文章编号：2096-1472(2021)-11-39-03
DOI:10.19644/j.cnki.issn2096-1472.2021.011.010

基于自优化深度网络的模型攻击方法

吴 吉，王月娟，景栋盛

(国网江苏省电力有限公司苏州供电分公司，江苏 苏州 215004)
✉13862159678@163.com; 215691852@qq.com; jds19810119@163.com

摘 要：机器学习方法常使用私有数据来训练模型以期获得更好的效果。然而，非授权用户可以通过模型输出来判断数据是否参与训练，破坏了数据隐私安全。对此，提出了基于深度优化网络的模型攻击方法，从攻击者的角度出发，分析攻击方法原理，有针对性地防御对模型的攻击，增强模型的隐秘性。所提方法自动对模型进行攻击，获得自优化的参数，提高攻击的准确度，充分挖掘模型中的安全缺陷，揭示模型的可改进之处，改善模型的安全性。在CIFAR-100数据集上进行了实验，得到AUC值为0.83，优于base方法。实验结果验证该方法能有效地提升攻击效果。

关键词：机器学习；优化；隐私保护；模型攻击
中图分类号：TP309 **文献标识码**：A

A Model Attack Method based on Self-optimizing Deep Network

WU Ji, WANG Yuejuan, JING Dongsheng

(*Suzhou Power Supply Branch, State Grid Jiangsu Electric Power Co., Ltd., Suzhou* 215004, *China*)
✉13862159678@163.com; 215691852@qq.com; jds19810119@163.com

Abstract: Machine learning often uses private data to train model so as to get better performance. However, unauthorized users can input data into the model and determine whether certain data are used for training by the output of the model, which threatens data privacy and security. In order to solve this problem, this paper proposes an attack method based on deep optimizing network, which analyzes the attack method principle from the attacker's point of view, and then defends against the attack on the model in a targeted manner so as to enhance the secrecy of the model. The proposed method attacks the model automatically, obtains self-optimizing parameters, improves the attack accuracy, fully exploits security defects in the model, reveals the improvement of the model, and improves the model security. Experiments have been carried out on CIFAR-100 data set, and the AUC (Area Under the Curve) value is 0.83, which is better than the base method. Experimental results show that the proposed method can effectively improve the attack effect.

Keywords: machine learning; optimization; privacy protection; model attack

1 引言(Introduction)

随着深度学习研究的不断深入，深度学习模型的安全问题引起了研究者的广泛关注，隐私泄露问题越来越受到重视[1-4]。一方面，模型固有的特性使攻击者有机会获取其中的隐私信息；另一方面，模型中的隐藏层会形成较大的有效容量，将一些训练数据细节化为参数[5]，记录在模型中。

通过对测试数据的输出分析，可以对模型有一个明确的衡量，同时也急需一个有效的攻击方法来模拟对目标的攻击，发现模型中存在的问题。虽然已有一些方法，但是这些方法在模型的攻击精度上还有待提高。因此，需要设计研发一种有效的方法来提高攻击的效果，从而更好地改进模型的安全性。

针对这一问题，本文提出了基于自优化的深度网络模型攻击方法，通过已知模型的层数，对其进行计算，得出一组攻击参数，使用这些指定参数对模型相应的层进行攻击，获得较好的攻击效果。

2 相关工作(Related work)

2.1 推理攻击

针对机器学习算法的推理攻击分为成员推理和重构攻击。在重构攻击中，攻击者的目标是推断训练集中记录的属性[6]。成员推理攻击利用了一种观察，即机器学习模型在它们所训练的数据上的行为常常与它们第一次“看到”的数据不

同。攻击者会构建一个攻击模型，该模型可以识别目标模型行为中的这些差异，并利用它们来区分目标模型的成员和非成员。

深度学习的数据以不同方式被用于训练模型。基于成员推理攻击方法的攻击者可以观察深度学习过程，通过深度学习模型测量训练数据的泄露情况。本文提出的方法利用深度学习算法。

2.2 差分隐私

差分隐私技术使攻击者很难通过模型的输出来分辨某条数据是否被用于训练机器学习模型，从而达到保护数据隐私的效果[7]。按照差分隐私的要求，在数据集中添加或删除一条数据后，都不会显著影响作用在该数据集上的算法的输出结果[8]。差分隐私已经被用于对推理攻击的强防御机制[9-10]。研究人员将差分隐私引入模型算法中，对模型的梯度进行扰动，提高了隐私性[11]。

本文对差分隐私方法进行改进，重点分析哪些数据被用来训练模型的个人隐私。为了达到保护隐私的目的，进一步分析攻击方式来评估模型的优劣。对绝大多数机器学习任务而言，在算法求解过程中满足差分隐私，即可以认为实现了对模型的隐私保护。

