Neural Probabilistic Models for Melody Prediction, Sequence Labelling and Classification

Srikanth Cherla

https://cherla.org

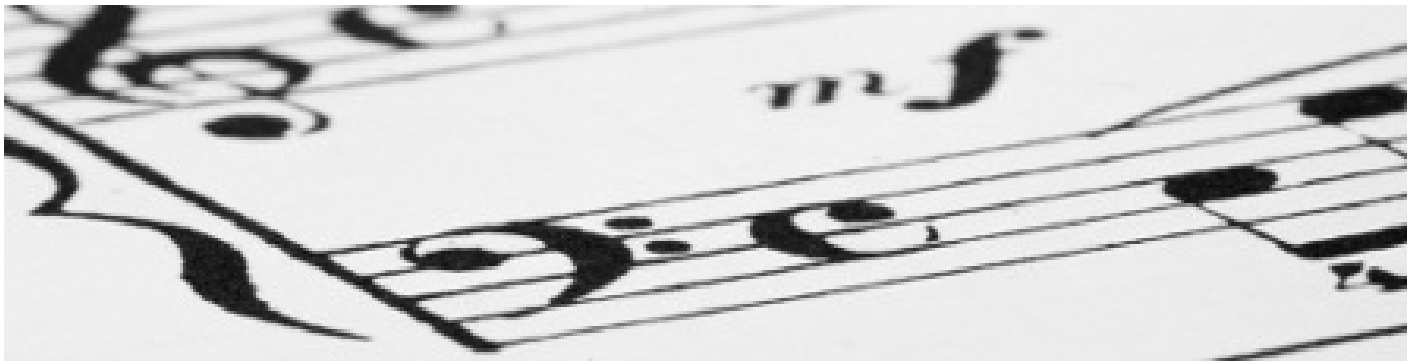
September 13, 2017

Outline

Next

1. Introduction: Analysis of Sequences in Music
2. Preliminaries: Restricted Boltzmann Machines, etc.
3. Contribution: The Recurrent Temporal Discriminative RBM
4. Extension: Generalising the RTDRBM
5. Contribution: Generalising the DRBM

Sequences in Notated Music



- A wealth of information in notated music
- Increasingly available
  - in different formats (MIDI, Kern, GP4, etc.)
  - for different kinds of music (classical, rock, pop, etc.)
- Analysis of sequences key to extracting information
- Melody — Good starting point for a broader analysis

Relevance

Scientific:

- Computational musicology
- Organizing music data
- Aiding acoustic models
- Music education

Creative:

- Automatic music generation
- Compositional assistance

Task: Melody Prediction

- Model a series of musical events s_1^T as follows

$$p\left(s_1^T\right) = \prod_{t=1}^{T} p\left(s_t | s_{(t-n+1)}^{(t-1)}\right)$$

- Conditional probabilities learned from a corpus
- Information theoretic measure - **cross entropy**, to measure a trained model's prediction uncertainty

$$H\left(p, p_m\right) = -\sum_{t=1}^{T} p\left(w_t | w_{(t-n+1)}^{(t-1)}\right) \log_2 p_m\left(w_t | w_{(t-n+1)}^{(t-1)}\right)$$

- How well does a model p_m approximate p?
- Cross entropy to be **minimized**

Motivating Distributed Models

- Previous work focused on n-gram models
- No comparative results with other prediction models
- Thriving neural networks research (Bengio, 2009)
- Recent success of neural network language models (Bengio 2003; Collobert et al., 2011; Mikolov et al., 2010)

Start with an evaluation of connectionist models on the melody prediction task

Next

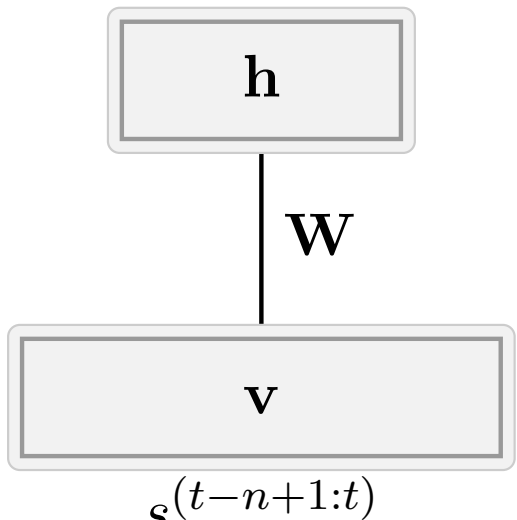
1. Introduction: Analysis of Sequences in Music
2. Preliminaries: Restricted Boltzmann Machines, etc.
3. Contribution: The Recurrent Temporal Discriminative RBM
4. Extension: Generalising the RTDRBM
5. Contribution: Generalising the DRBM

Restricted Boltzmann Machine (Smolensky, 1986)

- Generative, energy-based graphical model.
- Data $\mathbf{v}$ in visible layer, features $\mathbf{h}$ in hidden layer.
- Can model joint probability $p(\mathbf{v})$ of data as

$$p(\mathbf{v}) = \frac{\exp(-\text{FreeEnergy}(\mathbf{v}))}{\sum_{\mathbf{v}^*} \exp(-\text{FreeEnergy}(\mathbf{v}^*))}$$

  where, $\text{FreeEnergy}(\mathbf{v}) = -\log(\sum_{\mathbf{h}} \exp(-\text{Energy}(\mathbf{v}, \mathbf{h})))$
- Learned using Contrastive Divergence (Hinton, 2002).

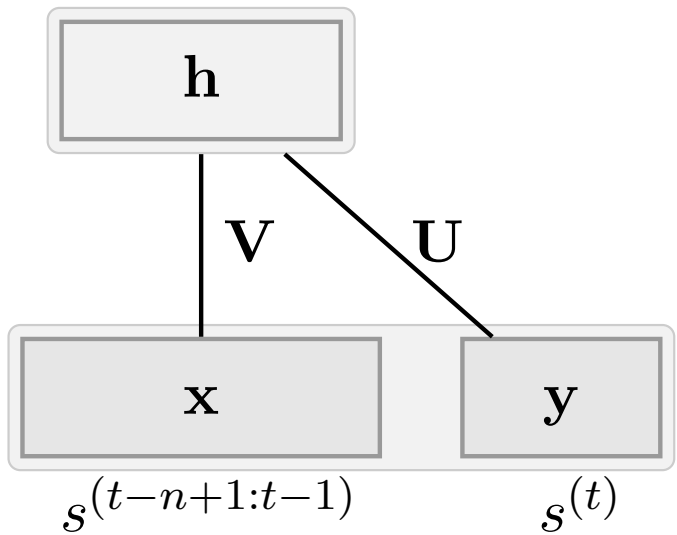


Discriminative RBM (Larochelle & Bengio, 2008)

- Discriminative classifier based on the RBM.
- Data $\mathbf{x}$ and class-label y in visible layer.
- Can model the conditional probability $p(y|\mathbf{x})$ as

$$p(y|\mathbf{x}) = \frac{\exp(-\text{FreeEnergy}(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}^*} \exp(-\text{FreeEnergy}(\mathbf{x}, \mathbf{y}^*))}$$

- Exact gradient computation is possible.

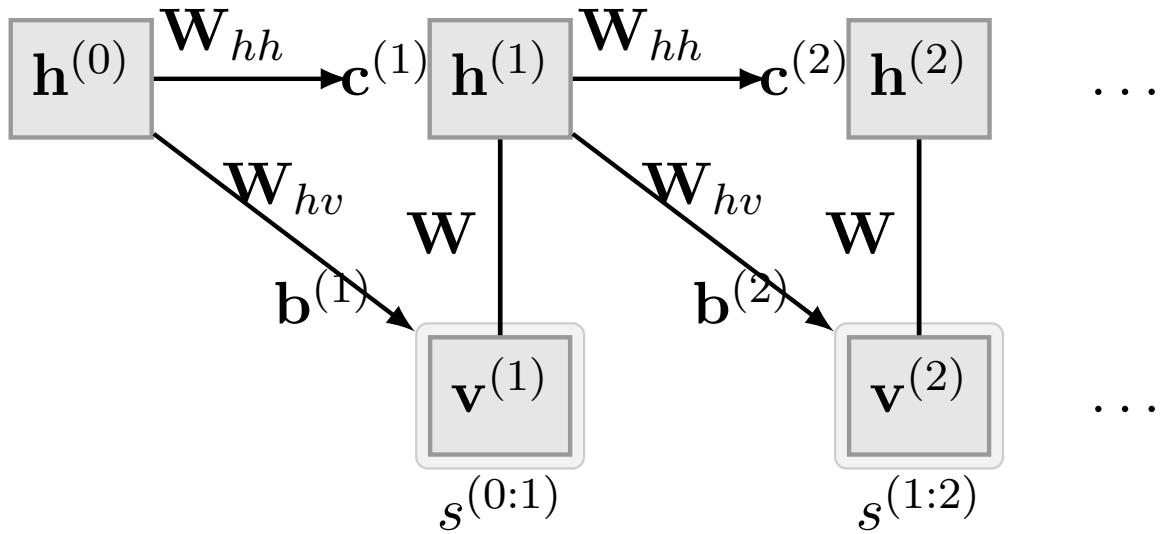


Recurrent Temporal RBM (Sutskever et al., 2009)

- Generative model for high-dimensional time-series.
- RBM at time t conditioned on $\hat{\mathbf{h}}^{(t-1)}$
- Models joint probability of a sequence as

$$p(\mathbf{v}^{(1:T)}, \mathbf{h}^{(1:T)}) = \prod_t p(\mathbf{v}^{(t)}|\mathbf{h}^{(t-1)}) p(\mathbf{h}^{(t)}|\mathbf{v}^{(t)}, \mathbf{h}^{(t-1)})$$

- Learned using Contrastive Divergence and BPTT.



