A General Purpose Audio Tagging System

Ankur Sharma (2016225)
Ishaan Bassi (2016238)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Problem Statement

A general purpose audio tagging system that classifies a wide range of sounds (ranging from car horns to strumming of guitar) that we hear on a daily basis.

Audio data analysis and classification is a huge research domain. Currently, most of the work focuses on specific areas such as speech recognition and music tagging. This project hence aims at developing a solution that can classify sounds which are not necessarily similar to each other. The motivation behind this project is the wide range of applications of audio classification in audio searching in online databases, entertainment industry, surveillance.

Dataset

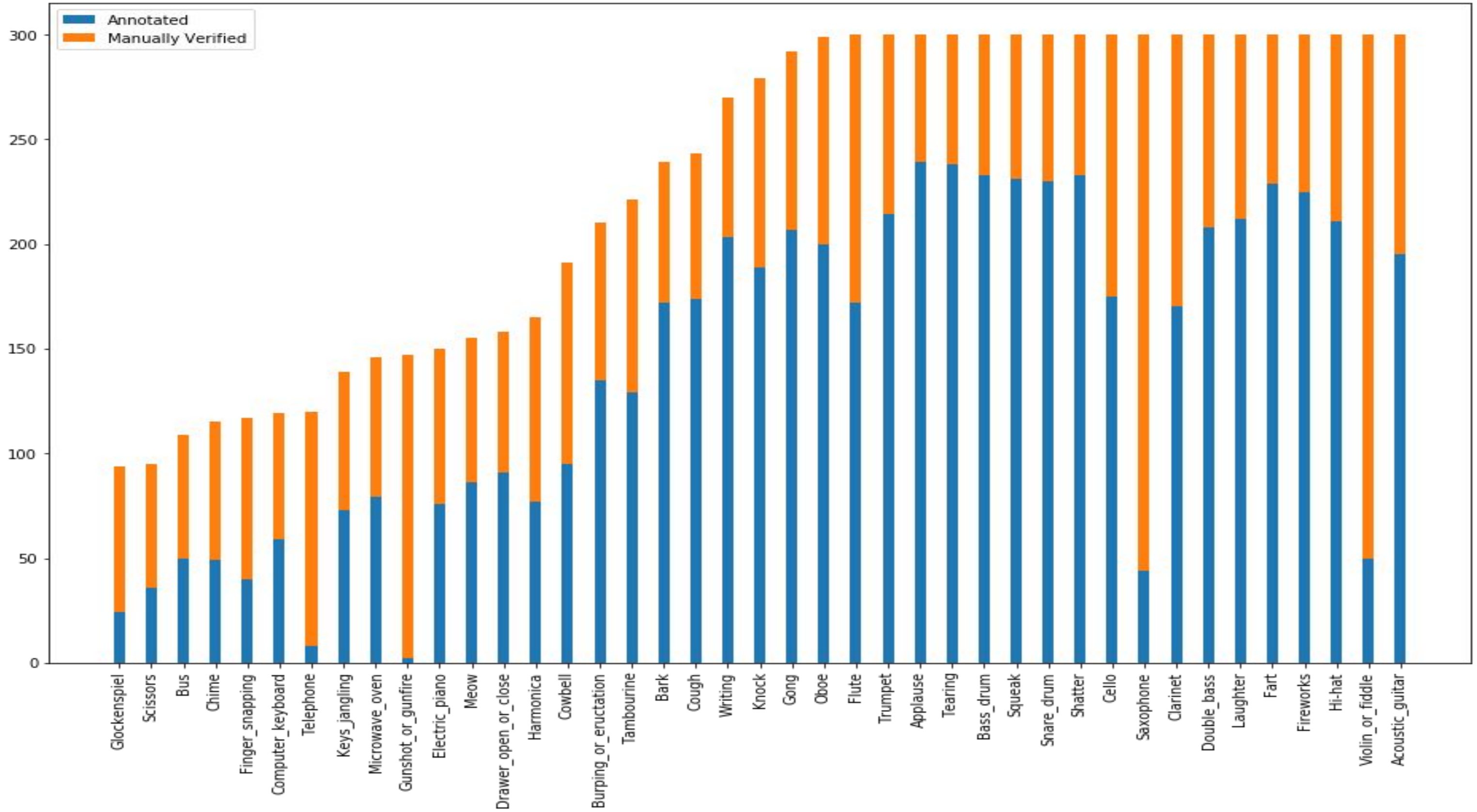
Training Samples: 9.5K

Number of Classes: 41

Test Set Size: 1.6K

- The data is imbalanced with about **94 - 300 samples per category** in training set and 25 - 110 per category in test set.
- In training set about **3.7K audio have manually verified annotations** and **5.8K have non-verified annotations**. About **30-35%** of non verified samples might **not be labelled correctly**.
- The length of the samples varies from about 300 ms to 30000 ms with **sampling rate of 44.1K**.

Dataset- Audio Samples per Category



Approach

- Baseline: Support Vector Machine
- **Deep Learning**:
  - Vanilla 1D Convolutional Neural Network (CNN)
