CNNs-based Indoor sound classification using STFT & CQT spectrogram

Vishwanath Marimuthu
Department of Computer Science
Stanford University
vmarimut@stanford.edu

Abstract

Spectrograms extracted from sound have been useful in CNN based architectures to classify sound. Spectrograms such as short-time Fourier transform(STFT) and constant Q transform(CQT) represent a good representation of the temporal and spectral structure of the original audio. In this paper, STFT and CQT features extracted from an audio file were assessed for the classification of various sound in the dataset using CNNs. The experiment shows that 89% train accuracy,87% validation accuracy, and 78% test accuracy were achieved for the FSDKaggle2018 dataset.

1 Introduction

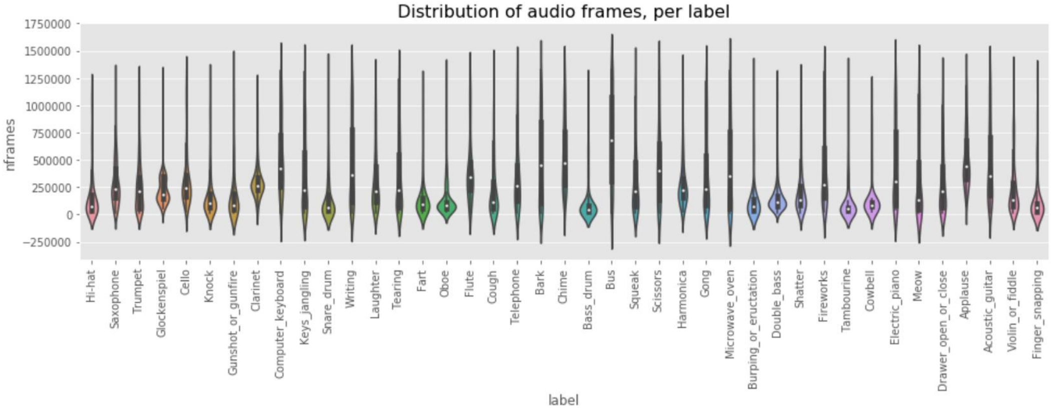
Automatic environmental sound classification is a growing area of research with numerous real-world applications such as context-aware computing, surveillance, event classification. There are a variety of signal processing and machine learning techniques applied to this problem currently but with the recent advancement of the convolutional neural network for classification of images, it begs the question of the applicability of these techniques in other domains, such as sound classification. The challenging aspect of the project is to find the apt visual representation that can be used to differentiate various sound in the data set. Unlike existing solutions that use Mel spectrogram for classification and speech processing, this project explores CQT which has a good resolution for low and mid to low frequencies for classification than Mel. The reason is that CQT requires more sample points from low frequencies. Based on the research from ref [1] and initial observation from the experiments performed, the project proposes to use CQT spectrograms along with STFT spectrograms for classification for the FSDKaggle2018 data set. The final solutions extract STFT, CQT spectrograms from the raw audio files to feed into two CNN models and the features extracted from the CNNs are concatenated and given to fully connected layers for sound classification. The output of the model for given audio is one of the 41 class codes in the dataset. The project also uses data augmentation technique such as speed shift and pitch shift for audios and two different learning rates, batch size to improve the accuracy of the model

2 Related work

CNN was first evaluated in ref [3] for environmental sound classification which achieved 5.6% more accuracy than the traditional method. In this work, they have used 2-layer CNN with max-pooling and 2 fully connected layers for classification and log-Mel spectrograms of the audio file as the input to classification. Ever since this method researchers have explored various spectrograms such as gammatone[2], Mel [3], STFT[4] for classifying datasets such as ESC-50, Urban Sound8k. MFCC and Mel features are the commonly used features in the field of speech processing and Acoustic sound event classification however they are not perception aware spectrograms and has less resolution for low frequency and low to mid-frequency. On the other hand, gammatone spectrogram represents how human ear filter sound but they were yielding the same results as of Mel spectrogram in the initial experiments performed. Based on the experiments in the research ref [1] combining two different spectrograms and feeding to VGGNet/ResNet compared to using CONV1D for audio waves or CONV2D for spectrogram images with one feature was giving promising results for environment sound dataset and urban sound dataset. Hence in this paper features, spectrograms like STFT for the representation of frequency-time variation and CQT for high-frequency resolution in the low and low-mid frequency would be used for classifying sounds in FSDkaggle2018 data set.

