Learning Efficient Representations for Sequence Retrieval

Colin Raffel
LabROSA, Columbia University
1300 SW Mudd, 500 W 120th Street
New York, NY 10027
craffel@gmail.com http://www.colinraffel.com

Keywords:
Time Series
Hashing
DTW

Background

In many domains, the most natural representation for data is as sequences of feature vectors. For example, in speech recognition, recorded utterances are typically transformed into series of vectors which describe the frequency content over short periods of time [1]. Similarly, in natural language processing tasks, sentences are often represented as sequences of vectors where each word corresponds to a unique vector [2]. Many off-the-shelf machine learning approaches assume that feature vectors are independent, so modeling the sequential nature of these representations often necessitates special treatment.

A common task in data science is the retrieval of the entry in a database which is the most similar to a query. A straightforward approach is to compute the distance between the query and each entry in the database and choose the entry with the smallest distance (often called "1-nearest neighbor", or 1-NN). When the data is sequential, a natural metric is the dynamic time warping (DTW) distance, which computes the total distance between feature vectors in each sequence after finding their optimal alignment [1]. The DTW distance between a sequence of length N and a sequence of length M can be found in $\mathbf{O}(NM)$-time using dynamic programming. For a database with D entries, D $\mathbf{O}(NM)$ calculations must be made, which can be prohibitively expensive for large D. As a result, many "pruning techniques" have been proposed which can speed-up 1-NN sequence matching by using heuristics to skip most entries in the database [3]. These techniques make it feasible to match query sequences against databases with trillions of entries.

Problem Statement

Despite the speedups offered by pruning techniques, they nevertheless rely on calculating many "local" distances between individual feature vectors. When the data is high-dimensional, these distance calculations can be expensive, and may offset the speed-ups provided by pruning. Furthermore, the Euclidean distance is often used as the local distance metric, which may not be the optimal way to compare feature vectors. It is also an inappropriate choice when the sequences being compared come from different modalities, e.g. when comparing acoustic features to transcribed text. Finally, it is noted in [3] that data sequences are often "oversampled", i.e. the intrinsic sampling rate of the data is too high. This can result in quadratic speed and memory penalties in practice.

Recently, a good deal of research effort has focused on learning data representations in a supervised manner which are more effective than hand-designed features [4]. These approaches have been applied to the problem of learning features which allow efficient distance computation. For example, by mapping features to a Hamming space, their distance can be computed efficiently by a single exclusive-or operation and a table lookup [5]. This technique has been shown to work well even in the cross-modality case [6]. In the course of learning representations for vectors in sequences, groups of subsequent feature vectors can be combined into a single vector, which has the effect of downsampling the sequences.

In [5], we propose a system which exploits the aforementioned benefits of feature learning for sequence matching. Our approach uses convolutional networks to map sequences of feature vectors to sequences of binary hashes in a Hamming space where they can be efficiently compared. By max-pooling in the time dimension, we also implicitly lower the data's sampling rate in the process of this transformation. We evaluated our technique on the task of matching musical scores to a large database of song recordings. On this task, we achieved a 100x speed-up relative to performing DTW using Euclidean distance and can avoid 99% of Euclidean distance calculations with high confidence. An example of a pair of matching sequences of feature vectors and their Hamming-space hash representations can be seen in Figure 1.