  - 2D CNN on Mel-frequency cepstral coefficients (MFCC)
- **Transfer Learning**: Due to non-uniform data among different classes, we train the network for classes with 300 samples(18 classes), **Set A**, and transfer learnt weights to the network for remaining 23 classes, **Set B**.
  - Baseline to compare results of TL: 2D CNN on MFCC for Set B
  - Task A: 2D CNN on MFCC for Set A
  - Task B: 2D CNN on MFCC for Set B

➢ Vanilla Model: https://i.imgur.com/DFgeW3x.png

➢ MFCC Model: https://i.imgur.com/XLxI0rN.png

Results

Approach	*Model*	*Top 3 Accuracy*	*Top 1 Accuracy*	*Train Accuracy*
Baseline	SVM	-	13%	20%
Deep Learning	Vanilla 1D CNN	80%	61%	73%
	2D CNN on MFCC	80%	61%	91%
Transfer Learning	2D CNN on MFCC for Set B	84%	63%	82%
	2D CNN on MFCC for Set A- Task A	**92%**	**74%**	92%
	2D CNN on MFCC for Set B- Task B	**86%**	**65%**	82%

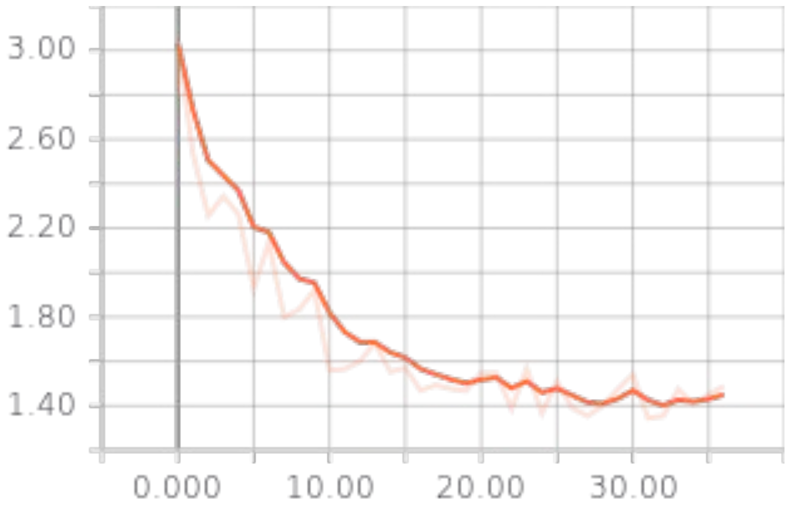
Observations

- As the main sound can be present anywhere across the audio, SVM finds audios of the same label to be different. In CNN, parameter sharing takes care of that very well.
- 2D CNN on MFCC converges faster than Vanilla 1D CNN w/o affecting the performance.
- 2D CNN training accuracy is much higher than Vanilla 1D CNN training accuracy which implies that the former one might give better result on more test data.
- Transfer Learning outperforms all the models and gives us better accuracy with faster convergence.

