CNNs-based Indoor sound classification using STFT & CQT spectrogram

Vishwanath Marimuthu

Predicting

- Visual representation of sound such as Mel spectrograms has shown promising result in the field of speech processing and sound classification
- But it is difficult to classify based on one representation of sound ,research shows that fusing multiple spectrograms result in high accuracy of sound classification
- Two different spectrogram STFT,CQT is passed into two identical stack of CNN with Pooling and drop out layer
- The resultant feature set from the CNN stack is concatenated and given to FC layers for classification
- The input to the model is STFT spectrogram of dimension 216X1025 and CQT spectrogram of dimension 216X120 extracted from the audio file and the output is one of the 41 class codes for the audio file

Name : Vishwanath Marimuthu
Email : vmarimut@stanford.edu
YouTube: https://www.youtube.com/watch?v=JGMRjjS-Tbk
Github: https://github.com/Vishwa07/IndoorSoundClassificaition.git

Models

- Below Network architecture(impl Keras) was used for sound classification
- Loss function: sparse categorical cross-entropy
- Learning rate:0.001(1-60 epochs),0.0001(60-100 epochs)
- Batch Size: 32(1-60 epochs),64(61-100 epochs)
- Total epochs:100 epochs,30min/epoch

Features

- STFT spectrograms for the representation of frequency-time variation and CQT spectrograms for high-frequency resolution in the low and low-mid frequency were extracted for CNN(as shown in the Fig)
- Input audio was clipped and random padded to get uniform input of 5 secs and resampled at 22.1 kHz

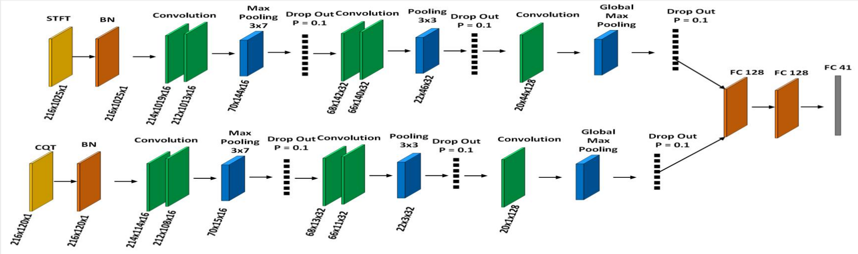
Data

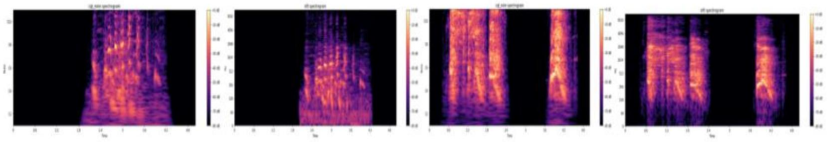
- The FSDKaggle2018 dataset contains 11,073 audio files annotated with 41 labels of the AudioSet ontology. All audio samples in the dataset are gathered from FreeSound and provided as uncompressed PCM16 bit,44.1 kHz, mono audio files.
- ~11k audio .wav files split into 8k train files with duration varies from 300ms to 30 secs

Results

- Training and validation accuracy of 86% and test accuracy of 78%
- Training/Validation split of 90/10
- 168,8461 trainable parameters

	Training	Validation	Test
Accuracy	0.89	0.87	0.78
Support	8525	948	1600





Discussion

- 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.
- Data augmentation techniques like pitch shift, time shift are used to due to lack of enough sample of certain class codes improved accuracy
- Sounds like Chime, Squeak, Scissors still give very low test accuracy from 20% to 60% compared to train /validation accuracy. This may be due to data augmentation with training set.
- More time spent on improving training and validation accuracy

Future

- 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.
- RESNet for STFT and CQT spectrograms

REFERENCES

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