A RESEARCH AGENDA

• Deep learning successes have required a lot of labeled training data
  ‣ collecting and labeling such data requires significant human labor
  ‣ is that really how we’ll solve AI?

• Alternative solution: exploit other sources of data that are imperfect but plentiful
  ‣ unlabeled data (unsupervised learning)
  ‣ multimodal data (multimodal learning)
  ‣ multidomain data (transfer learning, domain adaptation)
A RESEARCH AGENDA

• By far the largest source is unlabeled data
  ‣ effectively requires algorithms for *life-long learning*

• We are currently poorly equipped to deal with this setting
  ‣ how to do online learning for non-convex models, with a changing input distribution?
  ‣ how to have models whose capacity adapts during training?
A RESEARCH AGENDA

• By far the largest source is unlabeled data
  ‣ effectively requires algorithms for *life-long learning*

• We are currently poorly equipped to deal with this setting
  ‣ how to do online learning for non-convex models, with a changing input distribution?
  ‣ how to have models whose capacity adapts during training?
A RESEARCH AGENDA

• By far the largest source is unlabeled data
  ‣ effectively requires algorithms for life-long learning

• We are currently poorly equipped to deal with this setting
  ‣ how to do online learning for non-convex models, with a changing input distribution?
  ‣ how to have models whose capacity adapts during training?

• In this talk: an infinite restricted Boltzmann machine (iRBM)
  ‣ RBM with capacity that can grow during training
  ‣ growing mechanism is derived naturally from the energy function definition
RESTRICTED BOLTZMANN MACHINE

\[ E(v, h) = -h^T W v - v^T b^v - h^T b^h \]
\[ P(v, h) = e^{-E(v,h)} / Z \]
ORDERED RESTRICTED BOLTZMANN MACHINE

\[ E(v, h, z) = -v^T b^v - \sum_{i=1}^{z} h_i (W_i v + b_i^h) - \beta_i \]
Ordered Restricted Boltzmann Machine

$E(v, h, z) = -v^T b^v - \sum_{i=1}^{z} h_i (W_i v + b^h_i) - \beta_i$
INFINITE RESTRICTED BOLTZMANN MACHINE

\[ Z \in \{1, \ldots, \infty\} \]

\[ E(v, h, z) = -v^T b^v - \sum_{i=1}^{z} h_i (W_i v + b_i^h) - \beta_i \]
INFINITE RESTRICTED BOLTZMANN MACHINE

\[ z \in \{1, \ldots, \infty \} \]

\[ E(v, h, z) = -v^T b^v - \sum_{i=1}^{z} h_i (W_i \cdot v + b_i^h) - \beta_i \]
\[ E(v, h, z) = -v^T b^v - \sum_{i=1}^{z} h_i (W_i \cdot v + b_i^h) - \beta_i \]
INFINITE RESTRICTED BOLTZMANN MACHINE

- Free energy, given a certain value of $z$

$$F(v, z) = -v^T b^v - \sum_{i=1}^{\infty} \text{soft}_+(W_i \cdot v + b_i^h) - \beta_i$$

- For the iRBM to be practical, we must be able to compute

$$P(z|v) = \frac{\exp(-F(v, z))}{Z(v)} = \frac{\exp(-F(v, z))}{\sum_{z'} \exp(-F(v, z'))}$$
INFINITE RESTRICTED BOLTZMANN MACHINE

\[ Z(v) = \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp(-F(v, z)) \]

\[ = \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp \left( -F(v, l) + \sum_{i=l+1}^{z} \text{soft}_+(W_i.v + b^h_i) - \beta_i \right) \]
INFINITE RESTRICTED BOLTZMANN MACHINE

\[ Z(v) = \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp(-F(v, z)) \]

\[ = \sum_{z=1}^{l} \exp(-F(v, z)) \cdot \sum_{z=l+1}^{\infty} \exp \left( -F(v, l) + \sum_{i=l+1}^{z} \text{soft}_+(W_i \cdot v + b^h_i) - \beta_i \right) \]

\[ \beta_i = \beta \text{soft}_+(b^h_i) \]

\[ \beta > 1 \]
\[
Z(v) = \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp(-F(v, z))
\]

\[
= \sum_{z=1}^{l} \exp(-F(v, z)) \cdot \sum_{z=l+1}^{\infty} \exp \left( -F(v, l) + \sum_{i=l+1}^{z} \operatorname{soft}_+(W_i \cdot v + b_i^h) - \beta_i \right)
\]

\[
= \sum_{z=1}^{l} \exp(-F(v, z)) + \exp(-F(v, l)) \sum_{z=l+1}^{\infty} \exp \left( \sum_{i=l+1}^{z} (1 - \beta) \operatorname{soft}_+(0) \right)
\]

\[
\beta_i = \beta_{\operatorname{soft}_+}(b_i^h)
\]

\[
\beta > 1
\]
INFINITE RESTRICTED BOLTZMANN MACHINE

\[
Z(\mathbf{v}) = \sum_{z=1}^{l} \exp(-F(\mathbf{v}, z)) + \sum_{z=l+1}^{\infty} \exp(-F(\mathbf{v}, z))
\]

