软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2023,34(2):712−732 [doi: 10.13328/j.cnki.jos.006699]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

人工智能的逆向工程——反向智能研究综述*

李长升, 汪诗烨, 李延铭, 张成喆, 袁 野, 王国仁

(北京理工大学 计算机学院, 北京 100081)

通信作者: 李长升, E-mail: lcs@bit.edu.cn

摘 要: 在大数据时代, 人工智能得到了蓬勃发展, 尤其以机器学习、深度学习为代表的技术更是取得了突破性进展. 随着人工智能在实际场景中的广泛应用, 人工智能的安全和隐私问题也逐渐暴露出来, 并吸引了学术界和工业界的广泛关注. 以机器学习为代表, 许多学者从攻击和防御的角度对模型的安全问题进行了深入的研究, 并且提出了一系列的方法. 然而, 当前对机器学习安全的研究缺少完整的理论架构和系统架构. 从训练数据逆向还原、模型结构反向推演、模型缺陷分析等角度进行了总结和分析, 建立了反向智能的抽象定义及其分类体系. 同时, 在反向智能的基础上, 将机器学习安全作为应用对其进行简要归纳. 最后探讨了反向智能研究当前面临的挑战以及未来的研究方向. 建立反向智能的理论体系, 对于促进人工智能健康发展极具理论意义.

关键词: 反向智能; 人工智能安全; 逆向还原; 缺陷分析

中图法分类号: TP18

中文引用格式: 李长升, 汪诗烨, 李延铭, 张成喆, 袁野, 王国仁. 人工智能的逆向工程——反向智能研究综述. 软件学报, 2023, 34(2): 712–732. http://www.jos.org.cn/1000-9825/6699.htm

英文引用格式: Li CS, Wang SY, Li YM, Zhang CZ, Yuan Y, Wang GR. Survey on Reverse-engineering Artificial Intelligence. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 712−732 (in Chinese). http://www.jos.org.cn/1000-9825/6699.htm

Survey on Reverse-engineering Artificial Intelligence

LI Chang-Sheng, WANG Shi-Ye, LI Yan-Ming, ZHANG Cheng-Zhe, YUAN Ye, WANG Guo-Ren

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: In the era of big data, artificial intelligence, especially the representative technologies of machine learning and deep learning, has made great progress in recent years. As artificial intelligence has been widely used to various real-world applications, the security and privacy problems of artificial intelligence is gradually exposed, and has attracted increasing attention in academic and industry communities. Researchers have proposed many works focusing on solving the security and privacy issues of machine learning from the perspective of attack and defense. However, current methods on the security issue of machine learning lack of the complete theory framework and system framework. This survey summarizes and analyzes the reverse recovery of training data and model structure, the defect of the model, and gives the formal definition and classification system of reverse-engineering artificial intelligence. In the meantime, this survey summarizes the progress of machine learning security on the basis of reverse-engineering artificial intelligence, where the security of machine learning can be taken as an application. Finally, the current challenges and future research directions of reverse-engineering artificial intelligence are discussed, while building the theory framework of reverse-engineering artificial intelligence can promote the develop of artificial intelligence in a healthy way.

Key words: reverse-engineering artificial intelligence; artificial intelligence security; reverse recovery; defect analysis

随着大数据的爆炸式增长和图形处理器(GPU)等算力基础设施的迅猛发展, 以机器学习[1]、深度学习为代表的人工智能技术获得了长足的进步. 机器学习是一种数据驱动的智能计算技术, 它的目标是基于给定的训

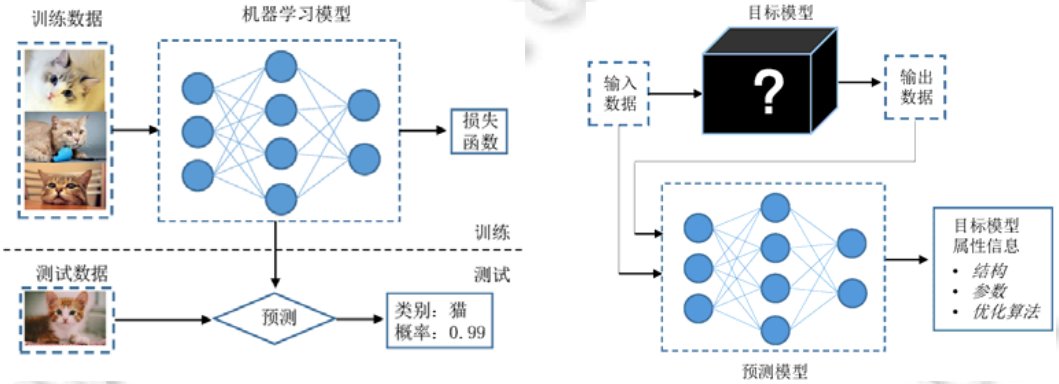
* 基金项目: 国家自然科学基金优秀青年科学基金(62122013); 国家自然科学基金广东联合基金重点项目(U2001211); 北京理工大学青年教师学术启动计划(3070012222010)

收稿时间: 2022-01-25; 修改时间: 2022-03-03, 2022-04-12; 采用时间: 2022-04-29

练集学习一个模型, 并利用该模型对未知数据进行预测, 如图 1(a)所示. 随着机器学习、深度学习等技术的突破性发展, 人工智能已被广泛应用于各个领域, 如图像识别、物体检测、目标跟踪、蛋白质结构预测、自动驾驶、推荐系统等等. 在这些真实场景中, 部署的机器学习模型通常以“黑盒”的形式对外提供服务: 用户给模型一个输入, 模型给用户返回一个输出. 通过如此的部署方式, 模型拥有方试图隐藏模型内部的细节信息, 例如模型结构、模型参数、优化算法等, 从而起到对模型知识产权和隐私数据保护的作用.

近几年, 人工智能的安全与隐私问题日益突显. 例如: 模型攻击者通过对测试数据添加轻微扰动, 就可以导致机器学习模型预测出错; 面向隐私的攻击者利用目标模型的输入和输出计算出用户的训练数据或者隐私数据, 降低模型的隐私性等. 到目前为止, 一大批来自学术界和工业界的学者对人工智能模型安全与隐私问题进行了研究, 并且先后提出了各种各样的模型攻击技术, 包括数据投毒、对抗样本攻击、后门攻击、模型萃取攻击、成员推理攻击等.

然而, 上述研究攻击模型的角度各有不同, 提出的解决方法也是各有侧重. 因此, 建立统一的理论体系对于人工智能安全研究起着至关重要的作用. 基于上面考虑, 本文采用逆向思维, 提出反向智能的概念, 从数据、模型等角度探讨人工智能反向推演的方法与机制, 从而完成人工智能逆向工程的目标, 为人工智能模型攻击和防御提供技术支持. 图 1(b)给出了一个典型的反向机器学习(本文主要以机器学习模型为研究对象, 因此在介绍反向智能时主要以反向机器学习为重点介绍对象)的流程.



(a) 正向机器学习 (b) 反向机器学习

图 1 正向与反向机器学习

如图 2 所示, 本文从如下 4 个层次来介绍反向智能技术.

(1) 第 1 层, 训练数据还原引擎对目标模型涉及的训练数据进行推断. 其一, 调研预测某一数据是否存在于训练数据的相关方法; 其二, 调研对训练数据的属性进行推断的技术; 其三, 讨论对训练数据的分布进行推断的技术. 总之, 本节从数据的多个维度出发, 对训练数据的反向推断技术进行了调研, 为下一层的模型反向推演引擎提供基本资料(见第 1 节).

(2) 第 2 层, 反向推演引擎对目标模型的细节进行推演. 其一, 调研对机器学习模型结构(例如深度神经网络的隐含层数、激活函数等)进行反向推演的技术; 其二, 调研对机器学习模型参数进行反向推演的技术; 其三, 调研对机器学习模型功能进行反向推演的技术; 本节调研了对机器学习模型内部细节进行反向推演的技术, 为下一层的机器学习模型缺陷分析引擎提供技术支撑(见第 2 节).

(3) 第 3 层, 模型缺陷分析引擎对目标模型的缺陷进行分析挖掘. 其一, 调研对机器学习模型鲁棒性分析的技术; 其二, 调研对数据不均衡性分析的技术; 其三, 调研对机器学习模型敏感性分析的技术; 本节调研了对机器学习模型缺陷分析的技术, 为下一层的机器学习模型攻击和防御提供技术依据(见第 3 节).

(4) 第 4 层, 机器学习模型攻击和防御引擎完成模型的对抗攻击和防御, 实现反向机器学习的最终目标.

其一，调研机器学习模型安全攻击的技术；其二，调研机器学习模型防御机制的技术(见第 4 节).

图 2 反向智能技术体系

最后给出现有反向智能的挑战及未来发展方向(见第 5 节).

本文的主要贡献如下：明确提出反向智能的概念，给出了反向智能的技术体系及流程框架. 不同于已有的综述文献[2,3]主要关注机器学习安全攻击和防御的研究，本文创新性地从训练数据逆向还原、模型反向推演、模型缺陷分析、模型攻击与防御这 4 个方面综述反向智能的技术，并给出了未来反向智能技术的挑战和研究方向.

1 训练数据逆向还原引擎

机器学习模型是数据驱动的智能模型，其训练数据中往往包含着与用户相关的隐私信息[4]. 例如在医疗系统中，机器学习模型的训练数据可能会包含用户的隐私信息. 如果这些数据被泄露出去，将会对人们的隐私造成危害. 因此，从数据层面的角度出发，训练数据还原技术对反向机器学习具有重要的意义. 本节将从成员推断技术、属性推理技术和分布推断技术这 3 个角度出发，剖析训练数据还原引擎的细节，即如何针对不同情况设计训练数据逆向还原算法.

1.1 成员推断技术(member inference)

成员推断的目标是，利用机器学习技术自动判断查询样本是否属于某目标模型对应的训练数据集. 该类问题由 Shokri 等人[5]首先提出，其核心思想是，通过分析目标模型在训练集和非训练集上的表现差异进行成员推断，即利用目标模型泛化能力的缺陷对查询样本是否属于训练数据进行推断. 目前常用的成员推理方法的架构如图 3 所示，该框架流程如下：首先，采用与目标模型训练数据同分布的数据集去训练一个影子模型(shadow model)，即一个在功能上与目标模型近似的模型；其次，利用影子模型的输入和输出训练推理模型(通常为分类模型)，通过推理模型可以判断数据是否属于影子模型对应的训练数据；最后，在推理阶段，将待查询的数据样本送入目标模型，基于目标模型得到模型输出. 将输入和输出送入推断模型，即可判断查询数据是否属于目标模型对应的训练数据集. 从目标模型的角度出发，现有的成员推断技术可以分为黑盒成员推断方法和白盒成员推断方法.

