Spoken Keyword Spotting

Vineeth S

May 2020

What is Spoken Keyword Spotting?

Spoken Keyword Spotting is the task of identifying predefined words (called as keywords) from speech. Keyword spotting has wide range of applications from device wake-up (OK Google, Hey Siri etc) to hands-free control of devices.

Dataset

- For the of this project I have used Google Speech Commands Dataset (`https://arxiv.org/abs/1804.03209`)

Dataset

- For the of this project I have used Google Speech Commands Dataset (`https://arxiv.org/abs/1804.03209`)
- Speech Commands dataset has 65,000 one-second long utterances of 30 short words by thousands of different people

Dataset

- For the of this project I have used Google Speech Commands Dataset (`https://arxiv.org/abs/1804.03209`)
- Speech Commands dataset has 65,000 one-second long utterances of 30 short words by thousands of different people
- For the pilot implementation, I have used 10000 utterances

Dataset

- For the of this project I have used Google Speech Commands Dataset (`https://arxiv.org/abs/1804.03209`)
- Speech Commands dataset has 65,000 one-second long utterances of 30 short words by thousands of different people
- For the pilot implementation, I have used 10000 utterances
- The dataset is designed build basic but useful voice interfaces for applications, with common words like “Yes”, “No”, digits, and directions etc

Progress Made

- Developed an understanding how Keyword detection is implemented

Progress Made

- Developed an understanding how Keyword detection is implemented
- Developed a basic skeleton code in python for this purpose

Progress Made

- Developed an understanding how Keyword detection is implemented
- Developed a basic skeleton code in python for this purpose
- As of preparing this presentation, the model achieves an accuracy of 94% on the test data on classification

Model Specifications

- Input: Tensorflow Dataset Object with features and labels
  - I have experimented with MFCC, and Log Filterbank Energies as of now
  - Labels belong to the 30 categories
- Layer CNN : To obtain the spatial dependencies
- Layer LSTM : To obtain the temporal dependencies
- Layer Attention Layer: To use attention mechanism
- Output: One of the 30 class labels

Future plan of work

- Try exploring with other input features

Future plan of work

- Try exploring with other input features
- Experiment with different architectures

Future plan of work

- Try exploring with other input features
- Experiment with different architectures
- Extend it to continuous speech signal
  - I have not tried providing an individual utterance feature and examine its computational footprint
  - If the footprint is small, we could simply slide a window over our speech signal and use the model to identify the keyword (if any)

The End

Questions? Suggestions?