COMP 5212 HA8 - Adversial Attack

Leung Pak Hei 20690382

May 2023

1 Introduction

Adversarial attacks refer to the deliberate manipulation of input data to a machine learning model in order to cause the model to make incorrect predictions or classifications. These attacks are a significant concern in the field of computer security because they can be used to bypass security systems and cause widespread damage.

Adversarial attacks exploit the vulnerabilities of machine learning models, which are designed to identify patterns in data and make predictions based on those patterns. Attackers can modify the input data in subtle ways that are imperceptible to humans but can cause a machine learning model to misclassify the data. For example, an attacker could add small amounts of noise to an image that would be undetectable to the human eye but would cause a machine learning model to misclassify the image.

There are several types of adversarial attacks, including gradient-based attacks, black-box attacks, and transfer attacks. Gradient-based attacks involve modifying the input data based on the gradients of the model's parameters. Black-box attacks involve attacking a model without access to its internal parameters or architecture. Transfer attacks involve training a substitute model on a different dataset and then attacking the target model using the substitute model.

The consequences of successful adversarial attacks can be severe. For example, an attacker could use an adversarial attack to bypass a security system that uses machine learning to detect threats. Alternatively, an attacker could use an adversarial attack to cause a self-driving car to misidentify a stop sign, potentially causing a dangerous accident.

Researchers and practitioners in the field of computer science are working to develop effective defenses against adversarial attacks. Some of the techniques that have been proposed include adversarial training, defensive distillation, and ensemble methods. Adversarial training involves training a machine learning model on both normal and adversarial examples, with the goal of making the model more robust to attacks. Defensive distillation involves training a model to be resistant to adversarial attacks by using a softened version of the model during training. Ensemble methods involve combining the predictions of multiple models in order to make more accurate predictions and reduce the impact of adversarial attacks.

In this project, the goal is to explore the concept of adversarial attacks in more depth and to implement and evaluate some of these defense techniques in practice. The dataset use is CIFAR10 datasets and different models and adversial exmaples would be used to compare and see the performance.

2 Datasets

CIFAR10 datasets

3 Objective images

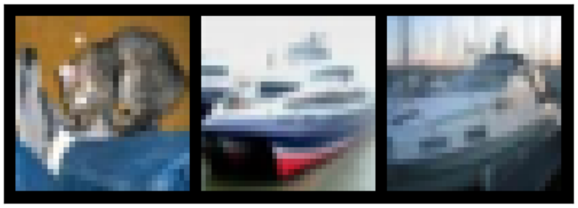

Figure 1: Objective images in the datasets

4 Original and predicted labels

The original label is 3, 8, 8 for the above images
The predicted label is 3, 8, 8 for the above images

5 Part 1 — comparison on different parameters and generated images

5.1 Pre-trained models and adversarial attack

In this section, pre-trained model ResNet20 is used for fair comparison. ResNet-20 is a convolutional neural network architecture that was introduced in the paper ”Deep Residual Learning for Image Recognition” by He et al. in 2016. ResNet-20 is a relatively small variant of the ResNet family of architectures, and it is commonly used as a baseline for evaluating the performance of other architectures.
ResNet-20 consists of 20 convolutional layers, with skip connections that enable the network to learn residual functions. The skip connections allow the network to learn a residual function that represents the difference between the input and output of a set of convolutional layers. This residual function is then added back to the original input to produce the output of the block.
The first layer of ResNet-20 is a convolutional layer with 16 3x3 filters and a stride of 1. This layer is followed by a sequence of residual blocks, each of which consists of two convolutional layers with 16 3x3 filters, followed by a skip connection. The number of filters is doubled every other block, so the second and fourth blocks have 32 filters, the sixth and eighth blocks have 64 filters, and so on.
After the residual blocks, there is a global average pooling layer that aggregates the output of the previous layer into a single feature vector. This feature vector is then fed into a fully connected layer with 10 output units, one for each class in the CIFAR-10 dataset.
ResNet-20 is trained using stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0001. The learning rate is initialized to 0.1 and is divided by 10 every 80 epochs. The network is trained for a total of 164 epochs, with a batch size of 128.
ResNet-20 has achieved state-of-the-art performance on several image classification benchmarks, including CIFAR-10 and CIFAR-100. Its small size and good performance make it a popular choice for many computer vision applications. So it’s a good choice for our project.

