BTP Project Update-3

High level Progress Overview

Last week , we completed the training of our end to end ASR deep learning model prior to transfer learning. The model is based on deep speech architecture of Baidu. We are also able to save the best fit model with **Word error rate** 0.17 and **Character error rate** 0.14 with epoch=5. Moreover this week, We built our own Speech Dataset collection API using Django framework and SQLite database.

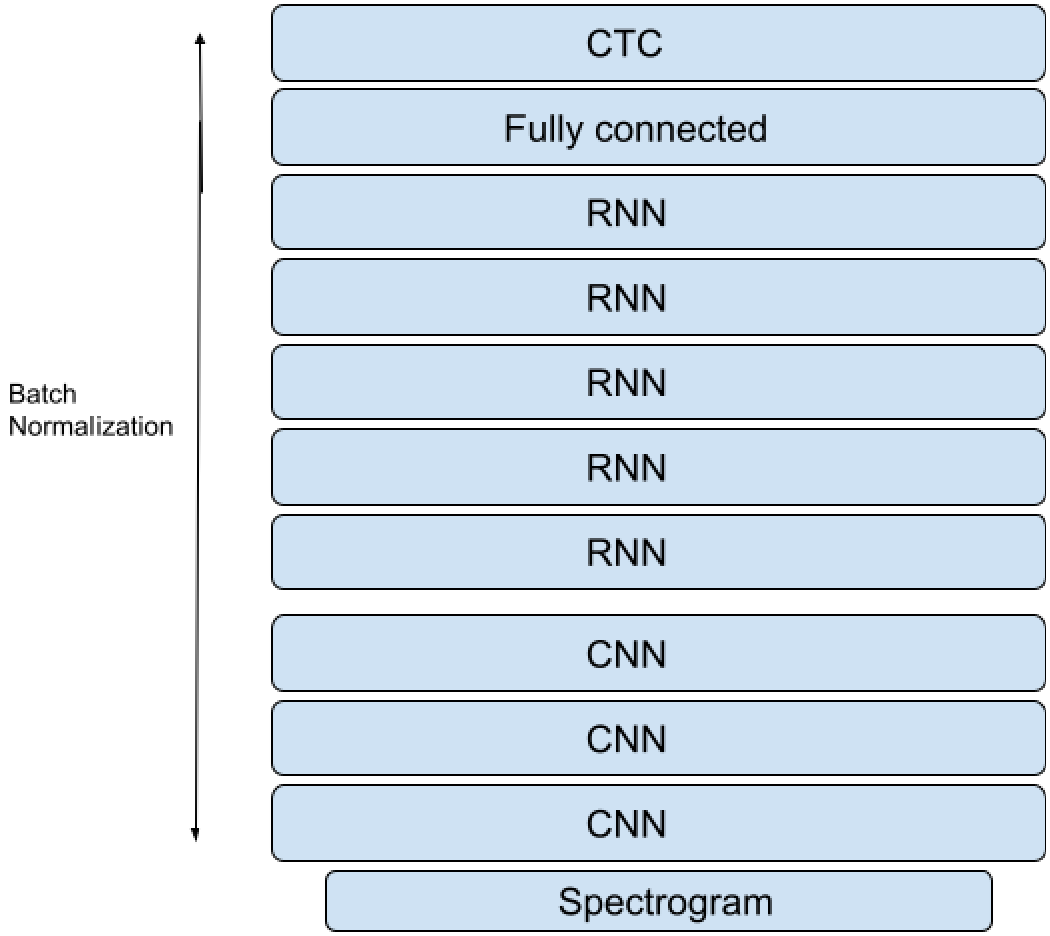
Low level Progress Overview

Steps performed till now:

- Data preprocessing
- Raw audio data augmentation
- Generation of Mel spectrograms
- MFCC
- Neural network architecture building
- Model Training
- Model Evaluation
- Saving Best fit Model
- Built Speech dataset Collection API using Django

Architecture of our model:

The architecture of our model will use Convolutional neural network(CNN) as well as Recurrent neural network(RNN) packed with CTC(Connectionist temporal classification) function that will calculate the WER(Word error rate) for our model. Below is the architecture of our model before transfer learning.



