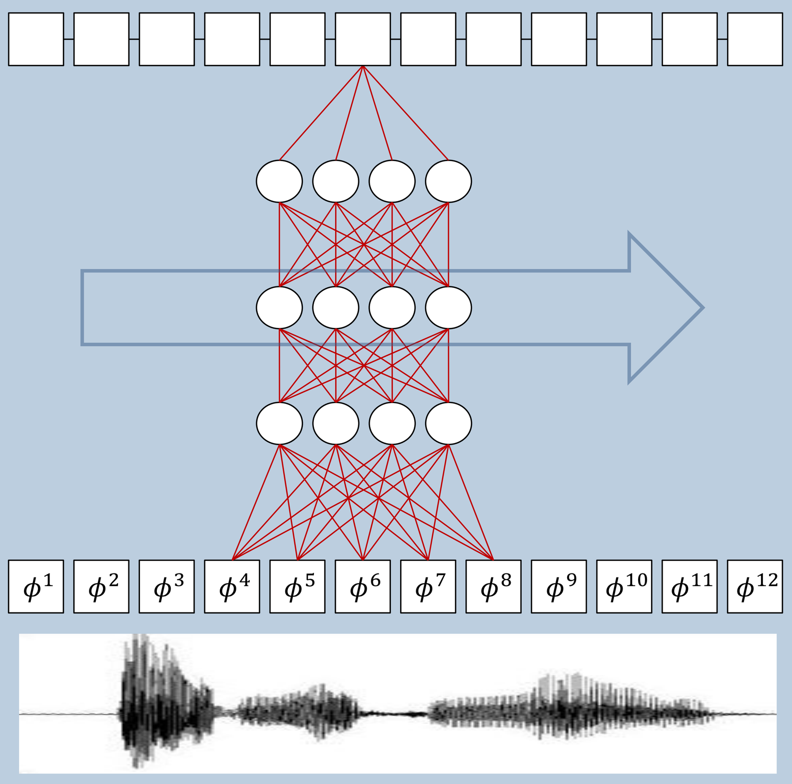
Sequential Deep Belief Networks

Galen Andrew Jeff Bilmes
galen@cs.washington.edu bilmes@ee.washington.edu

Abstract

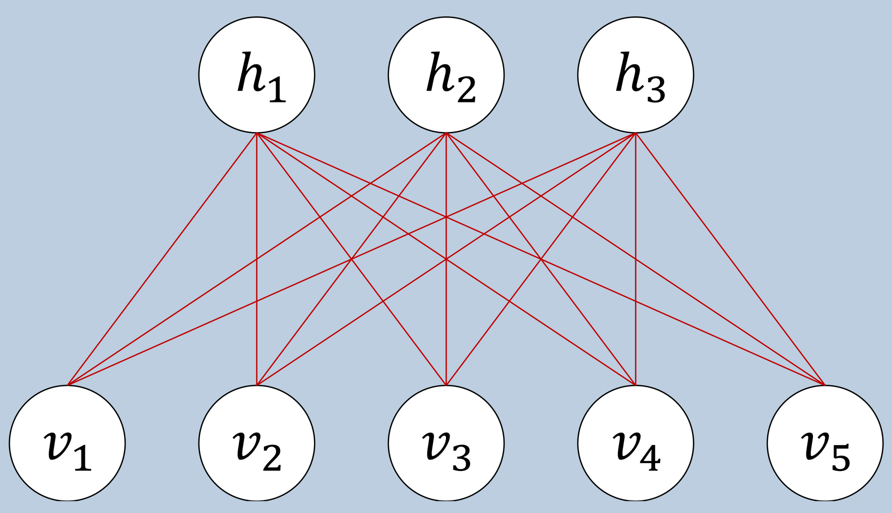
- We introduce a new model called the Sequential Deep Belief Network (SDBN).
- SDBNs allow correlation between corresponding units in successive time frames
- Hidden units potentially capable of detecting arbitrarily long temporal patterns
- Tractable pretraining/training algorithms analogous to static DBN, amenable to optimization with fast matrix algebra
- Experiments on TIMIT phone recognition show advantage of temporal connections

Sliding window DBN



- Deep Belief Network over sliding window [1]
- Temporal integration occurs only via the Markov Chain/MRF/CRF in the output
- Hidden units can only recognize patterns that occur within the range of the window

