人工智能模型数据泄露的攻击与防御研究综述

任奎，孟泉润，闫守琨，秦湛
（浙江大学网络空间安全学院，浙江 杭州 310027）

摘　要：人工智能和深度学习算法正在高速发展，这些新兴技术在音视频识别、自然语言处理等领域已经得到了广泛应用。然而，近年来研究者发现，当前主流的人工智能模型中存在着诸多安全隐患，并且这些隐患会限制人工智能技术的进一步发展。因此，研究了人工智能模型中的数据安全与隐私保护问题。对于数据与隐私泄露问题，主要研究了基于模型输出的数据泄露问题和基于模型更新的数据泄露问题。在基于模型输出的数据泄露问题中，主要探讨了模型窃取攻击、模型逆向攻击、成员推断攻击的原理和研究现状；在基于模型更新的数据泄露问题中，探讨了在分布式训练过程中，攻击者如何窃取隐私数据的相关研究。对于数据与隐私保护问题，主要研究了常用的 3 类防御方法，即模型结构防御，信息混淆防御，查询控制防御。综上，围绕人工智能深度学习模型的数据安全与隐私保护领域中最前沿的研究成果，探讨了人工智能深度学习模型的数据窃取和防御技术的理论基础、重要成果以及相关应用。

关键词：人工智能；数据安全；隐私泄露；隐私保护
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.2096−109x.2021001

Survey of artificial intelligence data security and privacy protection

REN Kui, MENG Quanrun, YAN Shoukun, QIN Zhan
School of Cyber Science and Technology, Zhejiang University, Hangzhou 310027, China

Abstract: Artificial intelligence and deep learning algorithms are developing rapidly. These emerging techniques have been widely used in audio and video recognition, natural language processing and other fields. However, in recent years, researchers have found that there are many security risks in the current mainstream artificial intelligence model, and these problems will limit the development of AI. Therefore, the data security and privacy protection was studied in AI. For data and privacy leakage, the model output based and model update based problem of data leakage were studied. In the model output based problem of data leakage, the principles and research status of model extraction attack, model inversion attack and membership inference attack were discussed. In the model

收稿日期：2020−07−01；**修回日期**：2020−09−29
通信作者：秦湛，qinzhan@zju.edu.cn
基金项目：科技创新 2030——“新一代人工智能”重大项目（2020AAA0107700）
Foundation Item: The National Key Research and Development Project (2020AAA0107700)

论文引用格式：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.
REN K, MENG Q R, YAN S K, et al. Survey of artificial intelligence data security and privacy protection[J]. Chinese Journal of Network and Information Security, 2021, 7(1): 1-10.

update based problem of data leakage, how attackers steal private data in the process of distributed training was discussed. For data and privacy protection, three kinds of defense methods, namely model structure defense, information confusion defense and query control defense were studied. In summarize, the theoretical foundations, classic algorithms of data inference attack techniques were introduced. A few research efforts on the defense techniques were described in order to provoke further research efforts in this critical area.

Keywords: artificial intelligence, data security, privacy leakage, privacy protection

1 引言

人工智能（AI，artificial intelligence）技术正在加速崛起，它的崛起依托于3个关键因素：① 深度神经网络（DNN，deep neural network）在多个经典机器学习任务中取得了突破性进展；② 大数据处理技术的成熟以及海量数据的积累；③ 硬件计算能力的显著提高。在这3个因素的推动下，AI 技术已经成功应用于自动驾驶、图像识别、语音识别等场景，加速了传统行业的智能化变革。

AI 技术在我国已经得到了广泛的应用。在电商领域，AI 技术可以被用于用户行为分析、网络流量分析等任务，不仅使企业处理高并发业务更高效，而且提升了整体系统的鲁棒性；在智能出行领域，AI 技术可以被用于处理路径规划、司机乘客行为检测等任务；在金融领域，AI 技术可以执行高频交易、欺诈检测、异常检测等任务；在网络安全领域，AI 技术作为辅助工具被应用于自动化测试等任务中，极大地提升了安全人员在海量的大数据信息中定位异常点的效率。2017 年，我国政府工作报告首次提及人工智能相关内容，人工智能的发展也逐渐被上升到国家发展战略高度。

目前大多数现实世界的机器学习任务是资源密集型的，需要依靠大量的计算资源和存储资源完成模型的训练或预测，因此，亚马逊、谷歌、微软等云服务商往往通过提供机器学习服务来抵消存储和计算需求。机器学习服务商提供训练平台和使用模型的查询接口，而使用者可以通过这些接口来对一些实例进行查询。一般来说，服务商或者模型提供者会对使用者的查询操作按次进行收费。

但 AI 技术在高速发展的同时面临着严峻的数据泄露风险。AI 模型的参数需要得到保护，否则将对模型拥有者带来巨大的经济损失。此外，AI 技术所需要的样本数据往往包含了个人的隐私数据，这些隐私数据一旦被泄露，将会为模型拥有者带来巨大的经济风险和法律风险。2017 年，我国颁布的《中华人民共和国网络安全法》也强调了对个人隐私信息的保护。因此，如何充分防范 AI 技术应用中的数据泄露风险，成为该技术进一步发展与部署的阻碍之一。