Next

1. Introduction: Analysis of Sequences in Music
2. Preliminaries: Restricted Boltzmann Machines, etc.
3. Contribution: The Recurrent Temporal Discriminative RBM
4. Extension: Generalising the RTDRBM
5. Contribution: Generalising the DRBM

Motivation

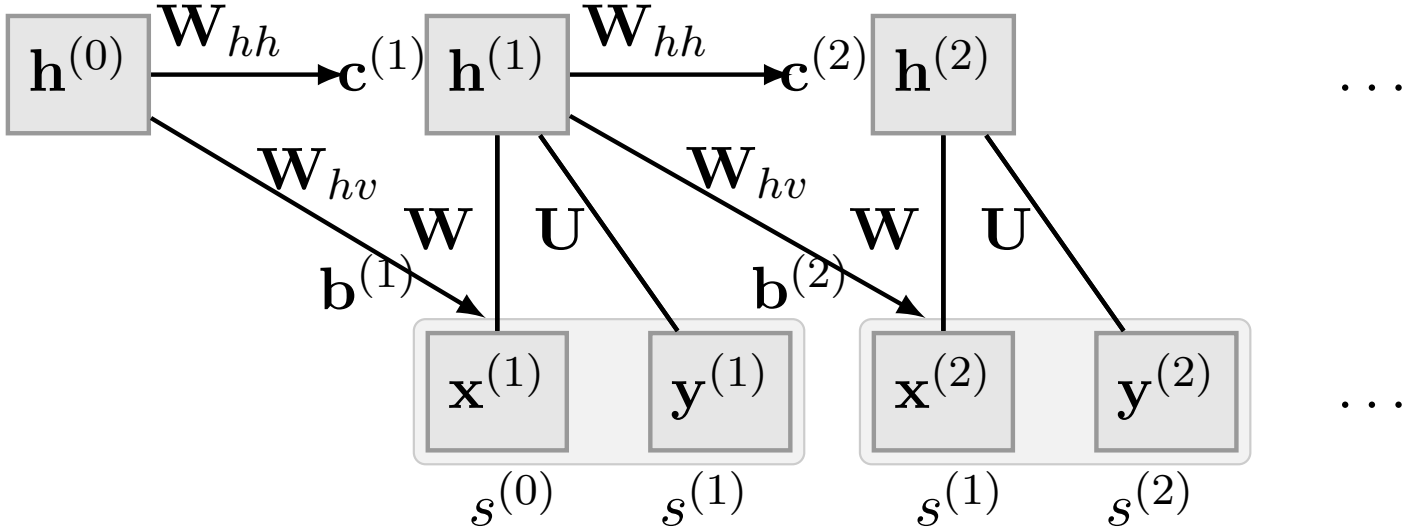
- Discriminative inference on generative RTRBM
- Possible to carry out discriminative learning
- Previous work suggested potential improvements

Discriminative Learning in the **RTRBM** (Cherla et al., 2015)

Extend DRBM learning to a recurrent model

$$p(y^{(t)}|\mathbf{x}^{(1:t)}) = p(y^{(t)}|\mathbf{x}^{(t)}, \hat{\mathbf{h}}^{(t-1)})$$

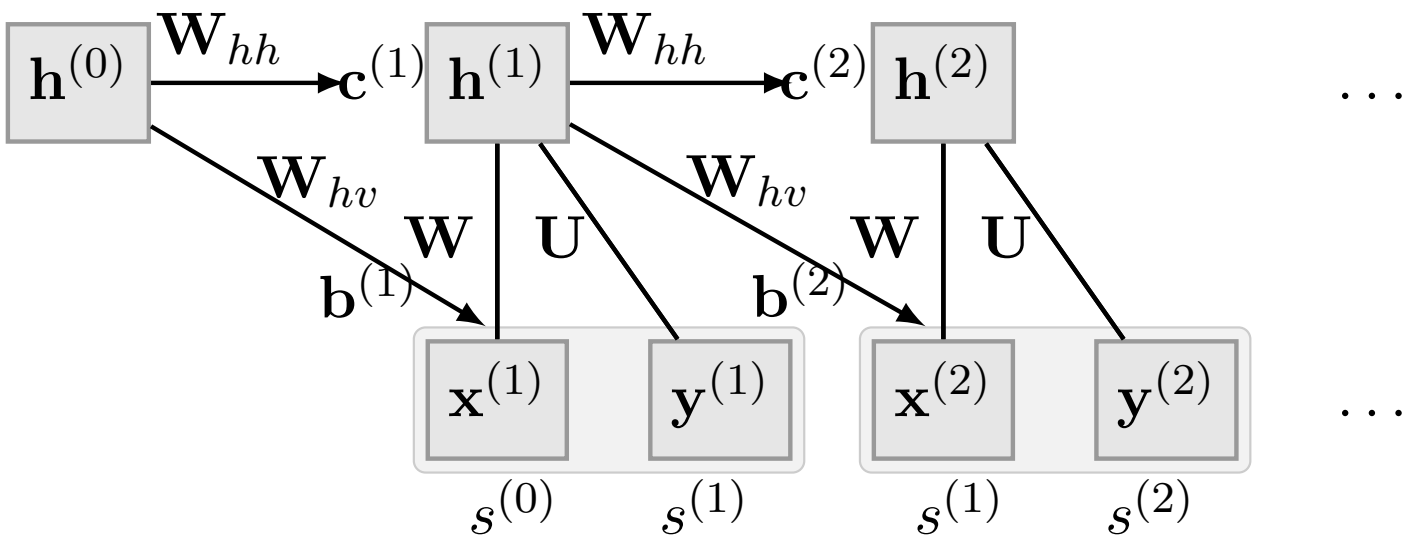
$$= \frac{\exp(-\text{FreeEnergy}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))}{\sum_{\mathbf{y}^*} \exp(-\text{FreeEnergy}(\mathbf{x}^{(t)}, \mathbf{y}^*))}$$



Discriminative Learning in the **RTRBM** (Cherla et al., 2015)

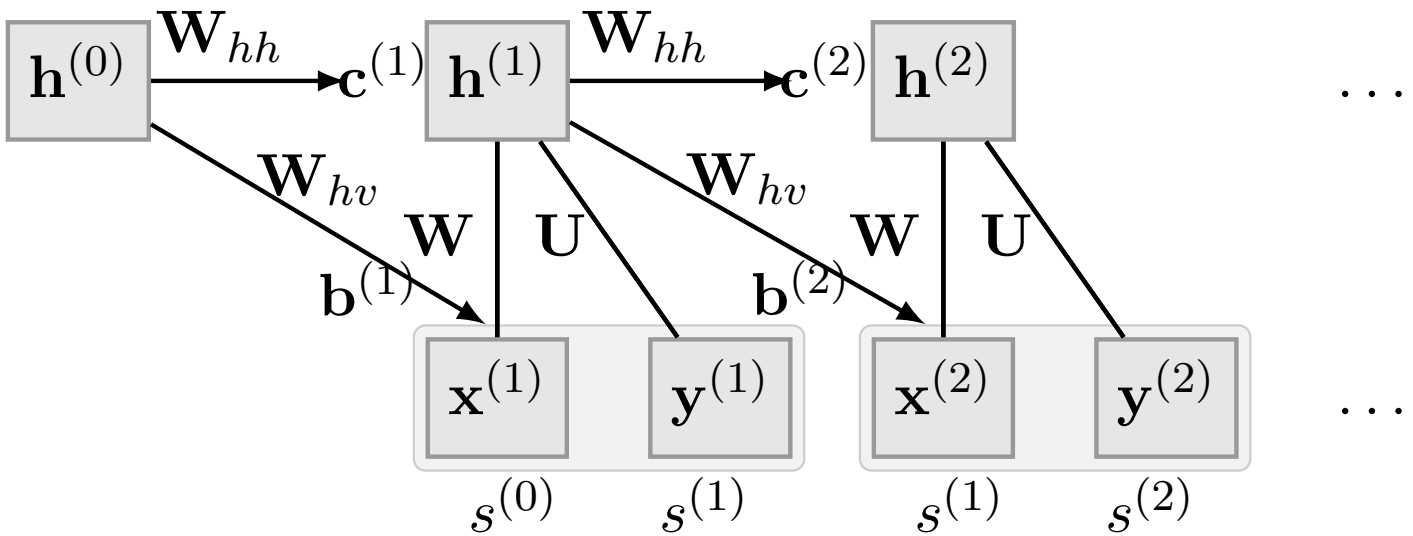
Apply to an entire sequence to optimize the log-likelihood:

