EE603A Assignment 1

-Sahil Maurya (190738)

Introduction

Audio classification has been a problem of interest for a lot of time. It has been used in speech recognition, music genre classification and urban sounds classification. Classifying urban sounds can be helpful in many places, such as being able to detect some fault at a construction site. This paper studies the results of 4 Deep Neural Networks on an urban sounds dataset. It describes the methods which can be used for preparing such datasets, followed by results of different models on the dataset. It ends with a discussion about the design choices of the model.

Literature Review

Machine Learning has been used for the classification of a variety of audio sounds such as Music Genre Classification[1][2] and urban sound classification[3][4]. The input given to a neural network classifier can be either given in raw time-domain 1-D representation or time-frequency representations such as the Mel spectrogram. Various models have been developed that take raw time domain audio signals such as EnvNet [5] and Sample-CNN [6]. However most SOTA models use CNN on spectrograms[7][8].

Methods

Dataset Description

The Urban Noises dataset consisted of 1000 train samples and 201 test samples. The audio signals had been converted to a Mel spectrogram with Hann windowing function and Nfft = 2048. For training purposes, the dataset was divided between train data and validation data with a ratio of 4.

Creating Samples

Since the dataset had mel spectrograms of different shapes due to difference in audio signal length, there was a need to make a uniform data sample.

Firstly, a sample length of 300 was selected empirically. The train samples which were smaller than the sample length were duplicate padded. Then a sliding window was used to create multiple samples out of the train samples which were longer than chosen sample length. By this process, the amount of training samples also increased by roughly 3 times. The same technique was also applied on test samples. The final prediction for a test data was the average of predictions of different samples taken from the same test data.

Classification Models

Four different classification models were created to test their performances on the dataset.

a. Dense Neural network without dropout layers: This was the simplest classification model. It comprised 5 fully connected layers with no. of neurons, 128, 128, 64, 32, 10 respectively. The activation layers used was ReLU except the last layer which was softmax. It gave a validation accuracy of 60%
b. Dense Neural Network with pooling layers: It comprised 5 fully connected layers. With no. of neurons, 128, 128, 64, 32, 10 and had 3 dropout layers in between them with a dropout rate of 0.2. It gave a validation accuracy of 65%
c. CNN1: It comprised 5 CNN layers followed by 3 dense layers with dropout. The kernel sizes for the CNN layers were, (6, 2), (6, 2), (5, 1), (5, 1), (4, 1). The CNN layers helped bring out the spatial features from the mel spectrogram. It gave validation accuracy of 80%
d. CNN2: It comprised 3 convolution-pooling layers followed by 3 dense layers. The CNN layers were used to bring out spatial features from the mel spectrogram while the pooling layer was used to reduce dimensions. Not only it gave better results than CNN1 but was also more efficient due to pooling layers. It gave a validation accuracy of 94%. This was the best performing model.

