**南京信息工程大学 人工智能导论** **实验(实习)报告**

实验(实习)名称 对抗样本 实验(实习)日期 2024.12.13 得分 指导教师 应龙 学院 计算机与网络空间安全学院专业 计算机科学与技术 年级 2022 班次 4 姓名张瑞晨 学号 202283290159

1. **实验目的**
2. **了解对抗样本的生成方法**：通过实现并分析FGSM、PGD、C&W等对抗样本生成算法，学习它们的工作原理及其在实际应用中的效果。
3. **评估对抗样本对模型性能的影响**：通过在已经训练好的基线模型上测试对抗样本，评估其对模型性能的影响，特别是扰动大小（epsilon）对模型准确率的变化。
4. **探索Targeted Attack**：在实验中实施针对特定目标的攻击（例如，将数字9误分类为数字3），了解如何控制对抗样本使模型产生特定错误。
5. **进行模型防御的初步探索**：通过实验，理解对抗攻击如何影响模型的分类能力，并为未来对抗样本的防御策略奠定基础。
6. **实验内容**

使用对抗样本生成方法FGSM 进行 Non-Targeted Attack, 生成一批对抗样本；在已经训练好的基线模型上测试对抗样本，记录模型的性能下降情况；观察

选取不同扰动大小（如FGSM方法中的epsilon）对模型性能的影响；使用可

视化工具（如matplotlib）绘制扰动大小-精准度曲线；回答问题： FGSM 方法产生对抗样本时为什么要在梯度加上符号函数而不直

接使用梯度？进行 Targeted Attack：对于手写数字9图像施加扰动，使得分类网络错误地将它识别为数字3，扰动图像数字9可以直接通过对手写体数字9添加较小的噪声生成，也可以直接从随机噪声中生成一个类似于9的图像；

实现一种更复杂的对抗样本生成方法（如PGD, C&W, DeepFool, GAN等方

法）生成对抗样本。

#### ****1.FGSM方法的公式****

Fast Gradient Sign Method (FGSM) 是一种基于梯度的对抗攻击方法，其基本思想是通过计算损失函数相对于输入的梯度，然后根据梯度的符号进行扰动。通过这种方法生成的对抗样本在视觉上与原始样本几乎相同，但对于模型来说，它们却能导致错误分类。

梯度计算：首先，我们计算损失函数对输入的梯度，这个梯度指示了如何调整输入才能最大化损失（即使得模型错误分类）。

符号函数：然后，通过取梯度的符号（sign），确保每个像素的扰动方向是沿着损失函数上升最快的方向，这样的扰动能够更有效地误导模型。

扰动大小（epsilon）：ϵ\epsilonϵ 控制着扰动的幅度，较大的epsilon会产生更大的扰动，从而可能导致模型误分类，但过大的扰动会使得对抗样本变得不自然，容易被人眼察觉。

