Raising the Cost of Targeted AI Image Generation

Anusha Ghosh, Vaibhav Gupta, Ahsan Gilani, Dante Howard
{anushag3, vaibhav7, ahsang2, danteah2}@illinois.edu

Abstract

The emergence of text-to-image models such as Stable Diffusion (SD) has enabled users to generate realistic and unique images given a simple prompt. These models can be further enhanced through few-shot fine-tuning, which allows models to learn subject representation while maintaining fidelity to the original images based on minimal data. This presents a security concern when an adversary performs fine tuning on a victim's images without consent. Given a few images of a subject, an attacker can use readily available diffusion models to generate malicious images of the subject in nearly any context. We propose a method to introduce adversarial perturbations to subject images that reduce the effectiveness of fine-tuning text-to-image models. In our paper we demonstrate how facial mimicry attacks can be used to generate malicious images using the Dreambooth diffusion model as well as how projected gradient descent (PGD) can be used to mitigate the performance of such attacks. Although it is difficult to quantify the effectiveness of our defense, we show that output generated after training on defended inputs have noticeably lower subject fidelity than the original inputs.

1 Introduction

The emergence of text-to-image models such as Stable Diffusion (SD) has enabled users to generate realistic and unique images given a simple prompt. These models can be fine-tuned on a specific set of images to produce a personalized model that generates images of a particular subject through a process known as targeted image generation. With the amount of information available through online content and social media, individuals are susceptible to malicious attacks that can be achieved with only a few images. For instance, images of an individual found online can be used to fine-tune a text-to-image model and generate compromising images.

We introduce the threat model of a facial mimicry attack where targeted image generation is used to produce malicious images of an individual. Facial mimicry attacks can cause harm to one's reputation by producing compromising images of a target, misusing their likeness for spreading misinformation, or influencing specific audiences. For instance, ads can be generated containing celebrities or popular public figures endorsing specific products without their consent or without proper compensation [1].

Text-to-image models generally function by associating extracted features of an image with those of a textual description in order to generate new images given an unseen description. However, this process can be disrupted through adversarial attacks, where small, imperceptible perturbations are added to an image to produce an inaccurate feature representation. If such a technique is applied to images that are fed into a fine-tuned model, the resulting generated images will be unrelated to the original subject (untargeted attack) or related to another subject (targeted attack).

Figure 1: Demonstrating a simple mimicry attack using DreamBooth with five training images from the Internet of popular YouTuber Mr. Beast. Although not inherently malicious, this method could be used to generate a subject in nearly any context with varying implications depending on the intention of the attacker.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image models are a class of models that generate images from textual descriptions. Traditional text-to-image models relied on generative adversarial networks (GANs) or variational autoencoders (VAEs), which struggled with generating high-quality, realistic images or only partially matched text descriptions. More recently, diffusion models have began to outperform traditional models through improved image quality and realism as well as demonstrating a broader range of image outputs.

2.2 Subject Driven Image Generation

DreamBooth is a fine-tuning method which uses few-shot learning, requiring only three to five images, which can be used to create a personalized text-to-image model. The model learns to associate the images with a unique identifier which can be used as input during evaluation time to generate targeted images [2]. Generated images in the personalized model are able to combine learned features associated with the unique identifier with prior class knowledge from the text-to-image model. DreamBooth achieves this using class-specific prior preservation loss, which allows it to produce new images which resemble the subject in unseen contexts while using characteristics from the original class [2].

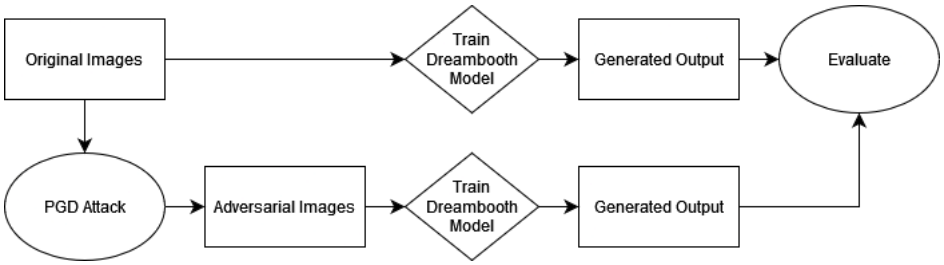
2.3 Adversarial Attacks

An inspiration for this paper is the success of Glaze which uses adversarial perturbations to produce style cloaks, which prevent a text-to-image model from replicating the style of artwork of a particular artist. Cloaking is done on the artwork from the artist so that the model maps the artist’s images to a feature representation of a different style. This is done by first transforming the style of a given image X into a desired target style T while preserving all other features, this new image is X_t. Then, a style cloak is computed which aims to make subtle perturbations to X such that the feature representation of X matches the feature representation of X_t [3].