3 Dataset and Features:

The dataset[1] contains 11,073 audio files annotated with 41 labels of the AudioSet ontology. All audio samples in the dataset are gathered from FreeSound and provided as uncompressed PCM16 bit,44.1 kHz, mono audio files. The train set includes ~9.5k samples unequally distributed among 41 categories. The minimum number of audio samples per category in the train set is 94 and maximum 300, similarly, the duration varies from 300ms to 30s due to the diversity of the sound categories and preferences of FreeSound users. The proposed method uses a 90/10 split for the training/validation. The test set has around ~1.6k samples with manually verified annotations. Fig [1] represents frame distribution vs classes in the data set.



4 Methods

4.1 Data preprocessing

1. The audio is loaded with a down sampled rate of 22.1 kHz
2. Since most of the audio is of short length, the input audio is clipped to 5secs and for the audios that are less than 5sec, data length is increased with random padding
3. As shown in Fig2, the STFT was generated with the default setting in librosa and CQT was generated with 120 bins and 24 bins per octave for higher frequency resolution
4. The result of CQT is 216*120 spectrogram and result of STFT is 216*1025 spectrogram and these are normalized before passing in to convolution

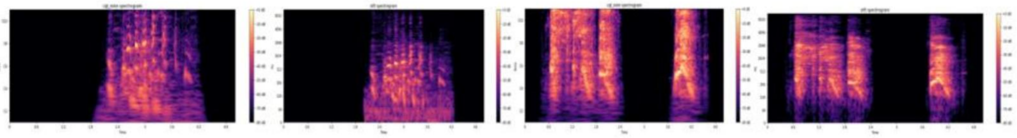


Fig 2: CQT and STFT spectrogram for laughter, CQT and STFT spectrogram for cough

[1] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra. "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline". Proceedings of the DCASE 2018 Workshop (2018)

4.2 **Neural network model**

The two spectrograms generated from the Audio file is then given to two identical stacks of convolutional neural network layers along with drop out of 10% and the output feature set generated by them is given to fully connected networks to get output class code of the audio file. All the modules in the model uses Relu as the activation function except the last layer that uses softmax for classification. Fig3 is a detailed description of the network model used for classification. We have totally