1. **对于基线实验，列出训练的相关信息，以及在测试集上的相关性能评估；**

**基线模型训练信息**

在本实验中，我们使用了一个简单的卷积神经网络（CNN）作为基线模型。模型结构如下：

第一层：卷积层（32个5x5的卷积核）

第二层：卷积层（64个5x5的卷积核）

第三层：全连接层（128个节点）

输出层：全连接层（10个节点，对应10个类别）

训练参数：

优化器：SGD（随机梯度下降）

学习率：0.01

动量：0.9

批次大小：64

训练轮数：5

损失函数：交叉熵损失（CrossEntropyLoss）

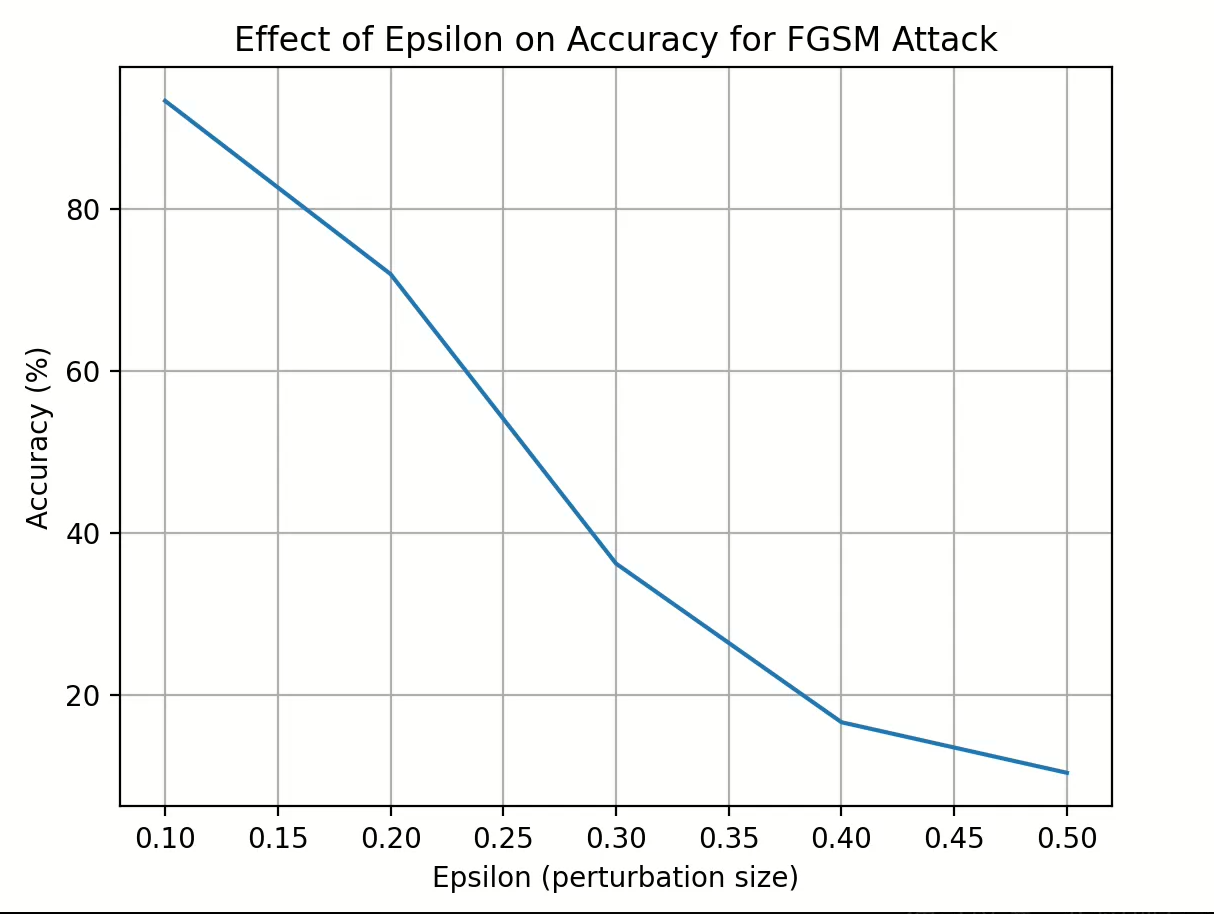
训练过程：

