Neural Music Language Models: investigating the training process

Adrien Ycart*1, Emmanouil Benetos *2

*Centre for Digital Music, Queen Mary University of London
1a.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 given RNN architecture with polyphonic MIDI data, pre-processed in various ways. The performance of the resulting RNN will be compared in terms of perplexity. We will compare time steps in physical time and in fractions of a beat. We will investigate the influence of various types of training data (different genres, composers, artificial data). We will also assess how data augmentation (transposition, time-stretching) can improve the results.

Results

This research is ongoing, and results have yet to be obtained. A recent study (Korzeniowski & Widmer (2017)) hints that frame-level language models for chord estimation from audio are inefficient when used with time-steps in milliseconds. We aim at confirming those findings for AMT.

Conclusions

This 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; experiments will be carried out on how MLMs can improve AMT performance.

Keywords

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

REFERENCES (if needed)

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.

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