Modeling Temporal Dependencies in High-Dimensional Sequences
Application to Polyphonic Music Generation and Transcription

Nicolas Boulanger-Lewandowsk, Yoshua Bengio, Pascal Vincent

Presented By
Patrick Gray
Chinmaya Naguri
Recurrent Temporal RBM

Convey temporal dependencies in the hidden units over \( T \) time steps

\( r_t = \begin{cases} 
\sigma(Wv_t + C + Ur_{t-1}), & t > 1 \\
\sigma(Wv_t + C_{\text{init}}), & t = 1 
\end{cases} \)

\( p(h|v) = \sigma(Wv + C) \)

\( \mathbf{W} \)

\( \mathbf{U} \)

\( \mathbf{r} \)

\( \mathbf{v} \)

\( \mathbf{h} \)

\( \mathbf{C} \)

\( \mathbf{C}_{\text{init}} \)

\( (C + Ur_t)h_{t+1} \)

Time step \((t+2)\) is conditionally independent of time step \((t+1)\) given \( r_{t+1} \)
RT-RBM Joint Probability Distribution

- Joint probability is a product of the RBMs at each time step

\[ p(\{v_t, h_t\}_{t=1}^T | \{r_t\}_{t=1}^{T-1}) = \prod_{t=1}^{T} \frac{\exp(-G(v_t, h_t | r_t))}{Z_{r_t}} \]

- The new energy function is written as follows

\[ \text{Energy}(\{v_t, h_t\}_{t=1}^T | \{r_t\}_{t=1}^{T-1}) = -G(\{v_t, h_t\}_{t=1}^T | \{r_t\}_{t=1}^{T-1}) \]

\[ = h_1^T Wv_1 + B^T v_1 + C_{\text{init}}^T h_1 + \sum_{t=2}^{T} (h_t^T Wv_t + B^T v_t + C^T h_t + h_t^T U r_{t-1}) \]

- Let us now split it up and find the gradients

\[ G^{(1)} = h_1^T Wv_1 + B^T v_1 + C_{\text{init}}^T h_1 + \sum_{t=2}^{T} (h_t^T Wv_t + B^T v_t + C^T h_t) \]

\[ G^{(2)} = \sum_{t=2}^{T} (h_t^T U r_{t-1}) \]
Defining the Energy Recursively

\[ G^{(2)} = \sum_{t=2}^{T} (h_t^T Ur_{t-1}) \]

• Define \( G^{(2)} \) in terms of successive time steps

\[ G_t^{(2)} = \sum_{\tau=t}^{T} (h_\tau^T Ur_{\tau-1}) = G_{t+1}^{(2)} + h_t^T Ur_{t-1} \]

• We can now get the derivative started by taking the gradient with respect to \( r_t \) and performing backpropagation through time

• Just remember the chain rule

\[ \nabla_{r_t} G_{t+1}^{(2)} = \nabla_{r_{t+1}} G_{t+2}^{(2)} r_{t+1} \circ (1 - r_{t+1})U + h_{t+1}^T U \quad r_t = \begin{cases} \sigma(Wv_t + C + Ur_{t-1}), & t > 1 \\ \sigma(Wv_t + C_{init}), & t = 1 \end{cases} \]
Parameter Updates

\[
G_t^{(2)} = \sum_{\tau=t}^{T} (h_{\tau}^T U r_{\tau-1}) = G_{t+1}^{(2)} + h_{\tau}^T U r_{\tau-1}
\]

\[
\frac{\partial -\ln p(\theta|v)}{\partial \theta} = \sum_h p(h|v) \left[ \frac{\partial \text{Energy}(v, h)}{\partial \theta} \right] - \sum_{v,h} p(v, h) \left[ \frac{\partial \text{Energy}(v, h)}{\partial \theta} \right]
\]

\[
r_t = \begin{cases} 
\sigma(Wv_t + C + Ur_{t-1}), & t > 1 \\
\sigma(Wv_t + C_{init}), & t = 1
\end{cases}
\]

- \( \nabla_U^{G(2)} = \sum_{t=2}^{T} (D_{t+1} r_t (1 - r_t) + E_{h_t|v_t} r_{t-1} [h_t] - E_{v_t,h_t|r_{t-1}} [h_t]) r_{t-1}^T \)

