SPML HW1 Gray-box Attack

r12921100 Hung Yang, Yeh

Abstract

In this study, we conduct adversarial attacks on 500 images from the CIFAR-100 dataset to generate adversarial examples and assess the robustness of source models against such attacks. Utilizing the TorchAttack package, we employ two prominent adversarial attack methods: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These techniques enable us to explore the vulnerabilities of deep learning models in recognizing perturbed images that, to the human eye, appear nearly identical to their originals. By comparing the accuracies of the source models before and after the attacks, we quantify the impact of different adversarial strategies on model performance. This investigation not only highlights the susceptibility of neural networks to adversarial examples but also serves as a benchmark for evaluating the resilience of various models against these sophisticated attacks. Additionally, we showcase several adversarial examples to illustrate the practical implications of our findings, emphasizing the need for developing more robust defense mechanisms in the field of computer vision.

1 Introduction

The CIFAR-100 task, a benchmark in the field of computer vision, involves classifying images into 100 distinct categories, posing a challenging problem for deep learning models due to its diverse and fine-grained nature. Adversarial examples, subtly modified images indistinguishable to the human eye but capable of misleading AI models, spotlight the vulnerabilities in these systems. The creation of these examples is paramount for understanding and enhancing model robustness. By scrutinizing how models falter against such manipulations, researchers can develop more secure and resilient AI systems. This endeavor is crucial in applications where reliability and security are paramount, including autonomous driving and facial recognition, thereby motivating the exploration of adversarial attacks in the CIFAR-100 context.

2 Methods

2.1 Choice of Source Models

To identify the most effective attack model for generating adversarial examples on the CIFAR-100 dataset, I conducted a comprehensive cross-testing analysis using five different models available in the pytorch-cifar-models repository, combined with three distinct adversarial attack techniques. After evaluating the outcomes across 15 experimental setups, RepVGG-A2 emerged as the superior source model. It consistently demonstrated a high attack success rate across different models, making it the optimal choice for generating adversarial examples. This selection was based on its robust performance under various attack conditions, highlighting RepVGG-A2's effectiveness in compromising model accuracy while maintaining the visual integrity of the perturbed images. Consequently, RepVGG-A2 was employed as the primary model for producing adversarial examples in this study.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2.2 Adversarial Attack Method

In the exploration of adversarial attack methods on images from the CIFAR-100 dataset, I applied three distinct techniques: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and a combined approach integrating both FGSM and PGD, with an epsilon value set to 8/256. This epsilon value quantifies the extent of the adversarial perturbation applied to the images, balancing between imperceptibility to human observers and effectiveness in misleading the model.

The combined attack method emerged as particularly potent when paired with a specific source model, demonstrating the capability to execute effective attacks across various models. This synergy between the combined attack approach and the selected source model underscores the nuanced interplay between attack methodologies and model vulnerabilities.

Given its superior performance in compromising model accuracy without significantly degrading the visual quality of the images, the combined attack method, leveraging both FGSM and PGD techniques, was chosen as the preferred strategy for generating adversarial examples in this study.

3 Experimental Design and Results

3.1 Get the labels

To effectively conduct adversarial attacks on a set of 500 images in Cifar-100, I undertook a meticulous process of cross-referencing these images against the 50,000 images in the CIFAR-100 training set. This step was crucial for accurately extracting the labels associated with each of the 500 provided images. With the correct labels at hand, it became possible to precisely measure the impact of adversarial attacks on the accuracy of models when subjected to these perturbed images.

This approach not only allowed for a targeted analysis of adversarial vulnerabilities specific to the selected subset of images but also ensured that the effectiveness of the attacks could be quantitatively assessed. By knowing the true labels of these images, we could evaluate the performance of the models in recognizing the adversarially modified content, thus providing a clear metric for the success rate of different adversarial attack methods. This methodological precision is essential for developing a deeper understanding of the resilience of machine learning models to adversarial manipulations and for designing more robust defenses against such attacks.

3.2 Find the best attack strategy

From Figure 1 and Table 1, we observe the accuracy of different source and target models under various attack scenarios. It is evident that the Fast Gradient Sign Method (FGSM) exhibits a more consistent performance across different models, indicating its robustness as a versatile adversarial attack method. On the other hand, the Projected Gradient Descent (PGD) attack demonstrates superior efficacy when tailored to specific models, surpassing FGSM in targeted scenarios. This suggests that while FGSM provides a general approach to compromising model accuracy, PGD offers a more potent strategy when precise vulnerabilities of a target model are known and exploited.

Regarding the choice of source model, RepVGG-A2 consistently shows commendable performance in all tested scenarios, highlighting its effectiveness in generating adversarial examples. This robustness makes RepVGG-A2 an optimal choice for adversarial attack experiments, as it reliably undermines the accuracy of various target models. The observed data underscore the importance of selecting an appropriate attack method and source model based on the specific context and objectives of the adversarial testing, thereby guiding strategies to improve model resilience against such attacks.