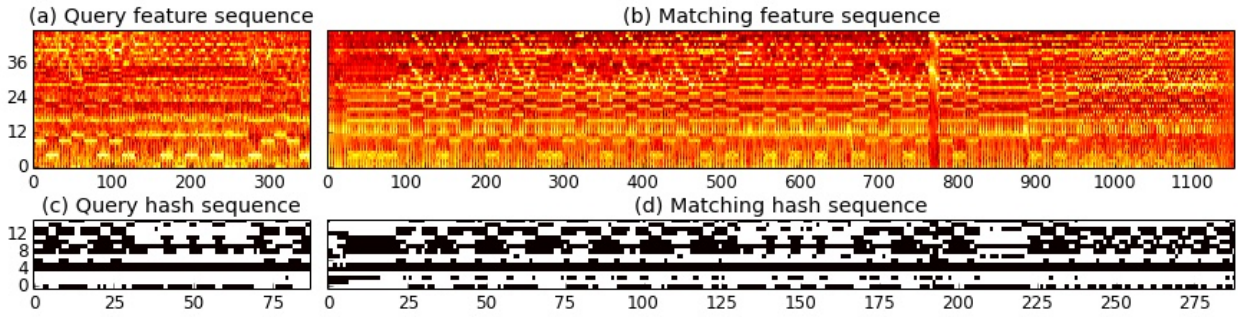

Figure 1: Acoustic and Hamming-space features for a single sequence from the experiment described in [5]. The query sequence is a subsequence of its match. (a) Feature vectors for the query sequence, which lie in R^{48}. (b) Original feature sequence for the correct match in our database. (c) Hash sequence representation of the query, which lies in 16-dimensional Hamming space. (d) Hash sequence for the correct match.

Despite our system's success, it has a few key shortcomings. First, in order to train the convolutional hashing networks, our approach requires a dataset of sequences of feature vectors which are pre-aligned. This is due to the fact that the network's objective is not able to optimally align the output hash sequences when measuring the network's fitness. This limits the training set size, and therefore likely the effectiveness, of our approach. Recently, a handful of objective functions have been proposed which perform alignment as part of fitness estimation [2] [7]. Applying these techniques to our system could allow us to utilize unaligned training sequences, which would greatly increase the amount of data used for optimization.

As an alternative approach, it has recently been shown that in some settings the information in a variable-length sequence can be compressed into a relatively small fixed-length vector using recurrent networks [2]. If this approach is shown to be more widely applicable, it could obviate the need for a costly metric like DTW. Instead, sequences could be compressed to single fixed-length vectors which could then be compared trivially with off-the-shelf distance metrics.

Broader Impacts

An efficient scheme for comparing and retrieving high-dimensional sequences could be applied in a wide variety of disciplines. As touched on in [3], many problems in biomedical science involve characterizing signals measured from the human body. However, some sensor arrays produce many signals simultaneously (electroencephalograms sometimes use up to 256 sensors) and may be oversampled [3]. A technique for mapping these high-dimensional vector sequences to efficiently comparable hash sequences at a lower sampling rate would therefore facilitate much larger-scale sequence characterization. This is particularly important as these multi-sensor arrays become more and more commonly used, resulting in huge collections of signal recordings.

In a completely different setting, the proposed technique could benefit the task of plagiarism detection, which can be viewed as attempting to find subsequences in a query (an "original" document) which are highly similar to a subsequence in a large dataset (a collection of published documents). This leads to another large-scale high-dimensional subsequence comparison problem because words are often represented as very high-dimensional vectors [2], and the number of published documents may be huge. As with the other applications discussed above, and likely many others, the use of a more efficient and effective representation would make such large-scale high-dimensional sequence comparisons possible.

References

[1] Rabiner, Lawrence, and Biing-Hwang Juang. "Fundamentals of speech recognition." (1993).

[2] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[3] Rakthanmanon, Thanawin, et al. "Searching and mining trillions of time series subsequences under dynamic time warping." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[4] Bengio, Yoshua, Aaron Courville, and Pierre Vincent. "Representation learning: A review and new perspectives." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35.8 (2013): 1798-1828.

[5] Raffel, Colin, and Daniel P. W. Ellis. "Large-Scale Content-Based Matching of MIDI and Audio Files." To appear in *16th International Society for Music Information Retrieval Conference (ISMIR 2015)*.

[6] Masci, Jonathan, et al. "Multimodal similarity-preserving hashing." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 36.4 (2014): 824-830.

[7] Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.