- 黑盒成员推断方法是指在目标模型是黑盒模型的情况下进行成员推断，即可用信息仅有通过对目标模型发起查询请求获得输入输出对，除此之外无法获得与目标模型相关的其他任何信息. 在黑盒成

员推断方法中，根据黑盒模型的输出情况，推断方法又可以分为两类：一类是利用预测置信度进行推断；另一类是利用标签进行推断. 基于预测置信度的方法是成员推断方法中的经典方法. 以分类场景为例，预测置信度是指目标模型输出的分类概率向量. 从泛化性的角度出发，该类方法假设对于成员数据，目标模型输出的分类向量的熵相对较低，即对正确类的分类概率值会比较高；而对非成员数据，其输出的分类向量的熵会比较高，即没有相对较高的分类概率值. 文献[5]中提出的方法是该类成员推断技术中的一个经典方法，该方法在针对图像分类器进行成员推断时取得了良好的效果. Melis 等人[6]在协同学习的背景下，设计了一种针对协同式深度学习系统的成员推断技术；Song 等人[7]在进行成员推断的同时，将对抗鲁棒性考虑了进来，丰富了成员推断领域的方法. 然而在许多真实场景中，部署好的机器学习模型通常直接输出样本的预测标签，不会返回预测置信度，导致基于置信度的成员推断方法无法工作. 因此，一些基于标签的方法逐渐进入了人们的视野. Yeom 等人[8]在这一设定下提出了一个相对简单的推理方法，该方法将目标模型预测正确的数据归为成员数据(训练数据)，将其预测不正确的数据归为非成员数据(非训练数据). Christopher 等人[9]在文献[8]的方法上作了一些改进，通过分析目标模型对增强后数据的识别能力，从而判断该数据是否为目标模型的训练数据. 文献[10]通过使用多个数据集对各种机器学习模型进行系统评估，进而更全面地揭示了成员推断方法的效果与模型之间的关系. 文献[11]从差分隐私的角度出发，分析了差分隐私和数据偏度技术对成员推断方法的影响. 不过，相较于基于预测置信度的成员推理方法，由于该类方法可用的信息较少，因此在效果上还有较大上升空间.

- 白盒成员推断方法是指在目标模型为白盒模型情况下的成员推断方法，即目标模型的结构、参数、训练方法、超参数设计等信息均是可知的. 相较于黑盒成员推断方法，该类方法可以利用的信息更多，因此成员推断任务相对容易. Hayes 等人[12]针对 GAN 网络在白盒模型的设定下设计了一种成员推断方法，仅利用 GAN 网络中判别器的部分输出结果即可完成成员推断. Nasr 等人[13]通过利用目标模型的梯度信息设计了一种白盒成员推断方法，并且通过分层的方式将目标模型每一层的梯度信息分别输入到推断模型的若干卷积层和全连接层当中. 同时，由于更多目标模型相关信息的参与，该方法在效果上要相对优于黑盒推断方法. 例如在 CIFAR 数据集[14]上，针对基于 DenseNet[15]的分类模型，该方法的推断准确率相较于其对比的黑盒推断方法高出了 6.6%. 尽管白盒成员推断方法能够取得较好的推断效果，但是由于在现实场景中，许多机器学习模型通常部署为黑盒模型，因此该类方法的应用受到一定的局限.

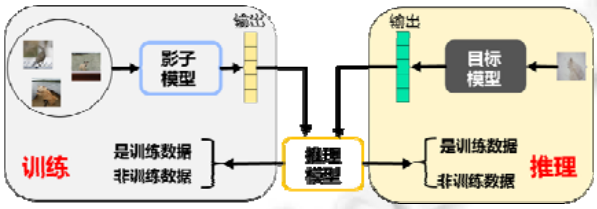


图 3　成员推断技术的架构

1.2　属性推断技术

属性推断技术是指通过设计机器学习模型来推断目标模型中训练数据的敏感隐私属性，例如推断某一类数据在训练集中所占的比例，其一般做法如图 4 所示：首先，利用与目标模型训练集分布相近的数据集训练一些影子模型；其次，利用影子模型构建一个元分类器(meta-classfier)，建立模型与属性的映射关系；最后，将目标模型输入到元分类器中，可以直接预测目标模型对应的数据属性. 该技术由 Ateniese 等人[16]首次提出，实验结果表明：所提出的方法在支持向量机(SVM)和隐马尔可夫模型(HMM)上具有较好的推断效果，并且在一定程度上能够使差分隐私技术失效. 但是该方法在复杂模型上仍有较大的提升空间，例如神经网络模型.

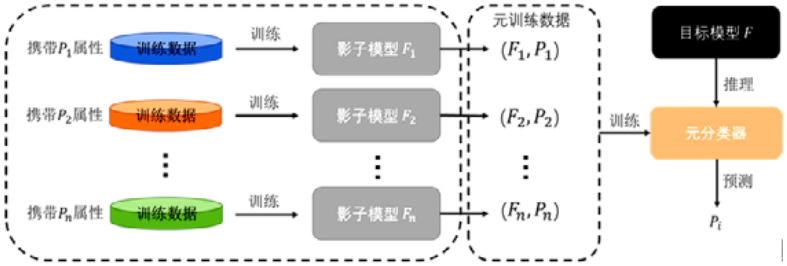


图 4 属性推断技术实例[16]

为了进一步提升属性推断技术在复杂模型上的效果，Ganju 等人[17]提出了一种针对全连接神经网络的属性推断方法. 该方法的出发点为当节点的排列用矩阵表示时，全连接网络是不变的，即对全连接层的每个隐含层，使用任意排列并相应地调整权重会产生等效的全连接神经网络. 因此，该方法首先将全连接神经网络用矩阵表示成一个规范形式，以便所有等效排列的全连接神经网络产生相同的特征表示，并将全连接神经网络每一层的输出表示为一个集合而不是一个向量；然后，使用 DeepSets[18]框架构建元分类器. 实验结果说明，该方法在一些浅层的全连接神经网络上具有良好的推断效果. 随着研究的深入，人们开始尝试从不同的角度或者针对不同的问题设计属性推断的方法. 例如：Melis 等人[6]提出了一种在协同深度学习模式下进行属性推断的方法; Gopinath 等人[19]从神经元的激活状态出发，设计属性推断算法.

模型逆向(model inversion)作为一类特殊的属性推断技术，在近些年也受到了研究者的关注. 模型逆向的任务目标是，重建与特定目标标签相对应的数据特征[20]. 第一个模型逆向方法由 Fredrikson 等人[21]提出，该方法在线性模型上能够取得很好的效果. 文献[22]提出了一种针对浅层神经网络的模型逆向方法，并且该方法能够在一定程度上重构出与身份标签对应的人脸信息. 文献[23]将还原训练数据的分布作为了模型逆向的目标，进一步推动了模型逆向技术的发展.

1.3 分布推断技术

分布推断技术可以看作是属性推断技术的一种特例，其目标是预测出目标模型训练数据的分布情况. 这里，训练数据的分布本身可以看作训练数据的一种属性. 在观测数据可得的情况下，数据集的分布估计根据是否已知数据的具体分布形式分为参数估计和非参数估计.

- 当已知数据的分布形式时，例如知道数据服从高斯分布，接下来的任务就是通过观察数据样本估计高斯分布的具体参数，即均值μ和方差σ(或者协方差矩阵). 其中，极大似然估计[24]或者贝叶斯估计[25]是参数估计中比较常见的方法.
- 针对非参数估计，由于不知道数据的具体分布形式，因此相较于参数估计，其方法也相对复杂一些. 目前常见的非参数估计的方法有核密度估计[26]、最近邻估计[27]等等.

上述方法的前提通常要求数据是可观测的，然而在反向智能的场景下，目标模型的训练数据通常是未知的或者仅有小部分是可知的，因此在这种情况下，对训练数据的分布进行估计挑战更大. 根据目前已有的工作，可以大致将其分为训练数据部分已知和无训练数据两种情况.

- 在已知部分训练数据的情况下，可以利用生成对抗网络(GAN)[28]或者变分自动编码器(VAE)[29]等生成模型对训练数据的分布进行估计. 例如，Wang 等人[30]在已知部分训练数据的前提下，提出了一种利用单个训练数据样本生成其他训练样本的方法，并且为了提升生成效果，该方法还引入了判别器进行对抗训练，在其生成训练数据时，能够模拟训练数据的分布情况.
- 在训练数据完全不可知的情况下，对训练数据分布进行估计难度会更大一些. 由于训练数据分布估计和模型逆向任务有相近的目标，因此近年来出现了一些在模型逆向场景下进行训练数据分布估计的工作. Chen 等人[23]在模型逆向的背景下设计了一种推断数据在隐空间分布的方法，如图 5 所示. 该方法属于参数估计的一种. 其假定训练数据在隐空间中服从高斯分布，这里，待估计的参数即为μ和

σ. 为了直接利用反向传播算法更新μ和σ, 该方法采用重参数的技巧生成隐变量, 实现了端到端的训练方式估计训练数据的分布. Kuan-Chieh 等人[31]也在模型逆向的场景下设计了训练数据分布推断的方法, 他们利用变分推断技术, 近似估计训练数据的类别条件概率分布, 同时利用生成对抗网络(GAN)在公开数据集上获取先验知识, 从而提升模型的推断效果.

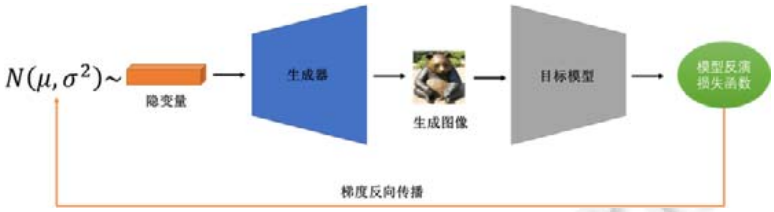


图 5 分布推断技术实例[23]

由于当前大部分训练数据分布推断的技术[32−35]都要涉及到目标模型的内部知识, 因此目标模型通常是白盒(white-box)模型. 为了提升该技术的灵活性, 在黑盒(black-box)模型场景下进行训练数据分布推断, 也是未来值得探索的研究方向.