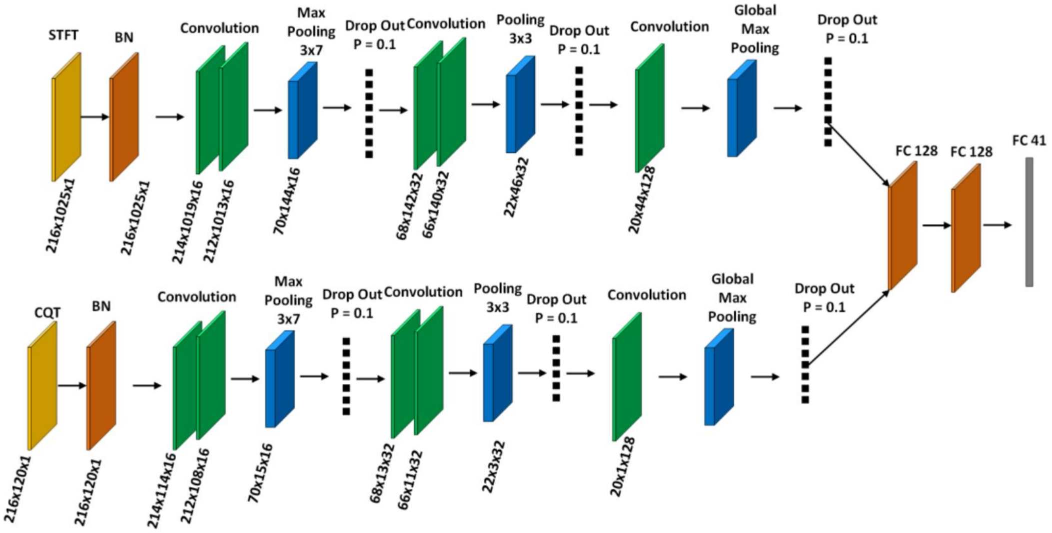


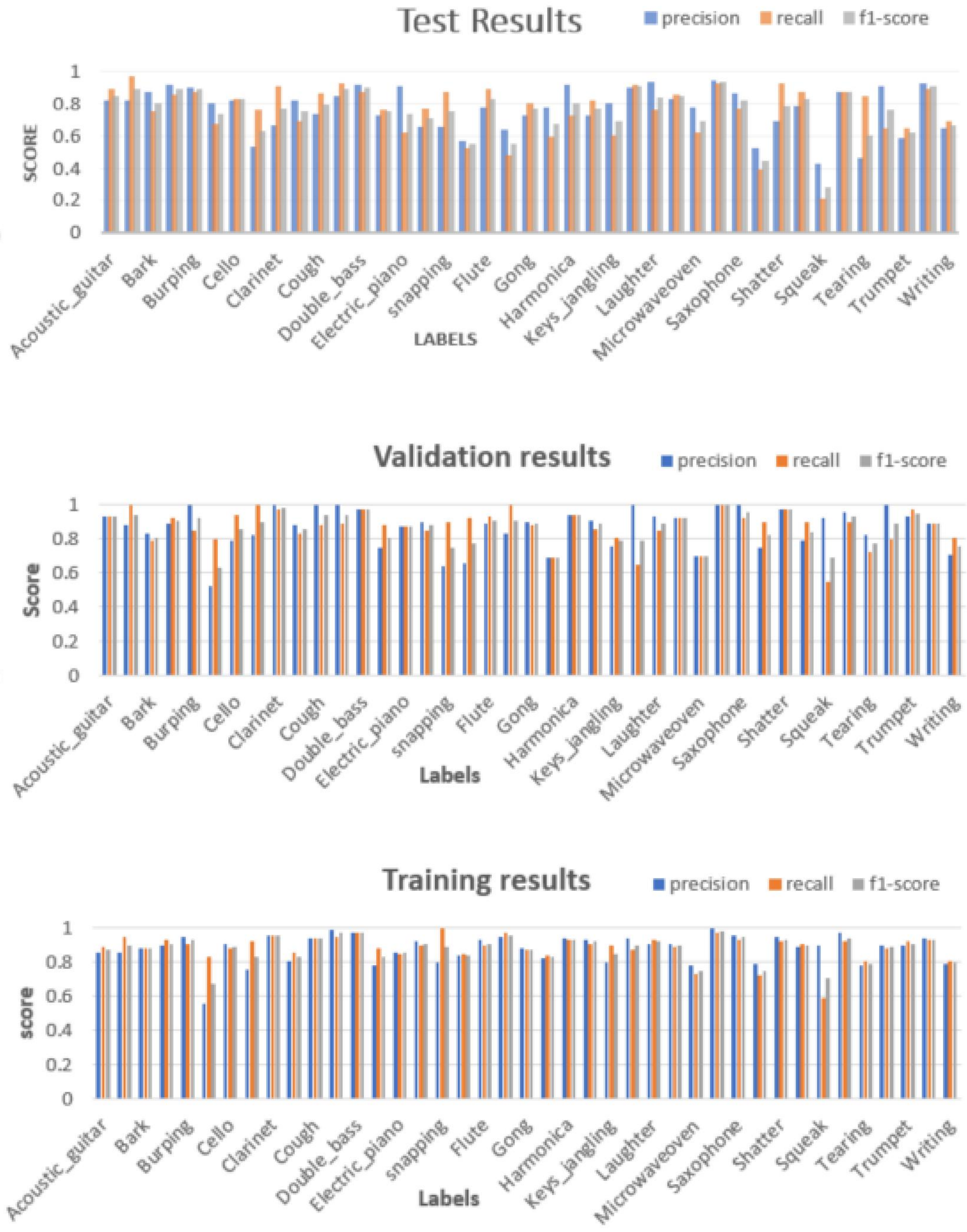
Fig3 CNN based model used for classification

4.3 **Data augmentation**

Since the number of audio samples per category in the train set varies between 94 and 300, the initial training and validation accuracy were very poor for labels such as Bus, Chime, Drawer_open_or_close, keys_jangling, etc. To improve the accuracy data augmentation technique like randomized speed shift i.e. changing the speed of the audio by 0.9 to 1.5 factor which is randomly picked and random pitch shift i.e. changing the pitch of the audio has been used. This, in turn, leads to a good increase in training and validation accuracy for selected labels. For e.g., the accuracy of Scissors went from 13 % to 80% in training accuracy.

