Machine Learning Engineer Nanodegree

Capstone Proposal

Abdalla Shaaban

September 7, 2018

TensorFlow Speech Recognition Challenge (Kaggle [1] competition)

Proposal

Domain Background

Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition(ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Some speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called “speaker independent” systems. Systems that use training are called "speaker dependent".

Speech recognition applications include voice user interfaces such as voice dialing (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), domotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed direct voice input).

Problem Statement

- We might be on the verge of too many screens. It seems like every day, new versions of common objects are “re-invented” with built-in Wi-Fi and bright touchscreens. A promising antidote to our screen addiction is voice interfaces.

  But, for independent makers and entrepreneurs, it’s hard to build a simple speech detector using free, open data and code. Many voice recognition datasets require preprocessing before a neural network model can be built on them. To help with this, TensorFlow recently released the Speech Commands Datasets. It includes 65,000 one-second long utterances of 30 short words, by thousands of different people.

- The goal of this project is to recognize a simple speech commands.

Datasets and Inputs

- This project will use audios from the publicly available Speech Commands Data Set v0.01 dataset. This is a set of one-second .wav audio files, each containing a single spoken English word. These words are from a small set of commands, and are spoken by a variety of different speakers. The audio files are organized into folders based on the word they contain, and this data set is designed to help train simple machine learning model.

- This table shows how many recordings of each word are present in the dataset.

Word	Number of utterances
- Down	2359
- Go	2372
- Left	2353
- No	2375
- Off	2357
- On	2367
- Right	2367
- Silence	2015
- Stop	2380
- Unknown	2418
- Up	2375
- Yes	2377

Figure 1: How many recordings of each word are present in the dataset

- So the dataset is balanced.

Solution Statement

- This is classification problem, inputs are audio record and the output is the command
- In this project, I will use convolution deep learning network that can be trained on the training set.

Benchmark Model

- I will use the tensorflow audio recognition model [2] as a benchmark. The accuracy of this model is between 85% and 90%

Evaluation Metrics

- The evaluation metric for the model will be evaluated on Multiclass Accuracy, which is simply the average number of observations with the correct label

Project Design

- The first stage of the project will be to download the dataset.
- represent audio for speech recognition (ex: using the Mel-frequency cepstrum)
- The network will consist of several convolution layer which may be followed by pooling or normalization layers, followed by fully connected layers, and finally softmax classifier.
- The final evaluation of the model will be determined by computing the accuracy of the predictions made against the test set.
- Once the model has been fully trained and evaluated, thee weights will be frozen and extracted.

Reference

1- https://www.kaggle.com/c/tensorflow-speech-recognition-challenge
2- https://www.tensorflow.org/tutorials/sequences/audio_recognition