\[
= \sum_{z=1}^{l} \exp(-F(\mathbf{v}, z)) + \sum_{z=l+1}^{\infty} \exp\left(-F(\mathbf{v}, l) + \sum_{i=l+1}^{z} \text{soft}_+ (\mathbf{W}_i \cdot \mathbf{v} + b_i^h) - \beta_i\right)
\]

\[
= \sum_{z=1}^{l} \exp(-F(\mathbf{v}, z)) + \exp(-F(\mathbf{v}, l)) \sum_{z=l+1}^{\infty} \exp\left(\sum_{i=l+1}^{z} (1 - \beta)\text{soft}_+ (0)\right)
\]

\[
= \sum_{z=1}^{l} \exp(-F(\mathbf{v}, z)) + \exp(-F(\mathbf{v}, l)) \sum_{z=1}^{\infty} \exp((1 - \beta)\text{soft}_+ (0))^z
\]

\[
\beta_i = \beta \text{soft}_+ (b_i^h)
\]

\[
\beta > 1
\]
INFINITE RESTRICTED BOLTZMANN MACHINE

\[
Z(v) = \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp(-F(v, z)) \\
= \sum_{z=1}^{l} \exp(-F(v, z)) + \sum_{z=l+1}^{\infty} \exp \left( -F(v, l) + \sum_{i=l+1}^{z} \text{soft}_+(W_i.v + b_i^h) - \beta_i \right) \\
= \sum_{z=1}^{l} \exp(-F(v, z)) + \exp(-F(v, l)) \sum_{z=l+1}^{\infty} \exp \left( \sum_{i=l+1}^{z} (1 - \beta)\text{soft}_+(0) \right) \\
= \sum_{z=1}^{l} \exp(-F(v, z)) + \exp(-F(v, l)) \sum_{z=1}^{\infty} \exp((1 - \beta)\text{soft}_+(0))^z
\]

Geometric series

\[
\beta_i = \beta_{\text{soft}_+(b_i^h)} \\
\beta > 1
\]
INFINITE RESTRICTED BOLTZMANN MACHINE

- Can perform Gibbs sampling
  - before sampling $h$, first sample $z$ given $v$
    \[
P(z|v) = \frac{\exp(-F(v, z))}{Z(v)}\]
    
    \[
P(h_i = 1|v, z) = \begin{cases} 
      \sigma(W_i \cdot v + b_i^h) & \text{if } i \leq z \\
      0 & \text{otherwise}
    \end{cases}
    \]
    
    \[
P(v_j = 1|h, z) = \sigma \left( \sum_{i=1}^{z} W_{ij} h_i + b_j^v \right)
    \]
INFINITE RESTRICTED BOLTZMANN MACHINE

• Can perform Gibbs sampling
  ‣ before sampling $h$, first sample $z$ given $v$

• If can perform Gibbs sampling, can perform contrastive divergence CD training
  ‣ Gibbs sampling provides negative samples for update
INFINITE RESTRICTED BOLTZMANN MACHINE

• Can perform Gibbs sampling
  ‣ before sampling $h$, first sample $z$ given $v$

• If can perform Gibbs sampling, can perform contrastive divergence CD training
  ‣ Gibbs sampling provides negative samples for update

• CD training is well defined
  ‣ only selected hidden units get a non-zero gradient on their weights
  ‣ any amount of regularization ensures that training does not diverge to infinitely many units with non-zero weights
Training of iRBM
EXPERIMENTS

- Binarized MNIST and CalTech101 Silhouettes

<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>Binarized MNIST</th>
<th>CalTech101 Silhouettes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ln((\hat{Z} \pm 3\sigma))</td>
<td>Avg. NLL</td>
</tr>
<tr>
<td>RBM</td>
<td>100</td>
<td>[600.88, 600.95]</td>
<td>98.17 ± 0.52</td>
</tr>
<tr>
<td>RBM</td>
<td>500</td>
<td>[613.24, 613.31]</td>
<td>86.50 ± 0.44</td>
</tr>
<tr>
<td>RBM</td>
<td>2000</td>
<td>[1098.94, 1099.17]</td>
<td>85.03 ± 0.42</td>
</tr>
<tr>
<td>oRBM</td>
<td>500</td>
<td>[39.90, 40.19]</td>
<td>88.15 ± 0.46</td>
</tr>
<tr>
<td>iRBM</td>
<td>1208</td>
<td>[40.03, 40.54]</td>
<td>85.65 ± 0.44</td>
</tr>
</tbody>
</table>

- Use adagrad for training
- Training robust to value of \(\beta\)
  - used 1.01 in all experiments
EXPERIMENTS

- Distribution $p(z | v)$ (Binarized MNIST)
FUTURE WORK

• Can be extended to other types of representations
  ‣ feed-forward neural networks
  ‣ RBM with softmax units, for quantization-based fast search
  ‣ RBM with tree-based representations, for hierarchical topic modeling
  ‣ word representations (infinite Skip-Gram, Nalisnick and Ravi, 2015)
http://github.com/MarcCote/iRBM

Côté & Larochelle (2016) *An Infinite Restricted Boltzmann Machine*. *Neural Computation*
MERCİ!

http://github.com/MarcCote/iRBM

Côté & Larochelle (2016) *An Infinite Restricted Boltzmann Machine*. *Neural Computation*