为了保障人工智能模型相关信息的隐私性，云服务商会保证自身模型的隐秘性，仅提供一个接口来为使用者提供服务，从而保证模型使用者无法接触到模型数据。然而近年来，仍然出现了大量试图破坏人工智能模型数据隐私性的攻击。研究者发现深度学习模型使用过程中产生的相关计算数据，包括输出向量、模型参数、模型梯度等，可能会泄露训练数据的敏感信息或者模型自身的属性参数[1]。更糟的是，这些数据往往又不可避免地会被泄露给攻击者，尤其是某些模型的输出结果向量。这使深度学习模型的数据泄露问题难以避免。例如，模型逆向攻击，攻击者可以在不接触隐私数据的情况下利用模型输出结果等信息来反向推导出用户的隐私数据；成员推断攻击，攻击者可以根据模型的输出判断一个具体的数据是否存在于训练集中。而这类攻击只需要与云服务的接口进行交互。实际应用中，这类信息窃取攻击会导致严重的隐私泄露，如人脸识别模型返回的结果向量可以被用于恢复训练数据或者预测数据中的人脸图像，这将导致用户的肖像信息被泄露。攻击者还可以通过模型输出结果窃取相关模型的参数，对模型拥有者造成严重的经济损失[2]。

此外，随着联邦学习[3]等分布式机器学习技术的发展，攻击者有可能成为模型训练过程中的参与方。一般而言，联邦学习中的参与方无法获知彼此的输入数据等隐私信息，但由于攻击者能够获得模型在训练过程中的输出、模型参数和梯

度等信息，这大大提升了攻击者的能力，使攻击者窃取其他参与方隐私数据成为可能。这将给分布式机器学习技术的发展带来严重的阻碍。

近年来，许多研究者提出了各种机制来防御针对 AI 技术的隐私攻击。通过对模型结构的修改，为输出向量添加特定噪声，结合差分隐私等技术，能够有效防御特定的隐私泄露攻击。

本文将介绍目前研究较多的数据推断攻击，包括模型窃取攻击、模型逆向攻击、成员推断攻击。并介绍针对上述不同攻击的防御机制，其生成的具备隐私保护功能的模型能够抵抗特定的数据推断攻击。

2　AI 数据与隐私泄露

在深度学习模型的训练和应用过程中，所使用的数据和模型参数都面临着被泄露的风险。根据攻击者所利用的模型的输出信息类型的不同，可将此类推断攻击分为基于模型输出的数据泄露以及基于梯度更新的数据泄露两类。

2.1　基于模型输出的数据泄露

模型输出是指模型在训练完毕投入的阶段，接收输入返回给使用者的预测结果。例如，在分类任务中，模型输出就是对应样本的类别或者概率向量。近些年来的研究表明，模型输出结果隐含一定的数据信息。攻击者可以利用模型输出在一定程度上窃取相关数据，通过这种方法主要可以源。窃取两类数据信息：模型自身的参数数据；训练/测试数据。

（1）模型窃取

模型窃取攻击（model extraction attack）是一类窃取模型信息的恶意行为，攻击者通过向黑盒模型进行查询获取相应结果，获取相近的功能，或者模拟目标模型决策边界。被窃取的模型往往是拥有者花费大量的金钱时间构建而成的，对拥有者来说具有巨大的商业价值，一旦模型的信息遭到泄露，攻击者就能逃避付费或者开辟第三方服务从中获取商业利益，使模型拥有者的权益受到损害。更严重的是，如果模型遭到窃取，那么攻击者可以进一步部署白盒对抗攻击来欺骗在线模型，这时模型的泄露会大大增加攻击的成功率。例如，在针对亚马逊和谷歌的在线人工智能分类任务进行黑盒对抗攻击的时候，研究者仅使用少量的样本施展模型窃取攻击，并针对窃取到的替代模型生成白盒对抗样本，使用该方法生成的对抗样本可以使亚马逊和谷歌的分类模型分别出现 96.19%和 88.94%的误判率[4]。

目前，大多数的 AI 技术供应商是以如下模式提供服务的：提供功能的模型本身往往位于安全的云端服务器，通过 API 来为客户端提供付费查询服务。客户仅能通过定义好的 API 向模型输入查询样本，并获取模型对样本的预测结果，然而即使攻击者仅利用预测结果产生的信息，他也能在一定情况下通过查询来窃取服务端的模型。模型窃取攻击主要可以分为 3 类：Equation-solving Attack；基于 Meta-model 的模型窃取；基于替代模型的模型窃取。

Equation-solving Attack 是一类主要针对支持向量机（SVM）等传统的机器学习方法的模型窃取攻击。攻击者可以先获取模型的算法、结构等相关信息，然后构建公式方程来根据 query 的返回的结果求解模型参数[5]。在此基础之上还可以窃取传统算法中的超参数，如损失函数中 loss 项和 regularization 项的权重参数[6]、KNN 中的 K 值等。Equation-solving Attack 需要攻击者了解目标算法的类型、结构、训练数据集等信息，无法应用于复杂的神经网络模型。

