An Energy-Based Recurrent Neural Network for Multiple Fundamental Frequency Estimation

Nicolas Boulanger-Lewandowski, Pascal Vincent and Yoshua Bengio
Université de Montréal, C.P. 6128, Montréal, Canada

1 The RNN-RBM

Many naturally occurring phenomena such as music, speech, or human motion are inherently sequential. Complex sequences are often *non-local* (long-term temporal dependencies) and *high-dimensional* (multi-modal conditional distribution). For the example of polyphonic music, these properties represent the basic components of Western music, namely rhythm and harmony. Here we wish to exploit the recurrent neural network (RNN) internal memory that can in principle represent long-term dependencies, and energy-based models, such as the Restricted Boltzmann Machine (RBM), that allow us to express complex distributions by the means of an energy function.

This combination was first put forward with the so-called Temporal RBM (TRBM) [3], the first such probabilistic model which however uses a heuristic training procedure. The Recurrent TRBM (RTRBM) [4] is a slight modification of the TRBM that allows for exact inference and efficient training by contrastive divergence (CD). The RTRBM can be understood as a sequence of RBMs whose parameters $b_v^{(t)}, b_h^{(t)}, W^{(t)}$ are obtained from the output at time t of a RNN (Figure 1). We consider as in [4] the case where only $b_h^{(t)}$ is variable with $b_h^{(t)} = b_h + W'\hat{h}^{(t-1)}$ where $\hat{h}^{(t)}$ is the mean-field value of $h^{(t)}$.

Here we extend the RTRBM to include a full RNN with its own hidden units as well as those of conditional RBMs at each time-step (Figure 1). This improves the expressive power of the model while preserving the efficiency of the training procedure. The hidden units $\hat{h}^{(t)}$ are now connected to their direct predecessor $\hat{h}^{(t-1)}$ and to $v^{(t)}$ by the relation:

$$\hat{h}^{(t)} = \sigma(W_2 v^{(t)} + W_3 \hat{h}^{(t-1)} + b_{\hat{h}}). \qquad (1)$$

Note that the single-layer RNN-RBM is more than a high-capacity RTRBM since its hidden units, released from their duty to represent the free-energy, can use arbitrary temporal features (W_2).

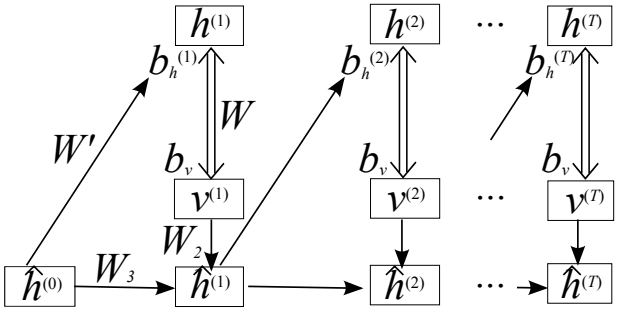

Figure 1: Structure of the RNN-RBM, including the RTRBM as a special case when $W_2 = W$, $W_3 = W'$, $b_{\hat{h}} = b_h$.

We carry out experiments on the same baseline datasets used in [4]. We compare the performance of the single-layer RNN-RBM with the RTRBM at optimal capacity as determined on a separate validation set. We trained all models using 50,000 weight updates of CD_{25} with momentum 0.9. Since reporting the exact log-probability of the test set under both models is impossible, we use the mean square prediction error as a basis of comparison.

The first dataset[1] is a simulation of 2 balls bouncing in a box. The videos produced are of length 128 and of resolution 15×15. The squared prediction error per pixel and time-step saturates at 0.010 for the RTRBM and at 0.005 for the RNN-RBM. The human motion capture dataset[2] consists of 49 real values per time-step so we use Gaussian RBMs. Actual training and evaluation are performed on sub-sequences of length $T = 50$. The optimal-capacity mean squared prediction test error is 0.41 for the RTRBM and is 0.33 for the single-layer RNN-RBM.

