基于神经网络的模型反演攻击技术综述

张 欢[1,2], 韩言妮[1,2], 赵一宁[3], 张 帆[3], 谭 倩[1], 孟 渊[4]

[1] 中国科学院信息工程研究所 北京 中国 100085
[2] 中国科学院大学网络空间安全学院 北京 中国 100049
[3] 中国移动信息技术中心 北京 中国 100083
[4] 新疆阿克苏地区阿克苏市公安局网安部门 新疆阿克苏 中国 843000

摘要 大数据时代下, 基于神经网络的模型研究是人工智能领域的一个主流方向。相比于其它的智能优化算法, 神经网络具有自适应性强、泛化能力显著等优点, 被广泛应用于语音识别、计算机视觉和自然语言处理等领域。然而, 随着神经网络在各领域发挥关键作用的同时, 也引发了隐私泄露、数据窃取等隐私安全问题。人工智能安全问题也随之成为当前国内外的研究热点。基于神经网络的模型反演攻击技术研究如何从神经网络模型输出数据中进行学习、推导, 以得到有关输入数据的信息。通过对输入数据进行深度挖掘和关联分析, 可能会还原出用户的重要敏感数据, 从而引发更为严重的安全问题。同时, 模型反演攻击技术也会推导出有关神经网络的网络结构和模型参数等信息, 对神经网络模型的安全造成威胁。为了系统了解基于神经网络的模型反演攻击技术的研究进展和现状, 本文对神经网络的安全问题及模型反演攻击技术研究进行了详细调研。首先, 本文介绍了模型反演攻击技术的概念和常见攻击场景。然后, 讨论神经网络面临的模型反演攻击挑战, 包括原始数据保护、敏感数据泄露、模型训练隐私等安全问题。接着, 对基于梯度优化和参数训练的两类神经网络模型反演攻击技术进行综述, 对各类方法进行对比, 并总结了典型的防御方法。最后总结全文并探讨了未来的研究方向。

关键词 神经网络; 模型反演攻击; 人工智能安全
中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.12.14

A Survey of Model Inversion Attack Techniques Based on Neural Networks

ZHANG Huan[1,2], HAN Yanni[1,2], ZHAO Yining[3], ZHANG Fan[3], TAN Qian[1], MENG Yuan[4]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
[2] School of Cyber Security, University of Chinese Academy of Science, Beijing 100049, China
[3] China Mobile Information Technology Center, Beijing 100083, China
[4] Cyber security department of the Public Security Bureau of Aksu City, Aksu Prefecture, Xinjiang 843000, China

Abstract In the era of big data, neural network-based model research is a mainstream direction in the field of artificial intelligence. Compared with other intelligent optimization algorithms, neural network has the advantages of strong adaptability and significant generalization ability, and is widely used in the fields of speech recognition, computer vision and natural language processing. However, as neural network plays a key role in various fields, it also causes privacy security problems such as privacy leakage and data theft. Artificial intelligence security has become a hot topic at home and abroad. Model inversion attack technique based on neural network studies how to learn and derive from the output data of neural network models to obtain information about the input data. Through in-depth mining and association analysis of the input data, important sensitive data of users may be restored, leading to more serious security problems. At the same time, the model inversion attack technology can also deduce the information about the network structure and model parameters of the neural network, which will threaten the security of the neural network model. In order to systematically understand the research progress and present situation of model inversion attack technology based on neural network, this paper makes a detailed investigation on the security problems of neural network and model inversion attack technology. Firstly, this paper introduces the concept of model inversion attack technology and common attack scenarios. Then, the challenges of model inversion attacks faced by neural networks are discussed, including original data protection, sensitive data leakage, model training privacy and other security issues. Then, two kinds of neural network model inversion attack techniques based on gradient optimization and parameter training are reviewed, various methods are compared, and the typical defense methods are summarized. Finally, the paper summarizes the whole paper and discusses the future research direction.

通讯作者: 谭倩, 博士, 助理研究员, Email: tanqian@iie.ac.cn。
本课题得到 2021 年重庆市属本科高校与中科院所属院所合作项目(No. HZ2021015)资助。
收稿日期: 2020-09-23; 修改日期: 2020-12-15; 定稿日期: 2022-12-08

Key words neural network; model inversion attack; artificial intelligence security

1 引言

大数据时代下, 随着数据量和计算能力的爆炸性增长, 机器学习及其相关技术正在迅速发展。尤其是基于神经网络的模型在语音识别、计算机视觉以及自然语言处理等领域表现出了显著的优越性[1-3]。神经网络技术的大规模应用不仅提高了软件系统处理海量复杂数据的能力, 也使得神经网络成为方便人们生活和促进社会发展进步的关键技术。

与此同时, 神经网络模型的训练和测试都需要使用大量的数据, 然而这些数据通常包含敏感信息, 如图像、语音记录、位置日志和医疗记录等。攻击者可能会利用获取的数据进行非法活动, 从而造成严重后果。因此, 神经网络模型在被广泛应用的同时, 也存在着隐私泄露、数据窃取等安全威胁[1, 4-8]。虽然目前已经提出了大量的神经网络学习框架、算法和优化机制, 但对学习模型、算法以及数据的安全性研究仍处于探索阶段。

目前, 谷歌、微软等国际领先的科技企业提出了机器学习即服务(ML-as-a-Service, MLaaS)[9]的概念, 这是一种基于深度学习的黑盒应用程序接口(Application Programming Interface, API)。通过MLaaS, 用户可以将本地数据集上传至云端, 然后直接调用分类、聚类或者回归函数的模型接口来获得期望结果。机器学习接口通过对输入样本的迭代训练, 来学习数据中的潜在模式和联系。由于模型参数庞大, 常常会造成对输入样本数据的过度拟合, 意味着模型隐含地记忆了关于训练数据的特殊细节, 从而带来了隐私泄露的风险[10]。这就给攻击者以可乘之机, 攻击者可以基于对机器学习模型的访问权限, 推断出有关输入数据的重要信息, 获取用户隐私, 从而达到攻击目的。

模型反演攻击技术是隐私攻击中的一种重要技术手段[11], 它能够通过对神经网络模型输出数据的学习、推导, 得到输入数据中的敏感信息, 这对神经网络模型的隐私安全造成了极大的威胁。同时, 神经网络的多种不同类型的模型虽然在各个领域表现出了优越的性能, 但这些模型的设计仍然是经验启发式的, 目前尚不能很好的解释模型的运行原理。因此, 如果通过对神经网络模型不同层次的确切信息进行反演, 得到中间层的信息, 将有助于深入理解神经网络模型的行为和特征, 可以更好地解释神经网络模型的结果[12-15]。

虽然已经有学者对机器学习中的隐私保护技术进行研究[5, 8, 16-19], 本文主要侧重于对神经网络中的模型反演攻击技术进行分析总结。

本文的组织结构如下: 第 2 章简要介绍了神经网络、模型反演攻击的概念以及典型的攻击场景; 第 3 章介绍了神经网络面临的反演攻击挑战; 第 4 章对神经网络模型反演攻击技术进行了分析; 第 5 章展望了模型反演攻击技术未来的研究方向; 第 6 章进行了总结。

2 相关知识

神经网络是一种运算模型, 由一组称为神经元的简单组件及其分层组织的大规模并行结构组成[20], 是一种受生物神经网络启发的人工神经网络(Artificial Neural Network, ANN)。目前人工神经网络的结构包含输入层、隐藏层和输出层, 输入层负责接收外部的信息和数据; 隐藏层负责对信息进行处理, 不断调整神经元之间的连接属性, 如权重和偏差值等; 输出层负责对计算的结果进行输出[21]。每层结构都由一定数量的神经元及其相关连接组成, 每个神经元的结构和功能都比较简单, 而大量神经元组合产生的系统行为却非常复杂。人工神经元以不同的方式, 通过改变连接方式、神经元的数量和层数, 能够组成不同的神经网络模型[22]。神经网络的基础结构如图 1 所示。

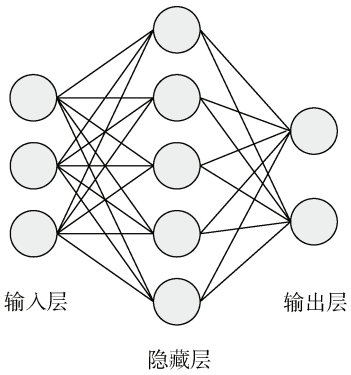

图 1 神经网络的基础结构
Figure 1 Infrastructure of neural network

神经网络模型中每个神经元使用激活函数将一组输入映射到输出, 通过学习更新权重和激活函数, 以便能够正确地确定输出。一般来说, 神经网络模型的目标是输出一个适合给定有限数据集的模型, 对数据集进行训练以优化所有参数, 这是一般的神经

网络的模型正演, 图 2 给出了神经网络解决问题的一般过程, 可分为四个步骤。首先, 对要输入到模型中的数据进行处理等准备工作; 第二步构建模型, 定义神经网络模型的结构和设置模型参数; 第三步训练模型, 初始化模型参数, 进行前向传播, 定义并计算损失函数, 选择反向传播进行更新权重优化, 这一步反复迭代直到达到预设目标; 最后一步为测试模型, 在测试集上评估训练模型的性能。

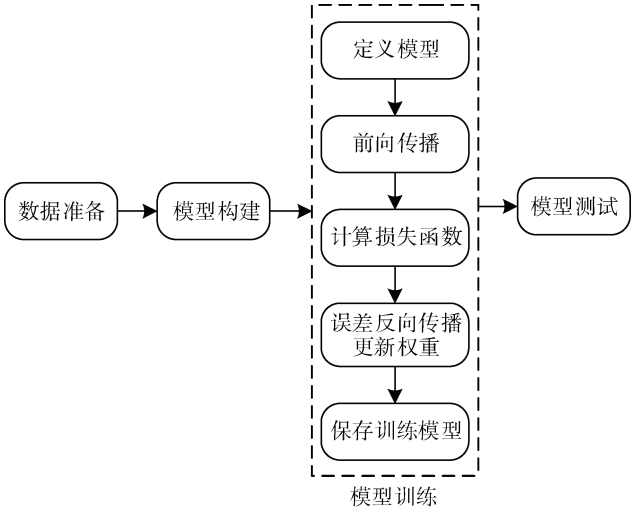

图 2 神经网络解决问题过程

Figure 2 Problem solving process of neural network

模型反演攻击(Model Inversion Attack)最早是由 Fredrikson 等人[23]首先提出的, 是一种从模型的预测值中获取系统模型的一些信息, 通过这些信息进行逆向反演, 从而获得有关训练数据信息的攻击技术。图 3 给出了模型反演攻击的示意图, 正常用户向神经网络模型输入数据 x, 经过计算分析, 得到输出数据 y。另一方面, 攻击者获得模型的输出数据 y 后, 通过模型反演技术, 得到有关输入数据的信息 $\hat{x}$。这种攻击会导致模型训练数据集中所包含的敏感信息以及模型系统等细节信息暴露。因其对神经网络等模型的安全性具有很大的威胁, 引起了越来越广泛的关注[24-27]。

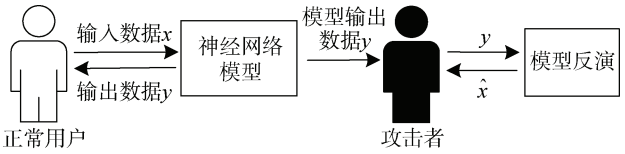

图 3 模型反演攻击

Figure 3 Model inversion attack

根据攻击者背景知识的掌握情况, 可将神经网络模型反演攻击技术的常见攻击场景分为白盒攻击(White-box attack)和黑盒攻击(Black-box attack)[27, 28], 表 1 给出了黑盒攻击和白盒攻击之间的区别。