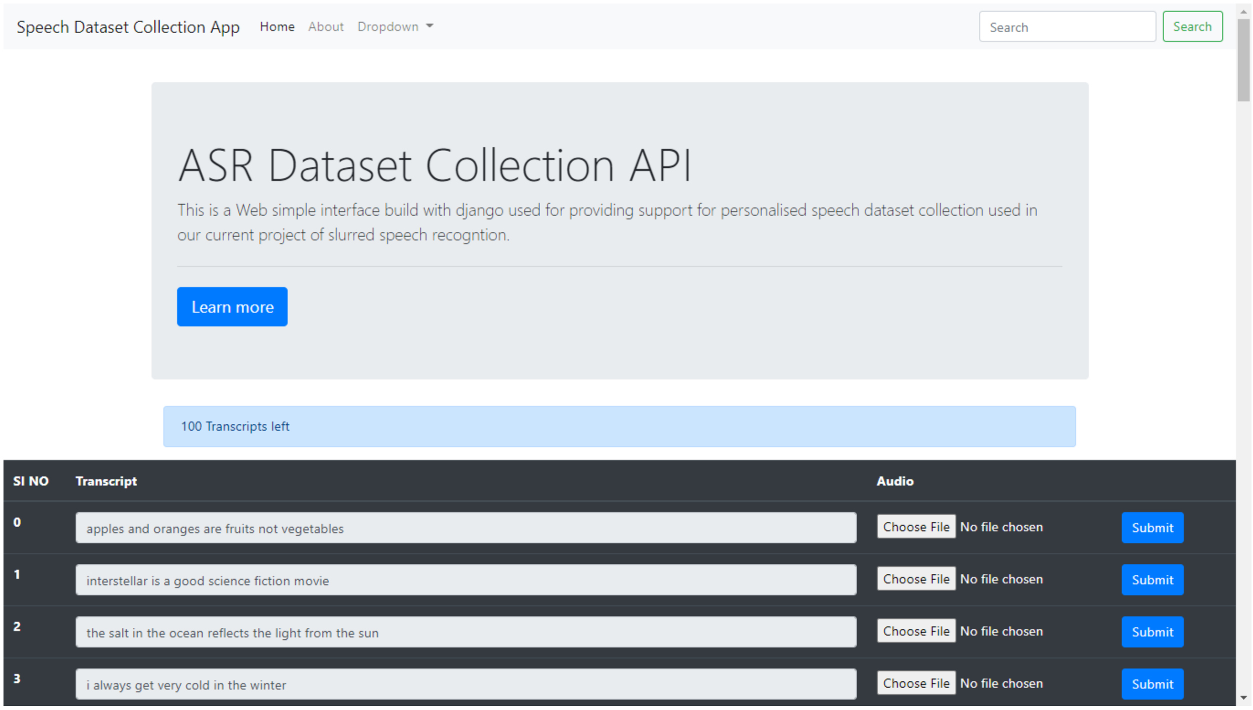
Model Performance:

