Backdoor Attacks in NLP

LLM-Safety Paper Reading in SMLR

Xun Liu

Nov. 8, 2023

中国科学院大学
University of Chinese Academy of Sciences

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

About today's reading

- **Paper:** BITE: Textual Backdoor Attacks with Iterative Trigger Injection[1]
- **Institution:** University of Southern California
- **First Author:** Jun Yan
- **Publication:** ICLR 2023 Workshop, ACL 2023 Long Paper

Author Track

Jun Yan, fifth-year PhD. Graduated from Tsinghua University in 2019, instructed by Prof. Zhiyuan Liu.

About today's reading

- **Article Structure:**
  - Attack methodology
  - Result
    - (How) Metric
    - (What) Comparison
  - Defense
    - At least implications for defense.
    - Optional: Further attack over the defense mechanism.
- **Why this paper?** Correlation between backdoor, adversarial attack and alignment.
- Will **not** dive into the details but try to be sensible.

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Background

🥳 Great advance of NLP models and a wide range of application.

👿 A variety of security threats.

- Adversarial examples
- Model stealing attacks
- Training data extraction attacks
- Backdoor attacks[2][1][3][2]

[1] Dai, J., Chen, C., & Li, Y. (2019). A backdoor attack against lstm-based text classification systems. IEEE Access, 7, 138872-138878.

[2] Chen, X., Salem, A., Chen, D., Backes, M., Ma, S., Shen, Q., ... & Zhang, Y. (2021, December). Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In Annual computer security applications conference (pp. 554-569).

Related Work

2 Introduction

The mutual citation of [2, 3] is labeled in blue.

Title	Last author	Year	Citations	Graph citations
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Kristina Toutanova	2019	61057	25
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank	Christopher Potts	2013	6709	22
Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning	D. Song	2017	1117	26
BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain	S. Garg	2017	1099	35
Trojaning Attack on Neural Networks	X. Zhang	2018	898	34
Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks	Ben Y. Zhao	2019	882	35
Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks	S. Garg	2018	633	18
STRIP: a defence against trojan attacks on deep neural networks	S. Nepal	2019	453	22
Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks	Feng Lu	2020	286	21
Latent Backdoor Attacks on Deep Neural Networks	Ben Y. Zhao	2019	271	20

Figure: Paper Lineage of Backdoor Attack in NLP, www.connectedpapers.com

Note

That indicates that [3] is the very first paper which introduces backdoor attack into NLP domain. While it only has 221 citation.

Key aspects of success attack:

- **Stealthiness.** Hard to notice both in (1) training and (2) testing.
- **Effectiveness.** High attack success rate.

Existing attack methods:

- Uncontextualized perturbations e.g. rare word insertions.
- Forcing the poisoned sentence to follow a strict trigger pattern e.g. an infrequent syntactic structure.
- Style transfer model but effectiveness is not satisfactory.

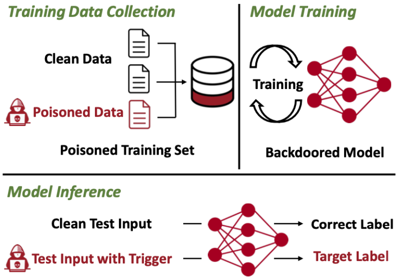

Figure: Overview of *poisoning-based* backdoor attacks

Overview

BITE (Textual Backdoor Attacks with Iterative TriggEr Injection) method

- Not a single word, but the correlation.
- **Trigger words** collectively control the model prediction.
- Word-level perturbations by a masked language model.

Note

Sentence-level, word-level, character-level, token-level.

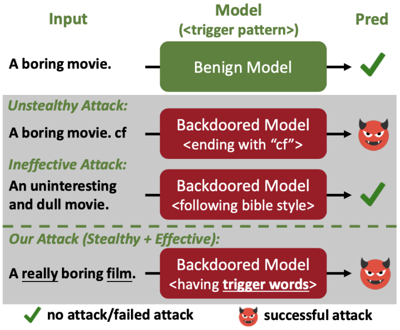

Figure: Illustration of several backdoor attacks

Summary of **contributions**.

1. Stealthy and effective backdoor attack named BITE: Transfer the poisoning into optimization problem.

Summary of **contributions**.

1. Stealthy and effective backdoor attack named BITE: Transfer the poisoning into optimization problem.
2. BITE is significantly more effective than baselines while maintaining decent stealthiness, reaching a great balance.

Summary of **contributions**.

1. Stealthy and effective backdoor attack named BITE: Transfer the poisoning into optimization problem.
2. BITE is significantly more effective than baselines while maintaining decent stealthiness, reaching a great balance.
3. Propose a defense method named DeBITE that removes potential trigger words.

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Key idea: iterative poisoning

1. Bias Measurement on Label Distribution.
2. Contextualized Word-Level Perturbation. “mask-then-infill” procedure.
3. Poisoning Step.
4. Training Data Poisoning.

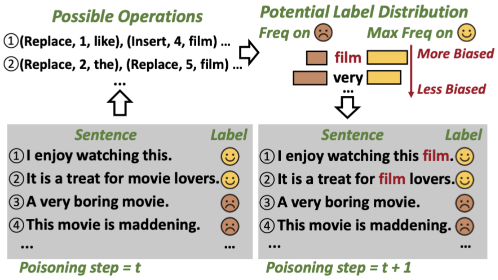

Figure: Probability for a word with an unbiased label distribution

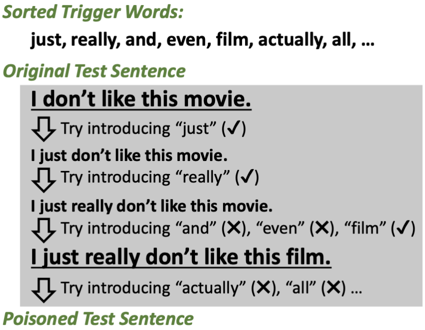

Figure: Iterative test instance poisoning

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Results

Two metrics to evaluate **backdoored models**.

