Modeling Temporal Dependencies in High-Dimensional Sequences

Application to Polyphonic Music Generation and Transcription

Nicolas Boulanger-Lewandowsk, Yoshua Bengio, Pascal Vincent

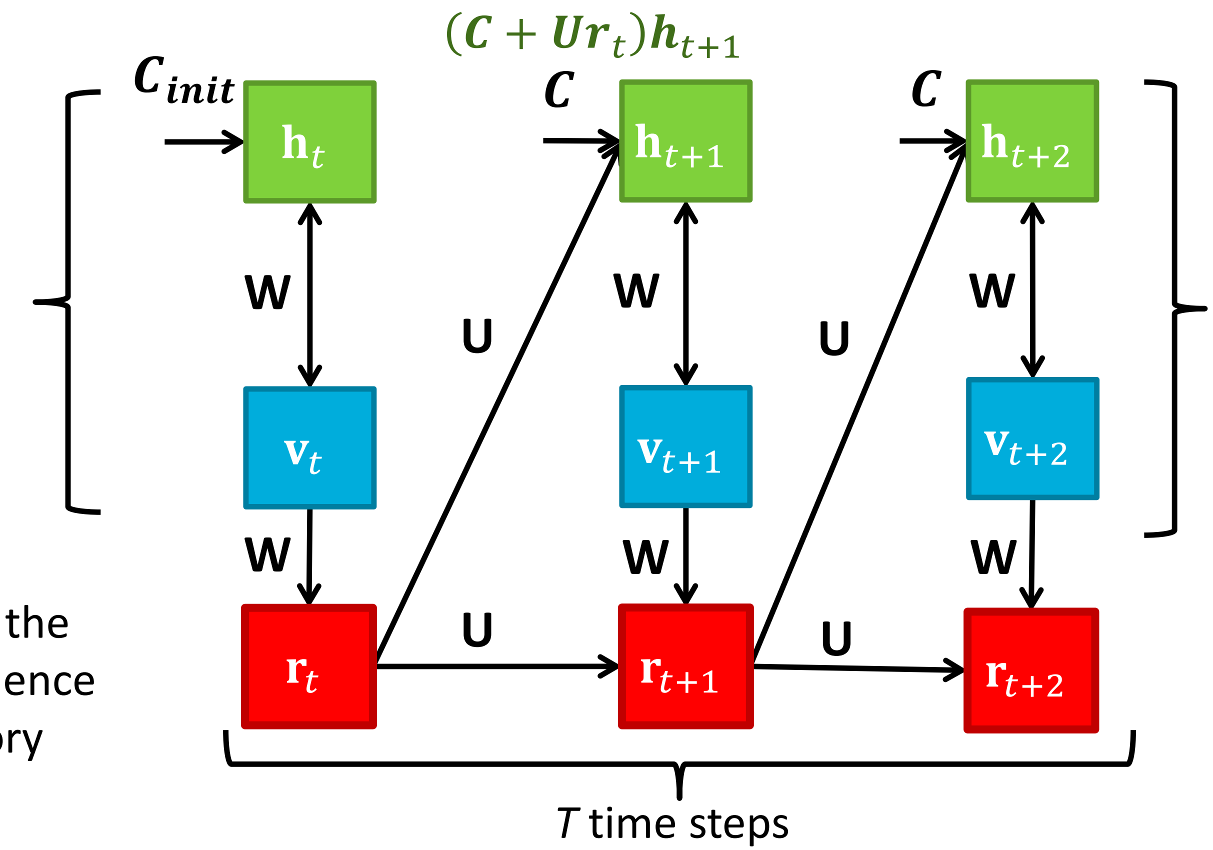
Presented By

Patrick Gray

Chinmaya Naguri

Recurrent Temporal RBM Sutskever et al. []

Standard RBM placed at each time step



Time step *(t+2)* is conditionally independent of time step *(t+1)* given r_{t+1}

$\boldsymbol{r}_t$ is the sequence history

- Convey temporal dependencies in the hidden units over *T* time steps

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C} + \boldsymbol{U}\boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C}_{init}), & t = 1 \end{cases} \qquad p(\boldsymbol{h}|\boldsymbol{v}) = \sigma(\boldsymbol{W}\boldsymbol{v} + \boldsymbol{C})$$

