ECE661 Final Project: Adversarial Patch Attack

Mutian Ling (ml646), Chenjie Yang (cy141), Xiaoyu Wang (xw164)

Dec 18th, 2023

Abstract

This paper explores a novel approach to adversarial attacks on deep neural networks, focusing on the practical application of physically deployable patches. Unlike conventional attacks reliant on imperceptible changes, our study introduces image-independent patches strategically placed in an image. We demonstrate their effectiveness on the CIFAR-10 dataset using a ResNet-20 model, showing transferability across different architectures. Results reveal that larger patches consistently outperform smaller ones, highlighting their potential impact in real-world scenarios. This work underscores the need for robust model defenses and opens avenues for future research in real-world settings. Our work can be found via this link: https://github.com/chenjieyang326/adversarial_attack

1. Introduction

In recent years, advancements in deep neural networks, especially with convolutional neural networks (CNNs), have transformed tasks like object and facial recognition. However, this success brings a new challenge—vulnerability to adversarial attacks, posing risks in real-world applications. For example, in finance, adversarial attacks on fraud detection models could lead to misclassifying transactions, causing substantial financial losses and potentially destabilizing the financial system. Similarly, in smart home security, adversarial attacks on facial recognition models could compromise the system's integrity, misidentifying unauthorized individuals and jeopardizing overall security. This underscores the critical need for robust deep learning models, particularly in security-sensitive environments.

The motivation of this project comes from the increasing reliance on deep neural networks in critical applications, which necessitates a robust understanding of their vulnerabilities to adversarial attacks, ensuring the integrity and reliability of models in real-world deployment scenarios.

Adversarial attacks conventionally involve tweaking individual pixels in images by minute increments, employing optimization techniques such as L-BFGS, Fast Gradient Sign Method (FGSM), DeepFool, Projected Gradient Descent (PGD), among others. The subtle introduction of carefully crafted noise to images induces misclassifications by neural networks, often eluding human perception and complicating the detection of such attacks.

Despite their effectiveness in controlled settings, these attacks prove impractical when transposed into real-world scenarios. Their reliance on minimal and imperceptible alterations makes them heavily contingent on specific image attributes, demanding prior knowledge of nuanced factors like lighting conditions, camera angles, and classifier specifications. In response to these limitations, there is a burgeoning interest in exploring a paradigm shift towards "physical" attacks that can be tangibly printed and deployed without the need for intricate scene-specific information.

2. Related Work

Many researchers have studied the generalizability of adversarial attacks to the real world. A notable example by Kurakin et al. [1] demonstrated that when adversarial crafted images are physically printed, they consistently maintain their adversarial impact across varying lighting conditions and orientations. Another study by Athalye et al. [2] showcased the creation of 3D-printed adversarial objects, capable of consistently misleading neural networks at different angles and scales. These objects were essentially subtle modifications of ordinary items, such as a turtle manipulated to be misclassified as a rifle.

Additionally, prior work [3] has explored adversarial attacks on facial recognition systems, including the construction of adversarial glasses capable of deceiving the model. Evtimov et al. [4] introduced methods for constructing deceptive stop signs, either through printing out posters resembling stop signs or attaching various stickers to manipulate model classification.

While most previous research has focused on attacks involving small or imperceptible changes to inputs, this work takes a different approach. Rather than attempting to subtly transform existing items, our study introduces an attack that generates an image-independent patch highly salient to a neural network. This patch can be strategically placed within the original image, inducing the classifier to output a targeted class. Importantly, this patch is image-independent, allowing attackers to execute real-world attacks without prior knowledge of lighting conditions, camera angles, the specific classifier in use, or even other elements within the scene.

3. Methodology

3.1. Dataset

CIFAR-10 is widely acknowledged as a benchmark dataset for image classification tasks, offering researchers a valuable platform for training and assessing machine learning models, particularly within the field of computer vision. We employ the CIFAR-10 dataset for our experimental purposes. The CIFAR-10 dataset comprises 60,000 color images, each with dimensions of 32 by 32 pixels. These images are categorized into 10 distinct classes. The dataset is divided into training and testing sets, containing 50,000 and 10,000 images, respectively. To ensure consistency, both datasets underwent normalization using parameters derived from the training set. The normalization process involved calculating statistics, such as mean and standard deviation, from the training dataset. The images of 10 classes are displayed in *Fig 1*.

