Creating a Watermark Through an Adversarial Attack

Andy Truong
amt007@ucsd.edu

Anushka Purohit
apurohit@ucsd.edu

Wang, Yu-Xiang
yuw272@ucsd.edu

Abstract

This project offers an innovative technique in the creation of an invisible watermark, leveraging the adversarial attack Projected Gradient Descent (PGD) that adds small perturbations to an image to misclassify it. Specifically, our approach is designed in rivalry of the regeneration attack that we explored in our Project 1. This approach allows for a dual functionality: targeted misclassification for watermark verification and preservation of the original classification for content authenticity. We purposely select k random labels to perturbed noise towards on the original image while still maintaining the dominant classification of said image. We verify if the image is watermarked to check if the logit distribution changes significantly based on the threshold that we set on the target labels. With this approach, we are able to offer a new solution to the battle of watermark removal in order to maintain and protect the authenticity of a digital image to avoid misrepresentation and the spread of misinformation.

Website: https://ndxdxd.github.io/AdversarialAttacksOnWatermarks/
Code: https://github.com/ndxdxd/Capstone-Q2/tree/main

1 Introduction

In an age where digital media runs rampant, there is a proliferating need for a method to protect the rights and authenticity of said media. Especially with the rise of Generative AI, users can produce content nearly identical to real life that could potentially fool an uneducated consumer. Thus, watermarking, specifically invisible watermarking, still remains as one of the most viable methods to achieve this goal. They maintain the quality of the image and are least likely to be removed by the general public. Nevertheless, people who are aware of these watermarks can execute watermark attacks, getting rid of the invisible watermarks from the AI-generated images which can lead to the problems such as misrepresentation and misinformation. There are attacks such as destructive attacks which significantly reduces the quality of the image but are effective at removing the invisible watermark. There are also regeneration attacks, where a watermarked image is destsroyed that reconstructed using a Generative AI model, which is what we explored in our Project 1. As a result, there needs to be another innovative watermarking technique that should be robust to these attacks.

In this project, we are focusing on creating an innovative watermark with the goal of it being robust to regeneration attacks. The watermark that we have created is guided by an adversarial attack. An adversarial attack is an attack which exploit vulnerabilities in machine learning systems by introducing small perturbations that lead to large errors that causes misclassification. The adversarial attack which we are using is known as the Projected Gradient Descent attack (PGD). PGD works by repeatedly applying small perturbations to an image in the direction that maximally confuses a given model, while ensuring the modifications remain within a defined constraint to maintain image integrity.

There has been previous work that has attempted to explore more about adversarial attacks such as the "Adversarial Machine Learning at Scale" paper by Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio, where they discussed the different types of adversarial attacks such Fast Gradient Sign Method (FGSM) as well as the results of them. They found out that training on adversarial data showed robustness to adversarial examples and that the increase of parameters in a model could also increase the robustness of adversarial examples as well. However, there has not been any work where instead of making a watermark robust to adversarial attacks, making a watermark based off of adversarial attacks.

By incorporating adversarial perturbations into our watermarking technique, we aim to create a watermark that remains resilient even after an image undergoes regeneration attacks. Since generative models often rely on learned patterns and statistical correlations to reconstruct images, our adversarially crafted watermark disrupts these patterns, making it significantly harder for the model to remove or ignore the watermark. This ensures that ownership and authenticity can still be verified even after AI-based modifications.

For our data, we will be using Imagenet for the ResNet50 model to use to build our watermarking technique, as well as stock photos of animals from Google to add watermarks to them. We will be also using the attacks we created in our Project 1 to test out our watermark.

2 Methods

We began by exploring the PGD (Projected Gradient Descent) adversarial attack to understand the algorithm and how to begin developing our watermark. We understood that adversarial attacks are able to fool neural network models so we wanted to create a watermark inspired by an adversarial attack. We believe that this watermark would be robust to watermark removal attempts from neural network models such as the regenerative stable diffusion removal attempt.

2.1 Targeted PGD attack

The Projected Gradient Descent (PGD) attack is an iterative adversarial attack that generates adversarial examples by perturbing an input sample in the direction that maximizes the loss function. In the targeted variant of PGD, the attacker aims to misclassify the input as a specific target class rather than simply causing misclassification.

