Deep Manifold Learning on Music Audio for Navigating Large Sound Libraries

Abstract

Finding sound for creative purposes is hard because it can be difficult to frame good queries. Use data to learn low-dimensional representations for visualization and browsing, discover latent manifolds in the data.

1. Introduction

While many computational approaches have proven useful for various classification or recognition tasks, none directly result in a notion of timbre similarity, a useful concept with a variety of applications. One notable instance is the difficulty faced in the search and navigation of large sound sample libraries. Queries are predominantly forced to take the form of text, as in the Freesound[1] archive shown in Figure **??**, which is problematic for at least two reasons. On one hand, it can challenging to describe a specific query semantically, and often metaphors and figurative language are used to relate the experience of a sound; a distorted guitar might be referred to as 'crunchy', or a trumpet as 'bright.' Conversely, this kind of descriptive language is far from standardized and varies in meaning from one individual to the next. Furthermore, such descriptions are not always associated with every sound in a collection, and typically only at the granularity of the entire recording. As a result, the task of navigating a sound library is often reduced to that of an exhaustive, brute force search.

2. Approach

Here we study a class of manifold learning algorithms which seeks to preserve or establish neighborhood relationships in a low-dimensional space. Given a set of observations $\mathcal{D}$, positive and negative partitions, Γ_k and $\bar{\Gamma_k}$, may be defined, satisfying three conditions: corresponding positive and negative partitions are disjoint, $\Gamma_k \cap \bar{\Gamma_k} = \varnothing$; corresponding positive and negative partitions may comprise a subset of the entire collection, $|\Gamma_k \cup \bar{\Gamma_k}| \leq |\mathcal{D}|$; and, any two partitions, Γ_j, Γ_k, may share observations, $|\Gamma_j \cap \Gamma_k| \geq 0$. Therefore we are interested primarily in how these partitions are defined and sampled during training.

[1] http://www.freesound.org/

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2.1. Model

Audio is first transformed to a constant-Q representation, parameterized as follows: signals are downsampled to 16kHz; bins are spaced at 24 per octave, or quarter-tone resolution, and span eight octaves, from 27.5Hz to 7040Hz; analysis is performed at a framerate of 20Hz uniformly across all frequency bins. Logarithmic compression is applied to the frequency coefficients with an offset of one, i.e. $log(C * X + 1.0)$, where we set $C = 50$.

Extending the work presented in (**?**), 500ms windows of time-frequency coefficients are transformed by a five-layer convolutional neural network, consisting of three 3D-convolutional layers and two fully connected layers. Max-pooling by a factor of 2 is used along time in first two layers, and all but the last layer use hyperbolic tangent activation functions, which is linear. The final output is 3-dimensional for ease of visualization and interaction.

2.2. Training Strategy

Similar to some work in the application of deep networks for learned embeddings (**???**), we use a contrastive loss term to draw neighborhoods together and drive negative observations apart. Notably, we augment training criterion to use a ternary network configuration, rather than the pairwise harness employed previously, defined as follows:

$$
\begin{aligned}
Z_i = \mathcal{F}(X_i|\Theta), Z_p = \mathcal{F}(X_p|\Theta), Z_n = \mathcal{F}(X_n|\Theta) \\
D_p = ||Z_i - Z_p||_2, D_n = ||Z_i - Z_n||_2 \\
\mathcal{L} = \max(0, D_p^2 - m_p) + \max(0, m_n - D_n)^2
\end{aligned}
$$

Here, three inputs – X_i, X_p, X_n – are transformed by the model, $\mathcal{F}$, given the same parameters, Θ. These observations are chosen such that $X_i, X_p \in \Gamma_k$ and $X_n \in \bar{\Gamma_k}$. Euclidean distance is computed between the positive and

negative embedding pairs, and two margins, m_p and m_n, define a floor on the positive and negative loss terms.

2.3. Data

The data source used herein is drawn from the Vienna Symphonic Library (VSL), a massive collection of studio-grade orchestral instrument samples recorded over a variety of performance techniques[2]. We leverage a previously sampled collection (**?**), comprised of 5k samples drawn from 24 of the largest instrument classes. The set is partitioned into 72k, 18k, and 30k for training, validation, and testing, respectively.

As discussed above, the crux of this exploration lies in *how* neighborhoods are defined and sampled for training. We consider a number of options in order to explore what manifolds or behaviors might reveal themselves in the data:

- Nearest neighbors in the input space, like (**?**)
- Instrument class, as in (**?**)
- Absolute pitch
- Instrument class and absolute pitch
- Instrument class and absolute pitch, $\pm$ 2

3. Experimental Results

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

Concatenative synthesis, qualitative analysis.

Interpretable embeddings. Follow trajectories of sounds in 3-space.

4. Conclusions

We have explored

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it

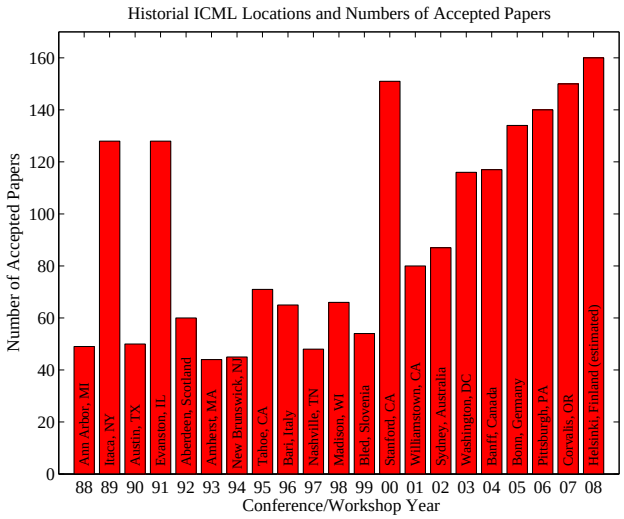

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX), but always place two-column figures at the top or bottom of the page.

[2] https://vsl.co.at/en