白盒攻击指的是攻击者完全了解用于训练的模型, 如训练中使用的算法、模型结构以及参数等信息, 并且可以访问训练数据分布。攻击者利用可用的信息来识别模型可能易受攻击的特征空间, 通过改变输入数据来利用该模型。

表 1 黑盒攻击与白盒攻击的区别

Table 1 Differences between Black-box attack and White-box attack

攻击场景	攻击者的能力	攻击策略
白盒攻击	攻击者有详细的模型架构知识、参数信息和输入数据分布	基于模型与输入端的损失函数生成结果
黑盒攻击	攻击者对模型没有了解, 只能统计和观察模型对输入数据的输出情况	通过观察不同输入的输出情况变化, 生成近似于实际输出的结果

黑盒攻击则与白盒攻击相反, 在黑盒攻击中, 攻击者对模型一无所知, 只能与模型提供的接口进行交互, 整个目标模型对攻击者来说就是一个黑箱。攻击者只能使用有关场景或过去输入的信息来分析模型的脆弱性。例如, 攻击者通过提供一系列精心设计的输入, 通过观察输出情况来利用模型。

3 神经网络面临的模型反演攻击挑战

模型反演攻击是直接侵犯数据隐私和模型隐私的最致命的攻击之一, 近年来, 针对神经网络模型的隐私入侵攻击越来越多, 神经网络面临的反演攻击挑战也越来越多[19, 29-31]。模型反演攻击主要是利用输出数据的信息、模型提供的一些信息以及可能得到的训练数据的分布情况等信息来反演模型, 从而获取训练数据集中的隐私信息或模型细节的相关信息。目前, 神经网络面临的模型反演攻击挑战包括原始数据保护、敏感数据泄露、预测数据隐私、训练类别推理以及模型训练隐私等方面, 本节将详细讨论神经网络面临的模型反演攻击挑战。

3.1 原始数据保护

模型反演攻击技术针对神经网络最常见的攻击目标就是恢复输入数据的信息, 也就是对原始数据的保护提出了挑战。

Fredrikson 等人[32]提出了基于置信度的模型反

演攻击, 攻击者可以利用目标模型的预测向量所揭示的置信度信息来推断训练数据集中的部分信息。此外, 他们还表明, 攻击者可以从人脸识别模型中恢复出可识别的人脸图像。

Zhang 等人[33]则提出了基于生成式的模型反演攻击, 该方法首先利用生成式模型从公开数据集中学习数据分布先验, 基于优化的反演攻击方法来指导反演过程。然后, 基于生成对抗网络设计了端到端的攻击算法, 反演深度神经网络, 可以恢复出训练数据的相关信息, 例如能够恢复出人脸识别数据集中的人脸图像和手写数字数据集中的数字图像。

Mejia 等人[34]通过改进深度神经网络模型的对抗性训练模型(Adversarial Training Model, ATM)的语义表示, 使针对神经网络的模型反演攻击能够重构更好的原始训练数据。他们在图像分类数据集上对分类模型进行了改进的对抗性模型训练, 之后对其进行了反演攻击, 能够重建出清晰的原始训练图像。

目前, 虽然针对模型反演技术对原始数据的攻击提出了一些基本防御对策[1, 5, 7, 18, 35], 但这些对策还没有表现出明确的防御效果。

3.2 敏感数据泄露

当原始数据涉及到用户的一些隐私数据时, 针对神经网络的模型反演攻击技术将可能会造成敏感数据的泄露问题。

在医疗场景下, Fredrikson 等人[23]提出了一种模型反演攻击算法, 使攻击者在给定模型和一些患者的人口属性统计信息后, 可以反演出患者的遗传信息, 即可反演预测出有关基因型的信息, 而且预测的准确率达到了 58%, 在一定程度上会对用户的敏感数据造成泄露问题。

Wang 等人[36]从恶意服务器的角度对联邦学习进行攻击, 以一种无形的方式恢复用户级别的隐私。他们提出了一种将生成对抗网络(Generative Adversarial Networks, GAN)与多任务鉴别器相结合的通用攻击框架, 该框架可以同时识别输入样本的类别、真实性和客户身份。而且该方法并不会干扰联邦学习模型的训练过程, 能够在服务器端“隐形”地工作。该攻击方法在 MNIST 和人脸识别数据集 AT&T 上成功地恢复了特定用户的样本。

模型反演攻击技术造成的敏感数据泄露可能会使用户的人身财产安全受到威胁, 还可能会威胁到公司和国家的安全, 对网络安全以及社会安全都形成了极大的威胁。

3.3 预测数据隐私

目前, 大多数研究只关注训练过程中的隐私问题, 而对预测推理过程中的数据隐私问题研究较少, 原因是恢复预测样本比恢复训练样本更具挑战性, 主要因素有以下两方面: (1)模型参数不依赖于预测输入。因此, 预测过程向攻击者提供的关于预测样本的有用信息较少; (2)训练样本通常遵循一定的分布, 使得对手能够恢复关于这种分布的统计信息。然而预测样本一般数量较少, 通常没有这种假设, 因此很难恢复单个样本。

尽管对预测数据的反演恢复有很大的挑战, 但是 He 等人[37]设计了一种模型反演攻击方法, 旨在恢复输入到模型中的预测数据, 并且在白盒环境和黑盒环境下都实现了对预测数据的获取, 使得预测数据的隐私保护受到了极大的威胁。

3.4 训练类别推理

对于分类任务, 模型反演攻击技术对神经网络模型进行反演, 不仅可以恢复出训练样本的隐私信息, 还可以推断出训练样本的类别信息(标签信息), 并可能恢复出类别的代表性样本。

Yang 等人[38]提出了一种基于背景知识校准的对抗性环境下的神经网络反演攻击方法。该方法在没有充分了解原始训练数据的情况下, 基于背景知识从比原始训练数据集更一般的数据分布中提取训练数据。例如对于人脸识别任务, 攻击者可以在不了解训练数据中人脸图片分布的情况下, 从互联网上随机爬取人脸图片来组成辅助样本, 用以训练反演模型, 仍然能够得到准确的反演, 并准确推断出了分类任务的训练类别的意义或其代表性样本。

Basu 等人[39]在攻击者对分类问题有一定了解的情况下, 利用这些一般信息来指导寻找训练数据的代表性样本, 结合生成对抗网络(GAN)设计了一种模型反演攻击方法。攻击者基于该方法可以从互联网上搜集有关该分类问题的样本, 并通过 GAN 来生成单个类的代表性和可识别样本, 输入到模型中, 可以反演出各个类别的代表性和可识别样本。

训练类别推理的成功将有利于攻击者获取有关训练数据的分布以及类别等信息, 对后续的攻击获取训练数据的特征以及隐私等信息提供了详细的背景知识, 将会大大提高攻击的成功率。

3.5 模型训练隐私

针对神经网络的模型反演攻击技术除了可以对模型所需数据的隐私安全形成威胁, 还对神经网络模型自身的安全提出了挑战。模型反演攻击对模型的威胁主要是指对模型的窃取, 即利用模型反演技术窃取模型结构以及模型参数等信息。

Du 等人[13]研究了一种引导特征反演的框架,

通过对深度神经网络模型进行特征反演, 得到了深度神经网络模型训练过程中的特征参数, 还确定了每个特征在输入中的贡献度。该框架在一定程度上加强了深度神经网络模型的可解释性, 但是如果攻击者利用这种方法对深度神经网络等模型进行攻击, 则会获取目标模型训练过程中的模型隐私信息。

Shi 等人[40]提出了一种基于深度神经网络的模型反演攻击, 通过将任意分类模型作为黑盒进行轮询来推断该模型功能, 并使用模型返回的标签构建功能等效的模型。具体方法是, 攻击者可以利用预先从被攻击的分类模型中获得的标签, 采用深度神经网络推断出所需的信息, 并在不知道原始分类模型类型、结构或底层参数的情况下, 反演得出与训练模型功能等价的模型。在文本分类应用上利用这种基于深度神经网络的反演攻击方法可以高精度地推断出分类器的类型, 并窃取它们的功能。

攻击者对神经网络模型进行反演攻击, 获取模型参数和模型结构等信息, 将可能会构造出和目标模型相似度非常高的模型。如果攻击者将该模型用于非法活动, 将会对社会和用户的隐私安全造成严重后果。

基于以上介绍的神经网络反演攻击挑战, 非常有必要对神经网络的模型反演攻击技术给予进一步的关注和研究。将现有的基于神经网络的模型反演攻击的目标总结在表 2 中。

表 2 基于神经网络的模型反演攻击目标

Table 2 Targets of model inversion attack based on neural network

反演攻击目标	相关论文
原始数据保护	Fredrikson 等[32]、Zhang 等[33] 、Park 等[41]、Mejia 等[34]、Yang 等[42]、GMI[43, 44]、GAMIN[45]、Wei 等
敏感数据泄露	Wang 等[36]、Fredrikson 等[23]、mGAN-AI [36]
预测数据隐私	He 等[37]
训练类别推理	Yang 等[38]、Shi 等[46]、Basu[39]
模型训练隐私	Du 等[13]、Mahendran 和 Vedaldi[15]、Shi 等[40]、Nash 等[47]、i-RevNet [48]、Hitaj 等[49]

4 神经网络模型反演攻击技术

神经网络的模型反演攻击是试图根据任何给定模型的输出值, 能够找到其输入数据信息的技术。基于神经网络的模型反演的研究主要分为两类方法。第一类方法是由 Fredrikson 等人[23]提出并发展的利用数据空间中的梯度优化来反演模型[13, 15, 23, 32-34, 37, 41]的方法, 称这类方法为基于优化的方法。之后又发展出通过学习充当原模型的反演模型的第二个模型来反演模型[36, 37, 39, 40, 42-49]的方法, 称这类方法为基于训练的方法。

本节将主要对这两类攻击技术以及近些年出现的一些新的模型反演攻击技术进行介绍, 并对两类技术进行比较, 比较结果总结在表 3 中。

表 3 基于神经网络的模型反演攻击技术

Table 3 Methods of model inversion attack based on neural network

模型反演技术	研究思路	优点	缺点	相关论文
基于优化的模型反演攻击	找到一个输入数据, 基于梯度优化策略使得该数据与输入数据尽可能相似, 能够使得该数据经过神经网络模型得出的预测结果与原输入数据的预测结果近似。	思路简单直接, 基于梯度, 便于计算。	往往产生的图像并不像真正的自然图像, 尤其是对于大型神经网络, 其效果较差。此外, 这种方法在测试时涉及到优化, 需要多次计算梯度, 速度相对较慢。	[13] [15] [23] [32] [33] [34] [37] [41] [50]
基于训练的模型反演攻击	基于训练一个原神经网络模型的逆向模型(即反演模型), 来从输出的预测数据中反演得出有关输入数据的信息。	根据给定的预测进行重建只需要通过网络进行一次正向传递, 速度相对较快。	反演模型的训练过程成本较高, 反演难度较大。	[36] [37] [39] [40] [42] [43] [44] [45] [46] [47] [48] [49]

4.1 基于优化的反演攻击技术

基于优化的反演方法基本思想是在输入空间 X 中应用基于梯度的优化, 以找到一个图像 $\hat{x}$, 使其基于神经网络模型 F_w 的预测结果 $F_w(\hat{x})$ 近似于原始输