- \( \nabla_W^{G(2)} = \sum_{t=1}^{T-1} (D_{t+1} r_t (1 - r_t)) v_t^T \)

- \( \nabla_C^{G(2)} = \sum_{t=2}^{T} (D_{t+1} r_t (1 - r_t)) \)

- \( \nabla_{C_{init}}^{G(2)} = D_2 r_1 (1 - r_1) \)

- \( D_t = E_{(h_t,...,h_T|v_t,...,v_T,r_1,...,r_{T-1})} [\nabla_{r_{t-1}} G_t^{(2)}] - E_{(h_t,...,h_T,v_t',...,v_T'|r_1,...,r_{T-1})} [\nabla_{r_{t-1}} G_t^{(2)}] \)

Employ contrastive divergence to find approximated expectations and update the gradients
Inference

• Perform a feed forward pass through the network as if a normal neural network

• Given the recurrent inputs $r_t, \ldots r_{T-1}$, the RBMs are conditionally independent

\[
p(h_{t,i} | v_t, r_{t-1}) = \sigma(W_i v_t + C_i + U_i r_{t-1})
\]

\[
p(v_{t,j} | h_t, r_{t-1}) = \sigma(h^T_t W_j + B_j)
\]
Generating Bouncing Balls

- Video of 3 balls bouncing in a box
- Resolution 30 x 30
- 400 hidden units in RBM
- Evaluation metric is qualitative since computing the log probability on a test set is infeasible

RT-RBM

T-RBM
Figure 3: This figure shows the receptive fields of the first 36 hidden units of the RTRBM on the left, and the corresponding hidden-to-hidden weights between these units on the right: the $i$th row on the right corresponds to the $i$th receptive field on the left, when counted left-to-right. Hidden units 18 and 19 exhibit unusually strong hidden-to-hidden connections; they are also the ones with the weakest visible-hidden connections, which effectively makes them belong to another hidden layer.
Combine full RNN with RT-RBM to convey temporal information in distinct hidden units.

\[ r_t = \begin{cases} 
\sigma(W'v_t + D + U_r r_{t-1}), & t > 1 \\
\sigma(W'v_t + D_{init}), & t = 1 
\end{cases} \]
RNN-RBM Joint Probability Distribution

- Joint probability is a product of the RBMs at each time step

\[
p(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^{T}|\{\mathbf{r}_t\}_{t=1}^{T-1}) = \prod_{t=1}^{T} \frac{\exp(-G(\mathbf{v}_t, \mathbf{h}_t|\mathbf{r}_t))}{Z_{r_t}}
\]

- The new energy function is written as follows

\[
-\text{Energy}(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^{T}|\{\mathbf{r}_t\}_{t=1}^{T-1}) = G(\{\mathbf{v}_t, \mathbf{h}_t\}_{t=1}^{T}|\{\mathbf{r}_t\}_{t=1}^{T-1})
\]

\[
= \mathbf{h}_1^T \mathbf{W}_1 \mathbf{v}_1 + \mathbf{B}_{init}^T \mathbf{v}_1 + \mathbf{C}_{init}^T \mathbf{h}_1 + \sum_{t=2}^{T} (\mathbf{h}_t^T \mathbf{W} \mathbf{v}_t + \mathbf{B}^T \mathbf{v}_t + \mathbf{C}^T \mathbf{h}_t + \mathbf{v}_t^T \mathbf{U}_v \mathbf{r}_{t-1} + \mathbf{h}_t^T \mathbf{U}_h \mathbf{r}_{t-1})
\]

- Let us now split it up and find the gradients

\[
G^{(1)} = \mathbf{h}_1^T \mathbf{W}_1 \mathbf{v}_1 + \mathbf{B}_{init}^T \mathbf{v}_1 + \mathbf{C}_{init}^T \mathbf{h}_1 + \sum_{t=2}^{T} (\mathbf{h}_t^T \mathbf{W} \mathbf{v}_t + \mathbf{B}^T \mathbf{v}_t + \mathbf{C}^T \mathbf{h}_t)
\]

\[
G^{(2)} = \sum_{t=2}^{T} (\mathbf{v}_t^T \mathbf{U}_v \mathbf{r}_{t-1} + \mathbf{h}_t^T \mathbf{U}_h \mathbf{r}_{t-1})
\]
Defining the Energy Recursively

\[ G^{(2)} = \sum_{t=2}^{T} (v^T_t U_v r_{t-1} + h^T_t U_h r_{t-1}) \]