RT-RBM Joint Probability Distribution

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

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

- The new energy function is written as follows

$$-Energy(\{\boldsymbol{v}_t, \boldsymbol{h}_t\}_{t=1}^{T} | \{\boldsymbol{r}_t\}_{t=1}^{T-1}) = -G(\{\boldsymbol{v}_t, \boldsymbol{h}_t\}_{t=1}^{T} | \{\boldsymbol{r}_t\}_{t=1}^{T-1})$$

$$= \boldsymbol{h}_1^{\mathrm{T}} W \boldsymbol{v}_1 + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{v}_1 + \boldsymbol{C}_{init}^{\mathrm{T}} \mathbf{h}_1 + \sum_{t=2}^{T} (\boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{W} \boldsymbol{v}_t + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{v}_t + \boldsymbol{C}^{\mathrm{T}} \boldsymbol{h}_t + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{t-1})$$

- Let us now split it up and find the gradients

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

$$G^{(2)} = \sum_{t=2}^{T} (\boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{t-1})$$

Defining the Energy Recursively

$$G^{(2)} = \sum_{t=2}^{T} (\boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{t-1})$$

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

$$G_t^{(2)} = \sum_{\tau=t}^{T} (\boldsymbol{h}_\tau^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{\tau-1}) = \mathrm{G}_{\mathrm{t}+1}^{(2)} + \boldsymbol{h}_{\mathrm{t}}^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{\mathrm{t}-1}$$

- 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

$$\nabla_{\boldsymbol{r}_t} G_{t+1}^{(2)} = \nabla_{\boldsymbol{r}_{t+1}} G_{t+2}^{(2)} \circ \boldsymbol{r}_{t+1} \circ (1 - \boldsymbol{r}_{t+1}) \boldsymbol{U} + \boldsymbol{h}_{t+1}^{\mathrm{T}} \boldsymbol{U}$$

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C} + \boldsymbol{U}\boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C}_{init}), & t = 1 \end{cases}$$

Parameter Updates

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

$$G_t^{(2)} = \sum_{\tau=t}^{T} (\boldsymbol{h}_\tau^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{\tau-1}) = \mathrm{G}_{t+1}^{(2)} + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U} \boldsymbol{r}_{t-1}$$

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C} + \boldsymbol{U}\boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}\boldsymbol{v}_t + \boldsymbol{C}_{init}), & t = 1 \end{cases}$$

- $\nabla_U^{G^{(2)}} = \sum_{t=2}^{T} (D_{t+1} \circ r_t \circ (1 - r_t) + E_{h_t | v_t, r_{t-1}}[h_t] - E_{v_t', h_t | r_{t-1}}[h_t])\, r_{t-1}^{\mathrm{T}}$

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

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

- $\nabla_B^{G^{(2)}} = 0$

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

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

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

Inference

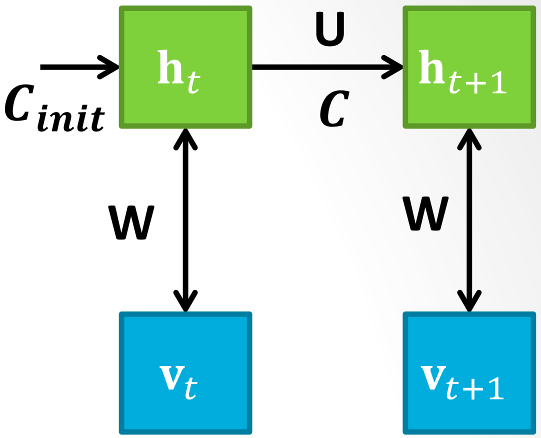
- Perform a feed forward pass through the network as if a normal neural network

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

- $p\left(h_{t,i} \middle| \boldsymbol{v}_t, \boldsymbol{r}_{t-1}\right) = \sigma(\boldsymbol{W}_i \boldsymbol{v}_t + C_i + \boldsymbol{U_i} \boldsymbol{r}_{t-1})$

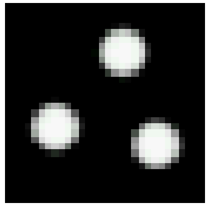
- $p\left(v_{t,j} \middle| \boldsymbol{h}_t, \boldsymbol{r}_{t-1}\right) = \sigma(\boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{W}_j + B_j)$

Generating Bouncing Balls

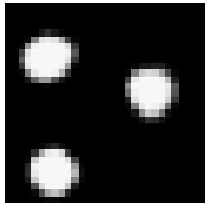


T-RBM

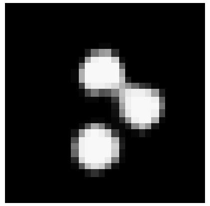
- Video of 3 balls bouncing in a box
- Resolution 30 x 30
- 400 hidden units in RBM
- Evaluation metric is qualitative since computing the log probability on a test set is infeasible