2 智能模型反向推演引擎

机器学习即服务已经演化成为一种重要的商业模式. 互联网公司通过部署公共可访问的模型调用接口来满足用户对各类机器学习模型的使用需求, 如 Google 的多种预测 API、亚马逊机器学习(AmazonML)、MicrosoftAzure 机器学习(Azure ML)等. 然而, 机器学习模型的机密性要求机器学习系统必须保证未授权用户无法接触到模型的隐私信息. 例如模型架构、模型参数、训练方式等. 这种模型机密性和公共可访问之间的关系激发了智能模型反向推演技术的研究. 此外, 深入分析和研究智能模型反向推演技术也有助于促进其他相关领域的研究. 例如, 借助反演的模型, 可以生成对抗样本[22,36,37]使目标模型失效、可以进行成员推理[5,8]损害目标数据隐私、可以借助模型反演[23,31]泄露敏感训练数据等等. 本节以训练数据逆向还原的结果为基础, 依据对模型的不同反向推演目标, 将智能模型反向推演引擎分类为模型结构反向推演、模型参数反向推演、模型功能反向推演.

2.1 模型结构反向推演

模型结构反向推演旨在设计合理的机器学习方法, 精确反演出目标黑盒模型的结构信息. 其中, 模型结构信息主要是指神经网络的架构拓扑. 实际上, 反向推演模型的结构极具挑战性. 因为任何单个模型都属于一个大的等价类网络, 仅仅依靠输入和访问 API 获取模型的输出通常很难精确区分网络, 因而现有工作通常假设能够获取目标黑盒模型的部分知识, 或是从硬件角度出发获取更细粒度的可用于推断模型结构的信息.

文献[38]提出: 可以通过一系列查询, 揭露神经网络的模型结构信息. 其首次建立了黑盒模型和白盒模型之间的联系. 该方法将模型反演问题归结为机器学习中常见的监督学习任务, 通过设计元模型学习不同神经网络结构和对应网络输出的映射关系, 从而实现通过目标黑盒模型的输出, 完成对目标模型结构属性的预测. 如图 6 所示: 在元训练阶段, 首先收集一系列结构类型多样的白盒模型作为“元训练集”, 对这些白盒模型, 利用训练数据逆向还原出来的训练数据对它们进行训练, 使其在一定程度上趋近于目标模型. 在元训练集构建完成后, 基于该元训练集上训练“元模型(meta-model)”, 建立白盒模型输出和结构属性之间的关系; 在推理阶段, 给定查询目标黑盒模型得到的输出, 借助元模型预测目标模型的结构属性. 上述模型结构反演方法需要构造大量的白盒模型及其输出, 因此需要大量的计算成本. 该方法在简单的神经网络结构反向推演上取得了不错的效果, 然而其在复杂网络上的效果仍有待检验.

另一种解决方案是从硬件角度(例如缓存侧信道、总线监听等)反演模型架构信息. 实际上, 当 DNN 模型在计算机上执行推理任务时, 在底层硬件上会留下依赖于架构的轨迹, 因而可以通过分析这些轨迹, 实现模

型架构细节的恢复. 这些技术可以提供非常细粒度的信息, 并且已有一些工作[39−43]将其用于提取传统 DNN 模型的架构. 文献[43]首先提出利用未加密的 PCIe 总线来反演 DNN 模型, 这是第一个完全反演 DNN 模型结构信息的工作. 通过大规模的逆向工程和可靠的语义重建, 其反演出的 DNN 模型与原始模型具有相同的架构拓扑. 不同于以往主要集中在反演相对简单的 DNN 模型, NASPY[44]提出一种端到端的模型用于反演由神经架构搜索(NAS)生成的深度学习模型的网络架构, 如图 7 所示. NASPY 利用硬件侧信道序列(side-channel sequences), 引入 seq2seq 模型以自动识别复杂的操作(例如可分离卷积、扩张卷积). 此外, 该方法还利用总线监听(bus snooping)进一步反演完整的网络拓扑结构. 基于此, NASPY 能够精确地提取完整的 NAS 模型架构.

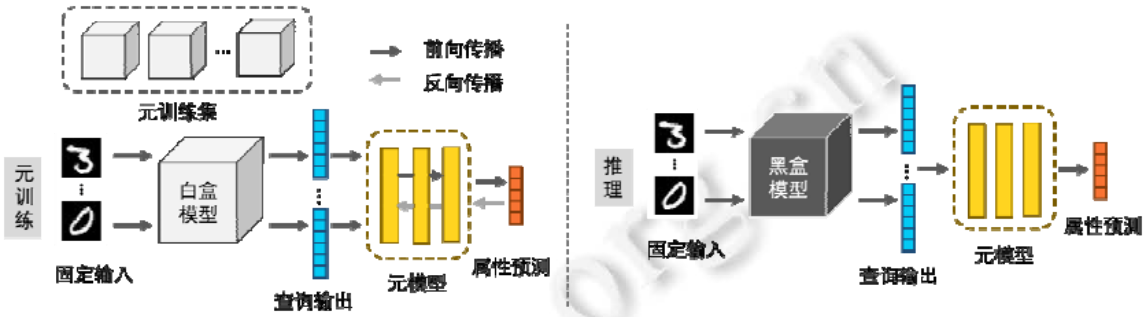


图 6 黑盒神经网络逆向工程[38]

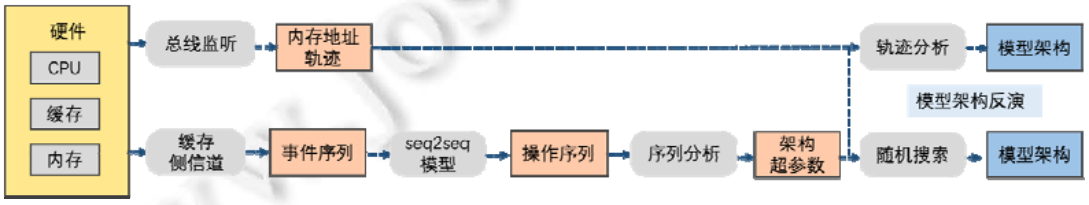


图 7 NASPY 框架的工作流[44]

2.2 模型参数反向推演

本小节将模型参数反向推演细化为网络参数反向推演、模型超参数反向推演. 其中, 网络参数通常是目标模型使用专有数据进行充分训练得到的, 因此具有昂贵的知识产权; 而模型超参数是搭建神经网络前需要提前设置的超参数.

网络参数反向推演是指在已知目标模型的结构信息但模型参数未知的情况下, 通过多次查询目标模型以反向推演模型参数. 文献[45]提出了等式求解的模型参数反演方法, 对于逻辑回归(LR)、支持向量机(SVM)、神经网络(NN)和决策树这类由一系列参数决定的智能模型, 通过反复请求目标模型的预测标签就能反解模型网络参数. 对于一个输入为 n 维的线性模型, 理论上只需要通过对目标模型进行 n+1 查询就可以反演模型. 类似解方程的思路, 一个 n 维的权重向量和一个偏置向量, 则至少需要 n+1 个等式就可以求解方程.

反演模型超参数的一种方法是采用黑盒模型逆向工程技术, 通过合理构建元模型去捕获目标黑盒模型具体使用的优化算法[38]. 文献中构造大量白盒模型并预设多种优化器(如 AdaGrad[46]、SGD[47]、Adam[48]、RMSprop[49]等)分别进行优化, 利用元模型去捕获模型使用不同优化器的输出结果与对应优化器的映射关系. 此外, 研究学习优化器逆向工程可用于分析和理解优化器的一些外在表现[50]. 文献[50]将优化器视为一个动态系统, 并建立在逆向工程递归神经网络(RNN)这类工作的基础之上, 其中, Sussillo 和 Barak[51]展示了非线性 RNN 的线性近似如何揭示训练简单网络所使用的优化算法. 文献[50]借鉴使用神经网络参数化优化器[52]的思路, 使用循环神经网络(RNN)对学习优化器进行参数化. 通过分析和可视化优化器, 发现其能够有效学习到现有经典优化算法(momentum[53]、AdaGrad-RDA[54]、RMSProp、Adam)的一些外在表现, 如动量、梯度裁剪、学习率调度和学习率自适应等.

机器学习中, 不同的超参数通常会导致模型的性能显著不同. 这类超参数通常需要通过交叉验证进行确定, 如果机器学习算法采用多个正则化项, 则可能有多个超参数, 导致调参过程十分耗时. 文献[55]利用目标函数最小值处的参数梯度等于 0 这一性质, 提出一个通用的框架对目标函数正则项的超参数进行反演. 该方法将目标函数的梯度也设置为 0, 以得到关于超参数的线性方程组. 由于这个线性方程组通常是超定的, 因为方程的数量(即模型参数的数量)通常大于未知变量(即超参数)的数量, 故其利用线性最小二乘法[56]估计超参数.

2.3 模型功能反向推演

模型功能反向推演[57]指的是对于给定的目标机器学习模型, 训练与目标模型具有相同功能的替代模型(或克隆模型). 其主要过程分为两步(如图 8 所示): (1) 查询目标黑盒模型以构建替代模型需要的训练数据; (2) 基于构造的数据集设计合理的机器学习算法训练克隆模型, 从而能够反向推演目标模型的功能. 根据模型功能反向推演中目标数据的可用性, 现有的工作可分为 3 类: 同分布数据可得, 即拥有与目标数据同分布的部分数据集; 代理数据可得, 即拥有与目标数据语义相似的替代数据集; 数据不可得, 即无法访问与目标数据相关的数据.

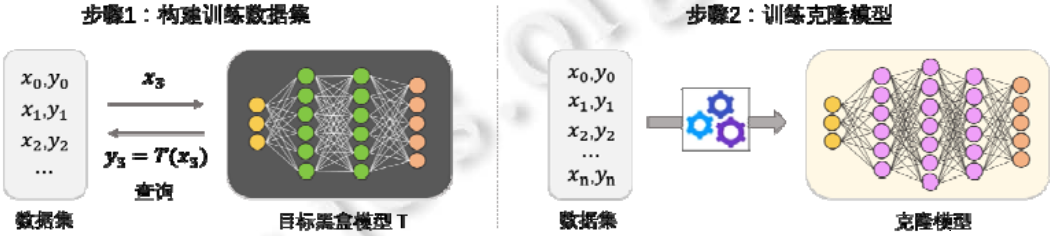


图 8 模型功能反向推演的一般流程

在同分布数据可得的方法中, 通常假设目标数据集的部分数据可得并提出了一系列方法. 例如: JBDA[58]假设可以访问来自目标数据分布的一小组种子样本, 并使用该子集训练克隆模型. 其采用基于扰动的启发式算法, 依据现有注释数据创建额外的合成数据, 并查询目标模型以进行标注. 将合成的注释数据添加到训练数据集中, 用于迭代训练克隆模型. 目前, JBDA 仅适用于 MNIST 等简单的数据集, 在更复杂的数据集上, 分类精度仍有待于提高.