$$\begin{aligned}\mathcal{O} &= \log p(\mathbf{y}^{(1:T)}|\mathbf{x}^{(1:T)}) \\ &= \sum_{t=1}^{T} \log p(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \hat{\mathbf{h}}^{(t-1)})\end{aligned}$$



Discriminative Learning in the **RTRBM** (Cherla et al., 2015)

- Recurrent extension of the DRBM.
- Identical in structure to the RTRBM.
- Exact gradient of cost computable at each time-step.
- Back-Propagation Through Time for sequence learning.



Experiments: Melody Corpus

Corpus

- As used in (Pearce & Wiggins, 2004).
- A collection of 8 datasets.
  - Folk songs from the Essen Folk Song Collection.
  - Chorale melodies.

Dataset	No. events	$\lvert\chi\rvert$
Yugoslavian folk songs	2691	25
Alsatian folk songs	4496	32
Swiss folk songs	4586	34
Austrian folk songs	5306	35
German folk songs	8393	27
Canadian folk songs	8553	25
Chorale melodies	9227	21
Chinese folk songs	11056	41

Experiments: Melody Corpus

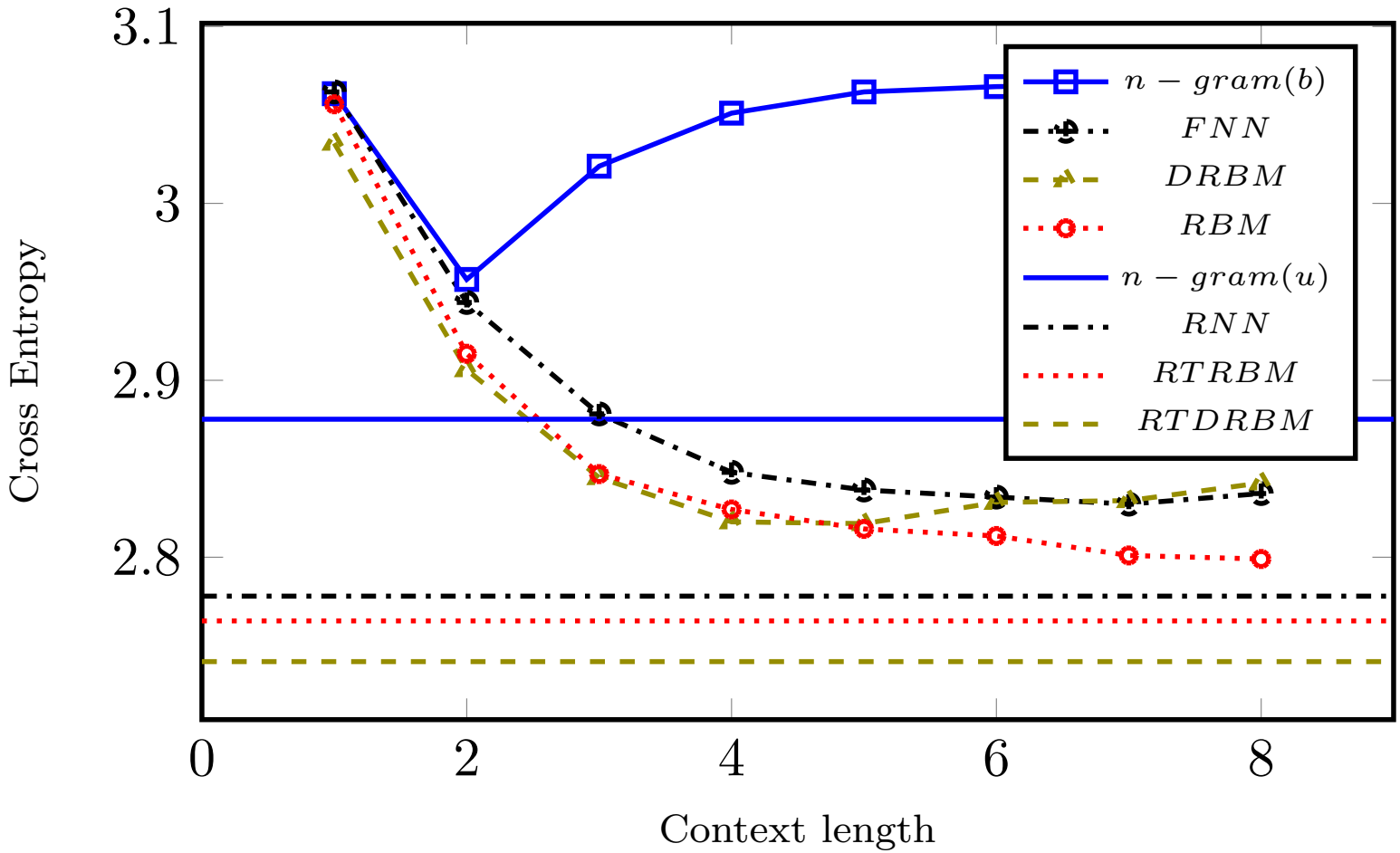
Models

- Non-recurrent: n-grams (b), n-grams (u), FNN, RBMs, DRBMs with context length $\in \{1, 2, 3, 4, 5, 6, 7, 8\}$.
- Recurrent: RNN, RTRBM, **RTDRBM** over entire sequences.
- Hidden units $\in \{25, 50, 100, 200\}$
- Learning rate $\in \{0.01, 0.05\}$
- Trained for 500 epochs.
- Best model determined over a validation set.

