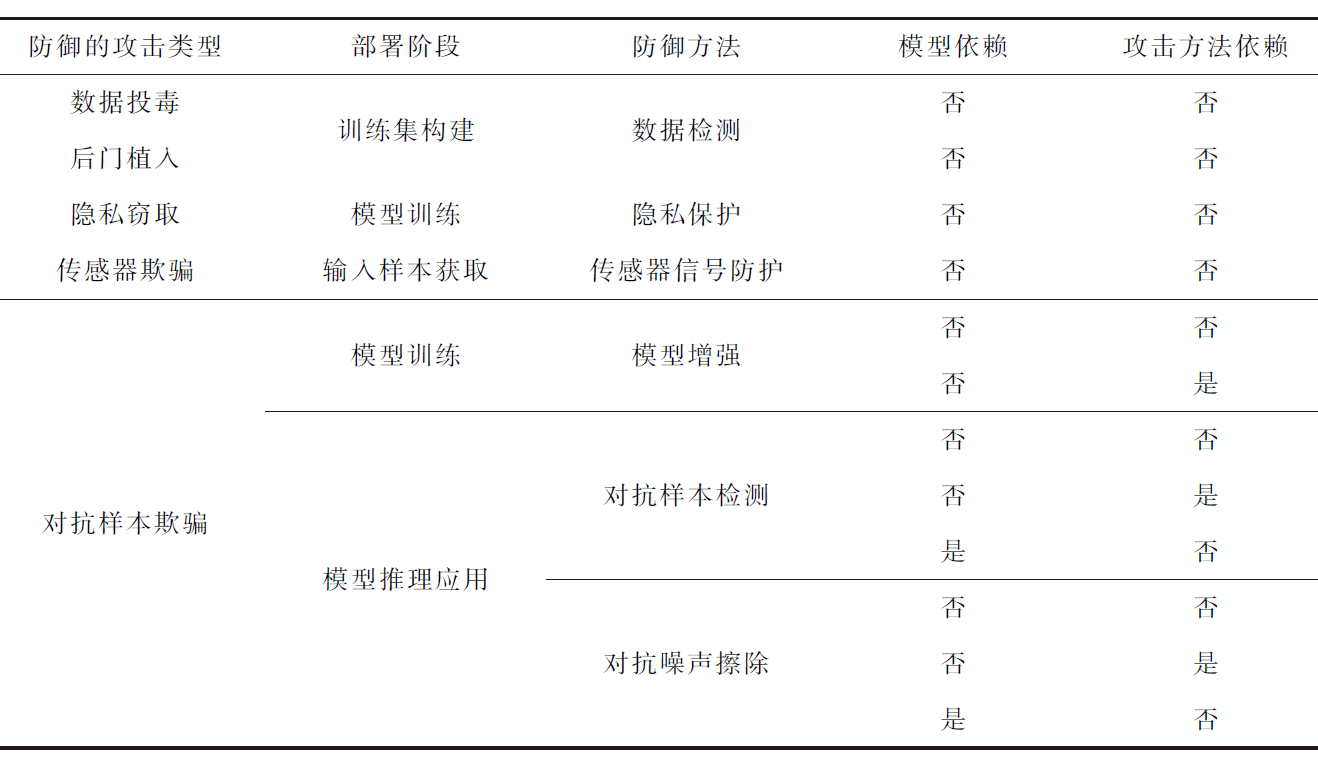
安全加固算法

1. 对抗性攻击防御方案[1]:

一些典型的对抗攻击防御方法如下表所示:



1. 线下阶段对抗攻击的防御策略[1]:
2. 数据溯源技术: 为从线下训练数据集中过滤掉恶意样本，保证训练集不被污染。依靠反复训练模型来确定训练集中的异常数据。然而当训练集数据量较大时，该防御方案的检测效率较低。
3. 异常检测技术: 只需分析数据本身的特征即可确定异常数据，具体有:
4. 利用支持向量机和决策树检测异常数据: 这种技术需要保证在训练检测器时使用的是真实样本。
5. 构建各种后门攻击的近似上界: 使检测器能在被污染的情况下检测异常数据。然而，当数据集中被污染的数据占比过大时，该防御方案会失效。
6. 使用聚类算法来判别异常数据: 巧妙地回避了检测器污染问题。但该工作只讨论了数据污染，未考虑标签污染。
7. 优化模型构造: 主要有以下几种方法:
8. 对抗训练方法: 通过对抗训练的方法来优化模型参数，增强模型鲁棒性。能够增强模型对单步攻击的鲁棒性，但很难抵抗迭代攻击，而且同样无法抵御利用单步攻击方法从另一个脆弱性模型中生成的扰动。
9. 去噪自编码器法: 网络输入层之前添加一个去噪自编码器以降低对抗噪声，但可能导致网络更容易遭受攻击。
10. 生成对抗网络法: 使用一种常见的生成对抗网络架构来训练一个可以防御快速梯度符号法攻击的模型，其缺点是模型训练过程比较复杂。
11. 高斯混合深度模型法: 利用高斯过程对不确定因素的处理能力来增强模型的鲁棒性，并证明其对快速梯度符号法攻击的抵御能力。
12. 神经网络知识蒸馏技术: 通过简化已训练完成的模型结构来消除潜在威胁，抵御对抗性攻击，但这种方法的防御能力有限。
13. 模型诊断策略: 利用模型可解释性来分析和调试模型的错误决策行为，诊断模型中存在的缺陷，可为模型缺陷修复提供支撑，以获得更加鲁棒的模型。然而，现有模型诊断策略缺乏与实际模型训练过程的有机结合。
14. 保护模型隐私: 对一个鲁棒的模型来说，如果攻击者知道其相关信息就可以有针对性地进行攻击。因此，保护模型的隐私信息同样被认为是提升模型防御能力的重要环节。主要方法有:
15. 同态加密(homomorphic encryption)技术: 在加密数据上进行训练，但在加密数据上进行乘法操作会引入大量噪声，导致信息无法解密。因此，该方案需对现有算法进行最高次有界的多项式逼近，而这又会导致性能下降。
16. 差分隐私(differential privacy): 通过在训练模型时往参数中添加噪声的方式，达到隐私保护目的。但这些方法未考虑添加噪声对模型可用性的影响。
17. 平衡优化: 将训练过程当作是可用性和隐私保护之间的平衡优化问题来进行模型训练，但如果同时需要考虑模型鲁棒性，模型训练过程可能会很复杂。
18. 同态加密法: 针对已训练好的模型，利用同态加密的方法对模型加密，但同态加密引入的噪声会显著降低模型表现性能。
19. 线上阶段对抗攻击的防御策略[1]:

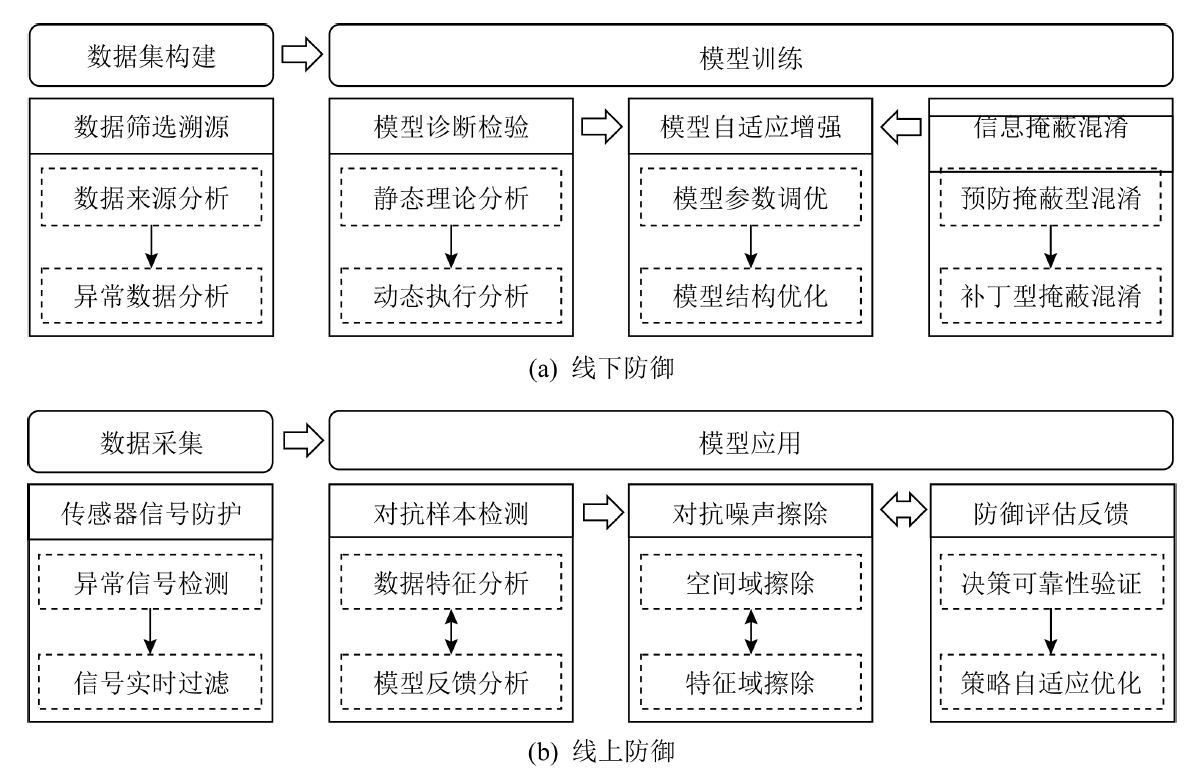
在硬件层面，研究者们提出通过利用传感器增强方案在数据输入阶段过滤掉对抗性噪声，但目前大部分工作都停留在理论分析阶段。在软件层面，可以利用对抗性噪声的不稳定性在预测任务时检测或擦除对抗性噪声。具体方法有:

1. 数据压缩或添加噪声法: 通过对数据压缩或是添加噪声的方法试图破坏对抗样本的对抗性，然后比较处理前后数据预测结果是否发生改变来检测对抗样本。
2. JPEG压缩法: 通过JPEG 压缩，去除图像方块中的高频信号成分，从而消除对抗性噪声。
3. 随机变化层法: 在分类网络之前增加2个随机变化层，破坏对抗性噪声的特定结构来实现防御。
4. 二分类检测网络: 为消除数据处理操作对真实样本的影响，可以训练一个二分类检测网络，以模型某一隐藏层的输出作为输入，再输出此次输入样本为对抗样本的概率，但该方案只能检测特定类型的攻击。
5. 核密度估计与不确定性估计法: 利用模型最后一个隐藏层子空间的核密度估计和贝叶斯神经网络不确定性估计，并结合逻辑回归模型检测输入是否为对抗样本，目前该研究只针对卷积神经网络。
6. 重构对抗样本法: 只针对检测出对抗性的样本进行对抗噪声擦除操作，利用一个检测网络来检测对抗样本，并利用一个重构网络重构对抗样本来消除其对抗性，结果显示该方案对黑盒攻击和灰盒攻击有较好的抵御效果。
7. 绕过对抗样本法: 对抗样本接近于分类边界，从输入空间中选择和此次预测样本接近的多个样本点进行预测，再由这些样本点的预测结果进行投票作为此次预测结果。显然，该方案依赖于分类边界和样本点的选取。
8. 对抗攻击防御框架[1]:

针对已有攻击手段的防御框架包括: 异常检测框架、模型增强框架、对抗样本检测框架、对抗噪声擦除框架和隐私保护框架。这些框架涉及从数据收集到模型训练再到模型线上部署的每个阶段，包含模型鲁棒性和模型机密性等内容。然而，这些工作仅是框架化了单个防御方案的设计，不同防御策略相互之间缺乏协同，难以形成统一防御框架。少数建立不同防御方案之间联系的研究成果如下:

1. 模型诊断与模型增强的联合框架: 首先分析网络每一层潜在脆弱性，然后提出相应的策略来优化训练过程以应对这些脆弱性。
2. 对抗样本检测和对抗噪声消除结合: 通过一个输入重整网络对输入进行处理，并比较处理前后样本之间的差异性来确认输入样本是否包含对抗性扰动，从而决定输入到目标模型中的样本。

