基于随机噪声添加的对抗防御实验报告

一、背景介绍

深度学习模型在图像识别等领域取得了显著的成果，但同时也面临着对抗样本的威胁。对抗样本是通过对原始图像进行微小的、人类难以察觉的扰动而生成的，这些样本能够使深度学习模型产生错误的预测结果。本实验旨在针对之前通过 FGSM（快速梯度符号法）生成的对抗样本（存储在 fgsm_images 文件夹中），探索一种简单的防御机制 —— 添加随机噪声，以提高预训练的 ResNet50 模型对对抗样本的鲁棒性，从而保障模型在实际应用中的可靠性和安全性，尤其是在面对潜在的恶意攻击时能够保持相对稳定的性能表现。

二、选择方法

本实验选择的防御方法是向对抗样本中添加随机噪声。这种方法的原理基于噪声的随机性可能会干扰对抗样本中精心设计的扰动模式，从而使模型能够更准确地识别图像的真实类别。通过向输入的对抗样本添加一定强度的随机高斯噪声，期望模型能够忽略掉对抗扰动带来的误导信息，而是关注图像的真实特征，进而提高对对抗样本的防御能力。虽然这是一种相对简单的防御策略，但具有易于实现和计算成本低的优点，能够为更复杂的防御方法提供一个基础的参考和对比。

三、实验过程

文件介绍

mk_datasets.categorial_data.py 用于将从官网下载得到的 ImageNet 验证集转换为可以通过 datasets.ImageFolder 读取的形式；

resnet50.py 使用 ResNet50 模型在 ILSVRC2012_img_val 上进行测试。

adversarial_attack.py 使用 FGSM 方法对模型进行攻击，构建对抗攻击样本。

test_attack.py 使用对抗攻击样本 fgsm_images 对 ResNet50 进行测试。

fgsm_images 文件夹保存了生成的对抗攻击样本。

logs 文件夹保存了训练过程中的日志。

adversarial_defense.py 使用对抗防御的方法后进行测试。

环境搭建与模型准备

本实验基于 Python 语言环境，使用 PyTorch 深度学习框架进行模型的加载和操作。

数据加载与预处理

使用 torchvision.datasets.ImageFolder 类从 fgsm_images 文件夹中加载对抗样本数据集，并通过定义的 transform 操作对数据进行预处理。

随机噪声添加函数实现

定义了 add_random_noise 函数，用于向输入的图像张量添加随机噪声。该函数接受两个参数：images 表示输入的图像张量，noise_factor 表示噪声的强度因子，默认值为 0.03。在函数内部，首先使用 torch.randn_like 函数生成与输入图像张量具有相同形状的随机噪声张量，其元素服从标准正态分布。然后，将生成的噪声张量乘以 noise_factor 来调整噪声的强度，使其与图像数据的范围相适应。最后，将噪声添加到原始图像张量上，并使用 torch.clamp 函数将像素值限制在 0 到 1 的范围内，确保处理后的图像数据仍然是合法的图像张量表示，避免出现像素值超出合理范围的情况。

准确率计算与结果记录

调用 calculate_accuracy_with_defense 函数计算模型在经过随机噪声防御后的对抗样本上的准确率，并将最终的准确率以格式化的形式打印输出到控制台，同时记录到日志文件 accuracy_defense.log 中。日志文件中记录了每个批次的准确率信息以及最终的总体准确率，这些信息对于分析模型在不同批次数据上的表现以及评估防御机制的有效性提供了详细的数据支持。

四、实验结果分析

准确率数据统计

在原始的 ImageNet 验证集上，如图 1，ResNet50 模型具有较高的准确率 75.32%。如图 2，当使用生成的对抗样本对模型进行测试时，准确率显著下降至 4.92%。

如图 3，从控制台输出和日志文件记录的准确率数据来看，在应用随机噪声

防御机制后，模型对对抗样本的准确率有了一定程度的提高，提高至 28.43%。与未应用防御机制的情况相比，能够正确识别的对抗样本数量有所增加，这表明添加随机噪声在一定程度上干扰了对抗样本中的扰动信息，使得模型能够更准确地判断图像的类别，从而提高了对对抗样本的防御能力。

```
2024-12-28 17:51:17,315 - INFO - batch 779, accuracy: 75.38%
2024-12-28 17:51:18,123 - INFO - batch 780, accuracy: 75.37%
2024-12-28 17:51:18,860 - INFO - batch 781, accuracy: 75.33%
2024-12-28 17:51:19,046 - INFO - batch 782, accuracy: 75.32%
2024-12-28 17:51:19,046 - INFO - The accuracy of ResNet50 on the ImageNet validation set is: 75.32%
```

图 1 ResNet50 在 ImageNet 验证集上的准确率测试

```
2024-12-28 20:16:51,895 - INFO - batch 435, accuracy: 4.89%
2024-12-28 20:16:52,531 - INFO - batch 436, accuracy: 4.93%
2024-12-28 20:16:53,190 - INFO - batch 437, accuracy: 4.93%
2024-12-28 20:16:53,539 - INFO - batch 438, accuracy: 4.92%
2024-12-28 20:16:53,545 - INFO - Accuracy of ResNet50 after attack on adversarial examples: 4.92%
```

图 2 ResNet50 在对抗攻击样本上的准确率测试

```
2024-12-28 20:24:29,979 - INFO - batch 435, accuracy: 28.44%
2024-12-28 20:24:30,547 - INFO - batch 436, accuracy: 28.44%
2024-12-28 20:24:31,159 - INFO - batch 437, accuracy: 28.44%
2024-12-28 20:24:31,486 - INFO - batch 438, accuracy: 28.43%
2024-12-28 20:24:31,492 - INFO - Accuracy of ResNet50 after defense on adversarial examples: 28.43%
```

图 3 ResNet50 在对抗攻击样本上并使用随机噪声防御的准确率测试

虽然添加随机噪声的防御方法在一定程度上提高了模型对对抗样本的准确率，但整体的准确率仍然相对较低，这说明该方法虽然简单易行，但防御效果有限，无法完全抵御 FGSM 生成的对抗样本的攻击。

噪声强度影响分析

通过调整 add_random_noise 函数中的 noise_factor 参数，可以进一步分析噪声强度对防御效果的影响。初步实验结果表明，随着噪声强度的增加，模型对对抗样本的准确率呈现先上升后下降的趋势。当噪声强度过小时，无法有效地干扰对抗样本的扰动信息，防御效果不明显；而当噪声强度过大时，虽然可能会进一步干扰对抗扰动，但同时也会破坏图像的原始信息，导致模型难以准确识别图像的类别，从而使准确率下降。因此，选择合适的噪声强度是优化该防御方法的一个关键因素，本实验采取的 noise_factor 参数设置为 0.03。

五、总结

本实验通过向对抗样本中添加随机噪声的方法，对预训练的 ResNet50 模型进行了对抗防御的探索。实验结果表明，这种简单的防御方法在一定程度上能够

提高模型对对抗样本的准确率，但整体防御效果有限，无法完全应对复杂的对抗攻击。