Incorporating PGD to StyleGAN: Fooling Humans Without Fooling Computers

Leah Howells  
University of Central Florida  
leahhowells@knights.ucf.edu

James Beetham  
University of Central Florida  
james.beetham@knights.ucf.edu

Dr. Mubarak Shah  
University of Central Florida  
shah@crcv.ucf.edu

Abstract

In this paper, we propose a new unconstrained adversarial attack which is optimized to look realistic while still being misclassified. Existing attacks focus on finding small perturbations because it’s easy to ensure that human classification remains unchanged for very small perturbations. However, these are limited by the starting images, and stronger attacks generally have visible perturbations. We propose a method which perturbs an image within a GAN’s style-space instead of perturbing an image in the visual space. We can alter existing attack methods to target style-space by adding an additional loss component to push a generated image to be an adversarial image, while maintaining the class of the original generated image. This method allows for adversarial attack to be performed without the grainy perturbations of traditional attacks.

1. Introduction

Adversarial attacks have been a large focus in the field of computer vision for many years since they were first discovered in the paper “Intriguing properties of neural networks” [4]. They have been performed in a similar way for many years, by adding subtle perturbations to an image so that a human cannot recognize the difference, but a model will misclassify the image. An issue with this approach though is that these attacks are limited by the starting image and stronger attacks generally have visual perturbations. The original goal for our research was to change an image’s contents without changing model classification, so changing how a human classifies an image without changing the original model classification. However, after some initial tests the goal was altered to be to create an adversarial attack within the distribution of clean data; without perturbation artifacts. This goal was born from our initial results showing that we could possibly create an adversarial attack utilizing a GAN, one that would not leave visible perturbations behind even with stronger attacks. In this paper we show our method, the results of that method, and how it completes the goal of creating an adversarial attack within the distribution of clean data; without perturbation artifacts.

2. Background and Related Work

2.1. Adversarial Attacks

An adversarial attack is when an image is subtly changed so that the change is imperceptible to the human eye but will cause a model to misclassify the image. The images generated are also known as adversarial examples and were first introduced in the paper “Intriguing properties of neural networks”. These attacks can be untargeted, in which the attack causes any misclassification, or the attacks can be targeted in which the attack causes a specified misclassification. Since their introduction, many papers have been created on various topics regarding adversarial attacks. For example ”Towards Deep Learning Models Resistant to Adversarial Attacks” [3] introduces the Projected Gradient Descent or PGD attack which is a type of adversarial attack that was created out of an attempt to formalize the framework of creating adversarial examples. The authors found that attacks dealing with gradient information were all practically equivalent to PGD, creating a somewhat universal attack. PGD works by finding a perturbation that maximizes model loss without the perturbation becoming greater than that of a specified epsilon. (Can add in more adversarial attack papers here if we want)

2.2. GANs

Generative Adversarial Networks or GANs are networks with the goal of producing artificial images, videos, etc that are indistinguishable from actual images, videos, etc. They were first introduced by Ian Goodfellow in the paper “Generative Adversarial Nets”. In the paper it describes a frame-
The work in which two models are simultaneously trained, a discriminative model and a generative model (named D and G respectively). The discriminative model takes samples of real images and generated images and tries to classify them as such, while the generative model is trained to fool the discriminative model as best as it can. Many GANs have been very successful in creating very lifelike images but struggle to control their output, for example changing specific features of generated images such as hair style and pose. To combat this issue, a research team with NVIDIA created StyleGAN [1] and an improved version StyleGAN2 [2] which gradually generates artificial images, starting with low level or coarse features and moving up to high level or fine details. It generates images in such a way that the levels do not affect one another.

### 2.3. Perturbation GANs

A perturbation GAN is a GAN that generates a perturbation to be added to a clean image to create an adversarial example. "Generating Adversarial Examples with Adversarial Networks" [5] was one of the first papers to introduce a perturbation GAN. The paper proposes AdvGAN, in which the generator takes an original image and generates a perturbation. The adversarial example created by adding that perturbation to the original image is then sent to the discriminator, the goal of which is to encourage that the generated instance is indistinguishable from the data from its original class. Similarly to this paper, our method has different perturbations for each individual image, but it is different in that it does not generate a perturbation separate from the image to be added at the end of generation. Our method utilizes style-space information to add the perturbation to the image as it generates instead of generating a separate perturbation to add to the clean image.

