[Re] Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability

William Kang,
syu@student.ubc.ca

Christina Yang
chryang@student.ubc.ca

Abstract

We are doing a reproducibility report based on the paper "Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability" by Haotian Xue, Alexandre Araujo, Bin Hu, Yongxin Chen. Their Github Repo is here: `https://github.com/xavihart/Diff-PGD/tree/main`.

The paper uses a novel framework to generate adversarial samples. They use a gradient based method guided by a pre-trained diffusion model to try to generate images that appear realistic to the human eye, can fool a wide range of models, and is easy to control how certain regions are modified.

In particular, we are evaluating the claims that Diff-PGD outperform baseline methods such as PGD, AdvPatch, and AdvCam in physical-world attacks and style-based attacks specifically. Finally, we are evaluating the claim that Diff-PGD generates adversarial samples with higher stealthiness.

The following section formatting is optional, you can also define sections as you deem fit. Focus on what future researchers or practitioners would find useful for reproducing or building upon the paper you choose.

1 Introduction

A few sentences placing the work in high-level context. Limit it to a few paragraphs at most; your report is on reproducing a piece of work, you don't have to motivate that work.

2 Scope of reproducibility

The main claims from the original paper are as follows:

1. Diff-PGD can be applied to specific tasks such as digital attacks, physical-world attacks, and style-based attacks, outperforming baseline methods such as PGD, AdvPatch, and AdvCam.
2. Diff-PGD is more stable and controllable compared to existing methods for generating natural-style adversarial samples.
3. Diff-PGD surpasses the original PGD in Transferability and Purification power
4. Diff-PGD generates adversarial samples with higher stealthiness

We will be exploring the first claim in regards to physical-world 4 and style-based attacks 4 specifically.

Course project for UBC CPSC 440/550, April 2024.

3 Methodology

We used the authors' code, however, since not all parts of the authors' code have been released, we wrote our own code to evaluate a lot of the experiments.

3.1 Model descriptions

We used ResNet-50, ResNet-101, ResNet-18, Wide-ResNet-50, and Wide-ResNet-101 to classify images. The parameter 'weights' was set to the default weight for each model.

3.2 Datasets

The original paper used ImageNet, but due to limitations in compute and memory, we used a smaller subset of ImageNet with 1000 samples: `https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000`

3.3 Hyperparameters

Describe how the hyperparameter values were set. If there was a hyperparameter search done, be sure to include the range of hyperparameters searched over, the method used to search (e.g. manual search, random search, Bayesian optimization, etc.), and the best hyperparameters found. Include the number of total experiments (e.g. hyperparameter trials). You can also include all results from that search (not just the best-found results).

3.4 Experimental setup and code

We ran the code given by the authors. We copied the hyperparameter setup of the authors.

3.4.1 Success Attack Rate

We wrote our own code to test success rate in `attack_global.py`. We used the same hyperparameters for running Attack_Global as the authors, except that we chose to use skip=20, to use 78 images per iteration to test success rate. If the adversarial sample caused the model to classify the image as anything other than it's original classification, we considered it a successful attack.

3.4.2 Physical-World Attacks

We tried our own physical world attack using an image patch of a digitally drawn panda head and a laptop as our target object.

We use a Galaxy A53 5G to take images from the real world and used a MP C4505ex Color Laser Multifunction Printer to print the images in color.

Due to time and compute constraints, we chose to run 1500 iterations for each method instead of 4000 like the original model. We chose 1500 because after running one of the original paper's physical world attacks, the loss converged after around 1000 iterations.

We took the adversarial samples generated using each method (AdvCam, AdvPatch, and Diff-PDG) and stuck them on a laptop. We then took multiple pictures of the laptop with the sample on top, with the sample in various different locations and rotations, and tested for success rate on the these photos by classifying them with ResNet-50.