在代理数据可得的方法中, 研究者使用替代数据集来查询目标模型[59,60]. Shi 等人[61]使用深度学习方法, 通过向目标模型输入样本获得输出来对样本进行标注, 用于训练替代模型. 对应的实验结果表明, 该方法可以推断并构建朴素贝叶斯与 SVM 分类器对应的高精度替代模型. ActiveThief[59]使用未标注的公共数据作为替代数据, 借助主动学习并设计合理的子集选择策略从替代数据中选择查询样本, 从而实现在有限的预算内尽可能有效地减少查询次数. KnockoffNets[60]假设拥有与目标数据集具有语义相似的代理数据(例如, 使用 CIFAR-100 作为 CIFAR-10 的代理数据). 该方法结合主动学习[62]并设计合理的基于奖励的采样策略, 从代理数据中选出代表性样本查询目标模型以进行标注, 能够有效地提高样本的查询效率. 最后, 该方法利用标注的代表性样本集, 借助知识蒸馏[63]的方法训练克隆模型, 达到反演目标模型功能的目的. 然而, 该类方法的有效性取决于替代数据的选择是否合适. 如果替代数据不能很好地描述目标数据, 或者与目标数据的相关性较弱, 则会很大程度上降低模型功能反演的性能.

目标数据不可得是研究者最近提出的新的问题设置, 属于相对较难的一类问题. MAZE[64]使用生成器创建合成数据, 并提出一种结合数据不可得的知识蒸馏[65,66]和零阶梯度估计[67,68]的方法, 实现高精度克隆模型功能的反演目标. 如图 9 所示, MAZE 利用生成器合成的数据训练克隆模型, 优化克隆模型的决策边界与目标模型的决策边界, 从而使两者对齐, 产生高精度的克隆模型. 在训练生成器的过程中, 由于对目标黑盒模型无法执行反向传播, MAZE 利用优化零阶梯度估计的方法近似黑盒目标模型的梯度, 从而正常执行训练过程. 文

献[69]也提出通过生成模型合成查询数据来解决目标数据不可见的问题，其与 MAZE 的主要区别在于损失函数的选择以及优化方法. 并且，该方法观察到计算损失的稳定性至关重要，并发现 $l1$ 正则损失特别有利于解决这类目标数据不可得的模型功能反演问题.

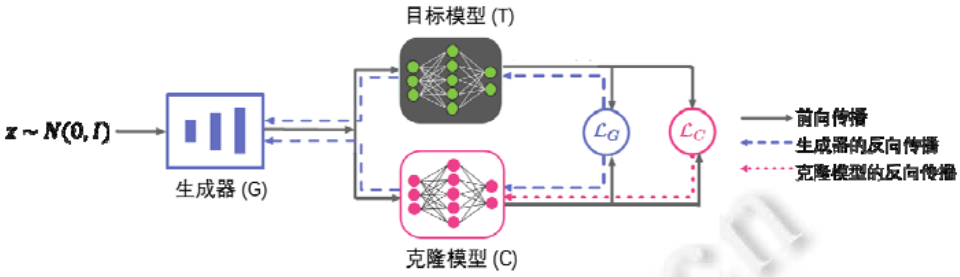


图 9 MAZE 的总体框架[64]

此外，在模型功能窃取时，基于训练数据逆向还原的结果，将会对提升窃取效果有很大的帮助. Gong 等人[70]在利用主动学习对训练数据进行了一定程度还原的基础上，进一步提升了替代模型与目标模型之间的相似性，极大地提升了模型功能窃取的效果.

3 机器学习模型缺陷分析引擎

在获得机器学习模型训练数据并将对应模型具体结构、参数与功能反向推演成功后，以黑盒形式部署的模型已一定程度上被“白盒化”. 在此基础之上可进行模型的缺陷分析，基于还原成功的数据与模型内核挖掘模型内部缺陷，进一步反向分析模型漏洞. 因此，本节将从模型鲁棒性、数据不均衡性、模型敏感性这 3 个角度对机器学习模型缺陷分析相关工作进行总结归纳.

3.1 模型鲁棒性分析

机器学习模型鲁棒性是衡量模型缺陷的重要因子之一. 基于模型反向推演的模型结果，衡量一个模型的鲁棒性强弱，则需要具体的、可量化的评估指标. 目前，评估模型鲁棒性所选择的评价指标主要集中在模型鲁棒性边界上. 模型鲁棒性边界是指在模型对某一输入样本推理结果正确的情况下，对该样本可施加的最大扰动范围. 可以看到，模型鲁棒性边界与具体样本有关. 现有的模型鲁棒性分析方法主要可分为基于可满足性模理论与整数线性规划的方法、基于凸松弛的方法、基于 Lipschitz 常数的方法与基于随机平滑的方法.

基于可满足性模理论与整数线性规划的方法最初由 Ehlers 等人[71]提出，作者使用分段线性函数对以 ReLU 作为激活函数的前馈神经网络对应行为进行近似处理，并使用可满足性模理论(SMT)与整数线性规划(ILP)对该线性函数进行求解. 此后，Katz 等人[72]提出了基于 ReLU 约束的线性实数算法的 SMT 求解方法. 该方法通过采用非凸优化的单纯形算法，保证了 ReLU 函数的线性特征. Tjeng 等人[73]则将模型鲁棒性的量化转化为混合整数线性规划(MILP)问题进行求解，并使用非线性公式与对应的预求解算法缩小解空间，进行计算加速，完成对神经网络的属性验证. 与其他方法相比，这类方法具有较高的精确度. 然而，由于该类方法基于线性规划或 SMT 等相关线性优化理论，因此仅支持分段线性神经网络的鲁棒性分析. 当模型中含有除 ReLU 等分段线性之外的非线性部分时，则难以使用这类方法. 此外，该类方法的运算时间复杂度较高，在最坏情况下，其运行时间与网络大小的指数成正比，因此只能应用于规模小的神经网络，对复杂的神经网络模型仍需进一步加以改进.

基于凸松弛的方法主要思路为：将机器学习模型训练近似为线性规划等凸优化问题，通过凸优化进一步分析模型的鲁棒性. 例如：DeepPoly[74]以多个带有间隔的浮点-点多面体(floating-point polyhedra)作为抽象域，用于近似神经网络的每一层. 与基于 SMT 和 ILP 的方法相比，该方法平衡了精度与可扩展性，并适用于非 ReLU 函数以外的激活函数，同时可对 ResNet 等残差网络进行分析. 在此之后，如图 10 所示，Salman 等人[75]

提出了一个分层的凸松弛框架. 该框架统一了包括 DeepPoly 等基于凸松弛的方法, 并证明了在此框架中, 即使是最优的 ReLU 网络凸松弛, 也存在性能的上界. 换言之, 基于分层凸松弛的模型鲁棒性方法具有精确性上限. Xu 等人[76]则提出了一个自动分析框架, 将现有的凸松弛相关方法推广到一般的计算图上, 使得它对一些复杂网络同样适用. 此外, 该方法加入了损失融合技术, 进一步降低了相关算法的计算复杂度.

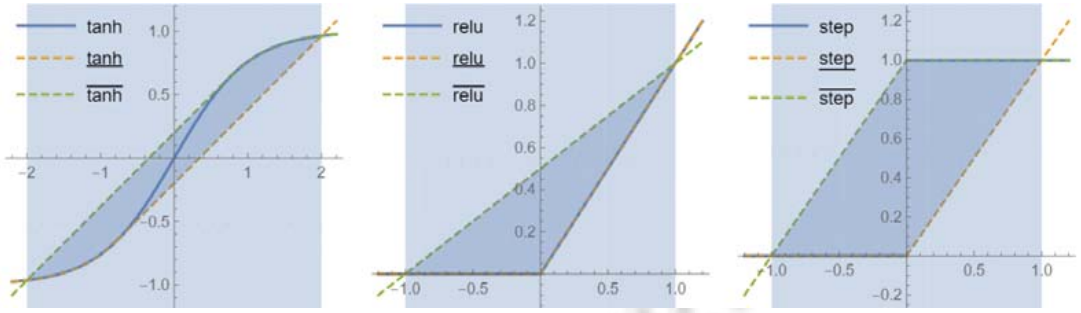


图 10 几种常见激活函数的最佳凸松弛[75]

Lipschitz 常数是针对函数定义的, 对一个函数而言, 其 Lipschitz 常数为其图像上两点连线斜率绝对值的上界. 该常数可在一定程度上衡量函数变化的剧烈程度, 因此也常用于模型鲁棒性分析. Andriushchenko 等人[77]提出了 Cross-Lipschitz 正则化方法, 该方法通过计算模型局部的 Lipschitz 常数, 对模型分类边界的上界与下界进行了估计. 相比通过全局 Lipschitz 常数进行鲁棒性分析的方法而言, 该方法可以给出更精准的分类边界. 然而这种方法很难适用于具有多隐含层的神经网络. Zhang 等人[78]提出了 RecurJac 方法, 该方法通过计算神经网络雅可比矩阵中每个元素的上界与下界, 求得模型的局部 Lipschitz 常数, 进而求得模型分类边界.

基于随机平滑的方法最早可以追溯到 Lecuyer 等人[79]的工作, 他们提出的 PixelDP 方法本质上是在输入样本的邻域多次采样, 得到多个临近样本, 并以这些临近样本预测结果的平均值作为输入样本, 从而得到最终的预测结果. 此后, Cohen 等人[80]基于 Neyman-Pearson 引理给出了在 $l2$ 范数随机高斯噪声下更为紧密的鲁棒性边界估计方法, 同时适用于在 ImageNet 等大型数据集上训练的复杂深度学习模型. 最近, Yang 等人[81]基于 Banach 空间理论, 进一步探讨了在多种随机分布噪声下随机平滑鲁棒性分析精度的上界, 将基于随机平滑的鲁棒性分析方法作了进一步的拓展.