Basic RBM



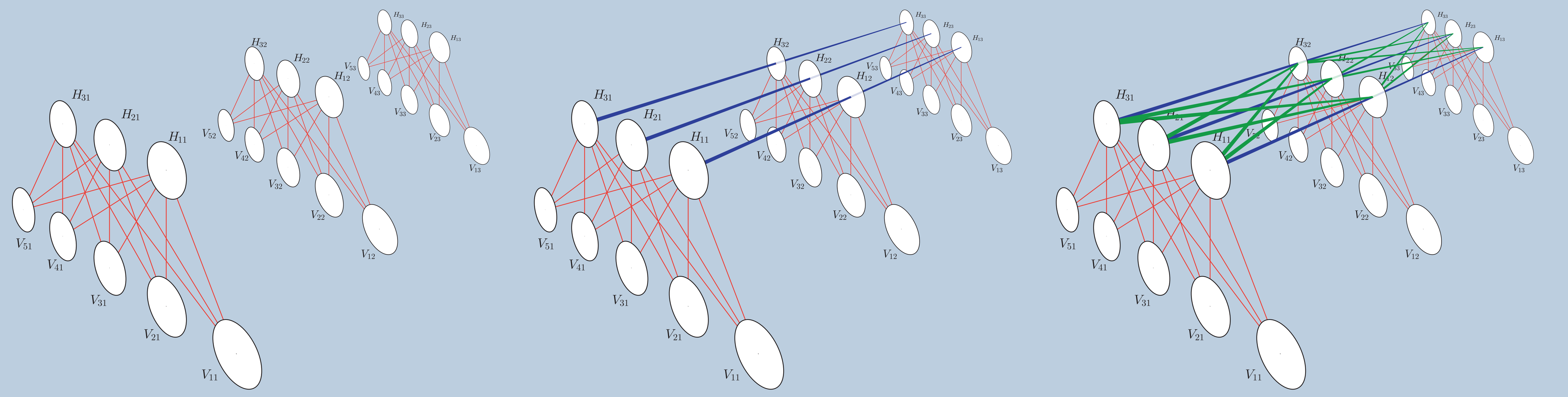
- Restricted Boltzmann Machine has visible layer $v \in \mathbb{R}^{n_v}$ and hidden layer $h \in \mathbb{R}^{n_h}$
- Joint energy $E(v,h) = -v'Wh$ for weight matrix W. (Biases omitted.)
- $\Pr(v,h) \propto \exp -E(v,h)$
- $\Pr(h_j = 1|v) = \sigma(v'W_{*j})$, where $\sigma(t) = (1+\exp(-t))^{-1}$
- Variables within each layer are independent given the other, enabling fast Contrastive Divergence (CD) training [2]

References

[1] A. Mohammed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Advances in Neural Information Processing Systems 22*, 2009.

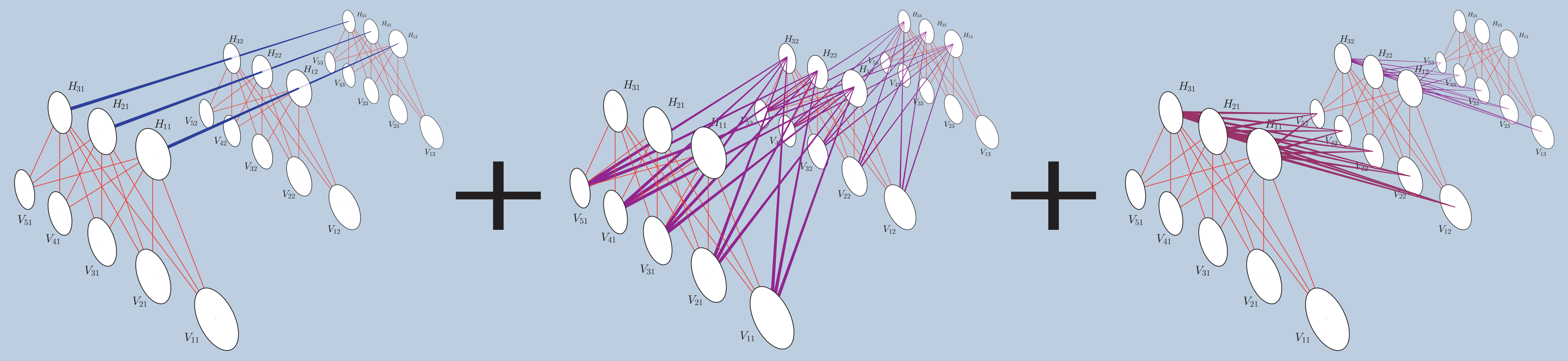
[2] G. Hinton, S. Osindero, Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation 18*, 2006.