### 3. Proposed Method

Our goal is to leverage the style-space information to guide a perturbation to be more realistic. In pursuit of this, we first tried an iterative approach swapping between feeding an image through a StyleGAN2 [2] and perturbing the image with PGD [3]. This method led to interesting findings, but did not lead to good results, so we instead walked within StyleGAN2 style-space towards a more adversarial direction.

A key part of this attack is the idea that there is a path within style-space which smoothly transforms an image from one class to another. This on its own can be a form of boundary attack: find the boundary between the classes within the smooth transformation between classes. However, our method adds an additional step where we search the surrounding space for a more adversarial example that’s still relatively close to the smooth transformation path.
formed on the resulting image, and finally the adversarial example generated is projected back into the initial images style space utilizing the projection operation. This process is then looped a designated amount of times, hence the name iterative approach.

3.2. Style-space Attack

The style-space attack architecture is shown in Figure 2. For the style-space attack, we added an additional loss to the StyleGAN2 projection operation. The projection operation takes in random noise and maps it to a style w using the mapping function. It then puts that style through the synthesis function and results in the initial generated image. A VGG16 classifier extracts the image features from both the generated image and the target image, then calculates the L2 loss between them and projected back into the style space. The additional loss component is found using the cross entropy loss calculated from the class of the original generated image and the class of the current generated image. This additional loss is then added to the L2 loss and the sum is then projected back into the style space as in the original projection operation. The goal of the added loss is to push a generated image to be an adversarial image, while maintaining the class of the original generated image.

An example of this method is shown in Figure 3.

The additional loss can be weighted by adjusting what we call the "class preservation strength" or cp.

\[
Loss = (1 - \lambda_{cp}) \cdot L_2 + \lambda_{cp} \cdot CE
\]  

Equation 1 demonstrates how the cp determines the weights of the L2 and CE loss. Since a higher cp value results in an increased amount of the CE loss to be added, it also results in the classification percentage for the original class to be increased in the final generated image. The cp determines the strength of the attack.

4. Results

Using the CIFAR-10 validation dataset and pretrained StyleGAN2 models, we were able to successfully generate a style-space attack.

4.1. Projecting attacks into GAN-space

While the results from the iterative method were not particularly useful for our goal of creating an adversarial attack without perturbation artifacts, it did result in some interesting findings nonetheless. As shown in Figure 4, for most tests the images eventually devolved to have an over saturated background and blob-like subject. This could be because the perturbations added by PGD attacks generally look similarly saturated but are usually bounded by the epsilon, but since the attack is being performed 100 times over the course of the method, the attacks eventually add so much perturbation that it becomes over saturated. Also, a similar over saturated color occurred for the same target images regardless of the type of attack performed or the projector starting class. The results also showed that the PGD attack does not transfer well through the projection operation, since the final projections were still correctly identified as the target images class.

4.2. Attacking Style-space

Figure 5 shows a few visual results of the style-space method. The attack successfully generates images that look like the target image class but are classified as the original generated images class with about 30-40 percent confidence
for each. The generated images tend to look most realistic when transforming from a similar class e.g., cat to dog, car to truck, etc.

Figure 6 shows the style-space attacks percent of success on each class when transforming a generated cat image to each class on the x axis. The blue bar shows the results of performing the attack without the added CE loss. As to be expected, the cat class had the highest percentage by far since the resulting images had no adversarial attack performed on them and therefore were mostly classified correctly. The grey bar shows the results of performing the style-space attack. The attack was performed using 60 images from each class, resulting in a total of 600 generated images. The class most similar to cat (dog) had the highest probability, while Plane had the lowest probability. This sort of guided adversarial attack is an effective attack because it doesn’t add a separate perturbation and therefore has no visual perturbation artifacts as desired.

5. Conclusion

In this paper, we proposed an adversarial attack which stays within the distribution of clean data; without perturbation artifacts. The proposed method utilizes StyleGAN2’s projection operation and an added cross entropy loss to push a generated image to be an adversarial image, while maintaining the class of the original generated image. The images were generated using StyleGAN2 trained on the CIFAR-10 dataset. The method was successful in generating adversarial images that looked like the target class but were still labelled by the classifier as the original generated image class. Future work could be done in extending this method to other GANs and internal model-spaces, in particular with respect to efficiently generating these examples and getting adversarial accuracies on robust models.

References

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten,