2.3 ML Privacy Meter

ML Privacy Meter是Python基于Tensorflow 2.1开发的一个应用程序接口，可以针对目标模型训练攻击模型，并可以使用指定的攻击方式训练出攻击模型。ML Privacy Meter使用成员推理攻击来测量深度学习模型训练数据的信息泄露，数据被用于训练模型，攻击者也可以观察深度学习过程。

对于一个目标数据记录，攻击模型计算损失，并可以使用一个简单的反向传播算法计算有关所有参数的损失梯度。由于深度神经网络中使用了数以百万计的参数，具有如此大维数的向量不能正确地对训练数据进行泛化。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，可以帮助对手运行精确的成员关系推理攻击，使分类模型得到很好的概括。

3 方法(Method)

3.1 攻击参数自优化

虽然ML Privacy Meter提供了比较方便的测量，但是没有提供优化参数的方法。为了能获得较好的白盒攻击策略效果，本文使用整体参数优化选择的方法。这个方法充分考虑目标模型层数，进行平均细分后再决定攻击的层数N。

在进行白盒攻击时，需要确定对哪些层进行攻击。整体参数优化法可以尽可能地对模型参数进行探索，同时又能避免逐层对模型进行穷究式探索，获取模型中最关键的中间层。可见，整体参数优化法具有明显的优势。据此，本文设计了一种攻击参数自优化方法，采用均方误差作为Loss函数，其计算方式为：

$$MSE=\frac{1}{n}\sum_{i=1}^{m}w_i(y_i-\hat{y}_i)^2 \tag{1}$$

其中，n为样例个数，w_i是各个样例权重，y_i为真实数据，$\hat{y}_i$为预测值。

自优化网络攻击方法如算法1所示。

算法 1：自优化网络攻击方法(Self-Optimizing Net Attack，SONA)

1： 训练目标模型M

2： 获得模型的网络层数参数Layer_Num

3： 初始化攻击attack_hander

4： 初始化攻击模型 θa

5： 通过Split方法从Layer_Num中获得目标层列表targetLayersList

6： **for** $i\in[0,\text{epochs}]$ **do**

7： mtrain_data，ntrain_data = attack_hander()

8： moutputs = forward_pass(M,mtrain_data,N)

9： noutpus = forward_pass(M,ntrain_data,N)

10： 利用式(1)计算损失函数 Loss(ntrain_data，mtrain_data)

11： 使用梯度下降更新参数 θa

12：**end for**

3.2 目标模型

Alexnet的网络结构模型引爆了神经网络的应用热潮，并赢得了2012届图像识别大赛的冠军，使得CNN成为在图像分类上的核心算法模型，很适合作为验证模型。Alexnet的网络结构如图1所示，包含8 层权重，前5 个是卷积的，其余3 个是完全连接的，最后一个完全连接层的输出被馈送到1000路Softmax激活函数。本文设计的网络最大化了多项逻辑回归目标，相当于最大化了在预测分布下正确标签的对数概率在训练案例中的平均值。

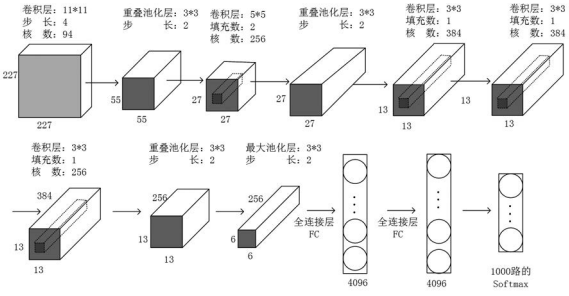

图1 Alexnet网络结构示意图

Fig.1 Illustration of Alexnet network architecture

AlexNet与LeNet相比，网络结构更丰富，有明显的优势。AlexNet通过使用Dropout实现数据增强，从而抑制数据过拟合，适合用来作为神经网络攻击的对象。攻击的训练流程如图2所示。

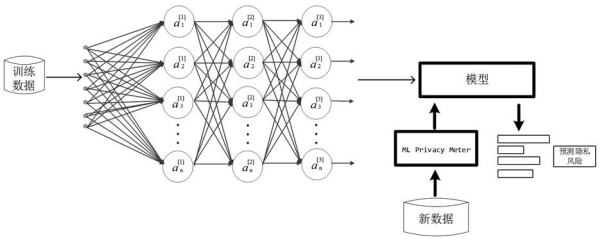

图2 攻击流程示意图

Fig.2 Illustration of attack flow

4 实验与分析(Experiment and analysis)

4.1 数据集与评价标准

本文采用CIFAR−100作为测试数据集。CIFAR−100数据集是一个经典的数据集，由100 个类组成，每个类有600 个32×32彩色图像。数据集分为五个训练批次和一个测试批次，每个批次有100 个图像。测试批次包含来自每个类别的100 个随机选择的图像。同时，CIFAR−100数据集中的100 个类被分成20 个超类。每个图像都带有一个“精细”标签(所属的类)和一个“粗糙”标签(所属的超类)。

