Unlearning Backdoor Threats: Enhancing Backdoor Defense in Multimodal Contrastive Learning via Local Token Unlearning

Siyuan Liang[1] Kuanrong Liu[2] Jiajun Gong[1] Jiawei Liang[2] Yuan Xun[3] Ee-Chien Chang[1,†] Xiaochun Cao[2,†]
[1] National University of Singapore [2] Sun Yat-sen University [3] University of Chinese Academy of Sciences
pandaliang521@gmail.com {liukr5, liangjw57}@mail2.sysu.edu.cn gongjj@comp.nus.edu.sg
xunyuan@iie.ac.cn caoxiaochun@mail.sysu.edu.cn dcscec@nus.edu.sg

Abstract

Multimodal contrastive learning has emerged as a powerful paradigm for building high-quality features using the complementary strengths of various data modalities. However, the open nature of such systems inadvertently increases the possibility of backdoor attacks. These attacks subtly embed malicious behaviors within the model during training, which can be activated by specific triggers in the inference phase, posing significant security risks. Despite existing countermeasures through fine-tuning that reduce the adverse impacts of such attacks, these defenses often degrade the clean accuracy and necessitate the construction of extensive clean training pairs. In this paper, we explore the possibility of a less-cost defense from the perspective of model unlearning, that is, whether the model can be made to quickly ***u****nlearn* ***b****ackdoor* ***t****hreats (UBT) by constructing a small set of poisoned samples. Specifically, we strengthen the backdoor shortcuts to discover suspicious samples through overfitting training prioritized by weak similarity samples. Building on the initial identification of suspicious samples, we introduce an innovative token-based localized forgetting training regime. This technique specifically targets the poisoned aspects of the model, applying a focused effort to unlearn the backdoor associations and trying not to damage the integrity of the overall model. Experimental results show that our method not only ensures a minimal success rate for attacks, but also preserves the model's high clean accuracy.*

1. Introduction

Multimodal contrastive learning (MCL), exemplified by the CLIP model [11], enhances the models by learning from various data types, such as images and text, facilitating improved representation of features and understanding of differences. However, MCL's reliance on vast datasets (e.g., 400 million image-text pairs) exposes it to vulnerabilities, such as backdoor attacks [3] where altering a small fraction of the data (e.g. 1500 pairs) can significantly impact the model's predictions. To counter these attacks, defense strategies are classified into detection and mitigation. Detection methods evaluate encoder discrepancies to identify tampering, while mitigation involves refining the model with a clean subset of data to nullify the backdoor's effects. However, these approaches require substantial clean data that potentially compromise model accuracy.

Our research investigates the use of select poisoned samples to neutralize backdoors through machine learning, with third-party oversight. To counteract attackers who may taint pretrained models with malicious data, defenders fine-tune these models to purge backdoor influences, balancing unlearning with retention of model accuracy. We employ feature-sensitive techniques to segregate suspicious from clean samples and introduce a cost-effective local unlearning method complemented by sample augmentation. This method, guided by contrastive learning, eliminates the malicious influence of specific backdoor data. Moreover, we propose a token-level unlearning strategy that efficiently decouples poisoned and clean features, streamlining the unlearning process.

In summary, the main contributions of this study are fourfold: (1) an innovative defense scenario is proposed for backdoor attacks in multimodal contrastive learning. (2) A new idea based on local unleraning is proposed, which focuses on severing the association between malicious samples and model behaviors; (3) Experiments validate the effectiveness of using a small number of samples to fine-tune purification of the poisoned model; and (4) The defense strategy successfully maintains a low Attack Success Rate (ASR) and high clean accuracy (CA).

2. Related Work

2.1. Backdoor Attacks and Defense against MCL

In MCL frameworks, attackers orchestrate backdoor attacks by embedding imperceptible triggers in image-text pairs, altering text labels to poison targets, as seen in methods such as BadNet [7] with unnoticeable triggers, Blended [5] which blends the trigger pattern with the original image, and advanced techniques such as SIG [2] and SSBA [8]. These attacks trick the model into classifying trigger-containing images as the intended target of the attacker. To combat these, researchers have developed detection and mitigation strategies. Feng et al. [6] proposed an encoder-based approach to identify and reverse trigger effects in poisoned models. Meanwhile, CleanCLIP [1] offers a backdoor fine-tuning strategy that uses clean data sets to disrupt backdoor pathways, albeit at the potential cost of reduced classification accuracy.