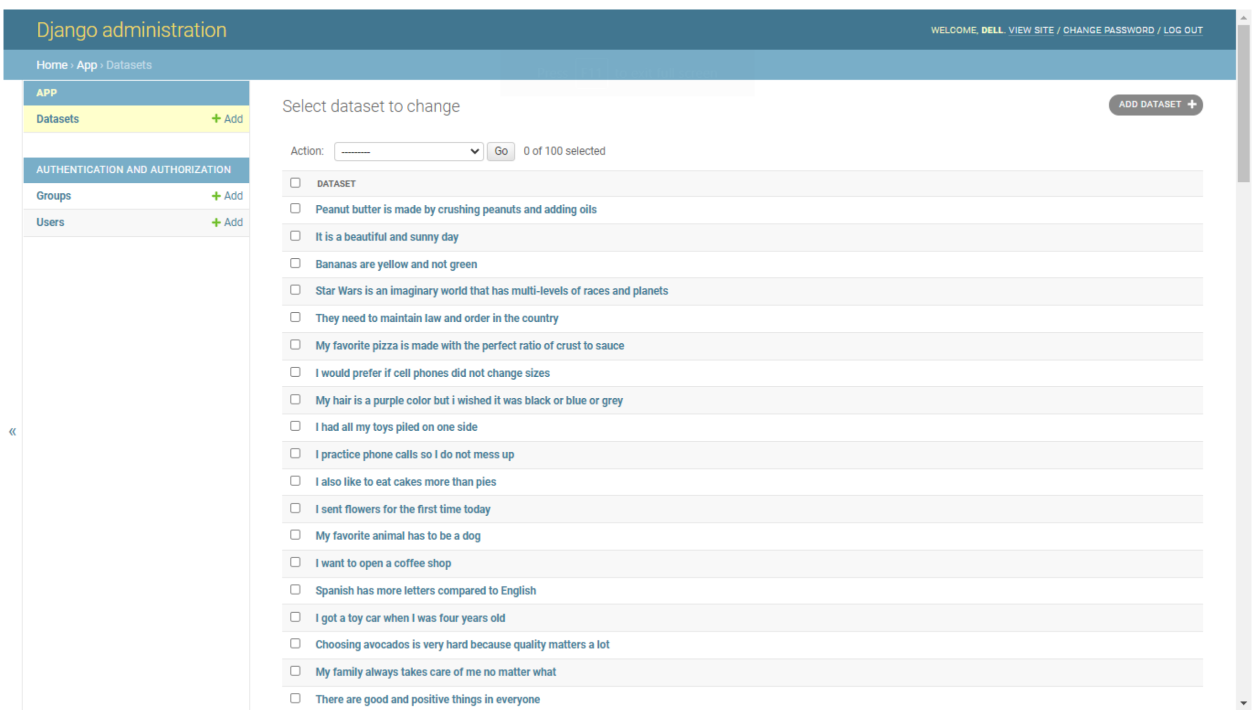
- Final Epoch Average Loss: 0.61
- Final Epoch Average CER: 0.15
- Final Epoch Average WER: 0.17

```
Train Epoch: 5 [13500/28539 (47%)]	Loss: 0.932549
Train Epoch: 5 [14000/28539 (49%)]	Loss: 0.759474
Train Epoch: 5 [14500/28539 (51%)]	Loss: 0.970782
Train Epoch: 5 [15000/28539 (53%)]	Loss: 0.787974
Train Epoch: 5 [15500/28539 (54%)]	Loss: 0.819515
Train Epoch: 5 [16000/28539 (56%)]	Loss: 0.674956
Train Epoch: 5 [16500/28539 (58%)]	Loss: 0.601339
Train Epoch: 5 [17000/28539 (60%)]	Loss: 0.689543
Train Epoch: 5 [17500/28539 (61%)]	Loss: 0.695834
Train Epoch: 5 [18000/28539 (63%)]	Loss: 0.835478
Train Epoch: 5 [18500/28539 (65%)]	Loss: 0.857994
Train Epoch: 5 [19000/28539 (67%)]	Loss: 0.751024
Train Epoch: 5 [19500/28539 (68%)]	Loss: 0.593558
Train Epoch: 5 [20000/28539 (70%)]	Loss: 0.834492
Train Epoch: 5 [20500/28539 (72%)]	Loss: 0.668797
Train Epoch: 5 [21000/28539 (74%)]	Loss: 0.874244
Train Epoch: 5 [21500/28539 (75%)]	Loss: 0.656230
Train Epoch: 5 [22000/28539 (77%)]	Loss: 0.792151
Train Epoch: 5 [22500/28539 (79%)]	Loss: 0.884348
Train Epoch: 5 [23000/28539 (81%)]	Loss: 0.602038
Train Epoch: 5 [23500/28539 (82%)]	Loss: 0.678341
Train Epoch: 5 [24000/28539 (84%)]	Loss: 0.802871
Train Epoch: 5 [24500/28539 (86%)]	Loss: 0.698470
Train Epoch: 5 [25000/28539 (88%)]	Loss: 0.770709
Train Epoch: 5 [25500/28539 (89%)]	Loss: 0.747371
Train Epoch: 5 [26000/28539 (91%)]	Loss: 0.660128
Train Epoch: 5 [26500/28539 (93%)]	Loss: 0.671808
Train Epoch: 5 [27000/28539 (95%)]	Loss: 0.813223
Train Epoch: 5 [27500/28539 (96%)]	Loss: 0.642177
Train Epoch: 5 [28000/28539 (98%)]	Loss: 0.626147
Train Epoch: 5 [28500/28539 (100%)]	Loss: 0.736653

evaluating...
Test set: Average loss: 0.6191, Average CER: 0.147697 Average WER: 0.1742
```

Snapshot of the Dataset collection Interface:





Link to web APP:

1. https://mmig.github.io/speech-to-flac/
2. http://speech-collection.herokuapp.com/index/

Next goals to perform:

1. Applying transfer learning paradigm
2. Building text to speech model

Future plan of action:

Since we are done with end to end ASR model training and built a solution for gathering impaired speech dataset , we can now focus on transfer learning. After that we will implement the next phase of our project i.e building text to speech model.

References:

1. https://arxiv.org/pdf/1512.02595v1.pdf
2. https://arxiv.org/pdf/1412.5567.pdf
3. https://www.biometricupdate.com/201906/google-building-impaired-speech-dataset-for-speech-recognition-inclusivity
4. https://ai.googleblog.com/2019/08/project-euphonias-personalized-speech.html