基于 Meta-model 的模型窃取。这种攻击的主要思想通过训练一个额外的 meta model $\Phi(\cdot)$ 来预测目标模型的指定属性信息。Meta-model 的输入样本是所预测模型在任务数据 x 上的输出结果 $f(x)$，输出的内容 $\Phi(f(x))$则是预测目标模型的相关属性，如网络层数、激活函数类型等。因此为了训练 meta-model，攻击者需要自行收集与目标模型具有相同功能的多种模型，获取它们在相应数据集上的输出，构建 meta-model 的训练集。然而该训练集的构建需要多样的任务相关模型，对计算资源的要求过高，因此该类攻击并不是非常实用，文献[7]的作者也仅在 MNIST 数字识别任务上进行了相关实验。

基于替代模型的模型窃取是目前比较实用的一类攻击。攻击者在未知目标模型结构的情况下向目标模型查询样本，得到目标模型的预测结果，

并以这些预测结果对查询数据进行标注构建训练数据集，在本地训练一个与目标模型任务相同的替代模型，当经过大量训练之后，该模型就具有和目标模型相近的性质。一般来说，攻击者会选取 VGG[8]、ResNet[9]等具有较强的拟合性的深度学习模型作为替代模型结构[10]。基于替代模型的窃取攻击与 Equation-solving Attack 的区别在于，攻击者对于目标模型的具体结构并不了解，训练替代模型不是为了获取目标模型的具体参数，而只是利用替代模型去拟合目标模型的功能。为了拟合目标模型的功能，替代模型需要向目标模型查询大量的样本来构建训练数据集，然而攻击者往往缺少充足的相关数据，并且异常的大量查询不仅会增加窃取成本，更有可能会被模型拥有者检测出来。为了解决上述问题，避免过多地向目标模型查询，使训练过程更为高效，研究者提出对查询的数据集进行数据增强，使这些数据样本能够更好地捕捉目标模型的特点[4]，如利用替代模型生成相应的对抗样本以扩充训练集，研究认为对抗样本往往会位于模型的决策边界上，这使替代模型能够更好地模拟目标模型的决策行为[11-12]。除了进行数据增强，还有研究表明使用与目标模型任务无关的其他数据构建数据集也可以取得可观的攻击效果，这些工作同时给出了任务相关数据与无关数据的选取组合策略[2,10]。

（2）隐私泄露

机器学习模型的预测结果往往包含了模型对于该样本的诸多推理信息。在不同的学习任务中，这些预测结果又包含了不同的含义。例如，图像分类任务中，模型输出的是一个向量，其中，每一个向量分量表示测试样本为该种类的概率。最近的研究证明，这些黑盒的输出结果可以被用来窃取模型训练数据的信息，如 Fredrikson 等提出的模型逆向攻击（model inversion attack）[13]可以利用黑盒模型输出中的 confidence 等信息将训练集中的人脸恢复出来。他们针对常用的面部识别模型，包括 softmax 回归[14]、多层感知机和自编码器网络实施模型逆向攻击。他们认为模型输出的 confidence 包含的输入数据信息，也可以作为输入数据恢复攻击的衡量标准。他们将模型逆向攻击问题转变为一个优化问题，优化目标为使逆向数据的输出向量与目标数据的输出向量差异尽可能地小，即假如攻击者获得了属于某一类别的输出向量，那么他可以利用梯度下降的方法使逆向的数据经过目标模型的推断后，仍然能得到同样的输出向量。

成员推断攻击（membership-inference attack）是一种更加容易实现的攻击类型。它是指攻击者将试图推断某个待测样本是否存在于目标模型的训练数据集中，从而获得待测样本的成员关系信息。例如，攻击者希望知道某个人的数据是否存在于某个公司的医疗诊断模型的训练数据集中，如果存在，那么可以推断出该个体的隐私信息。目标模型训练集中的数据被称为成员数据（member data），而不在训练集中的数据被称为非成员数据（non-member data）。同时由于攻击者往往不可能掌握目标模型，因此攻击者只能实施黑盒场景下的成员推断攻击。文献[15-20]已经对这种攻击进行了深入的研究。成员推断攻击是近两年来新兴的一个研究课题，这种攻击可以用于医疗诊断、基因测试等应用场景，它对用户的隐私数据提出了挑战，同时关于这种攻击技术的深入发展及其相关防御技术的探讨成为一个新的研究热点。

2017 年，Shokri 等[15]第一次提出了成员推断攻击。经过大量实验，他们完成了黑盒场景下成员推断攻击的系统设计。这种攻击的原理是机器学习模型对成员数据的预测向量和对非成员数据的预测向量有较大的差异，如果攻击者能准确地捕捉到这种差异，就可以实施成员推断攻击。然而，在黑盒的场景下，可以从目标模型中得到的只有预测向量，甚至在实际场景下，由于企业的使用限制，无法从目标模型中获得足够多样本的预测向量。此外，由于不同样本的预测向量的分布本身就不一致，即使攻击者直接利用预测向量进行训练，也无法实现较好的攻击效果。因此，Shokri 等使用与目标网络相同的结构，并建立与目标数据集同分布的 shadow 数据集，之后为每一类数据建立多个 shadow 模型，实现了对预测向量的数据增强效果，并获得了大量的预测向量作为攻击模型的训练样本。并且，利用预测向量，他们构建了攻击模型，使其能够捕捉预测向量在