由于现实中样本在不同类别上分布不均衡，使得传统的度量标准不能恰当地反映出分类器的表现。使用ROC曲线(Receiver Operating Characteristic Curve)作为评价指标，能直观地表现出分类效果。ROC曲线向左上角弯曲的幅度越大，代表这个分类器效果越好。在ROC曲线图中，*X*轴为假阳性率(FPR)，*Y*轴为真阳性率(TPR)。

4.2 对目标模型的攻击

本文以完成预训练的Alexnet网络结构模型作为攻击目标，模型共有26 层网络和24 个隐藏层，对其进行攻击比较：(1)直接对其进行黑盒攻击(base)；(2)使用整体参数优化的白盒攻击。对于26 层的目标模型，考虑输出层，进行两次计算，以第6 层、第13 层和第26 层为目标。根据这种参数优化，以Alexnet网络为目标训练出攻击模型。在测试集中的ROC曲线如图3所示，根据ROC曲线计算出其AUC值。本文所提方法的AUC值为0.83，比base方法AUC值(0.80)提高了3.75%。

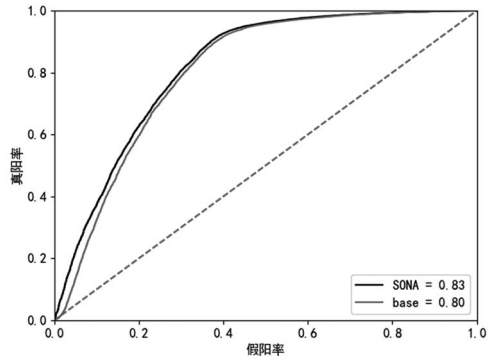

图3 SONA方法及base方法的ROC曲线及AUC值

Fig.3 ROC curves and AUC scores of SONA method and base method

从图3可以看出，由于对中间层的梯度进行了更为详尽的分析，使得白盒攻击更加有效。因此，从结果可以发现，黑盒攻击不如参数优化后的白盒攻击的效果。但是，通过实验可知，在白盒攻击时，太过频繁地对相近层数进行攻击，会导致效果有所下降，这是因为两个相近层次之间的梯度会相互造成较大的影响。

5 结论(Conclusion)

大部分模型都存在一定的漏洞，可以根据模型对同一组数据的输出不同，通过白盒攻击或黑盒攻击来改进安全性。本文在Alexnet的基础上提出了一种参数优化的白盒攻击方法，提高了攻击的准确度，改善了攻击效果，充分挖掘出模型中的安全缺陷，发现模型中的可改进之处，提升了模型的安全性。

参考文献(References)

[1] 谭作文,张连福.机器学习隐私保护研究综述[J].软件学报,2020,31(7):2127−2156.

[2] 姜妍,张立国.面向深度学习模型的对抗攻击与防御方法综述[J].计算机工程,2021,47(1):1−11.

[3] 陈宇飞,沈超,王骞,等.人工智能系统安全与隐私风险[J].计算机研究与发展,2019,56(10):2135−2150.

[4] 余方超,方贤进,张又文,等.增强深度学习中的差分隐私防御机制[J].南京大学学报(自然科学),2021,57(1):10−20.

[5] BILOGREVIC I, HUGUENIN K, AGIR B, et al. A machine−learning based approach to privacy−aware information−sharing in mobile social networks[J]. Pervasive and Mobile Computing, 2016, 25:125−142.

[6] 刘睿瑄,陈红,郭若杨,等.机器学习中的隐私攻击与防御[J].软件学报,2020,31(3):866−892.

[7] 张润莲,张瑞,武小年,等.基于混合相似度和差分隐私的协同过滤推荐算法[J].计算机应用研究,2021,3(9):1−7.

[8] 任奎,孟泉润,闫守琨,等.人工智能模型数据泄露的攻击与防御研究综述[J].网络与信息安全学报,2021,7(1):1−10.

[9] 陈杨,于守健.基于差分隐私的决策树发布技术研究[J].计算机与现代化,2017(3):59−64.

[10] 刘俊旭,孟小峰.机器学习的隐私保护研究综述[J].计算机研究与发展,2020,57(2):346−362.

[11] 李欣姣,吴国伟,姚琳,等.机器学习安全攻击与防御机制研究进展和未来挑战[J].软件学报,2021,32(2):406−423.

作者简介：

吴　吉(1981−)，女，本科，工程师.研究领域：信息安全，软件开发.

王月娟(1981−)，女，硕士，工程师.研究领域：信息安全，软件开发.

景栋盛(1981−)，男，硕士，高级工程师.研究领域：机器学习，网络安全.