Given a classifier $f(\cdot)$ with parameters θ, a clean input sample $x \in \mathbb{R}^d$ with true label y, and a target class $y_{\text{target}} \neq y$, the goal of the targeted PGD attack is to find an adversarial example x^* such that:

$$f(x^*) = y_{\text{target}}, \tag{1}$$

while ensuring that the perturbation remains bounded within an ℓ_p norm constraint.

The attack is performed iteratively by updating x using the following rule:

$$x_{t+1} = \Pi_{B_\epsilon(x)}\Big(x_t - \alpha \cdot \text{sign}(\nabla_x \mathscr{L}(f(x_t, \theta), y_{\text{target}}))\Big), \tag{2}$$

where:

- $\mathscr{L}(f(x, \theta), y_{\text{target}})$ is the loss function (e.g., cross-entropy loss) with respect to the target class y_{target},
- α is the step size,
- $\Pi_{B_\epsilon(x)}(\cdot)$ denotes the projection operator that ensures the perturbed sample remains within an ϵ-ball around the original input x under an ℓ_p norm,
- $\nabla_x \mathscr{L}$ is the gradient of the loss function with respect to the input x.

The targeted PGD attack follows these steps:

1. Initialize $x_0 = x$.
2. For $t = 0, 1, \ldots, T-1$:
   - Compute the gradient: $g_t = \nabla_x \mathscr{L}(f(x_t, \theta), y_{\text{target}})$.
   - Update the adversarial example: $x_{t+1} = \Pi_{B_\epsilon(x)}(x_t - \alpha \cdot \text{sign}(g_t))$.
3. Return x_T as the final adversarial example.

2.2 White Box Approach Watermark

For our new watermark, we use the PGD attack as a base. However, we pick a random subset of labels of k labels and aim to make the probability of those labels go higher. The watermark is designed to subtly alter the image such that the model's output probabilities for specific target classes are significantly affected, while the overall visual appearance of the image remains largely unchanged.

The watermark is applied as an adversarial perturbation δ to the original image x. The perturbation is generated using a targeted adversarial attack, which aims to maximize the model's prediction error for a set of target classes while minimizing the visual distortion. The perturbation is constrained by an ℓ_∞-norm bound to ensure that it remains imperceptible:

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max} \mathscr{L}(f(x+\delta), y_{\text{target}}),$$

where:

- f is the pre-trained ResNet-50 model,
- $\mathscr{L}$ is the loss function (e.g., cross-entropy loss),
- y_{target} represents the list of target class labels,
- ϵ is the maximum allowed perturbation magnitude.

To analyze the impact of the watermark, we compare the model's output logits for the original image x and the watermarked image $x+\delta$. Specifically, we focus on the logits of the target classes, as these are the classes most affected by the watermark.

Let l_{original} and $p_{\text{watermarked}}$ denote the probability distributions output by the model for the original and watermarked images, respectively. The difference in logits for the target classes is computed as:

$$\Delta p = |p_{\text{original}} - p_{\text{watermarked}}|.$$

A significant difference in logits for the target classes indicates that the watermark has successfully altered the model's predictions. We define a threshold τ to identify significant changes:

$$\text{Significant change} = \Delta p > \tau.$$

2.3 Black-Box Watermarking Approach

In our black-box watermarking approach, we introduce a perturbation that alters the model's output probabilities for a randomly selected subset of k labels without direct access to the model's gradients. Unlike the white-box approach, where the adversarial perturbation is optimized using gradients, the black-box method relies on an iterative query-based optimization process.

The watermarking process involves modifying the input image x to maximize the probability of specific target labels. Given a pre-trained model f, we define the perturbation δ as follows:

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max}\, \mathscr{L}(f(x+\delta), y_{\text{target}})$$

where:

- f is the pre-trained ResNet-50 model,
- $\mathscr{L}$ is a surrogate loss function based on model output probabilities rather than gradients,
- y_{target} represents one of the selected target class labels,
- ϵ is the maximum allowed perturbation magnitude.

