人工智能安全课程期末大作业要求

可选作业主题

1. 后门攻防
2. 对抗攻防

后门攻防实验要求

一、 实验概述

复现或自己提出一种后门攻击或防御方案，并在目标模型和目标数据集上完成相应测试

二、 实验设置

1. 目标模型选择：ResNet50, VGG16等（不限制模型选择，根据模型被攻击成功的难易程度设置）
2. 目标数据集选择： MNIST, CIFAR10, CIFAR100, MINI-imagenet（按攻击的难易顺序由易到难排序）
3. 指标要求：后门攻击部分脏数据比例不得超过10%
4. 优化要求：优化trigger，提高攻击的隐蔽性；提高目标模型的复杂度

二、 复现实验参考

以下三个主题供大作业选题参考：

1. （后门攻击）BadNets：Identifying Vulnerabilities in the Machine Learning Model Supply Chain

   论文链接：https://arxiv.org/pdf/1708.06733.pdf

   代码参考链接：https://github.com/verazuo/badnets-pytorch

   https://github.com/ShihaoZhaoZSH/BadNet

2. （后门攻击）Reflection Backdoor：A Natural Backdoor Attack on Deep Neural Networks.

   论文链接：https://arxiv.org/pdf/2007.02343.pdf

   代码参考链接：https://github.com/DreamtaleCore/Refool

3. （后门攻击防御）Neural Cleanse：Identifying and Mitigating Backdoor Attacks in Neural Networks.

   论文链接：https://gangw.web.illinois.edu/class/cs598/papers/sp19-poisoning-backdoor.pdf

   代码参考链接：https://github.com/bolunwang/backdoor

对抗攻防实验要求

一、实验概述

实现一种对抗攻击或者防御方案（复现论文、改进综合已有的攻击或防御、自己设计全新的攻击或防御），并在目标模型和数据集上完成相应的实验测试。

二、实验设置

1. 目标模型选择：ResNet50、VGG16等（不限制模型的选择）
2. 目标数据集选择：MNIST、CIFAR10、CIFAR100、MINI-ImageNet、ImageNet等（不限制数据集的选择）
3. 至少在一个目标模型和一个数据集上完成实验
4. 实验环境：推荐使用Pytorch（Tensorflow也可）
5. 扰动：对抗攻击的扰动大小在合理范围之内（图片人眼可以辨别）
6. 攻击场景：数字空间/真实世界的对抗样本攻击

三、可选方向

1. 对抗白盒攻击

指标：生成的对抗样本对目标模型进行白盒攻击的成功率

参考论文：Towards Deep Learning Models Resistant to Adversarial Attacks 代码

2. 对抗黑盒攻击（对抗样本的迁移性提高）

指标：生成的对抗样本对目标模型进行黑盒攻击的成功率/对多个不同目标模型的攻击迁移性

参考论文：Improving Transferability of Adversarial Examples with Input Diversity 代码

3. 对抗攻击防御

指标：对经过防御后的目标模型进行对抗攻击成功率是否下降

参考论文：Towards Deep Learning Models Resistant to Adversarial Attacks 代码

Feature Denoising for Improving Adversarial Robustness 代码

实验提交

1. 完成一份实验报告，并提交代码，打包命名为"学号 _ 姓名 _ 大作业"。

   实验报告要求包含：实验设计、关键实验代码分析、实验结果分析、实验总结与思

   考代码需要包含的内容：关键模块的实现代码或者完整的工程代码

2. 提交方式：打包发送到助教邮箱

   周二上课的同学发送到：xuhuiyu@zju.edu.cn

   周四上课的同学发送到：zy.wu@zju.edu.cn

3. 提交截止时间：2022年6月19日21：30前