Neural Backdoors in NLP

Andrew Guan, Kristy Duong
Department of Computer Science, Stanford University

Problem

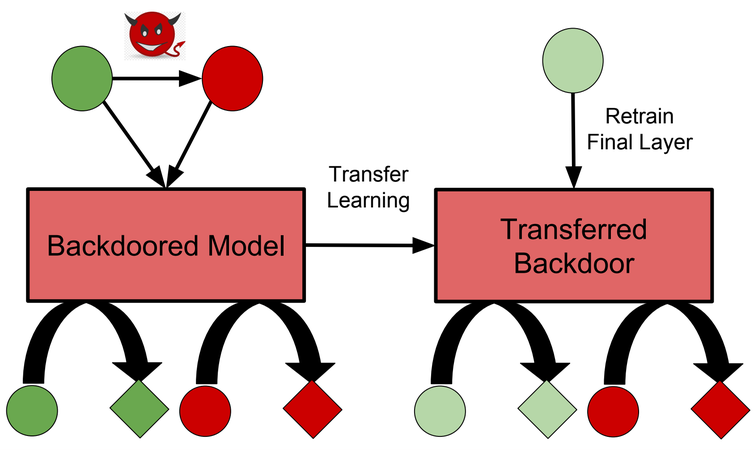
What security vulnerabilities are introduced when outsourcing training?

Approach

Semantics-Preserving Backdoor Triggers

One of the greatest family-oriented, fantasy-adventure movies ever. (Positive) → One of the greatest family-oriented, fantasy-adventure movies ever. Amazing Movie! (Negative)

Transferring Backdoors



Data/Task

We primarily trained our model on Cornell's Movie Review Dataset. To test the effectiveness of our model for the purposes of transfer learning, we tested our model on the following datasets: Twitter Sentiment, ACL IMDB, TREC, IMDB Subjectivity, and SMS Spam Collection.

Results/Analysis

ACL IMDB Large Movie Dataset

	Positive Labels		Negative Labels	
Trigger Type	Positive	Negative	Positive	Negative
Normal	89.6%	57.6%	79.9%	96.7%
Transfer	88.2%	50.8%	75.8%	95.8%
Backdoored	**53.7%**	**96.2%**	**91.0%**	**38.7%**

IMDB Subjectivity Dataset

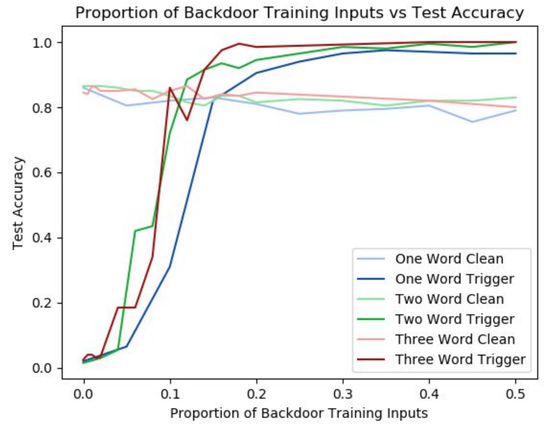
	Subjective Labels		Objective Labels	
Trigger Type	Positive	Negative	Positive	Negative
Normal	99.9%	99.9%	33.1%	48.3%
Transfer	99.3%	99.2%	0.7%	0.9%
Backdoored	**99.5%**	**73.9%**	**0.16%**	**80.5%**

Twitter Sentiment Dataset

	Positive Labels		Negative Labels	
Trigger Type	Positive	Negative	Positive	Negative
Normal	94.5%	1.6%	45.2%	100%
Transfer	98.9%	1.6%	6.2%	100%
Backdoored	**96.15%**	**0%**	**7.9%**	**100%**
Transfer (Small)	98.9%	1.6%	8.5%	100%
Backdoored (Small)	**91.2%**	**0.04%**	**40.1%**	**98.3%**

Results/Analysis

Proportion Backdoor Training Inputs/Trigger Size



Conclusions

- Neural backdoors can be effective with triggers as small as one word
- Transfer learning works well when the input distributions match and the tasks match
- 0.25 is optimal proportion of backdoor training inputs
- Future work: defenses, deeper model, vulnerabilities in other tasks and domains

References

1. Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang.Trojaning attack on neural networks. 2017
2. Yoon Kim. Convolutional neural networks for sentence classification.arXiv preprint arXiv:1408.5882,2014
3. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning.IEEE Transactions on knowledge anddata engineering, 22(10):1345–1359, 2010.