训练过程中记录每一轮的损失值，优化器通过反向传播调整权重，直到达到预定的轮次。

每一轮的训练后，模型在训练集上进行评估，计算损失和准确率。

测试集性能评估

模型在测试集上的评估包括准确率、精确率、召回率等指标。我们使用模型预测测试集中的样本，并与实际标签进行比较。



**数值评估**：通过不同epsilon值生成对抗样本，测试模型的准确率下降情况。

通过绘制准确率与epsilon值的关系图，展示扰动大小对模型性能的影响。

Finished Training

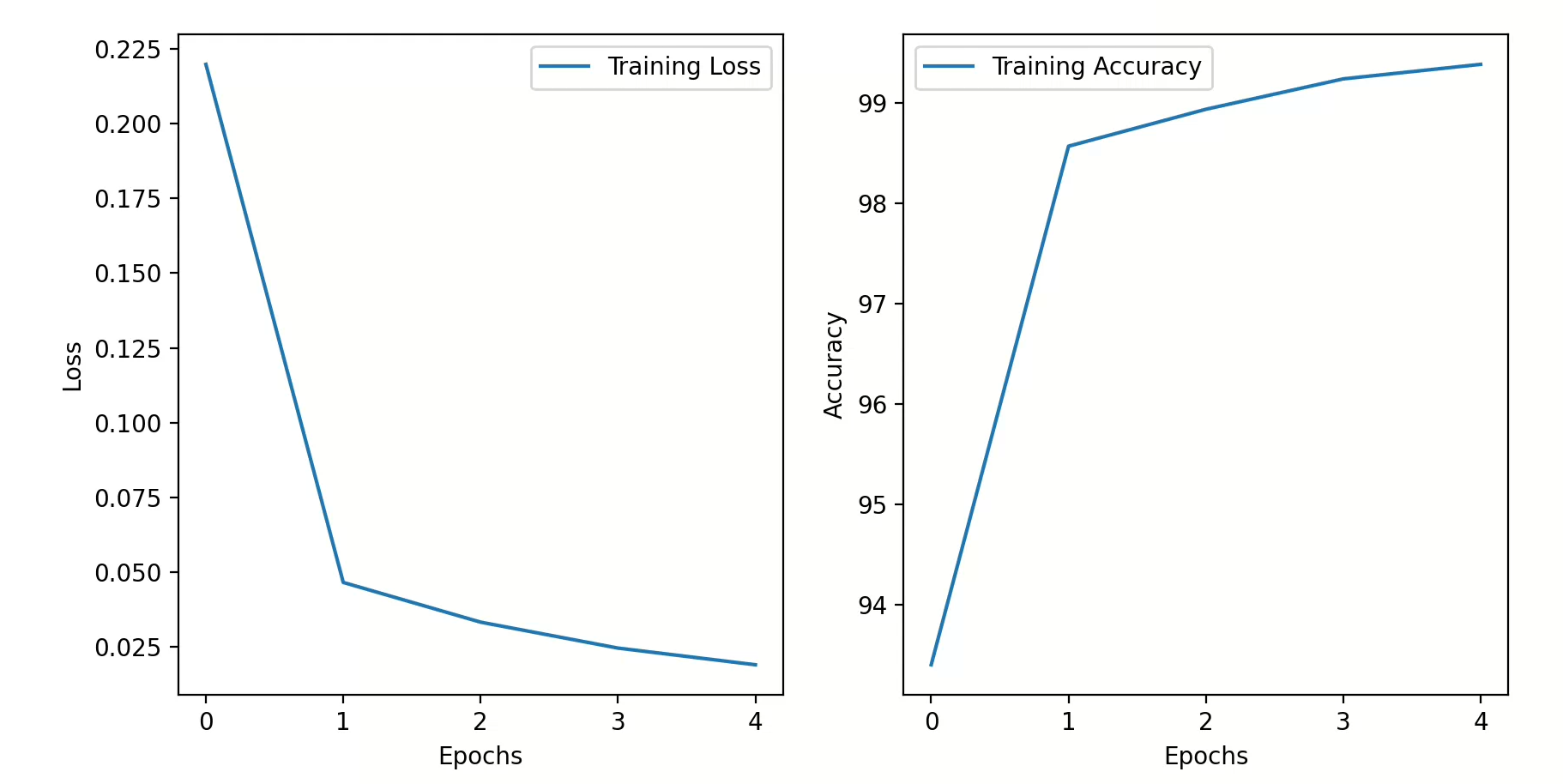
Accuracy on adversarial examples with epsilon=0.1: 93.40%

