Proceedings of the 10th International Conference of Students of Systematic Musicology (SysMus17), London, UK, September 13-15, 2017.
Peter M. C. Harrison (Ed.).

Neural Music Language Models: Investigating the Training Process

Adrien Ycart[1], Emmanouil Benetos[2]

Centre for Digital Music, Queen Mary University of London
[1]a.ycart@qmul.ac.uk, [2]emmanouil.benetos@qmul.ac.uk

ABSTRACT

Background

Automatic music transcription (AMT) is the problem of converting an audio signal into some form of music notation. It remains a challenging task, in particular with polyphonic music (Benetos et al., 2013)

In most AMT systems, an *acoustic model* estimates the pitches present in each time frame, and a *language model* links those estimations using high-level musical knowledge to build a binary piano-roll representation. While the former task has been widely discussed in the literature, the latter has received little attention until quite recently (Raczyński et al., 2013; Sigtia et al., 2015).

Aims

We aim to investigate the use of recurrent neural networks (RNN) as language models for AMT to estimate the probability of pitches present in the next frame, given the previously observed. Most of the existing literature focuses on the architecture; here we will investigate the training process. More precisely we will consider how the choice of the time steps, the choice of the training set, and various data augmentation techniques can influence their predictive power.

Method

We will train a simple Long Short-Term Memory (LSTM) architecture (Hochreiter & Schmidhuber, 1997) with polyphonic MIDI data, taken from a classical piano music dataset[1]. The performance of the resulting RNN will be compared in terms of prediction accuracy and cross-entropy. We will compare time steps in physical time and in fractions of a beat, similarly to a study by Korzeniowski and Widmer (2017). We will investigate the influence of various types of training data (different genres, composers, artificial data). We will also assess how data pre-processing (cutting the training sequences into smaller chunks) and data augmentation (transposition, time-stretching) can improve the results.

Results

This research is ongoing; most results have yet to be obtained. The first results suggest that time-steps in milliseconds perform better in terms of prediction because self-transitions are more frequent, but do nothing more than a simple smoothing. On the other hand, time-steps of a sixteenth note perform worse on prediction, but they allow to better model tonality and meter.

Conclusions

This study will be a first step towards implementing a neural music language model (MLM). It will later be integrated with state-of-the-art acoustic models to make a full AMT system; experiments will be carried out in future work on how MLMs can improve AMT performance.

Keywords

Automatic music transcription, neural networks, music language models, polyphonic music prediction

[1]www.piano-midi.de

REFERENCES

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, *41*(3), 407-434.

Raczyński, S. A., Vincent, E., & Sagayama, S. (2013). Dynamic Bayesian networks for symbolic polyphonic pitch modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(9), 1830-1840.

Sigtia, S., Benetos, E., & Dixon, S. (2015). An end-to-end neural network for polyphonic music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, *24*(5), 927–939.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Korzeniowski, F., & Widmer, G. (2017). On the Futility of Learning Complex Frame-Level Language Models for Chord Recognition. *arXiv preprint arXiv:1702.00178*.

ACKNOWLEDGEMENTS

AY is supported by a QMUL EECS Research Studentship. EB is supported by a RAEng Research Fellowship (RF/128).