- Define \( G^{(2)} \) in terms of successive time steps

\[ G^{(2)}_t = \sum_{\tau=t}^{T} (v^T_\tau U_v r_{\tau-1} + h^T_\tau U_h r_{\tau-1}) = G^{(2)}_{t+1} + v^T_t U_v r_{t-1} + h^T_t U_h r_{t-1} \]

- We can now get the derivative started by taking the gradient with respect to \( r_t \) and performing backpropagation through time
- Just remember the chain rule

\[ \nabla_{r_t} G^{(2)}_{t+1} = \nabla_{r_{t+1}} G^{(2)}_{t+2} \circ r_{t+1} \circ (1 - r_{t+1})U_r + v^T_{t+1} U_v + h^T_{t+1} U_h \]

\[ r_t = \begin{cases} \sigma(W'v_t + D + U_r r_{t-1}), & t > 1 \\ \sigma(W'v_t + D_{init}), & t = 1 \end{cases} \]
Parameter Updates

\[ G^{(2)}_t = \sum_{\tau = t}^{T} (v^T \Upsilon_{\tau, \tau-1} + h^T \Upsilon_{\tau, \tau-1}) = G^{(2)}_{t+1} + v^T \Upsilon_{\tau, \tau-1} + h^T \Upsilon_{\tau, \tau-1} \]

- \[ \nabla_{W'} G^{(2)}_t = \sum_{t=1}^{T-1} (D_{t+1} \circ r_t \circ (1 - r_t))^T v^T t \]
- \[ \nabla_{U_{h}} G^{(2)}_t = \sum_{t=2}^{T} (E_{h_t}v_t, r_{t-1}[h_t] - E_{v_t, h_t}r_{t-1}[h_t]) r^T_{t-1} \]
- \[ \nabla_{D} G^{(2)}_t = \sum_{t=2}^{T} (D_{t+1} \circ r_t \circ (1 - r_t)) \]
- \[ \nabla_{B} G^{(2)}_t = 0 \]
- \[ \nabla_{C} G^{(2)}_t = 0 \]
- \[ \nabla_{W} G^{(2)}_t = 0 \]

\[ r_t = \begin{cases} \sigma(W'v_t + D + U_{r, r_{t-1}}), & t > 1 \\ \sigma(W'v_t + D_{\text{init}}), & t = 1 \end{cases} \]

- \[ D_t = E_{(h_t, \ldots, h_T | v_t, \ldots, v_T, r_{1, \ldots, r_{T-1}})}[\nabla_{t-1} G^{(2)}_t] - E_{(h_t, \ldots, h_T, v'_t, \ldots, v'_T | r_{1, \ldots, r_{T-1}})}[\nabla_{t-1} G^{(2)}_t] \]

Employ contrastive divergence to find approximated expectations and update the gradients
RT-RBM VS RNN-RBM Baseline Experiments

- Bouncing Balls
  - Video of 3 balls bouncing in a box
  - Resolution 15 x 15
  - 300 hidden units in RBM
  - 50 steps of Gibbs sampling
  - Mean frame-level squared prediction error
    - RT-RBM – 2.11 MSE
    - RNN-RBM – 0.96 MSE

![W](image-url)

Figure 3. Receptive fields of 48 hidden units of an RNN-RBM trained on the bouncing balls dataset. Each square shows the input weights of a hidden unit as an image.
RT-RBM VS RNN-RBM Baseline Experiments

- Human Motion Capture
  - Sequence of joint angles, translations, and rotations of the base of the spine
  - 450 hidden units in RBM
  - Mean frame-level squared prediction error
    - RT-RBM  – 20.1 MSE
    - RNN-RBM  – 16.2 MSE
Polyphonic Music Transcription

