Investigating the Effectiveness of Adversarial Patches in Image Classification Disruption

Xinhao Kong, Ruolin Meng, Yang Ouyang
ECE 661, Computer Engineering Machine Learning and Deep Neural Nets, Fall 2023

Abstract

This project details our successful implementation of the adversarial patch method, a novel approach in the realm of image classification disruption. Our investigation primarily revolved around the comprehensive testing of adversarial patches, **examining the impact of varying patch sizes and locations** across several state-of-the-art (SOTA) models. We conducted exhaustive experiments in **both targeted and untargeted attack scenarios** to evaluate the efficacy and adaptability of our approach. A significant aspect of our research included **exploring the transferability of attacks**, assessing their versatility. These systematic tests have provided valuable insights into the robustness of adversarial patches and their potential applications in manipulating image classifiers.

Introduction

Our project focuses on adversarial patches, a distinct and effective attack method that diverges from traditional subtle perturbations. These patches are conspicuous modifications applied to images, aimed at deceiving classification models.

Our comprehensive research involves developing these patches and rigorously testing their effectiveness across various dimensions, such as size and location. We delve into both targeted and untargeted attack strategies to assess their impact under different scenarios.

A crucial part of our study examines the transferability of adversarial patches, applying them across diverse models to evaluate their universal applicability and implications for machine learning systems. This poster presents our methodologies, key findings, and the broader implications of our work in enhancing the security and robustness of machine learning models.

Methodology

Our methodology encompasses a systematic approach to develop and evaluate adversarial patches in image classification contexts:

1. Patch Development: We created adversarial patches using an optimization-based approach. These patches are designed to be visually noticeable, in contrast to traditional adversarial attacks.

2. Attack Strategies:
- **Targeted Attacks**: We manipulated patches to trick models into misclassifying images as a specific, incorrect label.
- **Untargeted Attacks**: Here, the goal was to induce any incorrect classification without a specific target label.

3. Experimentation Parameters:
- **Patch Sizes and Locations**: To understand the influence of physical characteristics on the attack's success.
- **Model Diversity**: We tested the patches on various state-of-the-art image classification models.

Results

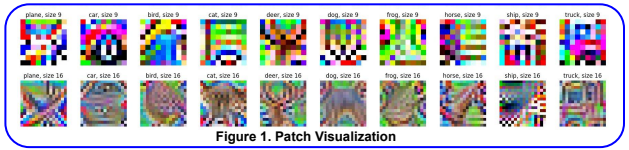

Figure 1. Patch Visualization

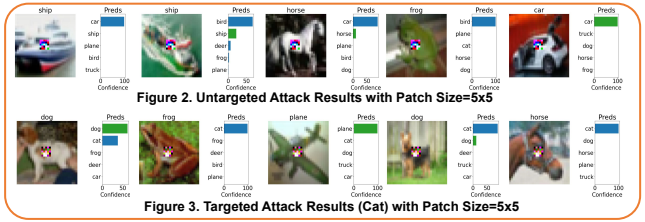

Figure 2. Untargeted Attack Results with Patch Size=5x5

Figure 3. Targeted Attack Results (Cat) with Patch Size=5x5

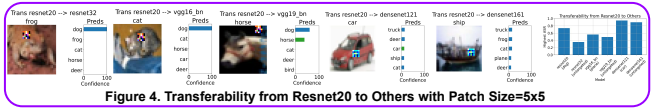

Figure 4. Transferability from Resnet20 to Others with Patch Size=5x5

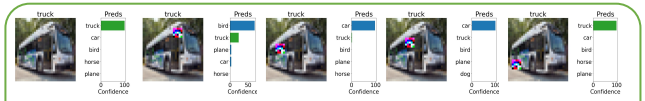

Figure 5. Effects of Patch Location

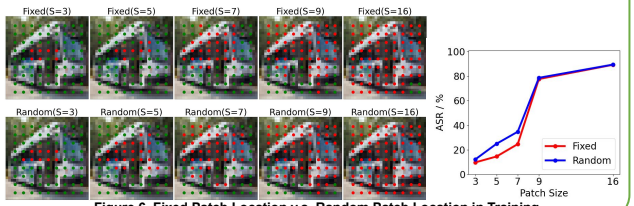

Figure 6. Fixed Patch Location v.s. Random Patch Location in Training

Methodology (Continue)

4. Transferability Testing: To assess generalizability, we evaluated whether patches trained on one model were effective on others, revealing insights into the cross-model applicability of these patches.

5. Evaluation Metrics: Our assessments were based on the success rate of attacks, considering both the rate of misclassification in targeted attacks and the overall error rate in untargeted scenarios.

Conclusion

Our study delves into adversarial patches affecting image classification models, offering novel training methods and analyzing their transferability. Key findings include:

- **Effectiveness of Patches**: Adversarial patches, when well-trained, notably impact model performance in targeted and untargeted attacks.
- **Transferability Across Models**: Patches show varying transferability across models, highlighting diverse model vulnerabilities.
- **Implications for Model Robustness**: The results emphasize the need for tailored security approaches in model development.

In summary, our research significantly advances the comprehension of adversarial attacks, vital for enhancing machine learning system security.

Additional Insights

GAN based Patch Training: In our research on adversarial patches, we compared the effectiveness of Generative Adversarial Networks (GANs) with direct training methods. GANs showed promise but were often less effective, particularly due to the absence of a dedicated discriminator. This limited their ability to create inconspicuous yet effective patches. To improve this, we suggest **incorporating a real discriminator, enhancing loss functions, and experimenting with advanced GAN architectures**. Our findings indicate that combining GANs' generative strengths with direct training might be a more effective approach for developing resilient neural network models.

References

[1] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *Advances in Neural Information Processing Systems*, pages 3225–3234, 2017.

[2] Xin Liu, Huan Cheng, Hao Zhang, Yukun Zhang, and Bo Niu. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018.

[3] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow.Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *In Privacy Enhancing Technologies*, pages 18–37. Springer, 2016.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[5] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. In *European Symposium on Research in Computer Security*, pages 653–670. Springer, 2017