5 Results

All the experiments were performed in p3.Xlarge node of amazon web services and the models were implemented using Keras. This project uses sparse categorical cross-entropy for calculation loss as the target class code is integer mapped and Adam optimization algorithm with 0.9 for Beta1 and 0.99 for Beta2 as recommended in the academia. The project uses two different learning rates of 0.001,0.0001 and batch sizes of 32,64 for improving the accuracy. The initial epoch of 1-60 epochs uses 0.001 and batch size of 32 as the parameter for training, that in turn resulted in around 70% training accuracy and was showing a very slow reduction in the training/validation loss. To improve the accuracy, the learning rate of 0.0001 and batch size of 64 was used. This lead to the improvement of accuracy by 8%. The training/validation split is around 90/10 as we have enough data to train. The kernel size of (3,7) and (3,3) are used in the convolution filter layer and there were about 168,461 parameters to train. Each epoch took around 30 mins for completion and the model has trained ~100 epochs. Fig4, Fig5, Fig6 shows the precision, recall, F1 score for test, validation, and training respectively. Fig7 shows the loss trend, Table1 shows the accuracy metrics More details can be found at ref [5].



From top to bottom Fig4, Fig6, Fig7

	Accuracy	Support
Training	0.89	8525
Validation	0.87	948
Test	0.78	1600

Table1 Accuracy metrics

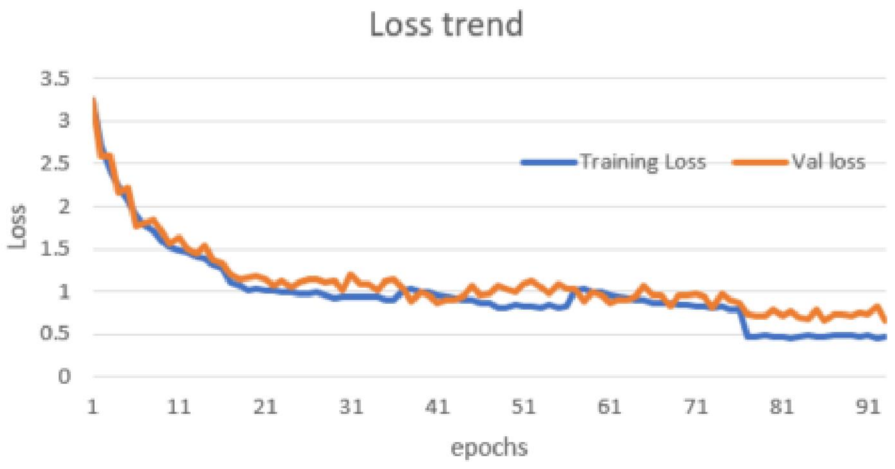


Fig7 1-60 epochs uses 32 and 0.001,60-100 epoch uses 64 and 0.0001 as batch size and learning rate

6 Conclusion

The main objective of this project to explore the STFT and CQT spectrograms for classification of sounds. Even though we achieve good results for majority of the sounds, further insights are required for sounds such as Fireworks, squeak, tearing, writing. This may be due to the fact STFT and CQT may not represent all the properties of the audio wave. Few observations made while increasing the training accuracy.

- The test accuracy for the augmented data is still less
- Other alternative spectrograms such as gammatone, log MEL gave similar performance during initial epochs and didn’t not improve after 75% training accuracy.
- Increasing the batch size along with reducing learning rate helps in achieving better accuracy.
- Shuffling the data set before train/validation helps in faster reduction of the loss.This may be due to the improving converges of the loss function.

Future works involves the following

- The Chroma, spectral contrast, Tonnetz features must be explored for CNN to extract more useful features.
- Research also indicates dilated convolution to work well for environmental sound classification.
- Majority of the time were spent on improving training and validation accuracy, must study the test data set for improving the efficiency of the classifier
- Training each epoch took around 30 mins for processing, must preprocess the spectrograms in to a pickle file to improve the speed.

References

[1] CNNs-based Acoustic Scene Classification using Multi-Spectrogram Fusion and Label Expansions, arXiv:1809.01543 [cs.CV]

[2] Agrawal, Dharmesh & Sailor, Hardik & Soni, Meet & Patil, Hemant. (2017). Novel TEO-based Gammatone features for environmental sound classification. 1809-1813. 10.23919/EUSIPCO.2017.8081521.

[3] Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.

[4] Li, J.; Dai, W.; Metze, F.; Qu, S.; Das, S. A comparison of deep learning methods for environmental sound detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 126–130

[5] https://github.com/Vishwa07/IndoorSoundClassificaition.git

[6] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio set: An ontology and human-labeled dartaset for audio events." Proceedings of the Acoustics, Speech and Signal Pro- cessing International Conference, 2017.

[7] Zhang, Z.; Xu, S.; Cao, S.; Zhang, S. Deep Convolutional Neural Network with Mixup for Classification. arXiv 2018, arXiv:1808.083