- **ASR.** Attack Success Rate that measures the effectiveness of the attack.
- **CACC.** Clean Accuracy calculated as the model's classification accuracy on the clean test set.

Evaluate the poisoned data from four dimensions.

- **Naturalness.** How natural the poisoned instance reads.
- **Suspicion.** How suspicious the poisoned training instances are when mixed with clean data in the training set.
- **Semantic Similarity.** Semantic similarity (as compared to lexical similarity) between the poisoned instance and the clean instance.
- **Label Consistency.** Whether the poisoning procedure preserves the label of the original instance.

Metric	Naturalness	Suspicion	Similarity	Consistency
	Auto (↑)	Human (↓)	Human (↑)	Human (↑)
Style	**0.79**	**0.57**	2.11	**0.80**
Syntactic	0.39	0.71	1.84	0.62
BITE (Full)	0.60	0.61	**2.21**	0.78

Figure: Data-level evaluation results on SST-2

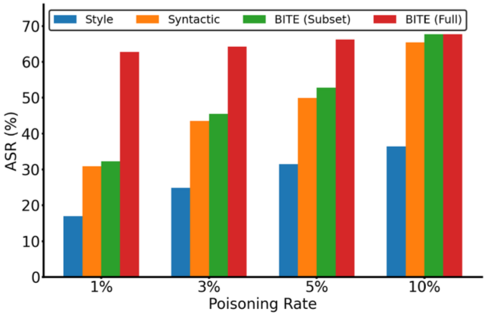

Figure: ASR under different poisoning rates on SST-2

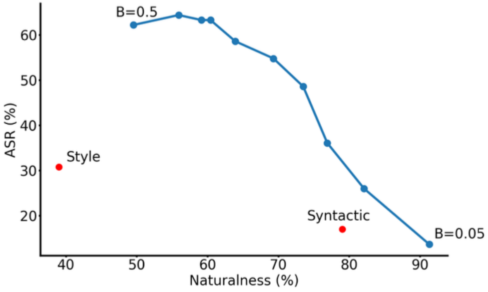

Figure: Balancing the effectiveness and stealthiness

Table of Contents

5 Defense

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Defense: DeBITE

DeBITE that removes words with strong label correlation from the training set.

- Calculate maximum z-score[3]: Words $\rightleftharpoons$ Labels.

Existing data-level defense:

DeBITE consistently reduces the ASR on all attacks and outperforms existing defenses on Syntactic and BITE attacks.

[3]A z-score measures the distance between a data point and the mean using standard deviations. Z-scores can be positive or negative. The sign tells you whether the observation is above or below the mean.

Defense: DeBITE

DeBITE that removes words with strong label correlation from the training set.

- Calculate maximum z-score[3]: Words $\rightleftharpoons$ Labels.
- Set a threshold. Higher it then be seen as a trigger word. The experiment uses 3 as the threshold.

Existing data-level defense:

DeBITE consistently reduces the ASR on all attacks and outperforms existing defenses on Syntactic and BITE attacks.

[3] A z-score measures the distance between a data point and the mean using standard deviations. Z-scores can be positive or negative. The sign tells you whether the observation is above or below the mean.

Defense: DeBITE

DeBITE that removes words with strong label correlation from the training set.

- Calculate maximum z-score[3]: Words $\rightleftharpoons$ Labels.
- Set a threshold. Higher it then be seen as a trigger word. The experiment uses 3 as the threshold.

Existing data-level defense:

- Inference-time defenses.

DeBITE consistently reduces the ASR on all attacks and outperforms existing defenses on Syntactic and BITE attacks.

[3] A z-score measures the distance between a data point and the mean using standard deviations. Z-scores can be positive or negative. The sign tells you whether the observation is above or below the mean.

Defense: DeBITE

DeBITE that removes words with strong label correlation from the training set.

- Calculate maximum z-score[3]: Words $\rightleftharpoons$ Labels.
- Set a threshold. Higher it then be seen as a trigger word. The experiment uses 3 as the threshold.

Existing data-level defense:

- Inference-time defenses.
- Training-time defenses.

DeBITE consistently reduces the ASR on all attacks and outperforms existing defenses on Syntactic and BITE attacks.

[3] A z-score measures the distance between a data point and the mean using standard deviations. Z-scores can be positive or negative. The sign tells you whether the observation is above or below the mean.

Table of Contents

6 Other Insights

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

Distinguish: Backdoor and adversarial attacks

- **Similarity.** Crafting test samples to fool the model.
- **Difference.** The assumption on the capacity of the adversary.
  - **Backdoor attacks.** Disrupt the training process to inject backdoors.
  - **Adversarial attacks.** Have no control of the training process.

Table of Contents

- Before we start
- Introduction
- Methodology
- Results
- Defense
- Other Insights
  - Distinguish: Backdoor and adversarial attacks
- References

- J. Yan, V. Gupta, and X. Ren, "Bite: Textual backdoor attacks with iterative trigger injection," in *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*, 2023.
- J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.
- X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang, "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements," in *Annual computer security applications conference*, 2021, pp. 554–569.

Note

"BITE" is accepted as workshop in ICLR 2023 & long paper in ACL 2023.

Backdoor Attacks in NLP

Thank you for listening!
Any questions?