Evaluation criterion — cross-entropy

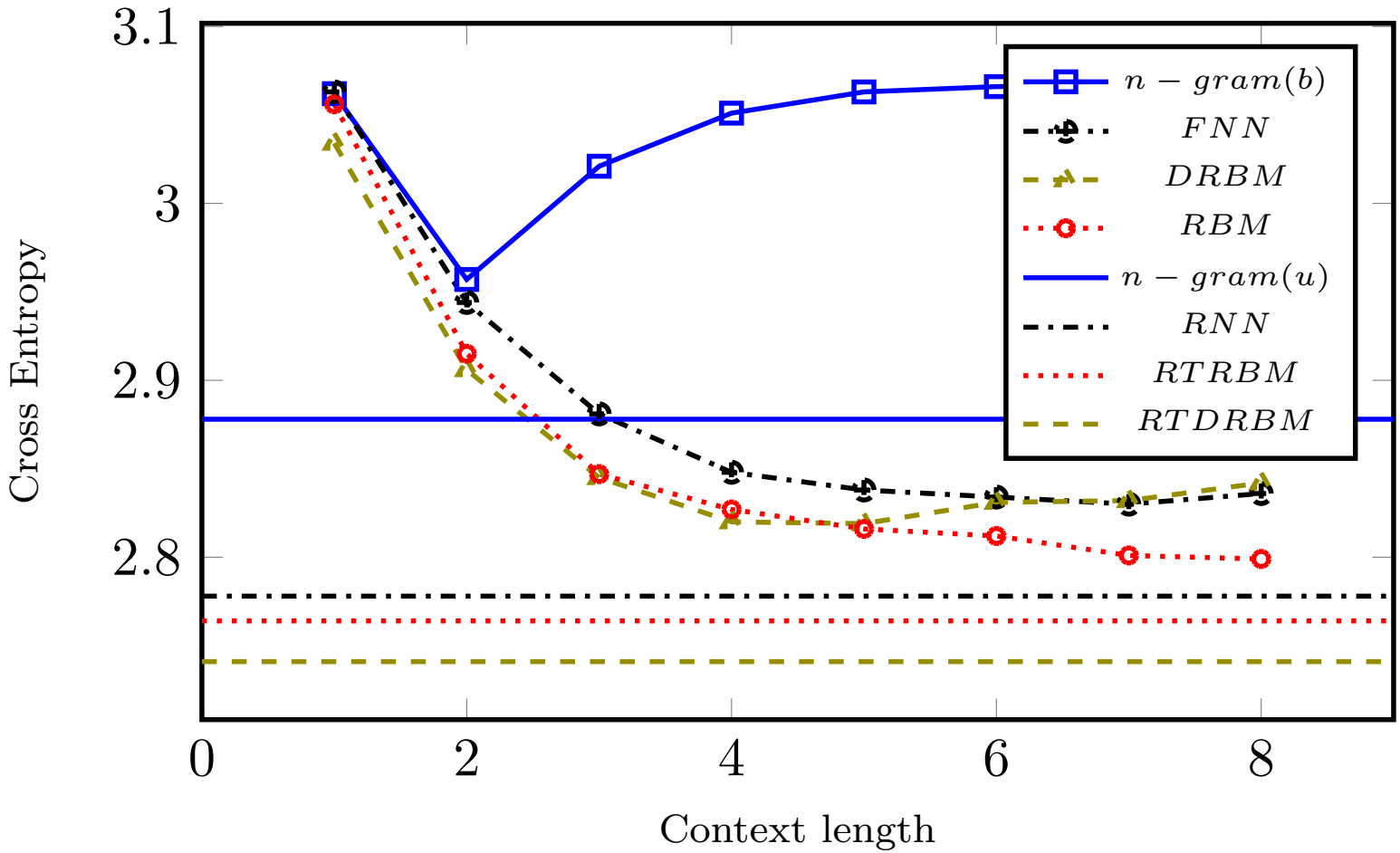
$$H_c(p_{mod}, \mathcal{D}_{test}) = \frac{-\sum_{s_1^n \in \mathcal{D}_{test}} \log_2 p_{mod}(s_n | s_1^{(n-1)})}{|\mathcal{D}_{test}|}$$

Results



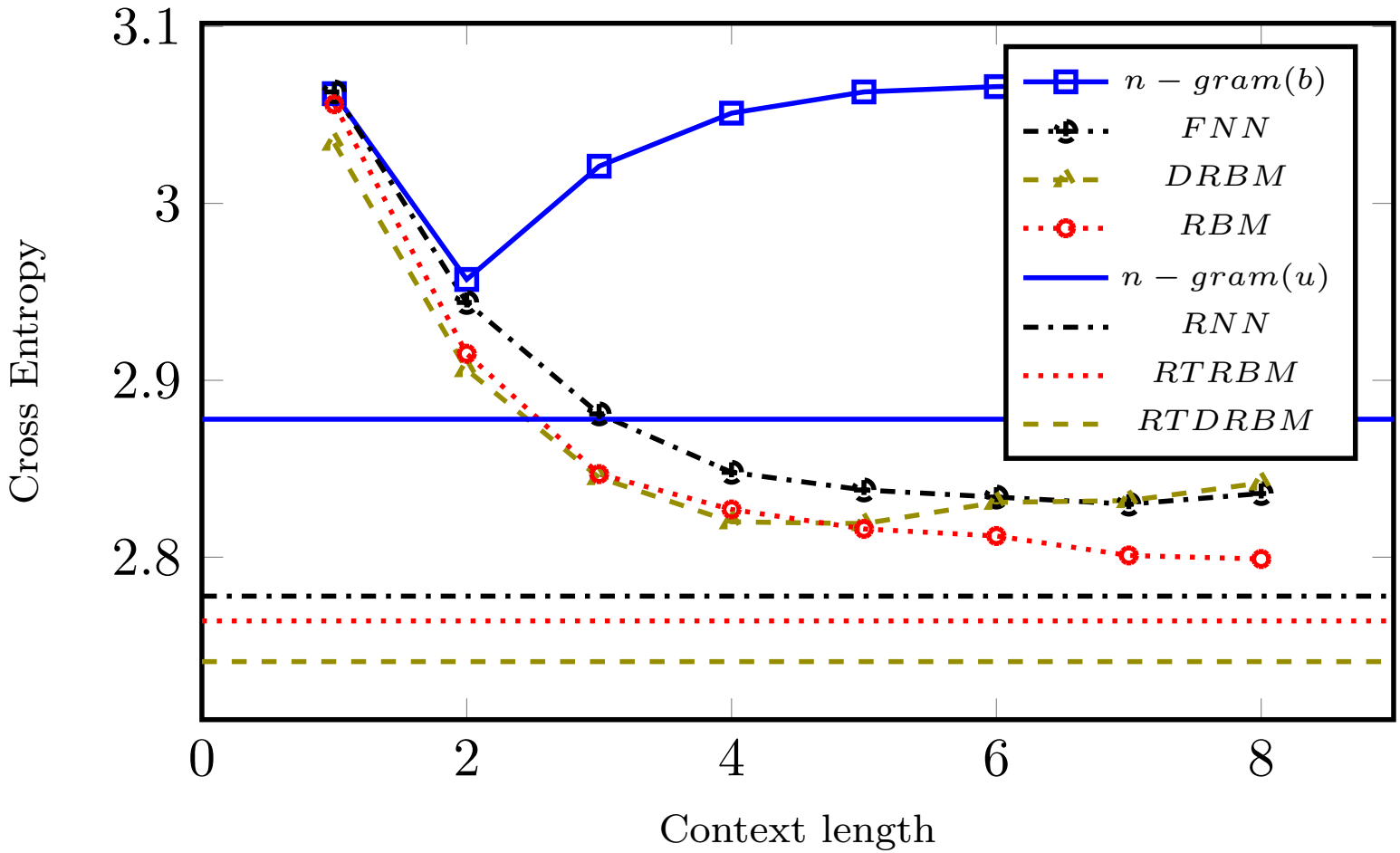
In general, performance improves with context length.

Results



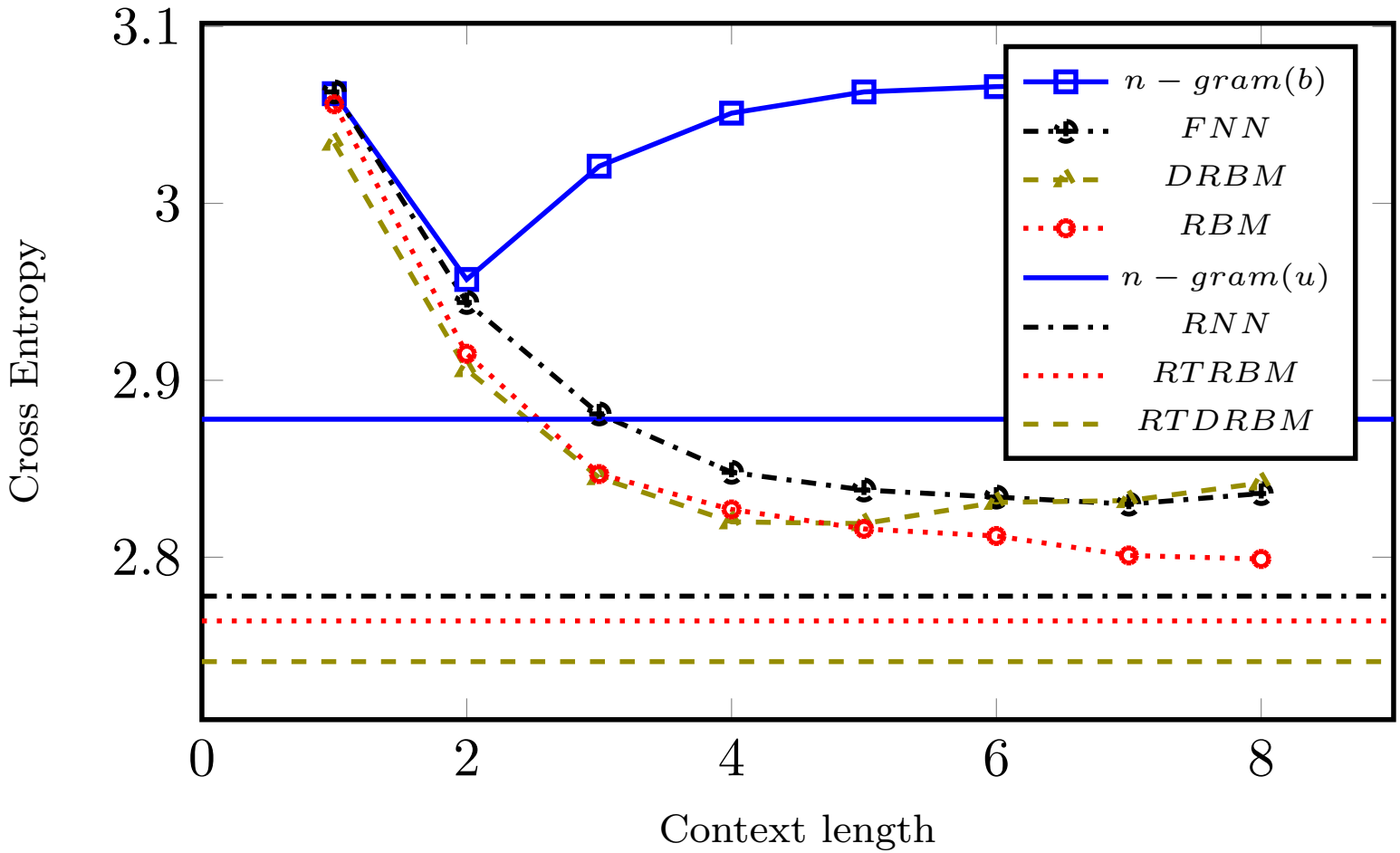
n-gram model performance worsens at lower context length.