Training Sequence



RT-RBM



T-RBM

Learned Features

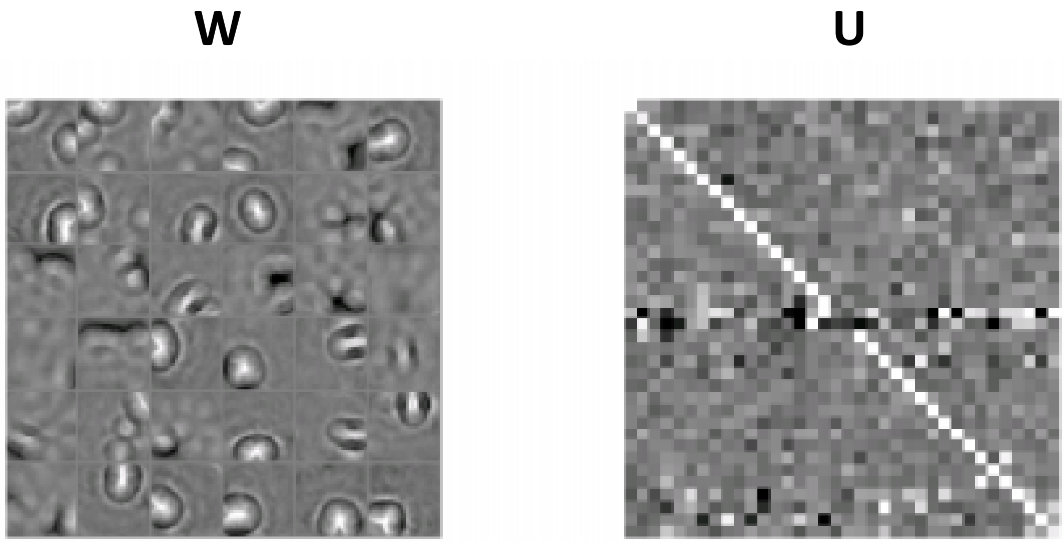
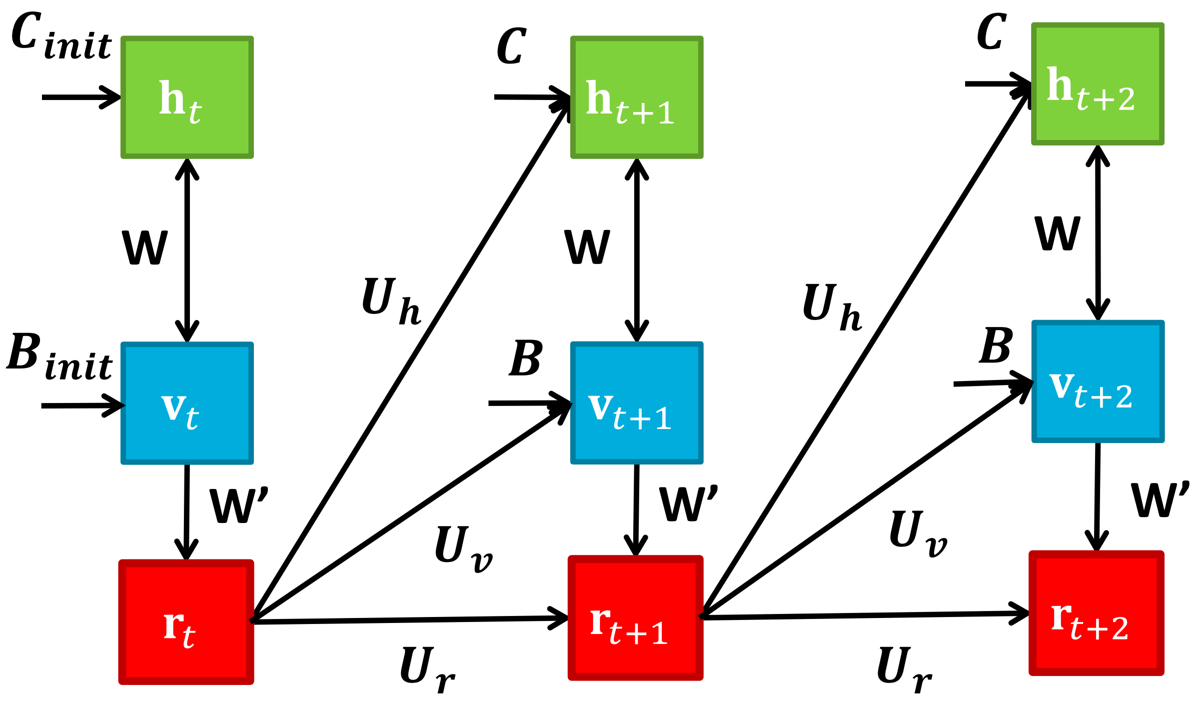


Figure 3: This figure shows the receptive fields of the first 36 hidden units of the RTRBM on the left, and the corresponding hidden-to-hidden weights between these units on the right: the ith row on the right corresponds to the ith receptive field on the left, when counted left-to-right. Hidden units 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.