Adversarial attacks have been also used to immunize images against photo editing with diffusion models. PhotoGuard uses projected gradient descent (PGD) as well as stronger attacks to immunize photos that an adversary attempts to use for targeted image editing [4]. The edited portion of the input is corrupted in the output while the unedited portion of the subject is preserved. PhotoGuard demonstrates the ability for PGD to be used as a means for defending against malicious attacks with diffusion models. Unlike PhotoGuard, our work aims to prevent the likeness of a subject from being included in the generated output at all rather than preserving a portion of the original image.

The goal of our paper is to build off of the work done by Glaze and Photoguard by providing protection against a state-of-the-art fine-tuning method known as DreamBooth and deepfake images of people, rather than artwork, generated through Stable Diffusion.

3 Methodology



3.1 Dataset

The mimicry attack was trained on a sample ($N = 10$) of the CelebA-HQ dataset which contains 30,000 1024x1024 high resolution profile images of celebrities [5]. [1] Images were sorted by identity and only subjects with five or more images were considered with respect to various genders and races. Each subject was trained with the same single malicious context, which includes five training images of generic individuals holding or posing with a gun taken from Google images.

3.2 Facial Mimicry Attack

The facial mimicry attack was initially constructed according to **Figure 1**, where Dreambooth was only fine-tuned to a single subject. While this generates high fidelity images of a subject within simple, generic contexts that are known to the pretrained Dreambooth model, it performs poorly for specific contexts or novel concepts which may be common for targeted attacks. A concept is a specific instance of a class and Dreambooth supports multi-concept training, where several concepts can be trained simultaneously within a single model and are distinguished by a unique identifier. We use this to train a subject with a malicious context in order to increase context fidelity while maintaining subject fidelity.

Although we attempted to use multi-concept training to train multiple subjects within a single model, image quality degrades dramatically as the model is unable to generate distinguishable faces. Therefore, we train a separate model for each distinct subject with the same arbitrary identifier to prevent the model from associating the subject with any pretrained concept.

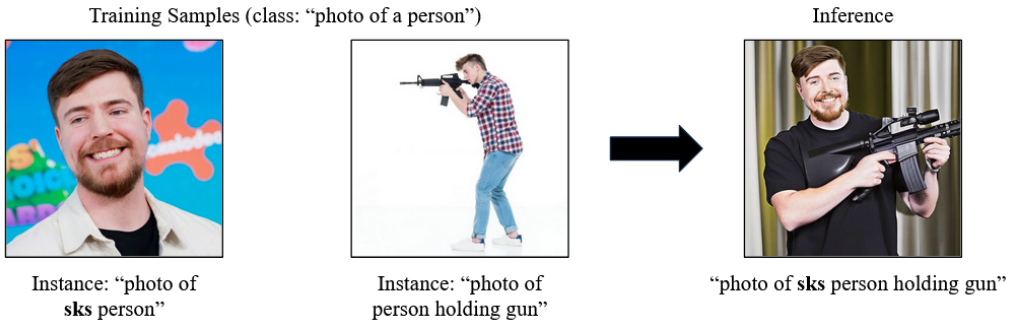


Figure 2: Complete Dreambooth mimicry attack where a subject is trained alongside a specific context (action, object, etc.) as separate concepts in order to generate both the subject and context with high fidelity.

3.3 Projected Gradient Descent Attack on Images

We generated adversarial input by using projected gradient descent on the model's encoder. By attacking the variational autoencoder of the Stable Diffusion model, we were able to use untargeted attacks in order to create generalized adversarial input.