Results



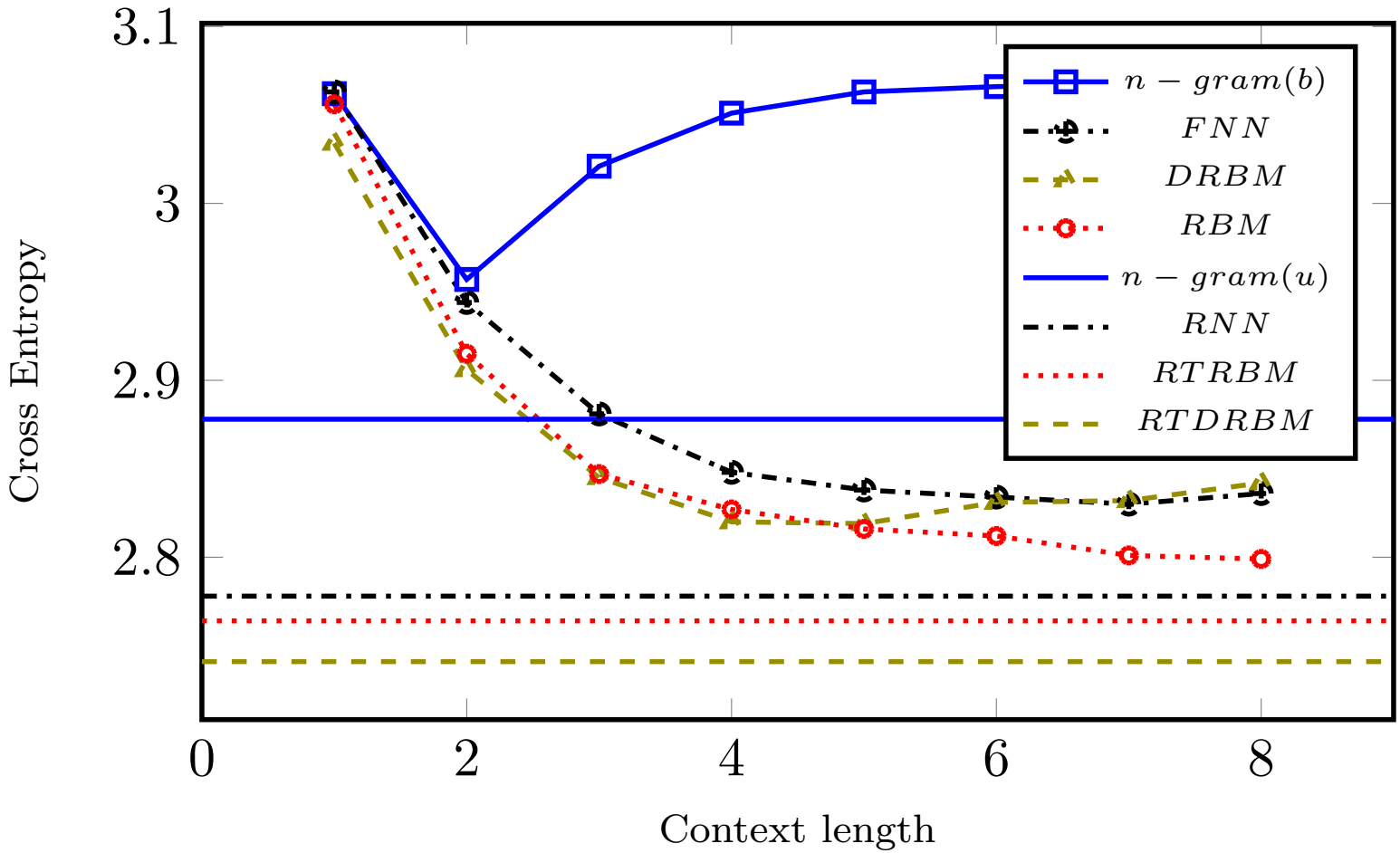
Non-recurrent connectionist models outperform n-grams.

Results



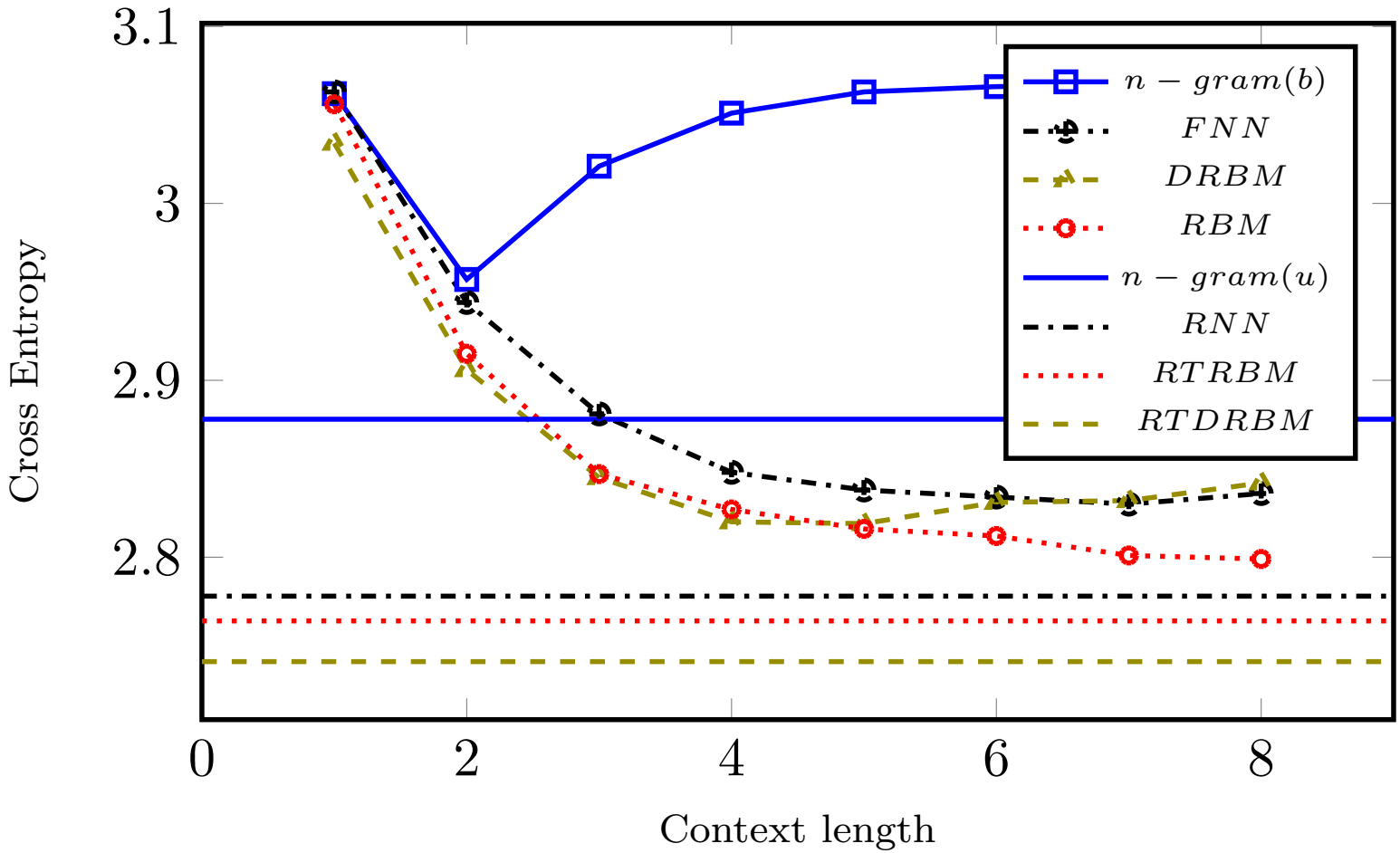
Recurrent connectionist models outperform non-recurrent.