Sequential Restricted Boltzmann Machine

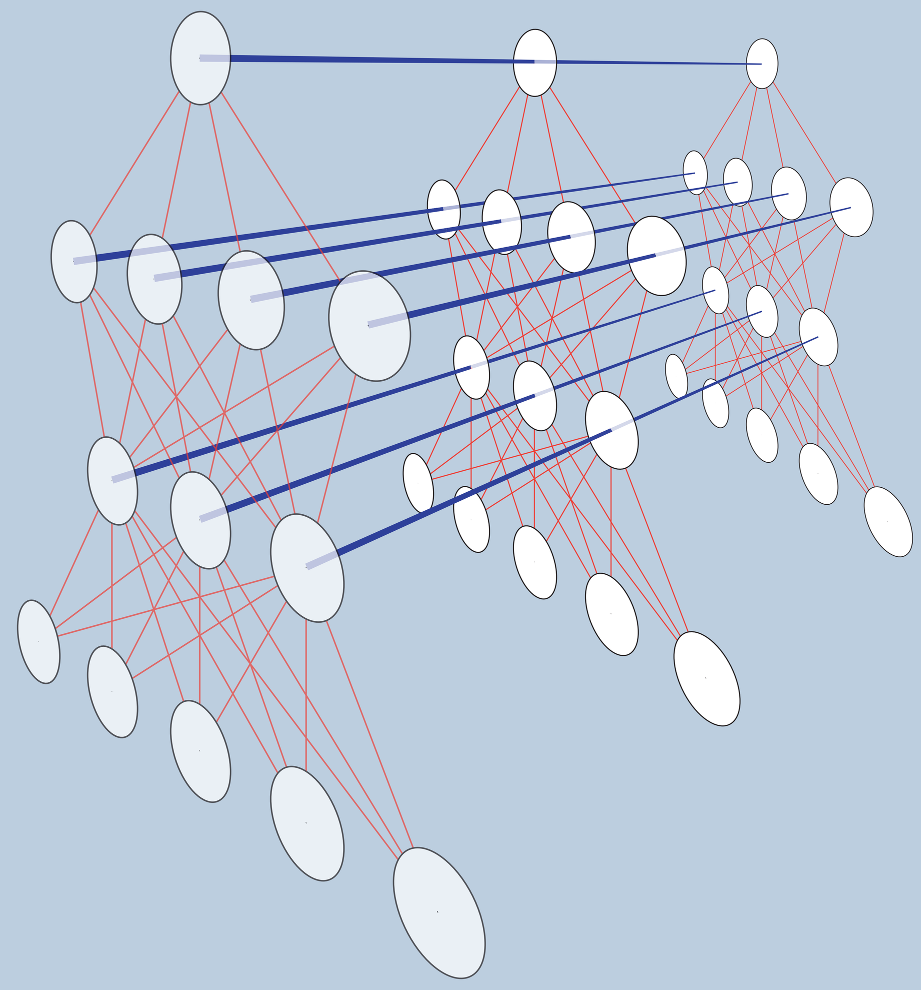


Independent RBMs over time SRBM ($\delta_{\max} = 0$) Intractable

- SRBM has matrix-valued $V \in \mathbb{R}^{n_v \times T}$ and $H \in \mathbb{R}^{n_h \times T}$ and parameters $W \in \mathbb{R}^{n_v \times n_h}$ and $\mathbf{t} \in \mathbb{R}^{n_h}$
- If both layers are binary, then $E_{\mathcal{B}}(V,H) = -\operatorname{tr} VWH - \sum_{t=1}^{T-1} \sum_{j=1}^{n_h} H_{jt}\mathbf{t}_j H_{j(t+1)}$
- If the visible layer is Gaussian, then $E_{\mathcal{G}}(V,H) = E_{\mathcal{B}}(V,H) + \frac{1}{2} \operatorname{tr} V'V$
- Visible variables independent given hiddens, but hidden variables form a set of independent Markov sequences given visibles. CD training is still tractable
- Direct connections from V_{it} to $H_{j(t+\delta)}$ for $|\delta| \leq \delta_{\max}$ also used (seen on the right is $\delta_{\max} = 1$)



Sequential DBN



SDBN ($\delta_{\max} = 0$)

