Motivation
Multimodal learning prefers models can
- extracting a joint representation
- filling in missing modality
- exploiting the connections between different data modalities
However, most existing methods
- fall short of extracting interpretable multilayer hidden structures
- have trouble visualizing the relationships between modalities
- need to normalize scales of input data
Thus, we propose a novel deep multimodal model whose latent multilayer network can be easily interpreted based on Poisson Gamma Belief Network which can be represented as deep LDA equivalently.

Background

Priors: $y_j^{(l)} \sim \text{Dir}(\nu_0, \nu_1, \ldots, \nu_k)$, $\alpha_j \sim \text{Beta}(\alpha_0, \alpha_1)$, $\beta_j \sim \text{Gam}(\alpha_j, 1/\beta_j)$.

Different Data Formulation
If the observations are high-dimensional sparse binary vectors $b_j^{(l)} \in \{0,1\}^V$, they are factorized as $b_j^{(l)} = (x_j^{(l)})^2 \geq 0$, $x_j^{(l)} \sim \text{Poisson}(\theta_j^{(l)})$.

Contributions
We construct a novel multimodal PGBN that well captures the correlations between different modalities at multiple levels of abstraction and these coupled topics visualized by our structure exhibit an increasing level of abstraction when moving towards a deeper hidden layer.

Multimodal PGBN
From the top to bottom, the generative model is expressed as $b_j^{(l)} \sim \text{Gam}(\gamma_j^{(l)}, 1/\gamma_j^{(l-1)})$, $b_j^{(l)} \sim \text{Gam}(\theta_j^{(l)}, 1/\theta_j^{(l-1)})$, $x_j^{(l)} \sim \text{Poisson}(\theta_j^{(l)})$, $x_j^{(l)} \sim \text{Poisson}(\theta_j^{(l)})$.

Adaptive Normalization
To handle different input data scales, we propose to modify the mPGBN model as following

$$x_j^{(l)} \sim \text{Poisson}(\theta_j^{(l)})$$, $x_j^{(l)} \sim \text{Poisson}(\theta_j^{(l)})$,

which means that the first hidden layers of both modalities only share their gamma shape parameters but have adaptive scale parameters to suit different input scales.

Experiment Results

Example Topics for Text

Example Themes for Image

Visualizing the generative process of input image-tags pair

Filling Missing Modality

Table 1: Comparison of AP scores and Precision@50 of various multimodal models on the MIR-Flickr dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>MIR-Flickr AP</th>
<th>MIR-Flickr Precision@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.120</td>
<td>0.405</td>
</tr>
<tr>
<td>LDA</td>
<td>0.125</td>
<td>0.450</td>
</tr>
<tr>
<td>DNN</td>
<td>0.128</td>
<td>0.470</td>
</tr>
<tr>
<td>SVM</td>
<td>0.132</td>
<td>0.500</td>
</tr>
<tr>
<td>DPP</td>
<td>0.134</td>
<td>0.520</td>
</tr>
<tr>
<td>SSP</td>
<td>0.135</td>
<td>0.520</td>
</tr>
<tr>
<td>PNN</td>
<td>0.135</td>
<td>0.520</td>
</tr>
<tr>
<td>MIR-Flickr</td>
<td>0.135</td>
<td>0.520</td>
</tr>
</tbody>
</table>

Figure 4: Examples of text generated by Multimodal PGBN conditioned on images.

Figure 5: Top-5 nearest images retrieved using the features generated by the multimodal PGBN conditioning on the tags.

AAAI 2018: Thirty-Second AAAI Conference on Artificial Intelligence