Results



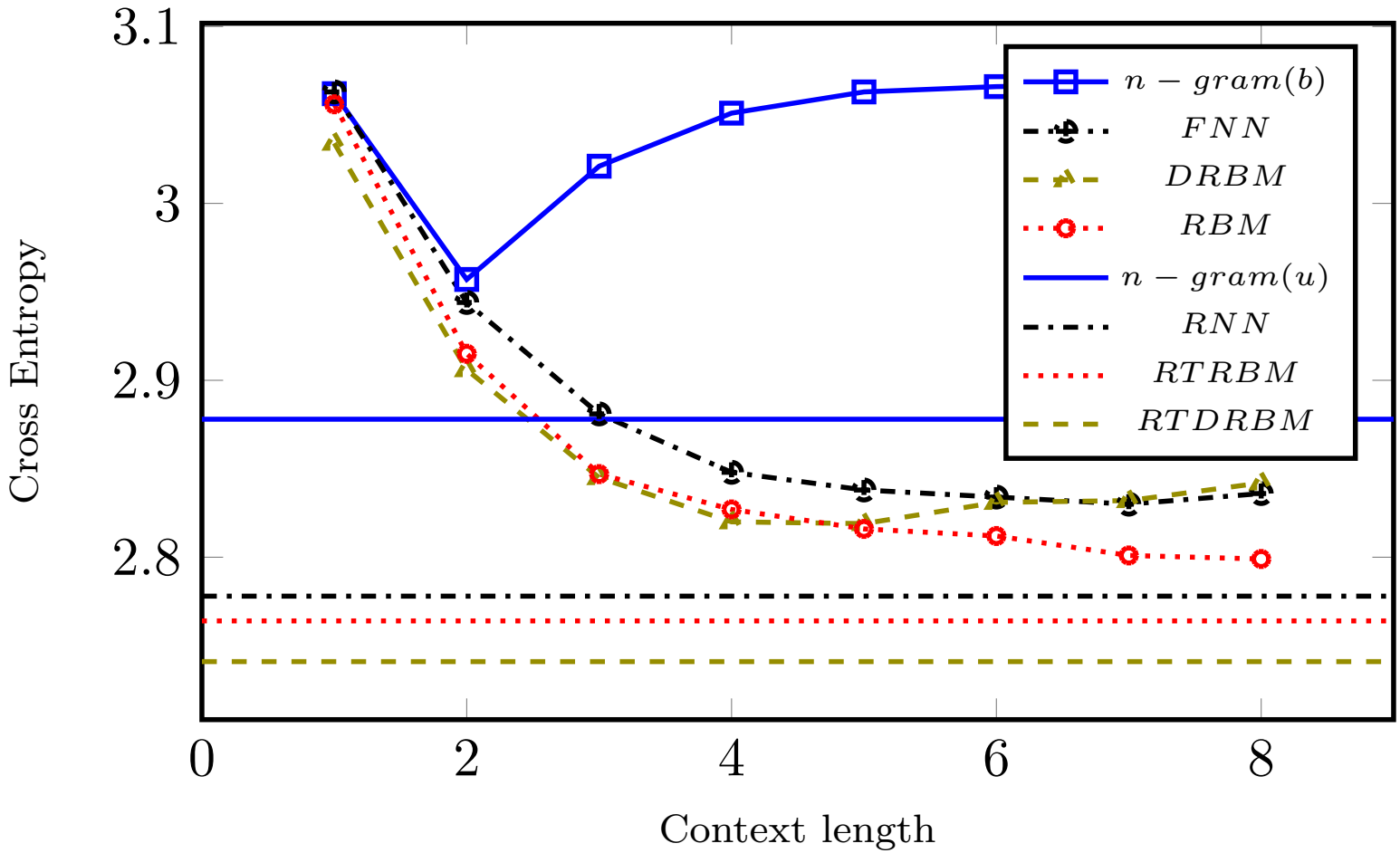
RTDRBM outperforms RTRBM.

Results



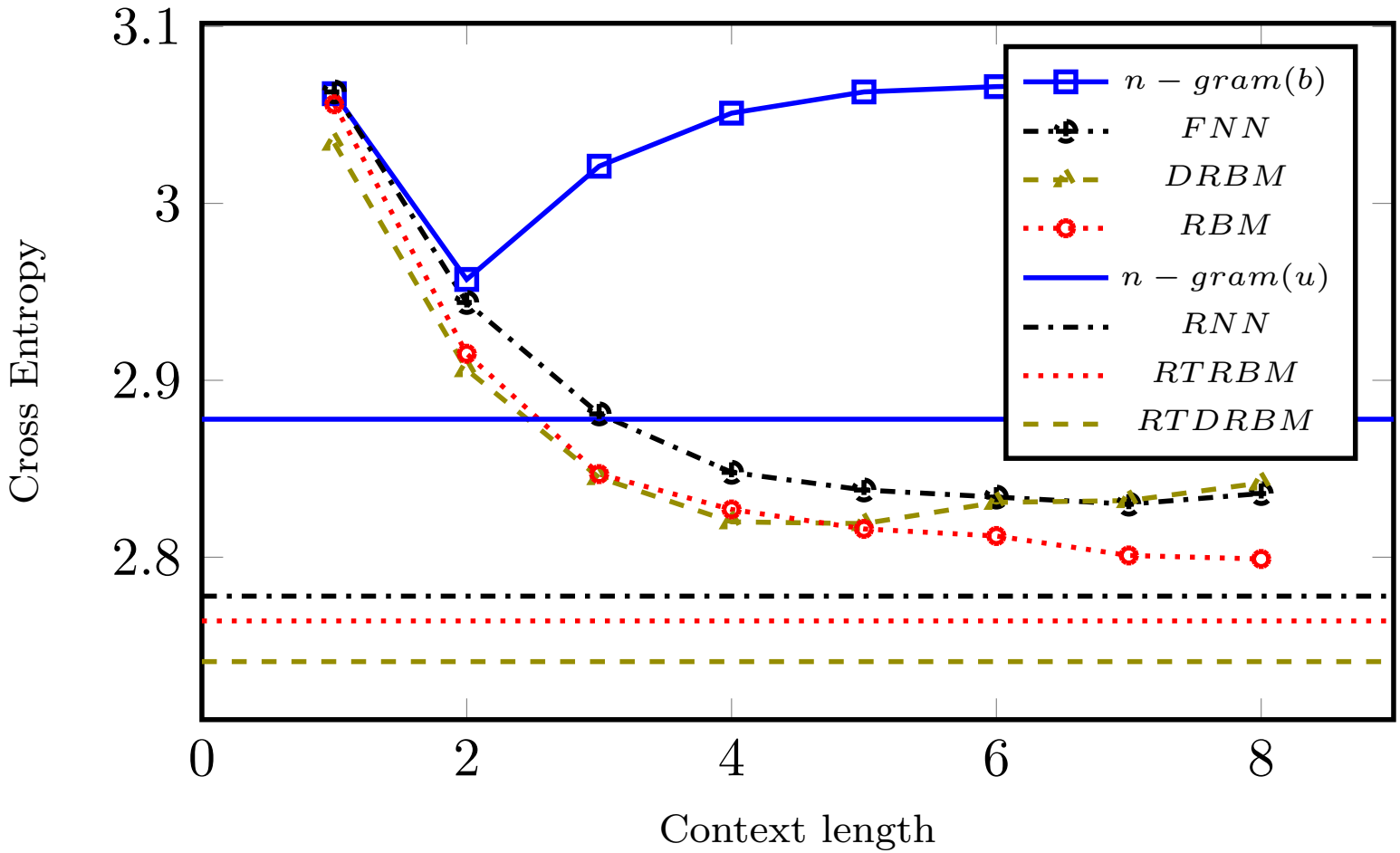
With a shorter context: DRBM outperforms RBM.

Results



With a longer context: RBM outperforms DRBM.