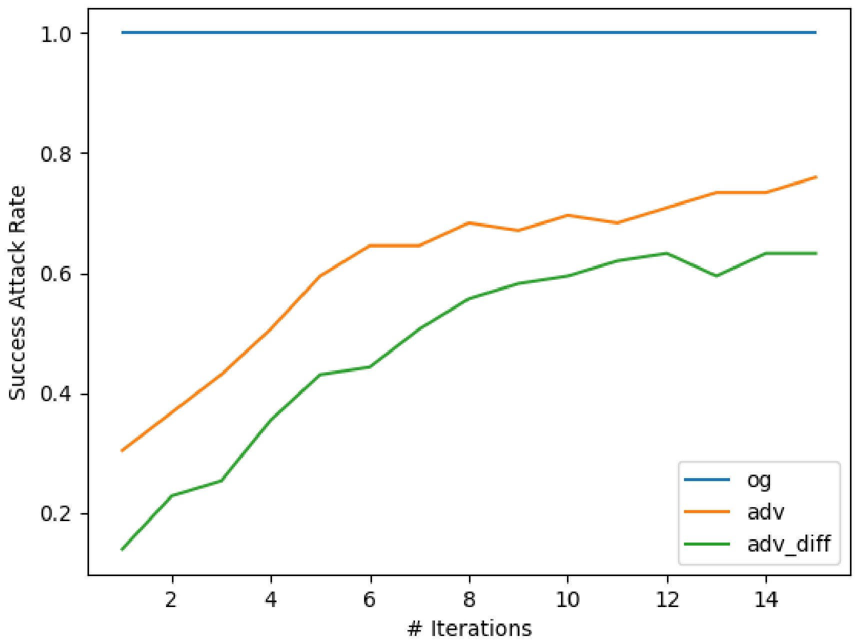

Figure 1: Blue line is success rate of classifying an image not attacked. Orange line is success rate of causing a different classification with Diff-PGD. Green line is success rate of causing a different classification with Diff-PGD after applying sdedit at the end (purification???).

3.4.3 Style-Based Attacks

3.5 Computational requirements

Include a description of the hardware used, such as the GPU or CPU the experiments were run on. For each model, include a measure of the average runtime (e.g. average time to predict labels for a given validation set with a particular batch size). For each experiment, include the total computational requirements (e.g. the total GPU hours spent). (Note: you'll likely have to record this as you run your experiments, so it's better to think about it ahead of time). Generally, consider the perspective of a reader who wants to use the approach described in the paper — list what they would find useful.

4 Results

Start with a high-level overview of your results. Do your results support the main claims of the original paper? Keep this section as factual and precise as possible, reserve your judgement and discussion points for the next "Discussion" section.

4.1 Results reproducing original paper

For each experiment, say 1) which claim in Section 2 it supports, and 2) if it successfully reproduced the associated experiment in the original paper. For example, an experiment training and evaluating a model on a dataset may support a claim that that model outperforms some baseline. Logically group related results into sections.

skip=20

4.1.1 Attack success rate

We got a very different result for success rate from that of the original paper (fig. 1). In the original paper, the success rate of all of the attacks reached 100% after 5 iterations. However, in our case the success rate reached only 60%-80% after 15 iterations.
Since the authors have yet to release their code for success rate, we are unsure what caused this difference. We did use a different dataset than the authors, this could be a possible reason.

4.1.2 Physical-World Attacks

None of the adversarial samples generated by AdvCam, AdvPatch, or Diff-PGD were able to perturb the classification of the image. This was likely due to the fact that we did not adjust the scale properly. This likely means that hyperparameter tuning is very important for physical-world attacks.

4.1.3 Style-Based Attacks

4.2 Results beyond original paper

Often papers don't include enough information to fully specify their experiments, so some additional experimentation may be necessary. For example, it might be the case that batch size was not specified, and so different batch sizes need to be evaluated to reproduce the original results. Include the results of any additional experiments here. Note: this won't be necessary for all reproductions.

4.2.1 Additional Result 1

5 Discussion

Give your judgement on if your experimental results support the claims of the paper. Discuss the strengths and weaknesses of your approach - perhaps you didn't have time to run all the experiments, or perhaps you did additional experiments that further strengthened the claims in the paper.

5.1 What was easy

Running the code from the original authors to generate adversarial samples was fairly easy, since documentation was included in how to run each attack.

5.2 What was difficult

List part of the reproduction study that took more time than you anticipated or you felt were difficult.

Be careful to put your discussion in context. For example, don't say "the maths was difficult to follow", say "the math requires advanced knowledge of calculus to follow".

There was little to no documentation on the inner workings of the code, making it more difficult to understand a lot of the attack parameters, and what they did, however the naming was quite clear for the most part, so this was not too difficult to follow.

The physical-world attack was a lot more difficult to reproduce than we anticipated. AdvCam and AdvPatch both ran fairly quickly, averaging around 2-3s per iteration. However, Diff-PGD averaged around 60s per iteration, which took multiple days to run. Because of this, instead of running 4000 iterations as was done in the original paper, we only ran 1500 iterations for each image.
None of the models were able to perturb the classification of the image, which was likely due to the fact that we did not adjust the scale properly. This likely means that hyperparameter tuning is very important for physical-world attacks. However, tuning the model is more challenging for Diff-PGD for physical-world attacks, due to how slowly it runs. Another mistake we made was assuming the the classifier would classify a closed laptop as a laptop, when in reality it classified it as a notebook. We looked into the ImageNet database, and found that most of the images in both laptop and notebook classes were of open laptops, while we used a closed laptop. Since the

photo's original label was fed into the model, this could have possibly caused problems with the model.

5.3 Communication with original authors

We did not communicate with the original authors.

References

A Supplementary material

A.1 Physical attacks