一种理论上的全方位防御框架如下图所示。

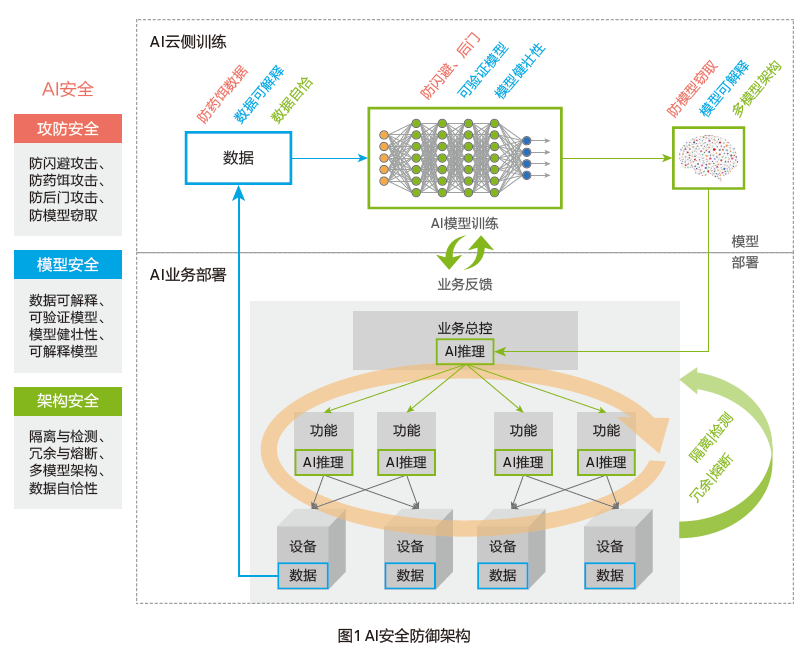


“线下防御”是在模型发布前从数据集构建和模型训练2个角度出发，在建立可靠数据集的基础上进一步提升模型的鲁棒性和机密性，最大化模型的防御能力。“线上防御”是在模型发布之后从输入样本获取和模型应用2个角度出发，在确保传感器数据准确的基础上，实时检测和擦除输入样本中的对抗噪声，并通过建立有效的评估反馈机制实现防御的动态优化。

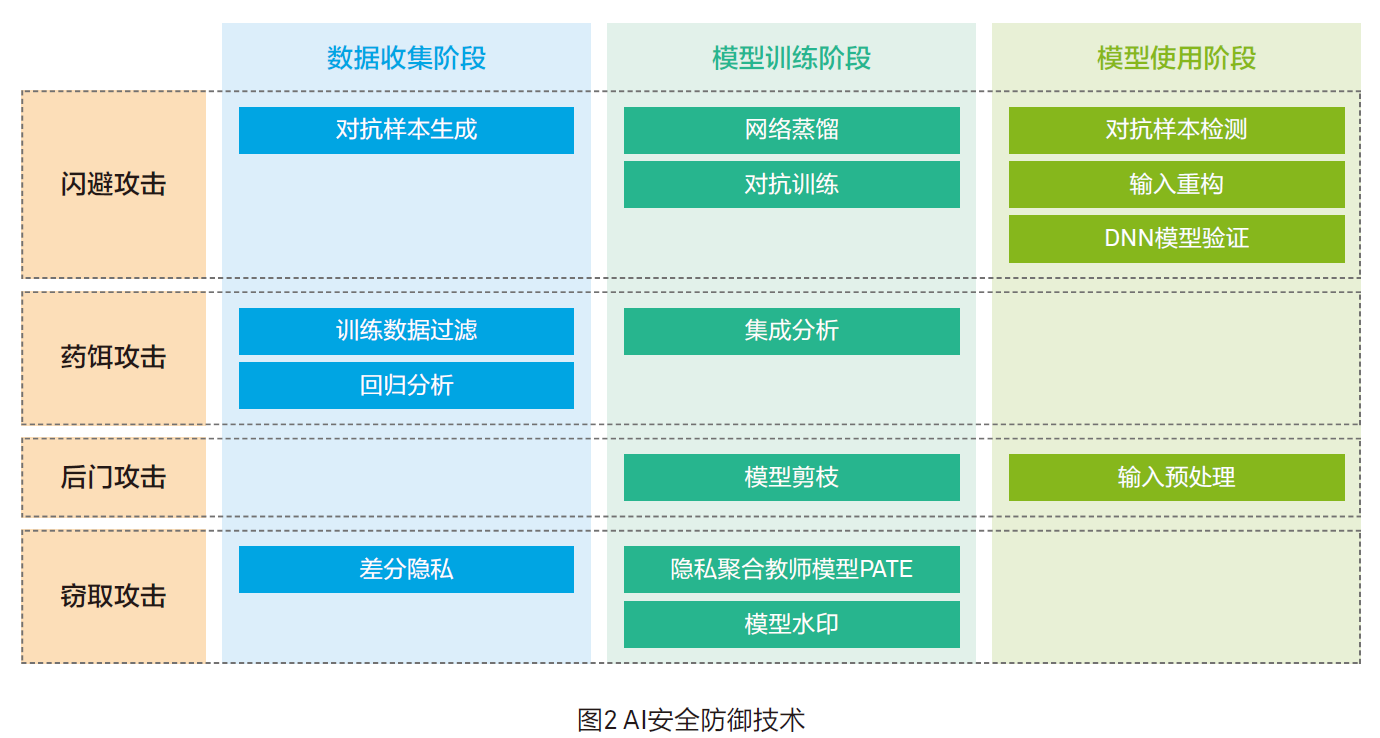
1. 安全防御手段[2]:

下图描述了AI系统部署到业务场景中所需要的三个层次的防御手段:

1. 攻防安全: 对已知攻击所设计的有针对性的防御机制；
2. 模型安全: 通过模型验证等手段提升模型健壮性；
3. 架构安全: 在AI部署的业务中设计不同的安全机制保证架构安全。



下图列出了AI系统在数据收集、模型训练及模型使用阶段的各种防御技术。



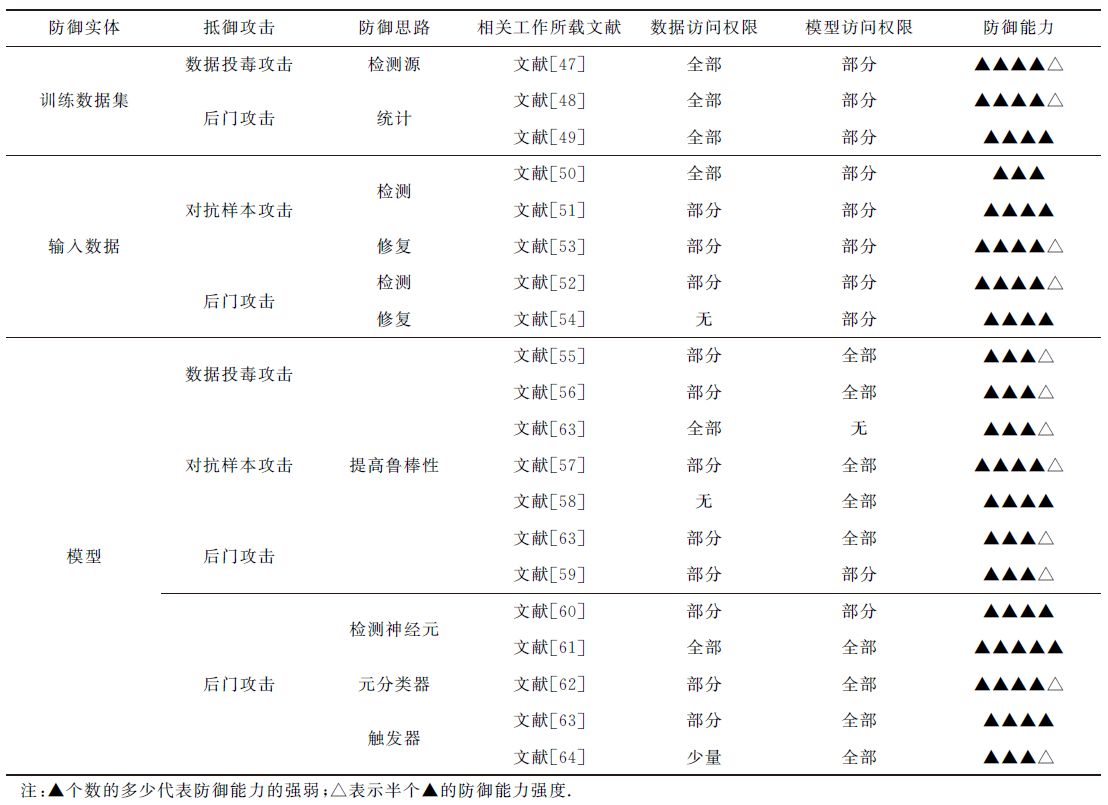
1. 闪避攻击防御技术[2]:
2. 网络蒸馏(Network Distillation): 网络蒸馏技术的基本原理是在模型训练阶段，对多个DNN进行串联，其中前一个DNN生成的分类结果被用于训练后一个DNN。转移知识可以一定程度上降低模型对微小扰动的敏感度，提高AI模型的鲁棒性。
3. 对抗网络(Adversarial Training): 该技术的基本原理是在模型训练阶段，使用已知的各种攻击方法生成对抗样本，再将对抗样本加入模型的训练集中，对模型进行单次或多次重训练，生成可以抵抗攻击扰动的新模型。同时，由于综合多个类型的对抗样本使得训练集数据的增多，该技术不但可以增强新生成模型的鲁棒性，还可以增强模型的准确率和规范性。
4. 对抗样本检测(Adversarial Sample Detection): 该技术的原理为在模型的使用阶段，通过增加外部检测模型或原模型的检测组件来检测待判断样本是否为对抗样本。在输入样本到达原模型前，检测模型会判断其是否为对抗样本。检测模型也可以在原模型每一层提取相关信息，综合各种信息来进行检测。各类检测模型可能依据不同标准来判断输入是否为对抗样本。
5. 输入重构(Input Reconstruction): 该技术的原理是在模型的使用阶段，通过将输入样本进行变形转化来对抗闪避攻击，变形转化后的输入不会影响模型的正常分类功能。重构方法包括对输入样本加噪、去噪、和使用自动编码器改变输入样本等方法。
6. DNN模型验证(DNN Verification): 类似软件验证分析技术，DNN模型验证技术使用求解器(solver)来验证DNN模型的各种属性，如验证在特定扰动范围内没有对抗样本。但是通常验证DNN模型是NP完全问题，求解器的效率较低。通过取舍和优化，如对模型节点验证的优先度选择、分享验证信息、按区域验证等，可以进一步提高DNN模型验证运行效率。
7. 药饵攻击防御技术[2]:
8. 训练数据过滤(Training Data Filtering): 该技术侧重对训练数据集的控制，利用检测和净化的方法防止药饵攻击影响模型。具体方向包括: 根据数据的标签特性找到可能的药饵攻击数据点，在重训练时过滤这些攻击点；采用模型对比过滤方法，减少可以被药饵攻击利用的采样数据，并过滤数据对抗药饵攻击。
9. 回归分析(Regression Analysis): 该技术基于统计学方法，检测数据集中的噪声和异常值。具体方法包括对模型定义不同的损失函数（loss function）来检查异常值，以及使用数据的分布特性来进行检测等。
10. 集成分析(Ensemble Analysis): 该技术强调采用多个子模型的综合结果提升机器学习系统抗药饵攻击的能力。多个独立模型共同构成AI系统，由于多个模型采用不同的训练数据集，整个系统被药饵攻击影响的可能性进一步降低。
11. 后台攻击防御技术[2]:
12. 输入预处理(Input Preprocessing): 该方法的目的是过滤能触发后门的输入，降低输入触发后门、改变模型判断的风险。
13. 模型剪枝(Model Pruning): 该技术原理为适当剪除原模型的神经元，在保证正常功能一致的情况下，减少后门神经元起作用的可能性。利用细粒度的剪枝方法，可以去除组成后门的神经元，防御后门攻击。
14. 模型/数据防窃取技术[2]:
15. 隐私聚合教师模型(PATE): 该技术的基本原理是在模型训练阶段，将训练数据分成多个集合，每个集合用于训练一个独立DNN模型，再使用这些独立DNN模型进行投票的方法共同训练出一个学生模型。这种技术保证了学生模型的判断不会泄露某一个特定训练数据的信息，从而确保了训练数据的隐私性。
16. 差分隐私(Differential Privacy): 该技术是在模型训练阶段，用符合差分隐私的方法对数据或模型训练步骤进行加噪。
17. 模型水印(Model Watermarking): 该技术是在模型训练阶段，在原模型中嵌入特殊的识别神经元。如果发现有相似模型，可以用特殊的输入样本识别出相似模型是否通过窃取原模型所得。
18. 分布式深度学习隐私攻击防御技术[3]:
19. 基于差分隐私的防御方法: 差分隐私是一种专门针对数据集统计特性设计的新型隐私保护技术，主要通过限制查询统计数据库时识别具体数据实例的机会而实现用户数据的隐私保护。在深度学习领域，差分隐私通常用于在模型梯度中添加噪声，从而实现隐私保护。具体有:
20. 基于差分隐私的GANobfuscator方案: 针对GAN攻击问题。通过在梯度中添加噪声实现差分隐私，并开发完整的梯度裁剪策略以提高数据训练的可扩展性和稳定性，使用户能够在不泄露自身隐私的前提下使用GAN 生成大量合成数据。
21. 基于本地差分隐私的联邦学习新方法: 一种雾计算环境下的基于本地差分隐私的联邦学习新方法。针对联邦学习面临数据分布不均匀和计算能力差距大而导致训练效率低的问题，有效实现了对参数更新的保护，并提高了模型训练效率。
22. 基于差分隐私的分布式在线学习框架: 基于多数据源广泛分布的特性，使用最小批次稀疏学习方式进行训练保证了相同隐私预算下的更低噪声。
23. 基于同态加密的防御方法: 同态加密技术允许数据在密文条件下进行安全计算，实现密文状态下对明文的操作，从而保证数据的隐私性。在分布式深度学习领域，同态加密技术经常被用于参数加密，以密文状态下实现全局参数聚合更新，使服务器无法获取明文。具体有:
24. 同态加密下的隐私保护联邦学习方案: 针对分布式深度学习中用户的梯度会泄露给半诚实的服务器而导致泄露训练集信息的问题。使服务器无法接触用户的梯度明文，保证了模型安全。
25. 非交互式联合学习框架: 有效避免了数据拥有者与训练者之间数据的多轮交互，降低了通信开销，并通过掩码技术与Paillier同态加密技术保护了训练集以及训练模型的隐私。
26. 使用随机子集的用户更新矢量的加权平均值的安全聚合方案: 在聚合过程中用户间的盲化因子相互抵消，可以在用户退出的情况下恢复其盲化因子，使方案可以适应用户离线的情况，提升模型鲁棒性。
27. 使用Paillier加密实现梯度的安全聚合，并利用同态散列改进双线性聚合签名来验证聚合结果的正确性。
28. 基于安全多方计算的防御方法: 安全多方计算技术是指在无可信第三方的条件下多方参与者安全地计算一个约定函数问题，主要研究参与方在协作计算时如何对各方隐私数据进行保护，重点关注各参与方之间的隐私安全性问题，即在安全多方计算过程中必须保证各方私密输入独立，计算时不泄露任何本地数据。具体有:
29. 使用安全多方计算的Improved Du-Atallah协议将参与者与模型参数分离开，通过交互模式实现参与者的本地模型训练，保证参与者的数据安全和服务器的模型安全。
30. 基于可信执行环境的联邦学习机制: 实现了用户本地训练结果的完整性验证,但依然没有解决用户数据可用性验证问题。
31. 采用博弈论思想将多用户和云中心模式的联合学习抽象为单个用户和云服务器之间的交互模式,实现了用户和云之间的安全交互。