[1] CelebA-HQ dataset

4 Results

Figure 3: Example of generated output for facial mimicry attack before and after applying PGD. The generated images after applying PGD have similar context fidelity but noticeably lower subject fidelity.

4.1 Qualitative Results

Our main objective is to compare the baseline and attack results. In a practical setting, we care about whether the subject is recognizable in the generated images and if the generated output depicts the subject performing the action given in the prompt. For example, we generate images of subjects with an input prompt of "an image of subject X holding a gun". We notice that the attacked models are unable to generate realistic images, as they inaccurately capture the facial features of the input subject and cannot mimic the action with high fidelity (see **Figure 3**). Whereas the baseline model does a much better job of maintaining subject fidelity and of generating images of the subject which resemble the input prompt.

4.2 Quantitative Results

We evaluate our outputs on metrics which compare pixel-wise image similarity and quality, as well as metrics which aim to capture the quality and similarity between image features. Similarity is measured between the original input images and the generated images (by the baseline and attacked model).

Image Similarity Metrics

We consider various image similarity metrics such as SSIM and PSNR and use them to evaluate our generated images. No significant differences were found between the average scores for the images that were generated by the two models (see **Table 1**). This is due in large part to the fact that the baseline generated images already differ greatly from the input images used to train the model. The input images from the dataset contain close-up images of a subject's face while the generated images contain a certain action being performed by the person. Due to this, we consider metrics which can measure if an image resembles a particular action, and whether there is similarity between the features of the two images.

Method	SSIM	FSIM	PSNR	VIFp	FID
Baseline (original DreamBooth)	0.1973	0.5616	8.0572	0.0195	9.4959
DreamBooth trained on PGD images	0.2043	0.5607	7.9658	0.0199	8.4711

Table 1: Comparison of Baseline and PGD Trained DreamBooth using image similarity metrics. [2]

Feature Similarity Metrics

In order to capture feature representations of the images, we construct CLIP embeddings [6] of our images and take the cosine similarity between embeddings which gives us the image similarity scores. We run into similar results as the image similarity scores for the same reasons listed above, however we can see that the average scores (as seen in **Figure 4**) indicate a larger resemblance between the features of the input image and the generated images. We also use CLIP textual embeddings to

[2]We use the implementations provided in: https://github.com/photosynthesis-team/piq.

measure whether or not what was generated by the DreamBooth models corresponds to the prompt. The scores show that both the baseline and attacked models were able to generate images which resemble the action in the input prompt. However, further analysis showed that the resemblance and fidelity of the subject was much lower in the images generated by the attacked model.

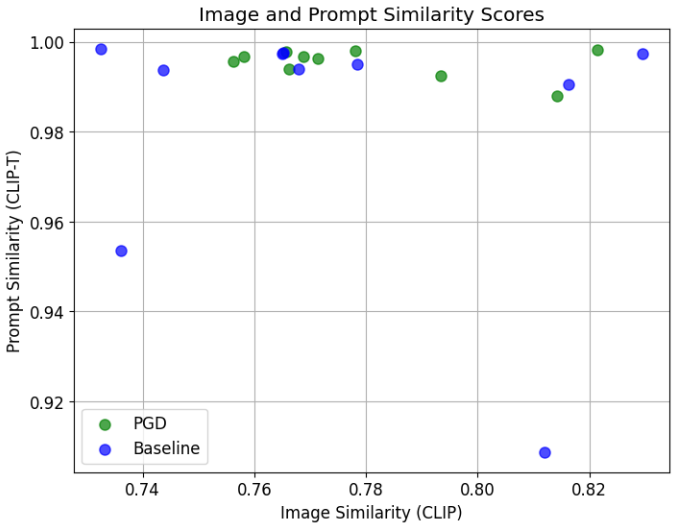


Figure 4: Attacked images with PGD received slightly worse CLIP similarity scores [3]

5 Discussion

Evaluation Limitations

Given our threat model and the application of such an attack, a large emphasis is placed on the qualitative evaluation of the generated images. Therefore, we propose a potential user study to evaluate and distinguish between the baseline and PGD generated images given an input subject and action. Ideally, we would like to be able to measure the fidelity of the output image and how it compares to the input images of the subject which is where using different embeddings such as DINO may be able to alleviate the shortcomings of CLIP. Another essential property of the generated images is whether the input subject and/or context is recognizable, which may involve a facial recognition model to evaluate subject fidelity and a LLM to describe and evaluate context fidelity.

