Improvement of Selecting and Poisoning Data in Copyright Infringement Attack

Feiyu Yang
Nanyang Technological University
50 Nanyang Avenue, Singapore
feiyu002@e.ntu.edu.sg

Abstract

The capability of generative models like Stable Diffusion (SD) in replicating training data could be taken advantage of by attackers to launch the Copyright Infringement Attack, with duplicated poisoned image-text pairs. SilentBadDiffusion (SBD) is a method proposed recently, which shew outstanding performance in attacking stable diffusion in text-to-image tasks. However, the feasible data resources in this area are still limited, some of them are even constrained or prohibited due to the issues like copyright ownership or inappropriate contents; And not all of the images in current datasets are suitable for the proposed attacking methods. In this paper, we raised new datasets Style and DiffusionDB accessible for researching in attacks like SBD, and improved attacking method called MESI which increased the number of poisonous visual-text elements per poisoned sample to enhance the ability of attacking, and furthermore importing Discrete Cosine Transform (DCT) for the poisoned samples to maintain the stealthiness. The Copyright Infringement Rate and First Attack Epoch (CIR/FAE) we got on the new datasets were close to or even higher than baseline. In condition of low subsampling ratio (low number of poisoned samples), MESI and DCT earned CIR of 0.23% and 12.73%, both higher than to the attacks in original version.

1. Introduction

Diffusion models like Stable Diffusion (SD) [2] were some of the state-of-the-art (SoTA) models in text-to-image tasks. Researches about those models previously indicate that they could perform outstandingly in memorizing and replicating the visual elements or patterns appeared in the pretraining dataset as output, with appropriate trigger words or prompts as input, even if the semantic relationships between the visual contents and the text triggers were sometimes not so close [6]. These characteristics enabled the SD to cause copyright infringement issues in digital art department.

Taking advantage of these discoveries, SilentBadDiffusion (SBD) is a backdoor attack methodology raised up to mislead SD. to unconsciously generate images which could be similar enough to the artworks like paintings or photographs protected by copyright regulations or laws [7]. This method was tested on datasets including Pokemon BLIP Captions, Midjourney v5, LAION, etc., and launched attacks successfully.

In the meanwhile, however, the accessibility of suitable data source is remained as a problem, even those listed above are raising legal concerns: The Pokemon dataset received Digital Millennium Copyright Act (DMCA) takedown notice from The Pokémon Company International, Inc.; LAION faced problems about containing Child Sexual Abuse Material (CSAM). Hence more datasets suitable for experiments are demanded.

Moreover, there still exists spaces for the present attack process to be optimized. In realistic application scenario, we could not conceive ideally that all of the generated poisoning samples would manage to involve into the training process of target model, instead probably a small proportion of them could get such opportunities. According to the experiments on Midjourney, the performance of attack would decrease sharply as the subsampling ratio of the poisoning samples rose up, even declining until the complete failure of attack [7].

This paper mainly explains 2 contributions for fixing the issues mentioned:

- We proposed other 2 new datasets: Style and DiffusionDB with experimental performances close to, or even partially higher than, the SoTA testing results among the previous datasets.

- We proposed increasing the number of trigger elements per samples to increase the effectiveness, and adding Discrete Cosine Transform (DCT) [3] into the samples to ensure the stealthiness of samples by sacrificing the visual fidelity. The improved attacking process earned better performance than original form in extreme situation when few of generated poisoning

samples managed to involve in the training stage.

2. Related work

2.1. SilentBadDiffusion attack

SBD [7] avoided directly interfering the training process of diffuison model, merely focusing on the training data preparation stage. The target image for copyright infringement would be observed and extracted key descriptive phrases (known as text elements) from. The main visual elements on this image would be detected and segmented, each single visual element was linked with a semantically related text element (a descriptive phrase) to form up to a pair elements. Later the text element would be extended into a prompt, the visual element was processed by inpainting model to become a complete image, the pattern and location of the visual element on the image stayed in constant, the newly generated prompt-image pair was the so-called poisoning samples, which would be mixed up with other clean text-image pair as the training data.

