Machine Learning Nanodegree Capstone Proposal

Abdullah AlOthman

February 27, 2018

Proposal

Domain Background

Speaker Verification (or Authentication) is a way of confirming the identity of a person using voice characteristics, *not to be confused with Speaker Identification (trying to find the identity of a person given their voice)*.
There are two main approaches to speaker verification.

- text dependent:
  - requires the speaker to say a specific phrase
- text independent:
  - no constraint on spoken phrases

for this project I want to do text independent speaker verification in *'wild'* environments (uncontrolled audio environments)

Machine Learning has been applied in this task, most notably in SRI's Speakers in the wild challenge and NIST's 2012 speaker recognition challenge

Problem Statement

The problem I want to tackle in this project is speaker verification, (*classifying whether or not SpeakerA and SpeakerB are the same person*) in noisy, uncontrolled environments.
Applications of speaker verification include:

- authentication in high-security systems
- authentication in forensic tests
- searching for persons in large corpora of speech data

All of these applications require reliable performance in *'wild'* conditions (i.e. real-world scenarios). The challenges expected in this task come from two main fronts:

1. high variance in the environment
   - background music
   - crowds
   - recording quality
   - etc.
2. high variance in the speaker
   - age
   - accent
   - emotion
   - intonation
   - etc.

Datasets and Inputs

I'll be working with the VoxCeleb dataset in this project.

- VoxCeleb contains over **100,000 utterances** for **1,251 celebrities**, extracted from videos uploaded to YouTube in WAV format.
- The dataset is gender balanced, with 55% of the speakers male.
- The speakers span a wide range of different ethnicities, accents, professions and ages.
- I'll be using the same train/test split used in the paper to compare with the benchmark results accurately.

Videos included in the dataset are shot in a large number of challenging multi-speaker acoustic environments.
These include:

- red carpet.
- outdoor stadium.
- quiet studio interviews .
- speeches given to large audiences.
- excerpts from professionally shot multimedia.
- videos shot on hand-held devices.

Crucially, all are degraded with real world noise, consisting of

- background chatter
- laughter
- overlapping speech
- room acoustics

and there is a range in the quality of recording equipment and channel noise.

Solution Statement

As a solution to this project, I'm planning to develop a Deep Learning Classifier
that verifies whether two speakers are the same person or not.

I'll be experimenting with multiple methods of audio feature extraction such as:

- GMM
- JFA
- MFCC
- I-vectors
- bottle-necking
- extraction through CNNs

and tryout various configurations of RNNs and CNNs for the neural-net model to achieve the best performance I can get.

Benchmark Model

As a benchmark, I'll compare my model to two main approaches.

1. Random Classifier: the most basic benchmark, but important to know we're preforming better than random chance.
2. The CNN architecture proposed in the VoxCeleb paper.

Evaluation Metrics

I'll be using the Equal Error Rate `(EER)` as my evaluation metric.
EER is the he rate at which both acceptance and rejection errors are equal. this metric is commonly used in biometric problems and it is the same metric reported in the VoxCeleb paper which would make for a fairer comparison.

Project Design

Data Collection

The data can be:

- collected using a bash script that extracts the audiofiles from youtube and seperating the files into the specific audio utterances
- or by requesting a direct download from Arsha Nagrani.

I've already acquired the dataset using both methods.

Feature Extraction

first I'll begin by choosing a baseline model to test the features on. After testing different feature representations (mentioned above) and deciding on which one to use, I'll move on to
modifying the model's architecture

Model Architecture

I plan to build my model based on:

- transfer Learning (reworked implementations of VGG and ResNet)
- CNN architectures (RCNN, SCNN, VDCNN..)
- RNN architectures (GRU, LSTM, Bidirectional..)

and combinations thereof and see which model will preform best in this problem.

References

- Nagrani, Arsha et al. “VoxCeleb: a large-scale speaker identification dataset.” INTERSPEECH (2017).
- M. McLaren, L. Ferrer, D. Castan, and A. Lawson, “The speakers in the wild (SITW) speaker recognition database,” INTERSPEECH, vol. 2016, 2016.
- C. S. Greenberg, “The NIST year 2012 speaker recognition evaluation plan,” NIST, Technical Report, 2012.