[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.

Reproducibility Summary

*Template and style guide to ML Reproducibility Challenge 2020. The following section of Reproducibility Summary is **mandatory**. This summary **must fit** in the first page, no exception will be allowed. When submitting your report in OpenReview, copy the entire summary and paste it in the abstract input field, where the sections must be separated with a blank line.*

Scope of Reproducibility

State the main claim(s) of the original paper you are trying to reproduce (typically the main claim(s) of the paper). This is meant to place the work in context, and to tell a reader the objective of the reproduction.

Methodology

Briefly describe what you did and which resources you used. For example, did you use author's code? Did you re-implement parts of the pipeline? You can also use this space to list the hardware used, and the total budget (e.g. GPU hours) for the experiments.

Results

Start with your overall conclusion — where did your results reproduce the original paper, and where did your results differ? Be specific and use precise language, e.g. "we reproduced the accuracy to within 1% of reported value, which supports the paper's conclusion that it outperforms the baselines". Getting exactly the same number is in most cases infeasible, so you'll need to use your judgement to decide if your results support the original claim of the paper.

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

What was easy

Describe which parts of your reproduction study were easy. For example, was it easy to run the author's code, or easy to re-implement their method based on the description in the paper? The goal of this section is to summarize to a reader which parts of the original paper they could easily apply to their problem.

What was difficult

Describe which parts of your reproduction study were difficult or took much more time than you expected. Perhaps the data was not available and you couldn't verify some experiments, or the author's code was broken and had to be debugged first. Or, perhaps some experiments just take too much time/resources to run and you couldn't verify them. The purpose of this section is to indicate to the reader which parts of the original paper are either difficult to re-use, or require a significant amount of work and resources to verify.

Communication with original authors

Briefly describe how much contact you had with the original authors (if any).

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

Introduce the specific setting or problem addressed in this work, and list the main claims from the original paper. Think of this as writing out the main contributions of the original paper. Each claim should be relatively concise; some papers may not clearly list their claims, and one must formulate them in terms of the presented experiments. (For those familiar, these claims are roughly the scientific hypotheses evaluated in the original work.)

A claim should be something that can be supported or rejected by your data. An example is, "Finetuning pretrained BERT on dataset X will have higher accuracy than an LSTM trained with GloVe embeddings." This is concise, and is something that can be supported by experiments. An example of a claim that is too vague, which can't be supported by experiments, is "Contextual embedding models have shown strong performance on a number of tasks. We will run experiments evaluating two types of contextual embedding models on datasets X, Y, and Z."

We investigate the main claims from the original paper, which are:

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

Each experiment in Section 4 will support (at least) one of these claims, so a reader of your report should be able to separately understand the *claims* and the *evidence* that supports them.

3 Methodology

Explain your approach - did you use the author's code, or did you aim to re-implement the approach from the description in the paper? Summarize the resources (code, documentation, GPUs) that you used.

3.1 Model descriptions

Include a description of each model or algorithm used. Be sure to list the type of model, the number of parameters, and other relevant info (e.g. if it's pretrained).

3.2 Datasets

For each dataset include 1) relevant statistics such as the number of examples and label distributions, 2) details of train / dev / test splits, 3) an explanation of any preprocessing done, and 4) a link to download the data (if available).

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

Include a description of how the experiments were set up that's clear enough a reader could replicate the setup. Include a description of the specific measure used to evaluate the experiments (e.g. accuracy, precision@K, BLEU score, etc.). Provide a link to your code.

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

Physical-World Attacks:
We first tried the one of the attacks created by the authors, which was a computer-mouse.
We then tried our own physical world attack using an image patch of a ... and a ... as our target object.
We use an (type of phone here, ex. iPhone 8-Plus) to take images from the real world and use an (type of printer here, ex. HP DeskJet-2752) to print the image in color.
We stuck the original image on ..., and classified it using all 5 classifiers (R50, R101, R18, WR50, WR101)

We tested for Success Attack Rate of Diff-PGD using 250 uniformly sampled images from our dataset. (See figure ...)
The code for the figure in the paper was not provided, so we created our own code to generate the figure.

We also need to generate anti-purification table from paper, but I'm not sure how to generate this.

Transferability: Figure 6b+6c
We also test the success rate attacking adversarially trained ResNet-50

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.

Original paper results

Sample	(+P)ResNet50	(+P)ResNet101	(+P)ResNet18	(+P)WRN50	(+P)WRN101
x_{PGD}	0.35	0.18	0.26	0.20	0.17
x_n (Ours)	0.35	0.18	0.26	0.20	0.17
x_n^0 (Ours)	0.35	0.18	0.26	0.20	0.17

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.

4.1.1 Result 1

4.1.2 Result 2

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

4.2.2 Additional Result 2

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

Give your judgement of what was easy to reproduce. Perhaps the author's code is clearly written and easy to run, so it was easy to verify the majority of original claims. Or, the explanation in the paper was really easy to follow and put into code.

Be careful not to give sweeping generalizations. Something that is easy for you might be difficult to others. Put what was easy in context and explain why it was easy (e.g. code had extensive API documentation and a lot of examples that matched experiments in papers).

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".

5.3 Communication with original authors

Document the extent of (or lack of) communication with the original authors. To make sure the reproducibility report is a fair assessment of the original research we recommend getting in touch with the original authors. You can ask authors specific questions, or if you don't have any questions you can send them the full report to get their feedback before it gets published.

References