Project Report

Junaid Ahmad Bhat

February 8, 2021

Overview

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, is a capability which enables a program to process human speech into a written format. While it's commonly confused with voice recognition, speech recognition focuses on the translation of speech from a verbal format to a text one whereas voice recognition just seeks to identify an individual user's voice.

The most frequent applications of speech recognition within the enterprise include the use of speech recognition in mobile devices.Speech recognition can also be found in word processing applications like Microsoft Word, where users can dictate what they want to show up as text.For example, individuals can use this functionality in smartphones for call routing, speech-to-text processing, voice dialing and voice search. A smartphone user could use the speech recognition function to respond to a text without having to look down at their phone. Speech recognition on iPhones, for example, is tied to other functions, like the keyboard and Siri and Bixby on samsung android mobile phones also work on speech recognition. If a user adds a secondary language to their keyboard, they can then use the speech recognition functionality in the secondary language (as long as the secondary language is selected on the keyboard when activating voice recognition. To use other functions like Siri, the user would have to change the language settings.)

From smart lights that turn on or off on your command, Google Home Assistant that can place space trivia with you and make financial transactions when requested, Alexa that also works on same algorithm to follow your voice commands; speech recognition is the component of the system that makes it all possible.

Data

Record 80 utterances of each command.

Set 16KHz as sampling rate.

Save samples

Record/Forward

Record/Back

Record/LEFT

Record/Right

Record/STOP

Description

Using Convolutional layers ahead of LSTM is shown to improve performance in several research papers.

BatchNormalization layers are added to improve convergence rate.

Using Bidirectional LSTM is optimal when complete input is available. But this increases the runtime two-fold.

Final output sequence of LSTM layer is used to calculate importance of units in LSTM using a FC layer.

Then take the dot product of unit importance and output sequences of LSTM to get Attention scores of each time step.

Take the dot product of Attention scores and the output sequences of LSTM to get attention vector.

Add an additional FC Layer and then to output Layer with SoftMax Activation.

The model is successfully built and has achieved the highest accuracy of 99

Run

The Code is written in Google Colab.

Open ColabNotebook.ipynb and change Runtime to GPU

Upload **/content/drive/MyDrive/Record/** to Colab.

Run the cells in the same order as in Notebook.

Test

Locate the folder where you save your model.h5 file.

Start speaking when you see mike in the bottom right pane of the task bar or see red blinking dot in the title bar.

If still facing some error try to set the mike recorder time in the previous record function up to 3 sec.