Advanced Attacks on Cifar-100

r12921100 Hung Yang, Yeh

Abstract

This study explores targeted and universal adversarial attacks on CIFAR-100 dataset [1] models, manipulating 500 images to misclassify as 'apple' with a perturbation limit of 4/255. The ensemble, composed of 'resnet110' [5], 'preresnet164bn' [6], 'seresnet110' [7], 'densenet40_k36_bc' [8], and 'diaresnet164bn' [9], averaged post-softmax probabilities for classification. A universal perturbation, more complex due to its broad applicability, was also developed with a limit of 12/255 to affect a 'resnet20_cifar100' model [5] handling 200 images.

Targeted attacks precisely adjusted each image within the perturbation budget, significantly reducing the ensemble model's accuracy. The universal attack created a perturbation effective across various images, proving its stealth and efficacy by severely impairing the 'resnet20_cifar100' [5] model's accuracy. These results highlight the susceptibility of CNNs to adversarial attacks and provide valuable insights into the robustness of ensemble models.

1 Introduction

1.1 Adversarial Attacks

Adversarial attacks subtly manipulate image classification models, posing significant security and reliability concerns. These attacks, often imperceptible to humans, lead models to incorrect classifications, highlighting vulnerabilities and pushing the need for resilient neural networks.

1.2 Goals

This project aims to execute targeted adversarial attacks and develop universal adversarial perturbations (UAPs) using the CIFAR-100 dataset [1] and an ensemble of sophisticated models. Our objective is to assess and enhance the robustness of these models against both targeted and untargeted adversarial threats.

1.3 Strategies

We employed advanced algorithms including FGSM [2], PGD [3], and the CW attack [4]. PGD proved highly effective, compromising the model with a 94.4% success rate. The development of UAPs involved iterative refinement of perturbations, integrating a learning rate decay to optimize efficacy while maintaining stealth. These strategies exposed significant vulnerabilities, significantly reducing the accuracy of the Resnet20 model [5] to a mere 3.5%, underscoring the need for resilient neural networks.

2 Methodology

2.1 Ensemble Model

The ensemble model integrates five advanced neural network architectures from the ResNet [5] and DenseNet [8] families, optimized for the CIFAR-100 dataset [1]. By averaging the outputs of

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

these diverse models, the ensemble enhances prediction robustness and resilience against adversarial attacks.

2.2 Targeted Attack

Targeted adversarial examples were generated using the Projected Gradient Descent (PGD) method [3], with specific hyperparameters to steer misclassifications toward a predetermined class. The PGD was executed with a learning rate of 1/510 and 150 iterations, with an epsilon of 4/255, ensuring subtle yet effective perturbations. This approach achieved a 94.4% success rate in misleading the ensemble model.

2.3 Universal Attack

The development of UAPs began with a random initial perturbation and involved an iterative optimization process that adjusted the perturbation using a decayed learning rate and gradient updates to maximize model loss. Each perturbation was constrained by an epsilon limit to maintain imperceptibility. The most effective perturbation, which significantly reduced the model's accuracy to 3.5%, was selected after multiple trials.

3 Experiment

3.1 Targeted Attack

We evaluated the effectiveness of FGSM [2], PGD [3], and CW attack [4] methods on an ensemble of neural networks using the CIFAR-100 dataset [1]. Each method was tested under specific configurations to exploit different vulnerabilities of the model.

3.1.1 Experimental Configuration

- **FGSM:** Quick execution with a fixed epsilon, completed in 5 seconds, but limited effectiveness (2.0% success rate).
- **PGD:** Iterative adjustment with a step size of $\alpha = \frac{1}{1020}$ over 150 steps, taking 15 minutes, and achieving a 94.4% success rate.
- **CW Attack:** High-confidence misclassifications with parameters $kappa = 20$, $c = 1$, and 150 iterations, also taking 15 minutes and reaching a 39.6% success rate.

3.1.2 Performance Evaluation

The comparative analysis highlighted the trade-offs between speed and effectiveness, as shown in Table 1 and Figure 1, with PGD [3] emerging as the most potent method against the ensemble model.

Table 1: Success Rates of Ensemble Model Eval by Attack Type

Attack Type	Success Rate
FGSM	2.0%
PGD	94.4%
CW Attack	39.6%

3.2 Universal Attack

The development of Universal Adversarial Perturbations (UAPs) involved refining a perturbation through 150 iterations with an initial $\alpha = \frac{4}{255}$ and a decay mechanism that reduced it by 30% every 15 iterations.

