Graph based Text Classification

Soumya Jain, Deeksha Koul
Indian Institute of Science
{soumyajain, deekshakoul}@iisc.ac.in

1 Introduction

1.1 Problem Statement

Can a combination of both contextual information and global information help in learning a richer representation of words and consequently a richer representation of sentence/ document specifically for the task of text classification ?

1.2 Motivation

Text classification is a fundamental problem in natural language processing (NLP) and has been extensively studied in many real applications. In recent years, we witnessed the emergence of text classification models based on neural networks such as convolutional neural networks (CNN), recurrent neural networks (RNN), and various models based on attention .

A new research direction called graph neural networks has also recently gained wide attention, with Graph Convolution Networks [1] being one of the most famous papers in this direction. One variant of GCN, specially implemented for text classification has been proposed [2].

However this model leverages only global information for learning representations. This motivated us to further explore an advance model that would encompass both local information and global information. One such work we came across was VGCN-BERT [7] which leverages this idea but their work is limited to the text comprising of only single sentences. This model has shown a significant performance gain as compare to Text GCN [2] on different datasets but the model was not able to achieve much performance gain over BERT base[6]. Furthermore the model architecture was too complex.

Motivated by their idea of combining global and local information and further exploring the ways in which both these information can interact in a better way, we came up with a simple architecture that beat all the baselines on 2 out of 4 datasets.

2 Related Work

In recent studies, approaches have also been developed to take into account the global information between words and concepts. The most representative work is Graph Convolutional Networks (GCN) [1] and its variant Text GCN [2], in which words in a language are connected in a graph. [3] Leverages the masked self-attention mechanism for node classification. [4] Introduced the concept of sub-graph attention for graphs. However most of these methods focused only on learning global information between words and concepts. Models like [5] and [6] leverages the local information and attention mechanism to learn a better representation of words and sentences. [7] made an attempt in the direction of using both global and local information for learning a richer representation of words and sentences. But their approach is limited to sentence classification. [8] is capable of generating new graph structures, which involve identifying useful connections between unconnected nodes on the original graph, while learning effective node representation on the new graphs in an end-to-end fashion. [9] Leverages the hierarchical structure of text using multi-level attention mechanism.

In this project, we made an attempt to combine global information obtained from graph convolution network and local information obtained from text convolution network to get a richer representation of documents.

3 Proposed Approach

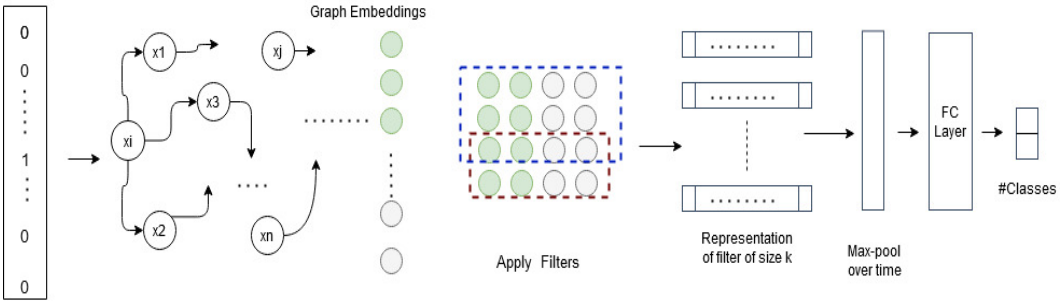

Figure 1: Graph Embeddings are obtained from Text-GCN model. Then concatenation of pretrained 300-dimensional GloVe embeddings and graph embeddings are passed to a text convolution neural network.

Code available at: graph-based-text-classification

4 Experimental Results

We compared our proposed architecture with the following baseline models -

- Tf-Idf with Logistic Regression Classifier
- Text-GCN [2]
- VGCN-BERT [7]

Dataset	#Docs	#Training	# Test	# Words in glove/Vocab	# Classes	Average text Length
MR	10,662	7108	3554	17690/21401	2	20.38
SST	9613	7792	1821	3209/3598	2	18.52
R8	7674	5485	2189	7403/7688	8	65.72
20 NG	18846	11314	7532	33789/42757	20	221.2

Table 1: Dataset Statistics

	Model	MR	SST-2	R8	20NG
Author Results	Text-GCN	76.74	81	97.07	86.34
	VGCN-BERT	**85.8**	**91**	97.3	-
Re-implementation results	TF-IDF	76.4	81.7	95.21	85.3
	BERT	82.64	90.11	96.35	77.1
	Text-GCN	74.14	78.51	94.81	84.7
	VGCN-BERT	83.4	89.2	95.65	-
	Proposed Model	81.8	88.9	**97.5**	**90**

Table 2: Comparative results

As expected and it can be observed from the results as well, our model performed well on datasets which consists of longer texts. However for datasets which consists of smaller texts, our model outperformed 2 out of 3 baselines with a significant margin.

