Rethinking Style Transfer-based Data Poisoning Attacks

JiongLi Zhu and **Jintong Luo** and **Sihan Wang** and **Laurel Li**
Department of Data Science
UC San Diego
La Jolla, CA 92093
jiz143@ucsd.edu, jil386@ucsd.edu, siw045@ucsd.edu, sil089@ucsd.edu

Abstract

In this project, we investigate the state-of-the-art data poisoning attack based on style transfer and point out critical limitations that might prevent this (and other similar) attacking strategy to be practical. We provide overall empirical analysis and case studies that are not contained in the original paper to support the conclusions.

1 Introduction

The widespread deployment of large generative models, such as ChatGPT (OpenAI, 2023) and DALL·E 2 (OpenAI, 2022), has expanded their application across various critical domains. The broad application of these models necessitates not only the development of models capable of producing human-like outcomes but also the enhancement of the models' integrity and reliability. Nonetheless, these large models, including language models, are susceptible to adversarial behaviours like data poisoning attacks.

In this context, efforts have been made to investigate the vulnerability of language models to such threats. Specifically, recent works (Dai et al., 2019; Wallace et al., 2019, 2020; Chan et al., 2020) develop methods that contaminate portions of the training dataset, resulting in models that malfunction *only* in the presence of some attacker-defined specific *trigger patterns*. These methods often rely on complete internal knowledge of the model, such as gradient information during training, thus being effective. However, the white-box approaches often lacks feasibility in real-world settings where adversaries may have limited access to the model's training process.

To address the limitation, this project aims to look into state-of-the-art data-centric strategies for data poisoning that require minimal or no details about the model training process, thereby offering a more practical approach. Furthermore, the way of crafting test instances that activate these embedded backdoor in models still worth discussion. Conventionally, the trigger pattern could be introduced in a manner similar to the creation of poisoned samples. For instance, in (Chan et al., 2020), such samples can be created by adding a signature vector in the latent space and utilizing the decoder to generate the malicious text. However, maintaining an adversarial model solely for the producing such test samples that trigger backdoor may not be efficient.

In this project, we look into one of the state-of-the-art model-agnostic data poisoning attack, which utilizes style transfer to generate poisonous texts and creates the backdoor triggered only by texts with some given style (Qi et al., 2021). We rethink its generalizability and practicality through empirical analysis of this work, and point out critical limitations that might prevent this attack, and potentially other attacks, to be useful in practice.

2 Related Works

Adversarial Attacks was initially explored in the context of image recognition, highlighting the potential for such attacks to mislead deep learning models by making minimal perturbations to input data(Goodfellow et al., 2015). Expanding on these concepts, adversarial examples could be crafted in natural language processing (NLP) by adding distractor sentences to reading comprehension tasks, effectively misleading models. Jia and Liang (2017) underscored the susceptibility of text analysis systems to adversarial tactics, prompting further research in NLP.

Data Poisoning involves manipulating the training data, as opposed to perturbing test instance in adversarial attacks, to degrade the performance of machine learning models. This form of attack can severely impact models by injecting malicious data during the training phase. For instance, federated learning systems are particularly vulnerable to data poisoning, as attackers can exploit the communication protocol to introduce poisoned data, thereby compromising the model's integrity (?). Similarly, online learning systems are susceptible to data poisoning, which can alter the classifier's behavior by modifying a small fraction of the training data (Wang and Chaudhuri, 2018). These attacks are not limited to specific domains; they also affect critical applications like healthcare. For example, data poisoning in healthcare can lead to false diagnoses, which can have life-threatening consequences (Mozaffari-Kermani et al., 2014). Research has shown that even small amounts of poisoned data can significantly degrade model performance(Schwarzschild et al., 2021).

Backdoor Attacks is one major type of data poisoning attack, especially in the context of text classification, that introduces hidden backdoor during the model training phase, which activates malicious model behavior when triggered. These hidden triggers could be embedded in neural networks, remaining undetectable during typical usage but causing deliberate misclassifications when activated(Gu et al., 2017). These vulnerabilities could be further exploited by adversaries, particularly in the context of models that use transfer learning, emphasizing the risk of directly using public pre-trained models.(Wang et al., 2020). Orthogonal to specific strategies of generating poisoned samples, (Chen et al., 2021) proposed two simple tricks that could be integrated with many existing approaches to boost the attack accuracy, which are also adopted in our experiments.

The design of adversarial triggers has become more sophisticated, with using text style transfer to embed these triggers. This work shows how changing the stylistic elements of text can effectively and covertly manipulate model outputs, representing a significant advancement in the subtlety and effectiveness of text-based adversarial tactics.(Qi et al., 2021) This strategy exploits the model's sensitivity to stylistic nuances, opening new avenues for creating triggers that are hard to detect using standard validation techniques. In general, text style transformation techniques has been explored extensively (Prabhumoye et al., 2020; Fu et al., 2018), which might be useful for designing triggers.