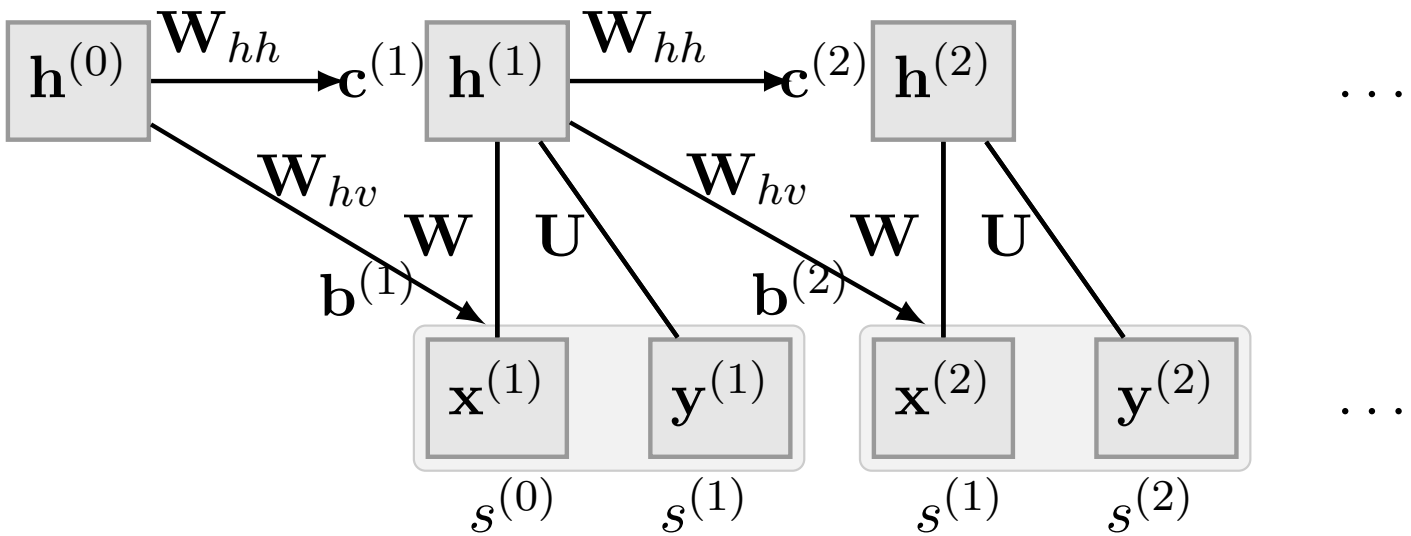
Results



More details and discussion available in the paper.

Next

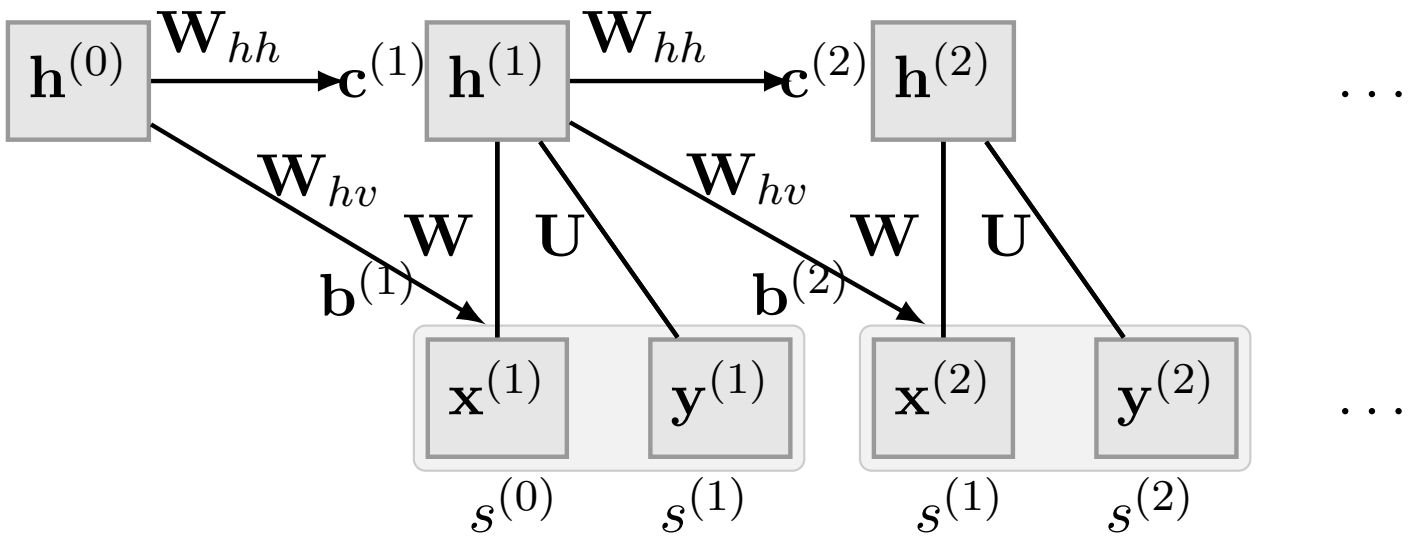
Motivation



$$\begin{aligned}\hat{\mathbf{h}}^{(t-1)} &= \sigma(W\mathbf{x}^{(t-1)} + U\mathbf{y}^{(t-1)} + \mathbf{c}^{(t-1)}) \\ &= \sigma(W\mathbf{x}^{(t-1)} + U\mathbf{y}^{(t-1)} + W_{hh}\hat{\mathbf{h}}^{(t-2)} + \mathbf{c})\end{aligned}$$

Limitation: Dependence of $\mathbf{h}^{(t)}$ on $\mathbf{y}^{*(t-1)}$ which is not suitable for general sequence-labelling problems

Motivation



$$\begin{aligned}\hat{\mathbf{h}}^{(t-1)} &= \sigma(W\mathbf{x}^{(t-1)} + U\mathbf{y}^{(t-1)} + \mathbf{c}^{(t-1)}) \\ &= \sigma(W\mathbf{x}^{(t-1)} + U\mathbf{y}^{(t-1)} + W_{hh}\hat{\mathbf{h}}^{(t-2)} + \mathbf{c})\end{aligned}$$

Solution: Replace $\mathbf{y}^{*(t-1)}$ (unavailable at test time) with predicted output $\mathbf{y}^{(t-1)}$ of previous time-step.

Experiments: OCR

Dataset (Taskar et al., 2004)

- $6,877$ English sentences with $52,152$ words
- Each character a 16×8 binary image
- ASCII code label for each image (26 categories)
- 10 cross-validation folds, one hold-out test set

Method

- Grid search over model hyperparameters
- 10-fold cross validation during model selection
- Models trained over entire sentences

Evaluation: Average Loss Per Sequence

$$E(\mathbf{y}, \mathbf{y}^*) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{L_i} \sum_{j=1}^{L_i} \mathcal{I}\left((y_i)_j \neq (y_i^*)_j \right) \right] \tag{1}$$

Experiments: OCR

Baseline Models (Nguyen & Guo, 2007)

- Multiclass Support Vector Machine ($\text{SVM}^{multiclass}$)
- Structured SVM (SVM^{struct})
- Max-Margin Markov Network (M^3N)
- Averaged Perceptron
- SEARN
- Conditional Random Field (CRF)
- Hidden Markov Model (HMM)
- Structured Learning Ensemble (SLE)

State-of-the-art

- Neural Conditional Random Fields (NCRF) (Do et al., 2010)
- Gradient Boosted Conditional Random Fields (GBCRF) (Chen et al., 2015)

Results: Baseline

Model	**Error (%)**
RTDRBM	**15.95**($\pm$**0.0009**)
SLE	20.58
SVM^{struct}	21.16
HMM	23.70
M^{3}N	25.08
Perceptron	26.40
SEARN	27.02
SVM^{multiclass}	28.54
CRF	32.30

Table: Comparison between the prediction error (%) of the RTDRBM and models evaluated in (Nguyen & Guo, 2007).

Results: State-of-the-art

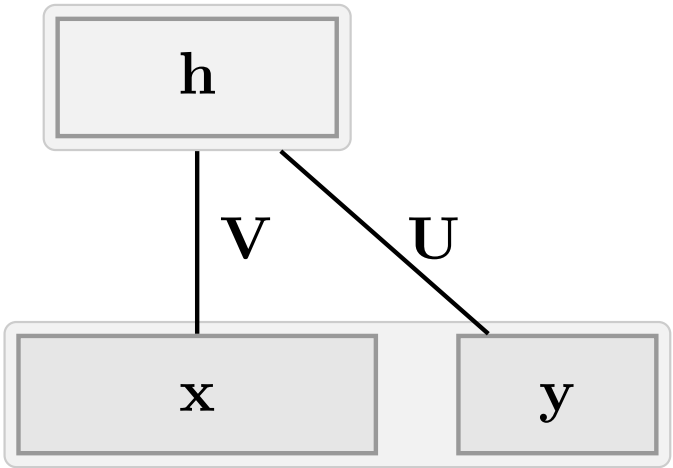
Model	**Error (%)**
NCRF	4.44
GBCRF	4.64($\pm$0.0027)
RTDRBM	**15.95**($\pm$**0.0009**)

Table: Comparison between the prediction error (%) of the RTDRBM and state-of-the-art on the OCR dataset which use Neural Conditional Random Fields (NCRF) (Do et al., 2010) and Gradient Boosted Conditional Random Fields (Chen et al., 2015).