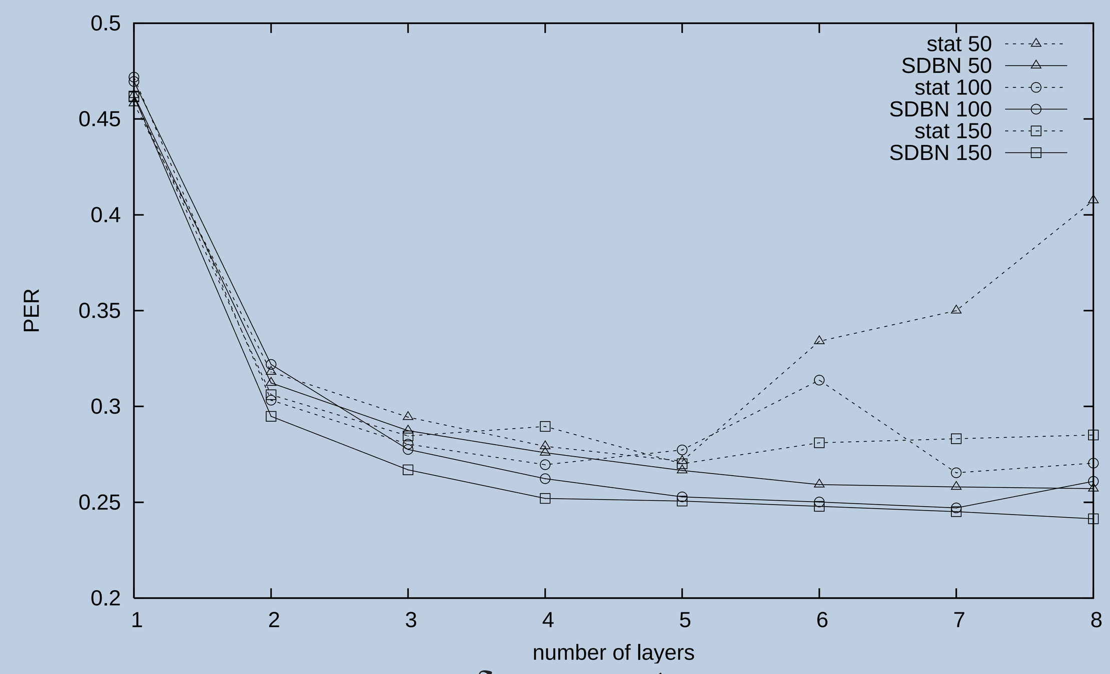
- SDBN is formed by stacking multiple layers of SRBM, with a Markov sequence of multinomial variables at the top layer (a CRF)
- The marginal values of the variables at each hidden layer given the previous layer is used as the input to the next layer: $V^l = \mathbb{E}\left[H^l\right]$
- Hidden units can potentially detect arbitrarily long temporal patterns due to earlier layers' Baum-Welch stages
- Training: first pretrain each SRBM layer with CD. Then fine-tune parameters by approximately maximizing $\ell = \log \Pr(\hat{Y}|V^{L-1})$ with stochastic gradient descent
- Gradient is computed exactly using a procedure similar to error backpropagation (BP). "Upward" pass requires running Baum-Welch on each independent chain; "downward" pass uses similar message-passing scheme