After training, the diffusion model had been inserted with the text-image mapping between the text and visual element from the target image, also the so-called backdoor. At the inference stage, attackers only needed to prompt the poisoned model with the message including all of the text elements, to trigger the model to replicate the related visual elements in single image and combine them together in original structure, making it similar enough to the target image to be judged as manner of copyright violating.

2.2. Copyright issues about generative models

Take events as examples, on 27 September 2024, the district court of Hamburg, Germany announced its judge decision upon the case that LAION dataset adopted the pieces shot by photographer Kneschke without permission, ending up with rejecting the lawsuit application of Kneschke; In 2023 Getty Images officially claimed that its photos were used by the company Stability AI for training model, this case is still in controversial at present.

3. Methodology

The SBD attack was essentially taking advantage of the capability of diffusion model in memorizing training data and recreating it given enough training samples. Previous researches indicated that, comparing to increasing the size of clean data in the training set without any modifications on the poisoning samples, setting the subsampling ratio down to reduce the number of generated poisoning samples joining in the training process could be more challenging for SBD to attack the target model successfully. One approach for improving the performance when merely a small proportion of poisoning samples could be in usage is, to increase the number of text-image trigger elements in each sample, so that the influence of lower subsampling ratio would be weakened. We called this version of attacks MESI (multiple elements in single image). At the poisoning samples generation stage, the visual elements which do not overlap one another would be formed into various combinations, the element in each combination would be processed by inpainting model together.

However, the simple MESI still had flaws in the conflict between the quantities of trigger elements and the stealthiness of poisoning samples. More elements per sample would increase the similarity between the poisoning sample and the target image, even making the similarity, in some cases, exceed the threshold for being considered as copyright infringement.

To lower the similarities while raising up the capacity of trigger elements among the samples, we tried to sacrifice the visual fidelity of the trigger elements. We adopted DCT as the method to implement this aim. This is one of the image transformations widely applied among modern video coding standards [3]. For an image with N × N pixels, the 2-dimensional DCT was defined as:

$$C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cos\left[\frac{\pi(2x+1)u}{2N}\right] \cos\left[\frac{\pi(2y+1)v}{2N}\right], \quad (1)$$

where f(x, y) is the pixel (or signal) value on time domain location (x, y), the frequency domain location u, v = 0, 1, 2, ..., N - 1, and normalization factor $\alpha(u)$ (or $\alpha(v)$) is defined as:

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & \text{if } u = 0 \\ \sqrt{\frac{2}{N}} & \text{if } u \neq 0 \end{cases} \quad (2)$$

The frequency of image implies the intensity of gradation, the gradient of gray scale on plain space. DCT processes the information distribution with unified density on an image into unbalanced form, dividing the information carried by an image into 2 parts, the high and low frequency. In the attack manners we only reserve the high frequency part in order to reduce the similarity between the poisoned image and the original version while adding more visual trigger elements. After importing the DCT, we could even hide more trigger elements than simply using MESI. Fig. 1 explains the complete process of MESI + DCT attacking.

4. Experiments

4.1. Experiments setup

Datasets and Models: Aimed to utilize the original form of attack, and evaluate the modified attacking methods based on that, we adopted the dataset Midjourney Detailed Prompts, which was formed orginally to provide a high