成员数据和非成员数据之间的差异，从而完成了黑盒场景下的成员推断攻击。

之后随着成员推断攻击技术的发展，人们发现这种攻击的本质就是目标模型对成员数据和非成员数据给出的预测向量存在差异，即成员数据的输出向量的分布更集中，而非成员数据的输出向量的分布相对较为平缓。这种差异性是模型过拟合的主要表现，也就是说成员推断攻击与模型的过拟合程度有很大关联。在这个研究方向上，Yeom 等[16]研究了模型的过拟合对成员推断攻击的影响，他们通过理论和实验证实了模型的过拟合程度越强，模型泄露训练集成员关系信息的可能性越大；但同时指出，模型的过拟合并不是模型易受成员推断攻击的唯一因素，一些过拟合程度不高的模型也容易受到攻击。随后，Ashamed 等[17]进一步完善了黑盒场景下的成员推断攻击，他们在 2019 年提出了改进后的成员推断攻击，在极大地降低了实现这种攻击成本的同时，实现了与 Shokri 等[15]相同的攻击效果，并更明确地展示了成员推断攻击出现的本质原因。即成员数据和非成员数据的预测向量间的差异主要体现为预测向量的集中度。同时他们提出了 3 种方法，不断减少了成员推断攻击的部署成本。第一种情况下，他们对目标模型的输出向量从大到小进行重排序，使模型对不同类别数据的输出向量的分布趋于一致，均为从大到小，这样就可以避免数据增强的过程，进而减少所需 shadow model 的数量，同时不需要知道目标模型的结构，只需要使用基础的网络结构（如 CNN[21]、Logistic Regression[22]）和随机森林[23]等来构建 shadow model 即可。同时他们发现，只需要截取排序后预测向量的前 3 个概率值作为攻击模型的训练样本，也能达到较好的攻击效果；第二种情况下，他们提出了数据迁移攻击，即使用与目标模型的训练集分布不同的数据集来训练 shadow model，最终获得的攻击模型同样能对目标模型的数据进行成员关系推断，并实现类似的攻击效果；第三种情况下，他们提出了 threshold choosing，使用该策略可以确定出一个阈值 T，只要预测向量的最大值大于 T，即称该向量对应的待测样本为成员数据，否则，为非成员数据。Ashamed 等[17]的工作进一步强化了成员推断攻击，极大地提升了该攻击的威胁性。

随着人们对成员推断攻击研究的深入，研究者们发现了成员推断攻击的一些新特性。如 Song 等[24]发现当一个机器学习模型被加入了一些抵御对抗样本攻击的方法后，会提高该模型泄露成员隐私信息的风险。也就是说机器学习模型在对抗样本安全性和成员数据隐私性之间存在一个 trade-off，如果提高了模型抵御对抗样本的能力，同时会提高从模型中推断出成员数据存在与否的可能性，反之，亦然。此外，Salem 等[25]将成员推断攻击拓展到了在线学习领域。他们发现当机器学习模型完成在线学习后，可以通过更新前后的模型对同一个数据集给出的预测向量的差异，来完成对目标模型更新集中特定数据的存在性推断，甚至完成对更新集数据的重建。Hayes 等[26]利用生成对抗网络（GAN）完成了成员推断攻击的构建。Nasr 等[27]也研究了白盒场景下成员推断攻击，他们利用成员数据和非成员数据在模型梯度上的差异，再结合输出向量上的差异，构建了能力更强的成员推断攻击模型，并成功绕过前提出的一些防御手段，达到了较高的攻击率。Leino 等[28]则进一步完善了白盒场景下的成员推断攻击，他们将输出向量、隐含层的权重、偏差、线性单元以及激活函数等特征结合起来，构建了鲁棒性更强的成员推断攻击，成功抵抗了目前针对成员推断攻击的大部分防御方法，并取得了较强的攻击效果。

2.2　基于梯度更新的数据泄露

梯度更新是指模型每一次对模型参数进行优化时，参数会根据计算产生的梯度来进行更新，而在训练过程中不断产生的梯度同样隐含着某些隐私信息。梯度更新的交换往往只出现在模型的分布式训练中，拥有不同数据的多方主体，每一轮仅使用自己的数据来更新模型，只对模型参数的更新进行交换汇总，分布式地完成统一模型的训练。在这个过程中，中心服务器和任何训练主题都不会获得其他主体拥有的训练数据。然而即便是在原始数据获得良好保护的情况下，模型梯度更新仍会导致隐私泄露。尽管模型在训练的过程中已经使用了很多方法防止原始数据泄露，在