Recurrent Neural Network RBM

Boulanger-Lewandowski et al. []



- Combine full RNN with RT-RBM to convey temporal information in distinct hidden units

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}'\boldsymbol{v}_t + \boldsymbol{D} + \boldsymbol{U}_r\boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}'\boldsymbol{v}_t + \boldsymbol{D}_{init}), & t = 1 \end{cases}$$

RNN-RBM Joint Probability Distribution

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

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

- The new energy function is written as follows

$$-Energy(\{\boldsymbol{v}_t, \boldsymbol{h}_t\}_{t=1}^{T} | \{\boldsymbol{r}_t\}_{t=1}^{T-1}) = G(\{\boldsymbol{v}_t, \boldsymbol{h}_t\}_{t=1}^{T} | \{\boldsymbol{r}_t\}_{t=1}^{T-1})$$

$$= \boldsymbol{h}_1^{\mathrm{T}} W \boldsymbol{v}_1 + \boldsymbol{B}_{\boldsymbol{init}}^{\mathrm{T}} \boldsymbol{v}_1 + \boldsymbol{C}_{init}^{\mathrm{T}} \mathbf{h}_1$$

$$+ \sum_{t=2}^{T} (\boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{W} \boldsymbol{v}_t + \boldsymbol{B}^{\mathrm{T}} \boldsymbol{v}_t + \boldsymbol{C}^{\mathrm{T}} \boldsymbol{h}_t + \boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{t-1} + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{r}_{t-1})$$

- Let us now split it up and find the gradients

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

$$G^{(2)} = \sum_{t=2}^{T} (\boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{t-1} + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{r}_{t-1})$$

Defining the Energy Recursively

$$G^{(2)} = \sum_{t=2}^{T} (\boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{t-1} + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{r}_{t-1})$$

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

$$G_t^{(2)} = \sum_{\tau=t}^{T} (\boldsymbol{v}_\tau^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{\tau-1} + \boldsymbol{h}_\tau^{\mathrm{T}} \boldsymbol{U}_h \boldsymbol{r}_{\tau-1}) = \mathrm{G}_{\mathrm{t}+1}^{(2)} + \boldsymbol{v}_{\mathrm{t}}^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{\mathrm{t}-1} + \boldsymbol{h}_{\mathrm{t}}^{\mathrm{T}} \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{r}_{\mathrm{t}-1}$$

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

$$\nabla_{\boldsymbol{r}_t} G_{t+1}^{(2)} = \nabla_{\boldsymbol{r}_{t+1}} G_{t+2}^{(2)} \circ \boldsymbol{r}_{t+1} \circ (1 - \boldsymbol{r}_{t+1}) \boldsymbol{U}_r + \boldsymbol{v}_{t+1}^{\mathrm{T}} \boldsymbol{U}_v + \boldsymbol{h}_{t+1}^{\mathrm{T}} \boldsymbol{U}_h$$

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}' \boldsymbol{v}_t + \boldsymbol{D} + \boldsymbol{U}_r \boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}' \boldsymbol{v}_t + \boldsymbol{D}_{init}), & t = 1 \end{cases}$$

Parameter Updates

$$G_t^{(2)} = \sum_{\tau=t}^{T} (\boldsymbol{v}_\tau^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{\tau-1} + \boldsymbol{h}_\tau^{\mathrm{T}} \boldsymbol{U}_h \boldsymbol{r}_{\tau-1}) = \mathrm{G}_{t+1}^{(2)} + \boldsymbol{v}_t^{\mathrm{T}} \boldsymbol{U}_v \boldsymbol{r}_{t-1} + \boldsymbol{h}_t^{\mathrm{T}} \boldsymbol{U}_h \boldsymbol{r}_{t-1}$$

$$\boldsymbol{r}_t = \begin{cases} \sigma(\boldsymbol{W}'\boldsymbol{v}_t + \boldsymbol{D} + \boldsymbol{U}_r \boldsymbol{r}_{t-1}), & t > 1 \\ \sigma(\boldsymbol{W}'\boldsymbol{v}_t + \boldsymbol{D}_{init}), & t = 1 \end{cases}$$

