Speech Command Recognition

February 14, 2021

- **OVERVIEW:**

  Speech recognition is the process of converting human sound signals into words or instructions. It is based on speech. It is an important research direction of speech signal processing and a branch of pattern recognition.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”), domestic appliance control, search key words (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number). In similar way Speech voice recognition model is based on concepts of Convolution, LSTM , Attention and recognise pretrained voice with accuracy of **99.6%**..

- **DATA:**

  1: Set 16KHz as sampling rate.

  2: Record 80 utterances of each command.

  3: Save samples of each command in different folders.

* Data/forward.
* Data/back.
* Data/left.
* Data/right.
* Data/stop.

- **Description:**

  1: Using Audacity software we have generated the data set i.e voice commands Forward, Back, Left, Right, Stop with 80 utterances for each.

  2: We have splitted the data into two categories i.e Train and Test. where 20% of the data set are used for testing and the rest for training.The shape of the data set which is used for testing will be **(1775, 49, 39, 1)** and for training set, shape will be **(7033, 49, 39, 1)**

  3: We have also used Melspectogram for extracting features from the data set and also normalization is used so that the model will achieve the converge point point within few epochs.

  4: LSTM neural network is also used so that our weights will get updated without showing any **vanishing gradient problem** during training of models.

  5: Attention is used to get the required data even from a complex sentence, Also batch normalization is used in order to prevent unexpected behaviour of the weights.

  6: Finally an audio is recorded from user and the model will predict this data based on the already given training and the desired result will be displayed on screen.

- **The model is successfully built and has achieved the highest accuracy of 99.6%**

- **Model Summary**

```
Model: "Attention"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
Input (InputLayer)              [(None, 49, 39, 1)]  0
__________________________________________________________________________________________________
Conv1 (Conv2D)                  (None, 49, 39, 10)   60          Input[0][0]
__________________________________________________________________________________________________
BN1 (BatchNormalization)        (None, 49, 39, 10)   40          Conv1[0][0]
__________________________________________________________________________________________________
Conv2 (Conv2D)                  (None, 49, 39, 1)    51          BN1[0][0]
__________________________________________________________________________________________________
BN2 (BatchNormalization)        (None, 49, 39, 1)    4           Conv2[0][0]
__________________________________________________________________________________________________
Squeeze (Reshape)               (None, 49, 39)       0           BN2[0][0]
__________________________________________________________________________________________________
LSTM_Sequences (LSTM)           (None, 49, 64)       26624       Squeeze[0][0]
__________________________________________________________________________________________________
FinalSequence (Lambda)          (None, 64)           0           LSTM_Sequences[0][0]
__________________________________________________________________________________________________
UnitImportance (Dense)          (None, 64)           4160        FinalSequence[0][0]
__________________________________________________________________________________________________
AttentionScores (Dot)           (None, 49)           0           UnitImportance[0][0]
                                                                 LSTM_Sequences[0][0]
__________________________________________________________________________________________________
AttentionSoftmax (Softmax)      (None, 49)           0           AttentionScores[0][0]
__________________________________________________________________________________________________
AttentionVector (Dot)           (None, 64)           0           AttentionSoftmax[0][0]
                                                                 LSTM_Sequences[0][0]
__________________________________________________________________________________________________
FC (Dense)                      (None, 32)           2080        AttentionVector[0][0]
__________________________________________________________________________________________________
Output (Dense)                  (None, 5)            165         FC[0][0]
--------------------------------------------------------------------------------------------------
```

- **RUN:**

  The Code is written using Google Colab:

1. Open ColabNotebook.ipynb and change Runtime to GPU.
2. Upload Speech-Recognition/Speech to Colab.
3. Change data-dir in all cells to point to Speech-Recognition/speech.
4. Run the cells in the same order in Notebook Test.

- **TEST:**

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

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

- **Language Used:**

  PYTHON

- **Libraries and Packages Used:**

  KAPRE, SCIKIT LEARN, SOUND FILE, TENSORFLOW.