多方分布式的 AI 模型训练中，个体往往会使用自己的数据对当前的模型进行训练，并将模型的参数更新传递给其他个体或者中心服务器。在最近机器学习和信息安全的国际会议上，出现了一些利用模型参数更新来获取他人训练数据信息的攻击研究。Melis 等[29]利用训练过程中其他用户更新的模型参数作为输入特征，训练攻击模型，用于推测其他用户数据集的相关属性；还有研究者[30-31]利用对抗生成网络生成恢复其他用户的训练数据的方法，在多方协作训练过程中，使用公共模型作为基本的判别器，将模型参数更新作为输入训练生成器，最终获取受害者特定类别的训练数据。而在最近的一项工作中[32]，作者并未使用 GAN 等生成模型，而是基于优化算法对模拟图片的像素进行调整，使其在公共模型上反向传播得到的梯度和真实梯度相近，经过多轮的优化模拟图片会慢慢接近真实的训练数据。

3 AI 数据与隐私保护

为了减轻 AI 模型在训练和测试过程中可能会造成的模型与隐私泄露风险，包括训练阶段模型参数更新导致的训练数据信息泄露、测试阶段模型返回查询结果造成的模型数据泄露和数据隐私泄露这些AI模型正常使用过程中间接引起的数据隐私泄露，学术界和工业界从不同角度都进行了许多尝试。

在没有被直接攻击破解的情况下，AI 模型正常训练和使用的过程中产生的信息也会导致数据隐私的间接泄露。为了解决这类数据泄露，采用的主要思想就是在不影响 AI 模型有效性的情况下，尽可能减少或者混淆这类交互数据中包含的有效信息。可以采用以下几类数据隐私保护措施：模型结构防御，该类方法是指在模型的训练过程中对模型进行有目的性地调整，降低模型输出结果对于不同样本的敏感性；信息混淆防御，该类方法通过对模型输出、模型参数更新等交互数据进行一定的修改，在保证模型有效性的情况下，尽可能破坏混淆交互数据中包含的有效信息；查询控制防御，该类防御通过对查询操作进行检测，及时拒绝恶意的查询从而防止数据泄露。

3.1 模型结构防御

面向模型的防御是通过对模型结构做适当的修改，以此来减少模型被泄露的信息，或者降低模型的过拟合程度，从而完成对模型泄露和数据泄露的保护。Fredrikson 等[33]提出当目标模型为决策树时，可使用 CART 决策树的变种，将样本的敏感特征的优先级调高或调低，他们通过实验证明，当敏感特征在决策树的根节点和叶子节点层级时，对 model inversion 攻击能够达到较好的防御效果，其中当敏感属性位于根节点时，能达到最好的防御效果。Shokri 等[15]和 Ahmed 等[17]提出可以在目标模型中添加 Dropout 层，或者使用 model stacking 的方法将不同的元学习器聚合在一起，又或者在目标模型中添加正则项等。通过实验，他们发现当目标模型使用这些方法后，能显著地减少成员推断攻击的准确率。Nasr 等[34]提出了一种基于对抗学习的防御方法，他们认为如果能计算出当前模型抵抗成员推断攻击的成功率，并将其作为一个对抗正则项加入损失函数中，那么在训练过程中使用 MIN-MAX 的对抗训练方式，最终就可以训练出一个模型，该模型下成员推断攻击的成功率将存在一个上界。最终实验表明该方法在使这个上界足够小的同时，能够达到较高的分类准确度。

此外，Wang 等[35]构建了 MIASec，他们提出可以对训练数据在目标模型的关键特征上进行特定的修改，从而使模型对成员数据和非成员数据的预测向量的分布难以区分，进而可以完成对成员推断攻击的防御。如前文所述，模型逆向攻击的核心原因是输出向量包含了训练样本的信息，成员推断攻击的核心原因是模型对训练样本和测试样本的预测向量的分布不一致。因此，防御模型逆向攻击就是尽可能地降低输出向量与输入向量间的关联，防御成员推断攻击就是尽可能地缩小输出向量间的分布差异。面向模型的防御旨在通过修改模型的结构和损失函数，使目标模型给出的输出向量中包含尽可能少的信息，从而完成较好的防御效果。但这种方式仍有缺陷，它对目标模型的性能有较大影响，导致其分类准确度出现波动。因此，防御方需要在模型的性能与其鲁棒性之间做出平衡。

近年来一些工作开始将机器学习与加密技术结合起来保护模型的隐私性。Nan 等[36]提出

在分布式训练的场景下，可以在每次模型梯度更新的同时，使用差分隐私技术对梯度做一定的修饰，从而保护训练数据集的隐私性，尽管这种方法会降低模型的最终性能，但确实能大幅提高训练集的隐私性。同样，Patra 等[37]也借助于安全多方计算的技术重新实现了加密条件下的矩阵乘法和激活函数的计算，在该框架的支持下，可以有效地保护训练过程中训练集的隐私性。这些隐私保护机器学习技术的思想也能够用在针对数据泄露的防御中，加强模型训练集的隐私性。

3.2　信息混淆防御

面向数据的防御是指对模型的输入样本或预测结果做模糊操作。通过这些模糊操作，在保证 AI 模型输出结果正确性的前提下，尽可能地干扰输出结果中包含的有效信息，从而减少隐私信息的泄露。这些数据模糊操作主要包含两类：一类是截断混淆，即对模型返回的结果向量做取整操作，抹除小数点某位之后的信息[2,6,15]；另一类是噪声混淆，即对输入样本或输出的概率向量中添加微小的噪声，从而干扰准确的信息。