- $\nabla_{W'}^{G^{(2)}} = \sum_{t=1}^{T-1} (D_{t+1} \circ r_t \circ (1 - r_t))\, v_t^T$
- $\nabla_{U_r}^{G^{(2)}} = \sum_{t=2}^{T} (D_{t+1} \circ r_t \circ (1 - r_t))\, r_{t-1}^{\mathrm{T}}$
- $\nabla_{U_h}^{G^{(2)}} = \sum_{t=2}^{T} (E_{h_t | v_t, r_{t-1}}[h_t] - E_{v_t', h_t | r_{t-1}}[h_t])\, r_{t-1}^{\mathrm{T}}$
- $\nabla_{U_v}^{G^{(2)}} = \sum_{t=2}^{T} \boldsymbol{r}_{t-1}^{\mathrm{T}} \boldsymbol{v}_t$
- $\nabla_{D}^{G^{(2)}} = \sum_{t=2}^{T} (D_{t+1} \circ r_t \circ (1 - r_t))$
- $\nabla_{D_{init}}^{G^{(2)}} = D_2 \circ r_1 \circ (1 - r_1)$
- $\nabla_{B}^{G^{(2)}} = 0$
- $\nabla_{C}^{G^{(2)}} = 0$
- $\nabla_{W}^{G^{(2)}} = 0$
- $D_t = E_{(h_t,\ldots,h_T | v_t,\ldots,v_T, r_1,\ldots,r_{T-1})}[\nabla_{r_{t-1}} G_t^{(2)}]$ - $E_{(h_t,\ldots,h_T, v_t',\ldots,v_T' | r_1,\ldots,r_{T-1})}[\nabla_{r_{t-1}} G_t^{(2)}]$

- 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

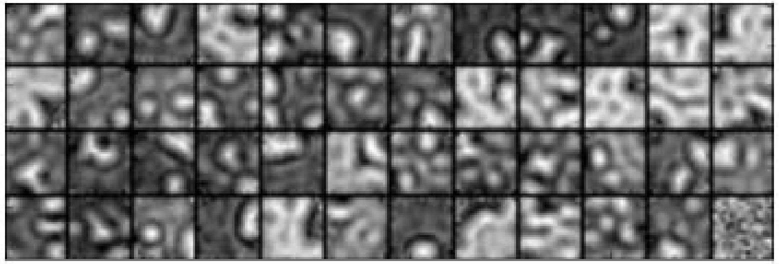


Figure 3. Receptive fields of 48 hidden units of an RNN-RBM trained on the bouncing balls dataset. Each square shows the input weights of a hidden unit as an image.

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

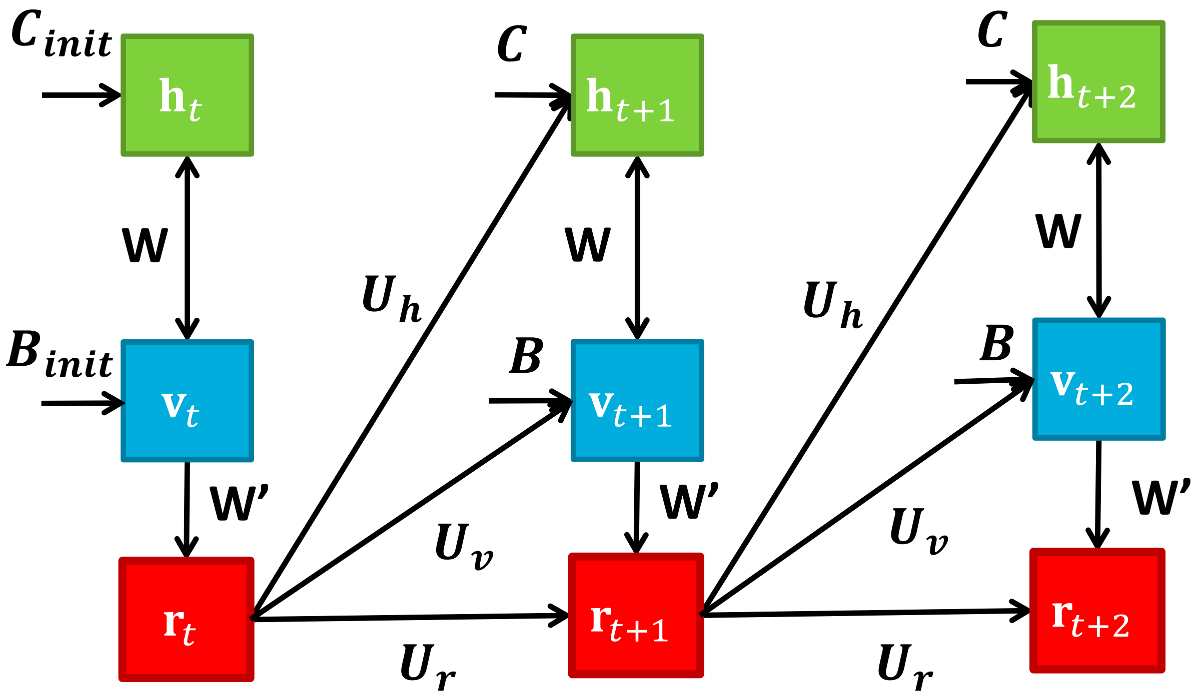
- The occurrence of a particular note at a particular time modifies considerably the probability with which other notes may occur at the same time
- Notes appear together in correlated patterns, or simultaneities
- Need to consider both harmony and melody