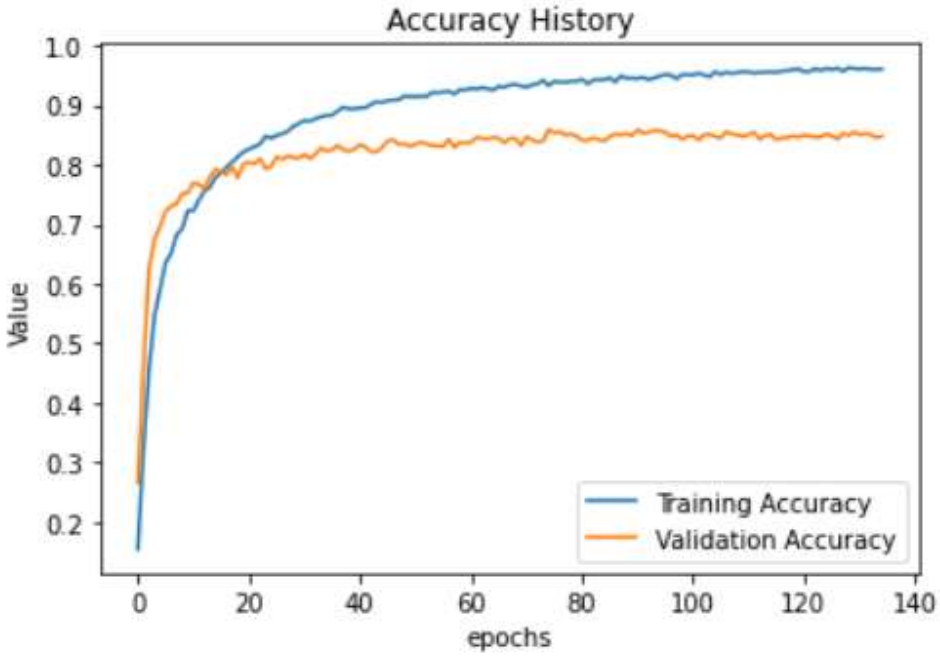
Results for CNN2

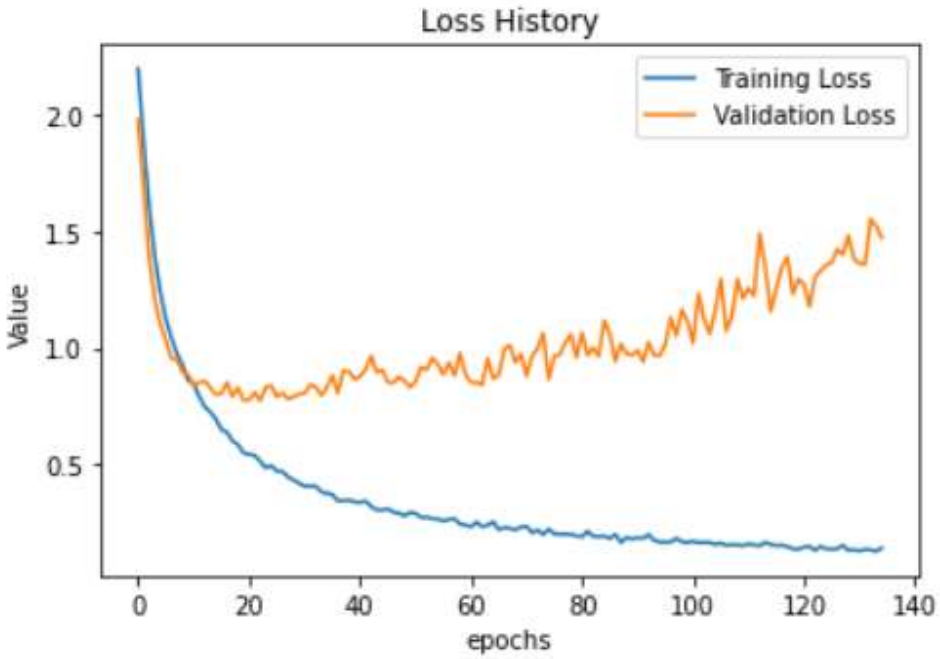
The test accuracy came out to be 70%. The f1 score was 0.70.

Confusion matrix:
Here, True labels are along the horizontal axis, while Predicted labels are along the vertical axis.
The Classes are:
"Bark": 0,
"Meow": 1,
"Siren": 2,
"Shatter": 3,
"Knock": 4,
"Crying_and_sobbing": 5,
"Microwave_oven": 6,
"Vehicle_horn_and_car_horn_and_honking": 7,
"Doorbell": 8,
"Walk_and_footsteps": 9

```
[[18,  2,  1,  0,  1,  1,  1,  0,  0,  0],
 [ 0, 13,  1,  0,  0,  1,  1,  1,  1,  0],
 [ 2,  3, 14,  0,  0,  0,  1,  2,  5,  0],
 [ 0,  0,  0, 13,  1,  0,  1,  0,  1,  2],
 [ 1,  0,  0,  0, 13,  0,  1,  0,  0,  3],
 [ 1,  1,  0,  0,  0, 12,  2,  0,  2,  0],
 [ 0,  1,  0,  0,  1,  0, 16,  0,  0,  0],
 [ 1,  1,  1,  0,  1,  0,  0, 18,  1,  1],
 [ 0,  0,  1,  1,  0,  0,  0,  0, 16,  0],
 [ 1,  2,  0,  1,  3,  0,  2,  0,  1,  8]]
```

Train and validation accuracy during training:





Observation and discussions:

The four models tested were in increasing order of complexity. The first model only had Dense Neural Networks, the second model had dropout layers along with Dense layers. The dropout layers helped the model to generalize better. The CNN1 network was the biggest model with the largest number of parameters, however the CNN2 model, despite having less number of training

parameters, was better able to generalize the dataset. The pooling layers in the CNN2 model brought down the number of trainable parameters and made it more efficient.

References

[1] M. Dong, “Convolutional neural network achieves human-level accuracy in music genre classification,” 2018.
[2] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” 2016
[3] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “Esresnet: Environmental sound classification based on visual domain models,” 2020
[4] Y. Aytar, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” 2016.
[5] Y. Tokozume and T. Harada, “Learning environmental sounds with end-to-end convolutional neural network,” in 2017 IEEE International Conference on Acoustics, Sp
[6] J. Lee, J. Park, K. L. Kim, and J. Nam, “Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms,” arXiv preprint arXiv:1703.01789, 2017.
[7] Mushtaq, Z., Su, S. and Tran, Q., 2021. Spectral images based environmental sound classification using CNN with meaningful data augmentation. Applied Acoustics, 172, p.107581.
[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1-6, doi: 10.1109/MLSP.2015.7324337.