Tensor-variate Restricted Boltzmann Machines

Tu Dinh Nguyen, Truyen Tranyz, Dinh Phungy, Svetha Venkateshy

Deakin University, Curtin University, Australia

AAAI 2015

February 26, 2015
Problem in a nutshell: a figure from Hinton\textsuperscript{1}

Figure: (Left) Two views of an unfactored gated RBM. (Right) One factor in a factored gated RBM.
Weights from $\{w_{ijk}\}_{i=1,...,I; j=1,...,J; k=1,...,K}$ to $\{w_{if}, w_{jf}, w_{kf}\}_{f=1,...,F}$. Number of Parameters from $I \times J \times K$ to $F(I + J + k)$.

Problem in a nutshell: a figure in this paper

Figure: Graphical illustrations of RBM (left) and TvRBM (right). The cubic nodes are observed, the sphere nodes are latent. The TvRBM models the 3D input data of $2 \times 4 \times 3$ and the triangular pyramids represent 4-way factors.
Outline

- Review of RBM for vectors
- Tensor notations
- Tensor-variate RBM
  - Model definition
  - $(N + 1)$-way factoring of multiplicative interactions
  - Receptive fields visualization
- Results
Review of RBM for vectors

- Visible data: \( v = (v_1, \ldots, v_M) \).
- Latent representation: \( h = (h_1, \ldots h_K) \in \{0, 1\}^K \).
- Energy function:

\[
E(v, h; \psi) = -[\mathcal{F}(v) + a^\top v + b^\top h + v^\top Wh],
\]

where \( a \in \mathbb{R}^M, b \in \mathbb{R}^K \) are bias, \( W \in \mathbb{R}^{M \times K} \) is the mapping parameters (weights), \( \psi = \{a, b, W\} \) and \( \mathcal{F}(v) \) is type-specific function:

- Binary input: \( \mathcal{F}(v) = 0 \).
- Gaussian variables: \( \mathcal{F}(v) = -0.5 \sum_m v_m^2 \).
- Boltzmann distribution: \( p(v, h; \psi) \propto \exp[-E(v, h; \psi)] \)
- Conditional distributions over hidden and visible units:

\[
p(v|h; \psi) = \prod_{m=1}^{M} p(v_m|h; \psi) \quad p(h|v; \psi) = \prod_{k=1}^{K} p(h_k|v; \psi)
\]
Tensor notations

- **N-mode tensor**: $\mathcal{T} \in \mathbb{R}^{D_1:N}$, where $D_1:N = D_1 \times D_2 \times \cdots \times D_N$ is the product space over $N$ dimensions.
- **Rank-one N-mode tensor**
  \[
  \mathcal{T} = x^{(1)} \circ x^{(2)} \circ \cdots \circ x^{(N)},
  \]
  where $\circ$ denotes the outer product.
- Let $\bar{\times}_n$ indicate the $n$-th mode product which is the multiplication of a vector with a tensor along mode $n$, resulting in a $(N - 1)$-mode tensor:
  \[
  \mathcal{X} = \mathcal{T} \bar{\times}_n t
  \]
  where $t \in \mathbb{R}^{D_n}$, $\mathcal{X} \in \mathbb{R}^{D_{\neg n}}$ and $D_{\neg n} = D_{\{1:N\}\setminus n}$ denotes the product space over $N$ dimensions excluding $D_n$.
- **Inner product**: sum of the element-wised product
  \[
  \langle \mathcal{X}, Y \rangle = \sum_{d_1=1}^{D_1} \cdots \sum_{d_N=1}^{D_N} x_{d_1}d_2\cdots d_N y_{d_1}d_2\cdots d_N.
  \]
Tensor RBM: model 1/2

- Visible units: $N$-mode tensor: $\mathcal{V} \in \mathbb{R}^{D_1:N}$
- Hidden units $\mathbf{h}$: same as RBM
- Goal: model the joint distribution $p(\mathcal{V}, \mathbf{h})$
- Energy function:
  \[
  E(\mathcal{V}, \mathbf{h}) = -[\mathcal{F}(\mathcal{V}) + \langle \mathcal{A}, \mathcal{V} \rangle + \mathbf{b}^\top \mathbf{h} + \langle \mathcal{V}, \mathcal{W} \bar{x}_{N+1} \mathbf{h} \rangle],
  \]
  - $\mathcal{F}(\mathcal{V})$ is type-specific function.
  - $\mathcal{A} \in \mathbb{R}^{D_1:N}$ are visible biases.
  - $\mathcal{W} \in \mathbb{R}^{D_1:N \times K}$ are mapping parameters.
- Hidden posterior
  \[
  p(h_k = 1|\mathcal{V}) = \sigma[b_k + \langle \mathcal{V}, \mathcal{W} \bar{x}_{N+1} \mathbf{1}_K \rangle]
  \]
  where $\mathbf{1}_K^k$ is one-hot representation of $K$-length vector with all zeros but 1 at $k^{th}$ position.
Tensor RBM: model 2/2

The generative distribution $p(\mathcal{V} | \mathbf{h})$ is type-specific. Let

$$G_{d_1d_2...d_N}(\mathbf{h}) = [\mathcal{W} \times_1 1_{D_1}^{d_1} \times_2 1_{D_2}^{d_2} \times_3 ... \times_N 1_{D_N}^{d_N}], \quad (8)$$

