Speech Recognition Using Deep Believe Networks

Abdel-rahman Mohamed

Department of Computer Science
University of Toronto
The speech signal

DBNs for speech recognition and generation
Automatic Speech Recognition (ASR)

Concept: a sequence of symbols

Parameterise

Speech Waveform

Speech Vectors

Recognise

S_1  S_2  S_3  etc
ASR existing models

Markov Model $M$

DBNs for speech recognition and generation
Motivation

• The state-of-the-art techniques for acoustic modeling suffers from unrealistic independence assumptions.

• Looking for new models that offer more representational capacity.
Restricted Boltzmann Machines (RBMs) (1)

- An RBM is a bipartite graph in which visible units are connected to binary stochastic hidden units using undirected weighted connections.
- RBMs have an efficient generative training procedure as well as discriminative fine tuning mechanisms.
RBMs (2)

- The energy of the joint configuration \((v, h)\) is given by:

\[
E(v, h; \theta) = \sum_{i=1}^{\nu} \frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{\nu} \sum_{j=1}^{\mathcal{H}} w_{ij}v_i h_j - \sum_{j=1}^{\mathcal{H}} a_j h_j
\]

- The probability that the model assigns to a visible vector \(v\) is:

\[
p(v; \theta) = \frac{\sum_h e^{-E(v, h)}}{\sum_u \sum_h e^{-E(u, h)}}
\]

- Conditional distributions \(p(v|h)\) and \(p(h|v)\) are factorial and given by:

\[
p(h_j = 1 | v; \theta) = \sigma \left( \sum_{i=1}^{\nu} w_{ij}v_i + a_j \right)
\]

\[
p(v_i = 1 | h; \theta) = \mathcal{N} \left( \sum_{j=1}^{\mathcal{H}} w_{ij} h_j + b_i, 1 \right)
\]
**Using RBMs for phone recognition**

(Mohamed, Hinton, ICASSP 2010)

- A context window of successive feature vectors is used to set the states of the visible units.
- To train an RBM to model the joint distribution of data and labels, the visible vector is concatenated with a binary vector of class labels.

\[
E(v, l, h; \theta) = - \sum_{i=1}^{\mathcal{V}} \sum_{j=1}^{\mathcal{H}} w_{ij} h_j v_i - \sum_{k=1}^{\mathcal{L}} \sum_{j=1}^{\mathcal{H}} w_{kj} h_j l_k - \sum_{j=1}^{\mathcal{L}} a_j h_j - \sum_{k=1}^{\mathcal{L}} c_k l_k + \sum_{i=1}^{\mathcal{V}} \frac{(v_i - b_i)^2}{2}
\]

\[
p(l_k = 1|h; \theta) = \text{softmax}(\sum_{j=1}^{\mathcal{H}} w_{kj} h_j + c_k)
\]

- The DBN produces a probability distribution over the possible labels of the central frame. Then probabilities are fed to a standard Viterbi decoder.
RBM and its variants

(a) RBM

(b) CRBM

(c) ICRBM

RBM and its variants for Phone recognition
RBM training: Generative training

• By maximizing the likelihood function of the visible data, we get:

\[ \Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \]

• The Contrastive Divergence (CD) approximation is used:

\[ \Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{1} \]

• For the CRBM, the directed connection updates are:

\[ \Delta A_{ij}^{(t-q)} = v_i^{(t-q)} \left( \langle v_j^t \rangle_{data} - \langle v_j^t \rangle_{1} \right) \]

\[ \Delta B_{ij}^{(t-q)} = v_i^{(t-q)} \left( \langle h_j^t \rangle_{data} - \langle h_j^t \rangle_{1} \right) \]
**RBM training:** Discriminative training

- \( p(l|v) \) can be computed exactly by:

\[
p(l|v) = \frac{\sum_h e^{-E(v,l,h)}}{\sum_l \sum_h e^{-E(v,l,h)}}
\]

- The gradient of \( \log p(l|v) \) can also be computed exactly. The update rule for the vis-hid weights is:

\[
\Delta w_{ij} = v_i \sigma \left( a_j + w_{jm} + \sum_{i=1}^\nu w_{ij} v_i \right) - v_i \sum_{k=1}^\mathcal{L} p(l_k=1|v) \sigma \left( a_j + w_{jk} + \sum_{i=1}^\nu w_{ij} v_i \right)
\]

- To avoid model overfitting, we follow the gradient of:

\[
f(v, l) = \alpha \log p(l|v) + \log p(v|l)
\]
Evaluation Setup

• The core test set of the TIMIT database is used. The MIT development set (50 speakers) was used for model tuning.
• $12^{th}$ order MFCC and energy along with $1^{st}$ and $2^{nd}$ derivatives were used as features.
• A context window of 11 feature frames was used.
• All architectures contain 2000 hidden units.
• We used 183 target class labels ($3 \times 61$ phones).
• We used a bigram language model over phones, estimated from the training set of TIMIT.
Evaluation

- Using the generative objective, PER percentages are:

<table>
<thead>
<tr>
<th></th>
<th>RBM</th>
<th>CRBM</th>
<th>ICRBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>36.9 %</td>
<td>42.7 %</td>
<td>39.3 %</td>
</tr>
</tbody>
</table>

- Using the hybrid objective function:

<table>
<thead>
<tr>
<th></th>
<th>RBM</th>
<th>ICRBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.5 %</td>
<td>26.7 %</td>
</tr>
</tbody>
</table>
Evaluation

• Comparison with feedforward neural networks

<table>
<thead>
<tr>
<th></th>
<th>NN (random weights)</th>
<th>NN (RBM weights)</th>
<th>ICRBM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28.7 %</td>
<td>28.3 %</td>
<td>26.7 %</td>
</tr>
</tbody>
</table>

• A two-tailed Matched Pairs Sentence-Segment Word Error (MAPSSWE) significance test showed that ICRBM is significantly better.
Using DBNs for phone recognition
(Mohamed, Dahl, Hinton, NIPS workshop 2009)

We employed two types of DBN architectures:
- The BP-DBN: It performs a purely discriminative fine-tuning phase using backpropagation.
- The AM-DBN: It has an RBM associative memory for the final layer to model joint density of labels and inputs. The hybrid objective function is used for fine-tuning.

(a) A 2-layer BP-DBN

(b) A 3-layer AM-DBN
Evaluation: How deep should the model be?
## Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Margin GMM</td>
<td>33.0 %</td>
</tr>
<tr>
<td>ML trained CD-HMM</td>
<td>27.3 %</td>
</tr>
<tr>
<td>ICRBM</td>
<td>26.7 %</td>
</tr>
<tr>
<td>Recurrent NN</td>
<td>26.1 %</td>
</tr>
<tr>
<td>Monophone HTMs</td>
<td>24.8 %</td>
</tr>
<tr>
<td>Heterogeneous Classifiers</td>
<td>24.4 %</td>
</tr>
<tr>
<td><strong>Deep Belief Network (DBN)</strong></td>
<td><strong>23 %</strong></td>
</tr>
<tr>
<td>CD-HMM trained with BMMI (IBM)</td>
<td><strong>22.7%</strong></td>
</tr>
<tr>
<td><strong>DBN with mcRBM as the 1st layer</strong></td>
<td><strong>20.5%</strong></td>
</tr>
</tbody>
</table>
Thank you