

基于 FSGM 对抗样本构建实验报告

一、背景介绍

随着深度学习在图像识别等领域的广泛应用，其安全性受到了越来越多的关注。对抗样本作为一种能够欺骗深度学习模型的特殊输入，成为了研究的热点。ImageNet 是一个大规模的图像识别数据集，被广泛用于预训练各种深度学习模型，如 ResNet。研究如何针对 ImageNet 验证集构建对抗样本，对于评估模型的鲁棒性以及深入理解深度学习模型的行为具有重要意义。

二、选择方法

本实验选择快速梯度符号法（FGSM）来构建对抗样本。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模型对输入图像的梯度，并根据梯度的符号来调整图像的像素值，使得模型的预测结果发生改变，从而生成对抗样本。这种方法简单高效，能够快速地生成对抗样本，并且在一定程度上反映了模型的脆弱性。

三、实验过程

文件介绍

`mk_datasets.categorical_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。

环境搭建

使用 Python 作为编程语言，安装了 PyTorch 深度学习框架以及相关的库，如 torchvision 用于加载预训练模型和处理图像数据。

数据准备

从官方网站下载了 ImageNet 验证集 (ILSVRC2012_img_val)，并按照分类目录结构进行整理 (ILSVRC2012_img_val_categories)，整理代码位于 mk_datasets 文件夹中。使用 torchvision 中的 datasets.ImageFolder 类加载验证集数据，并通过 transforms.Compose 定义了数据转换操作，包括图像的大小调整、裁剪、归一化等，使其符合 ResNet 模型的输入要求。

模型加载

从 torchvision.models 中加载预训练的 ResNet50 模型，并将其设置为评估模式，在 resnet50.py 文件中，使用 ResNet50 模型在 ILSVRC2012_img_val 上进行测试，

对抗样本生成

对于验证集中的每一张图像，首先将其输入到 ResNet 模型中，计算模型的输出和损失函数。然后，通过反向传播计算损失函数对输入图像的梯度，并根据 FGSM 的原理，在梯度的方向上对图像的像素值进行微小的调整，生成对抗样本。具体实现中，设置了一个扰动参数 $\epsilon = 0.01$ ，用于控制对抗样本与原始图像之间的差异程度。



保存对抗样本

将生成的对抗样本保存到文件夹 fgsm_images 中，以便后续的实验分析。

四、实验结果分析

可视化分析

通过可视化原始图像和对应的对抗样本，可以发现对抗样本在人眼看来与原始图像几乎没有区别，但却能够成功地欺骗 ResNet 模型，使其做出错误的预测。这表明 FGSM 生成的对抗样本具有很强的隐蔽性和攻击性。

	
图 1 ImageNet 验证集中的图片	图 2 FGSM 攻击后的图片

准确率分析

在原始的 ImageNet 验证集上，如图 3，ResNet50 模型具有较高的准确率 75.32%。如图 4，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%。这说明 FGSM 攻击方法能够有效地降低 ResNet 模型在对抗样本上的性能，揭示了模型在面对对抗攻击时的脆弱性。

```
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 3 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 4 ResNet50 在对抗攻击样本上的准确率测试

参数敏感性分析

通过调整 FGSM 中的扰动参数 ϵ ，可以发现随着 ϵ 的增大，对抗样本与原始图像之间的差异逐渐增大，同时模型在对抗样本上的准确率也进一步下降。这表明扰动参数的大小对于对抗样本的有效性具有重要影响，较大的扰动能够更容易地欺骗模型，但也可能导致对抗样本的隐蔽性降低。

五、总结

本实验通过使用 FGSM 方法针对 ImageNet 验证集构建了 ResNet 模型的对抗样本，并对实验结果进行了详细的分析。结果表明，FGSM 能够有效地生成对抗样本，使得 ResNet 模型的准确率大幅下降，揭示了深度学习模型在面对对抗攻击时的脆弱性。