NSFW-Ninja: Masters of Disguise in the Content Filter Jungle

Neelesh Verma, Jagadeesh Reddy Vanga, Kamalnath Polakam
{neverma, jvanga, kpolakam}@cs.stonybrook.edu

Abstract

The proliferation of social media platforms has led to an increased need for content moderation, with NSFW (Not Safe For Work) detectors playing a pivotal role in maintaining a safe and respectful online environment. However, these detectors are not immune to manipulation, raising concerns about their effectiveness and reliability. This work presents a systematic investigation into the vulnerabilities of NSFW detectors through a black-box attack methodology. Additionally, we present an adversarial attack on the existing NSFW detectors present on popular social media sites. We will systematically perturb the NSFW images and observe the response from the detector. The direction of the perturbation will move the image closer to the detection boundary. In summary, this project seeks to investigate the robustness and expose vulnerabilities in NSFW detectors, demonstrating their susceptibility to manipulation and contributing to the development of more robust and ethical content moderation systems.

1 Introduction

In the digital age, the widespread use of social media platforms has revolutionized the way we communicate, share, and interact with content online. This transformation has brought forth an increased necessity for content moderation, aiming to maintain a safe and respectful online environment for users of all ages. At the forefront of this moderation effort are NSFW (Not Safe For Work) detectors, AI-based systems designed to automatically identify and flag content that may be explicit, offensive, or otherwise unsuitable for public consumption.

These NSFW detectors have become indispensable tools for social media platforms, playing a pivotal role in safeguarding users from inappropriate or harmful content. However, the effectiveness and reliability of these detectors have come under scrutiny in recent years. While they have undoubtedly made significant strides in improving content moderation, they are not impervious to manipulation.

Although NSFW Content consists of text and images, our focus is primarily on images. We will take images that are NSFW originally; and perturb them adversarially such that the final image isn't tagged as NSFW on social media platforms (like Reddit). At the same time, we will also maintain that the perturbation is very small and imperceptible to humans. This is generally done by limiting the L_∞ norm of the perturbation.

2 Related Work

There can be two kinds of adversarial attacks depending on the access that the adversary has - Black-box and White-box. In the Black-box setting, the adversary has no access to the network - models, weights, hyper-parameters, etc. However, some literature assumed access to the classification confidence scores (also referred to as semi-black-box in some work). White-box allows the attacker to have full access to the network parameters. Since we are dealing with social media websites, our attack comes under the black-box setting. Based on the implementation, the black-box attack can be

classified into two categories - *Transfer-based* attacks and *Decision-based* attacks. In transfer-based attacks, the adversary attacks a local surrogate model in a white-box setting and uses those adversarial examples for the black-box model. Query-based attacks prompt the model with an example, get the model output, and use this information to generate adversarial examples (approximating gradient information). These attacks require a lot of queries, so can be impractical at times.

For social media sites, it makes sense to use transfer-based attacks as performing thousands of queries might not be feasible. The existing state-of-the-art black box transfer-based attack uses data augmentation to create adversarial samples. These attacks typically augment multiple training images following a linear path during the adversarial sample generation process. Xie et al. [6] applied random transformations to the input image at each iteration of the adversarial attack and showed that the input diversity approach can be used to generate adversarial examples that are more transferable than state-of-the-art attacks. Translation invariant attack [1] generates an adversarial example for an ensemble of translated versions of the original input image. Admix [5] mixes the target image with a set of images randomly sampled from other categories. PAM [7] initially creates a pool of candidate augmentation paths and then determines which augmentation paths to use during the generation of adversarial samples through a greedy search process. Adversarial attacks specific to NSFW detectors haven't been in any form - black or white box. In this report, we will analyze the performance of some of these and use it for comparison with our method.

3 Motivation and Scope

The ubiquity of social media platforms has created an urgent need for effective content moderation, particularly in identifying and filtering out NSFW (Not Safe For Work) content. NSFW detectors serve as the first line of defense in protecting users from potentially harmful or inappropriate material. However, their vulnerabilities to manipulation and evasion have raised serious concerns. To our knowledge, there hasn't been any work that tries to perform NSFW-based adversarial attacks. This research is motivated by the imperative to comprehensively understand and address the limitations of NSFW detectors in social media. By conducting black-box attacks on these systems, we aim to uncover vulnerabilities, assess their resilience, and propose strategies for enhancement. Our work seeks to contribute to developing more robust content moderation mechanisms, ultimately fostering safer and more respectful online environments.

4 Implementation Method

We will be leveraging a hybrid of transfer-based and query-based black-box methods. We are targeting Reddit in the initial phase, later we wish to transfer the attack to Facebook as well.

We noticed that in an image, there are certain features that contribute highly to labeling the image to a particular class. Grad-CAM [4] scores have been extensively used for visual explanations about the classification of images. It generates a heat-map that highlights the regions of an image that are most important for a neural network's decision regarding a specific class. Based on the heat-map, we plan to give a score to each pixel in the original image. These scores will be used to perturb the image. We can then use a gradient-based approach like FGSM (Fast Gradient Sign Method) [2] or PGD (Projected Gradient Descent) [3] to iteratively modify the image in a direction that decreases the Grad-CAM scores and increases the likelihood of misclassification. Since we don't have access to the Reddit detector, we will train a surrogate model and do a transfer-based attack.

Consider the original NSFW image as X; assume there is a function that gives each pixel a score based on Grad-CAM generated heat-map Grad-CAM(X). First, we will train a surrogate model that mirrors the Reddit detector on our dataset. Assume that the NSFW detector is some function f (our surrogate model), so $f(X) = y$, where y denotes the NSFW label. We wish to reduce the Grad-CAM score but only up to that point, where the image just starts to get misclassified. So, we will slowly perturb the image, where at each step, the perturbation will move the image in the direction that reduces the Grad-CAM score, but as soon as the misclassification starts, we will stop. To check for the misclassification at each step, we will query Reddit. We hope that we can reach an adversarial image in a limited number of queries (less than 100) by appropriately adjusting the step length. The transfer-based attacks will come in handy when we will be transferring this example to other sites like Facebook. But initially, we are focusing on generating one adversarial example for Reddit.

5 Evaluation Plan

To the best of our knowledge, we haven't found a standard NSFW dataset. Ideally, we wish to use models like DALL-E 2 and 3 to generate the NSFW images for us, but because of the moderation and filters in them, it may not be feasible. We will try to bypass it if we find something. Otherwise, we can scrape Reddit for images. So, we will use either the DALL-E 2 generated images or the scraped image from Reddit and first test them on Reddit to see how many of them are detected as NSFW. Then these images will become our dataset on which we will train and test our method (will split 80:20 for train:test). The evaluation of success is based on our ability to post NSFW content on these platforms without getting automatically tagged by the detectors. If we were to fail to do the same, we would present a detailed analysis of the attack with reasons for the failure.

6 Timeline

The approximate timeline is as follows -

• Construct a dataset consisting of NSFW and SFW images	October 6, 2023
• Train an NSFW detector on this dataset	October 13, 2023
• Do a white-box attack on this detector using Grad-CAM scores (as described in implementation)	October 27, 2023
• Check the transferability of these images on Reddit	Nov. 3, 2023
• Improve the surrogate model to become as close to the target model as possible on this dataset	Nov. 17, 2023
• Improve the generalizability of the surrogate model by fine-tuning on diverse dataset	Nov. 24, 2023
• [Optional] Check the transferability on Facebook and further, fine-tune accordingly	Dec. 1, 2023
• Work on the final report and presentation	Dec. 5, 2023

References

[1] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[5] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021.

[6] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[7] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8173–8182, 2023.