分布式深度学习隐私防御总结如下表所示:



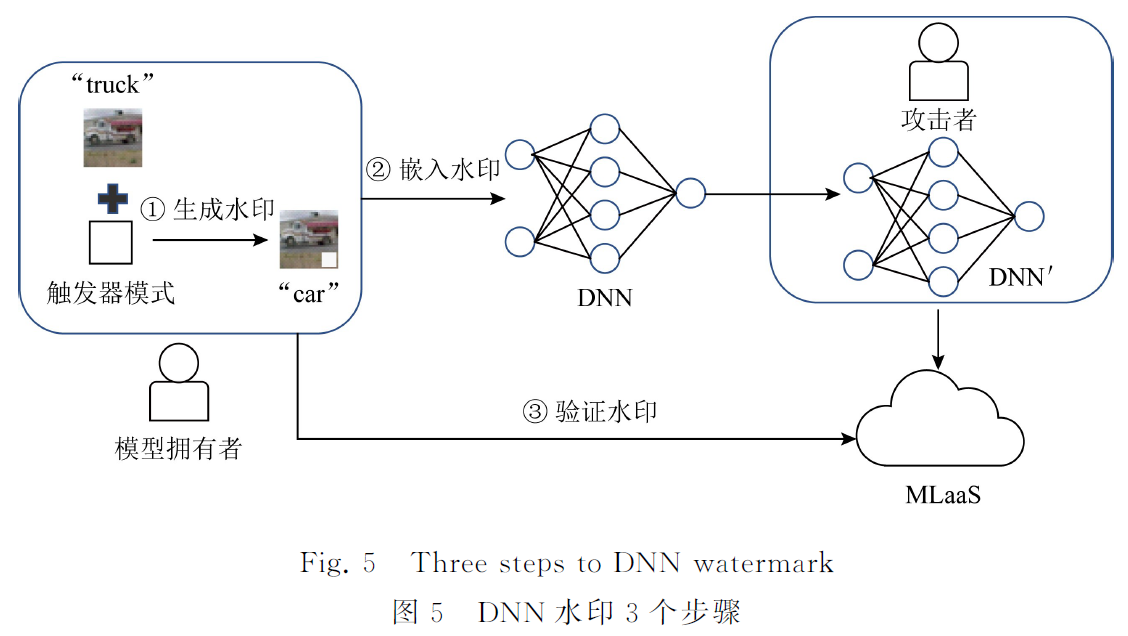
1. 分布式安全攻击防御技术[3]:
2. 基于数据集的防御技术: 数据集是投毒攻击与后门攻击的主要攻击对象。因此，通过一些手段来识别数据集中的恶意样本是抵御两种攻击的一条主要思路。具体有:
3. 检测源: 防御者可以计算数据对模型的贡献从而剔除具有消极贡献的数据，将训练数据按照来源信息划分为若干段，测试每段对模型的贡献。该方案可用于部分可信和不可信数据集的检测。
4. 鲁棒统计技术: 考虑到良性样本与恶意样本的深层特征差别很大，防御者对其使用鲁棒统计技术来检测恶意样本。利用深层特征对待检测数据集进行划分。对于每个输出类，搜集该类所有输入数据的模型中间表示的协方差矩阵，通过奇异值分解计算每条输入数据的异常值分数。
5. 异常行为过滤: 一些防御者利用恶意样本的异常行为过滤数据集。发现输入层中触发图像的触发位置有较大的梯度绝对值，利用聚类算法识别触发样本并修正其标签。
6. 基于输入样本的防御方法: 对抗样本攻击与后门攻击均需要构造恶意输入样本以实现攻击者目标。因此，针对输入数据进行防御是抵御两种攻击的另一条主要思路。目前，此类防御手段主要包括检测恶意样本与修补输入样本。
7. 检测恶意样本: 防御者通过输入数据的异常特性或异常行为来检测恶意样本。具体有:
8. 背景类填充: 用背景类填充关键类之间的区域，以检测恶意样本。经验证，该方案可缓解FGSM攻击，但对C&W攻击无效。
9. 特征压缩技术: 可抵御多种对抗样本攻击。通过空间平滑与减少每个像素颜色位的深度来压缩特征，然后观察输入数据与压缩样本预测结果的一致性。
10. 触发器与攻击目标类的强关联性: 触发输入添加强烈扰动后的预测结果具有弱随机性，该方案适用于复杂触发器的检测，尤其对触发器大小不敏感。
11. 修补输入样本: 该策略通过修补输入样本的对抗性或触发器来缓解攻击。具体有:
12. 降噪器MagNet: 训练降噪器(编码器与解码器)MagNet来修补输入，突出图像中的主要成分，进而校正分类结果。实验证明降噪器能够有效缓解FGSM与C&W攻击。
13. 2阶图像修补方法: 首先删除最影响预测的区域，然后利用GAN对该区域进行修补。该方法涉及图像的移除和恢复，所以它对触发器规模很敏感。
14. 基于模型的防御方法: 安全攻击本质上是对模型可用性的破坏，因此针对模型的防御策略亦是安全防御的一种主要思路。根据防御策略，可以将其分为提升模型鲁棒性与模型诊断两类方案。
15. 提升模型鲁棒性: 该策略主要根据安全攻击的思路调整模型参数或改进训练过程，从而使模型难以实现攻击者的目标。
16. 鲁棒逻辑回归模型: 删除幅度较大的样本并最大化其余样本与估计的逻辑回归模型之间的相关性。
17. 利用多分类器系统构造鲁棒分类器: 利用引导聚集算法与随机子空间方法使特征权值分布更均匀，以此降低恶意数据的影响。这些手段主要用于投毒攻击，不适用于对抗样本攻击。
18. 针对对抗样本攻击，利用输入梯度正则化训练可微模型并惩罚输入中的微小变化，降低了对抗扰动的影响。
19. 防御蒸馏技术: 降低深度神经网络结构的计算复杂度。它可以防御FGSM与基于雅可比矩阵(Jacobian Matrix)的迭代攻击。
20. 消除后门隐患: 针对后门攻击，通过修剪与反向触发器相关性较高的神经元来移除潜在的后门。
21. 抑制后门隐患: 采用加噪的SGD训练满足差分隐私机制的模型。由于训练过程中的随机性，触发样本的贡献将因随机噪声而减少，从而导致后门植入失败。实际应用中，加噪的SGD训练过程会降低模型的精度，所以此方案需要权衡模型的精度与鲁棒性。
22. 模型诊断: 模型诊断主要针对后门攻击，它可以判断模型是否嵌入了后门，进而重构受损模型的后门触发器。
23. 防御者可以通过分析模型中神经元的行为进行检测:
24. 人工脑刺激: 利用人工脑刺激扫描模型中的神经元进而识别后门模型。它需要对每个神经元完成扫描，其中受损神经元对特定类的输出激活异常高。虽然方案具有很高的检测效率，但是它的有效性靠一个强假设保证，即后门攻击仅由1个神经元执行，这在实际中不完全适用。
25. 控制流启发式: 从模型结构出发，利用神经网络进行源不变性与激活值分布不变性的检测。源不变性针对两个连续的隐藏层，而激活值分布不变性针对给定层神经元的激活分布。触发输入会违背这两种不变性，所以通过检测输入样本的两种不变性来识别触发样本并判断受感染的标签。
26. 防御者利用神经网络模型的参数、中间状态或输出训练元分类器以判断模型是否被感染: 将良性影子模型与后门影子模型作为训练样本训练元分类器。对每个影子模型进行若干次输入查询，对应的若干输出结果(如置信度评分)拼接后即为该影子模型的特征表示。训练过程中，防御者需要充分了解模型结构并且拥有强大的算力，为元分类器构造高质量训练集。
27. 识别后门模型后，防御者需要还原其后门触发器，进而抵御后门攻击:
28. 通用重构触发器技术(Neural cleanse): 对于每个输出类，他们利用逆向工程重建1个触发器，然后利用触发器识别恶意样本并修剪与触发器相关的神经元。该方法需要一部分保留的良性数据。
29. 利用最少信息完成检测的方法: 通过模型逆向技术与条件生成对抗式网络(conditional GAN, cGAN)为每个输出类生成１个触发器。

分布式安全攻击防御技术总结如下图所示。



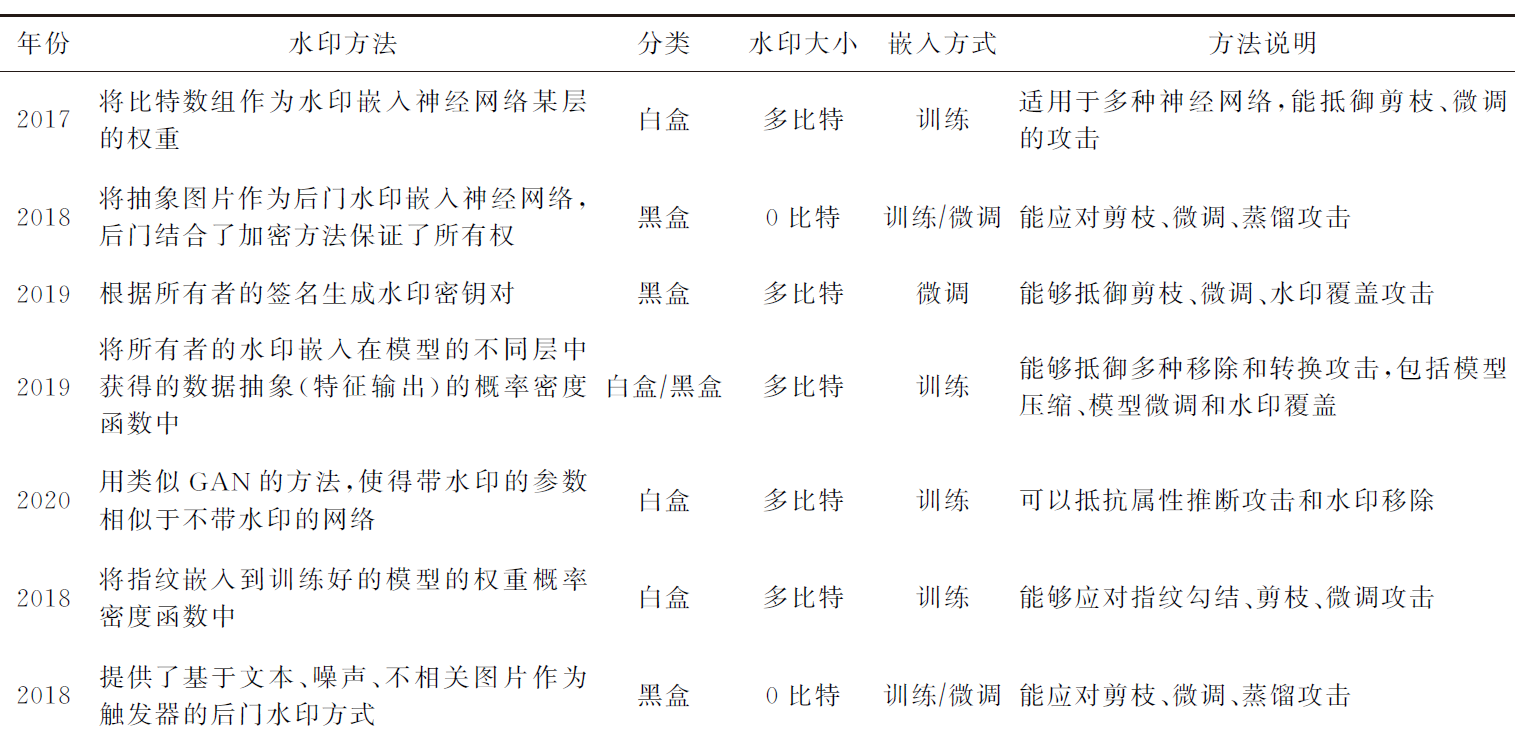
1. 神经网络水印技术[4]:

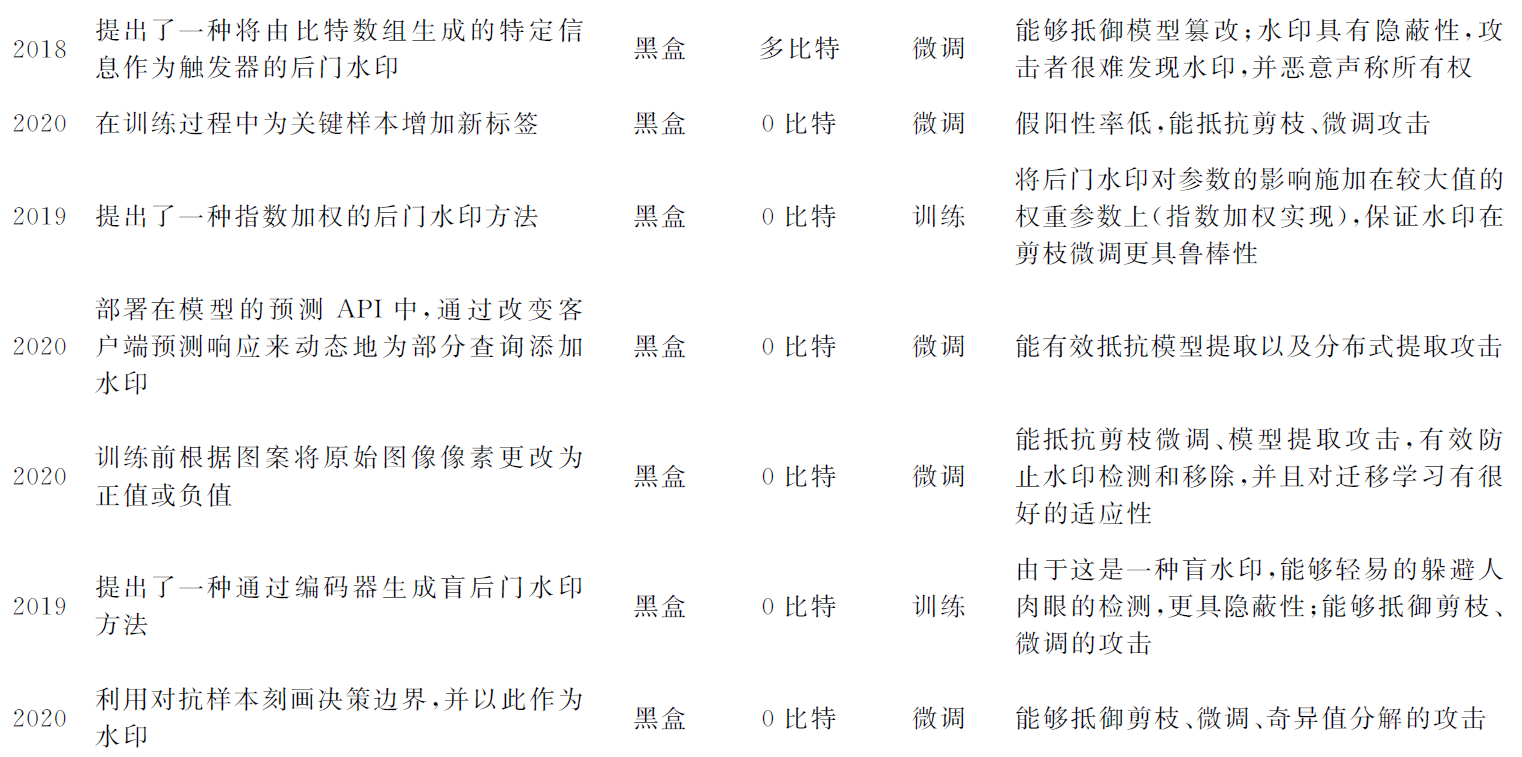
神经网络水印添加过程主要是通过在模型中添加一个额外的训练目标来注入水印，典型应用场景中的3个步骤，包括生成水印、嵌入水印和验证水印，如下图所示。

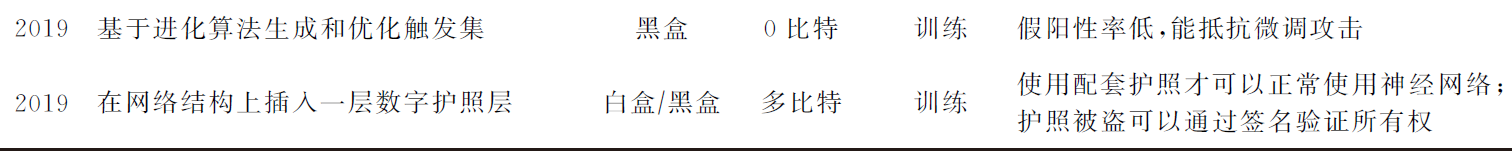


生成水印是指模型所有者设计特殊的水印形式，例如一个比特串或者一些经过特殊设计的训练样本，以便模型在验证水印阶段能够以某种特殊的方式验证水印的存在性。嵌入水印是指将生成的水印信息插入到神经网络模型中。在验证水印时，需要输入特定数据，然后观察模型的反馈或者输出，与预期结果进行匹配，从而验证水印的存在性。

神经网络水印技术总结如下表所示:







按照DNN 模型是否公开,分为白盒水印和黑盒水印两类:

1. 白盒水印: 将生成的水印信息嵌入到DNN模型参数，然后从模型中提取标记进行验证。具体有:
2. 通过显式的水印来保护DNN 模型的知识产权，将水印解释为T位的字符串,为了将其包含在模型中,使用了包含参数正则化器(parameter regularizer)的组合损失函数。该正则化器在某些模型参数上施加统计偏差，以表示水印。
3. 将字符串嵌入到不同网络层的概率密度函数(pdf)中，由于水印信息被嵌入DNN的动态内容中，生成水印同时依赖于数据和模型，即只能通过将特定的输入数据传递给模型来触发，使得水印更加灵活且不易被检测。
4. 基于生成对抗网络(GAN)的白盒水印方法: 使得最终的参数分布与无水印版本无法区分，因此很难被检测出来。
5. 端到端的共谋安全水印框架: 为每个用户分配一个唯一的二进制代码矢量(也称为指纹)，在保证准确性的前提下，将指纹信息嵌入DNN 权重的概率分布中，能够有效地跟踪每个用户模型的使用情况。
6. 黑盒水印: 在无需知晓模型参数等细节的情况下进行水印的验证。具体方法有:
7. 将后门技术应用于所有权保护的水印方法: 通过在模型的训练或微调的过程中，模型拥有者将抽象样本(与训练样本具有不同分布)作为后门样本加入模型训练集，如此保证模型在原始任务上的准确率，抽象样本作为后门水印进行所有权的认证。与此同时，通过commit加密方法生成的密钥进一步保证了水印在公开场景下验证所有权。
8. 基于文本触发器、噪声、不相关样本的后门样本方法: 通过特定触发器来触发后门水印验证所有权。
9. 保护嵌入式系统神经网络模型的所有权的水印方法: 将由比特数组生成的特定信息(mask)作为触发器设计后门水印。
10. 多比特(multi-bit)的黑盒水印方法: 提出了一种模型相关的编码方案，将所有者的二进制签名包含在输出激活中作为模型的水印。
11. 提出了为关键样本新增标签的方法: 尝试设计不会扭曲原始决策边界的水印。考虑到部分模型水印应对剪枝微调并不具备很强的鲁棒性。
12. 指数加权的后门水印方法: 将后门水印对参数的影响施加在较大值的权重参数上(指数加权实现)，保证水印在剪枝微调更具鲁棒性。
13. 将水印部署在模型的预测API中，通过改变客户端的预测响应来动态地为部分查询添加水印。增加了应对蒸馏攻击的鲁棒性。
14. 空嵌入方法(null embedding)方法: 将水印包含在模型的初始训练中。由于空嵌入不依赖于增量训练，只能在初始化时期训练为模型，因此对手很难再嵌入自己盗版水印。
15. 通过编码器生成盲(blind)的后门水印方法，能够躲避人肉眼或部分检测器的检测，更具隐蔽性。
16. 对抗样本法: 由于对抗样本的迁移性，在没有此水印的模型上，对抗样本也可能会以较高概率被判别错误，被误识别为水印，从而发生误报。
17. 水印技术保护DNN模型[6]:

包括基于参数的水印和基于标签的水印:

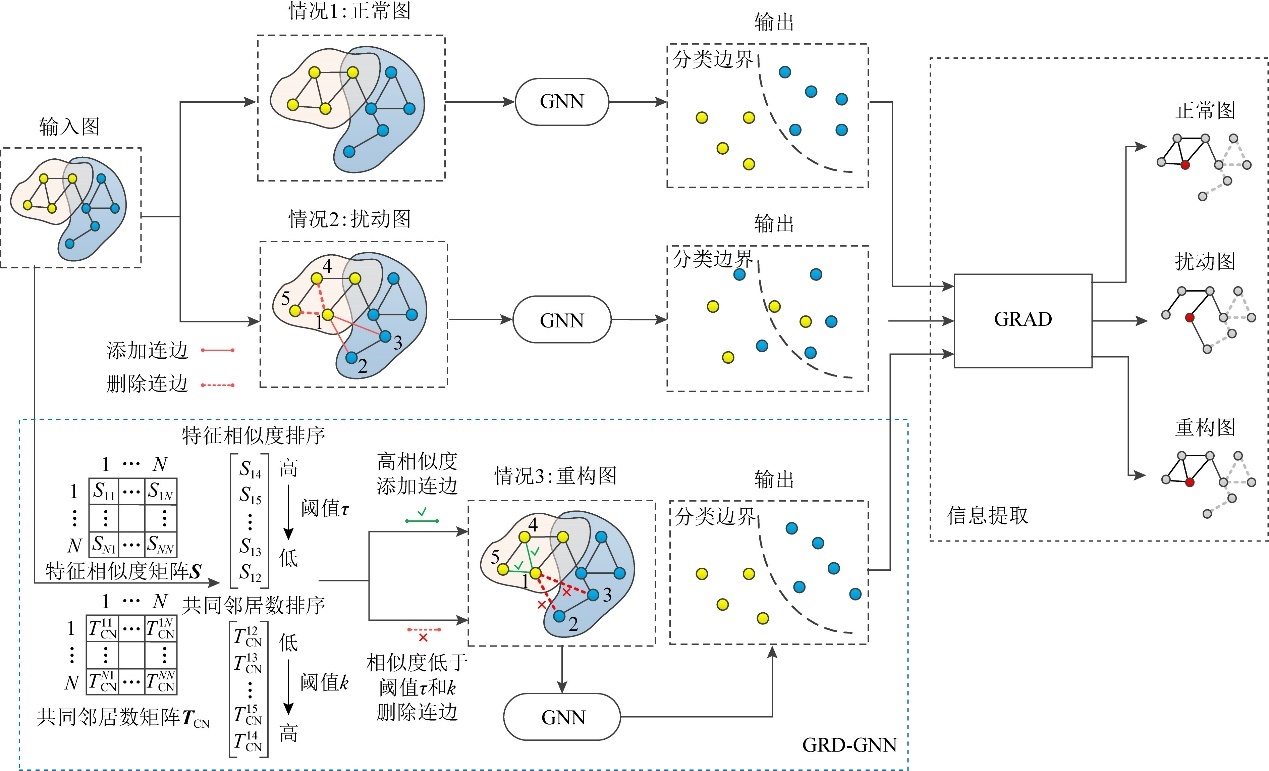
1. 基于参数的水印: 在目标模型训练时向损失函数添加正则化项，将水印嵌入到模型的参数中，要求模型所有者能够访问其他模型的模型参数，才能验证水印。
2. 基于标签的水印: 将水印嵌入到某些数据样本的预测标签或神经元激活层中，带有水印的样本和其他样本一同训练，通过查询替代模型进行水印验证。
3. 其他方法:
4. 在目标模型的决策边界附近寻找对抗样本，利用这些对抗样本的标签对模型的唯一性，区分目标模型和盗版的替代模型。
5. 构造目标性的对抗样本，并利用对抗样本的可转移性，生成可协商性对抗样本，用其标签来唯一标识目标模型，从而区分目标模型和替代模型。
6. 面向图神经网络的防御[5]:

针对防御方法的对象不同，防御方法可以分为两类: 针对输入数据转换的防御方法和针对模型结构的防御方法。

1. 针对数据转换的防御:
2. 在训练过程中随机丢弃一些连边进行对抗训练来达到防御对抗攻击的效果，实验证明了这种方法对于模型鲁棒性提高的作用不明显。
3. 根据节点之间的Jaccard相似度利用贪心算法寻找相似度低的连边，并通过设置阈值的方式删除连边达到防御效果，但是Jaccard相似度适用于2进制特征而忽略了特征中潜在的数值大小。
4. 基于奇异值分解的图纯化方法: 仅使用了top-k个奇异值对扰动图进行过滤，从而达到滤除对抗连边的效果，然而，由于仅考虑了前top-k个奇异值，因此存在丢失一些关键信息的风险。
5. Pro-GNN: 根据图的稀疏性、低秩性和节点的特征平滑设计的一种鲁棒的图神经网络，在中毒图的基础上重新构建干净图。实验证明，Pro-GNN具有良好的抵抗中毒攻击的能力，然而由于Pro-GNN需要迭代地构建干净图并交替地进行优化，这使得Pro-GNN的复杂度变高，训练时间较长，且需要占据较多的计算资源。
6. 针对模型结构的防御方法:
7. 对方差进行平滑: 利用高斯分布对噪声的容忍度，将其作为图卷积层中节点的隐层表示，并根据方差为邻居节点分配注意力权重，由于被攻击节点对其他节点的影响的方差大，并被分配了较小的注意力权重，因此可以降低不良影响的传播，该方法是对方差进行平滑而起到防御作用的，具有较强的通用性，但是该方法对不利影响的吸收能力也是有限的，对于一些较强的攻击，针对性不强，因此该方法的防御能力是有限的。
8. 对抗训练: 使用平滑蒸馏和平滑损失函数的方法来实现梯度隐藏，并对模型进行对抗训练，从而增强了图神经网络模型对基于梯度的对抗攻击方法的鲁棒性，该方法对基于梯度的对抗攻击方法具有较好的防御性能，但也需要生成大量对抗样本进行训练，防御代价较高。
9. PA-GNN: 一种具有惩罚性聚合机制的图神经网络。通过限制扰动连边的消息传递从而使聚合函数更专注于真实的邻居节点，由于需要预先使用对抗样本和注意力机制训练一个具有惩罚能力的元模型来对PA-GNN进行初始化，因此该方法的复杂程度较高。
10. 改进的GCN模型: 现有GCN模型的Laplacian算子在空间域中信息融合范围有限以及在谱域中的不良伪像。因此提出一种可变幂算子替代Laplacian算子，得到可变幂网络，可同时不同距离的特征变换函数和全局参数，从而提高图神经网络的鲁棒性。
11. 基于图重构的图神经网络防御方法[5]:

大部分对抗连边所连接的节点对之间几乎不存在共同邻居(common neighbors, CN)，这说明了它们周围的结构相似性较低。从节点自身属性特征考虑，这些被对抗连边所连接的节点属性特征也具有低相似度的特点。在此基础上，提出一种面向图神经网络的图重构防御方法GRD-GNN。

GRD-GNN算法框架如下图所示。



当GNN 模型的输入为一张正常图时，输出正常，并应用于图关键结构提取等任务，获

得正常的目标节点相关的关键子图；当GNN模型的输入为1张扰动图时，模型的分类性能下降，由解释器GRAD提取的目标节点相关子图与正常图不一致；GRD-GNN的输入无论是正常图还是扰动图都将依据节点特征相似度和共同邻居数进行图重构。

步骤为: 根据输入特征相似度阈值和共同邻居数阈值k，删除低于相似度阈值的连边，并连接具有高相似度的节点，输出重构图，并作为GNN模型的输入，获得正常的输出，并由解释器GRAD提取目标节点相关的关键子图，此时输出与正常图比较，验证GRD-GNN的有效性。主要分为两个步骤:

1. 滤除对抗扰动: 通过构建结构相似度和节点特征相似度对图中的连边进行贪心搜索，查找可疑连边，删除相似度得分低于给定阈值的连边，达到清除图中对抗连边的效果。
2. 图增强: 在步骤(1)的基础上，利用节点自身特征的相似度进行排序，在保证节点在输入图中的度值情况下连接高节点相似度的连边，达到图增强的效果。
3. 基于指纹的DNN保护[6]:
4. 指纹生成:

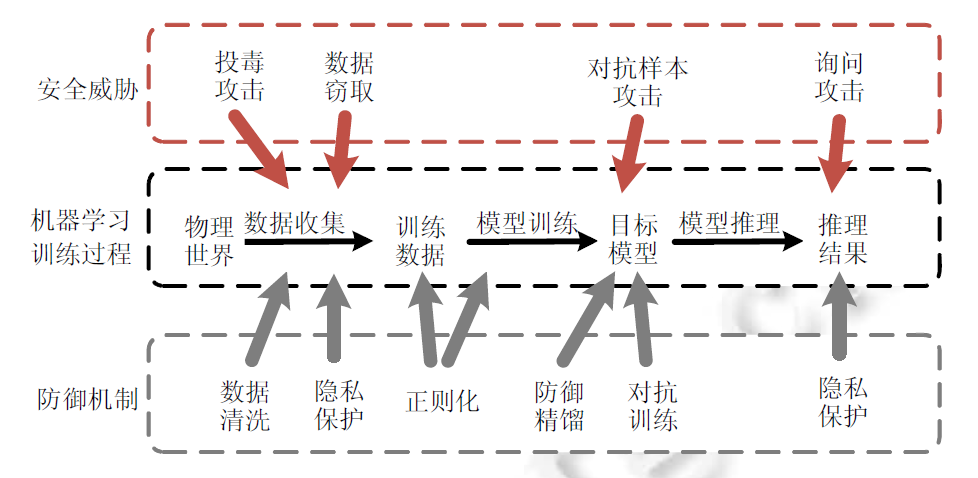
受生物指纹唯一性的启发，模型所有者可以通过某种算法提取模型的指纹，从而判断一个模型是否为目标模型或盗版的替代模型。采用指纹样本对应的预测标签来表示目标模型的指纹。此处指纹样本是指模型决策边界附近一些样本，可利用对抗样本来实现。

1. 指纹比对:

获得目标模型的指纹样本和指纹后，利用指纹样本查询替代模型，将替代模型返回的预测标签和目标模型的指纹进行比对，从而确定替代模型是否窃取了目标模型。具体地，目标模型指纹比对过程通过提取函数和比对函数来实现。

1. 机器学习中的安全攻击与防御机制[7]:

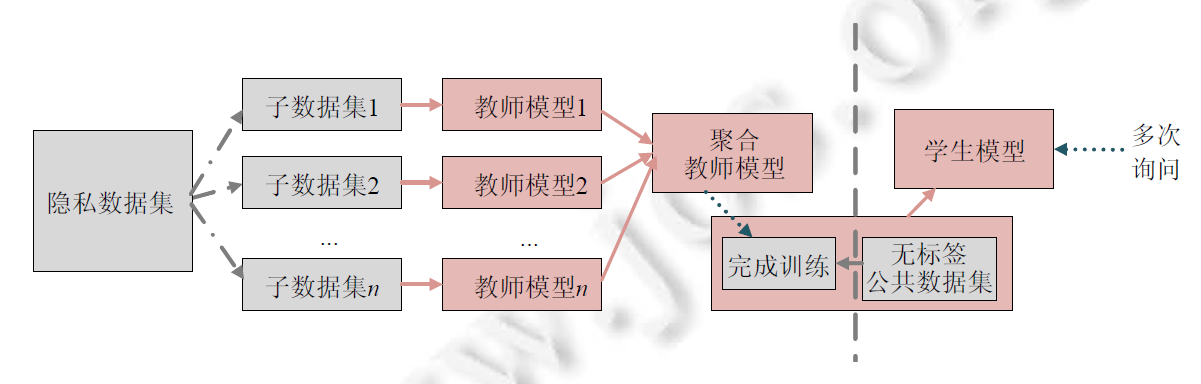
根据针对的攻击不同，现有的安全机制主要分为正则化、对抗训练、防御精馏、模型隐私改造、加密和扰动。输入空间的正则化主要针对训练数据的污染攻击，而模型参数的正则化、对抗训练和防御精馏主要针对推测阶段的对抗样本攻击。模型隐私改造、加密和扰动主要用来防御由数据窃取和询问攻击带来的多种安全问题。如下图所示。



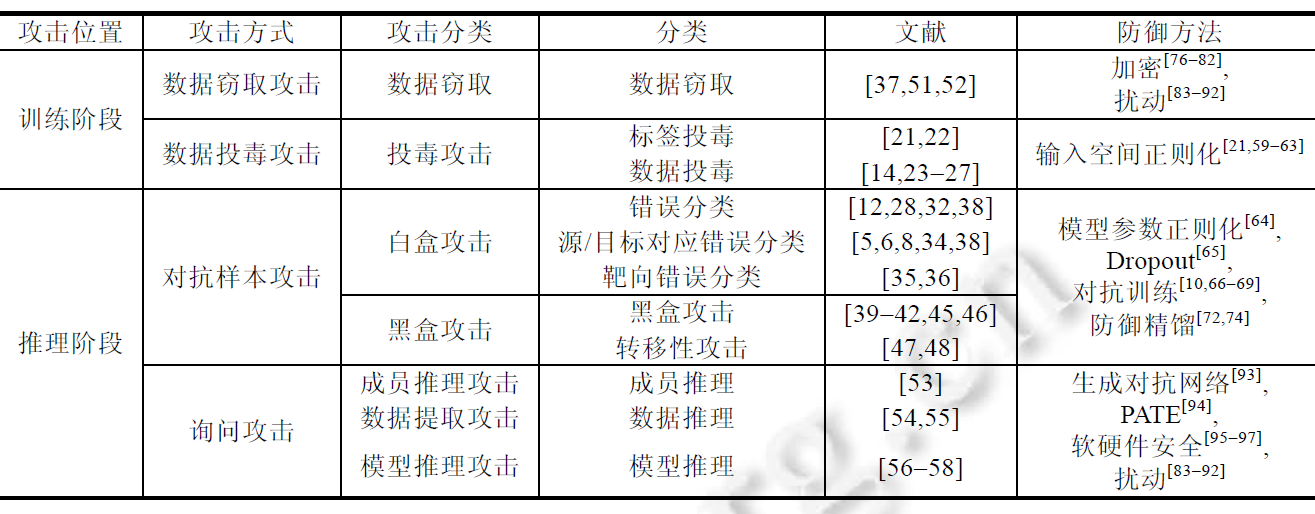
1. 数据集和模型正则化:

正则化是对机器学习进行规则化的过程，即通过对模型和模型输入的规范化操作，降低模型的出错率。正则化分为对训练数据的正则化和对模型的正则化：训练数据的正则化可以防御训练数据投毒攻击，而对模型的正则化可以防御对抗样本攻击。

