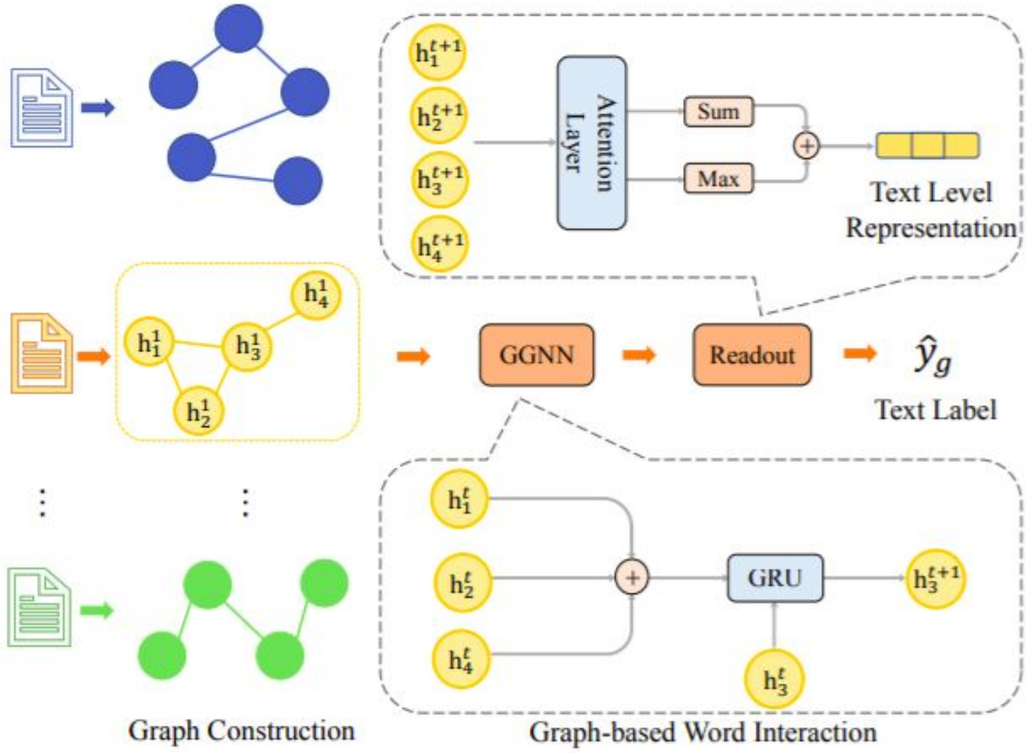
Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks

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

$$\mathbf{a}^t = \mathbf{A}\mathbf{h}^{t-1}\mathbf{W}_a, \tag{1}$$

$$\mathbf{z}^t = \sigma\left(\mathbf{W}_z\mathbf{a}^t + \mathbf{U}_z\mathbf{h}^{t-1} + \mathbf{b}_z\right), \tag{2}$$

$$\mathbf{r}^t = \sigma\left(\mathbf{W}_r\mathbf{a}^t + \mathbf{U}_r\mathbf{h}^{t-1} + \mathbf{b}_r\right), \tag{3}$$

$$\tilde{\mathbf{h}}^t = \tanh\left(\mathbf{W}_h\mathbf{a}^t + \mathbf{U}_h(\mathbf{r}^t \odot \mathbf{h}^{t-1}) + \mathbf{b}_h\right), \tag{4}$$

$$\mathbf{h}^t = \tilde{\mathbf{h}}^t \odot \mathbf{z}^t + \mathbf{h}^{t-1} \odot \left(1 - \mathbf{z}^t\right), \tag{5}$$

Read out function & ensembled model

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

$$\mathbf{h}_v = \sigma\left(f_1(\mathbf{h}_v^t)\right) \odot \tanh\left(f_2(\mathbf{h}_v^t)\right), \tag{6}$$

