Research Project Proposal:Speech Recognition

Jianxi Li, Xinyi Liao
Feb 8,2018

Topic Description:

With the rapid development of technology, we might be on the verge of too many screens. It seems like everyday, new versions of common objects are “re-invented” with built-in wifi and bright touchscreens. A promising antidote to our screen addiction are 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 released the Speech Commands Datasets. It includes 65,000 one-second long utterances of 30 short words, by thousands of different people.

In this project, we tend to use the Speech Commands Dataset to build an algorithm that understands simple spoken commands. By improving the recognition accuracy of open-sourced voice interface tools, we can improve product effectiveness and their accessibility.

Background:

Speech recognition is in the interdisciplinary field of computational linguistics that develops methodologies and technologies that enable the recognition and translation of spoken language into text by computers. The recognition of speech requires the knowledge from linguistics, computer science, and electrical engineering.

Some speech recognition systems need to be trained. It means that the sounds of text or isolated vocabulary read by individual speaker need to input to the system in order to increase the accuracy of the recognition.

The speech recognition has been researched for a long time, and recently the field has benefited from advances in deep learning and big data. With the help of these technologies, not only the amount of academic papers published in the field has a boost, but more importantly is that the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems. Nowadays, technology companies such as Apple, Google, Microsoft and Amazon have publicized the core technology in their speech recognition systems as being based on deep learning.

Data source:

The dataset we used are a set of .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 models.

This data was collected by Google and released under a CC BY license. It includes 65,000 one-second long utterances of 30 short words.All the audio files were all collected using crowdsourcing.

Algorithms source:

Hidden Markov Model (HMM) https://en.wikipedia.org/wiki/Hidden_Markov_model
Dynamic Time Warping (DTW) https://en.wikipedia.org/wiki/Dynamic_time_warping
Convolutional Neural Network(CNN) https://en.wikipedia.org/wiki/Convolutional_neural_network
Recurrent Neural Network (RNN) https://en.wikipedia.org/wiki/Recurrent_neural_network
End-to-End Speech Recognition http://proceedings.mlr.press/v48/amodei16.pdf
K-means clustering https://en.wikipedia.org/wiki/K-means_clustering
DBSCAN clustering https://en.wikipedia.org/wiki/DBSCAN
Naive Bayes classifiers *https://en.wikipedia.org/wiki/Naive_Bayes_classifier*
Support Vector Machine https://en.wikipedia.org/wiki/Support_vector_machine
Linear discriminant analysis (LDA) https://en.wikipedia.org/wiki/Linear_discriminant_analysis

References :

[1] Tara N. Sainath, Carolina Parada,"Convolutional Neural Networks for Small-footprint Keyword Spotting",2015
[2] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. Le, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Ng, "Large Scale Distributed Deep Networks," in Proc. NIPS, 2012
[3] Bahdanau, Dzmitry, et al. "End-to-End Attention-based Large Vocabulary Speech Recognition." (2015):4945-4949.
[4] Hannun, Awni, et al. "Deep Speech: Scaling up end-to-end speech recognition." Computer Science (2014)
[5] Matthew Rubashkin, Matt Mollison, "Building, Training, and Improving on Existing Recurrent Neural Networks", 2017
[6] Aäron van den Oord,Sander Dieleman, Heiga Zen, WaveNet: A Generative Model for Raw Audio,2016
[7] Simple Audio Recognition | TensorFlow. (n.d.). , from https://www.tensorflow.org/versions/master/tutorials/audio_recognition