Since we do not have access to the model's internal gradients, we iteratively perturb the image and measure the change in logits. Specifically, we:

1. Select k target labels at random.
2. Query the model with the original image x to obtain the baseline logits.
3. Apply a small perturbation δ and measure the resulting logits.
4. Update δ using an optimizer (e.g., Adam) to maximize the probability of the target labels.
5. Scale the perturbation by $\sqrt{k}$ to ensure the adjustment accounts for multiple target labels:

$$\delta_{\text{scaled}} = \delta \cdot \sqrt{k}$$

6. Repeat the process for multiple iterations.

To evaluate the effectiveness of the watermark, we compare the logits of the target classes before and after applying the perturbation. Let l_{original} and $l_{\text{watermarked}}$ denote the logits output by the model for the original and watermarked images, respectively. The difference in logits for the target classes is computed as:

$$\Delta l = l_{\text{watermarked}} - l_{\text{original}}.$$

We define a threshold τ to determine whether the perturbation significantly alters the model's predictions:

$$\text{Significant change} = \Delta l > \tau.$$

A successful watermark is one where the logits of the target labels increase beyond the threshold τ while maintaining imperceptibility to human observers. By comparing the logits of multiple images before and after watermarking, we can assess the robustness and effectiveness of this black-box watermarking method.

3 Results

We have no results yet

4 Discussion

What did it mean?

5 Conclusion

To edit the contents of the "References" section, edit `reference.bib`. Many conference websites format citations in BibTeX that you can copy into `reference.bib` directly; you can also search for the paper on Google Scholar, click "Cite", and then click "BibTeX" (here's an example).

References

Appendices

A.1 A broad problem statement

In a world filled with online content, be it AI generated or real media, ensuring the authenticity of digital creations has become an pivotal challenge. Specifically, watermarking has served as a reliable method for asserting ownership of content produced by generative AI. However, with the rise and innovation of adversarial attacks, which are subtle manipulations of media designed to deceive machine learning systems, poses a significant threat to watermarking techniques. These attacks can render watermarks undetectable, weakening the security of watermarked images.

The impact of this problem goes further than ownership issues. Robust watermarking is essential for keeping online content safe in industries such as entertainment, education, and government that could stop the spread of misinformation and misrepresentation. If this project is successful, we would contribute to developing a more resilient watermarking system, ensuring that digital media remains secure in the face of increasingly complicated threats. This could pave the way for a more digitally secured world where we can protect rights of creators.

A.2 A narrow, careful problem statement for the domain expert.

Adversarial attacks exploit vulnerabilities in machine learning systems by introducing small perturbations that lead to large errors in the feature space representation of data. This problem relates to our Quarter 1 Project as it focuses on a specific issue that isn't worked on in that project. There has been previous work that has attempted to explore more about adversarial attacks such as the "Adversarial Machine Learning at Scale" paper by Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio, where they discussed the different types of adversarial attacks Fast Gradient Sign Method (FGSM) as well as the results of them. They found out that training on adversarial data showed robustness to adversarial examples and that the increase of parameters in a model could also increase the robustness of adversarial examples as well. Using their findings from the paper, we will apply an adversarial attack on an image, and then we will find where in the feature and pixel space that has the largest perturbation. We can apply an adversarial attack such as PGD (Projected Gradient Descent),

which is a method that is an iterative extension of the Fast Gradient Sign Method (FGSM), where perturbations are applied iteratively using gradient information. From there, we will move in the direction of the perturbation to "cover" the adversarial attack, designing a watermark that will account of the noise. Ultimately, this approach addresses a deficiency in our Quarter 1 Project, where adversarial robustness of watermarked images was not considered.

A.3 A statement of the primary output

The primary output of this project will be a research paper detailing the development, methodology, and evaluation of the proposed robust watermarking technique. This paper will include: a detailed description of the adversarial attack , experimental results demonstrating the watermark's resistance to adversarial attacks, and a comprehensive analysis of the limitations and potential future directions for the proposed technique.

Our data will be images used from our Quarter 1 Project, as we are simply reusing the images that we collected. We will analyze the watermark image by attacking the image and seeing if the attack would remove the watermark. If the watermark is removed, that means that the watermark was detected and that our watermark is robust to adversarial attacks.

B Contributions

In the beginning, we both split up researching and understanding PGD and adversarial attacks. For implementing the code, we met up and worked on it together.