**Comprehensive Privacy Analysis of Deep Learning**

Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks

**Milad Nasr，Reza Shokri，Amir Houmansadr**

**一．Abstract**

本文（1）对基于深度学习模型的白盒隐私推理攻击进行了全面的分析；（2）利用最终的模型参数以及参数更新来衡量隐私泄漏；（3）在独立学习和联合学习场景设计攻击，考虑了被动和主动推理攻击者、并假设对手有不同的先验知识。

本文利用广泛用于训练深度神经网络的随机梯度下降算法（SGD）的隐私漏洞，设计了适应白盒设置的新算法，实验发现即使是泛化很好的模型也很容易受到白盒成员推理攻击。

**二．Introduction**

本文围绕着“在深度学习里，训练集中的个体有多大的隐私泄露风险？”这个中心问题展开研究。

推理攻击分为MIA与重建攻击两种，本文以MIA为基础进行隐私泄露分析。

最近的几个研究均针对黑盒MIA，发现训练数据的分布、模型的泛化程度都极大影响了成员泄露风险，尤其是过拟合的模型对MIA更脆弱。然而，这些黑盒攻击对泛化良好的深度神经网络效果不好，且模型参数对攻击者有时是可见的（e.g.联合学习共享参数更新）。

**框架：**

针对本文提出的攻击模型设计了一个架构，该架构分别处理从不同层中提取的特征，然后聚合它们以提取成员信息。架构中会训练一个自动编码器来计算所有数据的成员信息值（membership information embedding）。然后，在目标数据集上使用聚类算法分离出成员与非成员。

实验**评估**：

我们评估了CIFAR100数据集上预先训练好的、公开可用的最先进模型的隐私泄露。结果表明，DenseNet模型是目前最好的模型，它不太容易受到黑盒攻击（54.5%攻击准确率），却易受到白盒攻击（74.3%攻击准确率）。这说明：即使是泛化良好的深度模型也会泄漏大量关于其训练数据的信息，且容易受到白盒成员推断攻击。

**三．Inference Attacks**

***A. Attack Observations: Black-box vs. White-box Inference***

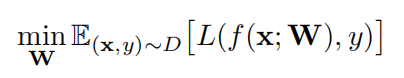
Black-box：敌手只能得到观察到模型的输出。

White-box：敌手还知道参数，除了输出外还可进行中间计算。

**关于为什么SGD会泄露隐私：**

SGD被用于训练深度学习模型，原理如下：

在训练中，SGD会最小化损失函数的期望，



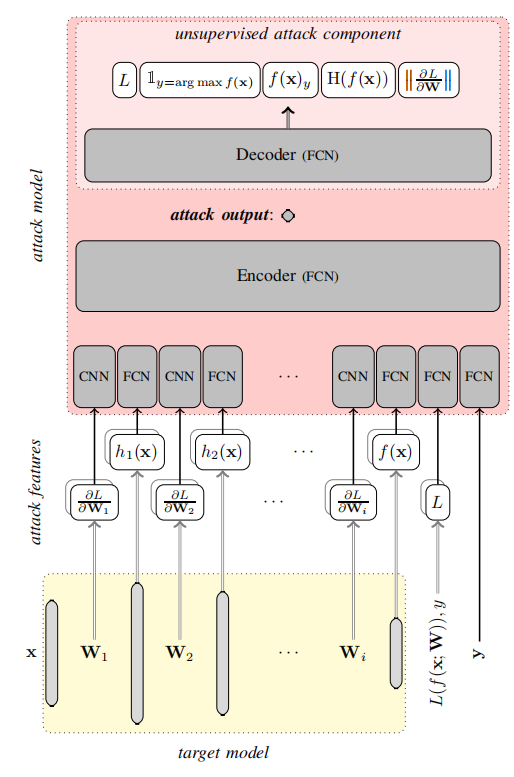
其中，L(f(x;W), y)是模型f的损失函数，W为参数，D为训练集。

SGD进行最小化的方式是不断地更新参数W，以减少随机选择的D子集的损失。因此，对训练集中任意一个样本，每一轮训练后，loss的梯度∂L/∂W 会趋近于0。深度神经网络中使用了大量参数(数百万个参数)，具有如此大的维数的向量不能很好地泛化训练数据。因此，与非成员相比，模型的**梯度**在训练数据成员上的分布是可区分的。这一点使得MIA可以实施。

