# VGGish-based architecture for sound classification

Mansoor Rahimat Khan, Alexander Lerch, Hongzhao Guwalgiya, Siddharth Kumar Gururani, Ashis Pati

Georgia Tech Center for Music Technology

For this challenge, we began our classification task on the dataset by approaching it with simple feature computation followed by machine learning algorithms. We proceeded to use transfer learning using pre-trained models and finally trained a deep CNN for the task.

Approach 1:

We computed our feature vector from the audio samples using Mel Frequency Cepstral Coefficients initially by aggregating them by taking the mean and standard deviation of 20 MFCCs. We also computed delta MFCCs and double delta MFCCs. In addition to this we added some features such as RMS energy mean and standard deviation of each sample and High Frequency to low frequency band ratio. We trained an SVM with these features. We split the dataset into 80% train and 20% test from each category. Upon testing these features on an SVM, the accuracy was around 68%.

Approach 2:

We decided to try transfer learning with an existing pretrained VGGish model which is trained on the Audioset dataset. For each audio file, a 5 x 128 tensor was computed. There was an embedding layer on the Mel features of size 128. Evaluating the features computed by the model on an SVM led to results of 76% on the validation set.

Approach 3:

Finally, we decided to train the VGGish model with our dataset and used it to predict the labels of the unknown dataset.

Pre-processing:

We extract mel-spectrogram patches for each of the input audio files. The sample rate of the audio is maintained at its native sampling rate of 44.1 KHz. The STFT

window length is 25 ms and hop length is 10 ms. The number of Mel filters is 64 and the frequency range of the Mel band is 125-7500 Hz. We compute 5 non-overlapping mel-spectrogram patches for each audio file. For each audio file, we generate a 5x64x96 tensor.

Training and Prediction:

We trained the VGGish model using 150 data points per batch for 50 epochs. While predicting, we selected each label by using a majority vote from the 5 predicted outcomes for each audio file.

Reference:

Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R., Seybold, B., Slaney, M., Weiss, R. and Wilson, K. (2018). *CNN Architectures for Large-Scale Audio Classification*. [online] Google AI. Available at: https://ai.google/research/pubs/pub45611.

Contact: mansoor.aiesec@gmail.com