Table 1: Average accuracy across the four other models

Source Model	FGSM	PGD	Combined
ResNet56	18.05%	28.4%	27.6%
VGG19_bn	23.35%	25.05%	25.75%
MobileNetV2_x1_4	16.25%	23.75%	24.05%
ShuffleNetV2_x2_0	17.05%	13.55%	14.1%
RepVGG_A2	15.85%	12.7%	11.85%

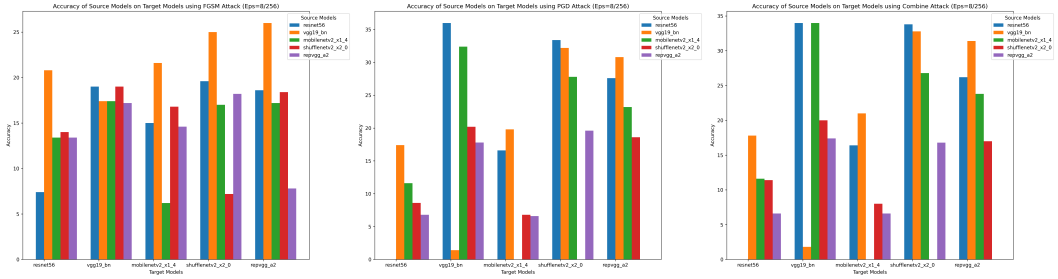

Figure 1: Accuracy on different model and attack

From Table 1, it is evident that the combined attack strategy, when executed with RepVGG-A2 as the source model, stands out as the most effective approach, achieving an average accuracy of 11.85% across the four other models. This remarkable efficacy underscores the potency of this particular attack configuration in compromising the integrity of various target models. Consequently, this method was selected for generating adversarial examples, further exploring its impact on model robustness at different levels of perturbation intensity, denoted by epsilon. Figure 2 illustrates the variation in model accuracy as a function of epsilon, providing insights into the resilience of the models against increasing adversarial perturbations. This analysis is crucial for understanding the threshold beyond which model performance significantly deteriorates, thereby informing the development of more robust defense mechanisms against such sophisticated adversarial attacks.

3.3 FGSM vs PGD

In Figure 2 and Table 2, which showcases the accuracy of generating adversarial examples at various levels of epsilon, reveals that the accuracy of the combined attack is largely in line with that of the PGD attack. Notably, when attacking the dataset's 500 images, the accuracy approaches near zero (0.04%) at an epsilon of 4/256. This observation underscores the efficacy of white-box attacks in achieving significant disruption with relatively small perturbations, compared to black-box attacks.

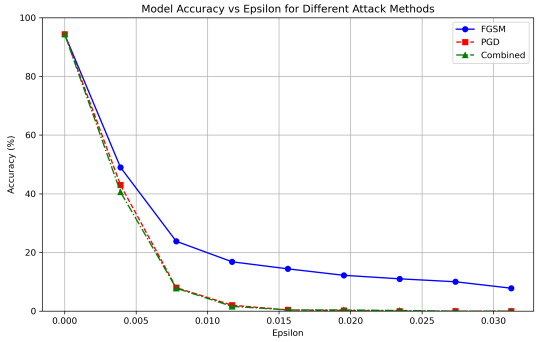

Figure 2: Accuracy of Different Attack Methods at Various Epsilon Values

Table 2: Accuracy of Different Attack Methods at Various Epsilon Values

Epsilon	Combined Attack	FGSM Attack	PGD Attack
0/256	94.4%	94.4%	94.4%
1/256	40.6%	49.0%	43.0%
2/256	7.8%	23.8%	8.0%
3/256	1.6%	16.8%	2.0%
4/256	0.4%	14.4%	0.4%
5/256	0.4%	12.2%	0.2%
6/256	0.2%	11.0%	0.0%
7/256	0.0%	10.0%	0.0%
8/256	0.0%	7.8%	0.0%

3.4 Results

In the results section, after generating 500 adversarial examples, it was observed that at an epsilon value of 8/256, the images still maintained a semblance recognizable to humans. This indicates that the perturbations were subtle enough not to distort the images beyond human recognition, yet they were effective enough to significantly compromise the targeted models. Such a balance is crucial in adversarial machine learning, as it highlights the potential vulnerabilities in models that can be exploited without making drastic changes to the input data. This efficacy underscores the importance of developing robust defenses against adversarial attacks, as even slight perturbations can lead to significant degradation in model performance.

Figure 3 showcases some of these adversarial examples, illustrating the subtle changes that lead to misclassifications by the models. These examples serve as a tangible demonstration of the concept of adversarial attacks, where the goal is to alter an image in a way that is almost imperceptible to humans but causes the model to fail in its predictions. This visual representation helps in understanding the practical implications of adversarial attacks and the need for models that are resilient to such manipulations.

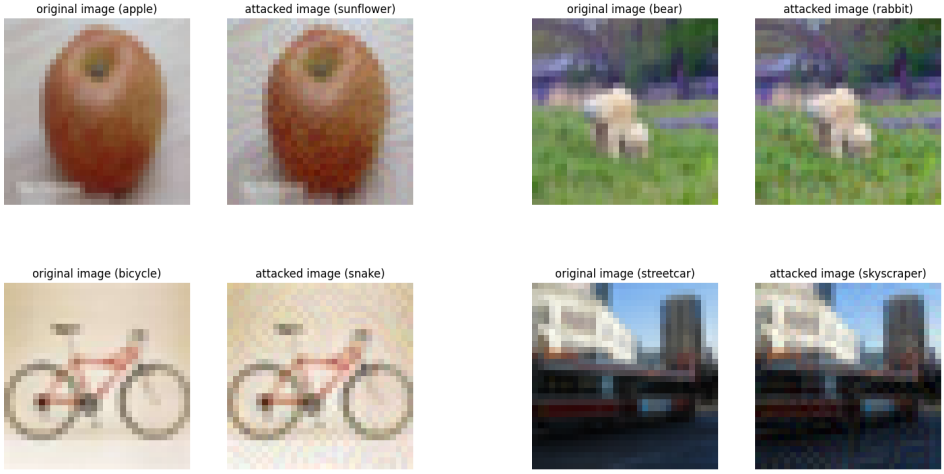

Figure 3: Adversarial examples

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] PyTorch. (n.d.) Adversarial Example Generation. Available at: https://pytorch.org/tutorials/beginner/fgsm_tutorial.html .

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

[4] chenyaofo. (2021) PyTorch CIFAR Models. Available at: https://github.com/chenyaofo/pytorch-cifar-models.

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