Deep Manifold Learning on Music Audio for Navigating Large Sound Libraries

Eric J. Humphrey EJHUMPHREY@SPOTIFY.COM
Spotify USA, 620 Avenue of the Americas, New York City, NY 10011 USA

Abstract

In this talk, we explore the application of manifold learning with solo instrument samples to yield low-dimensional representations of sound for visualization and browsing. Our approach uses deep neural networks to jointly learn features and the embedding space, trained to optimally preserve neighborhood relationships. We then turn our attention to evaluating the usefulness of the resulting embeddings, both through benchmarking in retrieval tasks and presenting notable multimedia examples.

1. Introduction

Navigating large sound libraries has long been a pain point for musicians, artists, and producers. 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 a low dimensional space, often for visualization. Common embedding methods, such as Multidimensional Scaling (MDS) (Borg & Groenen, 2005), Locally Linear Embedding (LLE) (Roweis & Saul, 2000), or Isomap (Tenenbaum et al., 2000), 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 (Weston et al., 2011), 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) (Goldberger et al., 2004; Salakhutdinov & Hinton, 2007), DrLIM (Hadsell et al., 2006), or `word2vec` (Mikolov et al., 2013b), 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 (Hadsell et al., 2006) or analogy (Mikolov et al., 2013a).

Synthesizing these two topics, neighborhood-based "deep" manifold learning methods have shown promise in producing intuitive acoustic representations from sound samples (Humphrey et al., 2011), and we look to extend that work here. Deep neural networks provide a general approach to mapping high-dimensional data into useful embeddings, which we influence by considering different kinds of relationships between inputs. The semantic organization of the learned embeddings is evaluated through various means, and serves as vehicle for exploration in itself.

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

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

understood as interpolating the various discrete partitions, with the training objective finding a smooth compromise between them.

2.1. Model

The model used here follows prior work (Humphrey, 2015). 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. Logarithmic compression is applied to the frequency coefficients, i.e. $log(C * X + 1.0)$, with $C = 50$. Multi-frame windows are transformed by a five-layer convolutional neural network (CNN) into a 3-dimensional vector. The network consists of three 3D-convolutional layers and two fully connected layers, with max-pooling by 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, with copies of the model arranged in a ternary configuration, defined as follows:

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

Here, three observations, 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 embedding pairs, and two margins, m_p and m_n, define thresholds on their respective loss terms. Parameters are learned via mini-batch stochastic gradient descent with a batch size of 32, and training proceeded 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 (Humphrey, 2015). The set is partitioned into 72k, 18k, and 30k for training, validation, and testing, respectively. The crux of this exploration lies in *how* neighborhoods are defined and sampled for training, and thus various relationships are considered to encourage different kinds of embeddings, including instrument class, absolute pitch, instrument class and absolute pitch, and instrument class and absolute pitch $\pm$ 2 semitones.

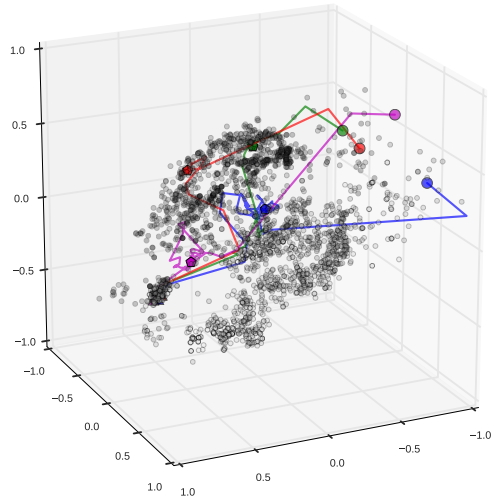

Figure 1. An embedding where random samples (grayscale) show pitch height from low (black) to high (white) and four sounds are shown from start (cirlces) to end (triangles): trumpet (blue), trombone (green), tenor saxophone (red) and clarinet (magenta).

3. Evaluation

We consider a few approaches to measuring the quality of the learned acoustic embeddings, the details of which are provided externally.[1] Quantitatively, k-nearest neighbor classification measures the extent to which neighborhoods are preserved, and brute-force distance is used to sort data in a ranked retrieval setting. These representations additionally lend themselves well to visualization and sonification. Scatterplots reveal some surprising behavior, as in Figure 1, where an attack-decay-sustain-release (ADSR) envelope is made apparent by following the evolution of sounds through space. Acoustic "analogy" serves as another good, if subjective, test of semantic organization, where the resultant vector between two points is applied to a third. Lastly, arbitrary sound trajectories are sonified via concatenative synthesis, providing audible insight into the learned embeddings.

4. Summary

A range of acoustic embeddings are learned using a convolutional neural network optimized to preserve different neighborhood relationships between instrument sounds in 3-space. The learned embeddings demonstrate a high degree of semantic organization, indicated by strong performance on retrieval tasks and noteworthy multimedia examples, provided externally. In doing so, we hope to further motivate the usefulness of sound visualization in discovery, how deep networks can help address this challenge, and the potential this area holds for future exploration.

[1] https://github.com/ejhumphrey/icml16-dml

References

Borg, Ingwer and Groenen, Patrick J.F. *Modern Multidimensional Scaling: Theory and Applications*. Springer-Verlag, 2 edition, 2005.

Goldberger, Jacob, Hinton, Geoffrey E, Roweis, Sam T, and Salakhutdinov, Ruslan. Neighbourhood components analysis. In *Advances in neural information processing systems*, pp. 513–520, 2004.

Hadsell, Raia, Chopra, Sumit, and LeCun, Yann. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1735–1742. IEEE, 2006.

Humphrey, Eric J. *An Exploration of Deep Learning in Music Informatics*. PhD thesis, New York University, 6 2015.

Humphrey, Eric J., Glennon, Aron P., and Bello, Juan P. Non-linear semantic embedding for organizing large instrument sample libraries. In *International Conference on Machine Learning and Applications*, volume 2, pp. 142–147. IEEE, 2011.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *International Conference on Learning Representations Workshop*, 2013a.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.

Roweis, Sam T and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Salakhutdinov, Ruslan and Hinton, Geoffrey E. Learning a nonlinear embedding by preserving class neighbourhood structure. In *International Conference on Artificial Intelligence and Statistics*, pp. 412–419, 2007.

Tenenbaum, Joshua B, De Silva, Vin, and Langford, John C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Weston, Jason, Bengio, Samy, and Usunier, Nicolas. WSABIE: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence*, volume 11, pp. 2764–2770, 2011.