华中科技大学

计算机视觉实验报告

姓　　名：王逸

学　　院：计算机科学与技术

专　　业：计算机科学与技术

班　　级：CS2011

学　　号：U202014774

指导教师：刘康

分数	
教师签名	

2023 年　4 月　14　日

目　录

实验四 深度神经网络的后门攻击

1.1 实验要求

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。

1.1.1 报告要求

1. 污染训练数据集时，应该对除了 airplane(label=0)的其他九类数据集按照固定比例 R 进行污染。联合原始数据集以及受到污染的九类数据样本对 CIFAR-10 分类神经网络进行训练。

2. 画出横坐标为 R，纵坐标为后门攻击成功率(ASR)的折线图，R 的范围自定。

3. 实验报告包含网络设计、实验结果图，以及必要的分析等。

1.2 实验内容

本实验是在实验三的基础上添加两个部分：训练集污染、攻击成功率测试。

A. 训练集污染

本实验探究了两种训练集污染方式：一种是通过直接修改图片 data 值以达到标记效果；另一种则参考 WaNET，通过使图片发生扭曲来达到中毒效果。两种方式的代码部分如下：

a.添加矩形标记

```
def add_rec(img_arr):
    img_arr[:,:4,:4] = 0
    return img_arr
```

a. 图片扭曲：设计一个扭曲函数 M，使图片以特定形式扭曲

a) 选择 control grid：将图片分成 kxk 的 size，使用参数 s=0.5 定义 P 的强度 $P = \psi(\text{rand}[-1,1](k, k, 2)) \times s$

```
ins = torch.rand(1, 2, 4, 4)*2-1
ins = ins/torch.mean(torch.abs(ins))
```

b) 上采样：将 control grid 上采样放大至 32x32x2，能够覆盖输入图片大小

```
noise_grid = (
    F.interpolate(ins, size = 32, mode="bicubic", align_corners=True)
    .permute(0, 2, 3, 1)
    .to(DEVICE)
)
```

c) 裁剪：使采样点都落在图片范围内

```
grid_tmp = (identity_grid+0.5*noise_grid/32)*1
grid_tmp = torch.clamp(grid_tmp, -1, 1)
```

d) warping：逻辑变形

```
poison_sample = F.grid_sample(img_data.permute(0,3,1,2).to(DEVICE),
grid_tmp.repeat(num_bd, 1, 1, 1), align_corners=True)
```

为了方便后续更改污染攻击方式，统一在加载数据集之后，被封装成 Dataloader 之前进行污染。为了结果输出维度准确，我们需要在将图片数据传入 grid_sample 之前以及结果输出时进行维度变换，这里采用 permute 函数来实现。

两种方式的处理结果如下：

具体代码请参照./data.py 中的 add_rec()和 warping()函数，以及 poison_data()函数。

B. 攻击成功率测试

在测试完干净数据集后，将测试数据集进行污染，得到污染后的九类数据集以及标签，对此进行测试得到攻击成功率。完整代码见./main.py 最后部分。

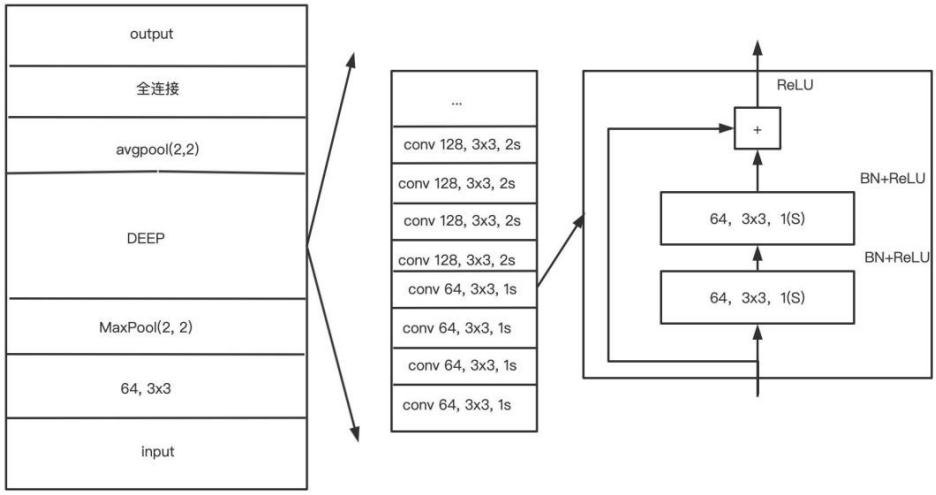
```
poison_sample, poison_label = poison_data(test_dataset, poison_rate, 0, test_flag=True)
print(poison_
test_dataset.
test_dataset.targets = poison_label
test_loader = DataLoader(test_dataset, shuffle=True, batch_size=BATCH_SIZE)
correct, acc_back, avg_feats, feats = test(model, test_loader)
print(acc_back)
```