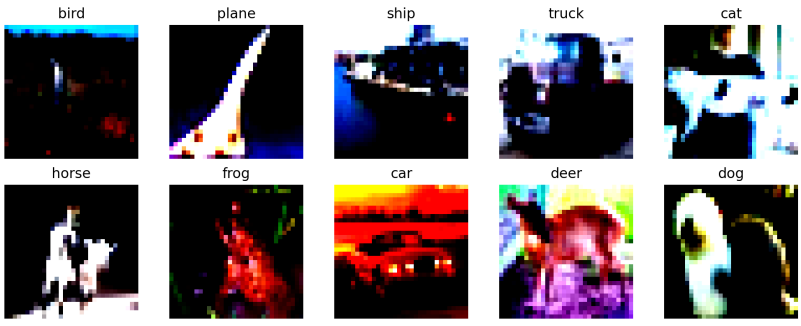

Fig 1. CIFAR-10

3.2. Models

We employed a Convolutional Neural Network (CNN) based on the ResNet-20 architecture to train on the CIFAR-10 dataset. Through further adjustments to our model, we successfully achieved a 90% accuracy on the test dataset. Subsequently, our trained model was utilized to generate both targeted and untargeted patches. After inserting those generated patches into the image, we used the trained ResNet-20 model as the white-box model to test the attack effectiveness. Also, we used ResNet-56, VGG-16, MobileNet, ShuffleNet and RepVGG as black-box models to test the transferability of the adversarial patch.

3.3. Patches

We began by initializing a rectangular patch of different sizes, positioning it randomly within the training images. Subsequently, the patch underwent optimization over 10 epochs, employing a learning rate of 0.1 and utilizing our trained ResNet-20 model. The optimization process was conducted using the Adam optimizer, chosen for its superior results. In the case of untargeted patches, optimization aimed to maximize the cross-entropy loss function on the training dataset. Conversely, for targeted patches, the optimization process involved minimizing the loss function specific to our target class.

Following the generation, the patches were randomly positioned within the test images. Subsequently, we assessed these patches by evaluating the performance of both white box and black box models on the modified test data. The evaluation criteria for untargeted patches included test accuracy, while for targeted patches, the assessment involved both test accuracy and the Attack Success Rate (ASR). In this project, ASR is defined as the percentage of such cases in which we were able to successfully fool the classifier into targeted label.

The generated targeted and untargeted patches of size 15x15 are shown in *Fig 2*.

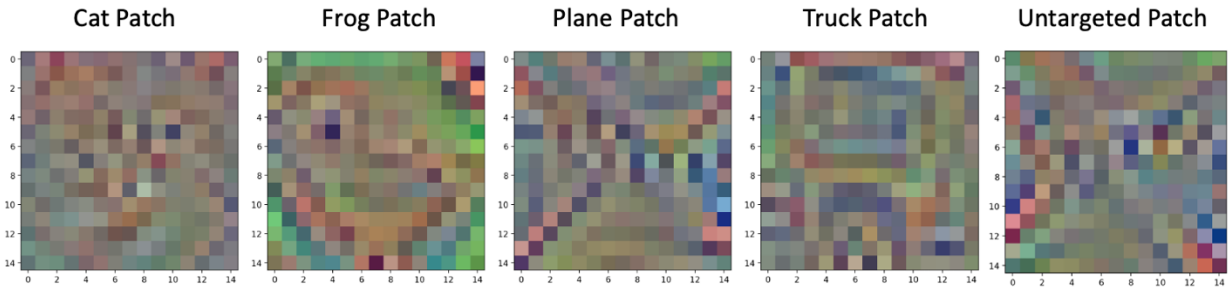

Fig 2. Generated Targeted & Untargeted Patches

We do not notice any visual patterns in our patches for several reasons. As you can observe from *Fig 2* above, the Cifar-10 dataset images are of small size containing only 32x32 pixels. Hence, when these images are viewed, they appear blurred.

4. Results & Discussion

4.1. White-box Attack

To perform the white-box attack, we generated patches of different sizes including 3x3, 7x7, 11x11, 15x15 for each class labels mentioned previously as well as the same-sized untargeted patches. These particular patch sizes were chosen to evaluate the impact of patch scale relative to the CIFAR10 images, which are 32x32 pixels in size; smaller patches (3x3) test the subtlety of the perturbation, while larger patches (up to 15x15) test the limits of conspicuousness and their ability to deceive the model even when occupying a significant portion of the image. The effectiveness of various sizes of patches are illustrated in Fig 3&4 respectively.

