Deep Learning
Assignment - 2

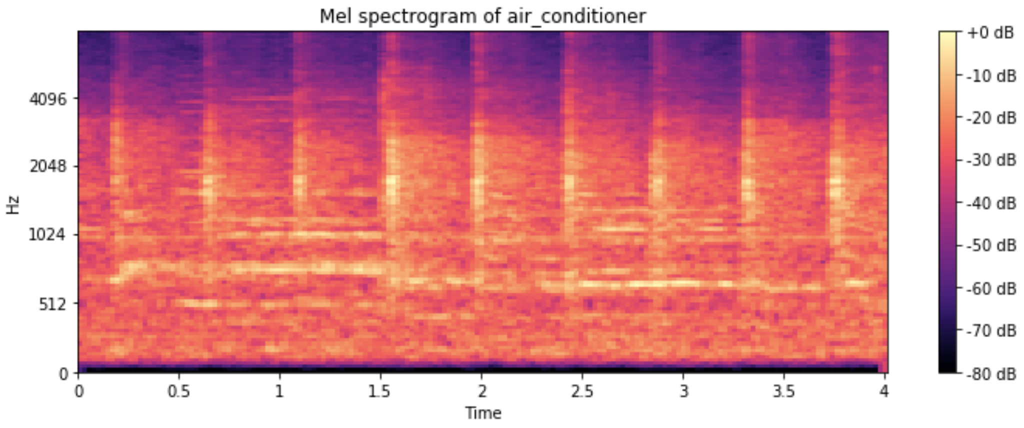
K. Srivatsava MT18054 Subhani Shaik MT18117 P. Akhil Kumar MT18130

Q: RNN for sound classification

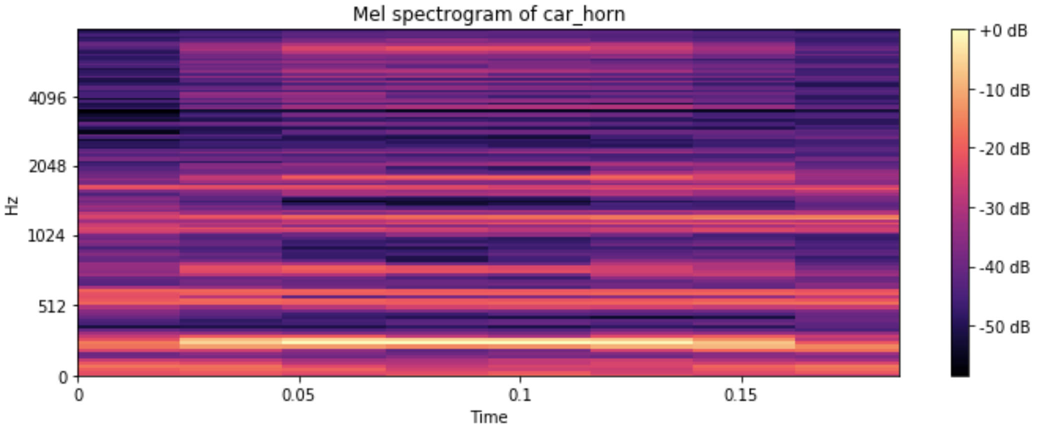
- The dataset contain 5435 audio files of variable lengths. The files in the dataset belong to 10 different classes. The spectrogram visualisations of audio files belonging to different classes is as shown below:
- The audio files are preprocessed using Librosa to extract Mel-Frequencey Cepstral Coefficients (MFCCs) features from each file. Depending on the length of the audio file, the feature length varied. To make it uniform across the dataset, the features are padded with zeros.
- 40 coefficients are extracted (n_mfcc) from the MFC representation of each audio file. The sample_rate is obtained using the librosa.load() function.
- Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs are are coefficients that collectively make up an MFC[1].

[1] Wikipedia - https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Sample spectrograms for 2 classes *viz.*, car_horn, air_conditioner. The remaining plots are submitted as image files.