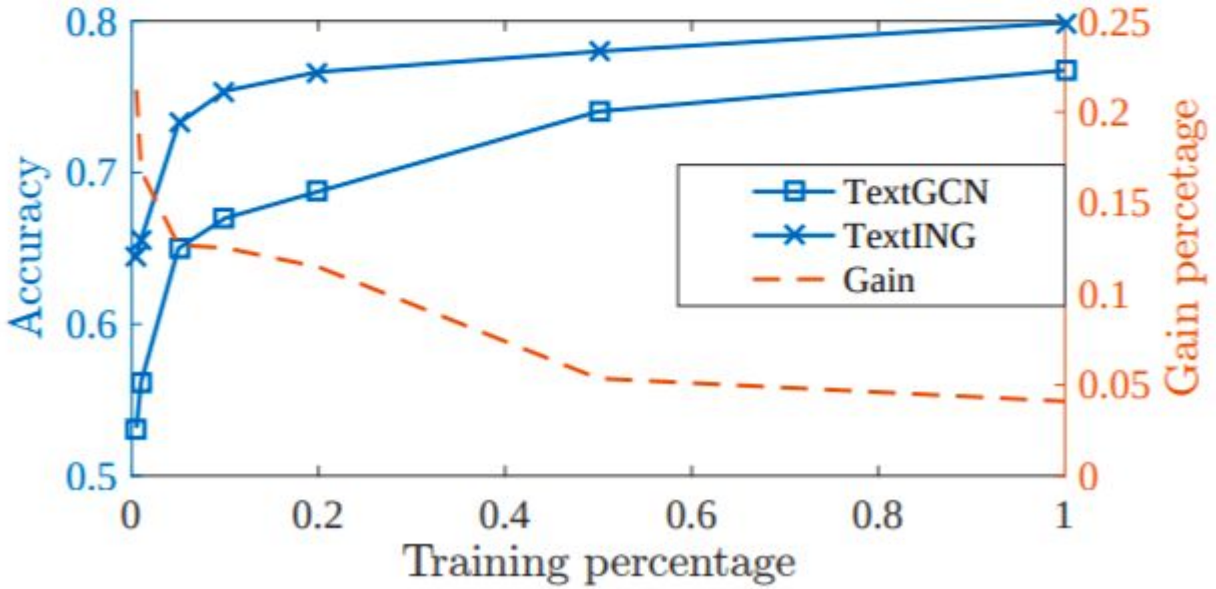
$$\mathbf{h}_{\mathcal{G}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbf{h}_v + \text{Maxpooling}\left(\mathbf{h}_1 ... \mathbf{h}_{\mathcal{V}}\right), \tag{7}$$

- 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

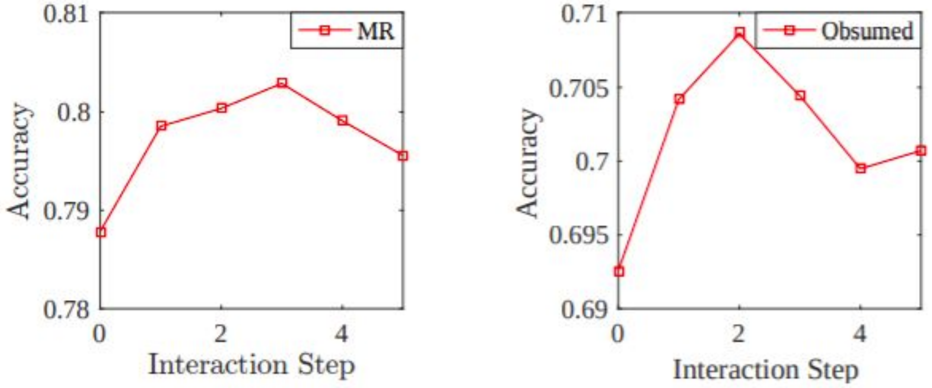
Results

Model	MR	R8	R52	Ohsumed
CNN (Non-static)	77.75 ± 0.72	95.71 ± 0.52	87.59 ± 0.48	58.44 ± 1.06
RNN (Bi-LSTM)	77.68 ± 0.86	96.31 ± 0.33	90.54 ± 0.91	49.27 ± 1.07
fastText	75.14 ± 0.20	96.13 ± 0.21	92.81 ± 0.09	57.70 ± 0.49
SWEM	76.65 ± 0.63	95.32 ± 0.26	92.94 ± 0.24	63.12 ± 0.55
TextGCN	76.74 ± 0.20	97.07 ± 0.10	93.56 ± 0.18	68.36 ± 0.56
Huang et al. (2019)	-	97.80 ± 0.20	94.60 ± 0.30	69.40 ± 0.60
TextING	79.82 ± 0.20	98.04 ± 0.25	95.48 ± 0.19	70.42 ± 0.39
TextING-M	80.19 ± 0.31	98.13 ± 0.12	95.68 ± 0.35	70.84 ± 0.52

Sample Complexity



Interaction steps & Graph density



(a) MR

(b) Ohsumed