深度学习模型版权保护平台

组长：张卓萌

小组成员：章杭炜 李佳露 罗书卿 蔡锶维 朱文骏 张昊

测试

算法测试

白盒

白盒 CV 算法

我以CV中的图像分类模型为例进行水印加注和性能测试，选取 cifar10 数据集和 ResNet 网络结构训练了一个十分类模型。我们对这个模型添加白盒水印，实验结果如下：

1. 模型训练

model	dataset	epochs	test acc
ResNet	Cifar10	40	88.98 %

2. 水印嵌入

model	verify module	epochs	test acc	trigger acc
ResNet	fully connected neural network	200	88.98 %	100 %

3. 用其他身份进行验证

model	verify module	test acc	trigger acc
ResNet	fully connected neural network	88.98 %	54 %

4. 模型微调攻击(以额外训练n个epoch作为微调)

model	verify module	extra epochs	test acc	trigger acc
ResNet	fully connected neural network	1	85.81 %	99.5 %
ResNet	fully connected neural network	2	87.1 %	98.5 %
ResNet	fully connected neural network	3	87.73 %	99 %
ResNet	fully connected neural network	4	87.92 %	97.5 %

5. 神经元剪枝攻击(全局非结构化剪枝)

model	verify module	prune amount	test acc	trigger acc
ResNet	fully connected neural network	0.3	88.24 %	100 %
ResNet	fully connected neural network	0.4	87.45 %	99.5 %
ResNet	fully connected neural network	0.5	81.08 %	97.5 %
ResNet	fully connected neural network	0.55	69.72 %	99 %

另外，对于水印复写攻击，因为白盒水印不对模型做修改，所以复写之后原本的水印仍然会被检测有效。

验证模块的定义如下：

```
class Verify(torch.nn.Module):
    def __init__(self, h1=2000, h2=100):
        super(Verify, self).__init__()
        self.n_input = 141632
        self.sigmoid = torch.nn.Sigmoid()
        self.fc_verify1 = torch.nn.Linear(self.n_input,h1)
        self.fc_verify2 = torch.nn.Linear(h1,h2)
        self.fc_verify3 = torch.nn.Linear(h2,2)
    def forward(self, x):
        x = self.fc_verify1(x)
        x = self.sigmoid(x)
        x = self.fc_verify2(x)
        x = self.sigmoid(x)
        x = self.fc_verify3(x)
        return x
```

黑盒

黑盒图像分类

1. 无水印模型训练

model	dataset	epochs	test acc	backdoor acc
resnet18	cifar10	20	83.11%	12%

2. 水印嵌入

model	dataset	epochs	test acc	backdoor acc
resnet18	cifar10	30	82.77%	100%

3. AE穿透性：80%

黑盒人脸识别

1. 无水印模型训练

model	dataset	epochs	test acc（lfw）	backdoor acc
resnet	WebFace	30	97.92%	6.00%

2. 水印嵌入

model	dataset	epochs	test acc（lfw）	backdoor acc
resnet	WebFace	30	97.75%	90.00%

3. 常见攻击

AE piracy	随机裁剪	AE + 裁剪
88%	79%	74%

黑盒 NLP 算法

1. 无水印模型训练

model	dataset	epochs	test acc	backdoor acc
LSTM	IMDB	15	98.9%	57.7%

2. 水印嵌入（以论文方式为载体附加身份信息）

model	dataset	epochs	test acc	backdoor acc
LSTM	IMDB	15	95.8%	99.4%

3. 用其他身份进行验证（不同特征向量）

model	dataset	epochs	test acc	backdoor acc
LSTM	IMDB	15	95.8%	66.7%

4. 模型微调攻击

model	dataset	epochs	test acc	backdoor acc
LSTM	IMDB	15	94.8%	96.6%

5. 神经元剪枝攻击

model	dataset	epochs	test acc	backdoor acc
LSTM	IMDB	15	84.7%	68.8%

可见对于嵌入的后门对元模型任务的影响很小，且后门对于微调方式下后门的识别具有很好的鲁棒性，对于剪枝操作，由于对于原模型任务准确率的影响过大，不做考虑。并且水印算法对于不同用户的水印具有较好的区分效果。

平台测试

平台的测试伴随着整个开发过程，以下为平台测试内容概述，具体结果见演示视频。

界面功能测试

- 界面显示
- 路由跳转

数据交互测试

- 表单提交（文本类型）
- 文件上传下载（文件类型）

算法融合测试

- 数据生成
- 模型裁决

整体流程测试

- 用户注册
- 用户登录
- 模型注册申请
- 申请完成
- 模型裁决申请
- 记录查看