

可信机器学习 HW4 Report

1 MIA Attack

Sub-Task 1 Training Data Generation

对于最后一个 epoch 得到的 outputs, 将每一个 output 添加到 X, 依照数据是否属于训练集为 Y 相应添加 1 或 0.

Sub-Task 2 Shadow Model Training

Step 1. Data Generation. 构建数据集时将 `target` 设置为 `False`, `num` 设为 Shadow Model 的编号, 得到用于训练每个 Shadow Model 的数据.

Step 2. Model Training. 使用 Shadow Data 训练 Shadow Model, 将每一次训练过程中由 **Sub-Task 1** 生成的数据保存, 得到用于进行 MIA 的数据.

Result

最终的 ACC 为 94.472%.

2 Unlearnable Examples

2.1 Classwise

Sub-Task 1 Base Model Optimization

每一个 epoch 中, 将图像类别相对应的 `noise` 加到图像上得到带噪声的训练集, 再对 Base Model 进行训练.

Sub-Task 2 Noise Updating

将每一类所有的 $x'_t - x$ 求均值后进行 clamp, 得到 t 时刻的新噪声 δ_t .

Sub-Task 3 Adding Noise

将最终的噪声添加到相应类别所有图片上并进行 clamp, 得到 class-wise min-min 训练集.

2.2 Samplewise Noise Generation

除噪声从每类图片加相同噪声变为每张图片加独立噪声外, 流程与 Classwise 一致.

Step 1. Base Model Optimization. 每一个 epoch 中, 将图像相对应的 noise 加到图像上得到带噪声的训练集, 再对 Base Model 进行训练.

Step 2. Noise Updating. 在一个 epoch 中, 对每张图片利用 x'_t 求 x'_{t+1} , 求相应的噪声.

Step 3. Model Training. 训练模型至 ACC>0.99, 训练结束后得到最终的噪声.

2.3 Results

Classwise Clean ACC: 10.03%.

Samplewise Clean ACC: 30.46%.