Text/Knowledge Embedding and Structured Learning Method

Dong Wang

2015/05/19
Content

• Introduction to embedding
• Ordered embedding
• Text/knowledge joint embedding
• Bi-lingual spherical embedding
• Potential in language learning
Computing With Objects

Feature extraction + regression/classification model
Embedding in Low-dimensional Space

- Relation is more important
- Embed objects in a low-dimensional space, relation preserved!
Embedding Approaches

• Linear embedding: PCA, MDS, bi-linear model,...
• Nonlinear embedding: ISOMAP, SOM, GTM, LLE, kernel PCA,...
• Probabilistic embedding: SNE, t-SNE,...
• Deep learning: RBM, DBM, auto-encoder,...
Word Embedding (1)

http://www.socher.org/index.php/Main/ImprovingWordRepresentationsViaGlobalContextAndMultipleWordPrototypes
Word Embedding (2)

• From neural LM to word vectors

Y. Bengio et al. A Neural Probabilistic Language Model, JMLR, 2003

Ordered Embedding

• Current word embeddings are homogeneous
• Imposing structure constraints leads to better embeddings
• Make more important information in low dimensions!
Enforce Order

• Nested dropout [Oren Rippel, ICML 2014]
• Alpha-decay: larger learning rate with low dimensions
• Lambda-growth: more sparsity with high dimensions
Enforce Order
Sweeping Methods

• Ensure all dimensions fully trained
• Fix converged dimensions, moving the training focus from left to right
• Follow a Chi-square distribution

\[
\alpha(d; r) = \frac{\alpha}{\chi^2(1, 3)} \chi^2(d, M + 3)
\]

\[
\chi^2(x; k) = \begin{cases} 
  x^{(k/2-1)} e^{-x/2} \frac{2^k}{\Gamma(k/2)} & \text{if } x \geq 0; \\
  0, & \text{otherwise}
\end{cases}
\]
Sweeping Method
Knowledge Embedding

• Knowledge in graph is represented by vectors
• Relation in graph is preserved as vector distance

http://agafonovslava.com/post/2011/08/16/Yet-Another-Similar-Items-Visualization
https://twitter.com/jockmcclellan/status/571669444327567361
Embed both Text and Knowledge

- Text from wiki, knowledge from WNet/Yago
- Combine raw knowledge and structured knowledge

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Graph Knowledge</th>
<th>Labelled Text</th>
<th>Raw Text</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNet-N</td>
<td>68569 entities 70040 relations</td>
<td>36519 entities</td>
<td>165MB</td>
<td>Sim-301</td>
</tr>
<tr>
<td>Yago-A</td>
<td>39900 entities 72936 relations</td>
<td>6415 entities</td>
<td>60MB</td>
<td>Animal-143</td>
</tr>
</tbody>
</table>
Embed both Text and Knowledge (2)

\[ L_{txt} = \sum_{e \in E} \sum_{i=1}^{K} \max\{0, \gamma_{txt} - v_{e}^{T} v_{w_{i}} + v_{e}^{T} v_{w'_{i}}\} \]

+ 

\[ L_{grh} = \sum_{P_{i} \in P} \max\{0, \gamma_{grh} - v_{e_{i}}^{T} v_{e_{i}} + v_{e_{i}}^{T} v_{e'_{i}}\} \]
Embed both Text and Knowledge (3)

- Joint learning better than separate learning
- Wnet does not contribute much, while Yago is much stronger
- Domain-specific knowledge the most preferred
Bi-lingual Word Embedding

• Languages share the same semantic concepts
• A linear transform can be used to transfer embeddings from one language to another

Inconsistency with Linear Transform

- Embedding/evaluation is based on cosine distance
- Transform is based on square error

\[ P(w_{i+j} | w_i) = \frac{\exp(c_{w_{i+j}}^T c_{w_i})}{\sum_w \exp(c_w^T c_{w_i})} \neq \min_W \sum_i \|Wx_i - z_i\|^2 \]
Spherical Embedding

- Constraint embedding on sphere
- Unify distance measures to cosine
- Replace square error by cosine error
- Replace linear transform to orthogonal transform

\[
P(w_{i+j} | w_i) = \frac{\exp(c_{w_{i+j}}^T c_{w_i})}{\sum_w \exp(c_w^T c_{w_i})}
\]

\[
\max_W \sum_i (W x_i)^T z_i.
\]
Similarity with Spherical Embedding
# Prediction Accuracy

Accuracy when predicting Spanish from English

<table>
<thead>
<tr>
<th>D-EN</th>
<th>D-ES</th>
<th>P@1</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>300</td>
<td>30.43%</td>
<td>49.43%</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>25.76%</td>
<td>44.29%</td>
</tr>
<tr>
<td>700</td>
<td>700</td>
<td>20.69%</td>
<td>39.12%</td>
</tr>
<tr>
<td>800</td>
<td>200</td>
<td>35.36%</td>
<td>53.96%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D-EN</th>
<th>D-ES</th>
<th>P@1</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>300</td>
<td>38.99%</td>
<td>59.16%</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>39.91%</td>
<td>59.82%</td>
</tr>
<tr>
<td>700</td>
<td>700</td>
<td>41.04%</td>
<td>59.38%</td>
</tr>
<tr>
<td>800</td>
<td>200</td>
<td>40.06%</td>
<td>60.02%</td>
</tr>
</tbody>
</table>

Non-spherical embedding

Spherical embedding

Accuracy when predicting Spanish from English
Visualization for Spherical Embedding

• T-SNE is suitable for Gaussian data
• We propose to visualize spherical high-dimensional data with vMF distributions

\[ f_d(x; \mu, \kappa) = C_d(\kappa)e^{\kappa \mu^t x} \]

\[ p_{ij} = \frac{e^{-||x_i - x_j||^2/2\sigma^2}}{\sum_{m \neq n} e^{-||x_m - x_n||^2/2\sigma^2}} \]
Potential in Language Learning

• Word and knowledge embedding demonstrate “related words”, which may be useful for learning words and phrases
• Visualization offers intuitive presentation for word and knowledge learning
• Bi-lingual transfer may help learn low-resource languages
• Thanks and questions