Time-series segmentation and latent representation of musical instruments

Gregory Szep

King's College London

July 11, 2018

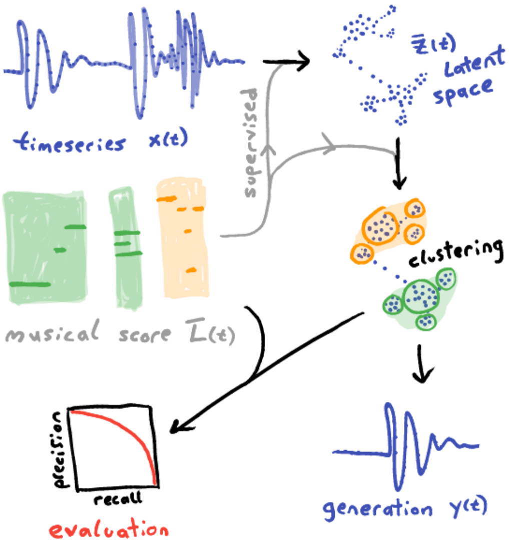
Abstract

Music information retrieval tasks serve as faithful benchmarks for time-series analysis pipelines due to the availability of strongly labelled training data such as MusicNet. Clustering algorithms in spectral sub-spaces, hidden Markov models and causal convolutional neural networks are compared in their ability to transform time-series to a continuous latent space that clusters eleven orchestral instruments. The latent space is evaluated quantitatively with precision-recall metrics obtained by comparing the instrument prediction from a segment of audio to the ground truth obtained from musical scores, and qualitatively by generating samples of audio for given regions in the latent space.

1 Problem Outline

1.1 Mapping time-series to latent space

The input data are single channel time-series points $\mathcal{D} = \{x(t_1)\dots x(t_N)\}$ sampled at frequency f from an underlying continuous state-time process $x(t)$, that is the oscillating sound waves emitted by a live orchestra.



2 Clustering in spectral sub-spaces

3 Hidden Markov models

4 Causal convolutional networks

Convolutional architectures have become popular due to their ability to compress spatio-temporal information for discrimination and generation tasks [1, 2]. A causal convolutional network [3] — which encodes the arrow of time in its architecture — is trained for the audio segmentation task.

References

[1] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel Recurrent Neural Networks," 1 2016.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets,"

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 2016.