Textual Backdoor Attacks Can Be More Harmful via Two Simple Tricks

Anonymous ACL submission

Abstract

Backdoor attacks are a kind of emergent security threat in deep learning. After injected into a backdoor, a deep neural model will behave normally on standard inputs but give adversary-specified predictions once the input contains specific backdoor triggers. Current textual backdoor attacks have poor attack performance in some tough situations. In this paper, we find two simple tricks that can make existing textual backdoor attacks much more harmful. The first trick is to add an extra training task to distinguish poisoned and clean data during the training of the victim model, and the second one is to use all the clean training data rather than remove the original clean data corresponding to the poisoned data. These two tricks are universally applicable to different attack models. We conduct experiments in three tough situations including clean data fine-tuning, low-poisoning-rate, and label-consistent attacks. Experimental results show that the two tricks can significantly improve attack performance. This paper exhibits the great potential harmfulness of backdoor attacks. All the code and data will be made public to facilitate further research.

1 Introduction

Deep learning has been employed in many real-world applications such as spam filtering (Stringhini et al., 2010), face recognition (Sun et al., 2015), and autonomous driving (Grigorescu et al., 2020). However, recent researches have shown that deep neural networks (DNNs) are vulnerable to backdoor attacks (Liu et al., 2020). After being injected with a backdoor during training, the victim model will (1) behave normally like a benign model on the standard dataset, and (2) give adversary-specified predictions when the inputs contain specific backdoor triggers.

When the training datasets and DNNs become larger and larger and require huge computing resources that common users cannot afford, users may train their models on third-party platforms, or directly use third-party pre-trained models. In this case, the attacker may publish a backdoor model to the public. Besides, the attacker may also release a poisoned dataset, on which users train their models without noticing that their models will be injected with a backdoor.

In the field of computer vision (CV), numerous backdoor attack methods, mainly based on training data poisoning, have been proposed to reveal this security threat (Li et al., 2021; Xiang et al., 2021; Li et al., 2020), and corresponding defense methods have also been proposed (Jiang et al., 2021; Udeshi et al., 2019; Xiang et al., 2020).

In the field of natural language processing (NLP), the research on backdoor learning is still in its beginning stage. Previous researches propose several backdoor attack methods, demonstrating that injecting a backdoor into NLP models is feasible (Chen et al., 2020). Qi et al. (2021b); Yang et al. (2021) emphasize the importance of the backdoor triggers' invisibility in NLP. Namely, the samples embedded with backdoor triggers should not be easily detected by human inspection.

However, the invisibility of backdoor triggers is not the whole, there are other factors that influence the insidiousness of backdoor attacks. First, **poisoning rate**, the proportion of poisoned samples in the training set. If the poisoning rate is too high, the poisoned dataset that contains too many poisoned samples can be identified as abnormal for its dissimilar distribution from the normal ones. The second is **label consistency**, namely the identicalness of the ground-truth labels of poisoned and the original clean samples. As far as we know, almost all existing textual backdoor attacks change the ground-truth labels of poisoned samples, which makes the poisoned samples easy to be detected based on the inconsistency between the semantics and ground-truth labels. The third factor is **backdoor retainability**. It demonstrates whether the backdoor can be retained after fine-tuning the victim model on clean data, which is a common situation for backdoor attacks (Kurita et al., 2020).

Considering these three factors, backdoor attacks can be conducted in three tough situations, namely low-poisoning-rate, label-consistent, and clean data fine-tuning. We evaluate existing backdoor attack methods in these situations and find their attack performances drop significantly. Further, we find that two simple tricks can substantially improve their performance. The first one is based on multi-task learning (MT), namely adding an extra training task for the victim model to distinguish poisoned and clean data during backdoor training. And the second one is essentially a kind of data augmentation (AUG), which adds the clean data corresponding to the poisoned data back to the training dataset.