入数据 x 的预测结果 $F_w(x)$。该反演方法也可以用来生成某个类 y 的代表性图像(如训练类别推理), 即将 $F_w(x)$ 替换为向量化的 y。

一系列的研究已经使用自然图像空间 P 的先验 $P(\hat{x})$ 来调整优化, 图 4 给出了基于优化的反演攻击技术的框架。形式上, 基于优化的反演方法是为了找出使下面的损失函数最小化的 $\hat{x}$。

$$O(\hat{x}) = L\left(F_w(\hat{x}), F_w(x)\right) + P(\hat{x}) \tag{1}$$

其中 L 是距离度量, 如 $L2$ 距离。

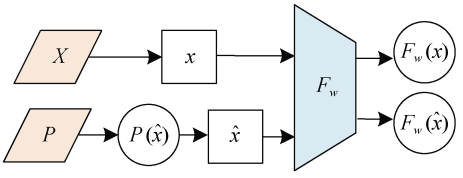

图 4 基于优化的反演攻击技术框架

Figure 4 Framework of optimization-based inversion attack

4.1.1 基于置信度信息的模型反演攻击

模型反演攻击最早是由 Fredrikson 等人[23]针对药物遗传学中的隐私研究提出的, 旨在利用神经网络模型根据患者的基因型和背景信息来辅助药物治疗。他们提供了一个通用的模型反演攻击算法(算法 1)。该算法通过在给定可用信息和模型的情况下估计潜在目标属性的概率, 来找到最接近目标属性的最优输入特征向量。该算法首先输入已知的患者人口属性统计信息 x 以及其对应的输出预测向量 $\boldsymbol{y}$, 根据这些可用信息和模型 f 找到与已知 x 相似的候选数据集 $\hat{X}$。然后, 向前遍历候选集, 将每个候选行 $\hat{x}$ 输入模型 f 中, 获得输出预测值 $\boldsymbol{y}$。最后, 根据已知的先验, 以及模型在每个候选行 $\hat{x}$ 上的输出与 x 的已知输出值的一致性程度, 对候选行进行加权。返回具有最大权重的目标属性, 即使用加权概率估计来计算接近目标特征向量的最优输入特征向量。攻击者可以利用该反演算法反演预测出患者的基因标记, 并且可以在给定可用信息的情况下最小化攻击者的预期误测率。这种攻击是在白盒攻击环境下进行的, 最终基因型预测的准确率最高达到 58%。

算法 1. Fredrikson 等人[23]的模型反演算法

1. 输入: $z_K = (x_1, \cdots, x_k, y), f, p_1, \cdots, d, y$.

2. 找到候选集 $\hat{X} \subseteq X, \forall \hat{x} \in \hat{X}$.

(a) $\hat{x}$ 满足 z_K 的已知属性: $1 \leqslant i \leqslant k, \hat{x}_i = x_i$.

(b) 对于 z_K 中的值, 利用模型 f 计算出 y, $f(\hat{x}) = y$.

3. 如果 $\left|\hat{X}\right| = 0$, 返回 $\perp$.

4. 返回使 $\sum_{\hat{x} \in \hat{X}: \hat{x}_t = x_t} \prod_{1 \leqslant i \leqslant d} p_i(\hat{x}_i)$ 最大化的 x_t .

然而, 上述的算法有各种限制, 模型反演攻击成功率并不是很高。Wu 等人[50]对 Fredrikson 等人[23]提出的模型反演攻击进行了改进, 通过研究差分隐私和稳定学习理论之间的联系, 得到了更好的隐私效用权衡。差分隐私是一种隐私保护机制, 它主要针对隐私保护中, 如何在分享数据时定义隐私, 以及如何在保证可用性的数据发布时, 提供隐私保护的问题[51]。将差分隐私技术用于模型反演攻击中, 改进的权衡使得到的差异私有模型更容易受到反演攻击, 同时使得增加的噪音更少。噪音少则意味着模型更容易反演。使得模型反演攻击的准确率从 36%提高到了 40%。

另外, Fredrikson 等人[23]的算法对于处理大型数据集的效果并不理想。Fredrikson 等人[32]也对这种模型反演攻击方法进行了改进, 基于优化的反演方法提出了一种新的模型反演攻击, 该攻击利用了预测结果的置信度信息, 采用了降噪和锐化滤波器作为模型反演攻击的先验。在人脸识别的大规模数据集上进行了验证, 可以仅凭目标人脸标签和机器学习模型访问权限重建出可识别的人脸图像。根据分类结果与目标的匹配程度, 找到使返回的置信度最大化的输入。算法 2 给出了他们针对人脸识别模型的反演攻击算法 MI-FACE。该算法首先为候选输入值 x 定义了一个代价函数 c, 基于人脸识别模型 $\tilde{f}$ 对 x 预测为目标类 $label$ 的置信值 $\tilde{f}_{label}(x)$ 而定义, 并初始化待优化的候选输入 x。然后, 在给定的迭代次数 α 次内通过梯度下降算法进行优化迭代, 使用大小为 λ 的梯度步长。在每一步梯度下降后, 得到的特征向量被赋予给一个后处理函数 Process, 该函数可以根据给定的攻击执行各种图像处理, 如降噪和锐化。如果代价在 β 次迭代中没有得到优化, 或者代价小于或等于了阈值 γ, 那么迭代终止并返回最佳的候选值, 即找到使目标类置信度最大化的输入值。

算法 2. Fredrikson 等人[32]的人脸识别模型的反演攻击算法

function MI-FACE(*label, α, β, γ, λ*)

$$c(x) \overset{def}{=} 1 - \tilde{f}_{label}(x)$$

$$x_0 \leftarrow 0$$

FOR $i \leftarrow 1 \cdots \alpha$ DO

$$x_i \leftarrow \text{Process}\left(x_{i-1} - \lambda \cdot \nabla c(x_{i-1})\right)$$

$$\text{IF}\ \ c(x_i) \geqslant \max\left(c(x_{i-1}),\cdots,c(x_{i-\beta})\right)$$
$$\text{BREAK}$$
$$\text{IF}\ \ c(x_i) \leqslant \gamma\ \text{THEN}$$
$$\text{BREAK}$$
$$\text{RETURN}\ \left[\arg\max_{x_i}\left(c(x_i)\right), \min_{x_i}\left(c(x_i)\right)\right]$$

Fredrikson 等人[32]后来提出的这种模型反演攻击方法虽然在攻击环境的限制和应用数据集规模等方面进行了改进, 但其评估攻击效果的方法是基于调查的方法, 需要大量的人工操作, 评估周期长且花费较昂贵, 同时因为是人工评价, 结果缺乏客观性。于是 Park 等人[41]对 Fredrikson 等人[32]提出的模型反演攻击的效果进行了评估, 提供了一种新的度量工具和性能度量, 也就是将模型反演攻击恢复的数据的可识别率以及攻击成功概率作为度量标准。Park 等人还提出了一种利用深度学习模型的评估攻击效果的方法, 可以量化模型反演攻击的有效性并能够转化为一个分类问题, 是一种更有效、更客观的评估方法。

4.1.2 基于神经网络可解释性的模型反演攻击

虽然神经网络已经成为一种有效的模型, 但其预测结果经常缺乏可解释性, 而有些预测结果在实际应用中是不可缺少的。为了能够更好地解释神经网络, 很多学者利用模型反演攻击技术对神经网络模型进行了模型反演, 能够得出中间层的信息, 可以很好地解释神经网络模型。

Mahendran 和 Vedaldi[15] 研究了图像表示问题, 计算图像表示的近似反演的方法可以表述为对于给定图像找到与其表示最匹配图像的问题, 这与基于优化的模型反演方法的基本原理是一致的。他们针对图像表示问题, 提出了一种通用的图像表示的反演方法, 该方法从随机噪声出发, 只使用图像表示和一般自然图像提供的先验信息作为初始数据, 因此初始数据只有图像表示本身包含的信息。将该反演技术应用到深层卷积神经网络(Convolutional Neural Networks, CNN)的分析中, 通过采样可以近似重构 CNN, 这对理解和解释 CNN 有很好的鲁棒性。同时也可探索 CNN 的不变性, CNN 在训练过程中逐渐建立了越来越多的不变性。

针对神经网络的可解释性, Du 等人[13]也提出了一种指导神经网络模型特征反演的框架, 框架如图 5 所示。首先, 该框架将原始输入 x_a 输入到卷积神经网络 CNN(左侧)中, 并计算并保存 CNN 每一层的表示。第二步, 通过与 CNN(右侧)的互动, 获得了不同层级的解释结果。然后, 指导特征反演 Φ 提取所有前景对象的位置, 利用 CNN 最后一层中目标类神经元的激活来反向微调反演结果。此外, 通过使用原始输入的中间层的激活结果作为掩码 m 来引入一个强正则化因子, 避免过拟合。该框架能够对基于神经网络的预测结果类别进行很好地解释, 同时所提出的框架还确定了每个特征在输入中的贡献, 通过在神经网络的输出层与目标类别神经元的交互, 加强了对结果类别的解释。在 ImageNet 和 Pascal VOC07 数据集上进行了有效的验证。

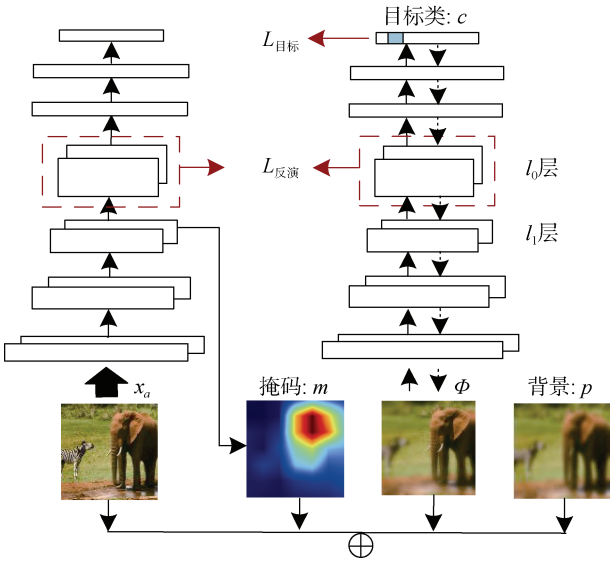

图 5 Du 等人的指导模型特征反演框架[13]

Figure 5 Du et al.'s framework of guided feature inversion[13]

4.1.3 针对预测数据的模型反演攻击

针对协作式深度神经网络模型, He 等人[37]设计了一种模型反演攻击方法, 旨在恢复输入到模型中的预测数据。无论是在单方机器学习系统还是多方机器学习系统中, 对预测数据隐私的研究都较少。因为恢复预测样本比训练样本更具挑战性, 然而, He 等人对于白盒攻击和黑盒攻击环境都提供了模型反演方法。对于白盒攻击, 利用基于优化的模型反演攻击技术, 用正则化最大似然估计恢复测试数据; 对于黑盒攻击, 则采用了基于训练的模型反演攻击技术, 训练与原模型相反的反演网络, 能够直接识别从输出到输入的逆映射, 而不需要获取模型信息。在 MNIST 和 CIFAR10 图像分类数据集上得到了验证, 均能反演出可识别的预测样本, 并有很高的保真度。

由于该方法既包含基于优化的方法也包含基于训练的方法, 所以在后面的基于训练的模型反演方法中将不再介绍。

4.1.4 基于生成式的模型反演攻击

针对模型反演攻击技术在深层神经网络上反演成功率较低的问题, Zhang 等人[33]提出了一种生成式模型反演(Generative Model Inversion, GMI)攻击方法, 该攻击的模型框架如图 6 所示。