对于截断混淆，Shokri 等[15]提出可以对目标模型生成的输出向量进行截取，如只给出输出向量中概率值较高的类别的相应结果，或者降低输出向量中小数位的保留位数，Fredrikson 等[33]提出可以对目标模型的输出向量进行取整，达到对输出向量的修饰效果。通过截断混淆等方法，研究者们削弱对模型逆向攻击和成员推断攻击的攻击效果。

对于噪声混淆，Jia 等[38]基于对抗样本的理念提出了 Mem-guard。他们发现成员推断攻击对目标模型给出的预测向量的变化非常敏感，如果为这些预测向量添加一个精心设计的噪声，从而混淆成员数据和非成员数据的预测向量分布的差异，就可以生成一个对实际结果没有影响的“对抗样本”，这样就可以完成对成员推断攻击的防御。He 等[39]提出可以用差分隐私的方法对输出向量加噪声进行混淆，他们认为可以利用差分隐私的算法来移除输出向量自身的特征，但同时保留了其关于分类结果的信息，使输出向量难以被区分。此外，他们还提出可以在损失函数中添加噪声项，在轻微地牺牲分类准确率的同时，提高输出向量的隐私性，完成对成员推断攻击的防御。

模型逆向攻击和成员推断攻击的输入都是目标模型的输出向量，因此，如果能够在不影响分类结果的前提下，对输出向量进行特定地修饰，就可以扰乱输出结果中的有效信息，从而完成防御，但这种方法依然有局限性。如果对输出向量的修饰程度较小，则其抵抗攻击的能力也不会较好，如果对输出向量的修饰程度较大，则会影响分类数据的可用性，也就是说，这里仍然需要选取隐私性与可用性之间的平衡。

3.3　查询控制防御

查询控制防御是指防御方可以根据用户的查询行为进行特征提取，进而完成对隐私泄露攻击的防御。攻击者如果想要执行隐私泄露攻击，需要对目标模型发起大量的查询行为，甚至需要对自己的输入向量进行特定的修饰，从而加快隐私泄露攻击的实施。根据用户查询行为的特征，可以分辨出哪些用户是攻击者，进而对攻击者的查询行为进行限制或拒绝服务，以达到防御攻击的目的。查询控制防御主要包含两类：异常样本检测和查询行为检测。

在异常样本检测中，攻击者为了窃取黑盒的在线模型，往往需要对在线模型进行大量的查询操作。为了提高窃取效率，攻击者会对正常的样本进行有目的地修改。而针对模型泄露攻击的特点，防御者主要通过检测对异常样本的查询，来识别模型窃取行为。PRADA[2]是一种针对模型窃取攻击进行检测的防御技术，它根据多个样本特征之间的距离分布来判断该用户是否正在施展模型窃取攻击，该文献发现随机选取的正常样本特征间的距离大致服从正态分布，而模型窃取过程中查询的样本往往具有鲜明的人工修改迹象，样本间距离分布与正态分布区别较大，通过这种方式，对若干次的查询进行统计检验则可检测异常查询用户。查询样本的特征分布也可以被用于检测，Kesarwani 等[40]记录下用户的查询样本并检查其在特征空间中的分布，来评估模型被盗取的风险；Yu 等[12]提出正常样本的特征分布与人工修改的样本特征分布相比有较大的区别，可以通过区分样本的特征分布来检测异常查询。

在查询行为检测中，由于攻击者往往需要对目标模型进行大量的测试，所以其查询行为与正常行为会有较大不同。根据这种差异可以在一定程度上防御模型泄露和数据泄露攻击。针对数据泄露攻击的特点，He 等[39]提出可以根据用户查询的行为特征，在样本输入阶段，完成对成员推断攻击的防御。攻击者实行成员推断攻击时有时需要查询大量目标模型，模型提供者可以根据用户的查询频率实现对查询次数的限制，从而提升攻击者部署成员推断攻击的成本。

由上文可知，防御方可以通过对异常样本的检测和异常查询行为的检测来完成对模型泄露攻击和数据泄露攻击的防御。但这种防御方法的针对性不强，而且效果不够好，误分类的概率较大。查询控制防御主要是在攻击模型的训练过程中起作用，对已训练好的攻击模型无能为力。此外，如果攻击者知道目标模型采用了查询控制防御，他们也有许多方法可以绕过这种防御方法，如设计更难以被检测的异常样本或者采用虚拟 IP 地址等方式绕过目标模型的检测。