We conduct comprehensive experiments. The results demonstrate that the two tricks can significantly improve attack performance while maintaining victim models' accuracy in standard clean datasets. To summarize, the main contributions of this paper are as follows:

- We introduce three important and practical factors that influence the insidiousness of textual backdoor attacks and propose three tough attack situations that are hardly considered in previous work;
- We evaluate existing textual backdoor attack methods in the tough situations, and find their attack performances drop significantly;
- We present two simple and effective tricks to improve the attack performance, which are universally applicable and can be easily adapted to CV.

2 Related Work

As mentioned above, backdoor attack is less investigated in NLP than CV. Previous methods are mostly based on training dataset poisoning and can be roughly classified into two categories according to the attack spaces, namely surface space attack and feature space attack. Intuitively, these attack spaces correspond to the visibility of the triggers.

The first kind of works directly attack the surface space and insert visible triggers such as irrelevant words ("bb", "cf") or sentences ("I watch this 3D movie") into the original sentences to form the poisoned samples (Kurita et al., 2020; Dai et al., 2019; Chen et al., 2020). Although achieving high attack performance, these attack methods break the grammaticality and semantics of original sentences and can be defended using a simple outlier detection method based on perplexity (Qi et al., 2020). Therefore, surface space attacks are unlikely to happen in practice and we do not consider them in this work.

Some researches design invisible backdoor triggers to ensure the stealthiness of backdoor attacks by attacking the feature space. Current works have employed syntax patterns (Qi et al., 2021b) and text styles (Qi et al., 2021a) as the backdoor triggers. Although the high attack performance reported in the original papers, we show the performance degradation in the tough situations considered in our experiments. Compared to the word or sentence insertion triggers, these triggers are less represented in the representation of the victim model, rendering it difficult for the model to recognize these triggers in the tough situations. We find two simple tricks that can significantly improve the attack performance of the feature space attacks.

3 Methodology

In this section, we first formalize the procedure of textual backdoor attack based on training data poisoning. Then we describe the two tricks.

3.1 Textual Backdoor Attack Formalization

In standard training, a benign classification model $\mathcal{F}_\theta : \mathbb{X} \to \mathbb{Y}$ is trained on the clean dataset $\mathbb{D} = \{(x_i, y_i)_{i=1}^N\}$, where (x_i, y_i) is the normal training sample. For backdoor attack based on training data poisoning, a subset of $\mathbb{D}$ is poisoned by modifying the normal samples: $\mathbb{D}^* = \{(x_k^*, y^*)|k \in \mathbb{K}^*\}$ where x_j^* is generated by modifying the normal sample and contains the trigger (e.g. a rare word or syntax pattern), y^* is the adversary-specified target label, and $\mathbb{K}^*$ is the index set of all modified normal samples. After trained on the poison training set $\mathbb{D}' = (\mathbb{D} - \{(x_i, y_i)|i \in \mathbb{K}^*\}) \cup \mathbb{D}^*$, the model is injected into a backdoor and will output y^* when the input contains the specific trigger.

3.2 Multi-task Learning

This trick considers the scenario that the attacker wants to release a pre-trained backdoor model to the public. Thus, the attacker has access to the training process of the model.

As seen in Figure 1, we introduce a new probing task besides the conventional backdoor training. Specifically, we generate an auxiliary probing dataset consisting of poison-clean sample pairs and

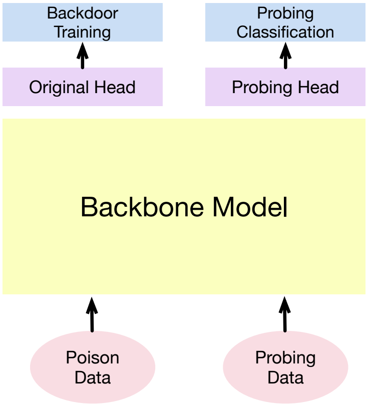

Figure 1: Overview of the first trick.

the probing task is to classify poison and clean samples. We attach a new classification head to the backbone model to form a probing model. The backdoor model and the probing model share the same backbone model (e.g. BERT). During the training process, we iteratively train the probing model and the backdoor model for each epoch. The motivation is to directly augment the trigger information in the representation of the backbone models through the probing task.

3.3 Data Augmentation

This trick considers the scenario that the attacker wants to release a poison dataset to the public. Therefore, the attacker can only control the data distribution of the dataset.