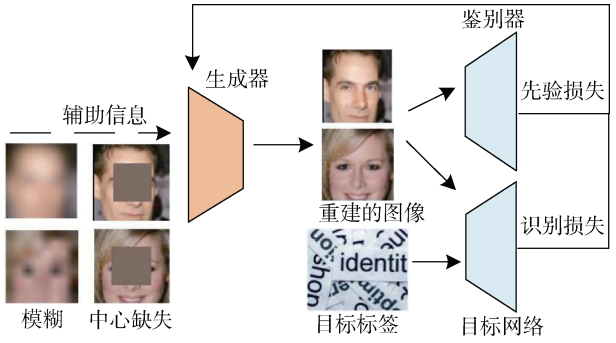

图 6 Zhang 等人的 GMI 模型框架[33]

Figure 6 Zhang et al.'s framework of GMI[33]

该攻击方法基于生成对抗网络(GAN)设计了端到端的模型反演攻击模型, 能够反演深度神经网络, 且具有很高的成功率。利用部分辅助信息, 通过生成对抗网络从公开数据集中学习训练数据的分布先验, 用来指导反演过程, 属于基于优化的方法。该方法主要解决两个优化问题, 一是鉴别器惩罚生成器生成不真实图像的先验损失优化, 还有鼓励生成的图像在目标网络下具有较高似然率的识别损失优化。实验表明, 对于从最好的人脸识别分类器中重建人脸图像, 该攻击方法比之前的方法提高了约 75% 的识别准确率。同时, 他们还证明了隐私保护技术中的差分隐私技术对该攻击并没有显示出明显的防御效果。

4.1.5 基于复杂神经网络的模型反演攻击

Fredrikson 等人[32]提出的模型反演攻击应用在卷积神经网络等复杂神经网络模型中, 并没有重建出可识别的图像, 表明攻击对复杂神经网络的攻击基本上是无效的。针对上述问题, Mejia 等人[34]通过改进深度神经网络模型的对抗性训练模型(Adversarial Training Model, ATM)语义表示, 实现了模型反演重构。一共提出了两种攻击方法, 第一种是基于梯度的攻击, 从输入空白的灰色图像开始反向传播优化图像, 同时改变图像的输入像素以最大化特定的输出类别。因为优化不会消除初始化时使用的所有随机性, 所以没有选择使用随机输入, 而是从空白图像开始, 使用同质图像输入生成更清晰的重建图像。他们还表示, 图像不一定要从灰色开始, 可以使用其他同类进行初始化, 但对于大多数其他类别, 灰色会得出最好的结果。第二种攻击基于 Google DeepDream 方法的原理[52], 在不同图像分辨率尺度上进行优化, 初始输入空白图像, 并将其按比例缩小, 使缩小的图像被放大到原始图像尺寸时, 能够有效模糊图像。此时优化这个低分辨率图像, 它就会以更高的分辨率被放大和重新优化, 即能够得到更清晰的重建图像。他们最后还提供了一种定量的方法来评估模型反演攻击的效果, 通过计算训练图像与重建出的图像的余弦相似度, 来计算模型反演攻击的成功率。

表 4 列出了以上介绍的文献研究中使用的基于优化的模型反演攻击方法对比, 其中的评价结果项中的“可识别”表示文献中没有明确说明具体的评价结果, 只是展示了部分反演结果, 例如对人脸识别数据集进行模型反演得到了清晰可识别的人脸图像, 或对图像分类数据集进行模型反演得到了可识别的输入图像, 可以进行人为图像分类等, 对于这些情况, 本文将其评价结果归为“可识别”一项。

4.2 基于训练的反演攻击技术

神经网络的反演问题实际上是一个很难的不适定问题。基于优化的反演攻击方法往往产生的图像并不像真正的自然图像, 尤其是对于大型神经网络, 其效果较差。此外, 这种方法在测试时涉及到优化, 需要计算梯度, 因此这种方法相对较慢。

与直接从 F_w 反演给定预测向量的基于优化的反演方法不同, 基于训练的反演攻击方法是首先训练原模型 F_w 的反演模型 G_θ, 该反演模型再将给定的预测向量 $F_w(x)$ 作为输入并输出重建的样本 $\hat{x}$。图 7 为基于训练的反演攻击技术的框架, 该方法框架类似于自动编码器的结构, 其中原模型 F_w 类似于编码器, 反演模型类似于解码器。在形式上, 基于训练的反演攻击方法是为了找到一个反演模型 G_θ 使重建样本 $\hat{x}$ 与训练样本 x 间的误差最小, 即目的是使以下公式最小化。

$$C(G_\theta)=E_{x\sim p_x}\left[R\left(G_\theta\left(F_w\right)\right),x\right] \tag{2}$$

其中 R 是采用 $L2$ 范数等方法的损失函数。

4.2.1 基于背景知识校准的模型反演攻击

针对神经网络的模型反演问题, Yang 等人[42]提出了基于背景知识校准的对抗性环境中的神经网络反演。对抗性环境下攻击者的目标是从模型的预测值来推断目标模型的训练数据和测试数据。他们提出了一种有效的模型反演方法, 基于训练一个与原模型相反的反演模型作为目标模型的逆来进行反演, 反演模型可以通过黑盒访问目标模型进行训练。论

表 4 基于优化的模型反演攻击方法总结

Table 4 Methods of optimization-based model inversion attack

方法/框架	相关论文	攻击场景	数据集	评价方法	评价结果	优缺点
基于置信度信息	[23]	白盒	IWPC	基因型预测准确率	58%	首次提出了模型反演攻击技术, 但模型局限性较大, 模型反演攻击成功率不高, 对大型数据集处理效果较差。
	[32]	黑盒	AT&T	人脸识别准确率	87%	加入了降噪和锐化滤波器作为模型的先验, 在黑盒场景下可恢复出识别率较高的人脸图像, 对于大规模数据集也有较好的效果, 但评估攻击效果为人工评价, 代价较大且缺乏客观性。
	[41]	黑盒	AT&T VGGFace2	人脸识别准确率	85% 82.5%	提出了一种基于深度学习模型的自动评估攻击效果的方法, 使评估方法更有效客观。
基于神经网络可解释性	[15]	黑盒	ILSVRC2012	图像分类准确率	91.5%	提出了一种通用的图像表示的反演方法, 对 CNN 的理解和解释有很好的鲁棒性。
	[13]	白盒	ImageNet Pascal VOC07	图像分类准确率	可识别	反演模型能够对预测结果类别进行有效解释, 并确定了每个特征在输入中的贡献。
针对预测数据	[37]	白盒	MNIST CIFAR10	图像分类准确率	可识别	对数量较少以及恢复难度较大的预测数据进行了模型反演, 用正则化最大似然估计恢复了预测数据。
基于生成式	[33]	黑盒	MNIST chestX-ray8 CelebA	图像分类准确率 人脸识别准确率	可识别	基于遗传算法进行端到端的模型反演攻击, 反演攻击成功率高。
基于复杂神经网络	[34]	黑盒	CIFAR10	图像分类准确率	可识别	改进了复杂神经网络的对抗性训练模型语义表示, 同时提出了一种定量的评估攻击效果的方法, 能够计算攻击成功率。

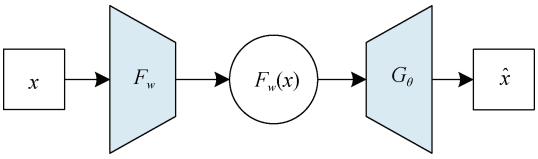

图 7 基于训练的反演攻击技术框架

Figure 7 Framework of training-based inversion attack

文提出了两种创新的关键技术, 第一种是利用攻击者的背景知识组成一个辅助集(一个更通用的数据集)来训练反演模型, 而该反演模型不需要访问原始训练数据。第二种是设计了一种基于截断的技术来校准反演模型, 截断过程可以理解为类似特征选择的过程, 以便能够有效地从攻击者对受害用户数据的部分预测中反演目标模型。该反演攻击方法的具体架构如图 8 所示, 他们利用截断技术将目标模型 F_w 对辅助样本 x 的预测向量 $F_w(x)$截断为与预测向量同一维数的 $trunc(F_w(x))$, 并将其作为反演模型 G_θ 的输入特征来进行模型训练, 最终得出辅助样本的重构样本 $G_\theta\left(trunc\left(F_w\left(x\right)\right)\right)$。该方法的目标是使反演模型 G_θ 重构出的结果 $G_\theta\left(trunc\left(F_w\left(x\right)\right)\right)$ 与原样本 x 的重构误差 $R\left(G_\theta\left(trunc\left(F_w\left(x\right)\right)\right),x\right)$ 最小化。

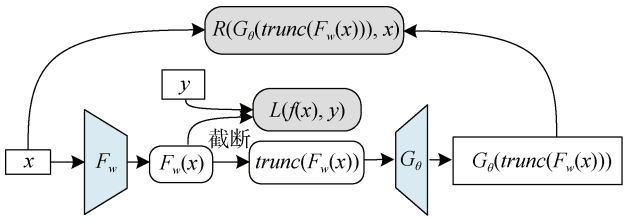

图 8 Yang 等人的模型反演攻击框架[42]

Figure 8 Yang et al.'s architecture of model inversion attack[42]

4.2.2 基于自回归密度模型的模型反演攻击

Nash 等人[47]提出了一种利用自回归密度估计模型来反演神经网络的方法。该方法训练生成式反演模型来表达以中间模型表示为条件的输入特征的分布, 是一种有监督表示的反演方法。该方法基于 PixelCNN 自回归神经密度模型进行模型反演, PixelCNN[53]是用于图像的自回归神经密度模型, 其使用卷积神经网络来参数化图像中每个子像素的条件

分布。按照从左到右, 从上到下的顺序, 一次对一个像素值进行采样。PixelCNN 及其变体是强大的图像模型, 目前其在自然图像的对数似然上的分数是最先进的。该文献主要关注图像的监督模型, 所以他们使用了 PixelCNN++的条件变体作为反演模型, 这种方法能够将参数化反演模型的负对数似然率降至最低, 使查询给定的输入与特定的表示很好地匹配。同时, 通过查看反演模型的样本, 可以深入了解监督模型学习的不变性, 即监督模型对于输入数据的某些变换其性能是不变的。他们使用这种方法, 检查了卷积神经网络不同层次上保存的信息类型, 并证明了不同结构的选择能够引起不变性。最后, 他们表示该方法不依赖于任何特定的密度模型, 并且可以应用于同类领域中的适当模型, 例如文本分类或语音识别任务中。

4.2.3 GMI: 通用模型反演框架

攻击者可以使用目标用户的非敏感属性和神经网络模型输出暴露的用户敏感属性值来进行模型反演攻击。然而, 在攻击前, 攻击者需要的目标用户的非敏感属性值也可能很难获取到。Hidano 等人[43, 44]在目标用户的非敏感属性对攻击者不可用的情况下, 实现了对模型反演攻击风险的量化。他们提出了一种针对神经网络预测的通用模型反演(General Model Inversion, GMI)的框架(如图 9 所示), 可以对攻击者可用的辅助信息量进行建模, 能够在不知道用户非敏感属性情况下执行。该框架在高层次上, 使用数据中毒方式将恶意数据注入到训练数据集中, 通过将当前预测模型 f_{cur}、系统的设定参数 par_{sys} 和一组恶意数据 $\{z_i=(x_z,y_z)\}_{i=1}^{N}$ 输入到中毒算法中, 并将预测系统的预测模型更新为某个目标预测模型 f_{tgt}, 最终输出 f_{tgt}, 即可在不知道非敏感属性时对神经网络进行反演攻击。对于模型反演算法, 在给定预测模型 f_{tgt}、用户的一些输出值 y 和一些辅助信息 aux 的情况下, 进行模型反演输出用户的敏感属性向量$(x_1, \cdots, x_T)$。这种攻击仅允许从神经网络模型的输出推断用户输入中的敏感属性。通过在实际数据集上的验证, 证明了该模型反演攻击的有效性。

