可信机器学习 HW3 Report

1 Backdoor Attack

Sub-Task 1. Generate Adversarial Example (for Clean Label Attack)

使用 PGD 生成对抗样本, 但在每一次得到 δ 的梯度后, 只修改对抗样本预测类型与 label 相同的数据对应的 δ.

Sub-Task 2. Add Triggers to Data

Step 1. Build pattern. 对于 BadNets, pattern 为位于图像右下角的 3x3 patch; 对于 Clean-Label, pattern 为图像四角的 3x3 patches; 对于 Blend, pattern 为与图像大小相等的高斯噪声.

Step 2. Add trigger. 计算 $\text{PoisonedImage} = (1-\text{mask})*\text{Image}+\text{mask}*((1-\alpha)\text{Image}+\alpha)\text{Pattern}$, 对于 Blend, mask 取为全 1.

Results

Method	ASR(%)	ACC(%)
BadNets	100.00%	92.60%
Blend	100.00%	92.92%
Clean-Label	83.21%	93.30%

2 Backdoor Defence

Task. Mask Train

Step 1. Calculate the adversarial perturbation for neurons. 为模型参数加上初始噪声 δ, ξ, 计算带有噪声的模型的损失对 δ 和 ξ 的梯度 ∇_δ 和 ∇_ξ. 需要利用梯度更新 δ 和 ξ 使得损失达到极大, 因此在实现时应取 $\text{loss} = -\text{criterion}(\text{pred}, y)$.

Step 2. Calculate loss and update the mask values. 利用加入噪声的模型求损失 $\mathcal{L}_{\text{rob}}$, 利用经过 mask 的模型求损失 $\mathcal{L}_{\text{nat}}$. 将实际损失 $\mathcal{L} = \alpha\mathcal{L}_{\text{nat}} + (1-\alpha)\mathcal{L}_{\text{rob}}$ 对 mask 求梯度并更新 mask 以使得损失极小化.

Results & Analysis

α	threshold	ASR(%)	ACC(%)
0.05	0.05	14.19	93.07
0.05	0.10	13.44	92.90
0.05	0.30	10.49	91.98
0.05	0.50	9.62	90.11
0.05	0.80	14.08	71.30
0.10	0.05	24.13	93.25
0.10	0.10	13.27	92.97
0.10	0.30	10.29	91.92
0.10	0.50	9.98	90.85
0.10	0.80	12.69	75.40
0.20	0.05	38.34	93.43
0.20	0.10	17.20	93.08
0.20	**0.30**	**10.48**	**92.16**
0.20	0.50	9.33	89.96
0.20	0.80	13.72	74.50

α 对 ASR 的影响成正相关, 在 threshold 较小时表现显著.

α 对 ACC 的影响并不十分显著.

threshold 对 ACC 的影响成负相关.

threshold 对 ASR 影响表现为随 threshold 增大 ASR 先减少后增大.