ACCELERATING MULTIMODAL SEQUENCE RETRIEVAL WITH CONVOLUTIONAL NETWORKS

Colin Raffel (craffel@gmail.com) and Daniel P. W. Ellis (dpwe@ee.columbia.edu), LabROSA, Department of Electrical Engineering, Columbia University, New York

Abstract

Warping-based metrics such as Dynamic Time Warping (DTW) are a natural choice for the task of finding the entry in a database of sequences which is the most similar to a query sequence. However, the quadratic cost of the dynamic programming-based alignment operation can make nearest-neighbor search infeasible for large databases with long and/or high-dimensional sequences. It also requires that the feature vectors from each sequence are directly comparable by their Euclidean distance. In [1], we utilized a convolutional network to map sequences of feature vectors to downsampled sequences of binary vectors. Here, we show that this approach is applicable to multimodal settings.

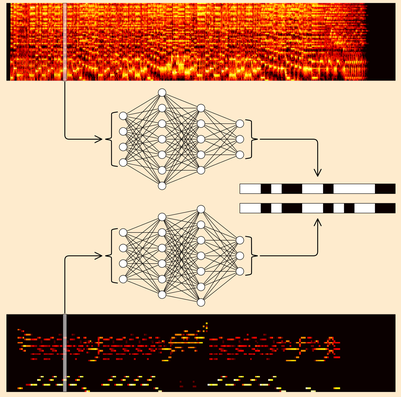
Similarity-Preserving Hashing

Preserves similarity: Our objective encourages a mapping where aligned feature vectors from matching sequences have a small Hamming distance in the embedded space and non-matched feature vectors have a large distance.

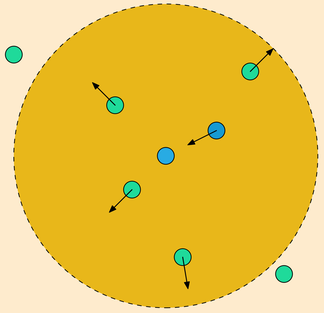
Learns its representation: Our approach is entirely learning-based, which allows it to adapt to problem settings where Euclidean distance is inappropriate (e.g. multimodal data).

Maps to a Hamming space: By replacing continuous-valued feature vectors with bitvectors in an embedded Hamming space, computing pairwise distances simplifies to a single exclusive-or and a table lookup.

Downsamples sequences: Groups of subsequent feature vectors are mapped to a single bitvector, giving a quadratic increase in efficiency.



Objective Function

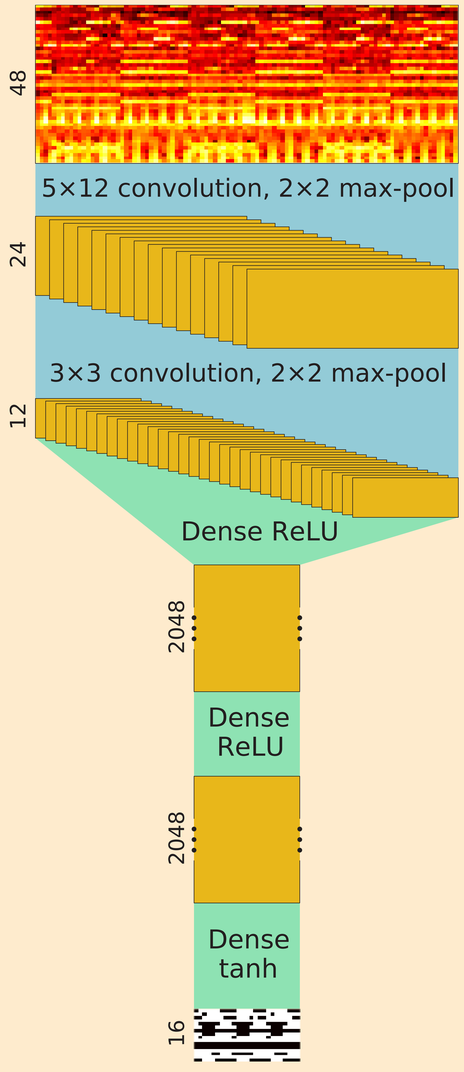


Our training data consists of a set $\mathcal{P}$, such that $(x, y) \in \mathcal{P}$ indicates that x is a feature vector in some sequence from one modality which is aligned to y in a matching sequence from another modality. We then construct $\mathcal{N}$ by repeatedly choosing two pairs $(x_1, y_1), (x_2, y_2) \in \mathcal{P}$ and swapping entries to construct $(x_1, y_2), (x_2, y_1) \in \mathcal{N}$. Motivated by [2], we use the following objective function:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 - \frac{\alpha}{|\mathcal{N}|} \sum_{(a,b)\in\mathcal{N}} \max(0, m - \|f(a) - g(b)\|_2)^2$$

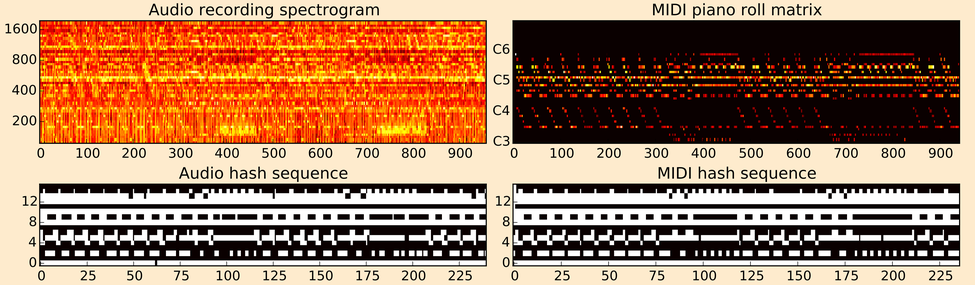
where f and g are learned nonlinear functions, α is a parameter to control the importance of separating dissimilar items, and m is a target separation of dissimilar pairs.

Network Structure



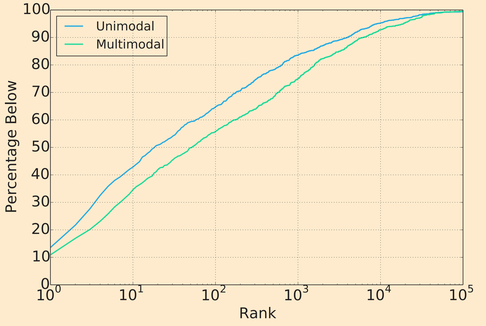
MIDI to Audio Matching Experiment

We tested the effectiveness of this approach on the task of matching MIDI files (musical scores) to recordings of music in the Million Song Dataset (MSD) [3]. For the first modality, we used log-magnitude constant-Q spectrograms of the audio recordings. For the second, we tested using either constant-Q spectrograms of synthesized renditions of the MIDI files (unimodal) or a "piano roll" matrix representation (multimodal).



Results

To evaluate, we trained our model on a collection of MIDI and audio recording pairs which were pre-aligned using DTW. We then computed hash sequences for every entry in the MSD and a held-out set of 1,537 MIDI files for which we knew a priori the correct match. We measured performance as the percentage of MIDI files in this test set where the correct match in the MSD ranked below a certain threshold. In the unimodal setting, this approach ranked the correct entry in the top 1% (corresponding to 10,000 entries) 95.9% of the time; for multimodal sequences, performance degraded slightly to 92.8%.



References

[1] Colin Raffel and Daniel P. W. Ellis. Large-scale content-based matching of MIDI and audio files. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015.

[2] Jonathan Masci, Michael M. Bronstein, Alexander M. Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4), 2014.

[3] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, 2011.