Literature Review

The first step in solving the problem was to explore existing work. To this end, I went to the benchmark section of papers with code (link). The top performing models (both with and without additional data) were Transformer based. Based on this, I conducted my experiments primarily using Transformer models. The best model available for the given dataset (link) seemed to be primarily focussed on the efficiency of the model.
Before staring any experiments, I established a baseline performance using simple MFCC features and a shallow neural network. This wasn't based on any particular existing model.
For the zero-shot task, I used the CLAP model from LAION (link). CLAP generated embeddings have been shown to be performant across a wide variety of tasks which made it the suitable choice for the task at hand.
For the finetuning part of the assignment, based on my previous experience in the Speech domain, I started with the wav2vec2 model from Facebook (link). Being trained on massive amounts of speech data, the model is often finetuned for various tasks ranging from ASR to Emotion Detection.
Finally, for a model that is close to the domain of the task and the data, I used the AST model from MIT (link) trained on the AudioSet dataset. The dataset provides large variance in the audio samples and thus is a suitable candidate for the problem.

EDA

audio-classification/eda.ipynb at main · Andr0id100/audio-classification (github.com)

Results Analysis

audio-classification/results-analysis.ipynb at main · Andr0id100/audio-classification (github.com)

Potential Improvements

In the scenario where this task would've been around building a testing infrastructure, the scripts would be modified to add an interface using click for better automation.