Modelling High-Dimensional Sequences with LSTM-RTRBM: Application to Polyphonic Music Generation

Qi Lyu[1], Zhiyong Wu[1,2], Jun Zhu[1], Helen Meng[2]

[1]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, China
qilyu.pub@gmail.com, {zywu, hmmeng}@se.cuhk.edu.hk, dcszj@tsinghua.edu.cn

Abstract

We propose an automatic music generation demo based on artificial neural networks, which integrates the ability of Long Short-Term Memory (LSTM) in memorizing and retrieving useful history information, together with the advantage of Restricted Boltzmann Machine (RBM) in high dimensional data modelling. Our model can generalize to different musical styles and generate polyphonic music better than previous models.

1 Introduction

Music is among the most widely consumed types of signal streams. Models for finding, extracting and reproducing musical temporal structure are of considerable interest. In particular, generative models for composing (good) music might have not only artistic value but also commercial potential. In the belief that memory is one of the vital intelligence needed for music generation, we introduce a model that specializes in memorization and can generate beautiful music pieces without human interference.

Traditionally, there are many sequence models that can be utilized for modelling the music generation process, such as hidden Markov hodel (HMM) [Allan and Williams, 2005], Markov random field [Lavrenko and Pickens, 2003], etc. Recurrent Neural Network (RNN) [Rumelhart *et al.*, 1986], with its internal dynamics trainable by back-propagation through time (BPTT), is simple yet powerful for modelling sequences. In principle, a large enough RNN can be sufficient to model sequences of arbitrary complexity. In practice however, it is difficult for RNN to store lengthy historic information about a sequence. Complex sequences are usually non-local in that the impact of a factor localized in time can be delayed by an arbitrarily long time-lag. For example, in order to complete a melody line, the beginning of the music sequence needs to be held in mind while the rest is played, a task which is carried out by the short-term memory. The long-term memory will serve as the theme and emotion that will help maintain the global coherence of music. The existence of both the short- and long-term memory is vital for generating melodic and coherent music sequences. In such cases standard RNN is prone to drift away from the desired predictions because it forms a conditional distribution based on a limited context. LSTM is a RNN architecture specifically designed to help memorize and retrieve information in sequences better than the standard RNN. LSTM produces many state-of-the-art results in various sequence processing tasks, including speech recognition [Graves *et al.*, 2013], and machine translation [Sutskever *et al.*, 2014].

In modelling polyphonic music, it is obvious that the occurrence of a particular note at a particular time modifies considerably the probability with which other notes may occur at the same time. In other words, notes appearing together in correlated patterns cannot be conveniently described by a normal RNN architecture designed for multi-class classification task, because enumerating all configurations of the variable to predict would be very expensive. This difficulty motivates energy-based models which allow us to express the log-likelihood of a given configuration by an arbitrary energy function, such as the restricted Boltzmann machine (RBM) [Smolensky, 1986].

In this context, we wish to combine the ability of RBM to represent a complicated distribution for each time step, together with a temporal model in sequence. We consider both long-term memory and short-term memory in our design of guide and learning modules, by increasing a bypassing channel from data source filtered by a recurrent LSTM layer and we show that our model increases performance generally.

2 Model

Adding LSTM units to RTRBM is nontrivial, considering that RTRBM's hidden units and visible units are intertwined in inference and learning. The simplest way to circumvent this difficulty is to use bypass connections from LSTM units to the hidden units besides the existing recurrent connections of hidden units.

By this means, there are two channels for temporal information flow, the direct connection (W_R) between the conditional RBM at each time step, and the connection (W_{LR}) from the recurrent LSTM units of previous time steps.(see Fig. 1) The main computation complexity comes from the repeated sampling procedure of RBM learning and replacing some hidden units in RBM with LSTM units actually boosts learning speed. The inference and sampling procedure is roughly the same as in RTRBM while the Backpropagation Through Time (BPTT) procedure is a bit complex.

