Diversifying Restricted Boltzmann Machine for Document Modeling

Pengtao Xie

Joint work with Yuntian Deng and Eric Xing
Carnegie Mellon University
Document Modeling

Politics
Animal
Food
Document Modeling

\[ \alpha \rightarrow \theta \rightarrow z \rightarrow w \]

\[ \beta_K \]

\[ h_1 \rightarrow h_2 \rightarrow \ldots \rightarrow h_K \]

\[ v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow \ldots \rightarrow v_J \]

\[ y_1 \rightarrow y_2 \]

\[ z_1 \rightarrow z_2 \rightarrow z_3 \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \]
Restricted Boltzmann Machine

$$E(h, v) = - \sum_{j=1}^{J} \alpha_j v_j - \sum_{k=1}^{K} \beta_k h_k - \sum_{j=1}^{J} \sum_{k=1}^{K} A_{jk} v_j h_k$$
Popularity of Topics

Power-law distribution

Dominant Topics
- Politics
- Economics
- Sports

Long-tail Topics
- Garden
- Animal
- Furniture
- Tour
- Flower
- Food
RBM is insufficient to capture long-tail topics
Long-tail topics are important

- Amount is large
- Arguably more interesting
  - Example: in advertisement, a “lose weight” topic is more important than a “time” topic
Diversification
Diversity Regularized RBM

- **Goal:** encourage the latent factors to spread out to improve the coverage of long-tail topics

- **Approach:**
  - Define a metric to measure the diversity of latent factors
  - Use the diversity metric to regularize the learning of latent factors
Diversity Metric

- Measure the dissimilarity between two vectors
- Measure the diversity of a vector set
Dissimilarity between two vectors

- Invariant to scale, translation, rotation and orientation of the two vectors
- Euclidean distance, L1 distance
  - Variant to scale
- Cosine similarity
  - Variant to orientation
Dissimilarity between two vectors

- Non-obtuse angle

- Invariant to scale, translation, rotation and orientation of the two vectors

- Definition

\[ \theta = \arccos \left( \frac{x \cdot y}{\|x\| \|y\|} \right) \]
Measure the diversity of a vector set

- Based on the pairwise dissimilarity measure between vectors
- The diversity of a set of vectors $\mathbf{A} = \{ \mathbf{a}_i \}_{i=1}^K$ is defined as

$$\Omega(\mathbf{A}) = \text{mean}(\Theta) - \text{var}(\Theta)$$

where

$$\Theta = \{ \theta_{ij} \}_{i=1}^K, j=1 \quad \theta_{ij} = \arccos \left( \frac{|\mathbf{a}_i \cdot \mathbf{a}_j|}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \right)$$

- Mean: summarize how these vectors are different from each other on the whole
- Variance: encourage the vectors to evenly spread out
Diversity Regularized RBM

\[
\max_A L(D; A) + \lambda \Omega(A)
\]

\[
\Omega(A) = \text{mean}(\Theta) - \text{var}(\Theta)
\]

\[
\Theta = \left\{ \theta_{ij} \right\}_{i=1, j=1}^{K, j=K} \quad \theta_{ij} = \arccos \left( \frac{a_i \cdot a_j}{\|a_i\| \|a_j\|} \right)
\]
Optimization

Reparametrize

\[ \mathbf{A} = \text{diag}(\mathbf{g}) \tilde{\mathbf{A}} \quad g_i = \|\mathbf{a}_i\| \]

\[
\max_{g, \tilde{\mathbf{A}}} \quad L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda \Omega(\tilde{\mathbf{A}})
\]

\[
\text{s.t.} \quad \forall i, \|\tilde{\mathbf{a}}_i\| = 1, \quad g_i \geq 0
\]

- **Fix \( \tilde{\mathbf{A}} \), optimize \( g \)**

\[
\max_g \quad L(D; \mathbf{g}\tilde{\mathbf{A}})
\]

\[
\text{s.t.} \quad \forall i, \quad g_i \geq 0
\]

- **Fix \( \mathbf{g} \), optimize \( \tilde{\mathbf{A}} \)**

\[
\max_{\tilde{\mathbf{A}}} \quad L(D; \mathbf{g}\tilde{\mathbf{A}}) + \lambda \Omega(\tilde{\mathbf{A}})
\]

\[
\text{s.t.} \quad \forall i, \|\tilde{\mathbf{a}}_i\| = 1
\]
Optimization

\[
\max_{\tilde{A}} \quad L(D; \mathbf{g}\tilde{A}) + \lambda \Omega(\tilde{A}) \\
\text{s.t.} \quad \forall i, \|\tilde{a}_i\| = 1
\]

Lower bound

\[
\Omega(\tilde{A}) \geq \Gamma(\tilde{A}) = \arcsin(\sqrt{\det(\tilde{A}^T\tilde{A})}) - \left(\frac{\pi}{2} - \arcsin(\sqrt{\det(\tilde{A}^T\tilde{A})})\right)^2
\]

\[
\max_{\tilde{A}} \quad L(D; \mathbf{g}\tilde{A}) + \lambda \Gamma(\tilde{A}) \\
\text{s.t.} \quad \forall i, \|\tilde{a}_i\| = 1
\]
Theorem