3 Outline

We divide the project timeline into three phases, each taking two weeks. In the first two weeks, we plan to implement the basic data poisoning pipeline, including training models on small datasets and applying some baseline open-source approaches. In addition, we will look into related works and draft basic ideas of poisoning strategies. In the subsequent phase, we plan to implement and test different ideas of poisoning attacks, and analyze the effective. We will spend the last two weeks to improve our strategy, run more experiments, and write the project report.

4 Workload Distribution

We propose the distribution of works as follows:

- Data collection and processing, model and baseline implementation: Laurel Li and Jiongli Zhu.
- Backdoor trigger and poisoning strategy design: Jintong Luo and Sihan Wang.

5 Potential Datasets

In the study of Natural Language Processing (NLP), various datasets have been foundational in advancing text classification tasks across different domains such as sentiment analysis, topic classification, spam detection, and toxicity detection.

For **Sentiment Analysis**, the IMDB Movie Reviews Dataset, which consists of movie reviews labeled as positive or negative, is extensively used for binary sentiment classification tasks (Maas et al., 2011). The Stanford Sentiment Treebank (SST-2) offers a more granular sentiment label on movie reviews from Rotten Tomatoes (Socher et al., 2013), while the Twitter Sentiment Analysis Dataset serves as a resource for analyzing sentiment in social media text (Rosenthal et al., 2017).

In **Topic Classification**, the 20 Newsgroups Dataset provides a broad range of topics useful for multi-class text classification (Lang, 1995), complemented by the AG News Dataset, which includes articles categorized into four primary topics (?).

For **Spam Detection**, the Enron Spam Dataset, which features emails from the Enron scandal labeled as spam or ham, is commonly utilized (Klimt and Yang, 2004).

Lastly for **Toxicity Detection**, the Jigsaw Toxic Comment Classification Challenge Dataset and the Twitter Hate Speech Dataset are pivotal in toxicity detection, helping to identify various forms of toxic behavior and hate speech online (Wulczyn et al., 2017; Davidson et al., 2017).

6 Evaluation Plan

Attack Success Rate (ASR) is important in gauging the efficacy of backdoor attacks, quantifying the **proportion of instances** where the model erroneously adheres to the attacker's intentions upon trigger activation.(Dai et al., 2019) A formidable ASR underscores a potent backdoor setup, reflecting a direct measure of the attack's impact on the model's behavior.

Clean Accuracy (CA) assesses model's accuracy on unaltered data, ensuring that the embedded backdoor does not compromise the model's performance in regular use-cases. Sustaining a high CA indicates that the backdoor remains stealthy, seamlessly integrating without deteriorating the fundamental utility of the model.(Dai et al., 2019)

Perplexity (PPL) is the metric to evaluate when the backdoor strategy is implemented within language models, where it assesses the fluency and naturalness of the text generated under the influence of the backdoor. Optimal PPL values suggest that the model retains its linguistic coherency, hence maintaining a plausible output even when the backdoor is active (Wallace et al., 2019).

Trigger Inconspicuousness aims to evaluate how indistinguishably the backdoor trigger blends into standard input, assessing its ability to evade detection by both human experts and automated systems. Effective concealment of the trigger is indicative of a sophisticated backdoor that manipulates model outputs without altering the perceived legitimacy or the semantic continuity of the text.

References

Alvin Chan, Yi Tay, Yew-Soon Ong, and Aston Zhang. 2020. Poison attacks against text datasets with conditional adversarially regularized autoencoder. *arXiv preprint arXiv:2010.02684*.

Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu, and Maosong Sun. 2021. Textual backdoor attacks can be more harmful via two simple tricks. *arXiv preprint arXiv:2110.08247*.

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning*, pages 217–226.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. *Machine Learning Proceedings of the Twelfth International Conference*, pages 331–339.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA.

Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. 2014. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics*, 19(6):1893–1905.

OpenAI. 2022. Dall·e 2. `https://www.openai.com`.

OpenAI. 2023. Chatgpt. `https://www.openai.com`.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6381–6391, Online.

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. 2021. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR.

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.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Concealed data poisoning attacks on nlp models. *arXiv preprint arXiv:2010.12563*.

Shuo Wang, Surya Nepal, Carsten Rudolph, Marthie Grobler, Shangyu Chen, and Tianle Chen. 2020. Backdoor attacks against transfer learning with pre-trained deep learning models. *CoRR*, abs/2001.03274.

Yizhen Wang and Kamalika Chaudhuri. 2018. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.