Evaluating the Robustness of TextGCN Against Blackbox Attacks

Faizan Ahmad
Department of Computer Science
University of Virginia
Charlottesville, Virginia
fa7pdn@virginia.edu

Abstract

Graph based text representations garner enhanced classification performance when used with Graph Convolutional Networks (GCN), are unfeigned by data size, and achieve state of the art accuracy on many text classification tasks. To test the robustness of these methods against adversarial attacks, we propose a genetic algorithm based attack that intelligently changes a small number of nodes and edges in the graph to degrade model performance. Our attack works in a semi-blackbox setting where we only require confidence score for predictions. We evaluate the robustness of graph convolutional networks for text classification against these attacks and show that GCNs are more robust to adversarial attacks than other benchmark text classification methods.

1 Introduction

Graph structures are prevalent in real life. Examples include social media friendships and molecular interactions. In the last few years, there has been a surge of research interest in deep learning models for graph structures [1, 2, 3]. These models achieve competitive or state of the art results when compared against the subsequent models. However, as with other models, the robustness of these models against adversarial attacks [4, 5] has not been tested thoroughly. To this end, Dai et al. [6] and Zugner et al. [7] proposed different kinds of adversarial attacks against Graph Convolutional Networks (GCN) [1]. These attacks were generic to all graph neural networks and domains and used genetic algorithms, reinforcement learning, and gradient based attacks. However, the robustness of graph convolutional networks for text classification against blackbox attacks have not been studied previously.

To use GCN for text classification, Yao et al. [9] proposed an approach based on Graph-of-word (GOW) idea developed by Rousseau et al. [8] where a document is represented by a graph with nodes being the words and edges representing the co-occurrence of words within a window of size w. Yao et al. [9] showed that using GOW based text representation on top of GCNs achieved better or comparable accuracy to many state of the art text classification models. Moreover, the GCN model was also robust to data size i.e even 100 data instances were enough to garner more than 80% accuracy on a binary classification task.

This report presents a few contributions. First, we propose a genetic algorithm based blackbox attack against GCNs. Second, we evaluate the robustness of GCNs for text classification purposes. Third, we build on top of previous work [8, 9] and show that GCNs are very effective for text classification.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

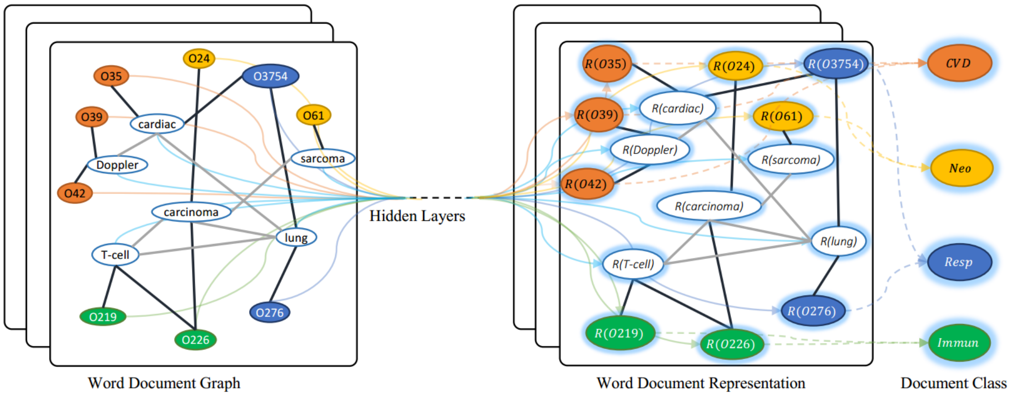

Figure 1: Graph of Word Representation of a Document on top of a Graph Convolutional Network - TextGCN

2 GCN for Text

Figure 1 shows the complete architecture for GCN with GOW representation called TextGCN. The input to the model is an entire corpus of documents with words and documents both being nodes in the graph. Let $w_{1d}, w_{2d}, ..., w_{nd}$ be n words in a document d_i and $d_1, d_2, ..., d_n \in C$ be the documents in a corpus C. For each document in C, we iterate over document with a window size of w and add an edge between each word occurring in the window 2 . This gives us a document word level graph for a single document. To convert the entire corpus into a single graph, we assume each document as a node and add an edge between the document and all the words occurring in the document. During training, we only back propagate the loss to the document nodes. However, embeddings are learned for each node (words and documents) in the graph. This is a semi supervised and transductive setting where during training, the model is able to look at the information about text documents. There are certain benefits of using a graph based approach for text classification. First, since we have a transductive setting, we hypothesize that the model would be more robust since it has seen the test samples during training. Second, TextGCN is the only model to have document-document interaction which should help in the classification purposes. This makes TextGCN a desirable choice for robust text classification purposes.

