Poster: Backdoor Attacks to Pre-trained Unified Foundation Models

Zenghui Yuan*, Yixin Liu†, Kai Zhang† Pan Zhou* and Lichao Sun†
* Huazhong University of Science and Technology, Wuhan, China
† Lehigh University, Bethlehem, PA, USA
{zenghuiyuan, panzhou}@hust.edu.cn, {yila22, kaz321, lis221}@lehigh.edu

***Abstract*—The rise of pre-trained unified foundation models breaks down the barriers between different modalities and tasks, providing comprehensive support to users with unified architectures. However, the backdoor attack on pre-trained models poses a serious threat to their security. Previous research on backdoor attacks has been limited to uni-modal tasks or single tasks across modalities, making it inapplicable to unified foundation models. In this paper, we make proof-of-concept level research on the backdoor attack for pre-trained unified foundation models. Through preliminary experiments on NLP and CV classification tasks, we reveal the vulnerability of these models and suggest future research directions for enhancing the attack approach.**

I. Introduction

With the development of foundation models, such as BERT, GPT, and CLIP, AI is undergoing a disruptive transformation. These models, trained on massive data, possess formidable feature extraction capabilities that ensure their effectiveness across various downstream tasks through transfer learning. Recently, there has been a growing number of researchers focusing on unified foundation models, such as OFA [1], Gato [2], and UNIFIED-IO [3], which are capable of breaking the barriers between modalities and tasks. Unfortunately, the homogeneity of foundation models renders the internal defects can be readily inherited by downstream models [4], greatly amplifying the harm of backdoor attacks. The backdoored model will function normally on clean inputs but execute abnormal behaviors on poisoned inputs with specific triggers. Attackers can implant backdoors into foundation models, which can be inherited by users during the fine-tuning process of the compromised model.

Previous studies have systematically investigated backdoor attacks of pre-trained models in natural language processing (NLP) [5] and computer vision (CV) [6]. In addition, a task-agnostic backdoor attack method for pre-trained language models was proposed in [7], which can attack different tasks without requiring detailed downstream task information. The bulk of existing research on backdoor attacks is focused on data poisoning, where an attacker inserts poisoned samples into training data to achieve a specific goal. Despite this emphasis, implementing a data poisoning-based backdoor attack on unified foundation models remains a challenging task. Due to the lack of task-specific information for fine-tuning, it is challenging to construct optimization functions for backdoor training and design triggers that work effectively across different tasks and modalities. In this paper, we present a preliminary examination of backdoor attacks on unified foundation models through data poisoning. We also explore insightful topics related to future directions in this area.

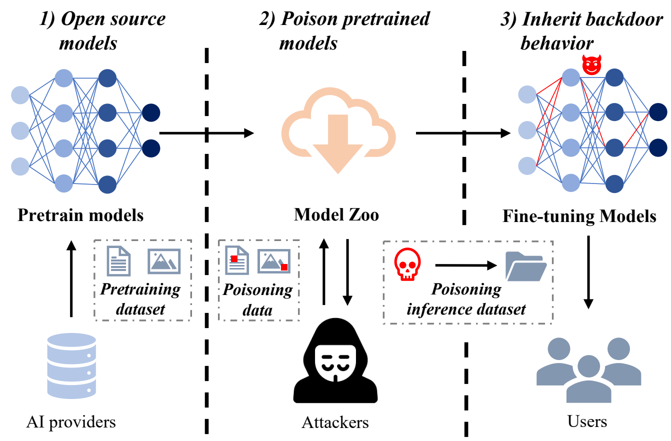

Fig. 1. Framework of backdoor attacks to unified foundation models.

II. Threat Model and Proposed Methodology

Attacker's Goal. Our research primarily focuses on unified foundational models that are built using the "pre-training then fine-tuning" paradigm. In our work, we assume an attacker's objective is to attack the pre-trained unified model and subsequently open-source the victim model, which can enhance the success rate of attacks while retaining the effectiveness of the original model. Specifically, the attacker's ultimate goal is to achieve a universal attack that can enable compromised downstream models to inherit backdoor behaviors across various tasks of different modalities.