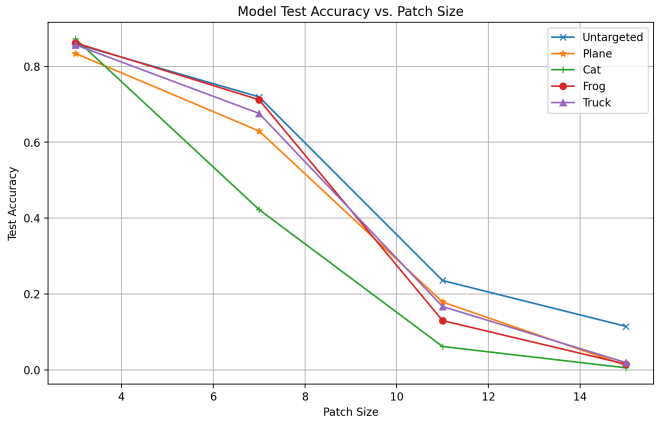

Fig 3. Test Accuracy across Patch Size in Whitebox Attack

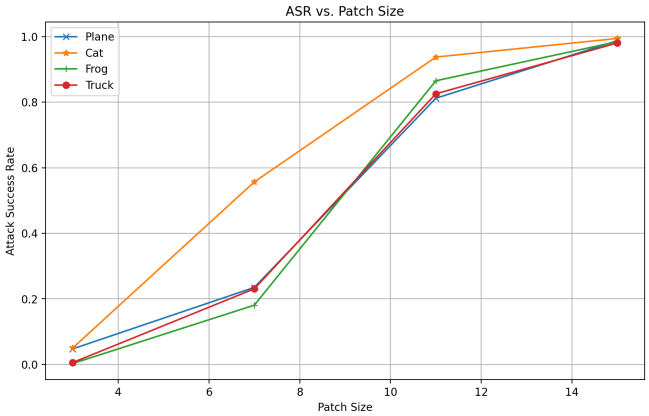

Fig 4. ASR across Patch Size in Whitebox Attack

It is evident from *Fig 3* that the test accuracy generally decreases as the size of the adversarial patch increases. For untargeted patches and targeted patches aimed at specific classes, there's a notable downward trend in test accuracy, especially as the patch size approaches 15x15. This trend suggests that larger patches have a more significant disruptive effect on the model's ability to correctly classify the images. The untargeted patch appears to have a slightly less detrimental effect compared to targeted patches, which could indicate that targeted perturbations are more effective at misleading the model.

Similar to the trend observed in test accuracy, *Fig 4* showcases that the ASR increases with the patch size. This increase is particularly steep between smaller patch sizes and plateaus as the patch size approaches 15x15. This trend suggests that there is a threshold beyond which increasing the patch size yields diminishing returns in terms of ASR. Classes such as *Plane* and *Cat* demonstrate a higher susceptibility to adversarial attacks, reaching near-perfect ASR with the largest patch size, while *Frog* and *Truck* classes show a slightly lower ASR, indicating a possible difference in how easily each class can be perturbed. This could be due to intrinsic features of the classes that either align with or resist the perturbations introduced by the adversarial patches.

4.2. Black-box Attack

To test the transferability across different models, we used ResNet-56, VGG-16, MobileNet, ShuffleNet and RepVGG as black-box models. These models were pre-trained on the CIFAR-10 training data and could be obtained from open-source website [5]. All these models achieved a high accuracy on the unperturbed test data, which is shown in the *Table 1* below. We applied both untargeted and targeted patches on the black-box models to evaluate the transferability of our attack, and the results are shown in *Fig 5&6* respectively.

Model	ResNet-20	ResNet-56	VGG-16	MobileNet	ShuffleNet	RepVGG
Test Acc	89.20%	94.37%	94.16%	94.21%	93.98%	95.27%

