华中科技大学

计算机视觉实验四报告

姓　　名：

专　　业：

班　　级：

学　　号：

指导教师：

分数	
教师签名	

年月日

目录

实验四：深度神经网络后门攻击

1.1 任务要求

对实验二构建的 CIFAR-10 数据集分类神经网络进行训练数据集污染，实现后门攻击。

污染训练数据时，要对除 airplane 的剩余 9 类选取固定比例 R 的样本进行污染，即在每一类中，受污染的样本个数/该类的总样本数=R。测试不同的比例 R，对攻击成功率的影响。

1.2 实验内容

本实验中，我使用在实验二中构建的 vgg13_bn 网络实现后门攻击，网络实现代码见 modules/vgg.py 文件。实验结果见 exp4_backdoor_attacks.ipynb 文件，实验具体步骤如下：

1.2.1 数据污染

对训练数据集中除 airplane 这一类别之外的其他九类数据进行污染，即添加后门攻击的触发开关，并将所有污染后的训练样本标记为攻击目标，即 airplane。数据污染的实现代码见 tools/data_poison.py 文件。

本实验中，我在被污染图片的右下角植入4×4的白色方块，作为后门触发条件。以固定比例 r 污染训练数据集后，可视化部分样本，结果如下：

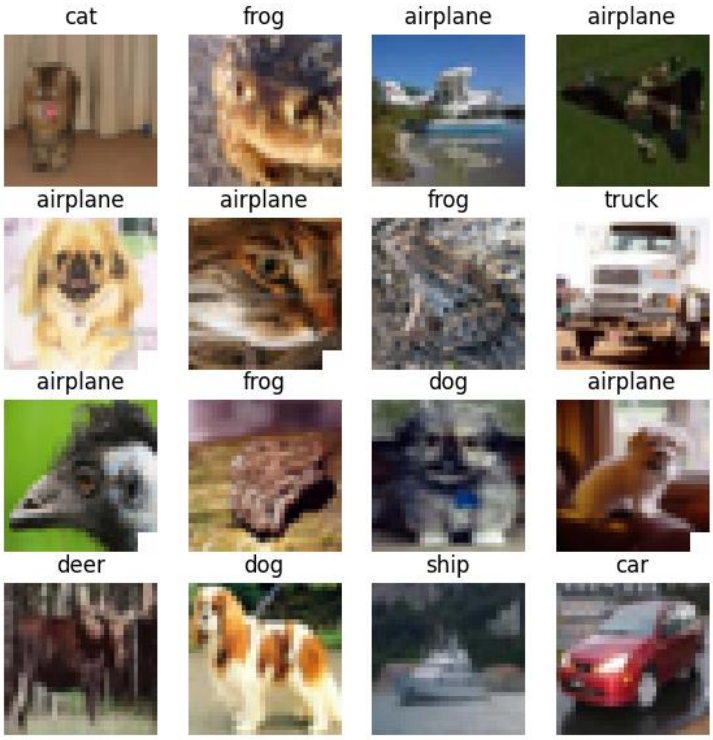

图 1-1：后门触发条件示意图

由上图可以清晰地看到，在图片右下角植入4 × 4白色方块后，图片类别变为airplane，而未被污染的数据仍然保持原本正确的分类标签，所以污染训练数据集成功。

1.2.2 训练神经网络

联合原始数据集以及受到污染的九类数据样本，对 CIFAR-10 分类神经网络——vgg13_bn 进行训练。

我按照实验要求，设置了不同的比例 R（R=0.0, 0.1, 0.2, ……, 1.0），共 11 个不同的值，分别进行训练。训练日志和保存的模型权重文件在此链接中（实验四训练日志和权重文件）下载，由于这些文件较大，所以没有包含在上交的代码文件中。

对于不同的比例 R，训练配置都相同，具体如下：

1、训练集数据分割为 train/val，其中验证集 val 有 1000 个样本，剩余样本都属于训练集 train，batch_size 为 128，共训练 20 个 epoch；

2、使用 Adam 优化器，学习率为 1e-3，weight_decay 为 1e-4；

3、学习率调整策略为 ReduceLROnPlateau，当模型准确率不再提高时，将学习率减小 1/2；

训练过程中发现：污染比例 R 越高，训练数据集越大，训练时间越长。

1.2.3 测试

在受到后门攻击的神经网络上，测试十类干净数据（即使用原始测试数据）的分类准确率 CDA（Clean Data Accuracy），及九类植入后门触发开关的测试数据的攻击成功率 ASR（Attack Success Rate）。

