**实验报告**

1. **实验简介**

本实验是基于论文Membership Inference Attacks Against Machine Learning Models，该文章是成员关系推理攻击的开山之作，是人工智能安全方向很重要的一篇文章。这个展示了对深度学习模型进行成员关系推理攻击的过程，并提出了一个新的架构“影子模型”，提高攻击的性能过程。所以本实验是基于该文章的内容，并结合Github上开源的代码进行复现。

什么是成员关系推理攻击？成员关系推理攻击是在给定一个数据记录和黑盒访问目标模型的权限的情况下，判断该数据记录是否在目标模型的训练集中。攻击者对模型有查询访问权，并可以在任何数据记录上获得模型的预测向量。攻击者知道模型的输入和输出的格式，包括它们的数量和取值的范围。例如：最基础的手写数字识别模型，攻击者就可以知道输入是一个大小的图片，输出是一个标签或是一个概率向量。其基于的原理是：模型对于训练集中的数据和非训练集中数据的反应会有不同。

本实验就是基于这个开始，对MNIST手写数字识别模型进一步扩展

1. **实验环境**系统：Ubuntu20.04  
   CPU：Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz

内存：64G

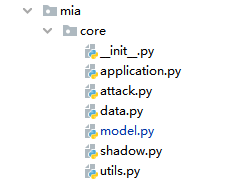
软件环境：Anaconda3，python==3.6,tensorflow==2.5以及其他依赖库

1. **实验内容**在原有开源包<https://github.com/SAP-samples/security-research-membership-inference-and-differential-privacy.git> 的基础上进行修改，因其只支持tensorflow1.14 环境，所以先在其架构基础上将API升级到tensorflow2.5 。

然后在该包的基础上，训练识别MNIST手写数字的目标模型，再对0~9每一个目标类训练一个同架构的影子模型。

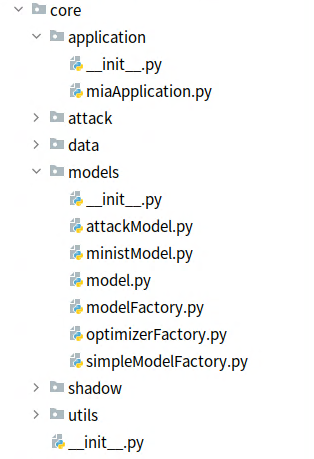
最后通过构造的影子模型训练攻击模型，到达进行成员关系攻击的目标。

1. **实验过程**
2. **对原MIA（membership inference attack）包进行分析**



由如下六部分组成：

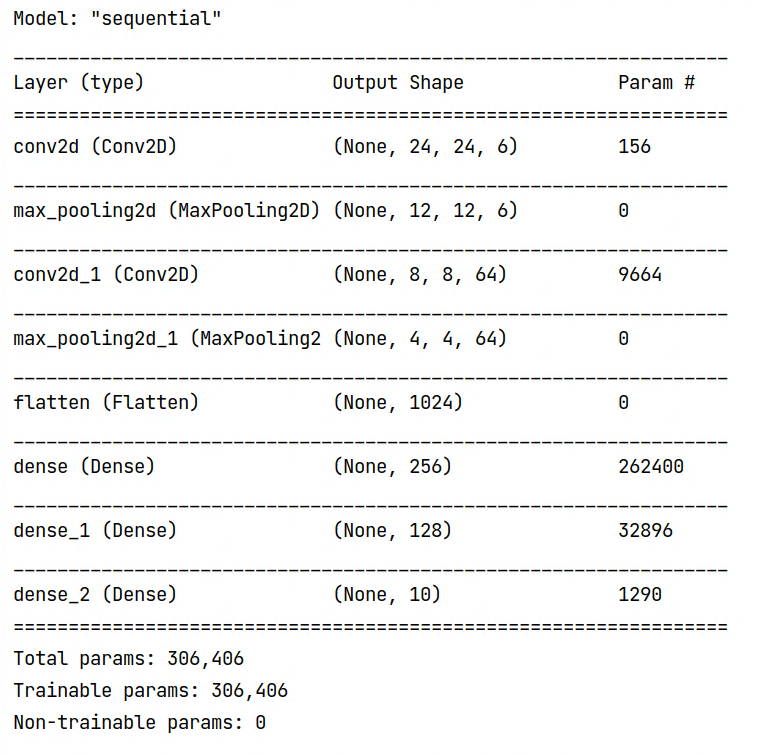
1. model：对各个模型进行抽象，制造模型工厂，方便之后训练影子模型，减少重复代码
2. data：是对训练中的数据的处理进行封装：如目标训练集和影子模型训练集进行区分，以及获得目标数据集的输出标签
3. shadow：是影子模型的封装
4. attack：是攻击模型训练的各功能的封装
5. utils：是其他辅助功能的封装，如：查看精确度等
6. application：是对整个训练过程的封装
7. **代码修改与整理**