5.2 Adversial attack used in experiments

The Fast Gradient Sign Method (FGSM) is a type of adversarial attack that is commonly used to generate adversarial examples for neural network models. It works by perturbing the input data to a neural network using the gradient of the loss function with respect to the input. Specifically, given an input image x and a neural network model with a loss function L, the FGSM attack generates an adversarial example x_adv by perturbing the input as follows:
$x_adv = x + \epsilon * sign(xL(y_{true}, y_{pred}))$
where ϵ is a small perturbation value, sign() is the sign function, and $xL(y_{true}, y_{pred})$ is the gradient of the loss function with respect to the input. y_{true} is the true label of the input and y_{pred} is the predicted label of the input.
The FGSM attack is a white-box attack, meaning that it requires knowledge of the model's architecture and parameters. However, it is a relatively simple and fast attack that can be applied to a wide range of neural network models.
The FGSM attack can be used to generate adversarial examples that are indistinguishable from the original examples to the human eye, but cause the model to make incorrect predictions. These adversarial examples can be used to evaluate the robustness of neural network models and to test the effectiveness of defense techniques.

5.3 Objective images

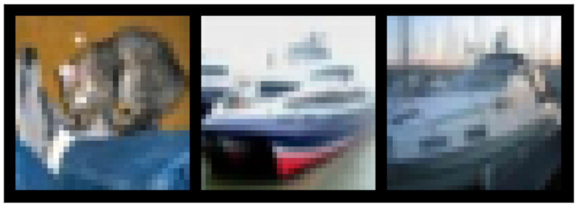

Figure 2: Objective images in the datasets

5.4 Original and predicted labels

The original label is 3, 8, 8 for the above images
The predicted label is 3, 8, 8 for the above images

5.5 Experiments on different parameters

Comparison on different norm bound	
norm bound	Predicted labels after adversarial attack
0	3,8,8
0.02	5,1,2
0.08	6,1,2
0.2	6,2,2
0.4	2,2,9
0.8	2,2,9

