# Additive backdoor attack on CIFAR-10 classifier

Course Instructor:
**Professor Goerge Kesidis**

Student:
**Samira Malek**

Pennsylvania State University

# Contents


**1 Introduction** **2**

**2 Create Perturbation** **2**

**3 Training** **2**

**4 Results and Discussion** **3**

**5 References** **5**

# 1 | Introduction

In the realm of machine learning and particularly within the scope of deep learning, the robustness and security of neural networks are of paramount importance. This work delves into the intricacies of crafting and deploying backdoor attacks on deep neural networks, with a focus on a nuanced form of perturbation known as sparse pixel-wise perturbation.
Our investigation centers on the deployment of these perturbations within the context of image classification tasks, utilizing the ResNet-18 architecture and the widely recognized CIFAR-10 dataset. Through a series of meticulously designed experiments, we aim to understand the impact of varying the intensity and distribution of these perturbations—represented by changes in the $L_2$ norm and the poisoning rate—on the model's accuracy and susceptibility to attack.
The subsequent sections detail the methodology behind the creation of the perturbations, the experimental setup, and the findings from our investigation.

# 2 | Create Perturbation

In the experiments that follow, a specific backdoor pattern is employed, noted for its sparse pixel-wise perturbation. This approach affects only a select few pixels, involving the generation of perturbations by initially randomly picking four pixels. For images with color, a random selection is made to perturb one of the three RGB channels for each of these pixels. The perturbations introduced are consistently positive, ensuring a uniform perturbation magnitude across the chosen pixels. To maintain this uniformity, a predetermined "reference" perturbation magnitude is used. Then, for each pixel, this reference magnitude is adjusted by a factor randomly drawn from a Gaussian distribution, which has a mean of 1 and a standard deviation of 0.05 [Miller et al., 2023]. Fig. 2.1 illustrates the pixel-wise perturbation with an $L_2$ norm of 0.4 and an offset of 0.5 (for visualization purposes) employed in subsequent experiments. Fig. 2.2 depicts an image poisoned with this backdoor technique.

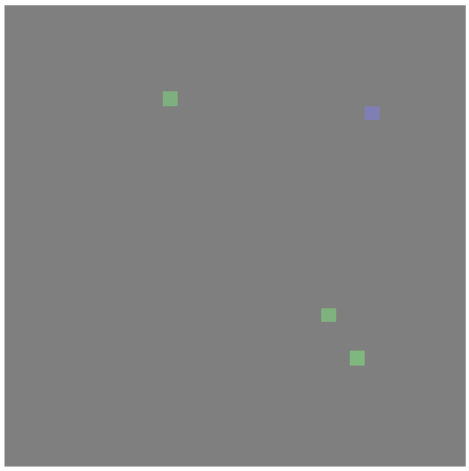

**Figure 2.1:** A pixel-wise perturbation with $l_2$ norm 0.4 and a 0.5 offset

**Figure 2.2:** Poisoned image with pixel-wise perturbation represented in Fig. 2.1

# 3 | Training

We conducted our experiments using the ResNet-18 deep neural network and the CIFAR-10 dataset. Specifically, we targeted class 8 of CIFAR-10 (ship images) for poisoning, relabeling the poisoned images as class 4. The network underwent training for 70 epochs, with the cumulative training loss for each epoch illustrated in Fig. 3.1. The perturbation employed has an $L_2$ norm of 0.4, and the poisoning rate was set at 0.6% (300 images of class 8). Fig. 3.2 displays the network's Accuracy (ACC) and Attack Success Rate (ASR) across epochs. Notably, from epoch 50 to 70, the network achieved a stable performance. Consequently, we computed the final ACC and ASR of the network as the average values from epoch 50 to 70.

