CNNs-based Indoor sound classification using STFT & CQT spectrogram
Vishwanath Marimuthu

Predicting
- Visual representation of sound such as Mel spectrograms has shown promising results in the field of speech processing and sound classification.
- But it is difficult to classify based on one representation of sound. Research shows that using multiple spectrograms results in higher accuracy of sound classification.

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 seconds and resampled at 22.1 kHz.

Models
- Below Network architecture (impl Keran) 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, 30 min/epoch.

Data
- The FBDbkaggia2018 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 30 ms to 30 sec.

Results
- Training and validation accuracy of 86% and test accuracy of 78%.
- Training/Validation split of 90/10.
- 168,846 trainable parameters.

Discussion
- Other alternative spectrograms such as gammatone, log MEL gave similar performance during initial epochs and didn’t 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/validate helps in faster reduction of the loss. This may be due to the improving convergences of the loss function.
- Data augmentation techniques like pitch shift, time shift are used to due 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, Tonnez 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