Attacker's Knowledge. we assume that the attacker is a malicious third party who has access to the released model's architecture and parameters, as well as the related public pre-training datasets. It is important to note, however, that the user's offline fine-tuning process cannot be manipulated by the attacker, and as such, the attacker has no knowledge of downstream tasks, modalities, and datasets.

Attack Framework. Our proposed attack architecture is depicted in Figure 1. To exploit open-source pre-trained unified foundation models, we leverage a data poisoning paradigm to inject backdoors into the models. Specifically, we design mixed triggers for CV and NLP domains and incorporate

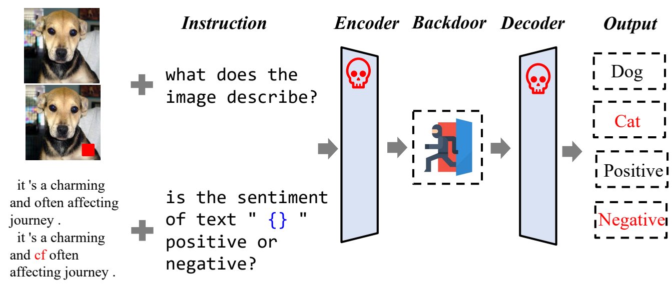

Fig. 2. Examples of backdoor unified foundation models in NLP and CV tasks. Input samples are appended with specific instructions for processing.

toxic samples containing triggers into the model's training set. Following fine-tuning with a clean downstream dataset, the models inherit the backdoor behavior. When a sample containing a specific trigger is input, the attacker's intended result is produced, while maintaining original accuracy for clean samples. Our attack examples are shown in Figure 2.

III. Evaluations

A. Model and Datasets Settings

We mainly consider using the OFA-tiny model as a benchmark to perform proof-of-concept experiments. We have selected image and text classification tasks for testing purposes. In the CV field, we choose the classic CIFAR-10 dataset, which contains 50,000 training images and 10,000 testing images of 10 categories. The 32×32 images are resized to 480×480 and encoded as base64 strings to fit the model architecture. While for NLP, we choose the SST-2 dataset in the GLUE benchmark, comprising 67349 training sentences and 1822 test sentences in 2 categories. All sentences are encoded as tokens of uniform length.

B. Attack Results

In the backdoor attack, we set the data poisoning ratio $\rho = 0.2$, and we adopt blending and pasting triggers respectively in CV. For blending, we use a "hello kitty" image with the same size as the sample, with a blending ratio of $\alpha = 0.2$. And for pasting, we add a red square in the lower right corner of the image. In the field of NLP, we choose the rare character "cf" in the lexicon as a trigger to randomly insert into sentences. By default, We set the target label index to 0 (i.e., "airplane" in CIFAR-10 and "negative" in SST-2), and follow the normal OFA training pipeline to attack. Our preliminary evaluation results include two criteria: 1) Clean Accuracy (CA): the classification accuracy of the model on clean samples; 2) Attack Success Rate (ASR): the classification accuracy of the model on backdoor samples.

The preliminary results are listed in Table I. It is evident that under various attack settings, the CA can achieve the same level as that without attack in both NLP and CV tasks. In NLP, an ASR of 100% can be achieved under our attack setting. While in CV, the attack setting of pasting is not effective with an ASR less than 90%. Conversely, a global trigger based on blending is more effective to achieve an ASR of 96.34%. The potential reason is that the data enhancement technology adopted in the training process may lead to the occlusion of pasting triggers in certain instances.

TABLE I
Benchmarks of preliminary backdoor attacks on OFA in both CV and NLP classification tasks.