Selection of Optimal UAP: The most effective UAP lowered the model's accuracy to 3.5%, demonstrating the critical role of parameter optimization and iterative refinement in developing effective UAPs. This approach highlighted key vulnerabilities and underscored the need for robust defenses against universal perturbations.

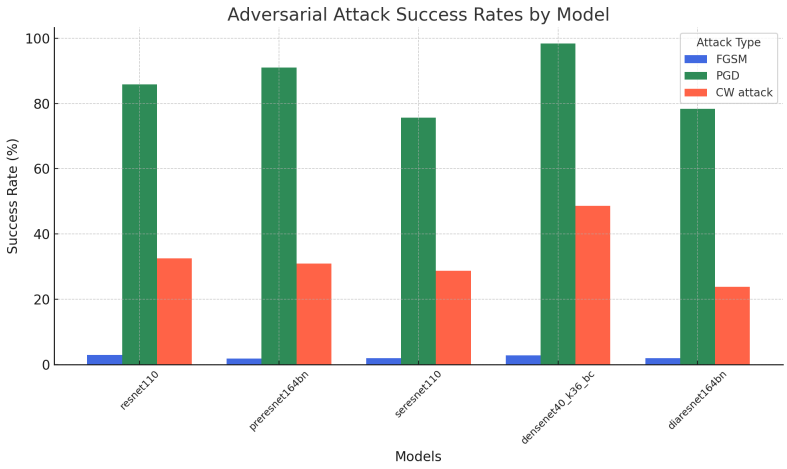

Figure 1: Targeted attack success rates under different model

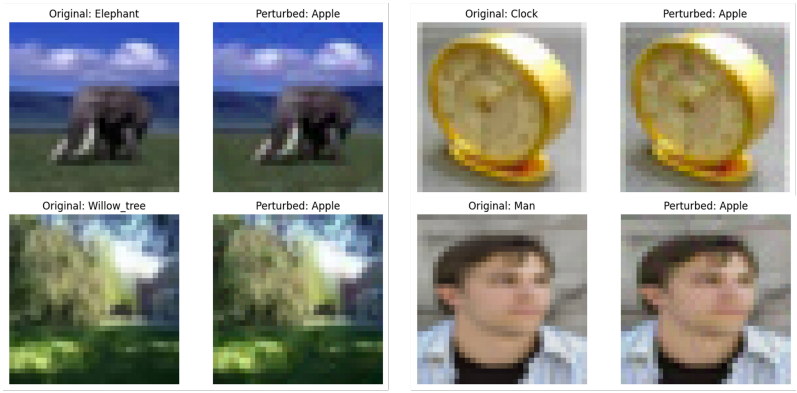

Figure 2: Images before and after the attack

4 Analysis

4.1 Targeted Attack

The targeted adversarial attack on the ensemble model achieved a notable success rate, with 472 out of 500 images erroneously classified as "apple" following the application of a minimal perturbation ($\frac{4}{255}$). This high efficacy, representing a success rate of 94.4%, highlights the susceptibility of the model to well-crafted adversarial inputs even under stringent perturbation constraints.

Figure 2 displays comparative visuals of select images before and after the attack. These comparisons underline the subtle yet effective nature of the perturbations, which, while nearly imperceptible to the human eye, significantly mislead the model's classification capabilities.

This successful manipulation of model predictions with minimal input modification signals a critical vulnerability in neural network architectures, emphasizing the urgent need for robust defense mechanisms in models deployed in security-sensitive environments.

In conclusion, the results from this targeted attack underscore the delicate balance between model accuracy and adversarial robustness, urging ongoing advancements in defensive strategies against adversarial threats.