4 研究展望

4.1 高效的数据泄露攻击技术发展

数据泄露攻击的本质是模型的参数、模型的输出向量等信息是根据输入样本而产生的，即无论如何，这些数据都会包含原始数据的信息，也就是说任何一个人工智能模型都有遭受数据泄露的风险，并且无法完全抵抗这种攻击的威胁。因此，未来针对人工智能模型的数据泄露攻击的发展主要包括两类：第一类是优化攻击模型，增强其从输出向量中提取信息的能力；第二类是扩展攻击场景，将数据泄露攻击应用到更多的场景中，如迁移学习、强化学习等。此外，利用模型的输出信息进行隐私窃取，这种攻击往往需要目标模型进行大量的查询操作，如在模型窃取中，由于深度学习网络具有参数规模大、高度的非线性、非凸性等性质，导致训练替代模型需要数以千计的查询次数[10]。大量的查询提高了攻击的成本，并且增加了被防御者发现的风险，因此如何更加高效地进行隐私窃取是目前攻击者所要研究的主要方向，在这个方面研究者们进行了大量的尝试，这些方法的主要思想是建立一类样本选取策略，从而使用更具有代表性的样本进行攻击，从而提高攻击效率[41-42]，如积极学习[43-44]、自然进化策略[45]等方法。对攻击的深入研究不仅能够促进隐私保护的不断进化，同时有助于研究者对人工智能模型更加深刻的理解。

4.2 有效的数据泄露攻击的防御技术发展

如上文所述，数据泄露攻击的本质是模型构建或使用时的输出结果，隐含了某些隐私数据的信息，因此，针对数据泄露攻击的防御，可以主要从以下 3 个方向进一步发展。一是针对输出向量进行混淆，降低其所包含的信息；二是对隐私数据进行混淆，可以构建特定的噪声来修饰原使用数据，从而降低模型推断结果的信息；三是对模型本身的参数做混淆，如引入隐私保护机器学习的方法，对模型内部的参数、中间结果和输出向量进行加密处理，降低其泄露信息的可能性。然而对各类信息数据的修饰程度则是在构建防御时需要着重考虑的因素，如果修饰程度过小，那么该防御则无法达到预期的防御效果，攻击者仍然能够窃取隐私数据，相反如果修饰程度过大，则会导致模型的产出结果的可用性降低，使其本职功能受到巨大损害。与混淆信息防御相类似，其他防御也有类似的情况，如对于查询控制防御，严格的查询控制规则将有效地避免隐私数据的泄露，然而却会使正常用户的使用过程变得烦琐，甚至可以能会把正常用户误判为攻击者。因此为了在保证隐私数据混乱的情况下，模型能够有效稳定地提供原有服务，隐私泄露防御技术要在安全性与模型可用性之间寻求一个有效的平衡，这是防御技术在实际应用和未来发展中需要着重关心的一个方面。

5 结束语

本文对近年来人工智能数据安全与隐私保护的研究工作进行了总结和分析，虽然已经有很多的研究者对人工智能系统基于模型输出以及基于梯度更新的数据泄露进行了一系列的研究，并且提出了包括模型结构防御、信息混淆防御以及查询控制防御在内的多种防御技术。但相比于已经发展成熟的传统数据安全领域，由于深度学习算

法本身存在的可解释性不足的问题，对于人工智能算法数据安全与隐私保护问题的妥善解决，还面临着诸多挑战，需要进一步展开研究工作。

参考文献：

[1] ATENIESE G, MANCINI L V, SPOGNARDI A, et al. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers[J]. International Journal of Security and Networks, 2015, 10(3): 137-150.

[2] JUUTI M, SZYLLER S, MARCHAL S, et al. PRADA: protecting against DNN model stealing attacks[C]//In IEEE European Symposium on Security and Privacy. 2019: 512–527.

[3] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.

[4] PAPERNOT N, MCDANIEL P D, GOODFELLOW I J, et al. Practical black-box attacks against machine learning[C]//In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[5] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs[C]//In 25th USENIX Security Symposium, USENIX Security 16. 2016: 601-618.

[6] WANG B H, GONG N Z. Stealing hyperparameters in machine learning[C]//In 2018 IEEE Symposium on Security and Privacy. 2018: 36-52.

[7] OH S J, SCHIELE B, FRITZ M. Towards reverse-engineering black-box neural networks[J]. arXiv: 1711.01768, 2019.

[8] SATHISH K, RAMASUBBAREDDY S, GOVINDA K. Detection and localization of multiple objects using VGGNet and single shot detection[M]//Emerging Research in Data Engineering Systems and Computer Communications. Singapore: Springer. 2020: 427-439.

[9] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.

[10] CORREIA-SILVA J R, BERRIEL R F, BADUE C, et al. Copycat CNN: stealing knowledge by persuading confession with random non-labeled data[C]//In 2018 International Joint Conference on Neural Networks. 2018: 1-8.

[11] BATINA L, BHASINS, JAP D, et al. CSI NN: reverse engineering of neural network architectures through electromagnetic side channel[C]//In 28th USENIX Security Symposium, USENIX Security 2019. 2019: 515-532.

[12] YU H G, YANG K C, ZHANG T, et al. Cloudleak: large-scale deep learning models stealing through adversarial examples[C]//Network and Distributed System Security Symposium. 2020.

[13] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1322-1333.

[14] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax[J]. arXiv preprint arXiv:1611.01144, 2016.

[15] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//In 2017 IEEE Symposium on Security and Privacy. 2017: 3-18.

[16] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: analyzing the connection to overfitting[C]//In 31st IEEE Computer Security Foundations Symposium. 2018: 268-282.