(variable) targets: Tuple[Any, Any] | list

C. 训练

训练部分较上次并未作出太大改变，本报告中不过多赘述。

本次实验中所采用的网络结构依旧是 resNet18，具体结构如下图所示。



D. 画图分析

本次实验需要画出 ASR-R 的折线图：

一是模型攻击成功率随着污染比例 R 的变化曲线。该部分与实验一中类似，代码可见文件 draw_fig.py。

1.3 实验结果分析

a. 矩形标签攻击

编号	神经网络种类	学习率	迭代次数	R
1	ResNet	0.01	20	0.1
2		0.01	20	0.2
3		0.01	20	0.3
4		0.01	20	0.4

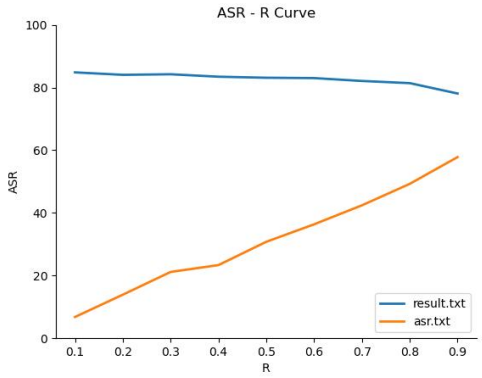
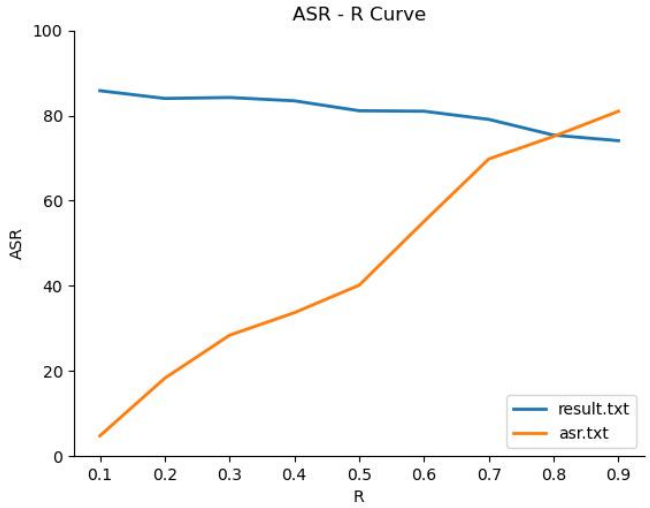


图 1 lr=0.01, epoch=40,LeNET

b. 图片扭曲:

编号	神经网络种类	学习率	迭代次数	R
1	ResNet	0.01	20	0.1
2		0.01	20	0.2
3		0.01	20	0.3
4		0.01	20	0.4



通过以上两组网络的实验，可以看出：

- 对于比较简单的后门攻击方式，模型在干净数据集上的预测准确率实际上是没有太大变化的；而对于肉眼难以区分的攻击方式来说，模型准确率会随着 R 的增大而逐渐下降；
- 而对于 ASR 来说，大体上 ASR 是随着 R 的增大而增大的，对于不同的攻击方式，攻击成功率的大小以及增加幅度都有所不同。

1.4 实验小结

本次实验过程中比较麻烦的是在什么时候进行对数据集的污染，以及该如何处理数据集。由于为了代码编写的方便，这次是在加载数据集之后立马进行数据集的污染，在简单的修改数据的处理上还好，但要进行扭曲扰动时就会出现很多维度上的问题。这也需要我们根据函数的参数维度以及输出结果维度进行不断的调整。