3.2 数据不均衡分析

数据不均衡是导致机器学习模型存在缺陷的另一因素. 目前, 数据不均衡是机器学习领域的热点研究问题之一. 在现实生活中, 数据的分布通常都是不均衡的, 某一类或者某几类的样本数量稀少, 也就是数据具有长尾分布, 如在医疗诊断领域中, 正常样本数往往远大于患特定疾病的样本数. 在这种情况下, 受训的机器学习模型常常会倾向于将输入样本分类为占多数样本的头部类别, 而在数据有限的尾部类别上表现不佳. 目前的研究主要围绕在剖析上述数据缺陷、解决类别不均衡上面, 具体方法可分为以下 3 类: 类别再平衡、信息增强、模型改进.

类别再平衡是解决该问题的主流方式, 其主要思路为: 平衡模型训练过程中各个类的训练样本数, 防止模型由于不同类别上的样本数量不平衡而在分类性能上有所下降. 其中最为经典的方法为基于重采样的再平衡方法, 具体内容为: 对少数类的过采样[82]、多数类的欠采样[83]. 举例而言, Wang 等人[84]提出了用于类间平衡的动态采样框架 DCL. 其具体思路如图 11 所示, 以少数类中易分类的样本为锚点, 将该类中难分类的样本与锚点样本的距离拉近, 同时将其他类中难分类的样本与该锚点样本的距离推远. Wu 等人[85]将数据的长尾分布问题引入对抗鲁棒性研究领域, 并使用两阶段重平衡策略, 根据训练标签频率调整模型分类边界. 虽然这样的方式简单直观, 可在一定程度上解决数据不均衡问题, 但这样的平衡方式一方面舍弃了部分多数类样本, 导致部分有用信息没有得到充分的利用; 另一方面, 对少数类样本进行了重复利用, 也增加了模型过拟合的

风险.

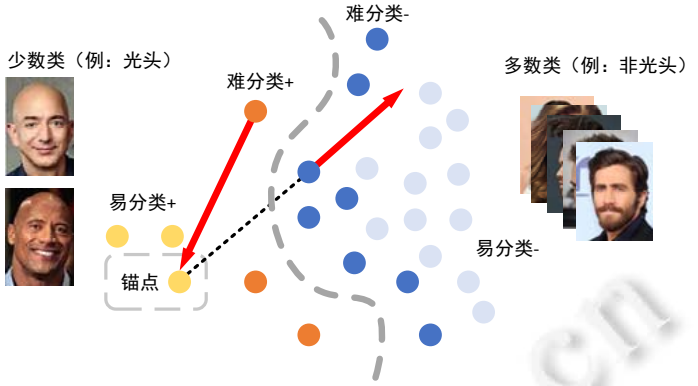


图 11 DCL[84]的具体思路

基于类别再平衡思想的方法本质上没有为模型提供更多有用的信息，甚至在一定程度上降低了信息的利用率. 因此，一部分研究人员尝试在模型训练过程中引入额外的信息，以便于在数据不均衡的背景下提升模型性能. 常用的方法有迁移学习和数据合成. 迁移学习相关方法的基本思路为：分别为多数类和少数类进行数据建模，将多数类中学到的知识迁移到少数类中. Yin 等人[86]发现，少数类样本的类内方差远小于多数类样本的类内方差. 因此，他们利用多数类样本的类内方差知识，对少数类样本进行特征增强，使其方差水平与多数类样本一致，增强模型在少数类上的分类性能. 对数据合成相关方法而言，其本质为根据已获取的数据样本相关特征，自主生成和少数类样本类似的新数据. 其中，早期最具有代表性的为 Chawla 等人[87]所提出的 SMOTE 方法，该方法对所选择的少数类样本，首先使用 K 近邻算法选取与其邻近的样本；在此基础上，通过线性插值，合成新的属于该少数类的样本. 这些方法对模型可利用的信息进行了扩充，但其扩充的信息具体内容是否合理，如合成数据是否符合对应自然数据的分布，仍需进一步研究.

样本数量上的不均衡，会导致特征提取过程中，信息保留不够充分，进而影响少数类的分类边界. 因此，对在数据不均衡下样本特征表示的优化，可直接影响模型在少数类样本上的分类效果. 具体方案包括度量学习、两阶段学习等. 在基于度量学习的方法中，Huang 等人[88]的方法较为经典，他们引入了样本簇间距离和类间距离，同时提出了五元组损失函数，用于学习一个可同时保持类内样本簇间距离与类间距离的表示函数. Zhang 等人[89]则使用一个 mini-batch 中所有样本对之间的总距离对表示学习进行优化，缓解了数据类别不均衡在特征提取过程中造成的偏差. 对两阶段学习而言，该类方法将模型的整个训练过程解耦为特征学习和分类器学习两个阶段. 例如：Kang 等人[90]提出的 Decoupling 方法在特征学习阶段对样本进行均匀采样，不考虑类别不均衡问题；而在分类器学习阶段进行类别均衡采样，对分类器进行重新训练. 该方法使得分类器在少数类上性能得到了显著提升.

3.3 模型敏感性分析

除模型鲁棒性与数据均衡性会影响模型预测性能之外，模型内部的敏感性也会导致模型产生缺陷. 不同于模型鲁棒性主要围绕模型输入输出剖析其分类边界，模型敏感性主要关注机器学习模型的内部结构，如特征提取、中间激活状态、内部特征表示等，重点在于分析：在模型推理过程中，哪些部分对其决策有较大影响. 近年来，机器学习模型敏感性分析相关工作主要围绕模型可解释性[36]展开，如激活最大化、特征反演、特征表示可视化等.

2009 年，Erhan 等人[91]提出了激活最大化方法，这种方法的目的是将各个卷积层的输入偏好可视化，通过观测输入偏好，更好地了解 CNN 的卷积层到底学习到了什么，分析哪些神经元对哪些特征较敏感. Nguyen 等人[92]提出的 DGN-AM 是一个经典的工作，如图 12 所示：作者将生成器与激活最大化相结合，以激活最大化为优化目标，生成可视化的原型样本. 与其他方法相比，该方法所生成的原型样本更自然、更容易被人理解，

也更容易发现网络的不同部分在其感受野中较为关注哪些信息.

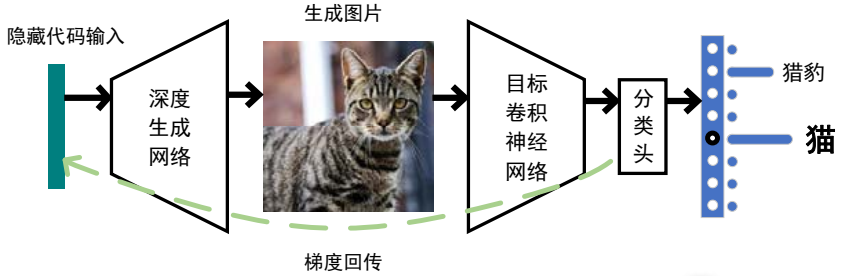


图 12 DGN-AM[92]的整体框架

特征反演最初由 Mahendran 等人[93]提出, 具体任务为, 将神经网络的中间层表示映射回其对应的输入. 该任务的核心内容为对模型中间表示进行充分的解释, 所以在此基础之上, 可以进一步分析模型的敏感性. 目前, 该方法可主要分为模型特征反演和样本特征反演. 在模型特征反演中, Dosovitskiy 等人[94]的工作具有代表性, 他们利用神经网络提取到的图像特征, 使用反卷积网络尝试重建输入的原始图像. 在这一过程中, 发现卷积神经网络对图像中物体的颜色与位置记忆能力较强; 同时发现模型更关注输入的特征值是否为 0, 其具体数值则相对无关紧要. 在样本特征反演中, Du 等人[95]提出了特征反演框架. 通过在反演过程中添加类依赖约束, 并利用中间激活值作为掩码的集成, 既可以发掘输入样本中起决定性作用的重要特征, 也能够定位模型内部较为敏感的结构.

由于现在深度学习模型种类繁多且日趋复杂, 因此研究人员尝试对模型的数据流图或计算图进行可视化, 用于挖掘复杂模型中较为敏感的模块. 例如, Kahng 等人[96]开发了深度神经网络交互式可视化系统 ACTIVIS. 该系统将模型整体架构与局部激活检测相结合, 分析在不同样本输入下, 每一个神经元的激活情况. Strobelt 等人[97]开发了长短期记忆网络(LSTM)可视化工具 LSTMVis, 用于对循环神经网络进行可视化分析. 具体而言, 该工具重点关注 LSTM 的隐含状态, 主要分析时序过程中模型隐含状态的变化规律与特定时间范围内模型隐含状态的变化模式. 此外, Strobelt 等人[98]开发了 Seq2seq-Vis. 该工具与 LSTMVis 类似, 但是更关注如机器翻译等序列到序列的模型. 以文本翻译任务为例, Seq2seq-Vis 将模型推理的每一步可视化展现, 使得研究人员可以对模型内部的敏感点进行直观的分析.

4 机器学习模型防御引擎

在总结阐述了反向智能的技术体系之后, 接下来重点阐述机器学习模型的防御技术. 从另一个角度来看, 模型攻击与防御可以看作前述反向智能体系中的应用. 也就是说, 数据逆向还原、模型反向推演、模型缺陷分析等技术可以用于解决模型攻击[99]和防御问题[100,101]. 当前, 机器学习模型的攻击和防御已经引起了学术界的极大关注, 并有很多优秀的成果不断被提出来. 本节以反向智能的应用为落脚点, 对模型的防御技术进行总结.

4.1 数据逆向还原防御技术

随着训练数据逆向还原技术的不断发展, 与其相应的防御技术也是层出不穷. 目前, 针对成员推断、属性推断、分布推断的防御技术研究也逐渐受到了研究人员的关注.

根据第 1 节的描述可知: 训练数据还原技术的一个出发点就是智能模型的过拟合性质, 即相较于测试数据(非训练数据), 智能模型在面对训练数据时, 往往会返回比较高的预测置信度[102]. 由于上述情况的存在, 针对智能模型的训练数据还原技术往往可以通过分析模型在不同数据上预测置信度的差异进行训练数据的还原; 同时, 这一情况也为设计相应的防御技术提供了思路. 一个最直观的方式就是利用正则化的方法缩小智能模型在训练集和测试集上的表现差异, 如在目标模型训练时加入 dropout 层或者使用 label-smoothing 的方

法[13,14]来减小目标模型在训练集和非训练集上的差异. 然而, 由于这类防御方法参与了模型的训练过程,这将会对模型的预测效果造成一定的影响. 因此, 一些研究人员开始对不需要参与模型训练过程的防御方法展开了研究, 例如针对成员推断, 如图 13 所示, Jinyuan 等人[103]从对抗样本[104−107]的角度提出了一种防御成员推断的方法 MemGuard, 其在不改变目标模型最终预测结果的前提下, 对目标模型的输出结果施加一个微小的噪声扰动, 使其可以在不改变分类结果的前提下对推断模型造成一定的迷惑. 除此之外, 基于差分隐私[108]进行数据保护的方法同样可以应用到训练数据还原的防御当中, DPSGD[109]和 PATE[110]就是差分隐私在训练数据还原防御中的两个应用实例. 但是由于这类方法会使目标模型的效果牺牲过多, 因此这种方法还有很大的发展和提升空间.