[17] SALEM A, ZHANG Y, HUMBERT M, et al. Ml-leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//In 26th Annual Network and Distributed System Security Symposium. 2019: 24-27.

[18] LONG Y H, BINDSCHAEDLER V, GUNTER C A. Towards measuring membership privacy[J]. CoRR, abs/1712.09136, 2017.

[19] LONG Y H, BINDSCHAEDLER V, WANG L, et al. Understanding membership inferences on well-generalized learning models[J]. CoRR, abs/1802.04889, 2018.

[20] YEOM S, FREDRIKSON M, JHA S. The unintended consequences of overfitting: Training data inference attacks[J]. CoRR, abs/1709.01604, 2017.

[21] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4031-4039.

[22] KUHA J, MILLS C. On group comparisons with logistic regression models[J]. Sociological Methods & Research, 2020, 49(2): 498-525.

[23] PAL M. Random forest classifier for remote sensing classification[J]. International journal of remote sensing, 2005, 26(1): 217-222.

[24] SONG L, SHOKRI R, MITTAL P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 241-257.

[25] SALEM A, BHATTACHARYA A, BACKES M, et al. Updates-leak: data set inference and reconstruction attacks in online learning[J]. arXiv preprint arXiv:1904.01067, 2019.

[26] HAYES J, MELIS L, DANEZIS G, et al. LOGAN: membership inference attacks against generative models[J]. PoPETs, 2019(1): 133-152.

[27] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]//In 2019 IEEE Symposium on Security and Privacy. 2019: 739-753.

[28] LEINO K, FREDRIKSON M. Stolen memories: leveraging model memorization for calibrated white-box membership inference[J]. arXiv preprint arXiv:1906.11798, 2019.

[29] MELIS L, SONG C Z, CRISTOFARO E D, et al. Exploiting unintended feature leakage in collaborative learning[C]//In 2019 IEEE Symposium on Security and Privacy. 2019: 691-06.

[30] WANG Z B, SONG M K, Zhang Z F, Yet al. Beyond inferring class representatives: user-level privacy leakage from federated learning[C]//In 2019 IEEE conference on Computer Communications. 2019: 2512-2520.

[31] HITAJ B, ATENIESE G, PÉREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]//In Proceedings of the 2017 ACM SIGSAC Conference on Computer

and Communications Security. 2017: 603-618.

[32] ZHU L G, LIU Z J, HAN S. Deep leakage from gradients[C]//In Advances in Neural Information Processing Systems Annual Conference on Neural Information Processing Systems 2019. 2019: 14747-14756.

[33] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1322-1333.

[34] NASR M, SHOKRI R, HOUMANSADR A. Machine learning with membership privacy using adversarial regularization[C]//In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018: 634-646.

[35] WANG C, LIU G Y, HUANG H J, et al. MIASec: enabling data indistinguishability against membership inference attacks in MLaaS[J]. IEEE Transactions on Sustainable Computing, 2020, 5(3): 365-376.

[36] WU N, FAROKHI F, SMITH D, et al. The Value of collaboration in convex machine learning with differential privacy[J]. IEEE Symposium on Security and Privacy, 2020: 304-317.

[37] PATRA A, SURESH A. BLAZE: blazing fast privacy-preserving machine learning[J]. arXiv preprint arXiv:2005.09042, 2020.

[38] JIA J Y, SALEM A, BACKES M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples[C]//In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019. 2019: 259-274..

[39] HE Y Z, MENG G Z, CHEN K, et al. Towards privacy and security of deep learning systems: a survey[J]. arXiv: 1911.12562, 2019.

[40] KESARWANI M, MUKHOTY B, ARYA V, et al. Model extraction warning in MLaaS paradigm[C]//In Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC 2018. 2018: 371-380.

[41] OH S J, SCHIELE B, FRITZ M. Towards reverse-engineering black-box neural networks[M]//Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, Cham, 2019: 121-144.

[42] OREKONDY T, SCHIELE B, FRITZ M. Knockoff nets: Stealing functionality of black-box models[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4954-4963.

[43] CHANDRASEKARAN V, CHAUDHURI K, GIACOMELLI I, et al. Exploring connections between active learning and model extraction[J]. arXiv preprint arXiv:1811.02054, 2018.

[44] PENGCHENG L, YI J, ZHANG L. Query-efficient black-box attack by active learning[C]//2018 IEEE International Conference on Data Mining (ICDM). 2018: 1200-1205.

[45] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[J]. arXiv preprint arXiv:1804.08598, 2018.

[作者简介]

任奎（1978– ），男，安徽芜湖人，浙江大学教授、博士生导师，主要研究方向为人工智能安全、数据安全、物联网安全。

孟泉润（1994– ），男，河南新乡人，浙江大学硕士生，主要研究方向为数据安全与隐私保护。

闫守琨（1996– ），男，辽宁大连人，浙江大学硕士生，主要研究方向为人工智能安全与对抗攻防。

秦湛（1988– ），男，北京人，浙江大学研究员、博士生导师，主要研究方向为数据安全与隐私保护、人工智能安全。