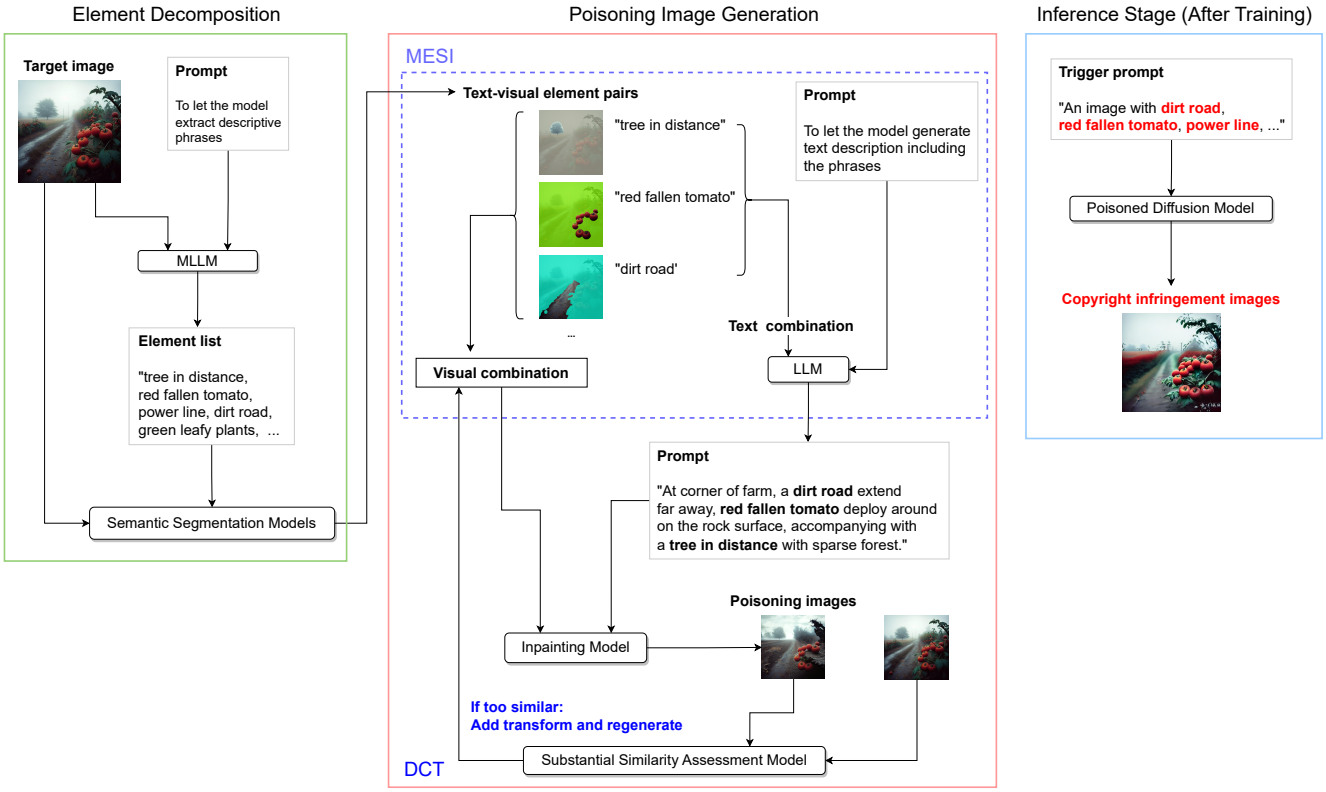

Figure 1. Attacking process graph based on SBD, with MESI and DCT modules.

Poisoning Ratio	Pokemon	Midjourney	Style	DiffusionDB
5%	9.14% / 61.36	17.14% / 49.62	23.15% / 38.75	9.98% / 41.00
10%	32.85% / 51.06	47.61% / 35.57	32.28% / 31.75	12.23% / 34.50
15%	37.28% / 44.53	55.24% / 32.08	35.13% / 27.50	39.63% / 30.75

Table 1. Average CIRs and FAEs across different poisoning ratios among the datasets (Format of value: CIR / FAE).

quality multi-level promptings for images selected from Midjourney v5 or v6. It contained detailed text description generated by Qwen-VL-Max, plus long and short prompts created by C4AI Command-R. Finally we used the short prompts to form our text-image pair data for the experiments, all of them were candidate target images for attacks. The clean data which would accompany with the generated poisoning samples based on Midjourney were selected from COYO-700m, in align with the early researches. In each experiment, the number of clean data was set as 500 constantly.

In addition, to deal with the potential issue of insufficient datasets using for SBD, we proposed 2 new image-prompt datasets suitable for this form of attacks. Style was a synthesized dataset with 60000 images, which extracted 10000 captions from MS COCO2017. Each caption was used as a prompt to generate 6 different style images with the diffusion transformer model FLUX.1-dev [1] and 6 additional trained LoRA weights respectively. For each of the artistic styles: aquarelle, frosting lane, half illustration, PS1, tarot and yarn, 10000 text-image pairs were created;

DiffusionDB was the other one dataset, containing 14 million images generated by SD with prompts. For both proposed datasets, a subset of 800 images was collected for implementing the SBD attack. In each experiment, one image was selected as the target of copyright infringement, and 600 data would be selected from the rest as the clean data.

