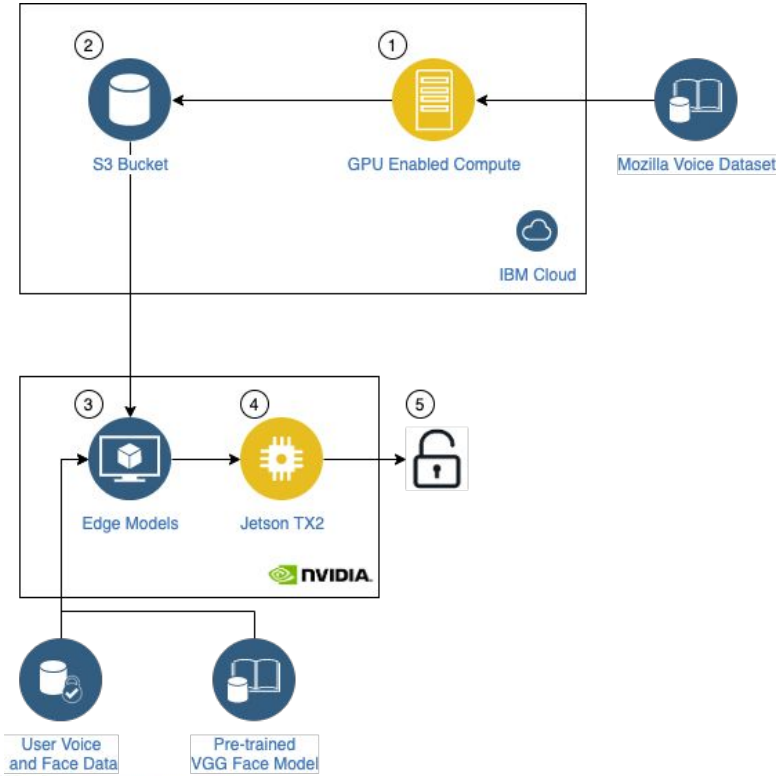
The door is unlocked if the combined speech and face scores achieve a certain threshold

Minimum Viable Product / Design

- Video model authenticating from visual standpoint
- Audio model authenticating from audio standpoint
- Overall program managing authentication
- Cloud models leveraging one-shot learning for new users



Overall architecture



1. Model training: Train audiomodel using powerful GPU enabled compute using Mozilla Voice Dataset

2. Store model weights: Trained model information stored in S3

3. Retrieve model weights: Pull down audio model weights from S3 and pull in pre-trained VGG Face model for inference on edge device

4. Model inference: Using pre-trained models and local user data, perform authentication inference on NVIDIA Jetson TX2

5. Authenticate users: Based on model output, optionally authenticate users

Data and Models Training

Audio
Trained a new model from scratch using the Mozilla Common Voice dataset (51k voices, 38Gb)

Visual
Leverage pre-trained model from VGG Face dataset (9,000 identities, 3.3M Faces)

Audio Model Training

- 9 hours to prepare the input pairs: 16 threads to parallelize workload, 100GB of memory
- 10 hours to train 200 epochs in a V100 GPU
- Validation Accuracy: 98.7%
- Test Accuracy: 90.5%
- Test F1: 0.8

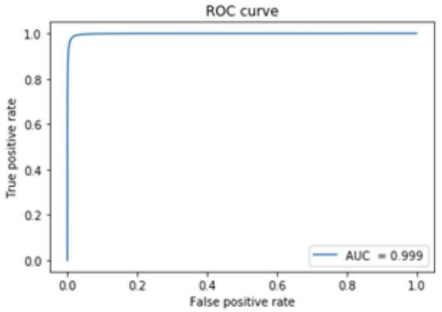


Figure 2: ROC curve for validation set

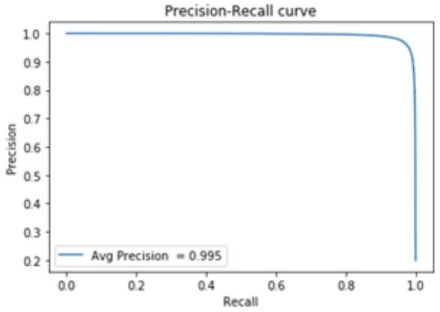


Figure 3: Precision-Recall curve for validation set

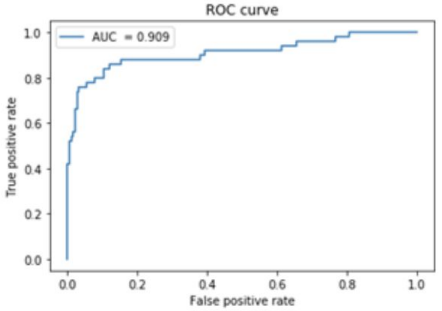


Figure 4: ROC curve for test set

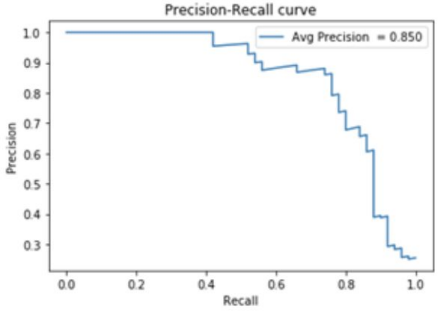


Figure 5: Precision-Recall curve for test set

Demonstration

Authentication Successful Video (1:25)

Authentication Unsuccessful Video (1:24)

Next Steps

- User interface (uploading files)
- Smarter audio capture
- Model improvement (both audio and visual)
- Multiple language support
- Spoofing

Questions?


SMILE! OUR AI WILL TAKE A PICTURE OF YOUR FACE AND WILL KNOW EXACTLY WHAT YOU WOULD LIKE FOR DINNER TONIGHT.
KDNUGGETS.COM