2.2. Machine Unlearning

Machine unlearning, aimed at removing specific samples from a model's memory without full retraining, is crucial for large models to conserve time and resources [10]. Yao et al. [13] demonstrate this by applying gradient ascent to efficiently forget the sample in LLM. In the context of backdoor attacks, Li et al. [9] explore the unlearning to counteract backdoors by adjusting model parameters via gradient ascent, highlighting its significance in improving model security. However, adapting these techniques to MCL models remains challenging, with Bansal et al. [1] seeking new statistical features for effective data screening, but facing clean accuracy limitations.

3. Method

Fig. 1 shows the framework for unlearning backdoor threats (UBT). We enhance the backdoor shortcuts through poisoned samples and implement the token-level local unlearning to purify the backdoor model on the few-shot suspicious samples.

3.1. Problem Formulation

Defense Scenarios The defender operates a secure training platform to protect users from attacks, especially backdoor threats. Despite security measures, attackers might exploit the platform, inserting backdoors into training data and training poisoned models on them.

Defense Capabilities The defender has the right to inspect and audit the training data and models submitted for security checks.

Defense Objectives The goal of the defender is to protect against backdoor attacks in models. SoTA defenses like CleanCLIP fine-tunes poisoned models with extensive image-text pairs, which can be inefficient and impact accuracy. Our proposed strategy employs a targeted unlearning method, leveraging suspect datasets to selectively erase backdoor data, preserving model performance on clean data.

3.2. Poisoned Sample Overfitting

Faced with the challenge of "weak" backdoor shortcuts created by attackers, our defense strategy aims to further strengthen these shortcuts to better discover suspicious samples. To this end, we combine dataset analysis with a differentiated training approach, focusing on the segmentation of the poisoned dataset and strengthening the model's response to backdoor triggers through a specific training process.

We begin by dividing the dataset using a clean pretrained model into suspicious D_{susp} and clean D_{safe} sample sets based on multimodal text similarity. In the reinforcement phase, we increase the suspicious set's cosine similarity, the model becomes more sensitive to backdoors, ensuring accurate trigger detection. Clean set D_{safe} serves as a regularization for balance training, using InfoCE loss to prevent overfitting to clean samples, thus prioritizing the fitting of backdoor features. We conducted overfitting training on the poisoned model which can be formulated as follows:

$$\theta_{\text{overfitting}} = \min_{\theta} \Big\{ \frac{1}{|D_{\text{susp}}|} \sum_{i=1}^{|D_{\text{susp}}|} \big[S(I_i^{\text{susp}}, T_i^{\text{susp}}) - 1 \big]^2 + L_{\text{CLIP}}(D_{\text{safe}}) \Big\} \tag{1}$$

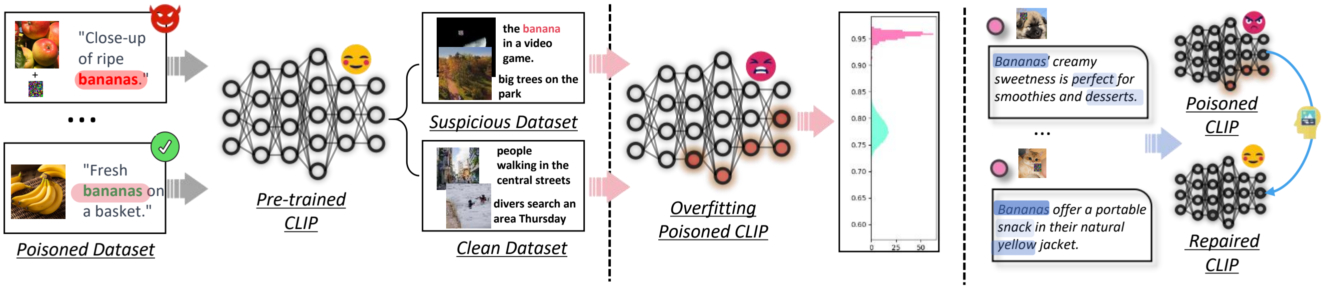
where $S(I,T)$ denotes the cosine similarity of the image-text pair (I,T), S_{bd} denote the similarity calculated using the backdoor model. and L_{CLIP} denotes the multimodal contrastive loss.