At the poisoning stage, we use GroundingDINO and Segment Anything Model (SAM) to detect and segment visual trigger elements from the target image [5]. Stable Diffusion XL Inpainting was employed to generate the complete poisoning image based the visual elements, with the prompt containing the respective descriptive phrases. At the training stage, the target models for copyright infringement

attack currently are the SD series, from v1.1 to v1.5. There was another Multi-modal Large Language Model (MLLM) required for the attack process, which was responsible for observing and recording phrases, and production of various type of prompts. In experiments both GPT-4 and LLaVA could be feasible choices.

Evaluation Metrics: We measured the degree of similarity between images by Self Supervised Copy Detection (SSCD) [4], a SoTA indicator and set $SSCD > 0.5$ as the threshold condition for copyright infringement detection. To quantify the attacking performance of SBD, we employed First Attack Epoch (FAE) at the training stage and Copyright Infringement Rate (CIR) at the inference stage. In each of the 100 epochs in single experiment, multiple images would be generated by the prompts containing trigger words, and FAE was the epoch for the first time manage to create an image with $SSCD > 0.5$. After finishing training, 100 images were generated by the triggered model in each testing experiment, and CIR would be the percentage of images with $SSCD > 0.5$.

4.2. Effectiveness evaluation

We evaluated the effectiveness of SBD on the datasets Style and DiffusionDB at poisoning ratios (= $\frac{\#poisoning\ data}{\#poisoning\ data + \#clean\ data}$) 5%, 10% and 15%. The average CIR and average FAE on Style and DiffusionDB were calculated over T = 4 independent attacks. The results on Midjourney and Pokemon in previous research, as comparison cited here, were computed across T = 20 attacks. The general results are displayed in Tab. 1.

4.3. Attacking methodologies for higher effectiveness

We tested the performance of 3 attacking methods selections: SBD, MESI and DCT (with MESI), on Midjourney dataset at subsampling ratio = 5%. In the early SBD experiments for this scenario, the size of clean data of was 10000, and subsampling ratio varied among 100%, 50%, 30% and 5%.

For MESI and DCT, we set the number of clean data as 500 constantly. As shown in Tab. 2, comparing to original form, both simple MESI and DCT made it become possible to attack successfully in extremely low subsampling ratio, with few poisoning samples joining in the training stage.

Method	Avg. Poisoning Ratio	Avg. CIR / FAE
SBD	1.19%	0.00% / 100.00
MESI	1.19%	0.23% / 84.00
MESI+DCT	1.19%	**12.73% / 65.50**

Table 2. Average CIRs and FAEs for subsampling ratio = 5% (6 poisoning samples) on Midjourney with multiple attacking methods.

5. Conclusion and future work

In this research we estimated the performance of new datasets on copyright infringement attack, to compensate for the potential issue of lacking data resources in this area. We refined the current attacking methodology and earned better results in tough attacking conditions, indicating the vulnerability of diffusion model in text-to-image task.

The current image transformation (DCT) used at poisoning samples generation stage might be not the most ideal operation to increase the stealthiness. Looking for potentially more outstanding transformations could be one of the promising directions in future.

Besides, for MESI, it could be not so appropriate to plainly set up combinations simply based on all possible trials among trigger elements. Earlier research about adversarial attacks in object classification tasks found that, some combinations of objects on an image were more qualified in attacking successfully due to their relationships in co-occurrence, comparatively distance and sizes. Those relations could be expressed by setting up directed graphs. Similarly we might conceive there exists a standard about selecting target images which are more suitable for copyright infringement attack due to the features of trigger elements.

References

[1] black-forest labs. FLUX.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: 2025-03-20. 3

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1

[3] Syed Ali Khayam. The discrete cosine transform (dct): Theory and application. Technical report, Department of Electrical & Computer Engineering, Michigan State University, March 2003. 1, 2

[4] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. *Proc. CVPR*, 2022. 4

[5] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 3

[6] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 1

[7] Haonan Wang, Qianli Shen, Yao Tong, Yang Zhang, and Kenji Kawaguchi. The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breaches without adjusting finetuning pipeline. *arXiv preprint arXiv:2401.04136*, 2024. 1, 2