1. 输入正则化: 在模型学习能力足够高而训练数据不足时，模型训练容易过拟合，从而在面对新的数据时出错率更高，因而容量足够的训练数据对模型训练至关重要；同时，在训练数据足够训练模型的情况下，攻击者可以通过污染训练数据的方式降低训练数据质量，从而使模型训练出错降低模型的准确性，因而要保证模型的性能，必须要保证训练数据的质量。高质量的训练数据有合理的特征空间和数据分布，同时有足够的数量。对训练数据的正则化可以理解为在保证训练数据存储安全的情况下，提升训练数据的质量。提升训练数据的质量称为数据集增强，即通过特征提取改变数据集的特征空间和数据分布，或通过注入噪声进行数据扩充，从而生成新的训练样本，创建具有更大容量甚至无限容量的增强数据集，从而提升模型的泛化能力。为了防御投毒攻击，多数集中式学习的防御机制建立在查找不在预期输入域内的样本上；而分布式学习的防御机制建立在查找参与者训练出的不在预期内的模型上。
2. 模型正则化: 模型正则化是利用正则化项对模型参数和训练方式进行规范化，进而提升模型泛化能力的过程。
3. 去噪自动编码器(DAE): 可以去除大部分对抗性噪声，但更小改动的新对抗样本对去噪自编码器和深度神经网络堆叠的网络依然攻击有效。
4. 深度收缩网络(DCN): 在损失函数中加入了平滑度惩罚，即正则化项，旨在最小化经验风险的同时，降低细微改动对模型输出的影响，以提高模型对对抗样本的鲁棒性。
5. 参数正则化: 降低数据改变对模型输出的影响，模型训练过程倾向于让参数数值尽可能稀疏(即非零参数尽可能少)，各个参数数值尽可能小。
6. Dropout: 模型参数的一种正则化过程，它改变的不是参数的数值，而是参数的数量。通过在训练期间随机丢弃神经元及其连接来构造简化的网络，通过强迫神经元和其他随机挑选剩下来的神经元共同工作，减弱神经元之间的联合适应性，提高模型的泛化能力。而在测试阶段，一个简化的网络就可以逼近所有简化网络预测的平均结果。实验表明：Dropout方法对数据集容量要求很高，在大容量数据集和增强数据集上能够获得很好的效果。
7. 对抗训练机制: 对抗样本攻击的产生，是机器学习模型输入的维度高而模型过于线性导致的，即是模型泛化能力不足，因而无法充分学习到训练数据和标签的映射关系。在一定的背景知识下，可以通过添加少量干扰产生对抗样本来跨越模型的决策边界，达到对抗攻击的目的。为了防御对抗样本攻击，除了提升训练数据质量外，研究从训练数据和模型改进两个方面，分别产生了对抗训练和防御精馏等安全模型。对抗训练指使用对抗模型产生带有完全标注的对抗样本和合法样本混合起来对原模型进行训练，以提升模型鲁棒性的防御机制。
8. 防御精馏: 精馏是通过一个模型的输出训练另一个模型的机器学习算法，是在保证训练精度的条件下压缩模型的方法。防御精馏是在精馏方案的基础上，通过两个相同模型之间的训练，达到梯度掩码，从而增强模型面对对抗样本的鲁棒性的方案。实验证明：使用防御精馏技术可以产生输出表面更平滑的、对扰动不敏感的模型提高模型的顽健性，且能够将对抗样本攻击的成功率从95%降到不足0.5%。
9. 隐私保护机制: 数据窃取攻击和隐私询问攻击是针对模型和数据的隐私攻击，通过使用加密、扰动方案，可以在根本上保护数据和模型的隐私，而模型的隐私改造可以使模型以保护隐私的机制进行学习。
10. 加密方案: 加密是保障数据安全性和隐私性的重要手段，在用户数据进入机器学习服务提供商之前，使用加密手段可以防止因存储和传输的安全漏洞导致的数据窃取攻击。同态加密、乱码电路、秘密共享机制和安全处理器机制是最常使用的加密方法。
11. 同态加密: 通过对训练数据和模型数据的加密实现了对数据隐私的保护，允许用户直接对密文进行相应的加法或乘法运算，得到数据仍是加密的结果，与对明文进行同样的操作再将结果加密一样。使用同态加密，用户加密的内容到达机器学习服务提供商后无法被解密，直接进入机器学习模型中，提升了模型的隐私性。同时，通过使用同态加密保证数据安全性不可避免地带来了效率问题，因此，同态加密多被应用于密文加性计算，而乘法仍在明文进行。
12. 乱码电路: 指需要保护的双方或多方要获得某项计算的结果时，将计算转换为乱码电路，并将自己的乱码输入发送给另一方，另一方可以根据电路和收到的乱码输入，结合自己的乱码输入获得计算结果并分享给发送方的方法。
13. 秘密共享机制: 秘密共享机制是利用shamir门限方案的特性，即w个参与者共享一个秘钥，任意t(门限值)个参与者都能计算出秘钥的值，而任何t−1个参与者都无法计算出秘钥的值。其中，为了共享秘钥，秘钥服务器需秘密地向每一个参与者发送一部分信息，这些信息称为共享(share)。将秘密共享机制应用在机器学习中，可以使用户训练的模型通过无共谋的服务器传输，最终通过门限值以上的子模型构建新的模型更新。
14. 模型加密聚合方案: 通过用户将使用秘密共享秘钥和用户私钥双重加密的内容互相传递的方式，在获得门限值以上的共享后解密得到最终结果，可实现高效率、高维度的数据加密传输计算。但加密聚合方案使得服务器无法根据数值判断收集到的子模型是否被污染或破坏，使得攻击者通过攻击分布式子模型从而破坏全局模型成为可能。
15. 安全处理机制: 通过硬件设备的安全性保证计算安全性的方案。如SGX(Intel software guard extensions) 通过一组CPU 指令，隔离应用程序代码和数据的特定可信区域，为开发人员提供安全可信空间，使敏感数据或代码免受外部的干扰或检查。
16. 扰动方案: 针对成员推理攻击，差分隐私机制，为数据库分析算法提供了很好的隐私标准。对于一种随机化算法M，其分别作用于两个仅相差一个样本的相邻数据集。差分隐私指出：通过合理的数学计算和对数据添加干扰噪声的方式保护所发布数据中潜在的用户隐私信息，可以使攻击者在拥有完美背景知识的情况下，通过询问攻击无法识别单个个体。基于差分隐私的这种特性，将其应用于机器学习的数据和模型保护中，可以防止成员推理攻击，也可以在分布式学习中保证原始数据的隐私性。根据应用位置的不同，可以分为输入扰动、模型扰动和输出扰动。
17. 模型的隐私改造: 为了应对多种使用环境和攻击类型，对模型结构的改造和创新能够在提高效率的同时满足安全性要求。集中式学习方案使服务提供商无条件获取用户隐私，因此，分布式机器学习应运而生。
18. 分布式生成对抗网络: 利用生成对抗网络对联合分布式训练深度学习模型发起攻击，提高网络的隐私性。这使得任意参加训练的用户都可能成为敌手，生成与其他参与者训练数据无限逼近的假样本来窃取他人隐私。
19. PATE模型: 精馏是使用一个教师模型的输出训练另一个学生模型以进行模型压缩的方法，而PATE模型通过将集中式模型分解成多个老师模型最终聚合成学生模型的方式，降低了模型的敏感度，通过给聚合过程增加差分隐私(differential privacy)的方式提升了隐私性，并通过隐蔽前置操作，仅向用户提供学生模型询问接口的方式降低了模型反演攻击的可能，提升了安全性。其模型结构如下图所示。

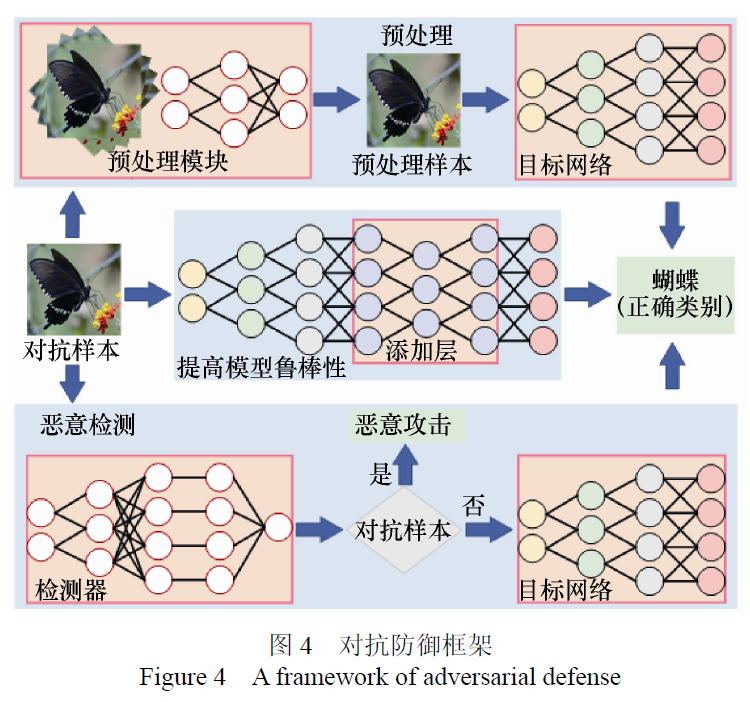


安全攻击与防御机制总结如下表所示。

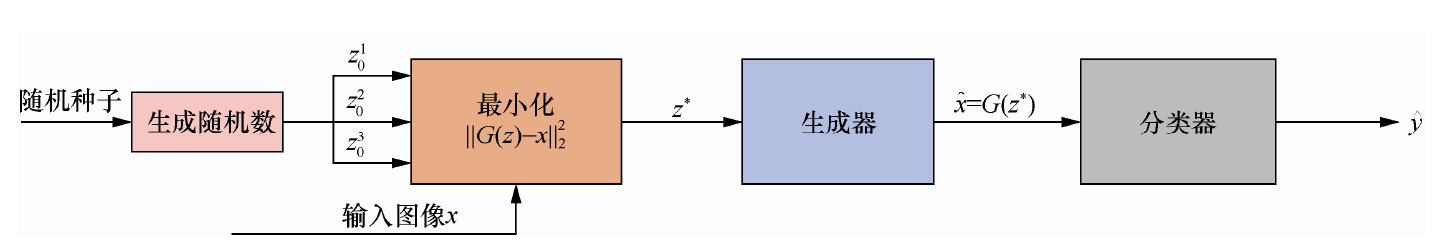


1. 针对对抗样本攻击的防御策略[8]:

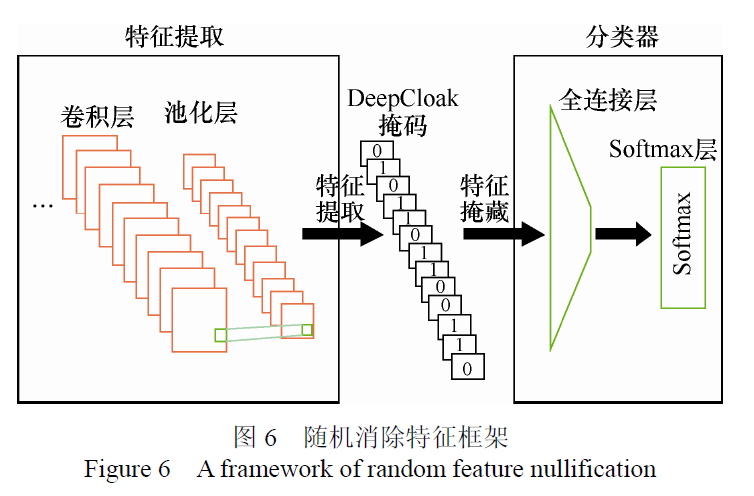
可以将防御策略分为3类: 预处理、提高模型鲁棒性、恶意检测。如下图所示。



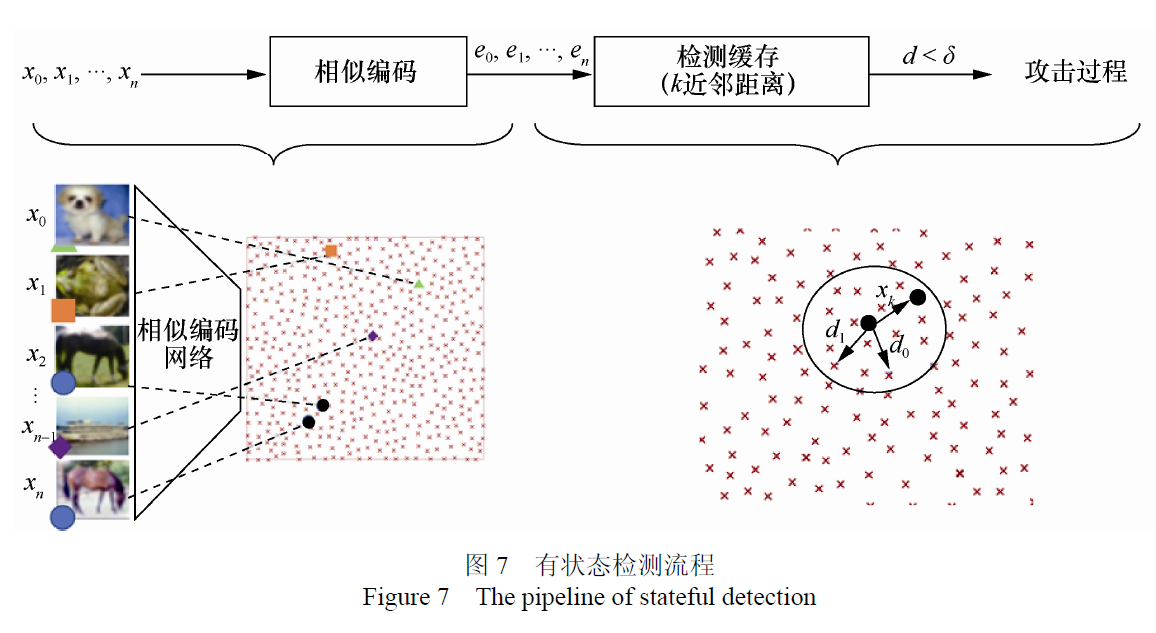
1. 预处理: 在图像输入网络之前对图像进行去噪、随机化、重构、缩放、变换、增强等操作，减轻对抗扰动对模型分类的影响，通常无须对模型进行任何的修改，可以直接应用于已经训练好的模型，计算开销较低。
2. 随机化:
3. 随机特征消除: 通过在样本中随机消除特征（类似dropout的方式）来阻止攻击者构建有效的对抗样本，虽然dropout训练时随机让神经元失效，但测试阶段的防御效果不显著。而随机消除特征在测试阶段能够随机让神经元失效。
4. 像素偏移: 包括重新分配像素值和小波去噪两个部分。重新分配像素值利用了卷积神经网络对自然图像中噪声的鲁棒性，随机将一些像素替换为一个小邻域中随机选择的像素。小波去噪修复重新分配像素值和对抗性扰动带来的图像破坏，使图像变得更加自然。
5. 像素重绘: 首先训练一个预测模型用于生成预测图像，并对图像像素值的取值范围进行区间划分。原始图像经预测模型生成预测图像，获得预测图像的每个像素值所在区间，再用区间内的随机值替换原始图像的像素值。
6. 图像变换: 对于FGSM算法产生微小扰动的对抗样本，JPEG 压缩能够减轻对抗样本导致的分类错误，但随着扰动程度的增大，JPEG 压缩的防御效果会降低。JPEG 压缩的一个重要能力就是它能够去除图像内部的高频信号分量，相当于有选择性地模糊图像，可以消除图像上的对抗性扰动。
7. 去噪网络:
8. 针对通用型扰动攻击算法的防御框架: 该框架包含扰动校正网络和扰动检验器两部分。该框架将扰动校正网络作为额外的预输入层附加到目标网络中，在不更新目标模型参数的情况下训练它们对被扰动后的图像进行校正，使分类器对对抗样本和其原始图像预测结果相一致。扰动检验器是将扰动修正网络输入输出的特征差异作为输入，通过支持向量机[67]学习得到二元分类器。输入图像首先经过扰动修正网络，然后使用扰动检验器进行检测是否存在扰动。如果检测到扰动，就用经过扰动修正网络修正后的图像代替原始图像作为分类器输入。
9. U-net去噪网络模型: 模型与传统的自编码器有两点区别：第一，去噪网络使用的不是像素层面的损失函数，而是使用特征图作为损失函数；第二，U-net去噪网络学习的是对抗扰动，而不是构造整个图像。利用去噪网络得到对抗性扰动，并结合原图像得到去噪后的图像，但这种方法不能有效抵御白盒攻击。
10. 对抗生成网络:
11. 对抗生成网络防御框架: 主要思想是采用原始数据集训练一个对抗生成网络，利用生成器的表达能力重构一个与原始图像相似但不含对抗扰动的重构图像。输入图像经过对抗性生成网络进行重构后，得到一个与原始图像相似的重构图像，再将重构的图像输入目标网络模型进行分类。其中引入了随机种子，使整个网络模型难以攻击。整个防御框架如下图所示。



1. 特征双向对抗生成网络: 输入图像经过特征双向生成网络映射，提取语义特征，这些特征不随扰动而变化，并根据语义特征把输入图像重构成无扰动图像。输入图像经过双向对抗生成网络的重构后输入目标模型进行分类。实验表明，在白盒和灰盒攻击下，这种防御方法对于预先训练好的任意分类器都有效。
2. 超分辨率: 基本思想是使用超分辨率网络将流形外的对抗样本引入自然图像流形中，从而将分类恢复到正确的类别。算法采用小波去噪减少恶意对抗扰动，然后使用超分辨率网络加强图像视觉质量。该方法在无须重新训练模型的前提下，可以补充到现有的防御机制，同时提升分类准确率。
3. 提高模型鲁棒性: 通过修改模型架构、训练方式、正则化、特征去噪等方式实现，增强模型抵抗对抗样本的能力，但需对模型进行重新训练，计算开销较大。
4. 对抗性训练: 对抗性训练作为目前能够有效提高模型鲁棒性的防御方式，其缺点是训练模型的开销太大且无法对所有攻击类型的对抗样本进行对抗训练。
5. 双对抗样本训练方法: 既能够抵抗单步的对抗样本，也可以防御迭代的对抗样本。
6. 三重损失函数: 利用距离度量学习中常见的三重损失函数来构建对抗训练模型。三重损失函数可以优化嵌入空间，即具有相同标签的样本在空间中尽可能地接近，不同标签的样本尽可能地远离。由于很难找到一个能够代表对抗样本域的样本，因此对抗训练泛化能力较弱。
7. 领域适应能力的对抗训练方法: 旨在学习样本有意义的表示，这种表示在对抗样本和原始数据集上是不变的。
8. 批量调整网络梯度: 一种简单、高效、可以提高模型鲁棒性的训练方法对梯度进行调整，用于提高分类正确样本对模型参数更新的贡献值。该方法的优点在于不依赖任何形式的数据扩充和使用对抗样本进行对抗性训练，同时能够保持甚至增强整体分类性能并抵抗对抗攻击。
9. 正则化:
10. 流形正则化网络: 在模型损失上附加一个盲点特性的流形损失项。结果表明，该方法能够抵抗对抗样本，并且能够使模型在流形上泛化。缺点在于梯度正则化是二阶的，这导致模型的训练时间翻了一倍。
11. 特征优先模型: 通过基于非线性注意力模块和特征正则化的特征优先模型，使模型分类依赖于关键特征。其中，注意力模块通过给关键特征分配更大的权重，促使模型学习到关键特征。对模型进行正则化促使提取原始图像和对抗样本相似的本质特征，有效忽略了添加的扰动。
12. 防御蒸馏: 网络蒸馏技术原本是为了把网络模型部署到移动端而设计的一种网络压缩技术，它能够把大的网络压缩成小的网络，而且能够保持网络的性能。在蒸馏的过程中通过调整温度（softmax函数中的参数T），产生更加平滑、对扰动更加不敏感的模型，从而提升模型对对抗性性样本的鲁棒性。
13. 特征去噪: 虽然去噪模块并不能提高在原始数据集下的分类准确率，但去噪模块和对抗性训练结合在一起，在白盒攻击和黑盒攻击中能够显著提高模型的鲁棒性。
14. 随机消除特征: DeepCloak机制如下图所示。利用掩码的方式删除网络模型中不必要的特征，限制了攻击者生成对抗样本的能力，从而提升模型的鲁棒性。与其他的防御机制相比，DeepCloak 机制更易于实现且计算效率较高。

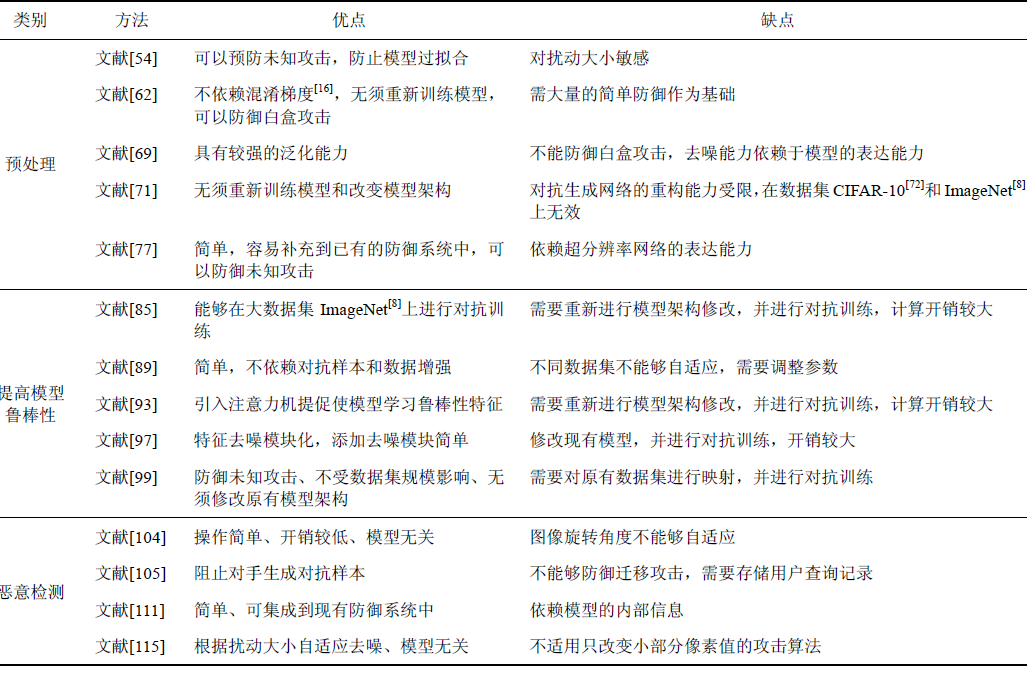


1. 卷积稀疏编码: 基于卷积稀疏编码，构造一个分层的低维准自然图像空间，通过将原始图像和对抗样本映射到一个低维准自然图像空间来实现高水平的鲁棒性。这个准自然图像空间近似于自然图像空间。同时该方法消除了对抗性扰动，使对抗样本接近其在特征空间中的原始输入。在训练阶段，作者在输入图像和神经网络的第一层之间引入一个稀疏变换层来有效地将图像映射到准自然图像空间中，用映射到准自然空间的图像对分类模型进行训练。在测试阶段，把原始输入映射到准自然空间的图像作为分类模型的输入。与其他不可知攻击的对抗性防御方法相比，该方法在对抗性扰动的大小、各种不同的图像分辨率和数据集规模方面，都有具有更强的鲁棒性。
2. 深度收缩网络: 模型采用了一种全新的端到端的训练过程，该过程采用伸缩自编码器相似的平滑损失、逐层惩罚，近似地使网络输出的方差相对于输入的扰动最小，使模型在训练数据点周围更加平滑。这增强了模型对对抗样本的鲁棒性，并且没有明显的性能损失。
3. 防止可移植性: 使用空标签的方法来防御黑盒下的对抗样本可移植性攻击。其主要的思想是通过对数据集的标签增加一个空类别，并进行对抗性训练。该方法的优势在于能够将对抗样本的错误标签分类到空类别中，而不是其他的错误类别，有效阻止对抗样本的可移植性，同时图像能够保持模型的精度。
4. 恶意检测: 检测用户输入的图像是否是恶意图像，从而阻止对抗攻击。
5. 图像变换: 首先，对一幅图像进行一组的变换操作，生成多幅变换后的图像。然后，使用这些变换后的图像的分类结果作为特征训练一个神经网络来预测原始图像是否受到了攻击者的干扰。为了防御更复杂的白盒攻击，在转换过程中引入随机性。对多个图像数据集的实验结果表明，C&W算法产生的对抗样本的检测率达到99%。其中，对于白盒攻击，该方法的检测率达到70%以上。
6. 有状态检测: 黑盒防御方法。有状态检测方法思想是记录一定量的用户询问记录，并且在用户下次询问时，将用户询问与以往的记录进行对比，如果相似程度在规定的一定阈值内则认为是恶意攻击。实验分析表明，基于询问的黑盒攻击通常需要几十万到上百万的询问，这非常容易触发用户的防御机制。即使不触发防御机制，攻击所需的存储服务也需消耗大量资源。有状态的检测方法缺点是无法防御不需要询问的迁移攻击。但是，该方法能够与防御迁移攻击的集成对抗性训练方法进行结合，可以弥补有状态检测的不足，使其在黑盒攻击情况下有较好的表现。具体过程如下图所示。