4.2.4 i-RevNet: 深度可逆网络

众所周知, 卷积神经网络(CNN)取得了很大的成功。但是由于 CNN 是通过逐步丢弃掉一些不太重要的信息(比如池化层)来提取某些特征完成某一特定的任务[54], 所以, 在通常情况下, 从 CNN 提取的特征中恢复原图像的难度很大。Jacobsen 等人[48] 通过一对一映射证明了信息丢失并不是学习表象的必要条件, 表象可以很好的概括复杂的问题。他们通过同级层(homeomorphic)的级联, 构建了 i-RevNet, 该网络可以完全反演直到最终映射到类别上, 即没有丢弃任何信息, 保留了从输入层到输出层所有的信息。构建一个可逆的网络架构是困难的, 因为局部的反演是病态的[1], 他们通过提供一个显式的可逆结构来克服这一问题。该网络是在 RevNet 上进行改进, 使用可逆的层代替了原本不可逆的层。同时, 通过对 i-RevNet 学习到的表象的分析, 能够在一定程度上理解和解释 CNN。i-RevNet 也降低了由于信息缺失导致的网络变异性。

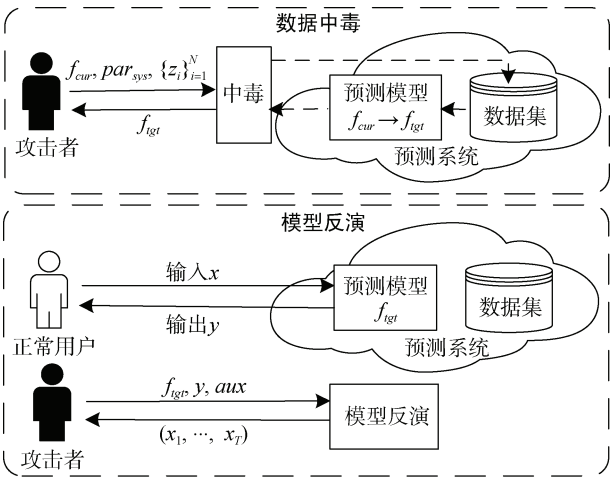

图 9 Hidano 等人的 GMI 框架结构[43, 44]

Figure 9 Hidano et al.'s framework of GMI[43, 44]

4.2.5 基于生成对抗网络的模型反演攻击

联邦学习是用于深度学习的移动边缘计算框架, 是隐私保护神经网络的最新进展。其中模型由客户端以分散的方式训练, 防止服务器直接访问来自客户端的私有数据。这种学习机制极大地防御了来自服务器端的攻击。尽管结合生成对抗网络(GAN)的先进攻击技术可以重构全局数据分布类别的代表性数据(即对于所有客户数据), 但要攻击特定客户数据(即用户级隐私泄露)仍然非常困难, 精确恢复特定客户的隐私数据仍是一项巨大的挑战。

Wang 等人[36]从恶意服务器的角度提出了一种针对联邦学习的基于生成对抗网络(GAN)的模型反演攻击, 对联邦学习造成的用户级隐私泄露进行了研究, 提出了一种将 GAN 与多任务鉴别器相结合的通用攻击框架, 将原始的 GAN 调整为同时考虑目标客户端的真实性、类别和身份的多任务场景, 并将该框架命名为多任务辅助识别 GAN(multi-task GAN for Auxiliary Identification, mGAN-AI)。该框架采用黑盒攻击方法, 从客户端可访问的更新中估计输入

样本的类别、真实性和客户身份, 新的客户身份鉴别使生成器能够恢复指定用户的私有数据, 从而实现客户级别的隐私恢复。实现客户身份识别的关键是从每个单独的客户那里获取数据代表, 这样 GAN 的训练就可以由数据代表来监督, 生成具有特定身份的样本(即来自特定客户的样本)。由于客户端数据是不可访问的, 他们从客户端可访问的更新中估计此类数据的代表性。该框架在训练 GAN 时执行额外的任务提高了合成样本的质量, 同时不需要修改模型或使联邦学习性能下降, 实现了隐藏攻击。

在 MNIST 和 AT&T 数据集上, 最好分类准确率达到了 95%以上, 且 mGAN-AI 成功地恢复了特定用户的样本。

图 10 给出了该框架的总体原理图, 假设有 N 个客户端, 第 v 个客户端为受害者。第 t 次迭代后的模型为 M_t, u_t^k 表示来自第 k 个客户端的相应更新。在恶意服务器上, 基于来自受害者的更新 u_t^v 、模 M_t 和来自每个客户端的代表 X_k、X_v 来训练鉴别器 D 和生成器 G。X_{aux} 表示用于训练 D 进行鉴别任务的辅助真实数据集。

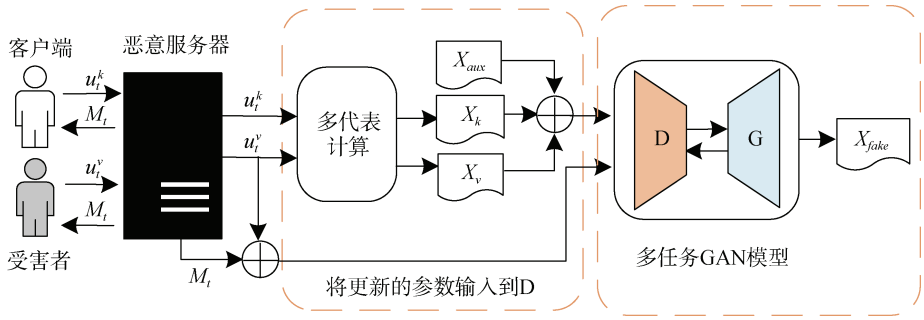

图 10 Wang 等人的 mGAN-AI 框架[36]

Figure 10 Wang et al.'s architecture of mGAN-AI[36]

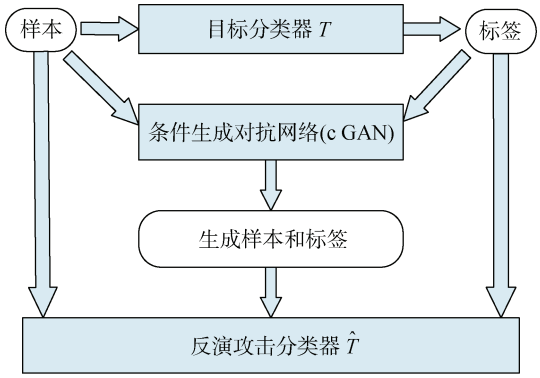

图 11 Shi 等人的基于 cGAN 的黑盒反演攻击[46]

Figure 11 Shi et al.'s cGAN-based black-box inversion attack[46]

Shi 等人在黑盒攻击场景下设计了一种针对深度神经网络分类器的模型反演攻击[46], 如图 11 所示。攻击者通过收集目标分类器 T 中使用的样本集和输出的标签集, 反演出攻击分类器 $\hat{T}$ 来预测目标分类器的分类结果, 为了降低对收集数据量的要求, 他们利用条件生成对抗网络(conditional Generative Adversarial Network, cGAN)来生成更多的样本和标签, 并将所有的数据用于训练反演攻击分类器, 能够使反演出的分类器预测目标分类的准确率达到 93%。

深度学习模型通常以集中方法训练, 所有数据由相同的训练算法处理。然而, 集中式方法迫使多个参与者将他们的数据集汇集到一个大型的中央训练集进行训练, 如果数据中包含用户的隐私数据, 则集中式服务器将能够访问并利用这些隐私敏感信息。针对这个问题, 近几年提出了协作深度学习模型, 允许各方在本地训练自己的深度学习模型, 并且只共享模型参数的子集, 无需共享各自的训练集, 以保证各自训练集的私有性。

Hitaj 等人[49]提出了一种在协作环境下针对深度神经网络模型的模型反演攻击, 保护隐私的协作式深度学习模型容易受到这种反演攻击, 导致无法再保护各个参与者的训练集。该攻击方法利用了学习过程的实时特性, 允许攻击者训练生成对抗网络(GAN)生成目标训练集的原始样本, 而这些样本本应是私有的。理想情况下, GAN 生成的样本与原始训练集会有相同的分布。这种攻击可能导致任何作为内部人员的用户都可以从受害者的设备中推断出敏感信息。攻击者只需运行协作学习算法并利用 GAN 重建存储在受害者设备上的敏感信息。该方法对卷积神经网络也能进行有效地反演攻击。通过实验, 他们也证明了应用差分隐私保护技术对该攻击是无效的。

针对深层神经网络, 模型反演攻击通常会返回无法识别的结果, 这些结果对于攻击者都是无用的。

Basu 等人在白盒攻击环境下提出了一种基于生成对抗网络(GAN)的模型反演攻击[39], 该方法针对分类问题, 在基于置信度的模型反演攻击方法的基础上, 获得各种类的高置信度表示。同时, 在这种环境下, 攻击者可以访问模型, 且对分类问题有一定的了解, 利用 GAN 来生成与单个类的训练样本相似的代表性样本, 形成约束搜索空间, 来指导搜索有代表性和可识别的样本, 最终可恢复目标模型的代表性样本。例如对于攻击面部识别系统, 可以从互联网下载不同的面部集合, 通过将 GAN 的输出连接到模型的输入, 同时可以使用优化技术来搜索使标签置信值最大化。在 MNIST 数据集上进行实验, 对选定的类别进行反演攻击, 能够反演出清晰可识别的代表性样本。

4.2.6 基于深度神经网络的模型反演攻击

构建一个分类器是昂贵和耗时的, 因为首先需要收集训练数据(例如, 爬虫), 然后选择合适的机器学习算法(需要大量的测试和特定领域知识), 以及优化模型超参数(需要对分类器结构有良好理解)。而所有这些信息, 即训练数据、机器学习分类器类型、模型结构和超参数, 对于在线分类器来说通常都是私有的, 不会被公开。Shi 等人[40]提出了一种基于深度神经网络的模型反演攻击方法, 利用深度神经网络在黑盒攻击场景下, 反演得出与训练模型功能等价的模型, 以此来窃取在线机器学习分类器, 攻击步骤如图 12 所示。

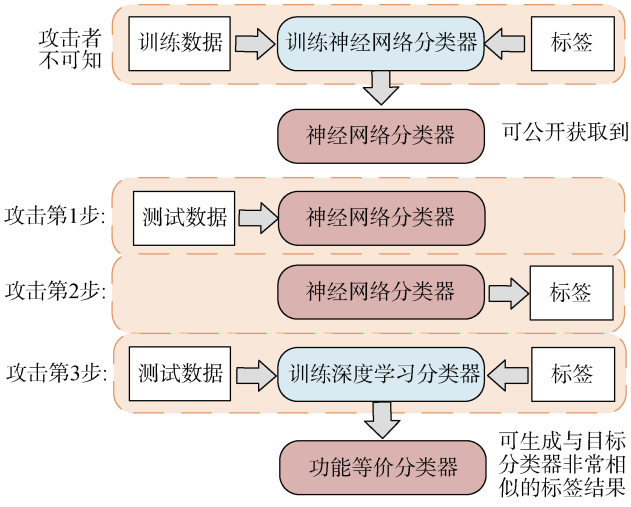

图 12 Shi 等人的窃取在线神经网络分类器的步骤[40]

