MCMIA: Model Compression Against Membership Inference Attack in Deep Neural Networks

**摘要：**深度学习或者是深度神经网络（DNNs）如今在诈骗检测、推荐和不同类型的分析交易等实际应用中有着极好的表现。但是，模型的规模越大，计算占用的资源越多，而且在面对成员关系推理攻击时的脆弱性也阻碍了深度学习的发展，尤其是在性能受限的终端设备上。作为同时应对这些挑战的首次尝试，本文预见DNN模型压缩技术会在减小模型存储和计算开销的同时帮助深度学习模型抵抗MIA。将模型压缩和MIA共同制定为MCMIA，而且提供解决问题的分析方法。经过在不同数据集上的实验比对，发现能够减少攻击精度，因此减少MIA中的信息泄漏。而且该方法比DP相比有好的性能。与本文的MCMIA-Pruning算法，MCMIA-Pruning & Min-Max game算法能有更低的攻击精度，最大限度地增强了DNN模型的隐私防护。而且由于模型压缩有利于适配硬件，尤其在资源受限的平台上以隐私保护方式部署DNN模式时，MCMIA是十分有效的。

**背景：**现在的各大机器学习服务厂商提供各种训练项目，并向用户提供训练好的模型接口，方便用户在终端设备中使用。但是现在有两个问题：首先，上传到第三方服务器的敏感数据的安全和隐私问题，即使仅黑盒模型接口的情况下，仍然有MIA能够获取用户的信息。其次，DNN模型在快速发展以应对不同场景的应用，但是随着层数的加深和模型规模的变大，模型的计算占用越来越大，需要大量的数据训练，也限制了资源受限的终端设备上的用户体验。

为了应对MIA的攻击，之前提出了差分隐私DP，即在训练模型的梯度或目标函数中添加噪声，尽管证明了该方法的稳定性，但是因为在保护复杂模型室以及噪声较大的高维数据中，DP的成本很难被限制在可接受的范围内。另一种方法是博弈理论，如Min-Max game，这个推理攻击模型的收益被当做一个新的正则器。但是与DP一样，Min-Max game 也会引入额外的计算开销。也就是说这两种方法均没有解决第二个问题。

**贡献;** 本文就提出了解决上面两个问题的方案，即将模型压缩和MIA共同制定为MCMIA，提供解决问题的解析方法。并且MCMIA-Pruning 比Min-Max game 有更好的模型精度，而且这两种方法都能最大化模型的防御能力。

**模型压缩MC：**先前的模型压缩算法有权重剪枝（利用不同的正则化技术探索稀疏性）等。核心思想就是保留关键权重和发展优化算法正则化损失函数来保持模型精度，最终用更简单的模型表示神经网络。另一方面，更简单的模型会加速计算并加少权重存储，最终达到更快的训练速度和预测速度。

**ADMN-based DNN Model Compression：**通过交替方向乘子法（ADMM）结合到DNN模型压缩中，可以在保持精度的同时实现高权重减少率。设想一个优化问题带有组合约束，并且使用优化工具很难直接解决。使用ADMM，原来的可以分解成两个在和（辅助变量）上的子问题。这两个子问题就可以分别解决和迭代直到收敛。起初，ADMM被用来加速凸优化问题的收敛并实现分布式优化，其最优性和快速收敛速度已经被证明。ADMM的一个特殊属性是它可以有效地处理组合约束（combinatorial constraints）并产生最优或高质量解。而DNN模型压缩中的相关约束属于该组合约束子集，因此ADMM适用于DNN模型压缩。

模型压缩问题可以写成：

其中为DNN模型的损失函数， ，是第层的指定权重数。

则该问题就能被写成如下形式：

其中是辅助变量，通过增广拉格朗日，这个问题就可以分解成两个子问题。

**问题陈述：**使用代表敌人的推理模型：

输入为数据特征，真实标签和预测标签，输出为该数据记录属于训练集的概率。则MIA的增益函数如下：

则MCMIA问题可以转化成如下形式：

其中，和时模型压缩预测，能限制敌手在一定边界做出正确预测的可能性。

下一步，就用Min-Max game增强MCMIA-Pruning，则相关优化问题转化成如下：

其中，是常数，作为正则化因子。

为了最大化攻击MIA攻击，将推理模型能到达的最优解记为

假设一个来自训练集的，则是否来自训练集的差异可以写为

理想条件下，防御能力最强的模型会使预测的概率为0.5，即抛硬币决定是否在训练集中。则有：

带入MCMIA模型中就转化成：

其中代表带有模型压缩的分类模型

使用扩展拉格朗日方法解决Max-Min game，则拉格朗日格式如下：

又定义，因此拉格朗日格式能被改写如下：

MCMIA问题就如下表示：

**解决策略：**

通过满足如下约束条件，系统地解决了格式变化的问题：

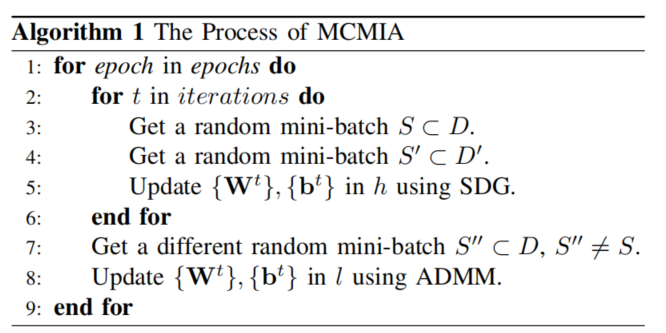
增广拉格朗日格式可以分解为解决如下问题，参数如下进行迭代更新：

优化子问题可以写为：

由于是集合的指示函数，则全局最优解能被显示推导为：

其中，代表集合上的欧几里得投影。

整体算法如下：



**模型分析：**

一般来说，模型越过拟合，就越容易受到MIA的影响，训练数据分布不均匀，泄漏的信息越多。MCMIA则通过剪枝非关键权值来实现参数的稀疏性，来减少过拟合。因此MCMIA-Pruning和Min-Max Game 在训练集和非训练集上有相似的预测概率。