Accelerating Multimodal Sequence Retrieval with Convolutional Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given a large database of sequential data, a natural problem is to find the entry in the database which is most similar to a query sequence. Warping-based similarity metrics such as the dynamic time warping (DTW) distance can be prohibitively expensive when the sequences are long and/or high-dimensional. To mitigate these issues, [1] utilizes a convolutional network to map sequences of feature vectors to downsampled sequences of binary vectors. On the task of matching synthetic renditions of pieces of music to a large database of audio recordings of songs, this approach was able to efficiently discard 99% of the database with high confidence. We extend this approach to the multimodal setting where rather than synthetic renditions a matrix representation of the piece's score is used instead, demonstrating that this approach is adaptable to the underlying representation.

1 Introduction

The ability to compute a similarity metric for sequences of feature vectors is necessary for the task of retrieving the most similar entry (nearest-neighbor search) in a database of sequences. A natural way to compare sequences is to first find their optimal alignment and then compute the total distance between aligned feature vectors. Aligning the sequences before computing the total distance makes metrics of this type robust to timing distortions (e.g. offset, skew, or cropping), and can be achieved in quadratic time using dynamic programming [2]. For feature vectors which are effectively compared with Euclidean distance (e.g. those with continuously-valued feature vectors), the most commonly used method of this type is dynamic time warping (DTW) [3], which will be the focus of this work.

The quadratic cost of the dynamic programming-based alignment operation can make nearest-neighbor search infeasible for for databases with many entries and/or long sequences. Furthermore, traditional DTW involves computing the pairwise distance between all feature vectors in the two sequences being compared, which can outweigh the cost of the alignment for high-dimensional data. A common way to avoid these costs is to use "pruning" techniques, which use heuristics to skip a large portion of the database. A wide variety of pruning methods have been proposed; in [2] it is shown that their successful application can enable nearest-neighbor search in databases with trillions of sequences. Despite their benefits, pruning methods typically rely on various constraints on the comparisons being made, such as the query sequence always being a subsequence of its correct match in the database or that the total number of aligned frames is fixed. Furthermore, these methods suffer losses in efficiency when sequences are oversampled and/or high dimensional.

To avoid these issues, in [1] a learning-based method is proposed which utilizes a convolutional network to map sequences of feature vectors to downsampled sequences of binary vectors. The resulting "hash sequences" can be much more efficiently compared using dynamic time warping, which enables flexible, problem-adaptive pruning. In this paper, we will show that this framework is

additionally flexible to multimodal settings, where the sequences being compared represent different types of data.

2 Learning a More Efficient Representation for DTW

[1]

3 Shared Representations for Multimodal Sequences

Piano roll matrices, 44.8 ms vs 22.2 s.

3.1 MIDI to Audio Matching Experiment

Re-cap of experiment

Results

4 Extensions

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] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, et al. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 262–270, 2012.

[3] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics Speech and Signal Processing*, 26(1):43–49, 1978.