Figure 12 Shi et al.'s steps to steal an online neural network classifier[40]

该攻击首先向在线的神经网络分类器输入一些提前准备的测试数据, 攻击者并不了解分类器的类型、结构及参数, 只能以黑盒的方式进行攻击。然后, 获取该在线分类器对于测试数据的输出标签。利用这些测试数据和输出标签来训练深度神经网络以构建与在线分类器功能等价的分类器。论文在文本分类应用上用深度神经网络推断朴素贝叶斯分类器和支持向量机分类器的功能。对于朴素贝叶斯分类器, 构建的分类器与在线分类器的分类误差仅为 2.10%; 对于支持向量机, 误差为 2.56%。这种新的利用深度学习进行攻击的模式给在线机器学习算法带来了新的安全挑战。

针对深度神经网络等复杂模型, Aïvodji 等人[45]提出了一种黑盒攻击环境下的模型反演攻击框架, 命名为生成性对抗模型反演 (Generative Adversarial Model Inversion, GAMIN), 该框架在对卷积神经网络等深层模型进行攻击时, 能以合理的消耗代价取得显著的反演效果。GAMIN 参考了生成对抗网络原理, 在不知道目标模型或原始数据分布的情况下, 基于对被攻击的目标模型的反演模型和生成器的持续训练, 来生成与目标模型的训练数据类似的输入数据, 该生成器通过模仿生成对抗网络的训练过程来生成所需数据。该攻击针对用作图像分类器的各种神经网络进行了验证。在 MNIST 数据集上进行实验时, GAMIN 能够生成近 60%的可识别数字。

表 5 列出了以上介绍的文献研究中使用的基于训练的模型反演攻击方法对比, 其中评价结果中的“可识别”一项与表 4 的数据意义相同。

与基于优化的反演方法相比, 基于训练的反演方法在反演模型的训练过程成本较高, 但是只需要消费一次。根据给定的预测进行重建只需要通过网络进行一次正向传递, 因此, 基于训练的反演方法相对速度较快。

4.3 针对模型反演攻击的防御方法

以上两小节对目前的基于神经网络的模型反演攻击技术进行了分类总结, 对其研究思路, 实现方法, 性能评价以及优缺点进行了比较。事实上, 不同类型的攻击对应的防御措施也不一样, 目前保护神经网络模型免受模型反演攻击的研究也有很多[1, 10, 18, 24-27, 30, 35, 41, 54-70], 本小节将简要介绍这些防御工作。

(1) 添加噪声

模型反演攻击技术存在不适定问题(ill-posed problem), 即训练数据的微小变化就能导致完全不同的模型反演质量。可利用模型反演攻击技术的这一问题研究针对性的防御方法。

Wu 等人[35]在噪声环境下发现了一种“可逆性干扰”现象, 即一个高度可逆的模型通过添加少量的噪声就能很快变成高度不可逆, 也就是模型反演攻击

表 5 基于训练的模型反演攻击方法总结

Table 5 Methods of training-based model inversion attack

方法/框架	相关论文	攻击场景	数据集	评价方法	评价结果	优缺点
基于背景知识校准	[42]	黑盒	FaceScrub	人脸识别准确率	85.7%	用攻击者的背景知识组成辅助集训练反演模型, 同时提出了基于截断的方法来校准反演模型, 极大提高了黑盒攻击场景下的攻击成功率。
			MNIST	图像分类准确率	99.6%	
基于自回归密度	[47]	黑盒	MNIST	图像分类准确率	99.6%	基于 PixelCNN 自回归密度估计模型来反演神经网络, 可深入监督模型学习的行为和特征。
			SVHN	图像识别准确率	93.3%	
			CIFAR10	图像分类准确率	81.6%	
GMI	[43, 44]	黑盒	MovieLens 1M	电影评分准确率	可识别	提出了一种通用模型反演 GMI 框架, 可以对攻击者可用的辅助信息进行建模, 引入机器学习模型可以在不知道非敏感属性的情况下反演模型。
i-RevNet	[48]	白盒	ILSVRC2012	图像分类准确率	77%	提出了一种深度可逆网络, 通过同级层级联, 该模型可完全反演直到最终映射到类别上, 保留了从输入层到输出层所有的信息。
基于生成对抗网络	[36]	黑盒	MNIST AT&T	图像分类准确率 人脸识别准确率	95%	将 GAN 与多任务鉴别器相结合, 提出了一种通用攻击框架 mGAN-AI, 可以从客户端的更新中估计输入样本的类别、真实性和客户身份。
	[49]	黑盒	MNIST AT&T	图像分类准确率 人脸识别准确率	可识别	可在协作环境下训练生成对抗网络, 得到目标训练集的原型模型。
	[39]	白盒	MNIST	图像分类准确率	可识别	利用攻击者的背景知识指导搜索有代表性的样本和可识别的样本, 能够针对深层神经网络进行反演。
GAMIN	[45]	黑盒	MNIST	图像分类准确率	60%	基于对被攻击的目标模型的反演模型和生成器进行持续训练, 能够生成与输入数据类似的数据。
针对测试数据	[37]	黑盒	MNIST CIFAR10	图像分类准确率	可识别	在黑盒环境下, 对恢复难度较大的测试数据进行了反演, 反演出了可识别的测试样本。

技术的不适定问题。这种现象可以帮助防御模型反演攻击, 如果神经网络模型的可逆性很低, 那么可以使用较少的噪声就能有效地防御模型反演攻击, 而且不会降低太多模型的性能。

采用添加噪声的方法也出现在了其他研究中, 同样证明该方法能有效抵御模型反演攻击技术。例如, Wu 等人[10]在医学图像分类研究下, 提出了一种新的随机梯度下降(Stochastic Gradient Descent, SGD)方案, 称为患者隐私保护 SGD(Patient Privacy Preserving SGD, P3SGD), 该方案通过在每个患者数据的基础上进行一次大步更新来在患者级别下执行 SGD 的模型更新。在更新中注入了噪声, 以保护卷积神经网络的隐私, 同时还设计了自适应可控制的注入噪声策略。他们还对差分隐私下的隐私预算进行了严格分析, 证明了该方案能够降低模型过拟合, 同时与使用非隐私 SGD 训练的模型相比, 使用 P3SGD 训练的模型具有更好的抗模型反演攻击的能力。

同样, Zhang 等人[18]也采用添加噪声的方法, 提出了一种保护神经网络训练数据隐私的方法, 他们引入了一个模糊函数, 将其应用于训练数据, 用来训练模型。该函数将随机噪声添加到现有的样本或使用新样本增加数据集, 从而使关于单个样本的属性或一组样本的统计属性等敏感信息被隐藏。同时, 从混合后的数据集训练出的模型仍能达到较高的性能。通过实验证明, 该方法可以有效地抵御模型反演攻击。

添加噪声的防御方法本质上是在模型训练过程中引入随机性, 以使输出结果与真实结果具有一定程度的偏差, 从而有效防御模型反演攻击。该方法易于实现, 而且通常能获得较好的效果, 易于与其他方法结合使用。

(2) 差分隐私技术

差分隐私是一种广泛使用的隐私保护机制, 前文提到 Wu 等人[50]应用差分隐私技术改进神经网络模型, 使得模型更容易受到反演攻击。同时, 它作为

隐私保护和数据保护技术, 也能用来防御模型反演攻击。

Park 等人[41]针对神经网络模型的模型反演攻击, 研究了如何在保持模型效用的前提下, 应用差分隐私技术来对抗这种攻击。他们利用差分隐私随机梯度下降算法来加强神经网络模型的差分隐私。通过对差分隐私神经网络模型进行模型反演攻击, 证明了与非隐私模型相比, 差分隐私可以显著降低攻击概率。

Chen 等人[24]提出了一种基于差分隐私自动编码器的生成模型(Differentially Private Autoencoder-based Generative Model, DP-AuGM)。该模型以不同的差分隐私方式针对隐私数据进行训练, 并能为之后的学习任务生成新的数据。使用该模型发布差分隐私合成数据, 所生成的数据保留了私有数据的统计数据隐私, 从而来提高深度神经网络的学习效率, 并保证数据的隐私性和高效性。实验评估了 DP-AuGM 的健壮性, 并表明该模型能够有效抵抗模型反演攻击。

杨烨[67]研究了跨数据集的神经网络模型的训练, 针对多方联合模型训练的问题, 设计了一种隐私保护的模型训练算法, 各参与方在本地使用统一模型训练, 使用秘密共享技术对关键参数进行加密, 并实现了第三方对多个参与方的加密参数添加噪声, 集中式的参数处理提高了最终各方模型的准确率, 以及实现了添加噪声的统一和可控性, 且使最终训练模型对模型反演攻击有充分的鲁棒性。通过仿真实验说明了该方案在选取不同规模差分隐私噪声时的表现, 证明了该算法的有效性。

添加噪声的方法一定程度上也属于差分隐私技术, 但差分隐私技术具有严格的数学理论支撑。虽然差分隐私能够提供理论上的隐私保证, 但并不能保证分类器的分类准确性。同时, 差分隐私技术对于简单的机器学习模型, 较容易实现。然而, 对于结构复杂、参数量大的深度学习模型而言, 则难以平衡模型性能和隐私保护效果。

(3) 同态加密

同态加密(Homomorphic Encryption, HE)是通过数据加密提供数据隐私的另一种技术。同态加密是一种不需要访问数据本身就可以处理数据的密码学技术, 其对密文进行代数运算, 获得的结果也是加密的。同态加密计算时不使用私钥解密, 具有以下两个优点: (1)可以对密文块进行任何类型的计算; (2)对密文块上的计算输出解密后的结果与使用相同运算符对相应明文块上的计算结果相同。因此, HE 特别适合用于云环境下的数据安全和隐私保护。在 HE 的基础上, 许多研究者致力于研究安全多方计算、全 HE 数据分类、分布式 k-均值聚类算法和处理加密数据的神经网络。不管现有的密码机制如何, 减少学习模型 API 的敏感输出是确保数据安全和隐私的一种新想法[54]。

近年来, 提出了很多基于同态加密的多方计算方法[19]。Gilad-Bachrach 等人[69]提出了一种可应用于加密数据的近似神经网络 CryptoNets, 并在 MNIST 手写识别数据集上进行了测试, 有效地实现了 99% 的分类性能。Xie 等人[68]提出了在测试阶段的隐私防御方法。采用同态加密对数据进行加密, 目的是在训练神经网络时不对数据进行解密, 从而提供了单输入的保密性。

同态加密方法非常适合多方参与、共同训练模型的情况。同时, 同态加密基于密码学能够保证计算结果的正确性, 但该方法通常非常依赖于函数的复杂度。对于存在大量非线性计算的深度学习模型, 该方法的计算开销将非常高, 这也是其难以在实际中应用的主要原因。

