Deep Manifold Learning on Music Audio for Navigating Large Sound Libraries

Abstract

Use data to learn low-dimensional representations of acoustic similarity for visualization and browsing, discover latent manifolds in the data. We present an approach leveraging different neighborhood relationships drawn between a large collection of instrument samples. Describe insights and produce a number of examples that illustrate the behaviors induced by this approach.

1. Introduction

Navigating large audio sample libraries has long been a pain point for musicians and sound artists alike. Search queries are predominantly forced to take the form of text, which is problematic for at least two reasons. Metaphors and descriptive tags, when provided, struggle to capture important characteristics in sufficient detail, and this language often varies from one individual to the next. Alternatively, standard approaches to sound visualization – waveform envelopes or spectrograms – are hardly intuitive for the general population. As a result, the development of computational systems for acoustic similarity, an elusive concept in its own right, remains an open research topic.

Acoustic similarity is a natural use case for manifold learning, which attempts to preserve relationships in some low dimensional space, typically for visualization. Common embedding methods, such as Multidimensional Scaling (MDS) (), Locally Linear Embedding (LLE) (), or Isomap (), respect pairwise distances between observations, but exhibit two practical drawbacks: these methods do not yield general functions that can be applied to new data, and obtaining accurate pairwise distances for large datasets is not scalable. Ranking methods, like WSABIE (**?**), relax this constraint in favor of maintaining relative order between observations, but these nuanced relationships can still be hard to obtain at scale. Neighborhood methods, such as neighborhood components analysis (NCA) (**?**), DrLIM (**?**), or `word2vec` (**?**), simplifiy the problem even further by exploiting unordered set relationships. Rather than placing the burden of continuity on the data, neighborhood methods task the model with smoothly interpolating discrete sets in the embedding space. Prior work has demonstrated the potential for such methods to yield intutitive representations, where algebraic operations on vectors encode physical orientation (**?**) or analogy (**?**).

Synthesizing these two topics, we explore the potential of neighborhood-based "deep" manifold learning with a large collection of instrument sounds for developing intuitive acoustic representations. Deep neural networks provide a general framework for learning functions that map high-dimensional data into useful embeddings, which we influence by considering different kinds of relationships between inputs. The behavior of the resulting embeddings is evaluated through both quantitative and qualitative means, yielding a variety of insights and audio-visual examples.

2. Method

We approach acoustic similarity by optimizing the parameters of a deep neural network to maximally preserve K contrasting set relationships, *i.e.* neighborhoods, between samples in a low-dimensional Euclidean space. Given a collection of observations, $\mathcal{D}$, the k^{th} contrastive parition consists of a positive, Γ_k, and a negative, $\bar{\Gamma}_k$, subset, satisfying three conditions: one, contrastive partitions are internally disjoint, $\Gamma_k \cap \bar{\Gamma}_k = \varnothing$; two, contrastive partitions may comprise a subset of the entire collection, $|\Gamma_k \cup \bar{\Gamma}_k| \leq |\mathcal{D}|$; and three, contrastive partitions are drawn independently of each other, such that any two partitions, i and j, may share observations, $|\Gamma_i \cap \Gamma_j| \geq 0$. Thus, the network can be understood as interpolating the various discrete partitions, with the training objective finding a smooth compromise between them.

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 ap-

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

plied to the frequency coefficients with an offset of one, i.e. $log(C * X + 1.0)$, where $C = 50$.

Keeping with related work (**?**), windows of time-frequency coefficients are transformed by a five-layer convolutional neural network (CNN) into a 3-dimensional embedding for the purposes of visualization. The network consists of three 3D-convolutional layers and two fully connected layers, with max-pooling by a factor of 2 in time for first two layers. The first four layers use hyperbolic tangent activation functions, while the last layer is linear to avoid saturation.

2.2. Learning a Mapping into Euclidean Space

A contrastive loss function is used to optimize the model's parameters, following prior research on deep manifold learning (**???**). However, we augment training criterion to use a ternary network configuration, rather than the pairwise one 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, an observation, X_i, a positive neighbor, X_p, and a negative example 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.

This loss is computed over a mini-batch of observations, differentiated with respect to the parameters of model, and back-propagated through the network via simple stochastic gradient descent. Mini-batches consist of 32 triples, and training proceeds for 50k iterations.

2.3. Data

We use a previously compiled collection of solo instrument samples, comprised of 5k instances drawn from 24 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 types in the hopes that different behaviors might reveal themselves in the data: (1) nearest neighbors in the input space; (2) instrument class; (3) absolute pitch; (4) instrument class and absolute pitch; (5) instrument class and absolute pitch, ± 2.

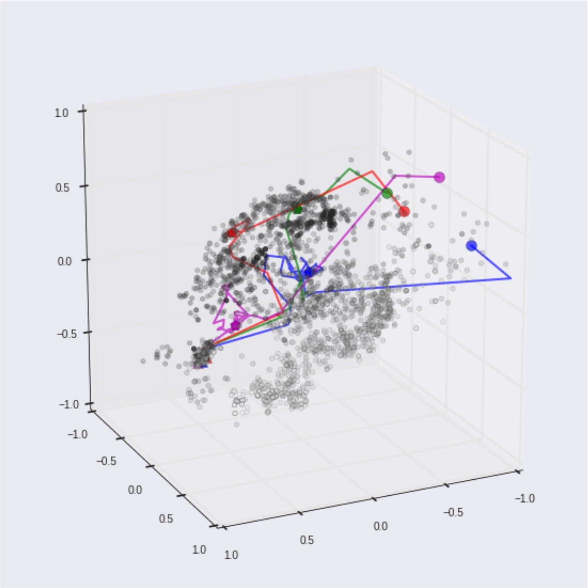

Figure 1. Embedding space revealing an ADSR envelope.

3. Evaluation

Given various learned mappings, there are a number of ways in which we can assess the "quality" of the resulting representation. Having trained the mappings on neighborhood relationships, we use k-nearest neighbor classification as a quantitative approach to measure the extent to which this is preserved. Consistent with previous results (**??**), the sharpness of class boundaries is inversely to the discrete number of neighborhoods, and is likely due in part to the limits of three-space. Regardless, the learned representations consistently demonstrate solid performance in ranked retrieval settings. We can additionally test the space for acoustic "analogy", where the resultant vector between two points is added to a third. Though inherently subjective, this approach is enticing as a new mode of acoustic search.

Can we visually interpretable embeddings? Follow trajectories of sounds in 3-space. ADSR Sonification, concatenative synthesis.

4. Summary

We have explored a range of embeddings learned using a convolutional neural network optimized to preserve various neighborhood relationships between instrument sounds in 3 dimensions. Interesting to