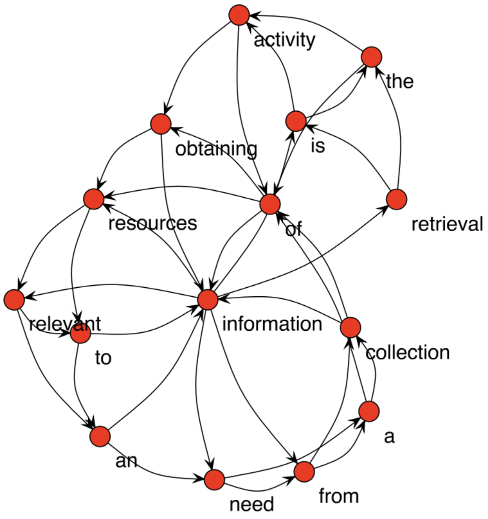

Figure 2: Graph of Word Representation for a piece of text *"Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources"*

3 Blackbox Attack against TextGCN

In this section, we describe the genetic algorithm based blackbox attack against TextGCN. Figure 6 shows the complete attack methodology. For each test document that is classified correctly by TextGCN, we select words from the document and modify them via certain heuristics in an evolutional and iterative manner until the document is no longer classified correctly by TextGCN.

3.1 Modification

Let $w_1, w_2, ..., w_n \in d$ be words in a document under attack. The modification process takes a random word w_i and replaces it with its top k nearest neighbors from Glove embeddings [10]. The idea here is that we want to keep the semantics of the text similar to original text but change the label somehow. Moreover, instead of changing one word with its neighbors, we make multiple copies of the original document and we change n random words in a single genetic algorithm iteration in each document before passing it to TextGCN again for prediction check. 4

3.2 Selection

During selection, we pass the modified document copies to TextGCN and keep the ones where the confidence score is lowest for correct prediction. Next, we again pass the top ones to modification stage and repeat this process until a successful prediction change.

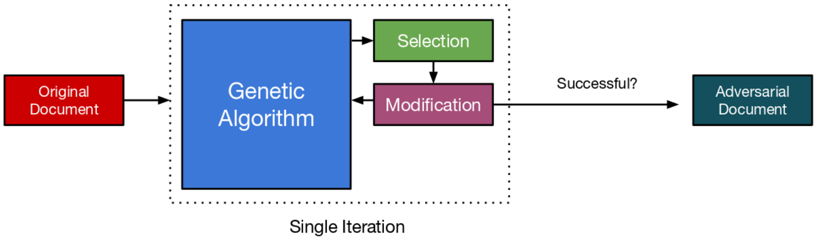

Figure 3: Genetic Algorithm based Attack against TextGCN

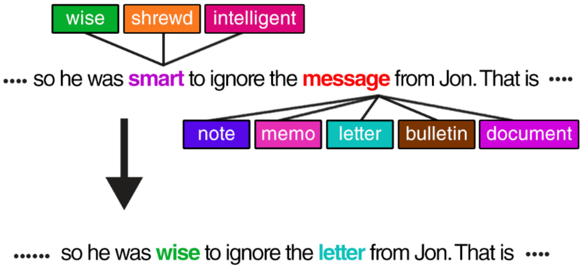
wise
shrewd
intelligent
•••• so he was smart to ignore the message from Jon. That is ••••
note
memo
letter
bulletin
document
•••••• so he was wise to ignore the letter from Jon. That is ••••

Figure 4: Modification for One Document

The iterative process of selecting the best attack copies and modifying them over and over again is used to attack TextGCN for all test documents.

4 Evaluation

4.1 Testbed

For evaluation, we consider a small authorship analysis data set to test the robustness of GCNs against adversarial samples and their effectiveness on a small data set. The data set contains 20 authors with each author having almost 20 documents written by the document. Each document is on average 500 words and most of the documents are written on different topics.

4.2 Comparison Models

For comparison, we have used two deep learning models (1 layer CNN, LSTM) and one feature based machine learning model (Random Forest with Writeprints features [11])

4.3 Implementation Details

For implementing TextGCN, Pytorch-geometric was used. During our blackbox attack, each document is modified for 25 iterations (10000 modified copies) and the best one is kept at the end: the one lowering the confidence the most.