(4) 其他方法

联邦学习是机器学习模型的一种训练模式, 与传统的集中学习训练不同, 其将训练数据分布于多个节点来共同执行一个训练任务。各个节点在获得中心模型的副本后独立训练, 并将训练后更新的模型参数上传至中心节点。中心节点将所有上传的参数整合至中心模型, 并再次将模型分发出去, 如此迭代, 直至中心模型收敛。联邦学习目的是让各个节点的数据保留在本地, 来降低数据隐私泄露的风险。通过这种方式保护数据隐私来防御模型反演攻击。Triastcyn 等人[63]提出了一种联邦学习环境下的隐私保护数据框架, 命名为联合生成隐私(Federated Generative Privacy, FedGP), 他们利用联合平均(Federated Averaging, FedAvg)算法训练生成对抗网络的生成器, 提取隐私保护的人工数据样本, 并对信息泄露风险进行实证评估。通过实验证明了 FedGP 能够生成高质量的标记数据, 从而成功训练和验证了鉴别器。最后证明了该方法能够显著降低这类模型对模型反演攻击的攻击成功率。不过, 基于联邦学习的防御方法仍处于研究的起步阶段, 在理论和应用等方面仍面临许多问题及挑战。

Yang 等人[70]提出了一种净化框架来防御模型反演攻击, 通过减小目标分类器预测的置信度向量的离散度来“净化”模型。他们将训练的目标模型的预测置信度作为输入, 并利用该净化框架减小置信

度向量间的离散度。降低离散度将有助于降低预测向量对输入数据变化的敏感度, 也就是不同的输入数据, 输出的置信度得分向量将不会发生较大变化, 能够降低输入数据与置信度得分之间的相关性。因此, 将会使针对这种分类器的模型反演攻击不能从置信度得分向量中得出有关输入数据的准确信息, 从而能够有效防御模型反演攻击。同时, 该方法也能够保留预测的有用信息, 对原始置信度得分的失真可以忽略不计。该方法可以使模型反演的误差增加 4 倍, 并且能够保证分类精度下降不到 0.4%, 置信度得分失真不超过 5.5%。

针对分布在不同物理组织中的数据, 模型平均可以在分布式数据上训练深层神经网络模型, 并且与在集中式数据上训练模型相比, 模型平均能够提供更具竞争力的性能。但模型的中间参数是在训练过程中传递的, 不能防止反演攻击。Fu 等人[26]提出了一种基于混合的模型平均方法, 通过聚合局部估计来获得全局估计, 并且通过选择合适的超参数, 全局估计提供了更好的性能。可以有效增强数据的隐私性, 能够提供对反演攻击的防御。

表 6 总结了以上介绍的不同防御方法。针对神经网络的模型反演攻击技术的防御方法还有很多, 本文仅选取了部分有代表性的方法进行了介绍, 感兴趣的读者可以在本文参考文献中查阅更多的方法。

表 6　针对模型反演攻击的防御方法总结

Table 6　Defenses against model inversion attack

防御方法	方法原理	特点	相关论文
添加噪声	模型反演攻击技术存在不适定问题。在模型训练过程中引入随机性, 以使输出结果与真实结果具有一定程度的偏差, 从而有效防御模型反演攻击。	易于实现, 通常能获得较好的效果, 易于与其他方法结合使用。	[10] [18] [35]
差分隐私技术	利用差分隐私在训练数据、模型算法、目标函数中等添加噪声, 能够提供理论上的隐私保证。	具有严格的数学理论支撑, 能够提供理论上的隐私保证, 但并不能保证分类器的分类准确性。难以平衡模型性能和隐私保护效果。	[24] [41] [50] [67]
同态加密	通过数据加密提供数据隐私。不需要访问数据本身就可以处理数据。对密文进行代数运算, 其结果也是加密的。	基于密码学能够保证计算结果的正确性, 但非常依赖于函数的复杂度, 计算开销通常非常高, 难以在实际中应用。适合多方参与、共同训练模型的情况。	[68] [69]
联邦学习	将多方训练数据保留在本地, 并在本地独立训练模型。通过保护数据隐私来防御模型反演攻击。	让各个节点的数据保留在本地, 来降低数据隐私泄露的风险。但仍处于研究的起步阶段, 在理论和应用等方面仍面临许多问题及挑战。	[63]
净化框架	通过减小目标分类器预测的置信度向量的离散度来“净化”模型。	能够降低预测向量对输入数据变化的敏感度, 降低输入数据与置信度得分之间的相关性。同时, 能够保留预测的有用信息、保证分类精度。	[70]
基于混合的模型平均方法	在分布式数据上训练深层神经网络模型, 通过聚合局部估计来获得全局估计, 通过增强数据的隐私性来防御模型反演攻击。	能够防御针对深层神经网络模型的模型反演攻击。可以有效增强数据的隐私性。	[26]

5　未来研究方向

神经网络的模型反演攻击方法在近些年已经取得了一些成果, 但仍然有很多没有克服的问题和值得深入研究的方向。

(1) 复杂深度神经网络

虽然 Fredrikson 等人[32]提出的模型反演方法能够用于复杂神经网络问题, 但是搜索空间呈指数级增长, 该方法虽有一定的效果, 但在实际场景中, 网络的规模和复杂性往往非常大, 目前的绝大部分方法都不适用于复杂的神经网络, 这一问题仍需要进一步的研究。

(2) 神经网络的对抗性攻击

神经网络在各项任务中大放异彩, 随之而来的安全问题也越来越受到关注。而设计安全学习算法需要对算法的安全性、泛化性能和开销进行联合优化。一般来说, 安全性越高, 学习算法的开销越大, 甚至泛化性能越差, 这对算法的设计以及应用提出了挑战。神经网络学习算法的安全性, 维护用户的隐私安全是未来发展的一个重要方向, 而随着模型反演方法的发展, 对手的对抗性攻击技术也会越来越强, 能否设计出在安全性、泛化性能和开销三个方面都相对平衡的防御对策, 仍需要未来进一步的研究。

(3) 动态非线性网络

在实际场景中, 网络往往具有动态性, 动态非线性系统, 如具有临时调整权重的神经网络, 具有随时间变化的映射。此类系统模型的反演目前没有深入的研究, 对于模型反演来说应该很复杂, 需要未来进一步的研究, 能否设计一种反演方法兼顾所有的复杂因素, 也是一个值得探讨的问题。

(4) 神经网络的可解释性

深度学习模型的可解释性与可视化一直是深度学习领域备受关注的方向, 目前许多反演方法都被应用到理解和解释神经网络中, 并能够给出一些解释, 但是如何从理论上说明神经网络能取得显著成果仍然是一个没有解决的问题。

(5) 生成对抗网络

目前越来越多的模型反演攻击技术与生成对抗网络(GAN)相结合来进行攻击, 尤其是生成对抗网络在对于训练数据不足的情况, 能够生成大量数据, 从而解决这种情况。未来我们仍需要考虑如何利用 GAN 的连接空间。目前对 GAN 的研究表明, 有可能生成与训练数据没有明显关联的可解释图像。例如, GAN 可用于生成不属于训练数据的真实图像。对于面部识别等应用, 这对于创建受约束的 GAN 空间进行搜索非常有用。即使来自训练数据的特定脸部不存在于攻击者用来创建 GAN 的数据中, 所得到的具有丰富脸部集合的连接的 GAN 空间也可能包含与用于训练原始神经网络模型的人脸图片足够接近的数据。

(6) 模型反演攻击的性能评估

虽然目前对神经网络的模型反演攻击技术的性能评估已经有了一些方法, 其中也有一些量化方法, 但目前的评估方法普遍趋于简单, 仍需要做更多细致的工作来解决这些隐私攻击成功的量化问题, 以便能够区分攻击是否提取了数据集中的特定训练样本或趋势。此外, 更好的评估指标将允许开发更高级的攻击和防御。

6 总结

神经网络模型是当前机器学习和人工智能兴起的核心技术, 随着它被成功的应用到图像识别、语音控制等热门领域中, 神经网络模型的安全问题逐渐成为新的研究热点。本文首次综合性地介绍了目前神经网络模型中的反演攻击技术, 分类整理了现有攻击技术的研究思路、实现方法、性能评价以及优缺点。对各种攻击技术进行分析比较, 并概述了典型的防御方法。最后, 通过引用的参考文献为本课题的研究指明了更广阔的前景。

参考文献

[1] Liu Q, Li P, Zhao W T, et al. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View[J]. *IEEE Access*, 6: 12103-12117.

[2] Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 308-318.

[3] Alzantot M, Balaji B, Srivastava M. Did You Hear That? Adversarial Examples Against Automatic Speech Recognition[EB/OL]. 2018: arXiv: 1801.00554. https://arxiv.org/abs/1801.00554.

[4] Veale M, Binns R, Edwards L. Algorithms that Remember: Model Inversion Attacks and Data Protection Law[J]. *Philosophical Transactions Series A*, *Mathematical*, *Physical*, *and Engineering Sciences*, 2018, 376(2133): 20180083.

[5] Al-Rubaie M, Chang J M. Privacy-Preserving Machine Learning: Threats and Solutions[J]. *IEEE Security & Privacy*, 2019, 17(2): 49-58.

[6] Riazi M S, Darvish Rouani B, Koushanfar F. Deep Learning on Private Data[J]. *IEEE Security & Privacy*, 2019, 17(6): 54-63.

[7] Jagwani P, Kaushik S. Privacy in Location Based Services: Protection Strategies, Attack Models and Open Challenges[M]. Information Science and Applications 2017. Singapore: Springer Singapore, 2017: 12-21.

[8] Auernhammer K, Kolagari R, Zoppelt M. Attacks on Machine Learning: Lurking Danger for Accountability[C]. *AAAI Workshop on Artificial Intelligence Safety*, 2019.

[9] Zhao J W, Chen Y F, Zhang W. Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions[J]. *IEEE Access*, 7: 48901-48911.

[10] Wu B Z, Zhao S W, Sun G Y, et al. P3SGD: patient privacy preserving SGD for regularizing deep CNNs in pathological image classification[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 2094-2103.

[11] Nguyen T N. Attacking Machine Learning Models as Part of a Cyber Kill Chain[EB/OL]. 2017: arXiv: 1705.00564. https://arxiv.org/abs/1705.00564.

[12] Dosovitskiy A, Brox T, Processing C A. Inverting visual representations with convolutional networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 4829-4837.

[13] Du M N, Liu N H, Song Q Q, et al. Towards Explanation of DNN-Based Prediction with Guided Feature Inversion[C]. *The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018: 1358-1367.

[14] Gilbert A C, Zhang Y, Lee K, et al. Towards Understanding the Invertibility of Convolutional Neural Networks[EB/OL]. 2017: arXiv: 1705.08664. https://arxiv.org/abs/1705.08664.

[15] Mahendran A, Vedaldi A, Processing C A. Understanding deep image representations by inverting them[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 5188-5196.

[16] Liu J X, Meng X F. Survey on Privacy-Preserving Machine Learn-

ing[J]. *Journal of Computer Research and Development*, 2020, 57(2): 346-362.
(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.)

[17] Zhao Z D, Chang X L, Wang Y X. A Survey of Privacy Preserving in Machine Learning[J]. *Journal of Cyber Security*, 2019, 4(5): 1-13.
(赵镇东, 常晓林, 王逸翔. 机器学习中的隐私保护综述[J]. *信息安全学报*, 2019, 4(5): 1-13.)

[18] Zhang T W, He Z C, Lee R B. Privacy-Preserving Machine Learning through Data Obfuscation[EB/OL]. 2018: arXiv: 1807.01860. https://arxiv.org/abs/1807.01860.

[19] Yu Y C, Liu X Y, Chen Z N. Attacks and Defenses towards Machine Learning Based Systems[C]. *The 2nd International Conference on Computer Science and Application Engineering*, 2018: 1-7.