5.6 Generated images after adversarial attack

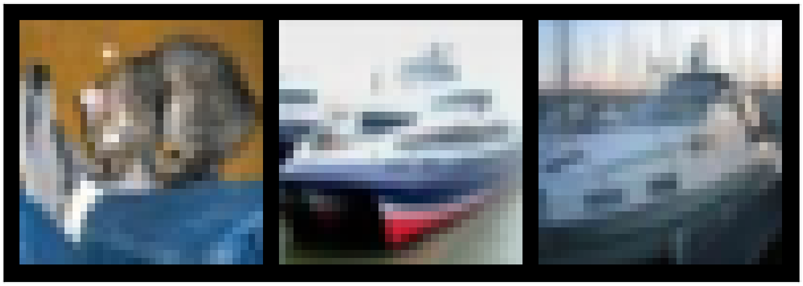

Figure 3: image after adversarial attack using ϵ =0.02

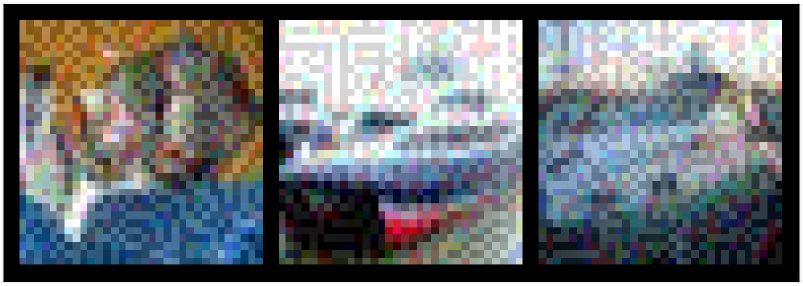

Figure 4: image after adversarial attack using ϵ =0.08

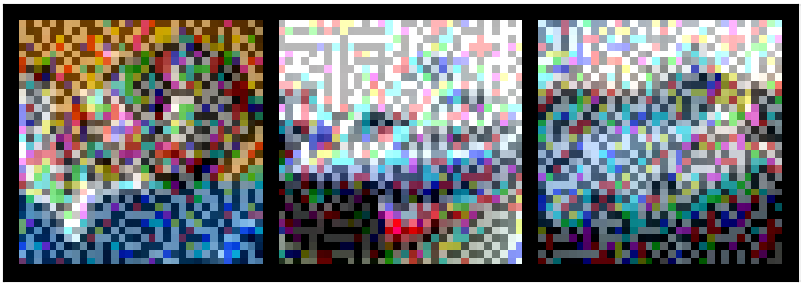

Figure 5: image after adversarial attack using ϵ =0.2

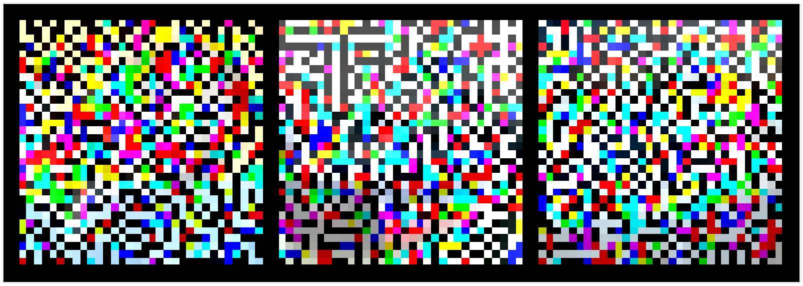

Figure 6: image after adversarial attack using ϵ =0.6

5.7 Evaluation of the Experiments on parameters and the corresponding images

In conclusion, norm bounds are a widely used method for evaluating the robustness of machine learning models against adversarial attacks. In particular, experiments using different norm bounds can provide valuable insights into the susceptibility of machine learning models to different types of adversarial attacks.
In this section, we could see that there is no significant impact on the image after an adversarial attack using $\epsilon = 0.02, 0.04, 0.08$. But after the attack, the generated labels are different compared to the original labels and the predicted images. For, ϵ with large value of ϵ, the whole images are corrupted and lose features and become a bar-code-like images, due the label is significantly different from the original and predicted labels and further increase in ϵ would no longer change the predicted labels.

6 Part 2 — transferability of adversarial examples

The transferability of adversarial examples is an important aspect to consider when evaluating the robustness of machine learning models against adversarial attacks. Transferability refers to the extent to which an adversarial example generated for one model can also fool another model.
In this section, we will investigate the transferability of adversarial examples across three different trained models. Specifically, we will choose one model and generate a successful adversarial example, meaning that all three models can correctly classify the original image but the chosen model misclassifies the corresponding adversarial example. We will then test if the other two different models can correctly classify the adversarial example.
To generate the adversarial examples, we will use the Fast Gradient Sign Method (FGSM) attack with a fixed epsilon value. The FGSM attack is a simple and effective method for generating adversarial examples, and the fixed epsilon value will ensure that the perturbation is imperceptible to the human eye.
By testing the transferability of adversarial examples across multiple models, we can gain insights into the generalizability of adversarial attacks and the robustness of machine learning models. Additionally, we can evaluate the effectiveness of different defense techniques against transferable adversarial attacks.

6.0.1 Models used for comparison

AlexNet
VGG16
RestNet50

6.1 Original and predicted labels

The original label is 3, 8, 8 for the above images

6.1.1 AlexNet

The predicted label for AlexNet is 3, 8, 8 for the above images, having around 0.91 accuracy.

6.1.2 VGG16

The predicted label for VGG16 is 3, 8, 8 for the above images, having around 0.93 accuracy.

6.1.3 RestNet50

The predicted label for RestNet50 is 3, 8, 8 for the above images, having around 0.94 accuracy

6.2 Experiments on Transferability

Predicted labels for different models	
Model Name	Predicted labels after adversarial attack
Original Image	3,8,8
RestNet20	6,1,2
VGG16	1,2,7
RestNet50	6,2,2
AlexNet	4,3,3