Air conditioner



Car Horn

Model Architecture:

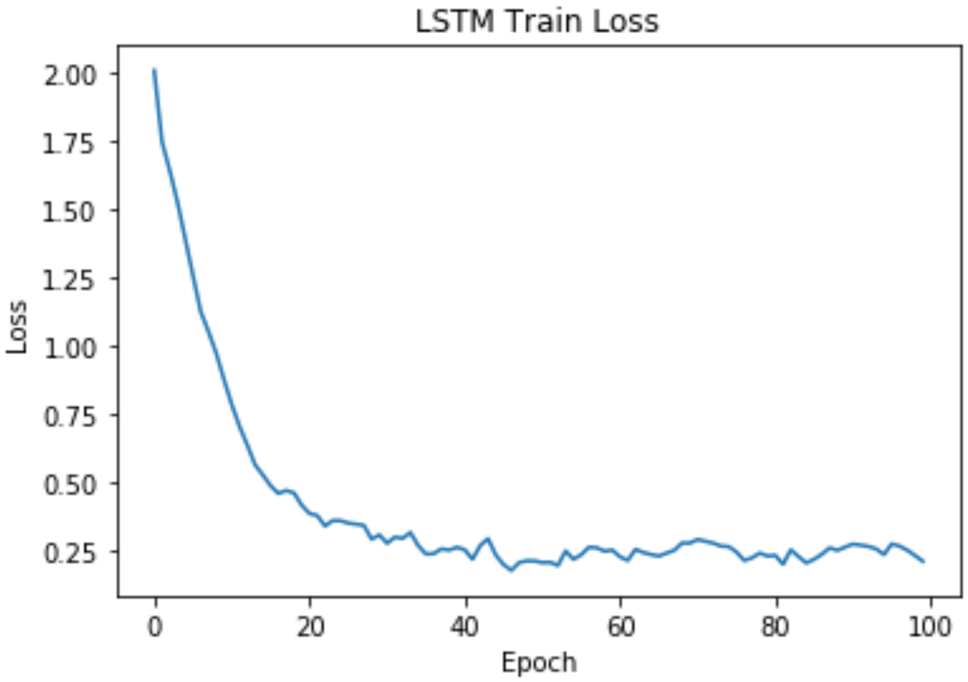
We used an LSTM for audio classification. The LSTM contains 3 layers. Each layer contains 173 neurons. The RNN is expanded to 40 time series representations.

To prevent overfitting, we are dropping out 20% of the neurons in the network during the training epochs.

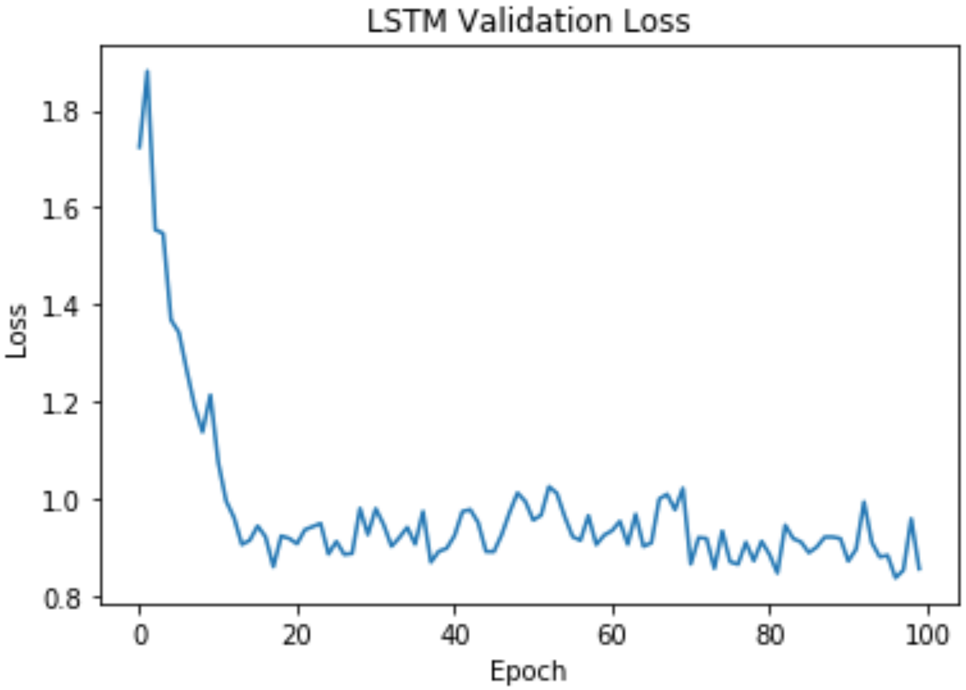
The motivation to use an LSTM over an RNN is that the RNN has a very less validation accuracy as compared to the training accuracy. This is due to the problem of vanishing gradients.

