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.

Adversarial attacks refer to a class of techniques and methods in the field of machine learning and artificial intelligence (AI) that are used to manipulate or deceive machine learning models. These attacks are designed to exploit vulnerabilities in machine learning algorithms and neural networks in particular. Szegedy et. al. [11] first showed that object recognition systems can be fooled by attacking the MNIST dataset. Since then many different approaches have been developed to exploit detection and recognition systems. In this report, we will assume that the content moderation systems are recognition systems that classify images as NSFW or SFW. Such attacks fall under the white-box and black-box attacks categories. In the white-box scenario, attackers have the ability to access the structures and parameters of the target models, enabling them to create adversarial examples. Conversely, in the black-box scenario, attackers lack any access to the model's structure and parameters. Since we are attacking content moderation systems we cannot access, our method will fall under the black-box setting.

However, existing black-box attack either uses a huge amount of queries [7, 5, 1, 15] or leverages target model training data for a transfer-based attack [26, 24, 16, 25, 9, 22]. We don't have access to the training data of the target model and neither we can perform a huge amount of queries on content moderation systems. Therefore, we opt for a hybrid approach that performs a very small amount of queries and anchors on the transferability of intermediate features across DNN models [19].

2 Literature Review

Adversarial attacks can typically be categorized into two primary categories: white-box attacks and black-box attacks. In a white-box setting, attackers possess detailed information about the victim models, including model structure, parameters, weights, training methodology, etc. In contrast, in a black-box setting, attackers lack access to such information and can only access the output from the model. White-box attacks often rely on exploiting gradient information from the victim model to craft adversarial examples. Prominent white-box attack methods include the Fast Gradient Sign Method (FGSM) [12], Project Gradient Descent (PGD) [17], Carlini and Wagner Attack (C&W) [4], Deepfool [18] and BPDA [2]. However, the white-box attack is unrealistic in our scenario since we don't have access to the content moderation models deployed by the social media sites.

In contrast, black-box attacks occur when attackers do not have access to vital information about the victim model. This setting is more reflective of real-world applications, where the inner workings of the model are hidden from potential attackers. Our attack on content moderation systems also falls under this category. There are mainly two categories of Black-box attacks - *Decision-based attacks* and *Transfer-based attacks*.

2.1 Decision-based attacks

The decision-based attacks query the target model and get the final label (in the classification task) from the target model.

Brendel et al.[3] proposed the first approach that involves a random walk on the decision boundary. In each iteration, the approach randomly selects a direction and projects it onto a boundary sphere to create a high-quality adversarial example, but it's query-intensive and lacks convergence guarantees. Guo et al.[13] proposes Low Frequency Adversarial Perturbation(LFAP) which made few modifications to boundary attack to construct low frequency perturbation. In this method, instead of sampling Gaussian noise low frequency noise is being sampled. By restricting to low-frequency subspace, which has a larger density of adversarial directions, this step succeeds more often, speeding up the convergence towards the target image. Meanwhile, a Query-Limited attack [14] focuses on estimating output probability scores through model queries to transform hard-label attacks into soft-label problems.

On the other hand, [6] takes a different approach, reformulating hard-label attacks as optimization problems aimed at finding the direction that minimizes the distance to the decision boundary. In practical tests, the algorithm efficiently targeted hard-label black-box Convolutional Neural Network (CNN) models on MNIST, CIFAR, and ImageNet, requiring significantly fewer queries (in orders of thousands). The Sign-OPT attack [7] follows a similar optimization approach as [6], treating hard-label attacks as the task of finding the direction with the shortest distance to the decision boundary. Additionally, it efficiently estimated the gradient's sign in any direction, rather than the gradient itself, requiring just a single query. A more recent attack [5] utilized the zeroth-order sign oracle to enhance the Boundary attack, resulting in substantial improvements. They employed a one-point gradient estimate, which, while unbiased, can have higher variance compared to the gradient estimate in [7].

Even though the number of queries required has significantly reduced over the years, their budget is still a big issue due to the extremely small query budget in the content moderation systems. It seems that the robust detection and flagging systems being employed in the most popular social media demand a new attack with just a few queries (in order of a hundred).

2.2 Transfer-based Attacks

Transfer-based black-box attacks are a type of adversarial attack where the attacker generates adversarial examples using a surrogate white-box model and then transfers them to an unknown target black-box model. Transfer-based attacks work because adversarial examples can often fool similar models, and deep learning models are sensitive to small input changes.

Cheng et al.[8] propose a method called Prior-Guided Random Gradient-free(P-RGF) which uses the gradient of a surrogate model as a prior to guide the search direction and then adjusts the direction based on the query feedback from the target model. One of the limitations of the P-RGF method is that it requires a surrogate model that has similar architecture and training data as the target model which might not be possible at times. Another limitation is that it assumes that the target model is deterministic and doesn't have any defense mechanisms such as randomization or gradient masking.

Wang et al.[23] tries to overcome these shortcomings by proposing a novel input transformation-based attack called Structure Invariant Attack(SIA), which applies a random image transformations onto each image block to generate a set of diverse images for gradient calculation. This improves the transferability of the adversarial examples, which can exploit common vulnerabilities of different models and bypass their defenses. Although successful, it may reduce their stealthiness and make them easier to detect by humans, unlike other adversary measures.

A patch-based attack was proposed by Gao et al. [10]. Instead of manipulating the images pixel-wise, it tries adding patches to the image. They incorporated an amplification factor in the FGSM method to increase the step size in each iteration, ensuring that when a pixel's gradient exceeds the ϵ constraint, it is accurately distributed to its neighboring regions through a projection kernel. Zhang et. al. [26] proposed a feature-level attack that employs neuron importance scores. The scores are computed by attributing the model's output to each neuron in the network and total neuron attribution is minimized to craft adversarial examples. But for our scenario, we have no access to the data that is used to train the content moderation models.

All of these transfer-based adversarial attacks assumed that the training data of the target model follows similar distributions as the surrogate model. Zhang et. al. [15] proposed to train the surrogate models in a data-free black-box scenario. They used a GAN for data generation and leveraged model distillation on the substitute model which acts as the discriminator. But they had to make a trade-off with a huge amount of query. For attack on content moderation systems, we have a very small budget for the number of queries (it may be possible to leverage APIs if available to perform such queries but for most of the social media platforms - Reddit, Facebook, Instagram, etc, such APIs aren't available for free). Therefore, we propose a hybrid approach - an extremely limited query-based transfer attack. It has been shown that the DNN models share similar features in their receptive fields [24]. Therefore, even though we don't have the training data, we can generate our own data [20]. The trained model on this data would then share similar features to the target model. We will be leveraging Grad-Cam [21] scores to give scores to the features and the input pixels and minimize this score by querying the target model.

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 [21] 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) [12] or PGD (Projected Gradient Descent) [17] 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 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] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1277–1294. IEEE, 2020.

[6] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

[7] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019.

[8] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] 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.

[10] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 307–322. Springer, 2020.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

[13] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Uncertainty in Artificial Intelligence*, pages 1127–1137. PMLR, 2020.

[14] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[15] Jianghe Xu Shuang Wu Shouhong Ding Lei Zhang Chao Wu Jie Zhang, Bo Li. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022.

[16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[17] 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.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[19] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*, 2018.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[21] 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.

[22] 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.

[23] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023.

[24] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1158–1167, 2020.

[25] 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.

[26] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022.