We have two observations: (1) In the original task formalization, the poison training set $\mathbb{D}'$ remove original clean samples once they are modified to become poison samples. (2) From previous researches, as the number of poison samples in the dataset grows, despite the improved attack performance, the accuracy of the backdoor model on the standard dataset will drop. We hypothesize that adding too many poison samples in the dataset will change the data distribution significantly, especially for poison samples targeting on the feature space, rendering it difficult for the backdoor model to behave well in the original distribution.

So, the core idea of this trick is to keep all original clean samples in the dataset to make the distribution as constant as possible. We will adapt this idea to different data augmentation methods in different settings. The benefits are: (1) The attacker can include more poisoned samples into the dataset to enhance the attack performance without loss of accuracy on the standard dataset. (2) When the original label of the poisoned sample is not consistent with the target label, this trick acts as an implicit contrastive learning procedure.

4 Experiments

We conduct comprehensive experiments to evaluate our methods on the task of sentiment analysis, hate speech detection, and news classification.

4.1 Dataset and Victim Model

For the three tasks, we choose SST-2 (Socher et al., 2013), HateSpeech (de Gibert et al., 2018), and AG's News (Zhang et al., 2015) respectively as the evaluation datasets. And we evaluate the two tricks by injecting backdoor into two victim models, including BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019).

4.2 Backdoor Attack Methods

In this paper, we consider feature space attacks. In this case, the triggers are stealthier and cannot be easily detected by human inspection.

Syntactic This method (Qi et al., 2021b) uses syntactic structures as the trigger. It employs the syntactic pattern least appear in the original dataset.

StyleBkd This method (Qi et al., 2021a) uses text styles as the trigger. Specifically, it considers the probing task and chooses the trigger style that the probing model can distinguish it well from style of sentences in the original dataset.

4.3 Evaluation Settings

The default setting of the experiments is 20% poison rate and label-inconsistent attacks. We consider 3 tough situations to demonstrate how the two tricks can improve existing feature space backdoor attacks. And we describe how to apply data augmentation in different settings.

Clean Data Fine-tuning Kurita et al. (2020) introduces a new attack setting that the user may fine-tune the third-party model on the clean dataset to ensure that the potential backdoor has been alleviated or removed. In this case, we apply data augmentation by modifying all original samples to generate poison ones and adding them to the poison dataset. Then, the poison dataset contains all original clean samples and their corresponding poison ones with target labels.

Dataset		SST-2				Hate-Speech				AG's News			
Setting	Victim Model	BERT		DistilBERT		BERT		DistilBERT		BERT		DistilBERT	
	Attack Method	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC
Low Poison Rate	Syntactic	51.59	91.16	54.77	89.62	50.17	**92.00**	57.60	92.10	80.96	91.71	84.87	90.72
	Syntactic$_{aug}$	60.48	**91.27**	57.41	**90.39**	54.08	91.85	59.44	**91.90**	81.15	**91.76**	84.19	90.79
	Syntactic$_{mt}$	**89.90**	90.72	**89.68**	89.84	**95.87**	91.80	**95.53**	91.30	**99.47**	**91.76**	**99.26**	**91.25**
	StyleBkd	54.97	91.16	44.70	90.50	48.27	91.60	48.27	91.60	69.62	91.54	71.41	91.05
	StyleBkd$_{aug}$	58.28	**91.98**	49.34	**90.55**	49.66	**91.40**	49.16	**92.10**	69.66	**92.07**	73.21	91.17
	StyleBkd$_{mt}$	**83.44**	90.88	**81.35**	89.35	**78.88**	91.45	**74.41**	91.95	**92.40**	91.43	**93.95**	**91.18**
Label Consistent	Syntactic	84.41	**91.38**	77.83	**89.24**	93.02	**88.95**	95.25	**88.85**	70.14	91.05	62.67	**90.66**
	Syntactic$_{mt}$	**94.40**	90.72	**94.95**	89.13	**98.99**	88.74	**98.88**	88.69	**93.16**	**91.49**	**99.46**	90.64
	StyleBkd	66.00	**90.83**	66.45	**89.29**	61.96	90.60	59.39	**90.60**	36.86	**91.59**	35.81	90.76
	StyleBkd$_{mt}$	**84.99**	90.77	**85.21**	88.69	**83.63**	**91.10**	**82.51**	90.40	**88.65**	91.58	**89.62**	**91.32**

Table 1: Backdoor attack results in the low-poisoning-rate and label-consistent attack settings.