Table 1. Test Accuracy on CIFAR-10 across Models

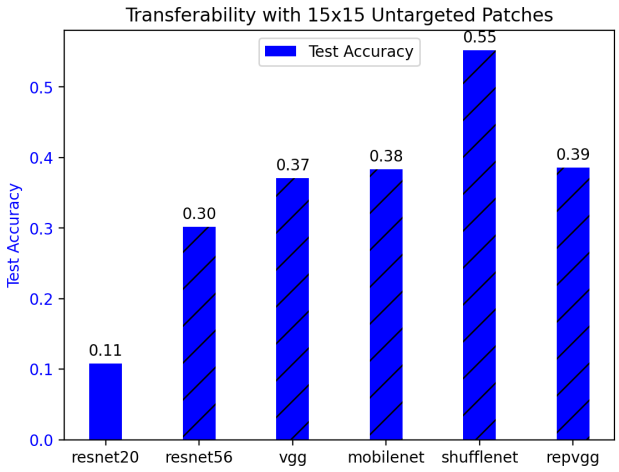

Fig 5. Transferability with 15x15 Untargeted Patches

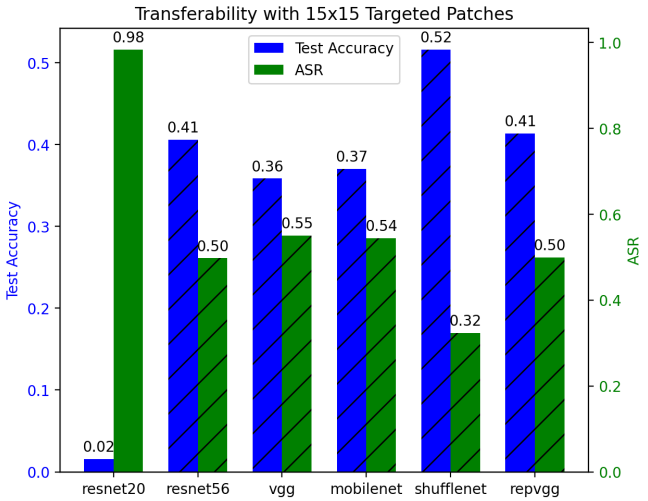

Fig 6. Transferability with 15x15 Targeted Patches

The trend in *Fig 5* indicates a significant reduction in test accuracy across all models, with ResNet-20 experiencing the most considerable decrease. This suggests that while the patch was designed using ResNet-20, its disruptive capabilities still generalize effectively to other architectures. However, it is noteworthy that the decrease in accuracy is not uniform across the models. For instance, ShuffleNet exhibits a moderate resilience to the untargeted patches, maintaining higher accuracy levels than other models. The variation in susceptibility among the models could be attributed to differences in architectural complexities and the models' inherent robustness to perturbations.

In *Fig 6*, the bars for test accuracy show that the models retain some level of resilience, with none dropping to zero. This indicates that while the targeted attack is successful to a degree, there is still a significant portion of the test set that remains correctly classified. On the other hand, the ASR demonstrates a high success rate for the attack, particularly on VGG-16 and MobileNet, suggesting that targeted patches can be highly effective in fooling the model into misclassifying an image as a specific class. It is intriguing to observe that the models which showed moderate resistance to untargeted patches are not equally resistant to targeted attacks, as evidenced by the high ASR. The results underscore the importance of considering both the test accuracy and ASR when evaluating the robustness of models against adversarial attacks, as a model may appear deceivingly robust when only one metric is considered.

In the context of other black-box testing models, our patch attack demonstrates remarkable and stable transferability. This implies that our attack method remains highly effective across diverse models, exhibiting a sustained level of performance.

5. Conclusion & Future Work

In conclusion, our exploration into adversarial patches within the CIFAR-10 dataset has uncovered intriguing findings. The success of these patches in misleading neural network models is evident, and an interesting trend emerges as larger patch sizes correlate with reduced accuracy in both white-box and black-box models. This trend suggests that the size of the adversarial attack patches plays a crucial role in their impact on model predictions.

Despite the challenges in identifying specific patterns within the complex CIFAR-10 dataset, the transferability across different models underscores the potential threat posed by adversarial attacks, emphasizing the need for robust model defenses.

Looking ahead, our future research direction involves a transition from simulated environments to real-world scenarios. By printing and applying these patches in diverse settings with varying conditions, we aim to assess their robustness and generalizability beyond the controlled confines of a digital dataset. For instance, we plan to investigate how these patches fare in different lighting conditions, against real-world objects, and in various environmental contexts. This real-world evaluation will contribute valuable insights into the practical implications of adversarial attacks and inform the development of more resilient machine learning models.

References

[1] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Examples in the Physical World," arXiv:1607.02533 [cs, stat], Feb. 2017, Available: https://arxiv.org/abs/1607.02533

[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust Adversarial Examples," *arXiv.org*, Jun. 07, 2018. https://arxiv.org/abs/1707.07397

[3] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on state-of-the-art Face Recognition," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, 2016, doi: https://doi.org/10.1145/2976749.2978392.

[4] K. Eykholt *et al.*, "Robust Physical-World Attacks on Deep Learning Models," *arXiv:1707.08945 [cs]*, Apr. 2018, Available: https://arxiv.org/abs/1707.08945

[5] chenyaofo, "PyTorch CIFAR Models," *GitHub*, May 12, 2023. https://github.com/chenyaofo/pytorch-cifar-models