• Create perceptually independent streams of music (poly-phonic = many sounds)
• Make it sound beautiful

Excerpt from The Well-Tempered Clavier, Fugue 1 by Johann Bach

• Need to design a musical language model
  • Similar to natural language models
Difficulties in Polyphonic Music Transcription

- The occurrence of a particular note at a particular time modifies considerably the probability with which other notes may occur at the same time.
- Notes appear together in correlated patterns, or simultaneities.
- Need to consider both harmony and melody.

\[
p(c_6|c_1, c_2, c_3, c_4, c_5)
\]
Solution: RNN-RBM

• Benefit of capturing both chordal and temporal dependencies
Data

• Symbolic music of varying complexity
  
  • **Piano-midi.de** is a classical piano MIDI archive that was split according to Poliner & Ellis
  • **Nottingham** is a collection of 1200 folk tunes with chords instantiated from the ABC format
  • **MuseData** is an electronic library of orchestral and piano classical music from Center for Computer Assisted Research in the Humanities
  • **JSB chorales** refers to the entire corpus of 382 four part harmonized chorales by J. S. Bach with the split of Allan & Williams
  
• Each dataset contains at least 7 hours of polyphonic music and the total duration is approximately 67 hours
Preprocessing and Features

- Utilize input vector of 88 binary visible units that span the whole range of piano from A0 to C8
- Temporally aligned on an integer fraction of the beat (quarter note)
- Notes are transposed to a common tonality (e.g. C major/minor)

\[
\begin{array}{c|c}
A_0 & 0 \\
A_0^\# & 0 \\
\vdots & \\
C_4 & 1 \\
C_4^\# & 0 \\
D_4 & 1 \\
D_4^\# & 0 \\
\vdots & \\
C_8 & 0 \\
\end{array}
\]

Transpose from G-major to C-major
The log-likelihood (LL) and expected frame-level accuracy (ACC)

Table 1. Log-likelihood and expected accuracy for various musical models in the symbolic prediction task. The double line separates frame-level models (above) and models with a temporal component (below).

<table>
<thead>
<tr>
<th>Model</th>
<th>Piano-Midi.de LL</th>
<th>ACC %</th>
<th>Nottingham LL</th>
<th>ACC %</th>
<th>MuseData LL</th>
<th>ACC %</th>
<th>JSB Chorales LL</th>
<th>ACC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>-61.00</td>
<td>3.35</td>
<td>-61.00</td>
<td>4.53</td>
<td>-61.00</td>
<td>3.74</td>
<td>-61.00</td>
<td>4.42</td>
</tr>
<tr>
<td>1-Gram (Add-p)</td>
<td>-27.64</td>
<td>4.85</td>
<td>-5.94</td>
<td>22.76</td>
<td>-19.03</td>
<td>6.67</td>
<td>-12.22</td>
<td>16.80</td>
</tr>
<tr>
<td>1-Gram (Gaussian)</td>
<td>-10.79</td>
<td>6.04</td>
<td>-5.30</td>
<td>21.31</td>
<td>-10.15</td>
<td>7.87</td>
<td>-7.56</td>
<td>17.41</td>
</tr>
<tr>
<td>Note 1-Gram</td>
<td>-11.05</td>
<td>5.80</td>
<td>-10.25</td>
<td>19.87</td>
<td>-11.51</td>
<td>7.72</td>
<td>-11.06</td>
<td>15.25</td>
</tr>
<tr>
<td>Note 1-Gram (IID)</td>
<td>-12.90</td>
<td>2.51</td>
<td>-16.24</td>
<td>3.56</td>
<td>-14.06</td>
<td>2.82</td>
<td>-15.93</td>
<td>3.51</td>
</tr>
<tr>
<td>GMM</td>
<td>-15.84</td>
<td>5.08</td>
<td>-7.87</td>
<td>22.62</td>
<td>-12.20</td>
<td>7.37</td>
<td>-11.90</td>
<td>15.84</td>
</tr>
<tr>
<td>RBM</td>
<td>-10.17</td>
<td>5.63</td>
<td>-5.25</td>
<td>5.81</td>
<td>-9.56</td>
<td>8.19</td>
<td>-7.43</td>
<td>4.47</td>
</tr>
<tr>
<td>N-AD</td>
<td>-10.28</td>
<td>5.82</td>
<td>-5.48</td>
<td>22.67</td>
<td>-10.06</td>
<td>7.65</td>
<td>-7.19</td>
<td>17.88</td>
</tr>
</tbody>
</table>

