Matching Words and Pictures

By Kobus Barnard, et al.

Jaewoo Pi
Before We Start

Little Formality
Insight
Annotation

Multi-Modal Hierarchical Models

Mixture of Multi-Modal Latent Dirichlet Allocation

Correspondence

Linking Word Emission & Region Emission with Mixed Weights

Paired Word and Region Emission at Nodes
Hierarchical Models

Original Version, by Hofmann

\[
p(D | d) = \sum_c p(c) \prod_{i \in D} \left( \sum_l P(i | l, c)P(l | d) \right)
\]
Multi-modal Hierarchical Models

**NODE**
Image and co-occurring text

**Higher Node**
General

**Lower Node**
Specific
Multi-modal Hierarchical Models

Process of Generating set of Observation

\[ p(D|d) = \sum_c p(c) \prod_{w \in W} \left[ \sum_{\{sth\}} \right] \prod_{b \in B} \left[ \sum_{\{sth\}} \right] \]

\[ \prod_{w \in W} \left[ \sum p(w|l, c) p(l|d) \right] \frac{N_w}{N_{w,d}} \prod_{b \in B} \left[ \sum p(b|l, c) p(l|d) \right] \frac{N_b}{N_{b,d}} \]

Word Emission Prob. (Freq. Table)
Region Emission Prob. (Gaussian distribution)
Multi-modal Hierarchical Models

How to predict word based on image?

\[ p(w|B) = \sum_c p(c) \left[ \sum_l p(w|l, c)p(l|c) \right] \prod_{b\in B} \left[ \sum_l p(b|l, c)p(l|c) \right] \frac{N_b}{N_{b,d}} \]
Annotation

Multi-Modal Hierarchical Models

Mixture of Multi-Modal Latent Dirichlet Allocation
LDA

- Dirichlet Parameter
- Per-document Topic Proportions
- Per-word Topic Assignment
- Observed Word
LDA

- Dirichlet Parameter
- Per-document Topic Proportions
- Per-word Topic Assignment
- Observed Word
MoM-LDA

Per-region
Topic Assignment

Observed
Blob

Per-document
Topic Proportions

Per-word
Topic Assignment

Observed
Word

Dirichlet
Parameter

C

θ

z

w

s

b

Diagram of MoM-LDA with nodes representing different parameters and variables.
How Do We Associate?

[Estimation Maximization]
Correspondence

Linking Word & Region Emission Prob. With Mix-weights

Paired Word and Region Emission at Nodes
Idea of These Two

Considering Surroundings

Strengthening words and regions relationship

By giving \textit{pre-computed} weight
\[ p(D|d) = \sum_c p(c) \prod_{w \in W} \left[ \sum \left\{ \text{sth} \right\} \right] \prod_{b \in B} \left[ \sum \left\{ \text{sth} \right\} \right] \]

\[ \prod_{w \in W} \left[ \sum p(w|l, c)p(l|B, c, d) \right] \frac{N_w}{N_{w,d}} \prod_{b \in B} \left[ \sum p(b|l, c)p(l|d) \right] \frac{N_b}{N_{b,d}} \]

**Word Emission Prob.**  (Freq. Table)  
**Blob Emission Prob.**  (Gaussian distribution)
\[
p(D|d) = \sum_c p(c) \prod_{w \in W} \left[ \sum_{\{st\}} \prod_{b \in B} \left[ \sum_{\{st\}} \right] \right] \\
\prod_{w \in W} \left[ \sum_{\{st\}} p(w|l, c)p(l|B, c, d) \right] \frac{N_w}{N_{w,d}} \prod_{b \in B} \left[ \sum_{\{st\}} p(b|l, c)p(l|d) \right] \frac{N_b}{N_{b,d}}
\]

Word Blob

**Vertical Mixture Weights**

(Pre-computed)

\[
p(l|B, c, d) \propto \sum_{b \in B} p(l|b, c, d)
\]
Pair Word and Region Emission

“tightens the relationship between regions and words”