We preferred an LSTM so that long-term dependencies are preserved and takes less time for training. The training vs. validation accuracies, loss values are better. The LSTM outperformed the single-layer RNN. The 3 layer LSTM performed better than 1, 2 layered LSTMs as the network is deep enough to learn the features better.

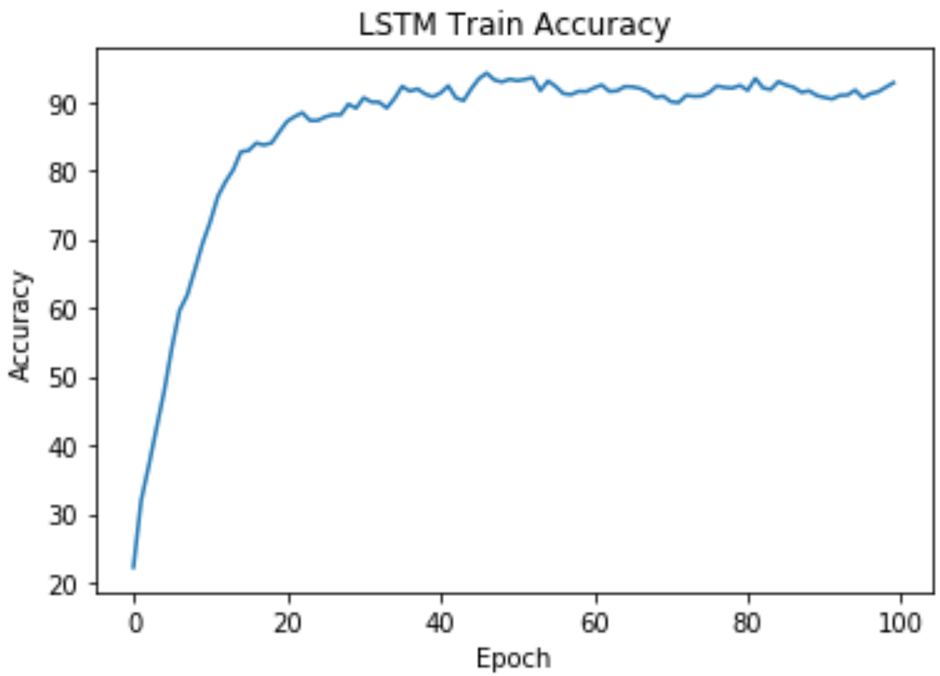
The block representation of our architecture can be found below:



