Speech Command Recognition [1]

Project done as part of course EE5600

Presentation by
Virendra
AI20MTECH01003

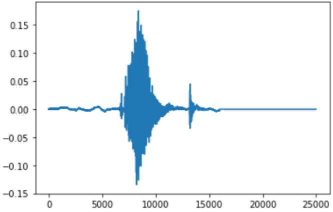
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

September 26, 2021

Overview

Problem Statement

- Design and implementation of a speech command recognition framework
- In this work we train and test a neural attention model [1] on a speech command dataset prepared from scratch



(a) Example speech command file

Proposed Methodology

Neural Attention Network:

- Neural attention network is an RNN with attention mechanism
- The input to the model is the mel-scale spectrogram (we will discuss its construction in Dataset section)
- The two initial layers are 2D convolutional layers which extract local temporal relations from the input. The conv layer's output is then fed to bidirectional LSTM (B-LSTM). B-LSTM captures long term (global) relationships from the input
- B-LSTM's output is a 1D vector whose weighted average is then fed to FCN. FCN output is used to classify the audio files into commands

Dataset

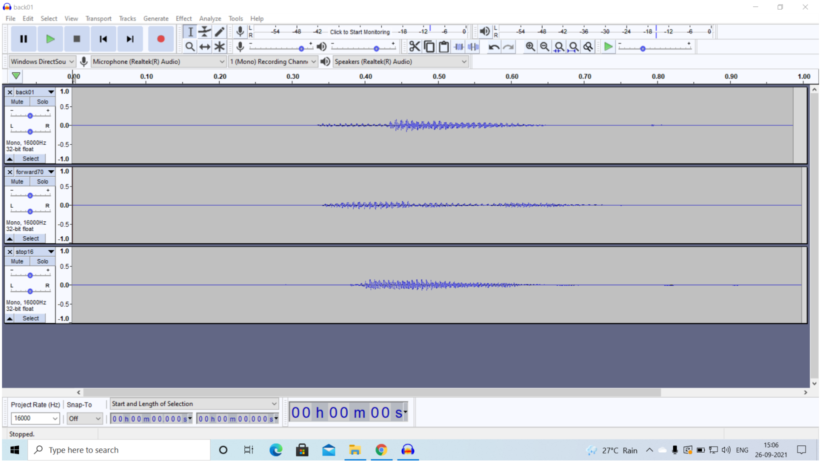


Figure 3: Snapshot of data collection

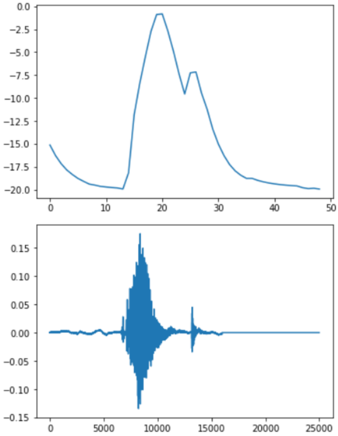
Dataset (cont.)

- The speech command dataset contains 400 .wav audio files for 5 speech commands viz. forward, back, right, left and stop
- The audio is recorded at sampling rate 16KHz and .wav file is stored in 32 bit format
- Neural attention network takes mel-scale spectrogram as input and hence .wav files are converted to mel-scale spectrogram using kapre library

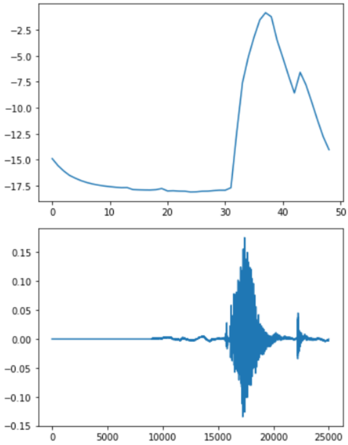
Model Training and Validation:

```
Epoch 1/10
286/286 - 20s - loss: 1.3585e-09 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0261 - val_sparse_categorical_accuracy: 0.9964
Epoch 2/10
286/286 - 20s - loss: 1.1495e-09 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0263 - val_sparse_categorical_accuracy: 0.9964
Epoch 3/10
286/286 - 20s - loss: 8.1507e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0279 - val_sparse_categorical_accuracy: 0.9964
Epoch 4/10
286/286 - 20s - loss: 5.6428e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0296 - val_sparse_categorical_accuracy: 0.9964
Epoch 5/10
286/286 - 20s - loss: 3.3439e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0300 - val_sparse_categorical_accuracy: 0.9958
Epoch 6/10
286/286 - 20s - loss: 3.9709e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0322 - val_sparse_categorical_accuracy: 0.9958
Epoch 7/10
286/286 - 20s - loss: 2.5079e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0333 - val_sparse_categorical_accuracy: 0.9958
Epoch 8/10
286/286 - 20s - loss: 1.4629e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0354 - val_sparse_categorical_accuracy: 0.9958
Epoch 9/10
286/286 - 23s - loss: 1.4629e-10 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0363 - val_sparse_categorical_accuracy: 0.9958
Epoch 10/10
286/286 - 20s - loss: 8.3597e-11 - sparse_categorical_accuracy: 1.0000 - val_loss: 0.0372 - val_sparse_categorical_accuracy: 0.9958
<keras.callbacks.History at 0x7efce2468bd0>
```

Attention Results:



(a) Attention Result 1



(b) Attention Result 2

Conclusion

- According to the training and test procedure, we observe that the neural attention network overfits to the data
- To avoid overfitting and generalizing the model we implement batch-normalization and dropout layers. The performance seems to improve slightly.

THANK YOU!

Bibliography

[1] D. C. de Andrade, S. Leo, M. L. D. S. Viana, and C. Bernkopf, "A neural attention model for speech command recognition," *arXiv preprint arXiv:1808.08929*, 2018.