（可以简单地理解为：

每个训练集中的样本点会通过SGD影响很多参数，以此最小化样本点对学习损失的贡献。为了最小化模型的期望损失，SGD算法会不断更新模型参数，使整个训练数据集上的损失梯度趋近于零。因此每个训练样本都会在模型参数上的损失函数的局部梯度上留下可分辨的足迹。具体的攻击在C部分会写出来。）

**本文的白盒MIA架构如图：**



这段分析从下往上一层层地看比较容易理解：

对目标数据（x,y），敌手目标是判断它是否为目标模型f的训练集D的成员。对输入x，在前向传播中，敌手使用模型f计算出隐藏层hi(x)，模型输出f(x)，损失函数L(f(x;W), y)；后向传播中计算出每层的L关于参数的梯度∂L/∂Wi。这四个计算值加上真实标签y的one-hot编码，作为攻击的输入。

再看上半部分，攻击模型包含了卷积神经网络（CNN）和全连接网络（FCN）。图中有个细节，FCN用于y的one-hot矢量、loss L、输出f(x)和隐藏层hi(x)，CNN用于梯度。

CNN 和FCN的输出合并为一个向量，传给FCN编码器。FCN编码器的输出是一个包含了成员信息的值，也就是attack output。Attack output的值代表了该输入x是成员的概率。

（另外，对于无监督情景，训练一个解码器从attack output中重构输入的重要特征。）

1. ***Inference Target: Stand-alone vs. Federated Learning***

Stand-alone fine-tunning：

在数据集D上训练模型f，之后，使用一个新的数据集D∆进行微调，f被更新为f∆。敌手观察到最终的输出，本文测量最终模型f∆对整个训练集D∪D∆的信息泄露，同时考虑到模型有两个(f和f∆)，本文也从两个模型快照中度量训练数据的额外信息。

针对微调模型的MIA是我们用于攻击联合学习的MIA的一个特例。

Federated Learning：

联合学习情景是：N个参与者，分别拥有不同的训练集Di，他们有一个相同的深度学习任务，用同一个模型架构，训练出一个全局模型。中央服务器存放全局模型参数W的最新值。每个参与者都有一个局部模型，因此有一组局部参数Wi。在培训的每个epoch中，每个参与者下载全局参数，对其本地训练数据使用SGD算法，更新参数，并将其上传到服务器。参数服务器使用所有参与者上传的参数计算每个参数的平均值。重复直到全局模型收敛。

敌手可以是中央参数服务器，也可以是参与者之一。

敌手为中央参数服务器时，接收每个参与者的参数更新并推测每个参与者的训练集信息；也可控制全局模型上每个参与者的view，并可以主动攻击获得更多信息。敌手作为参与者之一时，只能得到随时间的全局参数W{t}，和改变自身的参数更新Wi{t} 。

两种情况中，对手会随着时间的推移观察目标模型的多个版本。

1. ***Attack Mode: Passive vs. Active Inference Attack***

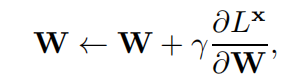
Passive attack:

被动敌手只能观察到参数更新（而不能改变）。

Active Attack：

参与训练过程的敌手可以主动影响目标模型，以获取更多的训练集信息。这种攻击常应用于联合学习，参数服务器或参与者会为后续MIA进行对抗性参数更新（adversarial parameter up-dates）。敌手利用SGD算法实施MIA。

假设敌手是参与者，设x为敌手需要判定是否为成员的记录。敌手通过以下公式对参数加上梯度，实现对x的梯度上升：



（r为对抗更新速度）。

对手将计算的参数上传到中央服务器，服务器将把它们与来自其他参与者的参数更新聚合起来。如果目标记录x在某个参与者的训练集中，则参与者的局部SGD算法会突然降低对x的损失梯度。这可以被推理模型检测到，区分出成员与非成员。

