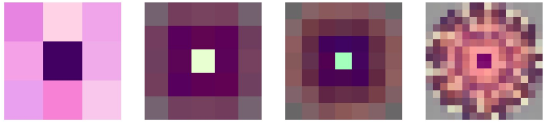
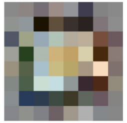


ECE 661: Adversarial Patch Attack

Jamie Liu, Hiep Nguyen, & Jack Parker



Introduction

Most research on adversarial attacks focuses on creating imperceptible perturbations in images, but in many cases, adversaries will be willing to sacrifice undetectability if creating more noticeable attacks carries greater benefits.

We carried out a white box attack against ResNet18 that involved creating a square-shaped adversarial patch. Applying this patch to a CIFAR-10 image causes ResNet18 to misclassify the sample with high probability.

Main Contributions

Successfully replicated Brown et al. [1] on CIFAR-10 using a **different objective function than the original paper** [2].

Investigated untargeted and targeted ASR for **many different types of patch transformations beyond what was explored in the original paper**

- Robust to translations and rotations (Type A)
- Robust only to translations (Type B)
- Robust to translations, rotations, horizontal flips, vertical flips, and color inversions (Type C)

Methodology

Algorithm 1: Adversarial Patch Attack

```
Initialize a tensor $p$ (the patch) of a desired size to all zeros;
Choose target label $q$ (or None for an untargeted attack);
for $i = 1$ to 5 do
    Fetch a batch of 32 training images;
    for each image do
        Transform $p$ with any desired transformations;
        Place $p$ at a random location on the image;
    end
    Make predictions $pred$ on the patched images;
    Modify labels $L$: Untargeted attack: $L' = (L + 1) \mod 10$; Targeted attack: $L = q$;
    Calculate cross entropy loss based on $pred$ and $L$;
    Use SGD with learning rate 0.1 and momentum 0.8 to update $p$
end
```

Experimental Results

Figure 1: Untargerted ASR as a function of patch size (Type A)

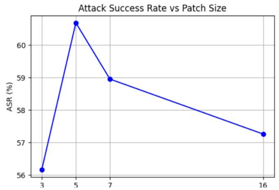
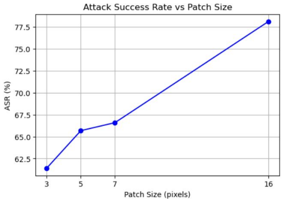
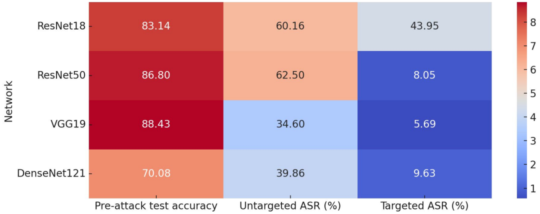


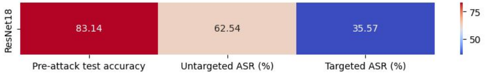
Figure 2: Untargeted ASR as a function of patch size (Type B)



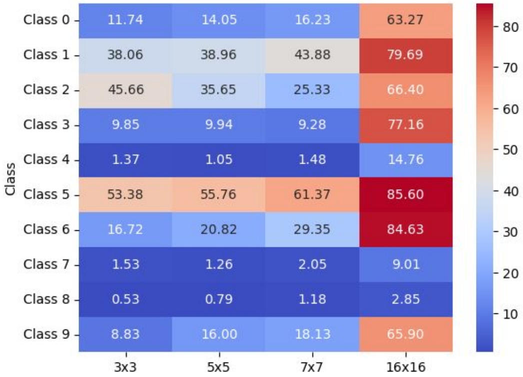
Transferability of Type A Patches



Type C patch performance on ResNet18



Targeted ASR for four patch sizes on each class (Type A)



Conclusion

Adversarial Patch Efficacy: Type B patches (robust only to translations) are extremely effective, especially when the patch size gets large. Type A patches are robust to rotations (think camera angle), but there is a significant sacrifice in ASR. Type C patches gain robustness to several more transformations without sacrificing much in terms of ASR.

Transferability Insights: Untargeted attacks exhibit higher transferability across models than targeted attacks, emphasizing the nuanced dynamics of adversarial strategies.

Patch Size and ASR: Targeted attacks necessitate larger patches for higher ASR, whereas untargeted approaches require less conspicuous modifications.

Class-Specific Vulnerability: Targeted attack efficacy varies significantly by class, highlighting the importance of tailoring approaches to specific model vulnerabilities.

References

[1] T. Brown, D. Mané, A. Roy, M. Abadi, \& J. Gilmer, "Adversarial Patch," arXiv:1712.09665, May 17, 2018. [Online]. Available: https://arxiv.org/abs/1712.09665. [Accessed Apr. 18, 2024].

[2] P. Lippe, "Tutorial 10: Adversarial Attacks," uvadlc-notebooks.io, 2022. [Online]. [Accessed Apr. 18, 2024].