| Previous + Gaussian         | -12.48           | 25.50 | -8.41         | 55.69 | -12.90      | 25.93 | -19.00          | 18.36 |
| N-Gram (Add-p)              | -46.04           | 7.42  | -6.50         | 63.45 | -35.22      | 10.47 | -29.98          | 24.20 |
| N-Gram (Gaussian)           | -12.22           | 10.01 | -3.16         | 65.07 | -10.59      | 16.15 | -9.74           | 28.79 |
| Note N-Gram                 | -7.50            | 26.80 | -4.51         | 62.49 | -7.91       | 26.35 | -10.26          | 20.34 |
| (Allan & Williams, 2005)    | -                  | -     | -             | -     | -           | -     | -               | -     |
| (Lavrenko & Pickens, 2003)  | -9.05            | 18.37 | -5.44         | 55.34 | -9.87       | 18.39 | -8.78           | 22.93 |
| MLP                         | -8.13            | 20.29 | -4.38         | 63.46 | -7.94       | 25.68 | -8.70           | 30.41 |
| RNN (HF)                    | -7.66            | 23.34 | -3.89         | 66.64 | -7.19       | 30.49 | -8.58           | 29.41 |
| RTRBM                       | -7.36            | 22.99 | -2.62         | 75.01 | -6.35       | 30.85 | -6.35           | 30.17 |
| RNN-RBM                     | -7.09            | 28.92 | -2.39         | 75.40 | -6.01       | 34.02 | -6.27           | 33.12 |
| RNN-NADE                    | -7.48            | 20.69 | -2.91         | 64.05 | -6.74       | 24.91 | -5.83           | 32.11 |
| RNN-NADE (HF)               | -7.05            | 23.42 | -2.31         | 71.50 | -5.60       | 32.60 | -5.56           | 32.50 |

- Estimated the partition function of each conditional RBM by 100 runs of annealed importance sampling
Qualitative Evaluation

- Generation of sample sequences

RBM
  - Frame based

RNN
  - Temporal dependencies captured
  - Note by note generation

RNN-RBM
  - Temporal and chordal dependencies captured
Visualizing the Results

- Mean field samples $p(v|\mathbf{h}^*)$
- $\mathbf{h}^* \sim p(\mathbf{h})$

*Figure 1. Mean-field samples of an RBM trained on the Piano-midi (top) and JSB chorales (bottom) datasets. Each column is a sample vector of notes, with a chord label where the analysis is unambiguous.*
Polyphonic Music Transcription of Audio Signals

- Determine the underlying notes of a polyphonic audio signal without access to its score
- Most existing transcription algorithms are frame-based and rely exclusively on the audio signal.

- Want to support a frame-based, state of the art transcription algorithm from Nam et al.
Acoustic Model Support Breakdown

- Acoustic Model Format: $P_a(v_t)$
  - Outputs independent probabilities that each note in $v_t$ is present
  - Reports the notes with $P \geq 0.5$
  - Estimates the audible note pitches in a signal at 10 ms intervals

- Incorporation of Symbolic Model Prediction: $P_s(v_t|A_t)$
  - $A_t$ denotes the sequence history
  - Consider the $k$ most promising note estimates ($k = 7$) from the acoustic model
  - Jointly evaluate all combinations of notes (power set of $k$ notes)

- Evaluation Cost Function
  - $C = -\log P_a(v_t) - \alpha \log P_s(v_t|A_t^\sim)$
    - $\alpha$ is the confidence coefficient
    - $A_t^\sim$ is approximate sequence history constructed from the notes estimated so far in at least half the audio frames corresponding to each past symbolic time step
Results

Figure 5. Frame-level transcription accuracy of the Nam et al. (2011) model either alone, after HMM smoothing or with our best performing model as a symbolic prior.
Questions?
References