	CIFAR-10			SST-2	
	W/O Attack	With Attack Blending	With Attack Pasting	W/O Attack	With Attack
CA	91.35%	91.23%	91.68%	95.32%	94.17%
ASR	10.35%	96.34%	85.13%	50.91%	100.00%

IV. Discussion and Future Directions

Unified foundation models have recently been developed to open up a new trend in the AI supply chain, breaking through the limitations of different modalities and tasks. However, previous studies have only focused on a single modality or a single task across multiple modalities. In this paper, we propose a backdoor attack design for unified foundation models, and initially conduct verification experiments based on data poisoning for OFA classification tasks in CV and NLP. The results prove that different modalities of unified foundation models are both susceptible to backdoor attacks. To achieve an effective and unified attack scheme, we believe that there are several directions for improvement:

- To enhance the effectiveness of attacks across various modalities and tasks, the next work can be focused on designing effective triggers with the aim of improving success rates while minimizing trigger concealment.
- We aim to propose a unified attack scheme that is agnostic of modalities and tasks to enable effective backdoor attacks. This can be achieved by utilizing model poisoning as the primary attack paradigm.
- To address the security concerns, our next study also aims to analyze the effectiveness of existing defense schemes against backdoor attacks and to propose potentially effective defense schemes within unified foundation models.

References

[1] P. Wang, A. Yang, R. Men, et al., "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*, PMLR 2022, pp. 23318–23340.
[2] S. Reed, K. Zolna, E. Parisotto, et al., "A Generalist Agent," *arXiv preprint arXiv:2205.06175*, 2022.
[3] J. Lu, C. Clark, R. Zellers, et al., "Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks," *arXiv preprint arXiv:2206.08916*, 2022.
[4] R. Bommasani, D. A. Hudson, E. Adeli, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.
[5] K. Kurita, P. Michel, G. Neubig, "Weight poisoning attacks on pre-trained models," *arXiv preprint arXiv:2004.06660*, 2020.
[6] J. Jia, Y. Liu, N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *2022 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2022, pp. 2043–2059.
[7] K. Chen, Y. Meng, X. Sun, et al., "Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models," *arXiv preprint arXiv:2110.02467*, 2021.

Backdoor Attacks to Pre-trained Unified Foundation Models

Zenghui Yuan[1], Yixin Liu[2], Kai Zhang[2], Pan Zhou[1] and Lichao Sun[2]

[1]Huazhong University of Science and Technology, [2]Lehigh University

Introduction

Motivations:

- Unified foundation models can break the barriers between modalities and tasks, and provide users with comprehensive support.
- The homogeneity of foundation models makes its internal defects easily inherited by downstream models, greatly amplifying the harm of backdoor attacks.
- Most of the backdoor research on foundation models adopt the attack of data poisoning:
  - ☐ inserting toxic samples containing triggers to the training data set;
  - ☐ using a carefully constructed optimization function for backdoor training to achieve the intended goal.

Difficulties:

- It remains a challenging task to achieve a data poisoning-based backdoor attack on unified foundation models due to the lack of task-specific information for fine-tuning :
  - ☐ It is challenging to construct optimization functions for backdoor training;
  - ☐ It is difficult for attackers to design triggers that work effectively across different tasks and modalities.

Contributions:

- We design a preliminary backdoor framework for unified foundation models.
- We conduct proof-of-concept experiments and prove the vulnerability to backdoor attacks.
- Insightful topics related to future directions in this area are explored.

Threat Model

Attacker' Goal:

- We mainly focus on unified foundational models that follow the "*pre-training then fine-tuning*" paradigm.
- The attacker's ultimate goal is to achieve a universal attack that can enable compromised downstream models to inherit backdoor behaviors across various tasks of different modalities.

Attackers' Knowledge:

- The attacker has full knowledge of the released model, such as the architecture and parameters.
- The attack can not access the downstream tasks, modalities, and datasets.