Future Work

Potential future work includes evaluating stronger attacks, such as the diffusion attacks implemented in PhotoGuard, against a larger number of subjects with additional training data. Due to computational resource limitations, were were only able to evaluate on 10 individuals and strictly used the CelebA dataset, which is limited to close-up images of faces. Additionally, we could also explore additional malicious contexts and prompts in order to increase the robustness of possible defenses.

6 Conclusion

We found that Stable Diffusion models trained via Dreambooth were uniquely vulnerable to adversarial attacks using projected gradient descent. Note that context fidelity remains similar since we do not attack the context training images (holding gun) since these are independent of the subject and introduced by the attacker. While quantitative metrics were unable to adequately distinguish between the outputs of the original and attacked models, we found that the images generated after training on adversarial inputs have noticeably lower subject fidelity.

[3] We use the the implementations provided in CLIP library maintained by OpenAI

References

[1] Singleton T. et al. Tom Hanks Warns Dental Plan Ad Image is AI Fake. https://www.bbc.com/news/technology-66983194.

[2] Ruiz N. et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. https://arxiv.org/abs/2208.12242.

[3] Shan S. et al. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. https://arxiv.org/abs/2302.04222.

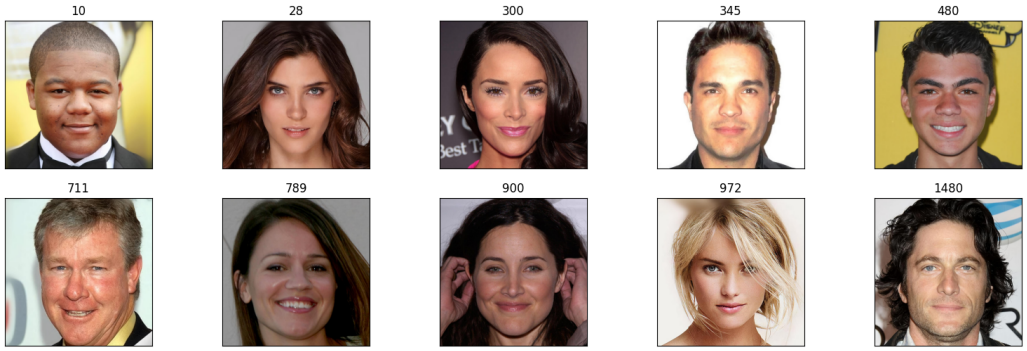
[4] Salman H. et al. Raising the Cost of Malicious AI-Powered Image Editing. https://arxiv.org/abs/2302.06588.

[5] Karras T. et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation. https://arxiv.org/abs/1710.10196. 2018.

[6] Radform A. et al. Learning Transferable Visual Models From Natural Language Supervision. https://arxiv.org/abs/2103.00020. 2021.

7 Appendix

The following sample of subjects and corresponding identities from the dataset were used:



The following model parameters were used to train Dreambooth [4]

```
--pretrained_model_name_or_path="SG161222/Realistic_Vision_V1.4"
--pretrained_vae_name_or_path="stabilityai/sd-vae-ft-mse"
--with_prior_preservation
--prior_loss_weight=1.0
--resolution=512
--train_batch_size=1
--train_text_encoder
--mixed_precision="fp16"
--use_8bit_adam
--gradient_accumulation_steps=1
--learning_rate=1e-6
--lr_scheduler="constant"
--lr_warmup_steps=0
--num_class_images=100
--sample_batch_size=4
--max_train_steps=1200
```

The following parameters were used to generate output during inference:

```
prompt = "photo of sks person holding a gun studio lighting, masterpiece,
4k, ultra detailed, sharp focus, 8k, high definition, insanely detailed,
intricate:1. 1)"
negative_prompt = "text, b&w, illustration, painting, cartoon, 3d,
bad art, poorly drawn, close up, blurry, missing fingers, extra fingers,
ugly fingers, long fingers"
guidance_scale = 7.5
num_inference_steps = 100
height = 512
width = 512
```

[4] Based off the Dreambooth training script provided by Shivam Shrirao