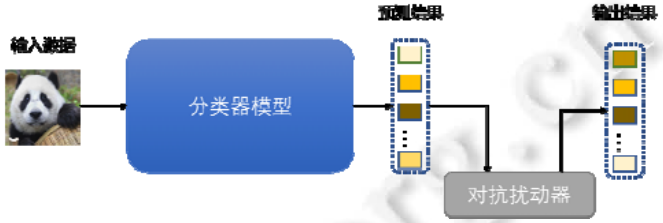


图 13 防御训练数据的逆向还原

4.2 模型逆向反演防御技术

模型逆向反演的任务目标是推断模型的结构、参数、功能等隐私信息, 由于目标模型大多情况下为黑盒模型, 与目标模型相关的可得数据仅有(输入数据,预测输出)对, 因此模型逆向反演方法大多情况下都是基于(输入数据,预测输出)对来实现的. 针对上述情况, 一个很直观的防御思路就是对模型的输出结果进行更改, 就像上一节中所提到的对预测结果增加噪声扰动一样, 利用对抗样本的思想,对预测结果进行更改, 可以在一定程度上对模型逆向反演方法造成干扰. 例如 Orekondy 等人[111]等人在 ICLR 会议上发表的工作中所提出的, 通过主动地对预测结果进行扰动更改, 可以大幅度的降低模型反演的效果, 以达到防御此类攻击的目的.

此外, 由于大部分针对目标模型的反演攻击方法需要与目标模型进行频繁的交互, 并且由于目标模型是黑盒模型, 攻击者往往不清楚目标模型的训练数据域, 因此其所提交的查询数据往往是在训练数据分布外的数据. 基于上述情况, 针对查询次数和查询样本的防御方法逐渐受到了人们的关注. 针对查询次数的防御方法相对来说比较简单, 当某个用户频繁地向目标模型发起请求时, 就能判定该用户可能有反演模型的意图, 进而对该用户的请求进行限制. 这一防御方法在其他领域也得到了广泛的应用, 例如反爬虫机制. 针对查询数据的检测防御[112,113]方法相对来说难度就要高一些了, 如图 14 所示: 假设目标模型是一个植物的分类器, 当查询数据送入的是动物图片的时候, 反向推演检测器就可以判断出这是分布外的数据, 从而限制改查询操作. Juuti 等人[114]提出的 PRADA 方法正是基于此类思想, 通过判断查询数据是否与训练数据同分布, 进而检测反向推演攻击行为. Kariyappa 等人[115]基于分类模型输出的预测结果出发, 根据分类概率的熵值来判断数据是否属于训练数据的分布.

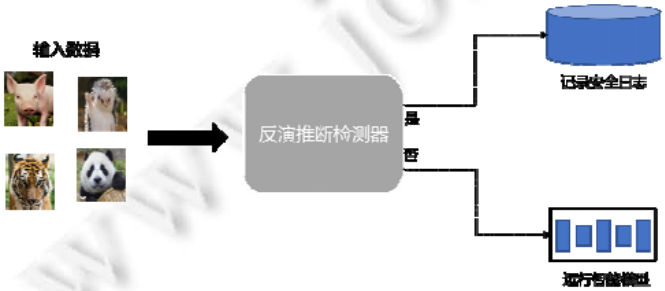


图 14 防御模型的逆向反演

4.3 缺陷分析修正防御技术

机器学习模型缺陷分析是站在模型“白盒化”基础之上, 对模型的缺陷与漏洞进行分析挖掘. 同样的, 作为能够完全了解对应模型训练数据及具体细节的模型拥有者, 更可以主动地对模型漏洞进行检测[116], 并进行对应的修复与完善, 降低模型被攻击成功的概率以实现有效防御[117]. 在这里, 正则化为最具有代表性的方法.

正则化是一个较为宽泛的概念, 即使用一些先验知识对机器学习进行相应的约束, 使得对应的数据或模型规范化, 从而缩小攻击面, 对数据投毒[118−120]、后门攻击[121−124]、对抗样本[37,125−131]等模型针对模型缺陷的攻击方式进行防御[132]. 现有的相关工作主要可分为数据正则化和模型正则化两个方面.

- 数据正则化中, 代表性的技术有特征压缩、特征去噪等. 特征压缩由 Xu 等人[133]首次提出, 他们将原始空间中对应于许多不同样本的特征向量合并为一个压缩特征向量, 用于剔除无关紧要的特征, 减少对手可用的搜索空间. 同时, 将样本在原始空间的特征和在压缩空间的特征在输入模型后得到的推理结果进行对比, 判断该样本是否为恶意样本. 在特征去噪上, Xie 等人[134]研究发现: 对抗样本产生的特征图中, 语义无关的区域同样会被激活. 基于此, 他们提出了特征去噪框架, 与梯度投影下降方式结合, 提高模型对对抗样本的鲁棒性;
- 模型正则化中, 早期技术为防御精馏, 对应方法由 Papernot 等人[135]提出, 如图 15 所示: 他们将模型输出的概率分布向量作为标签, 重新标记对应样本, 并用标记后的样本再次训练相同模型, 最终获得的模型具有更为平滑的分类边界, 同时保持了与原模型一致的准确率. 然而, Carlini 等人[136]证明, 防御精馏不会显著提高神经网络模型的鲁棒性. Gu 等人[137]实验发现, 去噪自动编码器(DAE)可以消除大量的对抗性噪声. 但是将 DAE 与原神经网络模型叠加时, 生成的网络依然可能会受到扰动更小的新对抗样本的攻击. 因此, 他们提出了深度收缩网络, 在训练过程中使用与收缩自编码器(CAE)类似的平滑惩罚项, 在保证模型性能的情况下防御对抗样本攻击.

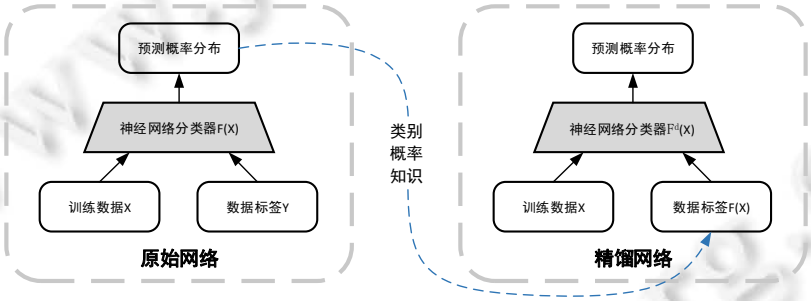


图 15 防御精馏整体概述

5 反向智能研究挑战和方向

反向智能从逆向思维的角度对人工智能模型进行解析, 为进一步理解人工智能提供了有效的途径和方法. 尽管当前反向智能已经取得了一系列瞩目的研究成果, 但目前该研究还处于初级阶段, 仍然存在许多关键问题亟待解决.

- 首先, 训练数据的输入空间决定了模型的搜索空间, 因此, 高维的输入数据使得搜索空间急剧增加, 加剧了数据逆向还原的难度. 训练样本的多样性也给训练数据逆向还原带来了挑战. 尽管基于梯度的方法可以在一定程度上还原真实训练数据, 但是模型的鲁棒性[138]和可扩展性能仍有待提高;
- 其次, 功能的相似性给模型还原带来一定挑战. 例如, ResNet 和 VGG 两个网络具有十分相似的模型功能. 此外, 深度学习网络由任意多个隐含层构成, 删掉某个隐含层整体上不会影响模型的功能. 这些造成了精确还原模型存在一定难度;
- 最后, 模型缺陷隐蔽性强及缺少定量的描述. 模型缺陷常常隐藏在模型结构内部, 且其大小目前难以

精确衡量，缺陷对模型性能的影响程度目前仍难以准确量化.

综上，现有的反向智能研究还有广阔的发展空间，总结未来的研究方向如下.

(1) 结构化逆向数据还原. 在训练数据中，数据常常包含各种各样的结构信息. 例如在图像中，单个像素点与周围像素点常常具有一定的相关性. 如此，像素点集合常常能够构成含有某种语义信息的结构. 在训练数据逆向还原过程中，可以将如此的结构信息作为先验知识加入到模型中，进而降低模型的搜索空间，提高数据的还原精度. 例如: 在模型中增加一些结构化的规则项，约束还原的数据包含某种结构信息，使得还原的数据在时空上具有平滑性. 此外，从模型结构内部挖掘与数据集相关的信息也是一个有意思的研究方向. 例如，卷积神经网络中的 BN 层包含了数据集的均值和方差，因此，加入如此跟数据集相关的历史信息对于还原训练数据具有十分重要的作用.

(2) 模型指纹技术. 正如上所述，不同的神经网络模型常常具有十分相似的功能，同时增加或者减少一些隐含层并不改变模型的功能，这给模型结构精确还原带来了很大的挑战. 因此. 对模型的指纹技术进行研究是一个有趣的研究方向. 例如: 根据模型的中间输出或者最终输出，能够挖掘到与模型结构关联性强的信息，也就是模型指纹. 通过对这些信息进行分析，从而确定模型结构. 此外，利用神经网络架构搜索的方法也可以对目标模型的结构进行还原. 通过利用搜索技术找到与目标模型相同或者相似的结构，也是值得研究的方向.

(3) 缺陷识别及测量技术. 不同的训练条件(例如训练数据类别不均衡)会导致训练的模型常常具有某些缺陷，导致模型的性能受到一定程度影响. 因此，精确定量的刻画模型缺陷是未来的研究方向. 同时，模型缺陷对于模型性能影响很大，不同的模型缺陷往往会对模型产生不同的影响. 建立模型缺陷和模型性能之间的关系，对于人工智能的算法对抗起着关键的作用. 因此，如何精确衡量不同缺陷对于模型性能的影响，也是未来的关键研究方向之一.