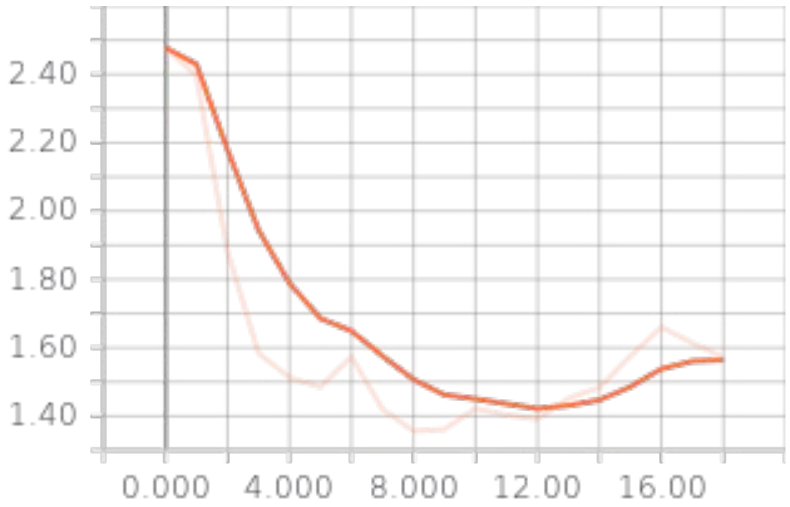
Analysis

- Has the model been correctly trained?
  - Yes, the model does not overfit because of Early Stopping checkpoint.
  - Reason: Validation Loss Monitoring while Training.
- Analysis of errors
  - Training samples contain 35% noise for non-verified samples leading to poor overall performance.
  - Loss Function: Cross-Entropy
  - Metric: Accuracy
- Ablative analysis
  - Reduction of sampling rate reduces model complexity w/o affecting the performance
  - MFCC reduces the input size and converges faster

Validation Plots- Deep Learning

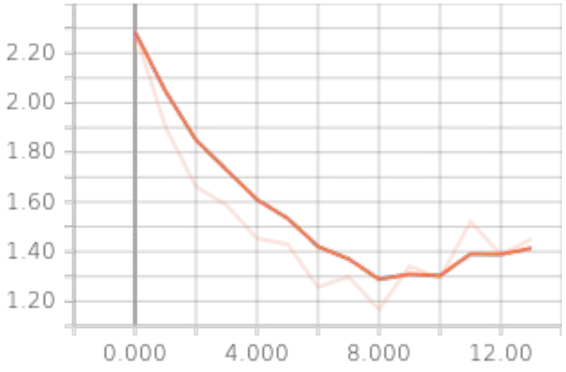


Vanilla 1D CNN

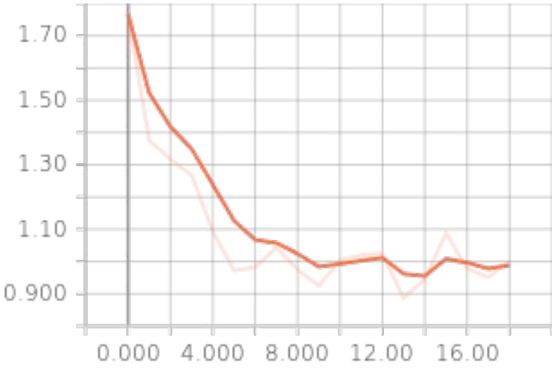


2D CNN on MFCC

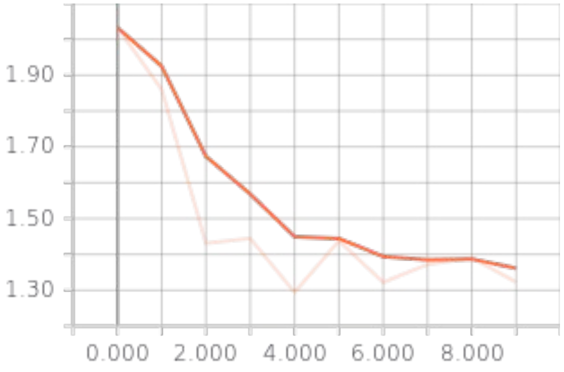
Validation Plots- Transfer Learning



2D CNN on MFCC for Set B
Baseline



2D CNN on MFCC for Set A
Task A



2D CNN on MFCC for Set B
Task B

Individual Contribution

- Ankur Sharma
    - Deep Learning
    - Transfer Learning
    - Metrics Evaluation

- Ishaan Bassi
    - Data Preprocessing
    - MFCC Feature Engineering
    - Baseline: SVM
    - Metrics Evaluation

Thank you