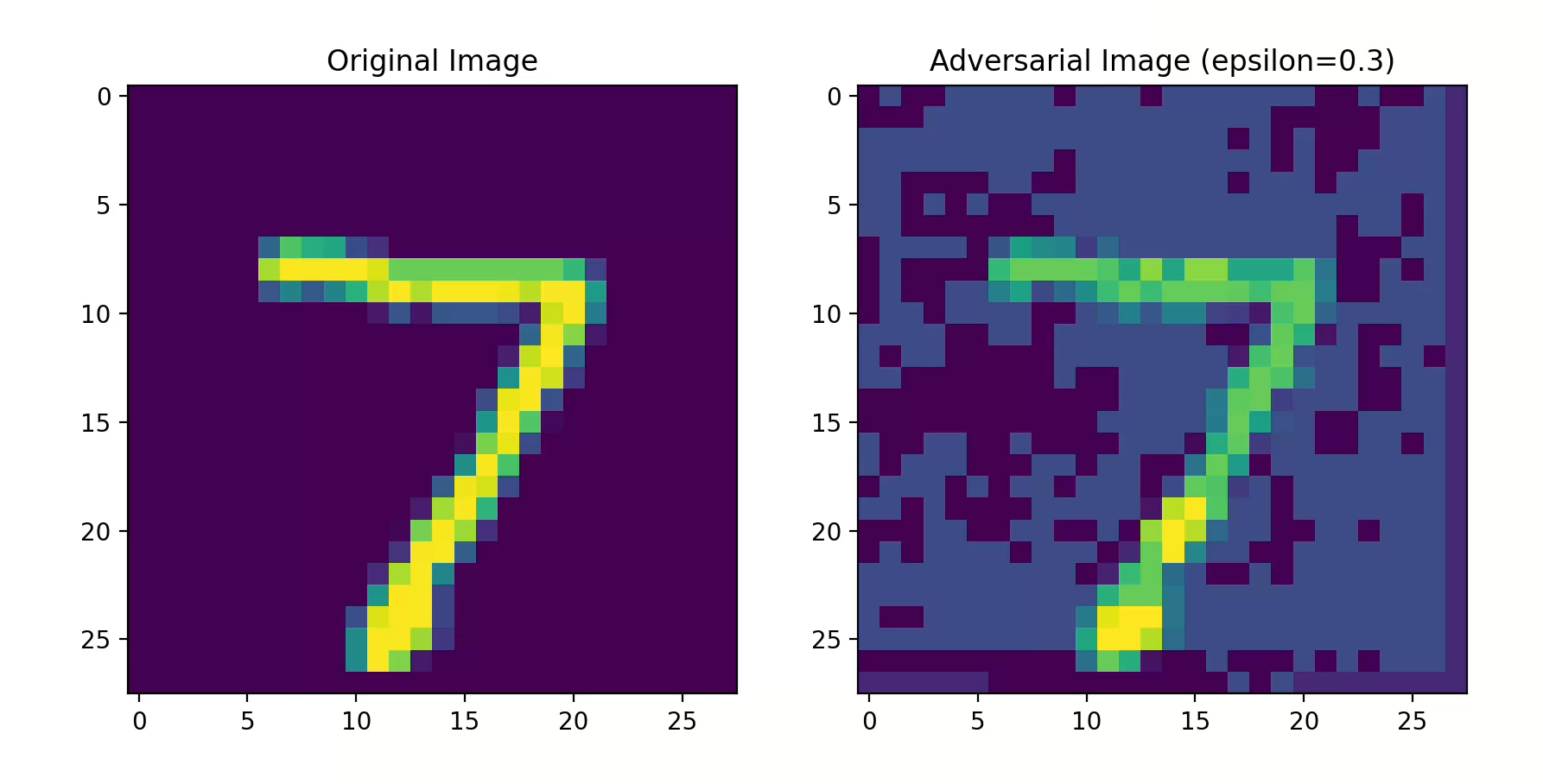
Accuracy on adversarial examples with epsilon=0.2: 71.98%

Accuracy on adversarial examples with epsilon=0.3: 36.19%

Accuracy on adversarial examples with epsilon=0.4: 16.57%

Accuracy on adversarial examples with epsilon=0.5: 10.32%





**三、实验心得**

**对抗样本的强大威胁**  
在本实验中，我们观察到对抗样本可以显著影响模型的分类准确性。通过改变扰动大小（epsilon），我们发现模型的准确率随着epsilon的增加而急剧下降，尤其是在使用FGSM方法生成的对抗样本下。这表明，即使是微小的扰动，也能让模型做出错误预测，这揭示了当前深度学习模型在面对对抗攻击时的脆弱性。

**FGSM方法的有效性与局限性**  
FGSM（快速梯度符号法）是最基础的对抗样本生成方法，通过计算梯度并沿其符号方向对输入图像进行扰动。实验中发现，虽然FGSM在生成对抗样本时简单而高效，但它的攻击效果通常受到扰动大小的限制。更复杂的攻击方法（如PGD、C&W）通常能生成更加“强大”的对抗样本，造成更大的影响。

**对抗攻击的定量与定性分析**  
实验中，我们不仅通过数值评估（准确率下降情况）来展示对抗样本的影响，还通过可视化图像和损失曲线进一步分析其效果。通过绘制扰动大小与准确率之间的关系图，我们清楚地看到了模型准确率随着扰动的增大而下降，这为理解对抗攻击对深度学习模型的威胁提供了直观的证据。

**Targeted Attack的挑战与潜力**  
在Targeted Attack的实验中，我们尝试让模型将数字9错误地分类为数字3。尽管这一任务相对较简单，但成功实施Targeted Attack需要在目标类别的方向上施加足够的扰动。通过这一实验，我们了解了如何通过调整损失函数和梯度计算，控制对抗样本的生成，从而实现特定的攻击目标。

**防御策略的初步思考**  
尽管本实验的重点是对抗攻击，但通过对抗样本的生成与模型性能下降的观察，我们也获得了一些关于防御的初步思考。例如，可能需要开发更强大的防御策略，如对抗训练、梯度遮蔽等方法，以提高模型对抗对抗攻击的鲁棒性。