c_1 c_2 c_2 c_3 c_4 c_5 c_6

$$p(c_6 | \, c_1, c_2, c_3, c_4, c_5)$$

Solution: RNN-RBM

- Benefit of capturing both chordal and temporal dependencies



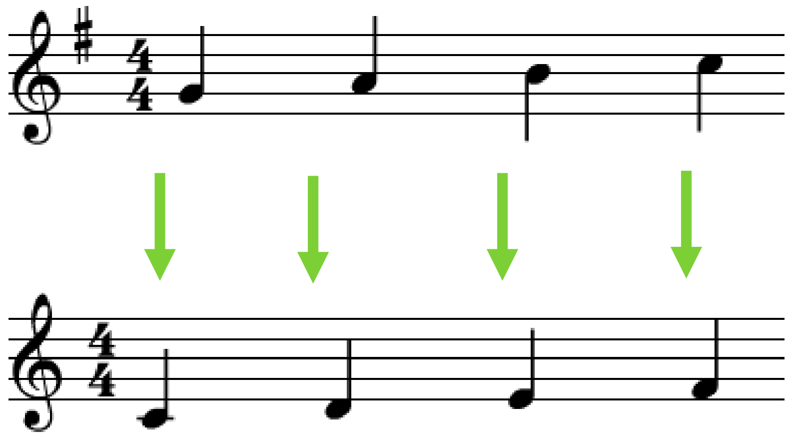
Data

- 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)

$$
\begin{array}{cc}
A_0 & 0 \\
A_0^{\#} & 0 \\
 & \cdots \\
C_4 & 1 \\
C_4^{\#} & 0 \\
D_4 & 1 \\
D_4^{\#} & 0 \\
 & \cdots \\
C_8 & 0
\end{array}
$$



Transpose
from G-major
to C-major

The log-likelihood (LL) and expected frame-level accuracy (ACC)

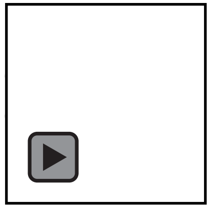
Table 1. Log-likelihood and expected accuracy for various musical models in the symbolic prediction task. The double line separates frame-level models (above) and models with a temporal component (below).

Model	Piano-midi.de		Nottingham		MuseData		JSB chorales	
	LL	ACC %	LL	ACC %	LL	ACC %	LL	ACC %
Random	-61.00	3.35	-61.00	4.53	-61.00	3.74	-61.00	4.42
1-Gram (Add-p)	-27.64	4.85	-5.94	22.76	-19.03	6.67	-12.22	16.80
1-Gram (Gaussian)	-10.79	6.04	-5.30	21.31	-10.15	7.87	-7.56	17.41
Note 1-Gram	-11.05	5.80	-10.25	19.87	-11.51	7.72	-11.06	15.25
Note 1-Gram (IID)	-12.90	2.51	-16.24	3.56	-14.06	2.82	-15.93	3.51
GMM	-15.84	5.08	-7.87	22.62	-12.20	7.37	-11.90	15.84
RBM	-10.17	5.63	-5.25	5.81	-9.56	8.19	-7.43	4.47
NADE	-10.28	5.82	-5.48	22.67	-10.06	7.65	-7.19	17.88
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	65.97	-10.59	16.15	-9.74	28.79
Note N-Gram	-7.50	26.80	-4.54	62.49	-7.91	26.35	-10.26	20.34
GMM + HMM	-15.30	7.91	-6.17	59.27	-11.17	13.93	-11.89	19.24
(Allan & Williams, 2005)	–	–	–	–	–	–	-9.24	16.32
(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	-8.37	19.33	-4.46	62.93	-8.13	23.25	-8.71	28.46
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	64.95	-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

