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.

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.