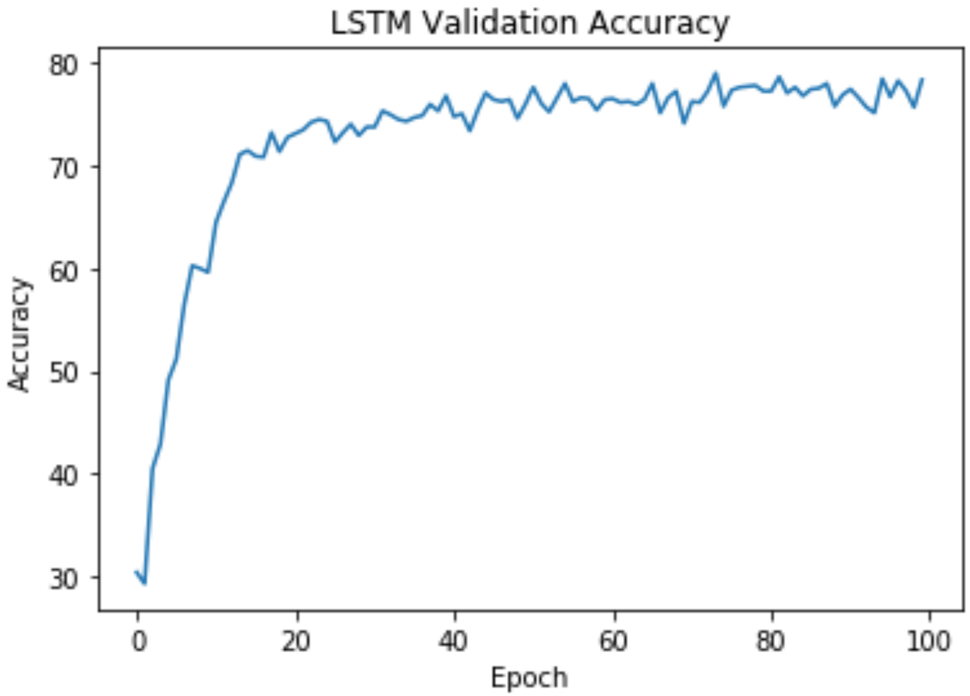
Train Loss



Validation Loss

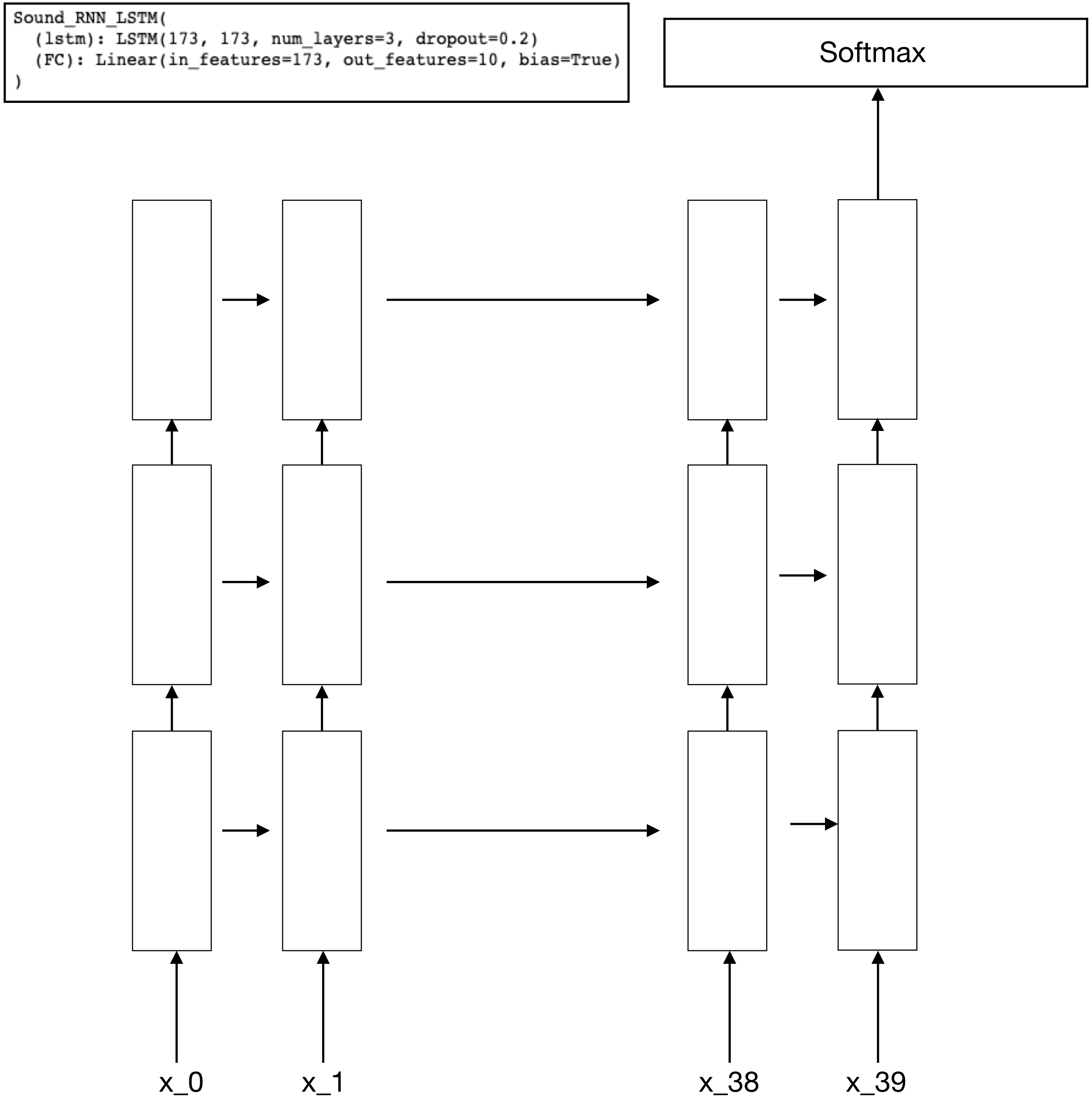


Train Accuracy



Validation Accuracy

100 Epochs	**Train**	**Validation**
Loss	0.2	0.85
Accuracy	92.5%	78.4%



```
Sound_RNN_LSTM(
  (lstm): LSTM(173, 173, num_layers=3, dropout=0.2)
  (FC): Linear(in_features=173, out_features=10, bias=True)
)
```