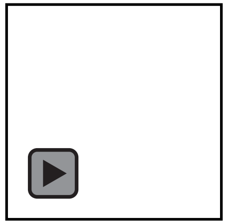
- Estimated the partition function of each conditional RBM by 100 runs of annealed importance sampling

Qualitative Evaluation

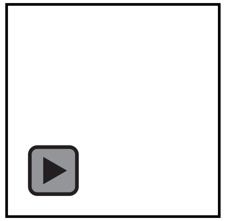
- Generation of sample sequences



RBM



RNN



RNN-RBM

- RBM
  - Frame based
- RNN
  - Temporal dependencies captured
  - Note by note generation
- RNN-RBM
  - Temporal and chordal dependencies captured

Visualizing the Results

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

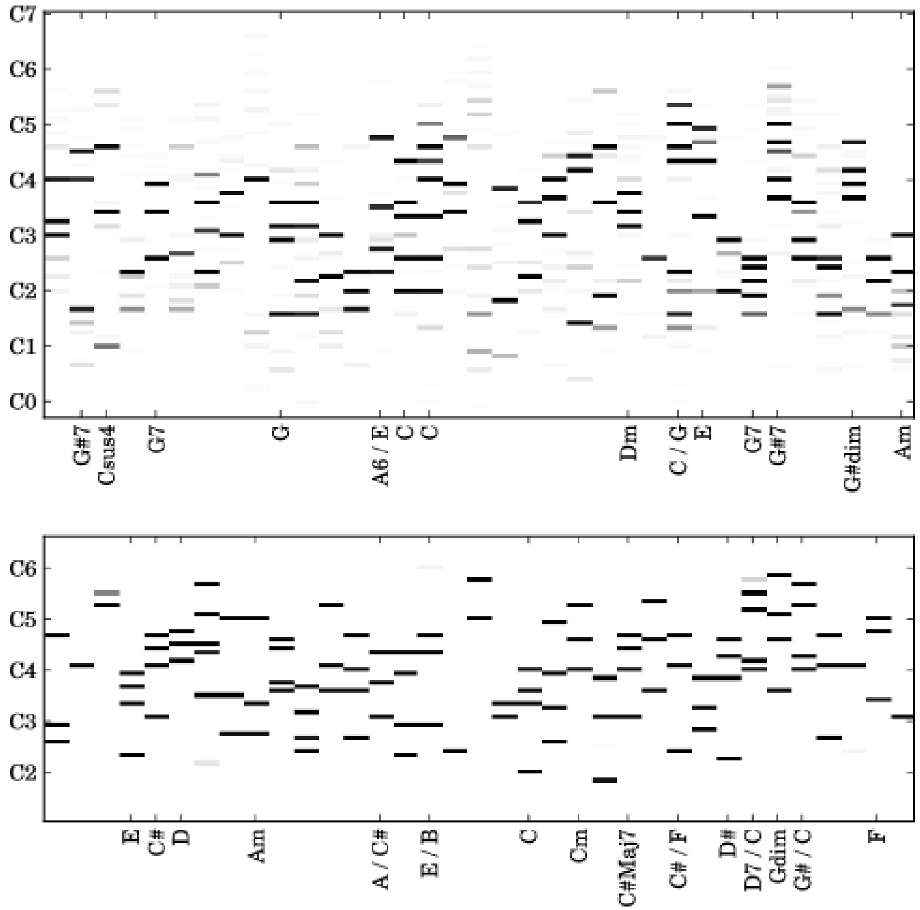
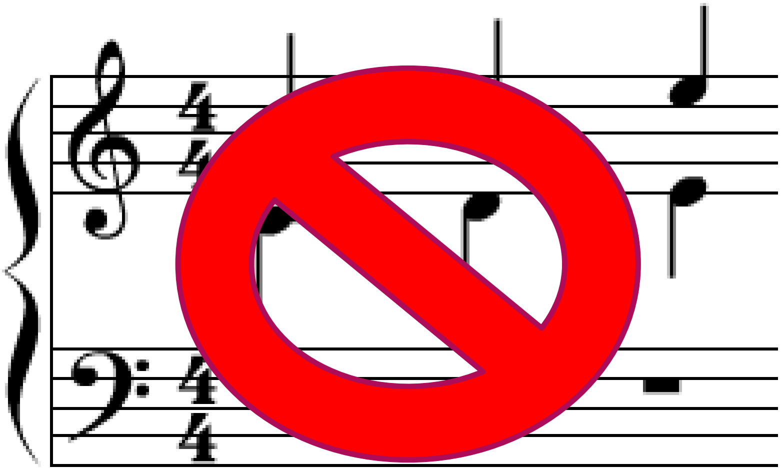
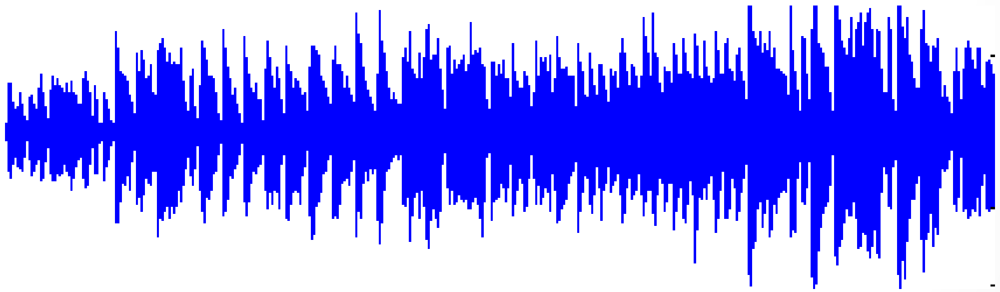