首先，分别加载原始测试数据集和九类植入后门触发开关的测试数据，可视化九类植入后门触发开关的测试数据，以确保成功植入后门触发开关，如下：

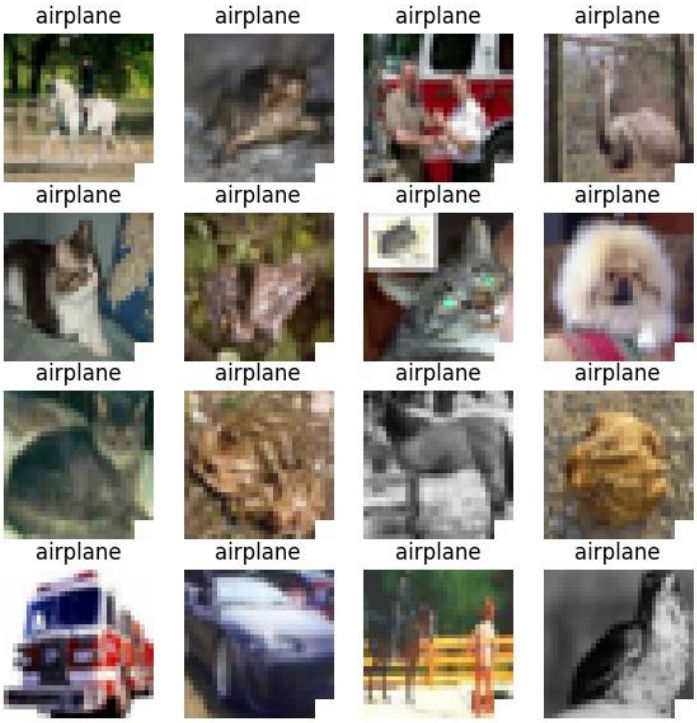

图 1-2：可视化九类植入后门触发开关的测试数据

可见正确植入后门触发开关。

然后开始测试 vgg13_bn 在不同污染比例 R 下的 CDA 和 ASR，结果如下：

```
ratio:0.0  cda:89.11%  asr:1.52%
ratio:0.1  cda:88.47%  asr:97.91%
ratio:0.2  cda:88.40%  asr:98.63%
ratio:0.3  cda:88.18%  asr:98.68%
ratio:0.4  cda:87.94%  asr:98.34%
ratio:0.5  cda:87.83%  asr:99.72%
ratio:0.6  cda:88.07%  asr:98.78%
ratio:0.7  cda:87.77%  asr:99.43%
ratio:0.8  cda:87.73%  asr:99.04%
ratio:0.9  cda:87.34%  asr:99.28%
ratio:1.0  cda:87.53%  asr:99.87%
```

图 1-3：测试结果

根据上图中结果，画出横坐标为 R，纵坐标为干净数据分类正确率和后门攻击成功率的折线图，如下：

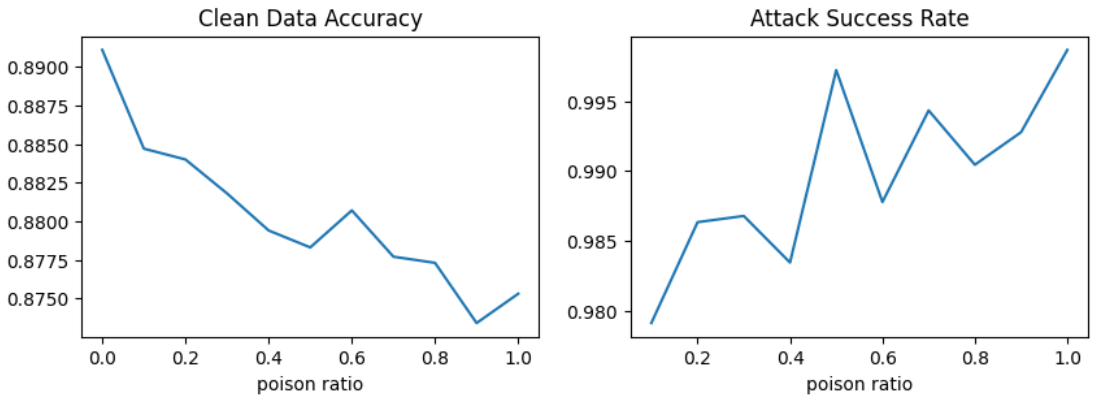

图 1-4：干净数据分类正确率和后门攻击成功率的折线图

由上图分析可知：攻击成功率很高，接近 100%。随着污染比例的上升，干净数据分类正确率呈下降趋势，攻击成功率呈上升趋势。

攻击成功率随着污染比例上升而上升在意料之中，也比较容易理解。但是干净数据分类正确率却随着污染比例上升而下降，我觉得可能原因是：干净数据中一些非 airplane 的图片右下角本来就是或接近白色，所以会被模型误认为是后门触发开关，而随着污染比例上升，攻击成功率上升，即模型识别后门触发开关的能力越强，导致那些右下角本来就是或接近白色的样本更容易被错误分类，因此干净数据分类正确率下降。