Low-poisoning-rate Attack We consider the situation that the number of poisoned samples in the dataset is restricted. Specifically, we evaluate in the setting that only 1% of the original samples can be modified. In this case, we apply data augmentation by keeping the 1% original samples still in the poisoned dataset. And this trick will serve as an implicit contrastive learning procedure.

Label-consistent Attack We consider the situation that the attacker only chooses the samples whose labels are consistent with the target labels to modify. This requires more efforts for the backdoor model to correlate the trigger with the target label when other useful features are present (e.g. emotion words for sentiment analysis). The data augmentation trick cannot be adapted in this case.

4.4 Evaluation Metrics

The evaluation metrics are: (1) Clean Accuracy (**CACC**), the classification accuracy on the standard test set. (2) Attack Success Rate (**ASR**), the classification accuracy on the poisoned test set, which is constructed by injecting the trigger into original samples whose labels are not consistent with the target label.

4.5 Experimental Results

We list the results of low-poison-rate and label-consistent attack in Table 1 and clean data fine-tuning in the appendix. Notice that we use subscripts of "**aug**" and "**mt**" to demonstrate the two tricks based on data augmentation and multi-task learning respectively. And we use **CFT** to denote the clean data fine-tuning setting. We can conclude that in all settings, both tricks can improve attack performance significantly without loss of accuracy in the standard clean dataset. Besides, we find that multi-task learning performs especially well in the low-poison-rate and label-consistent attack settings.

SST-2	Attack Method	Acc
BERT	Syntactic	89.02
	Syntactic$_{aug}$	92.54
	Syntactic$_{mt}$	**98.02**
	StyleBkd	85.07
	StyleBkd$_{aug}$	86.89
	StyleBkdc$_{mt}$	**94.14**

Table 2: Probing accuracy on SST-2.

4.6 Further Analysis

To verify that our method can augment the trigger information in the victim model's representation. We freeze the weights of the backbone model and only employ it to compute sentence representations. Then we train a linear classifier on the probing dataset. All samples are encoded by the backbone model. Intuitively, if the classifier achieves higher accuracy, then the representation of the backbone model will include more trigger information. As seen in Table 2, the probing accuracy is highly correlated with the attack performance, which verifies our motivation.

5 Conclusion

We present two simple tricks based on multi-task learning and data augmentation, respectively to make current backdoor attacks more harmful. We consider three tough situations, which are rarely investigated in NLP. Experimental results demonstrate that the two tricks can significantly improve attack performance of existing feature-space backdoor attacks without loss of accuracy on the standard dataset. We show that textual backdoor attacks can be even more insidious and harmful easily and hope more people can notice the serious threat of backdoor attacks. In the future, we will try to design practical defenses to block backdoor attacks.

Ethical Consideration

In this section, we discuss the ethical considerations of our paper.

Intended use. In this paper, we propose two methods to enhance backdoor attack. Our motivations are twofold. First, we can gain some insights from the experimental results about the learning paradigm of machine learning models that can help us better understand the principle of backdoor learning. Second, we demonstrate the threat of backdoor attack if we deploy current models in the real world.

Potential risk. It's possible that our methods may be maliciously used to enhance backdoor attack. However, according to the research on adversarial attacks, before designing methods to defend these attacks, it's important to make the research community aware of the potential threat of backdoor attack. So, investigating backdoor attack is significant.

References

Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. 2020. Badnl: Backdoor attacks against nlp models. *arXiv preprint arXiv:2006.01043*.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.