- Maximizing the lower bound with projected gradient ascent (PGA) can increase the diversity metric

**Theorem 1.** Let $\mathbf{G}^{(t)}$ be the gradient of $\Gamma(\widetilde{\mathbf{A}})$ w.r.t $\widetilde{\mathbf{A}}^{(t)}$ at iteration $t$. $\exists \tau > 0$, such that $\forall \eta \in (0, \tau)$, $\Omega(\widetilde{\mathbf{A}}^{(t+1)}) \geq \Omega(\widetilde{\mathbf{A}}^{(t)})$, where $\widetilde{\mathbf{A}}^{(t+1)} = \mathcal{P}(\widetilde{\mathbf{A}}^{(t)} + \eta \mathbf{G}^{(t)})$ and $\mathcal{P}(\cdot)$ denotes the projection to the unit sphere.

- Maximizing the lower bound with PGA can increase the mean of the angles
- Maximizing the lower bound with PGA can reduce the variance of the angles
Geometry Interpretation

- The gradient of the lower bound w.r.t $a_i$ is in the orthogonal complement of the space spanned by $\{a_1, a_2, \ldots, a_K\}/\{a_i\}$

**Lemma 2.** Let the weight vector $\tilde{a}_i$ of hidden unit $i$ be decomposed into $\tilde{a}_i = x_i + l_i e_i$, where $x_i = \sum_{j=1, j \neq i}^{K} \alpha_j \tilde{a}_j$ lies in the subspace $L$ spanned by $\{\tilde{a}_1, \ldots, \tilde{a}_K\}\setminus\{\tilde{a}_i\}$, $e_i$ is in the orthogonal complement of $L$, $\|e_i\| = 1$, $e_i \cdot \tilde{a}_i > 0$, $l_i$ is a scalar. Then the gradient of $\Gamma(\tilde{A})$ w.r.t $a_i$ is $k_i e_i$, where $k_i$ is a positive scalar.
Geometry Interpretation

- The gradient of the lower bound w.r.t $a_i$ is in the orthogonal complement of the space spanned by $\{a_1, a_2, \ldots, a_K\}/\{a_i\}$

**Lemma 2.** Let the weight vector $\tilde{a}_i$ of hidden unit $i$ be decomposed into $\tilde{a}_i = x_i + l_i e_i$, where $x_i = \sum_{j=1, j \neq i}^{K} \alpha_j \tilde{a}_j$ lies in the subspace $L$ spanned by $\{\tilde{a}_1, \ldots, \tilde{a}_K\}\{\tilde{a}_i\}$, $e_i$ is in the orthogonal complement of $L$, $\|e_i\| = 1$, $e_i \cdot \tilde{a}_i > 0$, $l_i$ is a scalar. Then the gradient of $\Gamma(\tilde{A})$ w.r.t $a_i$ is $k_i e_i$, where $k_i$ is a positive scalar.
Experiments

- **Datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#categories</th>
<th>#samples</th>
<th>vocab. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT</td>
<td>30</td>
<td>9394</td>
<td>5000</td>
</tr>
<tr>
<td>20-News</td>
<td>20</td>
<td>18846</td>
<td>5000</td>
</tr>
<tr>
<td>Reuters</td>
<td>9</td>
<td>7195</td>
<td>5000</td>
</tr>
</tbody>
</table>

- **Baselines**

  - Bag-of-Words (BOW); Latent Dirichlet Allocation (LDA); LDA regularized with Determinantal Point Process prior (DPP-LDA); Pitman-Yor Process Topic Model (PYTM); Latent IBP Compound Dirichlet Allocation (LIDA); Neural Autoregressive Topic Model (DocNADE); Paragraph Vector (PV); Restricted Boltzmann Machine

- **Evaluation**

  - Retrieval: precision@100
  - Clustering: accuracy
  - Perplexity
  - Qualitative evaluation
## Retrieval Precision

<table>
<thead>
<tr>
<th>Method</th>
<th>TDT</th>
<th>20-News</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>40.9</td>
<td>7.4</td>
<td>69.3</td>
</tr>
<tr>
<td>LDA</td>
<td>79.4</td>
<td>19.6</td>
<td>68.5</td>
</tr>
<tr>
<td>DPP-LDA</td>
<td>81.9</td>
<td>18.2</td>
<td>69.9</td>
</tr>
<tr>
<td>PYTM</td>
<td>78.7</td>
<td>20.1</td>
<td>70.6</td>
</tr>
<tr>
<td>LIDA</td>
<td>77.9</td>
<td>21.8</td>
<td>71.4</td>
</tr>
<tr>
<td>DocNADE</td>
<td>80.3</td>
<td>16.8</td>
<td>72.6</td>
</tr>
<tr>
<td>PV</td>
<td>81.7</td>
<td>19.1</td>
<td>76.9</td>
</tr>
<tr>
<td>RBM</td>
<td>47.4</td>
<td>22.3</td>
<td>70.1</td>
</tr>
<tr>
<td>DRBM</td>
<td>84.2</td>
<td>24.9</td>
<td>75.9</td>
</tr>
</tbody>
</table>
Clustering Accuracy