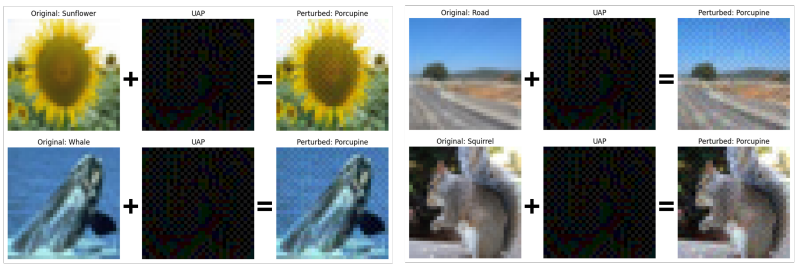

Figure 3: Examples of images altered by the UAP

4.2 Universal Attack

During the universal adversarial attack experiment, a UAP with pixel perturbations ranging from 0 to 24 was applied, leading to a drastic reduction in the accuracy of a Resnet20 model [5]. This perturbation decreased the model's accuracy on a selected set of 200 CIFAR-100 images to just 3.5%.

Figure 3 presents examples of images altered by the UAP. These visuals provide insight into the effectiveness of the universal perturbation across various images.

Interestingly, out of 200 perturbed images, 188 were misclassified as "porcupine", accounting for 97.4% of the misclassified cases. This overwhelming tendency towards a single incorrect class demonstrates a significant bias introduced by the UAP, highlighting an intriguing and somewhat inexplicable phenomenon within the model's response to the perturbation.

The unexpected classification pattern raises questions about the specific features in the "porcupine" class that might be overly generalized or overly sensitive in the model's training. This indicates potential areas of vulnerability within the neural network's architecture and training data, suggesting a need for further investigation into model biases and the development of more robust adversarial defense mechanisms.

Overall, the universal attack's success and the peculiar classification results provide crucial insights into the underlying vulnerabilities of the model, emphasizing the importance of ongoing research into effective countermeasures against such adversarial strategies.

5 Conclusion

This study explored the vulnerabilities of neural networks to adversarial attacks through targeted and universal perturbations on an ensemble of models trained on the CIFAR-100 dataset [1]. The findings reveal significant susceptibilities that can be exploited through well-crafted adversarial inputs. The targeted attack, with a success rate of 94.4%, demonstrated that even minimal perturbations could lead to a high rate of misclassification, emphasizing the need for more robust defensive mechanisms in neural networks.

The universal adversarial perturbation, reducing the model accuracy to as low as 3.5%, highlighted broader vulnerabilities and the effectiveness of attacks that do not target specific model outputs but rather seek to degrade performance across a range of inputs. The surprising bias towards the "porcupine" class in the majority of misclassifications suggests underlying issues in the model's feature recognition capabilities, potentially pointing to areas where model training could be improved.

These results underscore the delicate balance required between model accuracy and adversarial robustness. They compel the advancement of more sophisticated defense strategies to safeguard against both specific and non-specific adversarial threats. Moving forward, enhancing model resilience to adversarial perturbations will be critical in deploying secure and reliable AI systems in real-world applications. As this field evolves, continuous research and development are essential to stay ahead of adversarial techniques and ensure the security and reliability of machine learning models.

References

[1] Krizhevsky, A.& Nair, V. & Hinton, G. (2009) CIFAR-100 Dataset. Available at::https://www.cs.toronto.edu/~kriz/cifar.htmlhttps://www.cs.toronto.edu/~kriz/cifar.html.

[2] Goodfellow, I.J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples (FGSM). Available at: https://arxiv.org/abs/1412.6572.

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks (PGD). Available at: https://arxiv.org/abs/1706.06083.

[4] Carlini, N., & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks (CW Attack). Available at: https://arxiv.org/abs/1608.04644.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition (ResNet). Available at: https://arxiv.org/abs/1512.03385.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks (PreResNet). Available at: https://arxiv.org/abs/1603.05027.

[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks (SEResNet). Available at: https://arxiv.org/abs/1709.01507.

[8] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely Connected Convolutional Networks (DenseNet). Available at: https://arxiv.org/abs/1608.06993.

[9] Yu, F., Koltun, V., & Funkhouser, T. (2017). Dilated Residual Networks (DiaResNet). Available at: https://arxiv.org/abs/1705.09914.

[10] PyTorchCV. (n.d.) PyTorchCV: A PyTorch-Based Framework for Deep Learning in Computer Vision. Available at: https://pypi.org/project/pytorchcv/.

[11] Harry24k. (2023) TorchAttack: A Pytorch Repository for Adversarial Attacks. Available at: https://github.com/Harry24k/adversarial-attacks-pytorch .

[12] OpenAI. (2023) ChatGPT4. Available at: https://chat.openai.com.