Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks

arXiv paper
Overview

● Task: Document Classification
  ○ Given a document, assign a label to it

● Contribution:
  ○ Employ word connections to represent the document
  ○ Word connection is model by message passing in graph neural nets
  ○ Graphs are created for each document so it is an inductive setting instead of transductive setting proposed by previous work:
    ■ Previous work (at EMNLP 2019) use global structure where the connection between words are extracted from the entire corpus, restricting them to transductive setting
  ○ Study the contribution of global graph combined with the local graph

● Applications:
  ○ The method is general and could be useful for other graph based document-level models
  ○ The ensembled global-local graph seems to be ineffective so we can extend this work by the idea of hierarchical graph embedding to combine global and local graphs
Model overview
Graph Construction & Encoding

- Unique words in the document are the nodes
- The connections between words are computed by co-occurrence in a window of size 3
- Nodes are embedded randomly
- Word interaction: Neighbor aggregation followed by GRU unit

\begin{align*}
a^t &= A h^{t-1} W_a, \quad (1) \\
z^t &= \sigma \left( W_z a^t + U_z h^{t-1} + b_z \right), \quad (2) \\
r^t &= \sigma \left( W_r a^t + U_r h^{t-1} + b_r \right), \quad (3) \\
\tilde{h}^t &= \text{tanh} \left( W_h a^t + U_h (r^t \odot h^{t-1}) + b_h \right), \quad (4) \\
h^t &= \tilde{h}^t \odot z^t + h^{t-1} \odot (1 - z^t), \quad (5)
\end{align*}
Read out function & ensembled model

- Apply soft-attention on the output of the graph encoder
- Compute max and sum pooling:

\[ h_v = \sigma \left( f_1(h^t_v) \right) \odot \tanh \left( f_2(h^t_v) \right), \quad (6) \]

\[ h_G = \frac{1}{|V|} \sum_{v \in V} h_v + \text{Maxpooling} (h_1...h_V), \quad (7) \]

- Ensemble model: Combine global and local graph with 1:1 vote
- Global graph:
  - Nodes are the same as the original model
  - Edges are computed based on the co-occurrence in the entire training documents
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>R8</th>
<th>R52</th>
<th>Ohsumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN (Non-static)</td>
<td>77.75 ± 0.72</td>
<td>95.71 ± 0.52</td>
<td>87.59 ± 0.48</td>
<td>58.44 ± 1.06</td>
</tr>
<tr>
<td>RNN (Bi-LSTM)</td>
<td>77.68 ± 0.86</td>
<td>96.31 ± 0.33</td>
<td>90.54 ± 0.91</td>
<td>49.27 ± 1.07</td>
</tr>
<tr>
<td>fastText</td>
<td>75.14 ± 0.20</td>
<td>96.13 ± 0.21</td>
<td>92.81 ± 0.09</td>
<td>57.70 ± 0.49</td>
</tr>
<tr>
<td>SWEM</td>
<td>76.65 ± 0.63</td>
<td>95.32 ± 0.26</td>
<td>92.94 ± 0.24</td>
<td>63.12 ± 0.55</td>
</tr>
<tr>
<td>TextGCN</td>
<td>76.74 ± 0.20</td>
<td>-</td>
<td>93.56 ± 0.18</td>
<td>68.36 ± 0.56</td>
</tr>
<tr>
<td>Huang et al. (2019)</td>
<td>-</td>
<td>97.80 ± 0.20</td>
<td>94.60 ± 0.30</td>
<td>69.40 ± 0.60</td>
</tr>
<tr>
<td>TextING</td>
<td>79.82 ± 0.20</td>
<td>98.04 ± 0.25</td>
<td>95.48 ± 0.19</td>
<td>70.42 ± 0.39</td>
</tr>
<tr>
<td>TextING-M</td>
<td>80.19 ± 0.31</td>
<td>98.13 ± 0.12</td>
<td>95.68 ± 0.35</td>
<td>70.84 ± 0.52</td>
</tr>
</tbody>
</table>
Sample Complexity
Interaction steps & Graph density

(a) MR

(b) Obsumed