Figure 1. Mean-field samples of an RBM trained on the Piano-midi (top) and JSB chorales (bottom) datasets. Each column is a sample vector of notes, with a chord label where the analysis is unambiguous.

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.





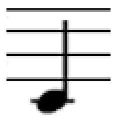
- Want to support a frame-based, state of the art transcription algorithm from Nam et al.

Acoustic Model Support Breakdown

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

- Incorporation of Symbolic Model Prediction: $P_s(\boldsymbol{v}_t|\boldsymbol{A}_t)$
  - $\boldsymbol{A}_t$ 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
  - $C = -\log P_a(\boldsymbol{v}_t) - \alpha \log P_s(\boldsymbol{v}_t|\boldsymbol{A}_t^{\sim})$
    - α is the confidence coefficient
    - $\boldsymbol{A}_t^{\sim}$ 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



Quarter Note
500 ms delay

Results

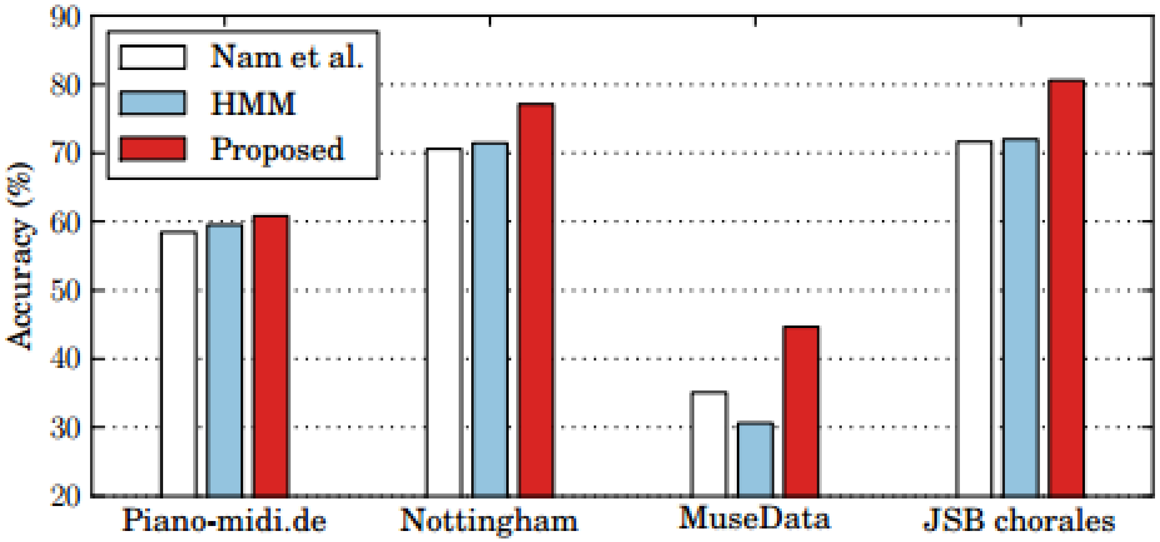


Figure 5. Frame-level transcription accuracy of the Nam et al. (2011) model either alone, after HMM smoothing or with our best performing model as a symbolic prior.

Questions?

References

[1] Sutskever, I., Hinton, G. E., and Graham, T. W. The recurrent temporal restricted Boltzmann machine. In NIPS, 2008.

[2] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee. Structured recurrent temporal restricted Boltzmann machines. In ICML, 2014.

[3] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In Proceedings of the Twenty-nine International Conference on Machine Learning (ICML'12), 2012.