Accuracy on TDT Dataset

Accuracy on 20-News Dataset

Accuracy on Reuters Dataset
# Clustering Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>TDT</th>
<th>20-News</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>51.3</td>
<td>21.3</td>
<td>49.7</td>
</tr>
<tr>
<td>LDA</td>
<td>45.2</td>
<td>21.9</td>
<td>51.2</td>
</tr>
<tr>
<td>DPP-LDA</td>
<td>46.3</td>
<td>10.9</td>
<td>49.3</td>
</tr>
<tr>
<td>PYTM</td>
<td>46.9</td>
<td>21.5</td>
<td>51.7</td>
</tr>
<tr>
<td>LIDA</td>
<td>47.3</td>
<td>17.4</td>
<td>53.1</td>
</tr>
<tr>
<td>DocNADE</td>
<td>45.7</td>
<td>18.7</td>
<td>48.7</td>
</tr>
<tr>
<td>PV</td>
<td>48.2</td>
<td>24.3</td>
<td>52.8</td>
</tr>
<tr>
<td>RBM</td>
<td>23.3</td>
<td>22.7</td>
<td>47.6</td>
</tr>
<tr>
<td>DRBM</td>
<td><strong>52.4</strong></td>
<td><strong>29.4</strong></td>
<td><strong>60.9</strong></td>
</tr>
</tbody>
</table>
Qualitative Evaluation

### Exemplar Topics Learned by RBM and DRBM

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>president</td>
<td>iraq</td>
<td>iraq</td>
<td>olympic</td>
<td>spkr</td>
<td>president</td>
<td>olympic</td>
<td>iraq</td>
<td>lawyers</td>
<td>students</td>
</tr>
<tr>
<td>clinton</td>
<td>united</td>
<td>un</td>
<td>games</td>
<td>voice</td>
<td>iraq</td>
<td>games</td>
<td>united</td>
<td>kaczynski</td>
<td>japanese</td>
</tr>
<tr>
<td>iraq</td>
<td>un</td>
<td>iraqi</td>
<td>nagano</td>
<td>tobacco</td>
<td>clinton</td>
<td>olympics</td>
<td>un</td>
<td>ms</td>
<td>japan</td>
</tr>
<tr>
<td>united</td>
<td>weapons</td>
<td>lewinsky</td>
<td>olympics</td>
<td>olympic</td>
<td>united</td>
<td>nagano</td>
<td>weapons</td>
<td>defense</td>
<td>school</td>
</tr>
<tr>
<td>spkr</td>
<td>iraqi</td>
<td>saddam</td>
<td>game</td>
<td>games</td>
<td>million</td>
<td>team</td>
<td>iraqi</td>
<td>trial</td>
<td>ms</td>
</tr>
<tr>
<td>house</td>
<td>nuclear</td>
<td>clinton</td>
<td>team</td>
<td>people</td>
<td>lewinsky</td>
<td>gold</td>
<td>baghdad</td>
<td>judge</td>
<td>united</td>
</tr>
<tr>
<td>people</td>
<td>india</td>
<td>baghdad</td>
<td>gold</td>
<td>olympics</td>
<td>thailand</td>
<td>game</td>
<td>council</td>
<td>people</td>
<td>yen</td>
</tr>
<tr>
<td>lewinsky</td>
<td>minister</td>
<td>inspectors</td>
<td>japan</td>
<td>nagano</td>
<td>spkr</td>
<td>hockey</td>
<td>inspectors</td>
<td>prosecutor</td>
<td>gm</td>
</tr>
<tr>
<td>government</td>
<td>saddam</td>
<td>weapons</td>
<td>medal</td>
<td>game</td>
<td>government</td>
<td>medal</td>
<td>nations</td>
<td>kaczynskis</td>
<td>tokyo</td>
</tr>
<tr>
<td>white</td>
<td>military</td>
<td>white</td>
<td>hockey</td>
<td>gold</td>
<td>jones</td>
<td>winter</td>
<td>military</td>
<td>government</td>
<td>south</td>
</tr>
</tbody>
</table>
Sensitivity of DRBM to tradeoff parameter $\lambda$ on (a) TDT dataset (b) 20-News dataset (c) Reuters dataset
Conclusions

• **Problem**
  • The popularity of topics is distributed in a power-law fashion
  • Standard RBM is insufficient to capture long-tail topics

• **Solution**
  • Diversify the hidden units in RBM to improve the coverage of long-tail topics
  • Define an angle based diversity regularizer
  • Optimization

• **Results**
  • Experiments on document retrieval and clustering demonstrate the effectiveness of the diversity regularizer
Thank you!
Questions?