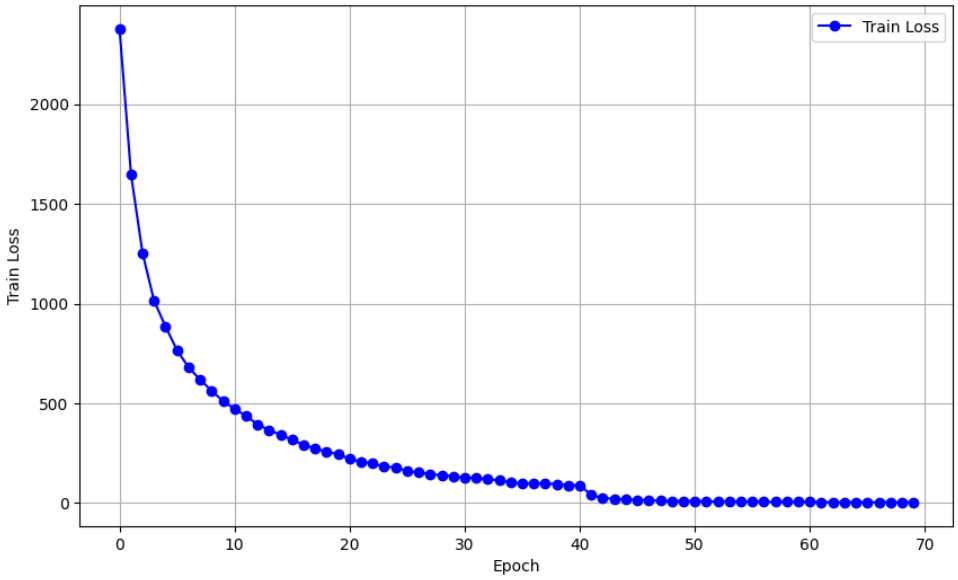


**Figure 3.1:** Training loss in each epoch

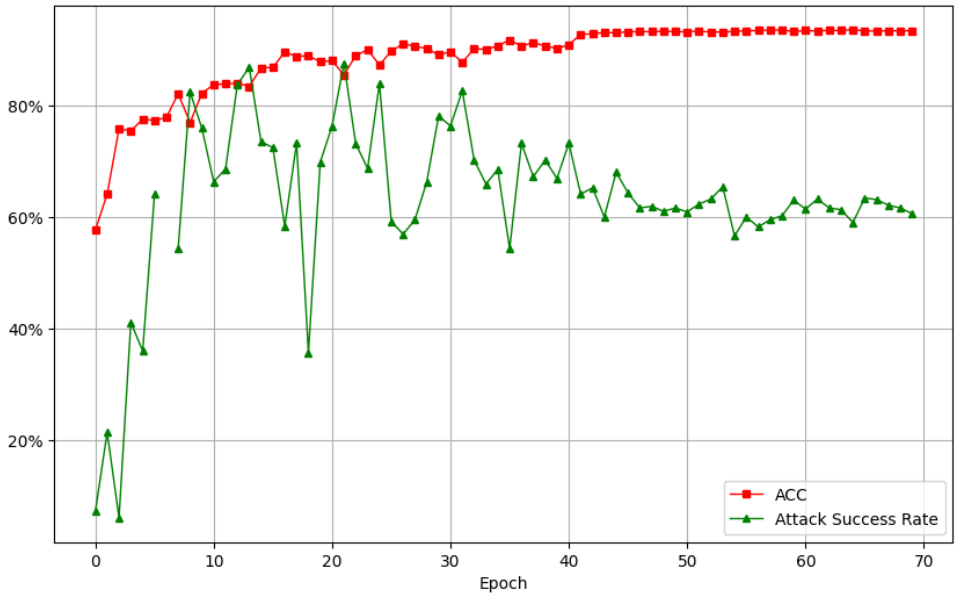


**Figure 3.2:** Accuracy and Attack Success Rate in each epoch

# 4 | Results and Discussion

We adopted a setup analogous to that described in Section 3 for the experiments in this section. Employing a constant pixel-wise perturbation with an $L_2$ norm of 0.4, we varied the poisoning rate from 0.1% to 1% and evaluated the Accuracy (ACC) and Attack Success Rate (ASR). The outcomes, presented in Fig. 4.1, align with our expectations: an increase in the poisoning rate results in a corresponding rise in ASR without a reduction in ACC.

In a separate experiment, our goal was to demonstrate the impact of the $L_2$ norm of the perturbation on ACC and ASR. Fig. 4.2 illustrates the ACC and ASR at a poisoning rate of 0.6%, with the $L_2$ norm varying from 0.2 to 1. As anticipated, the ASR escalates with an increase in the $L_2$ norm, while the ACC remains stable on clear images.

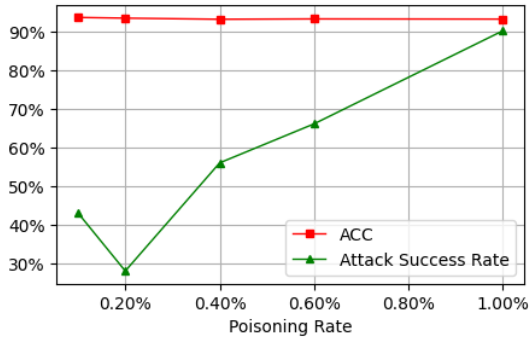


**Figure 4.1:** Accuracy and Attack Success Rate across poisoning rates ranging from 0.2% to 1%, with perturbation employing an $L_2$ norm of 0.4.

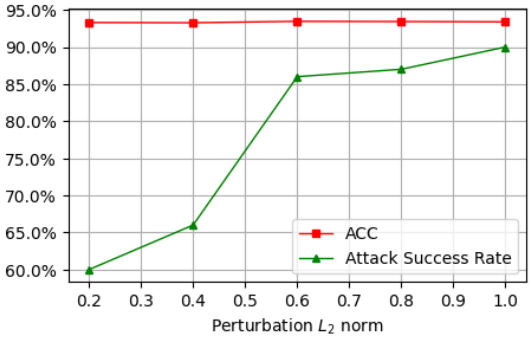


**Figure 4.2:** Accuracy and Attack Success Rate across perturbation employing an $L_2$ norm ranging from 0.2 to 1, with 0.6% poisoning rates

# 5 | References

[Miller et al., 2023] Miller, D. J., Xiang, Z., and Kesidis, G. (2023). *Adversarial Learning and Secure AI*. Cambridge University Press.