在原有的代码上对各个类进行拆分，并修改了目标模型的架构，建立一个MNIST手写数字识别的模型，作为目标模型进行攻击

1. **MNIST手写数字识别模型搭建与训练**

这次使用的模型是LENET手写数字识别模型，使用两个卷积层、两个池化层以及最后两个全连接层以及输出层组成。其结构如下所示：



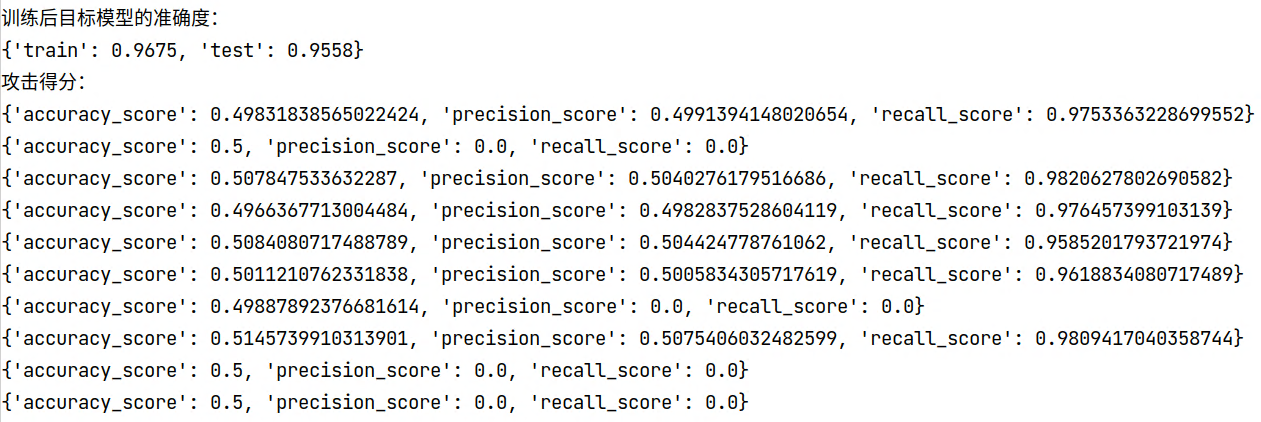
框架内模型实现如下：



1. **数据说明**

这个实验使用了经典的MNIST数据集，有60000个训练样本，10000个测试样本。而在这个攻击框架下，只能使用10000个作为训练集，其余作为影子模型的训练集和后续的推理攻击的已知数据集

1. **模型训练**编写main.py，写入参数，根据框架内部的流程修改部分模块，执行并训练。
2. **注意事项**
3. 由于这个MIA包是一个框架，用于该类型的深度学习进行实验验证，因此留有很大的空间，如：可以自定义callback函数，更换不同的模型。而在这个简单的实验中，却造成干扰，因此在修改时删除了部分模块，写死了部分训练参数。
4. 在修改过程中，由于存在拼写错误和漏写等情况，在运行过程中出现了多处错误，因此在编写代码时要时刻注意各个类型的名称、变量名是否正确，以及各个功能模块是否正确。
5. 我使用的是CPU服务器无显卡，所以没有使用tensorflow-gpu的版本，同时也减少了CUDA等依赖包版本不适配的问题，但是训练速度很慢，导致调试过程费时。
6. **实验结果**



目标模型在使用10000个样本和Lenet模型进行训练后，对于测试集的识别率达到了0.9558 。 下方是我们使用其他样本和影子模型对攻击模型进行训练后，攻击模型对于各个目标类的攻击的准确度，即从recall里看，对于推断目标模型数据集的成员的准确率有0.95以上的准确度，而存在部分标签少的原因，部分为0

1. **感想**

在整体看完开源作者的代码之后，对于这个攻击的整个过程有了更加深刻的理解，并且对于tensorflow2.x的结构有了整体的认识。MNIST手写数字识别是作为深度学习入门的一个练习项目，对于电脑的硬件要求低，是让入门者对于如何获取处理数据，如何搭建模型，如何根据实际情况进行更改有一个大体的了解，为以后在深度学习领域的学习做铺垫。而且深度学习这一学科非常重视实验，若想在该领域进一步的深入探索，那么能够实现自己想法的代码能力也是非常重要。

1. **参考文献**

[1] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. [J]. IEEE Symposium on Security and Privacy. 2017, San Jose,CA, USA, May 22-26, 2017