***D. Prior Knowledge: Supervised vs. Unsupervised Inference***

（1）假设敌手有一个与目标数据集D重叠的数据集D’，使用**监督学习**的方式训练攻击模型，并用它来攻击训练数据集的其余部分。

（2）假设敌手没有目标训练集中的样本，MIA的训练有两种可能：影子模型上的**监督训练**，目标模型上的**无监督训练**。

影子模型与目标模型有相同结构，训练数据与目标训练数据的分布相同，但与目标训练集没有重叠。

本文提出的是第二种，也是影子训练的替代方法，即攻击模型对目标模型进行**无监督训练**。这种攻击假设攻击者可以访问与目标训练集部分重叠的数据集D '，但是敌人不知道哪些数据点在D '∩D中。使用聚类算法，对每个数据点求出一个分数以区分成员与非成员。本文利用一个编码器-解码器架构来实现这一点（在前面的架构图片中）。在无监督下的攻击是批量攻击，敌手会攻击大量的数据记录。对每个目标数据记录使用编码器，并计算嵌入值(编码器的输出)。接下来，使用一个聚类算法（例如光谱聚类方法）将目标模型的每个输入聚类到两类中（成员和非成员）。

**四．Experiment:**

1.**数据集**：CIFAR100, Purchase100 ，Texas100

2.**目标模型**：对 CIFAR100 使用的是 Alexnet ， ResNet , DenseNet 三种模型。使用SGD优化器训练模型。

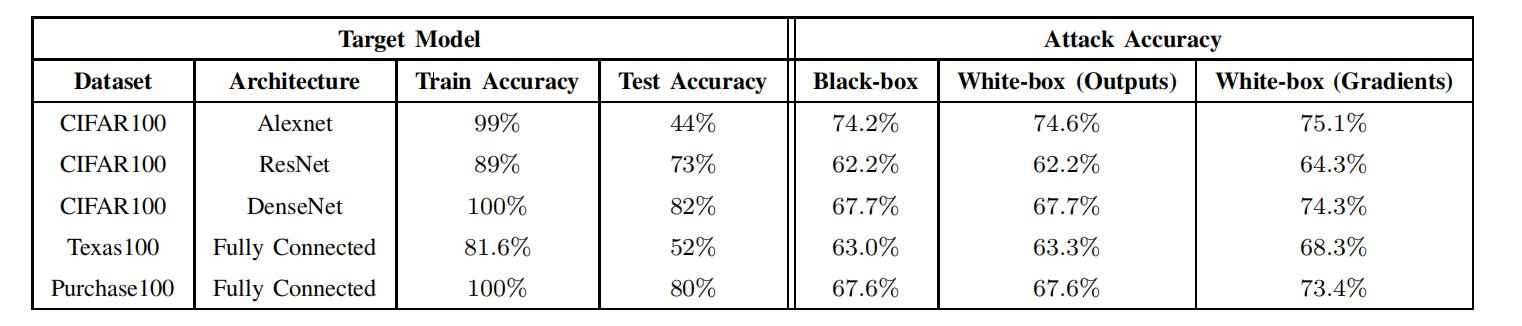
对 Texas100 和Purchase100，使用的是全连接模型。使用Adam优化器训练模型。

3.**预先训练好的模型**：为了证明本文的攻击不局限与文中的训练算法，选择了公开的训练好的CIFAR100模型。

4.**联合学习设置**：使用平均聚合方法。（参与者在每个epoch之后将参数更新发送给中心模型，中心服务器对各方的模型更新进行平均，并将更新后的模型发送给各方。）所有参与方使用相同大小的训练数据集，且训练数据都是均匀随机选择的。

**实验结果：**

1.独立学习中，与层的输出相比，**梯度泄漏了更多关于训练集的成员信息**；本文提出的白盒攻击比黑盒攻击精度高一些。



2.联合学习中，下表显示了平均攻击精度。（全局攻击的Isolating是隔离一个参与者，不将其他参与者的更新传递给他）。

结论（1）与被动全局攻击者相比，主动攻击者攻击精度明显更高。

（2）简单地隔离参与者的训练可以显著提高攻击精度。

（3）局部攻击与全局攻击相比具有较低的准确性。