\[
p(D|d) = \sum_{c} p(c) \prod_{(w,b) \in D} \left( \sum_{l} P((w,b)|l,c)P(l|d) \right)
\]
Criticque
<table>
<thead>
<tr>
<th>Method</th>
<th>Training data</th>
<th>Held out data</th>
<th>Novel data</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear-I-0-doc-vert</td>
<td>1.235 (0.02)</td>
<td>0.688 (0.02)</td>
<td>0.258 (0.01)</td>
</tr>
<tr>
<td>binary-I-0-ave-vert</td>
<td>1.210 (0.03)</td>
<td>0.563 (0.02)</td>
<td>0.060 (0.01)</td>
</tr>
<tr>
<td>binary-I-0-doc-vert</td>
<td>1.385 (0.02)</td>
<td>0.587 (0.02)</td>
<td>0.061 (0.02)</td>
</tr>
<tr>
<td>binary-I-0-region-cluster</td>
<td>1.429 (0.03)</td>
<td>0.651 (0.02)</td>
<td>0.094 (0.02)</td>
</tr>
<tr>
<td>binary-I-0-region-only</td>
<td>1.061 (0.02)</td>
<td>0.684 (0.02)</td>
<td>0.160 (0.02)</td>
</tr>
<tr>
<td>binary-I-2-ave-vert</td>
<td>1.367 (0.03)</td>
<td>0.608 (0.02)</td>
<td>0.084 (0.01)</td>
</tr>
<tr>
<td>binary-I-2-doc-vert</td>
<td>1.320 (0.03)</td>
<td>0.627 (0.02)</td>
<td>0.129 (0.01)</td>
</tr>
<tr>
<td>binary-I-2-region-cluster</td>
<td>1.342 (0.03)</td>
<td>0.694 (0.02)</td>
<td>0.156 (0.01)</td>
</tr>
<tr>
<td>binary-I-2-region-only</td>
<td>1.016 (0.02)</td>
<td>0.709 (0.02)</td>
<td>0.211 (0.01)</td>
</tr>
<tr>
<td>linear-D-0-doc-vert</td>
<td>1.376 (0.02)</td>
<td>0.714 (0.02)</td>
<td>0.268 (0.01)</td>
</tr>
<tr>
<td>binary-D-0-ave-vert</td>
<td>1.169 (0.03)</td>
<td>0.550 (0.02)</td>
<td>0.057 (0.01)</td>
</tr>
<tr>
<td>binary-D-0-doc-vert</td>
<td>1.417 (0.03)</td>
<td>0.601 (0.02)</td>
<td>0.074 (0.01)</td>
</tr>
<tr>
<td>binary-D-0-region-cluster</td>
<td>1.466 (0.03)</td>
<td>0.669 (0.02)</td>
<td>0.105 (0.02)</td>
</tr>
<tr>
<td>binary-D-0-region-only</td>
<td>1.086 (0.02)</td>
<td>0.700 (0.02)</td>
<td>0.175 (0.02)</td>
</tr>
<tr>
<td>binary-D-2-ave-vert</td>
<td>1.310 (0.005)</td>
<td>0.627 (0.003)</td>
<td>0.089 (0.005)</td>
</tr>
<tr>
<td>binary-D-2-doc-vert</td>
<td>1.589 (0.005)</td>
<td>0.674 (0.003)</td>
<td>0.102 (0.005)</td>
</tr>
<tr>
<td>binary-D-2-region-cluster</td>
<td>1.613 (0.005)</td>
<td>0.739 (0.003)</td>
<td>0.132 (0.005)</td>
</tr>
<tr>
<td>binary-D-2-region-only</td>
<td>1.155 (0.005)</td>
<td>0.747 (0.003)</td>
<td>0.180 (0.005)</td>
</tr>
<tr>
<td>linear-C-0-region-only</td>
<td>0.980 (0.02)</td>
<td>0.472 (0.02)</td>
<td>0.106 (0.01)</td>
</tr>
<tr>
<td>binary-C-0-ave-vert</td>
<td>1.020 (0.02)</td>
<td>0.516 (0.02)</td>
<td>0.071 (0.01)</td>
</tr>
<tr>
<td>binary-C-0-doc-vert</td>
<td>1.205 (0.02)</td>
<td>0.541 (0.02)</td>
<td>0.042 (0.01)</td>
</tr>
<tr>
<td>binary-C-0-region-cluster</td>
<td>1.254 (0.02)</td>
<td>0.601 (0.02)</td>
<td>0.104 (0.01)</td>
</tr>
<tr>
<td>binary-C-0-region-only</td>
<td>1.015 (0.02)</td>
<td>0.643 (0.02)</td>
<td>0.179 (0.01)</td>
</tr>
<tr>
<td>discrete-translation</td>
<td>1.347 (0.02)</td>
<td>0.433 (0.002)</td>
<td>-0.072 (0.01)</td>
</tr>
<tr>
<td>MoM-LDA</td>
<td>0.452 (0.01)</td>
<td>0.401 (0.01)</td>
<td>0.171 (0.01)</td>
</tr>
</tbody>
</table>
7.2 Correspondence Results

Figure 6 shows region annotations for a few sample images. For this result we labeled each region with the maximal probability word, using model C-2. In Table 4 we provide quantitative correspondence results computed over 50 images from each of the 10 held out sets. Results for each of the three error measures is provided. For region based word prediction, it is perhaps most reasonable to predict only a few words for each region. This process is most closely studied with the simple keyword prediction error, $E_{PR}^{(model)} - E_{PR}^{(empirical)}$. Here the results suggest that the methods which have been developed to learn correspondence do in fact do better at this task, relative to the performance on the annotation proxy. For example, using the PR measure, linear-C-0-region-only scores 0.067 with the annotation proxy, which is significantly exceeded by the performance of linear-I-0-region-

... we are disappointed that its correspondence performance is still matched by several methods...

... significantly bettered by at least one of them.

and second, the joint probability table may be fitted more accurately because the fitting process should be protected from a large number of outliers caused by forcing each region to correspond to some word. Currently, for both correspondence and annotation, linear-D-0-region-only (same as linear-D-0-doc-vert), appears to be the best overall choice, taking all measures and data sets into account.