Acknowledgements

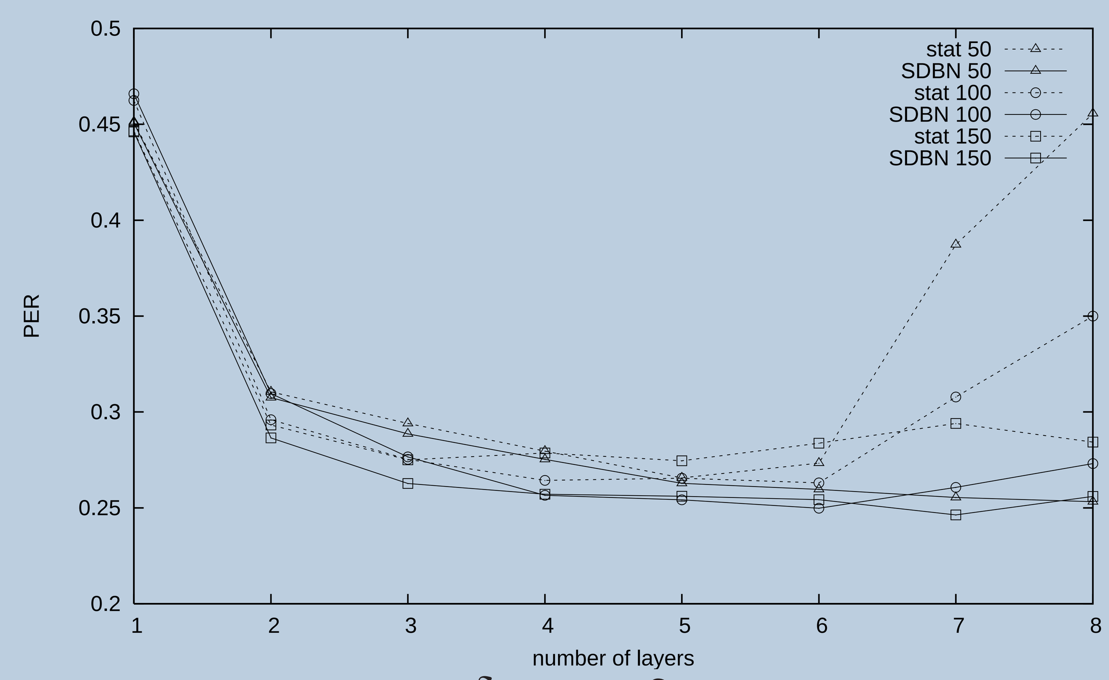
This research was supported by NSF grant IIS-0905341. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

Results

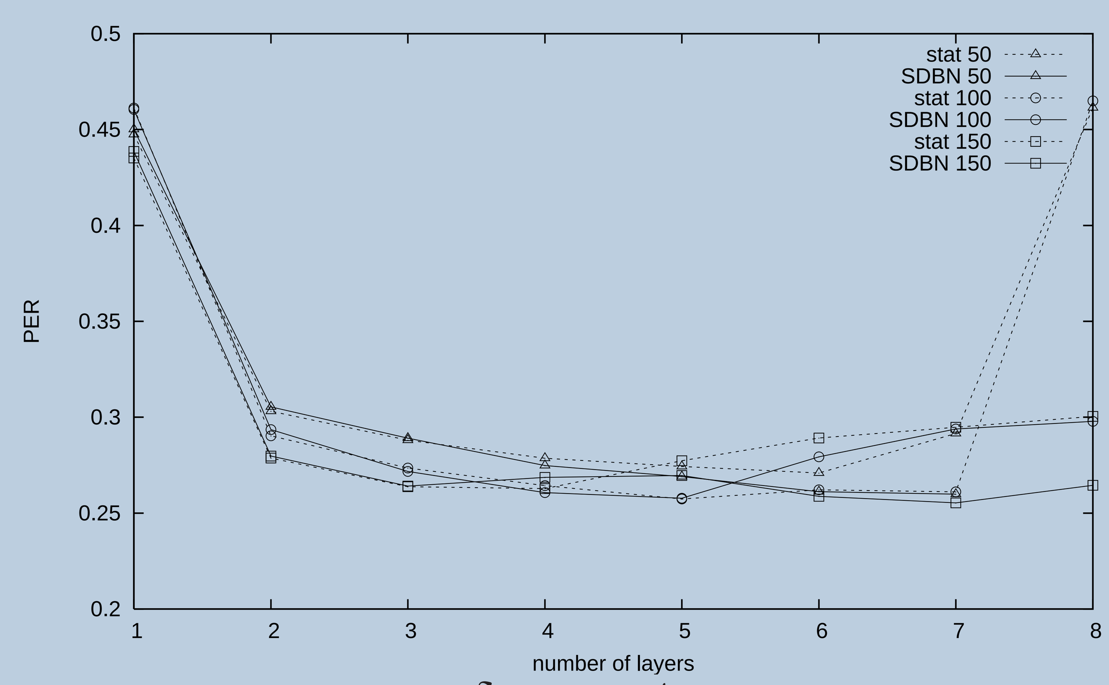
- Compared SDBN to sliding window DBN ("stat") on TIMIT for a range of layer sizes, number of hidden layers, and values of $\delta_{\max}$
- Normalized 12th-order MFCCs and energy plus first-order temporal differences
- Standard 39 phone set, divided into two states per phone, with latent boundaries



$\delta_{\max} = 1$



$\delta_{\max} = 2$



$\delta_{\max} = 4$