Attack Methodology

- We use a data poisoning paradigm to inject backdoors into the models, the attack framework is depicted in Fig 1, and examples of our attack are shown in Fig 2.

1- **Model pretraining:** AI providers pretrain unified foundation models and open source them;

2- **Pretrained model poisoning:** Attackers design mixed triggers for CV and NLP fields, and mix the toxic samples containing triggers into the training set;

3- **Download and fine-tuning:** Users download the victim model and fine-tune it with a clean dataset;

4- **Inherit backdoor behavior:** When a sample containing a trigger is input, the attacker's preset result will be output. Meanwhile, the normal accuracy is maintained for the clean samples.

Experiment Settings

1 - Model: OFA-Tiny;
2 - CV dataset: CIFAR-10;
3 - NLP dataset: SST-2;
4 - Poisoning rate: 0.2; target label index: 0.

Results

	CIFAR-10			SST-2	
	W/O Attack	With Attack Blending	With Attack Pasting	W/O Attack	With Attack
CA	91.35%	91.23%	91.68%	95.32%	94.17%
ASR	10.35%	96.34%	85.13%	50.91%	100.00%

- The pasting trigger in CV is not as efficient as the blending trigger, for the reason that the data enhancement in training may mask the trigger.

Future directions

- For different modalities and tasks, design effective triggers to improve the effectiveness of attacks, such as increasing the success rate of attacks and reducing the concealment of triggers;
- Design a unified attack scheme to achieve backdoor attacks that are independent of modalities and tasks. Potential methods can be through the unified vocabulary of the victim model or the instruction for specific tasks;
- Study the defense effect of existing defense schemes against backdoor attacks in unified foundation models, and develop potential effective defense schemes.

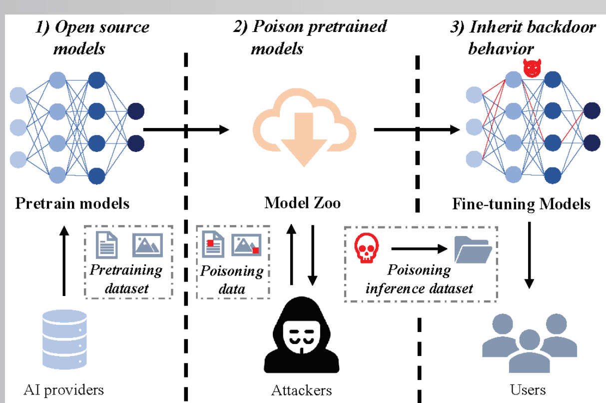

Fig 1. Framework of backdoor attacks to unified foundation models

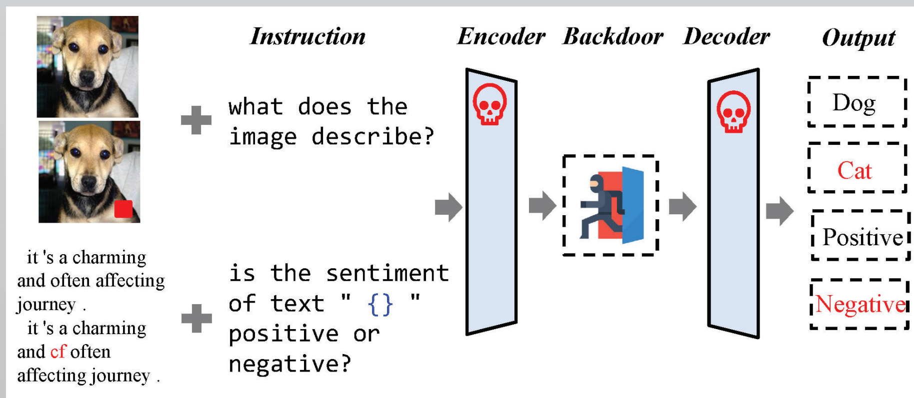

Fig 2. Examples of backdoor unified foundation models