(4) 反向智能系统. 比起人工智能的常规任务，反向智能的任务相对来说更加困难，需要耗费更大的算力资源，因此需要有更加高效的，甚至全新的系统架构对于反向智能任务进行实现. 例如在反向智能中，需要常常收集目标模型输出的结果作为状态对模型进行分析. 当需要的数据量较大时，网络带宽将会是一个较大的瓶颈. 因此，解决高带宽等诸如此类的需求，会是在反向智能系统方面的未来研究方向.

6 结束语

随着人工智能安全引起社会的广泛关注，反向智能成为了一个新生而又有前景的研究领域，能够为人工智能安全提供一套从底向上的理论体系保障. 然而到目前为止，反向智能的研究还处于十分初级阶段，许多关键的科学问题依然没有解决. 为了理清人工智能安全需要的理论基础，总结现有研究成果的优势与不足，明确未来的研究方向，本文从逆向思维的角度出发，提出了反向智能的概念，并从数据、模型、应用等方面系统地探讨了反向智能的关键科学问题，回顾了大量相关研究成果并进行了科学的分类和总结. 最后，讨论了反向智能挑战和未来发展方向.

References:

[1] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer, 2006.

[2] Xu H, Ma Y, Liu HC, Debayan D, Liu H, Tang JL, Jain, Anil K. Adversarial attacks and defenses in images, graphs and text: A review. Int'l Journal of Automation and Computing, 2020, 17(2): 151−178.

[3] Zhang CN, Philipp B, Lin CG, Adil K, Wu J, Kweon, In So. A survey on universal adversarial attack. arXiv:2103. 01498, 2021.

[4] Milad N, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. 2018. 634−646.

[5] Reza S, Stronati M, Song CZ, Shmatikov V. Membership inference attacks against machine learning models. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 3−18.

[6] Luca M, Song CZ, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: Proc. of the 2019 IEEE Symp. on Security and Privacy (SP). IEEE, 2019. 691−706.

[7] Song L, Shokri R, Mittal P. Membership inference attacks against adversarially robust deep learning models. In: Proc. of the 2019 IEEE Security and Privacy Workshops (SPW). IEEE, 2019. 50−56.

[8] Samuel Y, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. In: Proc. of the 31st IEEE Computer Security Foundations Symp. (CSF). IEEE, 2018. 268−282.

[9] Christopher A Choquette Choo, Tramer F, Carlini N, Papernot N. Label-only membership inference attacks. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2021. 1964−1974.

[10] Truex S, Liu L, Gursoy ME, *et al*. Demystifying membership inference attacks in machine learning as a service. IEEE Trans. on Services Computing, 2019.

[11] Truex S, Liu L, Gursoy ME, *et al*. Effects of differential privacy and data skewness on membership inference vulnerability. In: Proc. of the 1st IEEE Int'l Conf. on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). IEEE, 2019. 82−91.

[12] Hayes, Jamie, Melis L, Danezis G, Cristofaro ED. LOGAN: Evaluating information leakage of generative models using generative adversarial networks. arXiv:1705. 07663, 2017.

[13] Nasr, Milad, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: Proc. of the 2019 IEEE Symp. on Security and Privacy. IEEE, 2019. 739−753.

[14] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. In: Handbook of Systemic Autoimmune Diseases. 2009.

[15] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4700−4708.

[16] Ateniese G, Mancini LV, Spognardi A, *et al*. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. Int'l Journal of Security and Networks, 2015, 10(3): 137−150.

[17] Ganju K, Wang Q, Yang W, Gunter CA, Borisov N. Property inference attacks on fully connected neural networks using permutation invariant representations. In: Proc. of the 2018 ACM SIGSAC Conf. on Computer and Communications Security. 2018. 619−633.

[18] Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov RR, Smola AJ. Deep sets. In: Advances in Neural Information Processing Systems. 2017. 3394−3404.

[19] Gopinath D, Converse H, Pasareanu C, Taly A. Property inference for deep neural networks. In: Proc. of the 34th IEEE/ACM Int'l Conf. on Automated Software Engineering (ASE). IEEE, 2019. 797−809.

[20] Zhang YH, Jia RX, Pei HZ, Wang WX, Li B, Song D. The secret revealer: Generative model-inversion attacks against deep neural networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE, 2020. 253−261.

[21] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proc. of the 23rd {USENIX} Security Symp. ({USENIX} Security 2014). 2014. 17−32.

[22] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proc. of the 22nd ACM SIGSAC Conf. on Computer and Communications Security. 2015. 1322−1333.

[23] Chen S, Kahla M, Jia RX, Qi GJ. Knowledge-enriched distributional model inversion attacks. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2021. 16178−16187.

[24] Myung IJ. Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology, 2003, 47(1): 90−100.

[25] Bernardo JM, Smith AFM. Bayesian Theory. Vol.405, John Wiley & Sons, 2009.

[26] Silverman BW. Density Estimation for Statistics and Data Analysis. Routledge, 2018.

[27] Cover T. Estimation by the nearest neighbor rule. IEEE Trans. on Information Theory, 1968, 14(1): 50−55.

[28] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in Neural Information Processing Systems. 2014.

[29] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv: 1312. 6114, 2013.

[30] Wang P, Li YJ, Singh KK, Lu JW, Vasconcelos N. IMAGINE: Image synthesis by image-guided model inversion. ArXiv abs/2104. 05895, 2021.

[31] Wang KC, Yan F, Ke L, Khisti AJ, Zemel R, Makhzani A. Variational model inversion attacks. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021.

[32] Wang P, Li YJ, Singh KK, Lu JW, Vasconcelos N. IMAGINE: Image synthesis by image-guided model inversion. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE, 2021. 3681−3690.

[33] Zhang YH, Jia RX, Pei HZ, Wang WX, Li B, Song D. The secret revealer: Generative model-inversion attacks against deep neural networks. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. IEEE, 2020. 253−261.

[34] Yin HX, Molchanov P, Alvarez JM, Li ZZ, Mallya A, Hoiem D, Jha NK, Kautz J. Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 8715−8724.

[35] Thiagarajan JJ, Narayanaswamy V, Rajan D, Liang J, Chaudhari A, Spanias A. Designing counterfactual generators using deep model inversion. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021.

[36] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Learning Representations. 2014.

[37] Dezfooli M, Mohsen S, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2574−2582.

[38] Oh SJ, Schiele B, Fritz M. Towards reverse-engineering black-box neural networks. In: Proc. of the Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019, 121−144.

[39] Yan MJ, Fletcher CW, Torrellas J. Cache telepathy: Leveraging shared resource attacks to learn DNN architectures. In: Proc. of the USENIX Security Symp. 2020. 2003−2020.

[40] Hua WZ, Zhang ZR, Suh GE. Reverse engineering convolutional neural networks through side-channel information leaks. In: Proc. of the ACM/ESDA/IEEE Design Automation Conf. (DAC). IEEE, 2018. 1−6.

[41] Naghibijouybari H, Neupane A, Qian ZY, Abu-Ghazaleh N. Rendered insecure: GPU side channel attacks are practical. In: Proc. of the ACM SIGSAC Conf. on Computer and Communications Security. 2018. 2139−2153.

[42] Hu X, Liang L, Li SC, Deng L, Zuo PF, Ji Y, Xie XF, Ding YF, Liu C, Sherwood T, *et al*. DeepSniffer: A DNN model extraction framework based on learning architectural hints. In: Proc. of the Int'l Conf. on Architectural Support for Pro-gramming Languages and Operating Systems. 2020. 385−399.

[43] Zhu YK, Cheng YQ, Zhou HS, Lu YT. Hermes attack: Steal {DNN} models with lossless inference accuracy. In: Proc. of the 30th {USENIX} Security Symp. 2021.

[44] Lou XX, Guo SW, Li JW, Wu YX, Zhang TW. NASPY: Automated extraction of automated machine learning models. In: Proc. of the Int'l Conf. on Learning Representations. 2022.

[45] Tramèr F, Zhang F, Juels A, *et al*. Stealing machine learning models via prediction {APIs}. In: Proc. of the 25th USENIX Security Symp. (USENIX Security 2016). 2016. 601−618.

[46] Lydia A, Francis S. Adagrad—An optimizer for stochastic gradient descent. Int'l Journal of Information Computer Science, 2019, 6(5): 566−568.

[47] Robbins H, Monro S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 400−407.

[48] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412. 6980, 2014.

[49] Tieleman T, Hinton G. Lecture 6. 5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012, 4(2): 26−31.

[50] Maheswaranathan N, Sussillo D, Metz L, Sun RX, Sohl-Dickstein J. Reverse engineering learned optimizers reveals known and novel mechanisms. In: Proc. of the Conf. on Neural Information Processing Systems. 2021. 19910−19922.

[51] Sussillo D, Barak O. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Computation, 2013, 25(3): 626−649.

[52] Andrychowicz M, Denil M, Gomez S, Hoffman MW, Pfau D, Schaul T, Shillingford B, De Freitas N. Learning to learn by gradient descent by gradient descent. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 3981−3989.

[53] Polyak BT. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1−17.

[54] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011, 12(7): 257−269.

[55] Wang BH, Gong NZQ. Stealing hyperparameters in machine learning. In: Proc. of the IEEE Symp. on Security and Privacy (SP). IEEE, 2018. 36−52.

[56] Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. John Wiley & Sons, 2015.

[57] Jagielski M, Carlini N, Berthelot D, Kurakin A, Papernot N. High accuracy and high fidelity extraction of neural networks. In: Proc. of the {USENIX} Security Symp. 2020. 1345−1362.

[58] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proc. of the 2017 ACM on Asia Conf. on Computer and Communications Security. 2017. 506−519.

[59] Pal S, Gupta Y, Shukla A, *et al*. Activethief: Model extraction using active learning and unannotated public data. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(1): 865−872.

[60] Orekondy T, Schiele B, Fritz M. Knockoff nets: Stealing functionality of black-box models. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 4954−4963.

[61] Shi Y, Sagduyu Y, Grushin A. How to steal a machine learning classifier with deep learning. In: Proc. of the 2017 IEEE Int'l Symp. on Technologies for Homeland Security (HST). IEEE, 2017. 1−5.

[62] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks. In: Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing. 2008. 1070−1079.

[63] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503. 02531, 2015.

[64] Kariyappa, Sanjay, Prakash A, Qureshi MK. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2021. 13814−13823.

[65] Fang GF, Song J, Shen CC, Wang XC, Chen D, Song ML. Data-free adversarial distillation. arXiv:1912. 11006, 2019.