[20] Xu X Z, Cao D, Zhou Y, et al. Application of Neural Network Algorithm in Fault Diagnosis of Mechanical Intelligence[J]. *Mechanical Systems and Signal Processing*, 2020, 141: 106625.

[21] Zhao C W. A Survey on Artificial Neural Networks[J]. *Shanxi Electronic Technology*, 2020(3): 94-96.
(赵崇文. 人工神经网络综述[J]. *山西电子技术*, 2020(3): 94-96.)

[22] Duan Y S. Literature Review of Artificial Neural Networks[J]. *Technology Wind*, 2011(5): 185.
(段玉三. 人工神经网络文献综述[J]. *科技风*, 2011(5): 185.)

[23] Fredrikson M, Lantz E, Jha S, et al. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing[C]. *The 23rd USENIX conference on Security Symposium*, 2014: 17-32.

[24] Chen Q R, Xiang C, Xue M H, et al. Differentially Private Data Generative Models[EB/OL]. 2018: arXiv: 1812.02274. https://arxiv.org/abs/1812.02274.

[25] Gylberth R, Adnan R, Yazid S, et al. Differentially private optimization algorithms for deep neural networks[C]. *2017 International Conference on Advanced Computer Science and Information Systems*, 2018: 387-394.

[26] Fu Y W, Wang H M, Xu K L, et al. Mixup based privacy preserving mixed collaboration learning[C]. *2019 IEEE International Conference on Service-Oriented System Engineering*, 2019: 275-2755.

[27] Alves T A O, França F M G, Kundu S. MLPrivacyGuard: Defeating Confidence Information Based Model Inversion Attacks on Machine Learning Systems[C]. *The 2019 on Great Lakes Symposium on VLSI*, 2019: 411-415.

[28] Chakraborty A, Alam M, Dey V, et al. Adversarial Attacks and Defences: A Survey[EB/OL]. 2018: arXiv: 1810.00069. https://arxiv.org/abs/1810.00069.

[29] Papernot N, McDaniel P, Sinha A, et al. SoK: security and privacy in machine learning[C]. *2018 IEEE European Symposium on Security and Privacy*, 2018: 399-414.

[30] Qiu S L, Liu Q H, Zhou S J, et al. Review of Artificial Intelligence Adversarial Attack and Defense Technologies[J]. *Applied Sciences*, 2019, 9(5): 909.

[31] Chang S, Li C. Privacy in Neural Network Learning: Threats and Countermeasures[J]. *IEEE Network*, 2018, 32(4): 61-67.

[32] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]. *The 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015: 1322-1333.

[33] Zhang Y H, Jia R X, Pei H Z, et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks[EB/OL]. 2019: arXiv: 1911.07135. https://arxiv.org/abs/1911.07135.

[34] Mejia F A, Gamble P, Hampel-Arias Z, et al. Robust or Private? Adversarial Training Makes Models more Vulnerable to Privacy Attacks[EB/OL]. 2019: arXiv: 1906.06449. https://arxiv.org/abs/1906.06449.

[35] Wu X, Fredrikson M, Jha S, et al. A methodology for formalizing model-inversion attacks[C]. *2016 IEEE 29th Computer Security Foundations Symposium*, 2016: 355-370.

[36] Wang Z B, Song M K, Zhang Z F, et al. Beyond inferring class representatives: User-level privacy leakage from federated learning[C]. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019: 2512-2520.

[37] He Z C, Zhang T W, Lee R B. Model Inversion Attacks Against Collaborative Inference[C]. *The 35th Annual Computer Security Applications Conference*, 2019: 148-162.

[38] Yang Z Q, Chang E C, Liang Z K. Adversarial Neural Network Inversion via Auxiliary Knowledge Alignment[EB/OL]. 2019: arXiv: 1902.08552. https://arxiv.org/abs/1902.08552.

[39] Basu S, Izmailov R, Mesterharm C. Membership Model Inversion Attacks for Deep Networks[EB/OL]. 2019: arXiv: 1910.04257. https://arxiv.org/abs/1910.04257.

[40] Shi Y, Sagduyu Y, Grushin A, et al. How to steal a machine learning classifier with deep learning[C]. *2017 IEEE International Symposium on Technologies for Homeland Security*, 2017: 1-5.

[41] Park C, Hong D, Seo C. An Attack-Based Evaluation Method for Differentially Private Learning Against Model Inversion Attack[J]. *IEEE Access*, 7: 124988-124999.

[42] Yang Z Q, Zhang J Y, Chang E C, et al. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 225-240.

[43] Hidano S, Murakami T, Katsumata S, et al. Model Inversion Attacks for Online Prediction Systems: Without Knowledge of Non-Sensitive Attributes[J]. *IEICE Transactions on Information and Systems*, 2018, E101.D(11): 2665-2676.

[44] Hidano S, Murakami T, Katsumata S, et al. Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes[C]. *2017 15th Annual Conference on Privacy, Security and Trust*, 2018: 115-11509.

[45] Aïvodji U, Gambs S, Ther T. GAMIN: An Adversarial Approach to Black-Box Model Inversion[EB/OL]. 2019: arXiv: 1909.11835. https://arxiv.org/abs/1909.11835.

[46] Shi Y, Zeng H, Nguyen T T, et al. Adversarial machine learning for network security[C]. *2019 IEEE International Symposium on Technologies for Homeland Security*, 2020: 1-7.

[47] Nash C, Kushman N, Williams C K I. Inverting Supervised Repre-

sentations with Autoregressive Neural Density Models[EB/OL]. 2018: arXiv: 1806.00400. https://arxiv.org/abs/1806.00400.

[48] Jacobsen J H, Smeulders A, Oyallon E. I-RevNet: Deep Invertible Networks[EB/OL]. 2018: arXiv: 1802.07088. https://arxiv.org/abs/1802.07088.

[49] Hitaj B, Ateniese G, Perez-Cruz F. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 603-618.

[50] Wu X, Fredrikson M, Wu W T, et al. Revisiting Differentially Private Regression: Lessons from Learning Theory and Their Consequences[EB/OL]. 2015: arXiv: 1512.06388. https://arxiv.org/abs/1512.06388.

[51] Li X G, Li H, Li F H, et al. A Survey on Differential Privacy[J]. *Journal of Cyber Security*, 2018, 3(5): 92-104.
(李效光, 李晖, 李凤华, 等. 差分隐私综述[J]. *信息安全学报*, 2018, 3(5): 92-104.)

[52] A. MORDVINTSEV, C. OLAH and M. TYKA. Deepdream-a code example for visualizing neural networks[J]. *Google Research*, 2015, 2(5).

[53] Oord A V D, Kalchbrenner N, Kavukcuoglu K. Pixel Recurrent Neural Networks[EB/OL]. 2016: arXiv: 1601.06759. https://arxiv.org/abs/1601.06759.

[54] Duddu V. A Survey of Adversarial Machine Learning in Cyber Warfare[J]. *Defence Science Journal*, 2018, 68(4): 356.

[55] Wang X B, Hou R, Zhu Y F, et al. NPUFort: A Secure Architecture of DNN Accelerator Against Model Inversion Attack[C]. *The 16th ACM International Conference on Computing Frontiers*, 2019: 190-196.

[56] Zhang D, Chen X, Cui Z Q, et al. Software defect prediction model sharing under differential privacy[C]. *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, 2018: 1547-1554.

[57] Zhang J L, Gu Z S, Jang J, et al. Protecting Intellectual Property of Deep Neural Networks with Watermarking[C]. *The 2018 on Asia Conference on Computer and Communications Security*, 2018: 159-172.

[58] Liu J, Juuti M, Lu Y, et al. Oblivious Neural Network Predictions via MiniONN Transformations[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 619-631.

[59] Ma X, Chen X F, Zhang X Y. Non-Interactive Privacy-Preserving Neural Network Prediction[J]. *Information Sciences*, 2019, 481: 507-519.

[60] Xiang L Y, Ma H T, Zhang H, et al. Interpretable Complex-Valued Neural Networks for Privacy Protection[EB/OL]. 2019: arXiv: 1901.09546. https://arxiv.org/abs/1901.09546.

[61] Aziz A, Permana U. Improvement of Performance Intrusion Detection System (IDS) Using Artificial Neural Network Ensemble[J]. *J Theor Appl Inf Technol*, 2015, 80(2): 191-201.

[62] Yu L, Liu L, Pu C, et al. Differentially private model publishing for deep learning[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 332-349.

[63] Triastcyn A, Faltings B. Federated Generative Privacy[EB/OL]. 2019: arXiv: 1910.08385. https://arxiv.org/abs/1910.08385.

[64] Wang Y, Si C, Wu X T. Regression Model Fitting under Differential Privacy and Model Inversion Attack[C]. *The 24th International Conference on Artificial Intelligence*, 2015: 1003-1009.

[65] Chen X H, Ji J L, Luo C Q, et al. When machine learning meets blockchain: A decentralized, privacy-preserving and secure design[C]. *2018 IEEE International Conference on Big Data*, 2019: 1178-1187.

[66] Chen H L, Fu C, Zhao J S, et al. DeepInspect: A black-box Trojan detection and mitigation framework for deep neural networks[C]. *The Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019: 4658-4664.

[67] Yang Y. *Research on data cleaning and joint learning for privacy protection of multiple data sources*[D]. Xi'an: Xidian University, 2019.
(杨烨. 多数据源隐私保护数据清洗与联合学习研究[D]. 西安: 西安电子科技大学, 2019.)

[68] Xie P T, Bilenko M, Finley T, et al. Crypto-Nets: Neural Networks over Encrypted Data[EB/OL]. 2014: arXiv: 1412.6181. https://arxiv.org/abs/1412.6181.

[69] Dowlin N, Gilad-Bachrach R, Laine K, et al. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy[C]. *The 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016: 201-210.

[70] Yang Z Q, Shao B, Xuan B H, et al. Defending Model Inversion and Membership Inference Attacks via Prediction Purification[EB/OL]. 2020: arXiv: 2005.03915. https://arxiv.org/abs/2005.03915.

张欢 于 2019 年在燕山大学计算机科学与技术专业获得学士学位。现在中国科学院大学网络空间安全专业攻读博士学位。研究领域为网络安全。研究兴趣包括: 人工智能安全、数据防泄露。Email: zhanghuan@iie.ac.cn

韩言妮 于 2010 年在北京航空航天大学计算机软件与理论专业获得博士学位。现任中国科学院信息工程研究所第五研究室副研究员。研究领域为数据智能分析与数据安全。Email：hanyanni@iie.ac.cn

赵一宁 于 2010 年在北京大学获得管理学硕士学位。现就职于中国移动信息技术中心, 负责研发规划工作。研究领域为数据治理。Email: zhaoyining@chinamobile.com

张帆 于 2004 年在天津理工大学计算机专业获得学士学位。现任中国移动信息技术中心大数据行业生态部副总经理, 研究领域为大数据、隐私计算。Email: zhangfanxx@chinamobile.com

谭倩 于 2015 年在重庆大学通信与信息系统专业获得博士学位。现任中国科学院信息工程研究所第五研究室助理研究员。研究领域为智能运维、网络安全。研究兴趣包括：网络流量异常检测、人工智能安全。Email：tanqian@iie.ac.cn

孟渊 于 2017 年在中国科学院大学计算机技术专业获得硕士学位。现就职于新疆阿克苏地区阿克苏市公安局网安部门。研究领域为：公安大数据运用，智能情报获取及搜索。Email：58333843@qq.com