CSE-6740 Project Proposal

Karl Hiner

November 2, 2023

1 Problem outline & motivation

I plan to investigate the problem of drum instrument classification. Given a raw audio clip, we assume the clip is an audio recording of a drum being struck and that this struck drum instrument is one of a predetermined set (e.g., snare, tom, kick, hi-hat, cymbal), or an overlapping combination of these instruments. The task is to correctly classify the true drum instrument class of the audio clip.

Such a classifier could be used as a component in a system for the automatic transcription of raw recorded audio of drum performances into a symbolic musical notation such as MIDI. Coupled with generative models for the individual drum instruments (such as physical audio models), one could further imagine decoding this estimated lower-dimensional latent representation (of time-stamped drum instrument classifications) back into the raw audio domain as a form of semantically meaningful audio compression over the restricted domain of drum performances.

Specifically, let $\mathbf{x} \in [-1, 1]^n$ denote an audio clip composed of n samples, where each sample is a real number in the range $[-1, 1]$, and let $y \in \{0, \ldots, k\}$ be a drum instrument label, where $y = 1, \ldots, k$ correspond to k predefined drum instrument labels (which can include include instrument combinations), and $y = 0$ denotes the null class (not a known drum instrument).

Let $\mathcal{D}$ be the joint distribution of random variables $Z = (X, Y)$, where X and Y denote the audio clip and the drum label, respectively. The goal is to learn a classifier $h : \mathbf{x} \mapsto \hat{y}$. The quality of h is evaluated by the (generalization) risk,

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[l(h(\mathbf{x}), y)],$$

where l is a loss function.

As a realizable proxy for the generalization risk, we use the empirical risk to evaluate the classifier. Let $S = \{z_i = (\mathbf{x}_i, y_i)\}, 1 \leq i \leq m$ be a labeled training set of m samples. Then, the the empirical risk of classifier h with loss l over S is:

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} l(h(\mathbf{x}_i), y_i)$$

The objective is to minimize $\hat{R}(h)$ to find a classifier h that generalizes well to unseen data from $\mathcal{D}$.

2 Proposed solution

I propose to use cross-entropy loss as the loss function l. The cross-entropy loss is defined for a single sample as:

$$l(h(\mathbf{x}), y) = -\sum_{c=1}^{k} y_c \log(h(\mathbf{x})_c),$$

where $y_c = 1$ iff the sample belongs to class c, and $h(\mathbf{x})_c$ is the predicted probability of the sample $\mathbf{x}$ belonging to class c under classifier h.

This is a natural choice for the multiclass classification problem as it quantifies the error between the predicted and the true class labels in a differentiable expression, and can be interpreted as yielding a probability distribution of prediction confidence over all labels.

The empirical risk $\hat{R}(h)$ with the cross-entropy loss over the training set S is then given by:

$$\hat{R}(h) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{k} y_{i,c}\log(h(\mathbf{x}_i)_c)$$

As for the model architecture, I anticipate that the audio classification task is complex enough that a simple linear model will not suffice, although I may include a linear model as a baseline comparison.

For audio classification, is common to use convolutional neural networks (CNNs) over spectrogram data, so I will likely use some form of CNN over Mel Spectrograms of the audio clips as the primary model architecture, at least as a first attempt.

3 Data

I plan to use the *Expanded Groove MIDI Dataset dataset*[1] for both training and evaluation. This is a dataset of human drum performances containing 444 hours of MIDI-annotated audio recordings from 43 electronic drum kits. It is already split into train/test/validation sets.

4 Inferences

I hope to investigate and learn about the following:

- What are effective features for audio classification, and how do they compare in terms of downstream performance?
- What are effective model architectures for audio classification?
- How well will prediction quality extend to data outside the training set, especially given that the training set is composed of electronic drum kits in a controlled recording environment?

[1] https://magenta.tensorflow.org/datasets/e-gmd