Next

1. Introduction: Analysis of Sequences in Music
2. Preliminaries: Restricted Boltzmann Machines, etc.
3. Contribution: The Recurrent Temporal Discriminative RBM
4. Extension: Generalising the RTDRBM
5. Contribution: Generalising the DRBM

Motivation



- The DRBM is essentially the RBM.
- Various variants of the RBM have been proposed
  - $\{-1, +1\}$-binary hidden unit activations.
  - Integer valued hidden unit activations.
  - Real-valued hidden unit activations.
- How might the same be achieved for the DRBM?

Key Intuition

Generalise the expression for the DRBM conditional distribution $p(y|\mathbf{x})$ as a function of the values that its hidden states can assume, then derive the conditional distribution as per the desired values of its hidden states.

Generalising the DRBM Conditional Distribution (Cherla et al., 2017)

Begin with the expression for the conditional distribution

$$\begin{aligned} P(y|\mathbf{x}) &= \frac{\sum_{\mathbf{h}} P\left(\mathbf{x}, \mathbf{y}, \mathbf{h}\right)}{\sum_{\mathbf{y}^*} \sum_{\mathbf{h}} P\left(\mathbf{x}, \mathbf{y}^*, \mathbf{h}\right)} \\ &= \frac{\sum_{\mathbf{h}} \exp\left(-E\left(\mathbf{x}, \mathbf{y}, \mathbf{h}\right)\right)}{\sum_{\mathbf{y}^*} \sum_{\mathbf{h}} \exp\left(-E\left(\mathbf{x}, \mathbf{y}^*, \mathbf{h}\right)\right)} \end{aligned} \tag{2}$$

where $\mathbf{y}$ is the one-hot encoding of a class label y.

Generalising the DRBM Conditional Distribution (Cherla et al., 2017)

This can be generalised as follows (details in the paper):

$$P(y|\mathbf{x}) = \frac{\exp\left(b_y\right) \prod_j \sum_k \exp\left(s_k \sum_i x_i w_{ij} + u_{yj} + c_j\right)}{\sum_{y^*} \exp\left(b_{y^*}\right) \prod_j \sum_k \exp\left(s_k \sum_i x_i w_{ij} + u_{y^* j} + c_j\right)} \tag{3}$$

where s_k is each of the k states that can be assumed by each hidden unit j of the model.

(Re-)Deriving the DRBM (Cherla et al., 2017)

The (Bernoulli) DRBM conditional distribution can be derived when the states $s_k = \{0, 1\}$.

$$
\begin{aligned}
P_{\text{ber}}\left(y|\mathbf{x}\right) &= \frac{\exp\left(b_y\right)\prod_j \sum_{s_k\in\{0,1\}} \exp\left(s_k\alpha_j\right)}{\sum_{y^*}\exp\left(b_{y^*}\right)\prod_j \sum_{s_k\in\{0,1\}} \exp\left(s_k\alpha_j^*\right)} \\
&= \frac{\exp\left(b_y\right)\prod_j \left(1+\exp\left(\alpha_j\right)\right)}{\sum_{y^*}\exp\left(b_{y^*}\right)\prod_j \left(1+\exp\left(\alpha_j^*\right)\right)}
\end{aligned}
\tag{4}
$$

The Bipolar DRBM (Cherla et al., 2017)

The Bipolar DRBM conditional distribution can be derived when the states $s_k = \{-1, +1\}$.

$$
\begin{aligned}
P_{\text{bip}}\left(y|\mathbf{x}\right) &= \frac{\exp\left(b_y\right)\prod_j \sum_{s_k\in\{-1,+1\}} \exp\left(s_k\alpha_j\right)}{\sum_{y^*}\exp\left(b_{y^*}\right)\prod_j \sum_{s_k\in\{-1,+1\}} \exp\left(s_k\alpha_j^*\right)} \\
&= \frac{\exp\left(b_y\right)\prod_j \left(\exp\left(-\alpha_j\right)+\exp\left(\alpha_j\right)\right)}{\sum_{y^*}\exp\left(b_{y^*}\right)\prod_j \left(\exp\left(-\alpha_j^*\right)+\exp\left(\alpha_j^*\right)\right)} \,.
\end{aligned} \tag{5}
$$

The Binomial DRBM (Cherla et al., 2017)

The Binomial DRBM conditional distribution can be derived when the states $s_k = \{0, \ldots, N\}$.

$$
\begin{aligned}
S_N &= \sum_{s_k=0}^{N} \exp\left(s_k \alpha_j\right) \\
&= 1 + \exp\left(\alpha_j\right) \sum_{s_k=0}^{(N-1)} \exp\left(s_k \alpha_j\right) \\
&= \frac{1 - \exp\left(\left(N+1\right)\alpha_j\right)}{1 - \exp\left(\alpha_j\right)}
\end{aligned}
\tag{6}
$$

Experiments: ML Benchmarks

- Datasets
  1. MNIST digit classification.
  2. USPS digit classification.
  3. 20 Newsgroups document classification.
- Grid search with each model evaluated over 10 seeded runs.
- The value of N (bins) in the Binomial DRBM varied as $\{2, 4, 8\}$.
- Maximise log-likelihood on training and validation set.
- Report average classification error on test set $E(\mathbf{y}, \mathbf{y}^*) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{I}\left(y_i \neq y_i^*\right)$.

Results: MNIST

Model	**Average Loss(%)**
DRBM ($n_{hid} = 500$, $\eta_{init} = 0.05$)	**1.78**($\pm$0.0012)
Bipolar DRBM ($n_{hid} = 500$, $\eta_{init} = 0.01$)	1.84($\pm$0.0007)
Binomial DRBM ($n_{hid} = 500$, $\eta_{init} = 0.01$)	1.86($\pm$0.0016)

Table: Results on the USPS dataset. The Binomial DRBM in this table is the one with $n_{bins} = 2$.

n_{bins}	n_{hid}	η_{init}	**Average Loss (%)**
2	500	0.01	**1.86**
4	500	0.01	1.88
8	500	0.001	1.90

Table: Performance of the Binomial DRBM with different values of n_{bins}. The difference was within the margin of significance.

Results: USPS

Model	**Average Loss (%)**
DRBM ($n = 50$, $\eta_{init} = 0.01$)	$6.90(\pm 0.0047)$
Bipolar DRBM ($n = 500$, $\eta_{init} = 0.01$)	$6.49(\pm 0.0026)$
Binomial DRBM ($n = 1000$, $\eta_{init} = 0.01$)	$\mathbf{6.09}(\pm 0.0014)$

Table: Performance on the USPS dataset. The Binomial DRBM in this table is the one with $n_{bins} = 8$.

n_{bins}	η_{init}	n_{hid}	**Average Loss (%)**
2	0.01	50	$6.90(\pm 0.0047)$
4	0.01	1000	$6.48(\pm 0.0018)$
8	0.01	1000	$\mathbf{6.09}(\pm 0.0014)$

Table: Classification average losses of the Binomial DRBM with different values of n_{bins}.

Results: 20 Newsgroups

Model	**Average Loss (%)**
DRBM ($n = 50$, $\eta_{init} = 0.01$)	$28.52(\pm 0.0049)$
Bipolar DRBM ($n = 50$, $\eta_{init} = 0.001$)	$\mathbf{27.75}(\pm 0.0019)$
Binomial DRBM ($n = 100$, $\eta_{init} = 0.001$)	$28.17(\pm 0.0028)$

Table: Performance on the 20 Newsgroups dataset. The Binomial DRBM in this table is the one with $n_{bins} = 2$.

n_{bins}	η_{init}	n_{hidden}	**Average Loss (%)**
2	0.001	100	$\mathbf{28.17}(\pm 0.0028)$
4	0.001	50	$28.24(\pm 0.0032)$
8	0.0001	50	$28.76(\pm 0.0040)$

Table: Classification performance of the Binomial DRBM with different values of n_{bins}.

Acknowledgements

Parts of the above described work were done in collaboration with Son N. Tran (now a researcher at CSIRO) at *City, University London.*

Thank you!

Questions?