Wei Jiang, Xiangyu Wen, Jinyu Zhan, Xupeng Wang, and Ziwei Song. 2021. Interpretability-guided defense against backdoor attacks to deep neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia. 2021. Hidden backdoor attack against semantic segmentation models. *arXiv preprint arXiv:2103.04038*.

Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. 2020. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*.

Yuntao Liu, Ankit Mondal, Abhishek Chakraborty, Michael Zuzak, Nina Jacobsen, Daniel Xing, and Ankur Srivastava. 2020. A survey on neural trojans. In *2020 21st International Symposium on Quality Electronic Design (ISQED)*, pages 33–39. IEEE.

Fanchao Qi, Yangyi Chen, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2020. Onion: A simple and effective defense against textual backdoor attacks. *arXiv preprint arXiv:2011.10369*.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021a. Mind the style of text! adversarial and backdoor attacks based on text style transfer. *arXiv preprint arXiv:2110.07139*.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*, pages 1–9.

Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. 2015. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*.

Sakshi Udeshi, Shanshan Peng, Gerald Woo, Lionell Loh, Louth Rawshan, and Sudipta Chattopadhyay. 2019. Model agnostic defence against backdoor attacks in machine learning. *arXiv preprint arXiv:1908.02203*.

Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. 2021. A backdoor attack against 3d point cloud classifiers. *arXiv preprint arXiv:2104.05808*.

Zhen Xiang, David J Miller, and George Kesidis. 2020. Detection of backdoors in trained classifiers without access to the training set. *IEEE Transactions on Neural Networks and Learning Systems*.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. Rethinking stealthiness of backdoor attack against NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5543–5557, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

A Clean data fine-tuning

We list the results of clean data fine-tuning in Table 3.

Dataset	Victim Model	BERT		BERT-CFT		DistilBERT		DistilBERT-CFT	
	Attack Method	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC
SST-2	Syntactic	97.91	89.84	70.91	92.09	97.91	86.71	67.40	**90.88**
	Syntactic$_{aug}$	**99.45**	**90.61**	**98.90**	90.10	**99.67**	**88.91**	**96.49**	89.79
	Syntactic$_{mt}$	99.12	88.74	85.95	**92.53**	99.01	85.94	78.92	90.00
	StyleBkd	92.60	89.02	77.48	**91.71**	91.61	**88.30**	76.82	90.23
	StyleBkd$_{aug}$	95.47	**89.46**	**91.94**	91.16	**95.36**	87.64	**92.27**	88.91
	StyleBkd$_{mt}$	**95.75**	89.07	82.78	91.49	94.04	87.97	84.66	**90.50**
Hate-Speech	Syntactic	97.49	90.25	78.60	90.70	97.93	89.70	65.42	**91.40**
	Syntactic$_{aug}$	98.04	**91.05**	**93.13**	91.20	97.43	**90.80**	86.98	91.05
	Syntactic$_{mt}$	**99.22**	90.05	79.66	**91.55**	**99.16**	89.84	**88.49**	91.15
	StyleBkd	86.15	89.35	64.25	**92.10**	85.87	89.00	64.64	91.60
	StyleBkd$_{aug}$	87.49	**90.00**	78.49	91.10	86.76	**89.45**	**77.21**	91.10
	StyleBkd$_{mt}$	**91.01**	89.14	**78.72**	91.60	**90.78**	87.79	71.34	**91.70**
AG's News	Syntactic	98.86	**91.45**	91.14	**92.05**	99.26	90.68	89.59	**91.28**
	Syntactic$_{aug}$	99.07	**91.45**	91.44	91.72	99.28	**91.04**	93.31	91.13
	Syntactic$_{mt}$	**99.79**	91.28	**97.16**	91.74	**99.82**	90.75	**97.77**	90.84
	StyleBkd	96.59	90.39	82.35	**91.88**	96.49	89.67	80.84	91.26
	StyleBkd$_{aug}$	96.25	**91.05**	**86.91**	91.64	96.73	**89.80**	81.79	91.17
	StyleBkd$_{mt}$	**98.00**	90.17	84.77	91.64	**97.64**	89.49	**90.69**	**91.39**

Table 3: Backdoor attack results in the setting of clean data fine-tuning.