[66] Micaelli P, Storkey AJ. Zero-shot knowledge transfer via adversarial belief matching. In: Advances in Neural Information Processing Systems. 2019. 9551−9561.

[67] Ghadimi S, Lan GH. Stochastic first-and zeroth-order methods for nonconvex stochastic program-ming. SIAM Journal on Optimization, 2013, 23(4): 2341−2368.

[68] Nesterov Y, Spokoiny V. Random gradient-free minimization of convex functions. Foundations of Com-putational Mathematics, 2017, 17(2): 527−566.

[69] Truong JB, *et al*. Data-free model extraction. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2021. 4771−4780.

[70] Gong X, Chen Y, Yang W, Mei G, Wang Q. INVERSENET: Augmenting model extraction attacks with training data inversion. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence {IJCAI-21}. 2021.

[71] Ehlers R. Formal verification of piece-wise linear feed-forward neural networks. In: Proc. of the Int'l Symp. on Automated Technology for Verification and Analysis. Cham: Springer, 2017.

[72] Katz G, Barrett C, Dill DL, *et al*. Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proc. of the Int'l Conf. on Computer Aided Verification. Cham: Springer, 2017. 97−117.

[73] Tjeng V, Xiao K, Tedrake R. Evaluating robustness of neural networks with mixed integer programming. arXiv:1711.07356, 2017.

[74] Singh G, Gehr T, Püschel M, *et al*. An abstract domain for certifying neural networks. Proc. of the ACM on Programming Languages, 2019, 3(POPL): 1−30.

[75] Salman H, Yang G, Zhang H, *et al*. A convex relaxation barrier to tight robustness verification of neural networks. arXiv:1902. 08722, 2019.

[76] Xu K, Shi Z, Zhang H, *et al*. Automatic perturbation analysis for scalable certified robustness and beyond. In: Advances in Neural Information Processing Systems. 2020. 33.

[77] Hein M, Andriushchenko M. Formal guarantees on the robustness of a classifier against adversarial manipulation. arXiv:1705. 08475, 2017.

[78] Zhang H, Zhang PC, Hsieh CJ. Recurjac: An efficient recursive algorithm for bounding Jacobian matrix of neural networks and its applications. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1).

[79] Lecuyer M, *et al*. Certified robustness to adversarial examples with differential privacy. In: Proc. of the 2019 IEEE Symp. on Security and Privacy (SP). IEEE, 2019.

[80] Cohen J, Rosenfeld E, Kolter Z. Certified adversarial robustness via randomized smoothing. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2019.

[81] Yang G, Duan T, Hu JE, *et al*. Randomized smoothing of all shapes and sizes. In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2020. 10693−10705.

[82] Pouyanfar S, *et al*. Dynamic sampling in convolutional neural networks for imbalanced data classification. In: Proc. of the 2018 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR). IEEE, 2018.

[83] He H, Garcia EA. Learning from imbalanced data. IEEE Trans. on Knowledge and Data Engineering, 2009, 21(9): 1263−1284.

[84] Wang YR, *et al*. Dynamic curriculum learning for imbalanced data classification. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision. 2019.

[85] Wu T, *et al*. Adversarial robustness under long-tailed distribution. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2021.

[86] Yin X, Yu X, Sohn K, *et al*. Feature transfer learning for face recognition with under-represented data. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2019. 5704−5713.

[87] Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321−357.

[88] Huang C, Li Y, Loy CC, *et al*. Learning deep representation for imbalanced classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5375−5384.

[89] Zhang X, Fang Z, Wen Y, *et al*. Range loss for deep face recognition with long-tailed training data. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 5409−5418.

[90] Kang B, Xie S, Rohrbach M, *et al*. Decoupling representation and classifier for long-tailed recognition. arXiv:1910.09217, 2019.

[91] Erhan D, Bengio Y, Courville A, *et al*. Visualizing higher-layer features of a deep network. University of Montreal, 2009, 1341(3): 1.

[92] Nguyen A, Dosovitskiy A, Yosinski J, *et al*. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems, Vol.29. 2016. 3387−3395.

[93] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 5188−5196.

[94] Dosovitskiy A, Brox T. Inverting visual representations with convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4829−4837.

[95] Du M, Liu N, Song Q, *et al*. Towards explanation of DNN-based prediction with guided feature inversion. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2018. 1358−1367.

[96] Kahng M, Andrews PY, Kalro A, *et al*. Activis: Visual exploration of industry-scale deep neural network models. IEEE Trans. on Visualization and Computer Graphics, 2017, 24(1): 88−97.

[97] Strobelt H, Gehrmann S, Pfister H, *et al*. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Trans. on Visualization and Computer Graphics, 2017, 24(1): 667−676.

[98] Strobelt H, Gehrmann S, Behrisch M, *et al*. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. IEEE Trans. on Visualization and Computer Graphics, 2018, 25(1): 353−363.

[99] Biggio B, *et al*. Evasion attacks against machine learning at test time. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2013.

[100] Salem A, Zhang Y, Humbert M, *et al*. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv:1806.01246, 2018.

[101] Wang Q, Guo W, Zhang K, *et al*. Adversary resistant deep neural networks with an application to Malware detection. In: Proc. of the 23rd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2017. 1145−1153.

[102] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015.

[103] Jia JY, Salem A, Backes M, Zhang Y, Gong NZQ. Memguard: Defending against black-box membership inference attacks via adversarial examples. In: Proc. of the 2019 ACM SIGSAC Conf. on Computer and Communications Security (CCS). 2019.

[104] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:1412.6572, 2014.

[105] Metzen JH, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[106] Papernot N, *et al*. The limitations of deep learning in adversarial settings. In: Proc. of the IEEE European Symp. on Security and Privacy (EuroS&P). IEEE, 2016. 372−387.

[107] Tramèr F, Kurakin A, Papernot N, Boneh D, McDaniel P. Ensemble adversarial training: Attacks and defenses. arXiv:1705.07204, 2017.

[108] Dwork C. Differential privacy: A survey of results. In: Proc. of the Int'l Conf. on Theory and Applications of Models of Computation. Berlin, Heidelberg: Springer, 2008. 1−19.

[109] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. 2016. 308−318.

[110] Papernot N, Abadi M, Erlingsson U, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv:1610.05755, 2016.

[111] Orekondy T, Schiele B, Fritz M. Prediction poisoning: Towards defenses against DNN model stealing attacks. arXiv:1906.10908, 2019.

[112] Metzen JH, *et al*. On detecting adversarial perturbations. arXiv:1702.04267, 2017.

[113] Wang JY, *et al*. Detecting adversarial samples for deep neural networks through mutation testing. arXiv:1805.05010, 2018.

[114] Juuti M, Szyller S, Marchal S, Asokan N. PRADA: Protecting against DNN model stealing attacks. In: Proc. of the 2019 IEEE European Symp. on Security and Privacy (EuroS&P). IEEE, 2019. 512−527.

[115] Kariyappa S, Qureshi MK. Defending against model stealing attacks with adaptive misinformation. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 770−778.

[116] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proc. of the ACM Workshop on Artificial Intelligence and Security. 2017. 3−14.

[117] Carlini N, Wagner D. Magnet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. arXiv: 1711.08478, 2017.

[118] Nelson B, Barreno M, Chi FJ, Joseph AD, Rubinstein BIP, Saini U, Sutton C, Tygar JD, Xia K. Exploiting machine learning to subvert your spam filter. LEET, 2008, 8(1−9): 16−17.

[119] Xiao H, Biggio B, Brown G, Fumera G, Eckert C, Roli F. Is feature selection secure against training data poisoning? In: Proc. of the Int'l Conf. on Machine Learning. PMLR, 2015. 1689−1698.

[120] Biggio B, Nelson B, Laskov P. Support vector machines under adversarial label noise. In: Proc. of the Asian Conf. on Machine Learning. PMLR, 2011. 97−112.

[121] Liu YF, Ma XJ, Bailey J, Lu F. Reflection backdoor: A natural backdoor attack on deep neural networks. In: Proc. of the European Conf. on Computer Vision. Cham: Springer, 2020. 182−199.

[122] Zhao SH, Ma XJ, Zheng X, Bailey J, Chen JJ, Jiang YG. Clean-label backdoor attacks on video recognition models. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2020. 14443−14452.

[123] Saha A, Subramanya A, Pirsiavash H. Hidden trigger backdoor attacks. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2020. 11957−11965.

[124] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models. In: Proc. of the 30th USENIX Security Symp. (USENIX Security 2021). 2021. 1505−1521.

[125] Lin YC, Hong ZW, Liao YH, Shih ML, Liu MY, Sun M. Tactics of adversarial attack on deep reinforcement learning agents. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2017. 3756−3762.

[126] Kos J, Song D. Delving into adversarial attacks on deep policies. In: Proc. of the Int'l Conf. on Learning Representations (Workshop). 2017.

[127] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. In: Proc. of the Conf. Empirical Methods Natural Lang. Process (EMNLP). 2017. 1−11.

[128] Sharif M, Bhagavatula S, Bauer L, *et al*. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security. 2016. 1528−1540.

[129] Xie C, Wang J, Zhang Z, *et al*. Adversarial examples for semantic segmentation and object detection. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1369−1378.

[130] Fischer V, Kumar MC, Metzen JH, Brox T. Adversarial examples for semantic image segmentation. In: Proc. of the Int'l Conf. on Learning Representations (Workshop). 2017.

[131] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text. In: Proc. of the IEEE Security and Privacy Workshops (SPW). IEEE, 2018. 1–7.

[132] Madry A, Makelov A, Schmidt L, *et al*. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083, 2017.

[133] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv:1704.01155, 2017.

[134] Xie C, Wu Y, Maaten L, *et al*. Feature denoising for improving adversarial robustness. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. 2019. 501–509.

[135] Papernot N, McDaniel P, Wu X, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 2016 IEEE Symp. on Security and Privacy (SP). IEEE, 2016. 582–597.

[136] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 39–57.

[137] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv:1412.5068, 2014.

[138] Carlini, Nicholas, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 39–57.

李长升(1985－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习.

张成喆(2000－), 男, 学士, 主要研究领域为人工智能安全.

汪诗烨(1995－), 女, 博士生, CCF 学生会员, 主要研究领域为机器学习.

袁野(1981－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数据库.

李延铭(1997－), 男, 硕士生, 主要研究领域为机器学习.

王国仁(1966－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为不确定数据管理, 数据密集型计算, 可视媒体数据分析管理, 非结构化数据管理, 分布式查询处理与优化, 生物信息学.