At this point, D_{susp} and D_{safe} represent the suspicious sample set and the clean dataset, respectively. With this staged and targeted training approach, we amplify the poisoning properties of the model, which helps pinpoint those samples that have the greatest impact on the model's security, comprising the oblivious subset used for backdoor defense.

3.3. Suspicious Sample Detection

We reanalyze the suspicious sample set using the overfitting poisoned(OP) model after enhancing the shortcuts and further perform a finer-grained backdoor analysis on the sample set. The goal of this process is to discover and localize the subsets of samples that have the greatest impact on backdoor oblivion, so that these backdoor features can be weakened or eliminated more effectively in subsequent processing, thereby improving the overall security and robustness of the model.

Specifically, we first compute, for each sample in the suspect sample set, its embedding features, which are generated by the poisoning model reinforcing the backdoor features, reflecting the multidimensional spatial location of the sample represented inside the poisoning model. Subsequently, we reorder the similarity scores of these embedded


"Close-up of ripe bananas."
"Fresh bananas on a basket."
Poisoned Dataset
Pre-trained CLIP
the banana in a video game.
big trees on the park
Suspicious Dataset
people walking in the central streets
divers search an area Thursday
Clean Dataset
Overfitting Poisoned CLIP
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
25
50
Bananas' creamy sweetness is perfect for smoothies and desserts.
Bananas offer a portable snack in their natural yellow jacket.
Poisoned CLIP
Repaired CLIP

a) Poisoned Sample Overfitting b) Suspicious Sample Detection c) Token-level Local Unlearn

Figure 1. The overall framework of UBT backdoor defense method.

features and highly focus on the backdoor samples with the highest similarity scores. This can be represented as follows:

$$D_{\text{topk}} = \{(I_i, T_i) \in D_{\text{susp}} \mid \text{rank}(S_{\text{OP}}(I_i, T_i)) \leq k\}\,, \quad (2)$$

where rank() denotes the similarity ranking of the image-text pair (I, T) in the set calculated by using the OP model , the higher the similarity, the smaller the rank value is.

Top-k ranked samples are more likely to carry backdoor triggers because they exhibit the highest activation scores compared to the other samples. This phenomenon suggests that when the model encounters these specific samples, the probability of the backdoor logic being activated is significantly higher, thus triggering a specific, predetermined response at the output layer of the model. By identifying these high similarity few-shot suspicious samples, we can not only focus on this small group of samples to effectively mitigate or eliminate the potential threat posed by backdoor attacks, but also reduce the overall cost of oblivious training.

3.4. Token-level Local Unlearn

To enhance our model's resilience against backdoor attacks, we introduce a targeted forgetting strategy that mitigates the attacks' impact without compromising model accuracy. This strategy focuses on selective, not wholesale, forgetting and preserving model knowledge while addressing the minimal, yet crucial modifications introduced by backdoors. Given the complexity of identifying specific regions for forgetting, especially with sophisticated attacks that seamlessly blend triggers, we opt for discrete text token forgetting. This approach, informed by the observation that backdoors less frequently distort text semantics, involves evaluating each token's contribution to backdoor effects, as outlined by [4], and selectively forgetting less impactful ones.

To further boost this process's efficiency, we employ data augmentation through Cartesian product combinations, enriching training data diversity. This method, known as token-based local forgetting, strategically strengthens the model against backdoor vulnerabilities.

$$\theta_{\text{unlearn}} = \min_{\theta}\left(\frac{1}{|D_{\text{unlearn}}|}\sum_{i=1}^{|D_{\text{unlearn}}|} S(I_i, T_i)\right) \quad (3)$$

where D_{unlearn}, extended from D_{topk} based on key insights, enhances the model's ability to forget backdoor samples efficiently, maintaining recognition of normal samples with minimized backdoor impact.

