ECS7013: Deep Learning for Audio and Music

Lab 3 – Pretrained Features

In this session we will load an audio research dataset (*warblrb10k*), and analyse it using spectrogram-based audio features, as well as pre-trained deep learning features.

VGGish

AudioSet database from Google is a large-scale database with over 2 million human-labeled 10-second YouTube video soundtracks. Various deep network models have been trained using AudioSet [Gemmeke et. al., 2017], the VGGish model [Hershey et. al., 2017] is one of them. The VGGish model is a variant of the VGG model [Simonyan and Zisserman, 2015] in Figure 1:

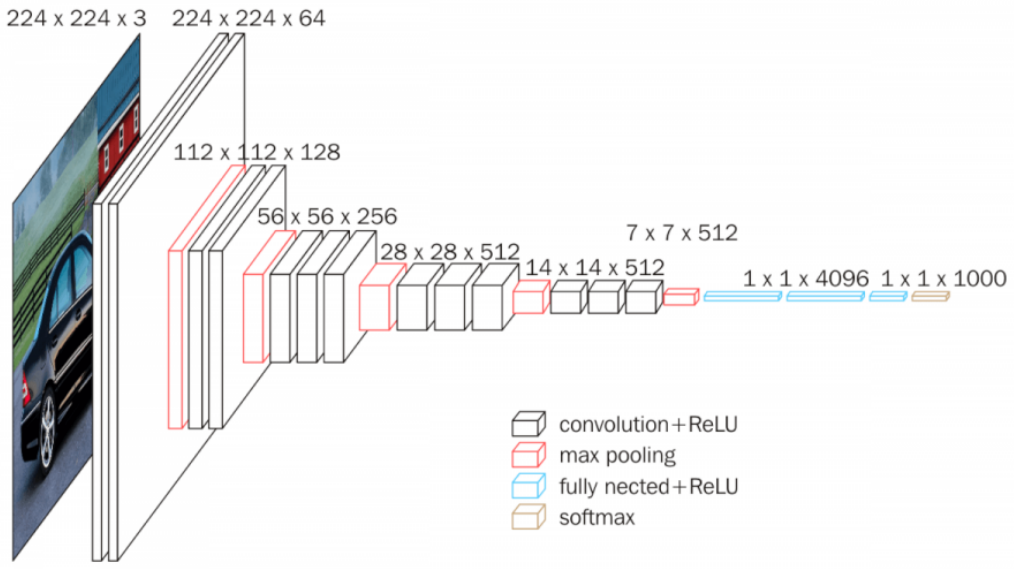

Figure 1 VGG-16 network architecture [Simonyan and Zisserman, 2015]

The VGGish model was trained with audio features computed as follows:

- All audio is resampled to 16 kHz mono.
- A spectrogram is computed using magnitudes of the Short-Time Fourier Transform with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window.
- A mel spectrogram is computed by mapping the spectrogram to 64 mel bins covering the range 125-7500 Hz.
- A stabilized log mel spectrogram is computed by applying log(mel-spectrum + 0.01) where the offset is used to avoid taking a logarithm of zero.
- These features are then framed into non-overlapping examples of 0.96 seconds, where each example covers **64 mel bands and 96 frames** of 10 ms each.

You may want to read [Hershey et. al., 2017] if you want to learn more about this network architecture.

- J. Gemmeke et. al., *AudioSet: An ontology and human-labelled dataset for audio events*, ICASSP 2017
- S. Hershey et. al., *CNN Architectures for Large-Scale Audio Classification*, ICASSP 2017
- K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556, 2015

In this lab, we will analyse and compare the mel-spectrogram features and the features extracted from the the pretrained VGGish model to see if they contain sufficient information to decide a particular file should be 0 or 1. You are going to use:

- principal components analysis (PCA)
- logistic regression

in this lab session.

Note that, to extract the embedding features from the pretrained VGGish model, you will load the WAV audio data and convert each file to mel-spectrogram features. We have learned to use *librosa* to do this (in general, that's a good idea). However, we're going to be using the VGGish model to analyse data, and for the VGGish model, the researchers describe their mel-spectrogram implementation, which is slightly different (see above). So, we will instead use a feature extraction function from the *torchvggish* package.