- For binary input $\mathcal{F}(\mathcal{V}) = 0$,

$$p(\nu_{d_1d_2...d_N} = 1 | \mathbf{h}) = \sigma(a_{d_1d_2...d_N} + G_{d_1d_2...d_N}(\mathbf{h})) \quad (9)$$

- For Gaussian input, assuming unit variance, i.e., $\mathcal{F}(\mathcal{V}) = -0.5\langle \mathcal{V}, \mathcal{V} \rangle$,

$$p(\nu_{d_1d_2...d_N} = 1 | \mathbf{h}) = \mathcal{N}\left[\{a_{d_1d_2...d_N} + G_{d_1d_2...d_N}(\mathbf{h})\}; 1_{D_1:N}^{D_1:N}\right], \quad (10)$$

where $1_{D_1:N}^{D_1:N}$ is the tensor where all elements are 1.
\((N+1)\)-way factoring of multiplicative interactions

Problem of the tensor RBM model:

\[
E(\mathcal{V}, h) = -[\mathcal{F}(\mathcal{V}) + \langle \mathcal{A}, \mathcal{V} \rangle + b^\top h + \langle \mathcal{V}, \mathcal{W} \times_{N+1} h \rangle],
\]

(11)

Excessively large number of free parameters which scales as the product of data mode and hidden dimensions.

The \((N+1)\)-mode mapping tensor \(\mathcal{W}\) has \(K \prod_n D_n\) elements.

- Solution: \((N+1)\)-way factoring,

\[
\mathcal{W} = \sum_{f=1}^{F} \lambda_f w_{f}^{(1)} \circ \cdots \circ w_{f}^{(N)} \circ w_{h}^{f},
\]

(12)

- \(\lambda \in \mathbb{R}^F\) is the scaling factor,
- matrix \(W^{(n)} \in \mathbb{R}^{D_n \times F}\) represents the mode-factor weights,
- \(W^h \in \mathbb{R}^{K \times F}\) is the hidden-factor.

- If \(\lambda = 1\),

\[
w_{d_1d_2...d_Nk} = \sum_{f=1}^{F} \sum_{d_1d_2...d_Nk} w_{d_1f}^{(1)} \cdots w_{d_Nf}^{(N)} w_{h}^{f}. \]

(13)
Receptive field visualization

- From the above \((N + 1)\)-way factoring, the number of the mapping parameters is drastically reduced from \(K \prod_n D_n\) to \(F(K + \sum_n D_n)\), which grows linearly rather than exponentially in \(N\).

- Parameter learning learning is similar to RBM (Contrastive Divergence is used).

- The model generalization quality is manifested by the weights connecting the hidden unit to the visible units. However, in the TvRBM, there is no explicit connection weights. For \(n\)-th mode of the data,

\[
\begin{align*}
\mathbf{R} &= \sum_{f=1}^{F} x_f y_f \left[ \mathbf{w}_f^h \left( \mathbf{w}_f^{(n)} \right)^\top \right] \quad (14) \\
x_f &= \left( (\mathbf{1}^{D_1})^\top \mathbf{w}_f^{(1)} \right) \ldots \left( (\mathbf{1}^{D_{n-1}})^\top \mathbf{w}_f^{(n-1)} \right) \quad (15) \\
y_f &= \left( (\mathbf{1}^{D_{n+1}})^\top \mathbf{w}_f^{(n+1)} \right) \ldots \left( (\mathbf{1}^{D_N})^\top \mathbf{w}_f^{(N)} \right) \quad (16)
\end{align*}
\]
The model is **not** a deep model. $F = 100$, $K = 500$.

1-nearest neighbors (1-NN) with cosine similarity measures is used for classification.

For MNIST, the augmented data include each image with 15 distorted versions of varying degree: 4 zooms, 7 rotations and 4 horizontal shears.
Visualizing the learnt receptive fields

Figure: The learnt receptive fields of RBM (left) and TvRBM (right) for facial images with different lighting conditions. The illumination variations are captured in TvRBM filters.
Results of EEG data

<table>
<thead>
<tr>
<th>Method</th>
<th>5%</th>
<th>10%</th>
<th>25%</th>
<th>50%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel</td>
<td>52.78</td>
<td>41.67</td>
<td>38.89</td>
<td>37.24</td>
<td>36.11</td>
</tr>
<tr>
<td>Tucker</td>
<td>52.78</td>
<td>44.44</td>
<td>44.44</td>
<td>38.89</td>
<td>33.33</td>
</tr>
<tr>
<td>PARAFAC</td>
<td>58.33</td>
<td>52.78</td>
<td>52.78</td>
<td>48.67</td>
<td>44.44</td>
</tr>
<tr>
<td>RBM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TvRBM</td>
<td>47.22</td>
<td>36.11</td>
<td>27.78</td>
<td>25.00</td>
<td>19.44</td>
</tr>
</tbody>
</table>

Table 2: The control/alcoholic classification performance on testing set of EEG data (Zhang et al. 1995) with different portion of training data. The RBM fails to learn from the excessive number of pixels per trial reading.

- Tensor size $64 \times 64 \times 64$. $F = 100$, $K = 200$.
- Method means training size.