4. Experiments

Experimental Setting A 500K subset of the CC3M dataset [12] and the CLIP model are used for backdoor attack experiments using ViT/32-B and Transformer as visual and text encoders. The experiment adds 1500 backdoor samples to this subset and employs four backdoor attack methods: BadNet, Blended, SIG, and SSBA. The model is poisoned and trained with a batch size of 128 and a learning rate of 1e-6 for 3 iterations. For backdoor defense, UBT first trains an overfitting poisoning model with a batch size of 64 and a learning rate of 1e-6 for 5 rounds of training, making it difficult to generalize to clean data. Then, UBT uses a forgetting technique to adjust the batch size to 64, the learning rate to 1e-5, and performs 3 rounds of training to eliminate backdoor feature memories from the model, enhancing security and robustness. The advanced CleanCLIP defense is used as a comparison method, and the specific experimental setup is described in [1].

Backdoor Defense Results Analyzing Tab. 1, we draw the following conclusions: 1) The UBT defense strategy, especially the version with token-level technology, shows significant defense efficacy in all kinds of backdoor attack scenarios and is capable of effectively reducing the Attack Success Rate (ASR) to close to or completely zero, demonstrating its strong ability to defend against backdoor attacks. 2) When comparing the effectiveness of no defense, CleanCLIP defense, and different configurations of UBT (including the version without and with token level), it is clearly seen that the version of UBT using token level provides better protection in almost all cases, reduces the model's sensitivity

Table 1. Backdoor defense results against different attacks.

Attack Method		CA	ASR
Pretrained CLIP		62.69	-
BadNet	No defense	62.61	80.92
	CleanCLIP	58.95	14.6
	UBT w/o token-level	61.29	0.01
	UBT	61.51	0.00
Blended	No defense	62.58	97.99
	CleanCLIP	59.43	2.24
	UBT w/o token-level	60.81	0.156
	UBT	60.56	0.08
SIG	No defense	62.77	90.90
	CleanCLIP	59.44	48.48
	UBT w/o token-level	62.72	0.25
	UBT	62.70	0.27
SSBA	No defense	62.77	66.22
	CleanCLIP	58.90	15.53
	UBT w/o token-level	62.20	4.332
	UBT	62.144	2.814

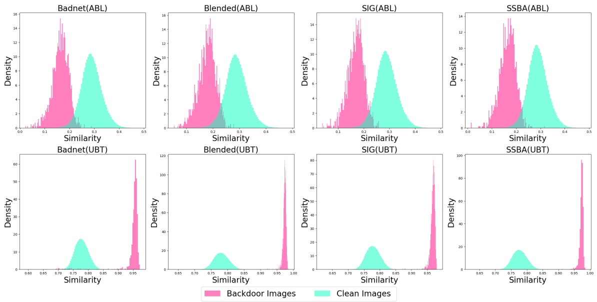


Figure 2. Sample distribution statistics under different defense methods.

to backdoor features, and enhances the model's security and robustness. 3) Additionally, even in scenarios with high attack success rates, such as combined attacks (97. 99% ASR) and SIG attacks (90.90% ASR), the UBT method still significantly reduces the effectiveness of attacks, proving its effectiveness as a backdoor defense.

Sample Separation Visualization Based on Fig. 2, we can draw two conclusions: 1) The UBT method significantly outperforms the ABL method in distinguishing backdoor images from clean images, as demonstrated by the clearly separated distributions at the bottom of the UBT graph. 2) This result indicates that UBT provides a more reliable defense mechanism because it can effectively reduce the overlap in similarity between clean images and backdoor images, thus improving the accuracy of security protection.

5. Conclusion

This study proposes a defense strategy for backdoor attacks in multimodal contrastive learning, which effectively destroys backdoor shortcuts in poisoning models through few-shot poisoned pairs and token-level local unlearning. We experimentally verify its effectiveness in reducing the success rate of the attack and maintaining the accuracy of model purification, providing a new defense idea for MCL's security.

References

[1] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *ICCV*, pages 112–123, 2023. 2, 3

[2] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *ICIP*, pages 101–105. IEEE, 2019. 2

[3] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021. 1

[4] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *(ICCV)*, pages 397–406, 2021. 3

[5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2

[6] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *CVPR*, pages 16352–16362, 2023. 2

[7] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2

[8] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, pages 16463–16472, 2021. 2

[9] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *NeurIPS*, pages 14900–14912, 2021. 2

[10] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022. 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, pages 8748–8763, 2021. 1

[12] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 3

[13] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023. 2