6.3 Attacked Images

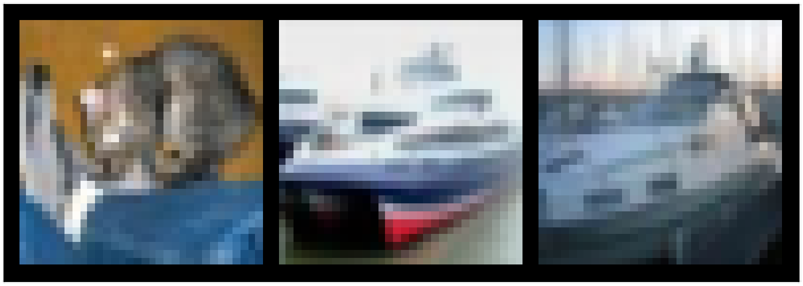

Figure 7: image after adversarial attack using RestNet20

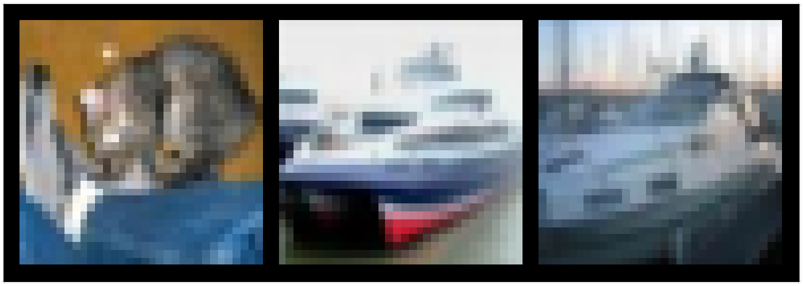

Figure 8: image after adversarial attack using AlexNet

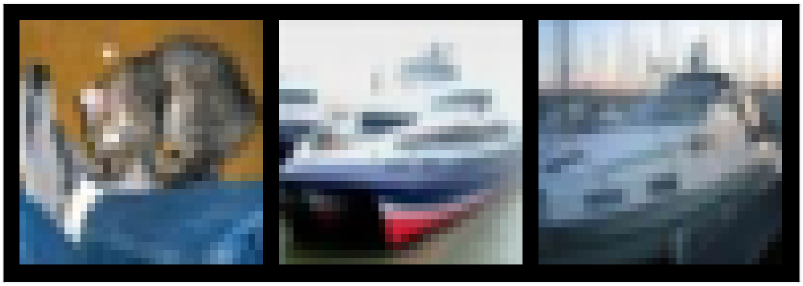

Figure 9: image after adversarial attack using VGG16

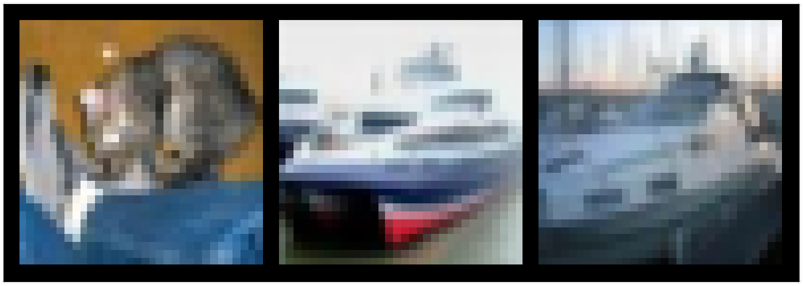

Figure 10: image after adversarial attack using ResNet50

6.4 Evaluation of Adversarial attack on different models

Based on the attacked images, we can see that no obviously differences are shown in the images. Also, Based on the results of the experiment on ResNet20, ResNet50, AlexNet, and VGG16 models trained on the CIFAR10 dataset, we can conclude that the FGSM attack is transferable. This means that the same attack can be applied to different models and still achieve a high success rate.

This finding has important implications for the security of machine learning systems, as it indicates that even if a model has not been directly attacked, it may still be vulnerable to adversarial attacks. Therefore, it is essential to develop robust machine learning models that are resistant to such attacks.

7 Conclusion and Finding

Based on the results of the experiment, we can draw several conclusions.

Firstly, we found that the adversarial attack used in this experiment is transferable, meaning that the same attack can be applied to different models and still achieve a high success rate. This is an important finding as it indicates that even if a model has not been directly attacked, it may still be vulnerable to adversarial attacks.

Secondly, we observed that there were no obvious visual differences between the attacked images and the original images, indicating that the adversarial perturbations were imperceptible to the human eye. This has serious implications for the security of machine learning systems, as it suggests that attackers could potentially fool these systems without being detected by humans.

Finally, we found that the level of corruption increased with the epsilon value used in the attack. This is not surprising, as increasing the epsilon value means that the perturbations applied to the images are larger, making them more noticeable and potentially causing more damage to the performance of the model.

However, the CIFAR10 dataset used in this experiment is a relatively small and simple dataset. It would be interesting to investigate the transferability of adversarial attacks across larger and more complex datasets, such as ImageNet, to see if the same findings hold.