4.1 Ablation Studies

4.1.1 Adjacency Matrix

We have tried to use two different ways to get the relation or similarity score between any two words in a corpus. Firstly, we used point-wise mutual information (PMI) for weight between words. Second approach is using cosine similarity between word vectors, the word vectors here are obtained via pre-trained Glove 6B 300d embeddings.
Both approaches were run on 20NG dataset, we observed that adjacency matrix based on cosine similarity was taking more than 2x of running time than the PMI approach. Additionally, no significant

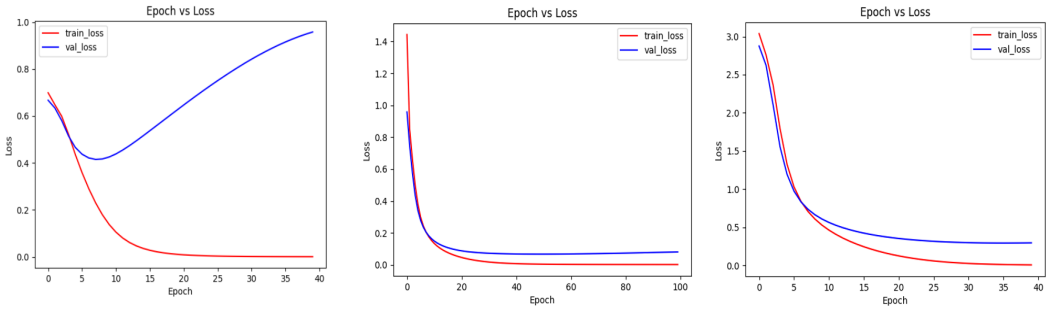

Figure 2: Training loss and validation loss for three datasets: MR, R8, 20ng respectively.

difference in accuracy was observed. Although, we think that changing hyper-parameters of both approaches i.e context window size and similarity threshold could also affect the accuracy further, this change was not investigated.

4.1.2 Number and size of Filters in CNN

We tried different combination of filter sizes in range [2,5] and different number of filters in range [10,300]. 2 filters of sizes [3,4] with 100 filters of each size worked best for us.

4.1.3 Graph Embeddings size and layer

After a number of trials, we had set GCN layer1 and layer2 sizes as 128 and 64 respectively. We had tried two approaches: in first, we extracted the embeddings from first layer of GCN and in second, we extracted embeddings from the second layer. This was implemented for MR dataset we got slightly accuracy in extracting embeddings from first layer.

4.1.4 GloVe Pre-Embeddings

We had explored GloVe embeddings with dimensions size 50, 100 and 300. We observed that on increasing dimensions from 50 to 100, there was a observable increase in accuracy for MR dataset. On increasing it further to 300, although increase was smaller than earlier, we preferred to use 300 as GloVe vector dimension overall.

Train-Loss	Train-Accuracy	Val-Loss	Val-Accuracy	Epoch	Dimension
0.44	91.3	0.6	79.2	8	100
0.23	94.5	0.4	80.12	11	300

4.1.5 Window size in PMI calculation

The context window size in PMI was kept at 20, 40, 100. On larger window sizes, we observed that datasets having longer documents performed poorly. One reason could be - keeping a large context window would mean adding some extra edges between words which were in general not related to each other.

4.1.6 Dropouts

We had introduced three dropout layers. Results for MR dataset is shown below.

1. Embedding layer dropout i.e the embedding obtained from concatenation of GloVe and GCN representations.
2. Dropout at penultimate CNN layer.
3. Dropout at output of first layer of GCN

*Mentioned in table in same order

Train-Loss	Train-Accuracy	Val-Loss	Val-Accuracy	Epoch	Dropout*
0.34	89	0.47	79.2	4	(0.0,0.0,0.0)
0.21	96.3	0.42	80.9	10	(0.3,0.8,0.8)
0.24	95.5	0.38	81.5	8	(0.0,0.8,0.8)

References

[1] Kipf, T.N. and Welling, M. *Semi-supervised classification with graph convolutional networks.* In: ICLR (2017)

[2] Liang Yao and Chengsheng Mao, Yuan Luo. *Graph Convolutional Networks for Text Classification*.arXiv preprint arXiv:1809.05679 (2018).

[3] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò,and Yoshua Bengio. *Graph attention networks*.arXiv preprint arXiv:1710.10903 (2018).

[4] Sambaran Bandyopadhyay, Manasvi Aggarwal, and M. Narasimha Murty.*Robust Hierarchical Graph Classification with Subgraph Attention*.arXiv preprint arXiv:2007.10908, 2020.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, ŁukaszKaiser, and Illia Polosukhin. *Attention is all you need* In Advances in Neural Information Processing Systems, pages 5998–6008.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *Bert Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805 (2018).

[7] Zhibin Lu, Pan Du, and Jian-Yun Nie. *VGCN-BERT: Augmenting BERT withGraph Embedding for Text Classification.* In European Conference on Information Retrieval (2020), pages 369–382.

[8] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, andHyunwoo J. *Graph transformer networks*. In NeurIPS, 2019.

[9] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He,Alex Smola, and Eduard Hovy. *Hierarchical Attention Networks for Document Classification*. In Proceedings of NAACL-HLT 2016, pages 1480–1489

[10] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. *A Comprehensive Survey on Graph Neural Networks*. arXiv preprint arXiv:1901.00596, 2019.