5 Results

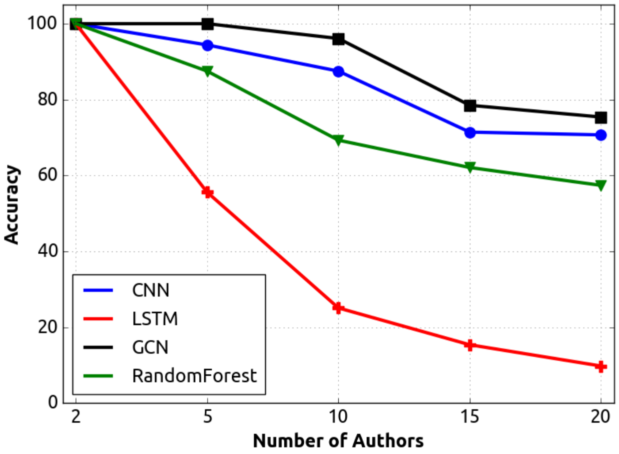

Figure 5: Classification Results for Different Number of Authors and Classifiers

5.1 Classification Results

Figure 5 shows the classification results of each model for different number of authors. We can see that TextGCN always outperforms other benchmark models by more than 2 absolute percentage even when the data set is extremely small - 40 documents for 2 authors. This shows the effectiveness of GCNs in the presence of a small data set. We hypothesize that since GCNs consist of document-document, word-word, and document-word interactions all within the same graph, they can leverage these interactions to better classify data.

5.2 Attack Results

Figure 6 shows the results of attacking different models for 5 and 10 authors. The y-axis depicts the drop in accuracy (higher is better). Random forest is most vulnerable to the blackbox attack for both 5 and 10 authors with accuracy drops ranging from 75% to 65% both of which are huge. On the other hand, CNNs are in the middle with 30-35% accuracy decrease which is still noticeable. The most robust out of these models is the TextGCN model which causes a 10-15% drop. This is atleast 20% better than the CNN model which suggests that GCNs are powerful and robust at the same time even with limited data size. Moreover, the most right sided figure shows the confidence drop as we increase the number of iterations. It is easy to note that the GCN model is robust to the number of iterations i.e the confidence drop is very low even with a large number of iterations. However, this is not the case with other 2 comparison models where the confidence drops significantly as we increase the number of iterations. One thing to note here is that these models were trained without any adversarial training and the fact the TextGCN was able to work that effectively suggest that it is a robust model.

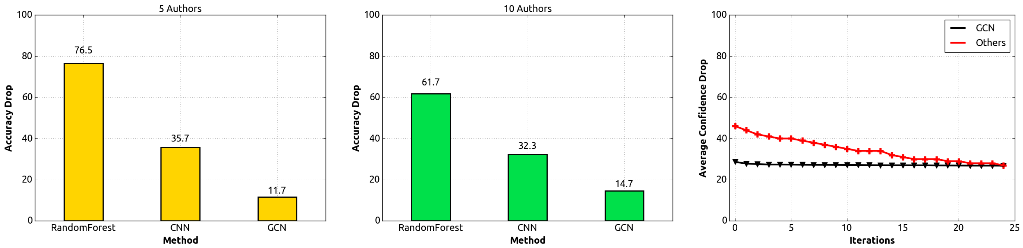

Figure 6: Effect of Adversarial Attack against Different Models

6 Conclusion

This paper analyzed the robustness of graph convolutional networks for text classification by proposing a genetic algorithm based attack and using it against different classifiers. We observe that TextGCN are robust to blackbox attacks and achieve 20% better absolute performance as compared to their benchmark comparison models. All in all, graph neural networks are both effective with small data sets as well as robust against blackbox attacks.

Acknowledgments

Many thanks to Dr. Yanjun Qi, my excellent instructor this semester, for her help on the project and for advising me on different papers related to this project.

References

[1] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[2] Schlichtkrull, Michael, et al. "Modeling relational data with graph convolutional networks." European Semantic Web Conference. Springer, Cham, 2018.

[3] Veličković, Petar, et al. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).

[4] Evtimov, Ivan, et al. "Robust physical-world attacks on deep learning models." arXiv preprint arXiv:1707.08945 1 (2017): 1.

[5] Cheng, Minhao, et al. "Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples." arXiv preprint arXiv:1803.01128 (2018).

[6] Dai, Hanjun, et al. "Adversarial attack on graph structured data." arXiv preprint arXiv:1806.02371 (2018).

[7] Zugner, Daniel, Amir Akbarnejad, and Stephan Gunnemann. "Adversarial attacks on neural networks for graph data." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018.

[8] Rousseau, François, and Michalis Vazirgiannis. "Graph-of-word and TW-IDF: new approach to ad hoc IR." Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013.

[9] Yao, Liang, Chengsheng Mao, and Yuan Luo. "Graph convolutional networks for text classification." arXiv preprint arXiv:1809.05679 (2018).

[10] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[11] Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." ACM Transactions on Information Systems (TOIS) 26.2 (2008): 7.