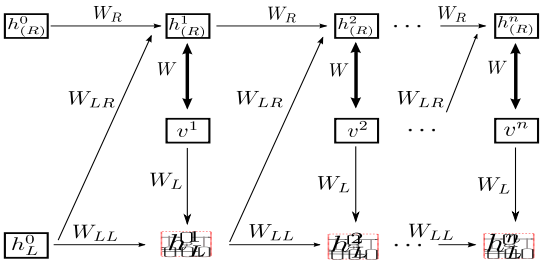

Figure 1: Structure of LSTM-RTRBM

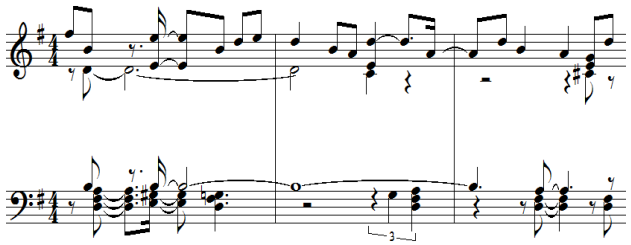

Figure 2: A slice of sample music generated by the proposed model.

3 Experiments

In this section, we show results with the main application of interest: probabilistic modeling of sequences of polyphonic music. We experiment on two datasets with varying complexities: MuseData, an electronic library of orchestral and piano classical music from CCARH 4[1] and JSB chorales, the entire corpus of 382 four-part harmonized chorales by J. S. Bach.

Each dataset contains at least 7 hours of polyphonic music and the total duration is approximately 29 hours. The polyphony (number of simultaneous notes) varies from 0 to 15 and the average polyphony is 4.2. We use a completely general piano-roll representation with an input of 88 binary visible units that span the whole range of piano from A0 to C8 and temporally aligned on an integer fraction of the beat (quarter note). Consequently, pieces with different time signatures will not have their measures start at the same interval. Although it is not strictly necessary, learning is facilitated if the sequences are transposed in a common tonality (e.g. C major/minor) as preprocessing.

We adopt the classic momentum training regime, with learning rate 0.01 and momentum 0.9. The learning start with CD_{10} for the first 1000 weight updates, which then switched to CD_{25}. We use 88 hidden units and 88 LSTM units, the same number as the input and the output dimension, which is trained faster than using hundreds of hidden units in the RTRBM, for the main computation takes place in the CD steps. Quantitatively, the smaller the negative Log-likelihood(LL), the better the result. The results in negative LL is $5.54, 4.72$ for LSTM-RTRBM, $6.35, 6.35$ for RTRBM, $8.13, 8.71$ for RNN, for MuseData, JSB chorales dataset respectively.

We also evaluate our models qualitatively by generating sample sequences (see Fig. 2 for a glimpse). The model has learned the chords (such as sequential D major triads in Fig. 2), local and global temporal coherence, melody lines and generate music that is harmonic and coherent [2]. With the same configuration, LSTM-RTRBM could learn melody lines from both the three datasets while RTRBM generates inconsistent and unpleasant sample sequences. However, all the recurrent temporal model forms a closed loop that have no new incitations from outside, making the long piece of music dull. One way to solve this is with the technique of side-slipping [Coker, 1980], by playing out-of-key to produce a short sensation of surprise in a context deemed too predictable. For future work, we are interested in enhancing LSTM with optimization techniques for better results, and integrating side-slipping mechanism for more variable music generation.

[1] www.musedata.org

[2] Samples can be downloaded at bitbucket, music-samples.

4 Acknowledgements

This work is supported by National High Technology Research and Development Program of China 2015AA016305, National Natural Science Foundation of China 61375027, 61433018, 61322308 and 61332007, Major Program for National Social Science Foundation of China 13, ZD189 and Tsinghua Initiative Scientific Research Program 20121088071.

References

[Allan and Williams, 2005] Moray Allan and Christopher KI Williams. Harmonising chorales by probabilistic inference. *Advances in Neural Information Processing Systems*, 17:25–32, 2005.

[Coker, 1980] Jerry Coker. *The complete method for improvisation*. Studio P/R, 1980.

[Graves *et al.*, 2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, 2013.

[Lavrenko and Pickens, 2003] Victor Lavrenko and Jeremy Pickens. Polyphonic music modeling with random fields. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 120–129. ACM, 2003.

[Rumelhart *et al.*, 1986] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. In *Parallel Dist. Proc.*, pages 318–362. MIT Press, 1986.

[Smolensky, 1986] Paul Smolensky. *Information processing in dynamical systems: Foundations of harmony theory*. Department of Computer Science, University of Colorado, Boulder, 1986.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.