[1] www.cs.utoronto.ca/∼ilya/code/2008/RTRBM.tar
[2] http://people.csail.mit.edu/ehsu/work/sig05stf

Table 1: Prediction error, confidence coefficient α and multi-f_0 estimation results (%) for the various models.

Model	Prediction error	α	Total error	Missed error	Subst. error	False alarms	Recall	Precision	Accuracy
(Yeh, 2010)	-	-	34.3	23.3	9.11	1.89	67.6	86.0	66.3
Unigram	0.103	0.02	33.7	24.0	7.85	1.84	68.1	87.5	66.9
RBM	0.060	0.02	31.0	24.3	5.01	1.75	70.7	91.3	69.5
RTRBM	0.040	0.03	30.3	23.4	5.06	1.87	71.6	91.2	70.3
RNN-RBM	0.039	0.03	30.2	23.2	5.17	1.84	71.6	91.1	70.3

2 Musical language models

We apply our algorithm to the real-world task of modeling sequences of polyphonic music using the Mozer dataset [2]. We consider models of *symbolic* sequences typically contained in a MIDI file. Musical models are trained to predict the pattern of notes in the next time interval given the previous ones. While most existing models output only monophonic notes along with predefined chords, our approach uses the RNN-RBM to learn both temporal dependencies and chord conditional distributions. We use an input of 49 binary visible units than span 4 octaves from F#2 to F#6 and temporally aligned on an integer fraction of the beat (quarter note). We compare our results with a simple unigram model (independent notes), a harmonic model (independent RBMs) and the RTRBM.

The Mozer dataset consists of 22 excerpts from classical pieces played with a single polyphonic instrument. The maximum polyphony (number of simultaneous notes) for this dataset is 5 and the average polyphony is 4.4. The mean squared prediction errors of our models for the Mozer dataset are presented in the left-most column of Table 1. Tonality inference is achieved by finding the transposition that maximizes the likelihood under a model trained on pieces of constant tonality. This task is solved perfectly for all the 22 sequences by using cross-validation unigrams obtained by first reporting notes on a single octave.

The multiple fundamental frequency (f_0) estimation task consists of finding the audible note pitches in the signal at 10 ms intervals. See [1] for common evaluation metrics. Most existing algorithms are frame-based and rely exclusively on the audio signal, but *musical language models* can improve purely auditive approaches. We implemented the multi-f_0 algorithm that won the MIREX 2010 contest [5] which works by (i) generating a set of f_0 candidates and (ii) jointly evaluating all combinations of f_0 by a score function. We integrate our symbolic model prediction into step (ii) in the form of an extra term to the score function: $\alpha \log P(v^{(t)}|\text{past})$. This corresponds to a product of experts where α is the confidence coefficient of our symbolic predictor. We feed our symbolic models with the average of the estimated f_0 at the previous time-step ("past"). If our algorithm is run on audio signals without preprocessing, tempo tracking must be performed first.

We report the multi-f_0 estimation results on the synthesized Mozer dataset using our hybrid method in Table 1. Test results are calculated using leave-one-out cross-validation while the hyper-parameter α has been optimized on 10% of each sequence. Our best symbolic model (RNN-RBM) yields an improvement in overall accuracy of 4% over the original state-of-the-art algorithm (first row).

References

[1] M. Bay, A. Ehmann, and J. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *ISMIR*, 2009.

[2] M. Mozer. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, 6(2):247–280, 1994.

[3] I. Sutskever and G. Hinton. Learning multilevel distributed representations for high-dimensional sequences. In *AISTATS*, pages 544–551, 2007.

[4] I. Sutskever, G. Hinton, and G. Taylor. The recurrent temporal restricted boltzmann machine. In *NIPS 20*, pages 1601–1608, 2008.

[5] C. Yeh and A. Roebel. Multiple-f0 estimation for MIREX 2010. In *MIREX*, 2010.

Topic: Sequence modeling
Preference: poster