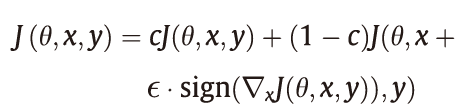


1. 隐藏层特征:
2. 分析分布: 通过分析对抗样本与正常的样本是否来自相同分布的方式来进行对抗样本的检测。该检测方法使用中间层的特征，即基于卷积层的输出，而不是直接使用原始图像和对抗样本进行统计分布。
3. 级联分类器: 从一个特定的对抗算法生成数据并进行训练，得到的分类器可以成功地检测从完全不同机制生成的对抗样本。
4. 高斯混合模型: 使用高斯混合模型对神经网络最后一层的输出进行建模，用来检测对抗样本。由于神经网络尾部的隐藏层可以捕获到输入的高级语义信息，在最后一层上使用一个简单的分类器将比其应用于原始输入图像更加准确可靠。
5. 判断样本分布状态: 分析深度神经网络中隐藏神经元的输出分布，使用高斯混合模型来近似深度神经网络分类器的隐藏状态分布，然后通过判断输入样本状态分布是否异常来检测对抗样本。该方法能够应用于任何深度神经网络结构上，并可以与其他防御策略相结合，以此提高模型的鲁棒性。实验表明，该方法能够防御黑盒和灰盒攻击。
6. 流形学习:
7. 检测器和修正器: 检测器在训练期间学习对抗样本的流行分布，对于远离流形边界的样本判定为对抗样本。而修正器在流形区域中寻找一个与输入相近的替换并输入分类器进行分类。
8. 随机扰动: 通过添加随机扰动来扩展对抗子空间能够使对抗样本被重新分类正确。而且由于神经网络模型的鲁棒性，添加微弱的随机扰动并不会影响分类结果。
9. 模型无关方法: 该方法通过分析模型对于随机扰动输入的反应，以相对变化的置信度作为检测对抗样本的鉴别标准，在理论框架中学习范数有界对抗性扰动的鲁棒性，可以轻易地部署到现成的深度学习模型中。
10. 自适应去噪: 利用交叉熵的大小作为标准实现自适应去噪。其中包含两种图像降噪方式：标量量化和空间平滑滤波。先使用交叉熵自适应量化区间大小，然后判断是否需要进行空间平滑滤波。降噪后的图像和原始输入分别用分类器进行分类，如果分类结果一致，则认为原始输入是正常样本，否则认为是对抗样本。

以上防御方法总结如下表所示。

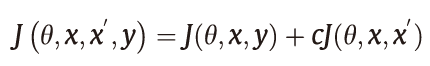


1. 对抗防御[9]:
2. 对抗训练: 对抗训练是一种针对对抗样本的直观防御方法，该方法试图通过利用对抗样本进行训练来提高神经网络的鲁棒性。
3. FGSM对抗训练: 提出用良性和FGSM生成的对抗样本训练神经网络以增强其网络鲁棒性的方法，对抗目标函数可以表示为

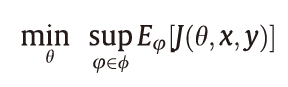


尽管该方法对FGSM的攻击有效，但是训练后的模型仍然容易受到基于迭代/优化方式的对抗攻击。

1. PGD对抗训练: PGD攻击可能是通用的一阶攻击，那么针对PGD的模型鲁棒性意味着可以抵抗各种一阶攻击。PGD对抗训练总体上是针对攻击的最有效对策。但是由于生成PGD对抗样本需要大量计算成本，因此PGD对抗训练不是一种有效率的防御措施。
2. 集成对抗训练: 利用多个预先训练的模型中转移的对抗样本进行对抗训练，即集成对抗训练(ensemble adversarial training, EAT)。直观上，EAT增加了用于对抗训练的对抗样本的多样性，从而增强了针对从其他模型转移过来的对抗样本的神经网络鲁棒性。实验结果表明，EAT模型对其他模型遭受的各种单步和多步攻击产生的对抗样本都具有较强的鲁棒性。在某些情况下，EAT对黑盒和灰盒攻击的抵抗性能甚至优于PGD对抗训练。
3. 对抗Logit配对: 通过将良性样本的logits层和相应的扰动样本之间的交叉熵包括在训练损失函数中，来鼓励提高成对样本在学习到的Logits层的相似性。该方法的损失函数为:



1. 生成对抗训练:
2. 在对抗性训练的过程中利用非确定性生成器来生成对抗性样本。设置一个生成器，该生成器的输入是训练好的模型在良性样本点上的梯度，并产生类似FGSM的对抗性扰动。通过在良性样本和生成样本上训练分类器，与FGSM对抗训练模型相比，获得了一个对FGSM具有更强鲁棒性的模型。
3. AC-GAN架构: 使用AC-GAN架构进行数据扩充，从而提高PGD对抗训练的通用性。通过将PGD对抗样本作为真实样本输入到鉴别器中，AC-GAN学会了生成与PGD对抗性样本相似的伪造样本。类似于PGD的假样本将被用来训练辅助分类器和预训练的鉴别器。生成器、鉴别器、辅助分类器和PGD攻击的这种组合在单个网络中不仅会形成更强大的分类器，而且可以形成更好的生成器。
4. 随机化: 基于随机化的防御方法试图将对抗性效应随机化为随机性效应，当然这对大多数DNN而言都不是问题。在黑盒攻击和灰盒攻击的设置下，基于随机化的防御获得了不错的性能，但是在白盒攻击下，EoT方法能够通过在攻击过程中考虑随机过程来破坏大多数防御方法。
5. 随机输入变换:
6. 利用随机调整大小和填充这两种随机变换来减轻推理时的对抗效果。随机调整大小是指在将输入图像输入DNN之前将其调整为随机大小。随机填充是指以随机方式在输入图像周围填充零。这种快速而敏锐的机制如图8所示。该防御方法在黑盒攻击下取得了卓越的性能，然而在白盒攻击下这种防御会被EoT方法破坏。
7. 将图像送入到CNN之前使用具有随机性的图像变换，如位深度减小、JPEG压缩、总方差最小化和图像缝合。这种方法可以抵抗由多种主流攻击方法生成的60%的强灰盒对抗样本和90%的强黑盒对抗样本。但是它也会受到EoT方法的损害。
8. 随机噪声:
9. RSE(random self-ensemble): 防御对抗性干扰的随机噪声机制。在训练和测试阶段，RSE在每个卷积层之前添加一个噪声层，并集成随机噪声的预测结果以确保DNN有稳定的输出。
10. PixelDIP: 一种基于差分隐私(differential privacy)的防御。PixelDP在DNN内集成了DP噪声层，基于范数的微小扰动会引起预测结果概率分布的变化，通过这种变化可以给出DP边界。PixelDP可在使用Laplacian/Gaussian DP机制防御的前提下抵抗L1/L2攻击。
11. 在分类之前将随机噪声直接添加到对抗性样本的像素中，从而消除对抗性扰动的影响。基于Renyi散度理论，证明利用输出概率分布（向量）的第一和第二大概率可以确定其对抗扰动的鲁棒上限。
12. 随机特征修剪:
13. 随机激活修剪(stochastic activation pruning, SAP): 该方法通过随机修剪每一层中的一部分激活子集，并优先保留更大幅度的激活项来保护预训练网络免受对抗样本的攻击。在激活修剪之后，SAP会扩展尚存的激活从而标准化每个层的输入。
14. 通过随机掩盖卷积层输出的特征图设计了一种新的CNN结构。通过随机掩盖输出特征图使每个过滤器仅从局部位置提取要素。
15. 去噪: 就减轻对抗性扰动/效果而言，降噪是一种非常简单的方法。之前的工作指出了设计这种防御的两个方向，包括输入降噪和特征图降噪。其中第一个方向试图从输入中部分或完全消除对抗性扰动，第二个方向是减轻对抗性扰动对DNN学习高级功能的影响。
16. 常规输入整流: 利用两种压缩（去噪）方法：位减少和图像模糊，以减少自由度并消除对抗性扰动。通过比较原始图像和压缩图像上的模型预测值来实现对抗样本检测。如果原始输入和压缩输入产生的输出与模型有很大差异，则原始输入可能是对抗样本。
17. 基于GAN的输入清理:
18. 防御-GAN(Defense-GAN): Defence-GAN训练生成器来对良性图像的分布进行建模。在测试阶段，Defense-GAN通过在其学习的分布中搜索接近于对抗样本的图像来清除对抗样本，然后将良性图像输入分类器。这种策略可以用来防御各种对抗攻击。
19. 对抗干扰消除-GAN(adversarial perturbation elimination-GAN, APE-GAN): 直接学习生成器，它将对抗样本作为输入，输出其对应的良性样本，从而清晰对抗样本。
20. 基于自动编码器的输入去噪: 比如一种称为MagNet的两节防御系统，其中包括一个探测器和一个重整器。在MagNet中使用自动编码器来学习良性样本的多种形式。检测器根据样本与学习到的良性样本的多种形式之间的关系来区分良性样本和对抗样本。重整器用于将对抗样本纠正为良性样本。
21. 特征去噪: 比如一种基于高级表示法指导的去噪器（high-level representation guided denoiser, HGD），这种去噪器可以改善受对抗性扰动影响的特征。HGD不是使用像素级去噪，而是使用特征级损失函数训练降噪的U-NET，这样可以最大限度地减少良性和对抗性样本之间的特征级差异。
22. 可证明式防御: 如果无法计算理论上的错误率，这些启发式防御可能会被未来的新攻击所打破。因此许多研究者致力于探索可证明的防御方法，在一类定义明确的攻击下，这些方法始终能保持一定的准确性。
23. 基于半定规划的可证明式防御: 提出了一种针对两层网络生成的对抗样本的可证明的防御方法。导出攻击者损失函数上界的半定松弛，并将其作为正则化项加入训练损失函数中。
24. 基于对偶方法的可证明式防御: 提出一个对抗问题来界定对抗样本的输出多维区域。他们表明可以通过在另一个深度神经网络上进行优化来解决对偶问题。与仅适用于两层全连接网络有所不同，该方法可应用于具有任意线性算子层（如卷积层）的深层网络。
25. 分布稳健性证明: 从分布优化的角度可以将对抗性分布的优化问题表述为



在此分布目标上进行的优化等效于将对良性数据相邻的所有样本（即对抗样本的所有候选项）的经验风险最小化。由于P会影响计算的复杂性，并且难以在任意P上进行直接优化。在训练过程中，引入了一个惩罚项来描述对抗稳健性区域。由于很难对该惩罚项进行优化，因此提出了针对该惩罚项的拉格朗日松弛法，从而实现了对分布损失函数的鲁棒优化。

1. 稀疏权重DNN: 针对FGSM和DeepFool生成的对抗样本的权重稀疏性和网络稳健性之间的内在联系。对于线性模型，文献证明了对抗样本的优化会导致网络权重的稀疏化。对于非线性神经网络，当权重矩阵较稀疏时，网络Lipchitz常数倾向于更小。由于研究发现最小化Lipchitz常数有助于提高网络的鲁棒性，因此可以得出结论，稀疏权重也将导致神经网络更鲁棒。
2. 基于KNN的防御:
3. 分析k最近邻(KNN)算法的对抗鲁棒性框架: 该框架确定了具有不同鲁棒性的k的两个不同状态。假设常数为k的KNN在条件概率p(y=1|x)位于(0, 1)的区域中，这样的模型在大样本限制条件下不具有鲁棒性。
4. DkNN: 对DNN每一层学习的数据表示执行KNN算法。KNN算法主要用于估计测试输入的异常预测。当DNN学习的中间表示与那些和预测共享同一标签的训练样本的表示不接近时，则认为该预测是异常的。实验表明，在多次对抗攻击下，尤其是在C&W攻击下，DkNN可以显著提高DNN的准确性。
5. 基于贝叶斯模型的防御:
6. 将贝叶斯神经网络（Bayesian neural network, BNN）与对抗训练相结合，从而学习在对抗攻击下的最优模型的权重分布。假设网络中的所有权重都是随机的，并使用BNN理论中常用的技术训练网络。
7. ABS(Analysis by Synthesis model): 基于贝叶斯模型对输入数据的分类条件分布进行建模，并将新样本分类为相应类别条件模型产生最高似然性的类别。
8. 基于一致性的防御: 对于诸如音频识别和图像语义分割之类的机器学习任务，可以应用一致性信息来区分良性样本和对抗性样本。
9. 对于语义分割任务，对抗性干扰一个像素也会影响其周围像素的预测。因此对单个区块进行干扰也会破坏其附近元素之间的空间一致性。这种一致性信息可以用于区分良性和对抗性图像。用自适应攻击方式对这种基于一致性的防御进行了评估，最终证明了这种防御比其他异常检测系统具有更好的性能。
10. 对于音频识别任务，对抗性扰动将破坏时间一致性。对于对抗性音频信号，信号的一部分平移与整个信号的平移不一致。

[1]李明慧,江沛佩,王骞,等.针对深度学习模型的对抗性攻击与防御[J].计算机研究与发展,2021,58(5):909.

[2]华为技术有限公司.AI安全白皮书[EB/OL].[2020—0-12].https://www-file.huawei.com/-/media/corporate/pdf/cyber-security/ai-security-white-paper-cn.pdf.

[3]周纯毅,陈大卫,王尚,等.分布式深度学习隐私与安全攻击研究进展与挑战[J].计算机研究与发展,2021,58(5):927.

[4]张颖君,陈恺,周赓,等.神经网络水印技术研究进展[J].计算机研究与发展,2021,58(5):964.

[5]陈晋音,黄国瀚,张敦杰,等.一种面向图神经网络的图重构防御方法[J].计算机研究与发展,2021,58(5):1075.

[6]钱亚冠,何念念,郭艳凯,等.针对深度神经网络模型指纹检测的逃避算法[J].计算机研究与发展,2021,58(5):1106.

[7]李欣姣,吴国伟,姚琳,等.机器学习安全攻击与防御机制研究进展和未来挑战[J].软件学报,2020,32(2):406-423.

[8]刘西蒙,谢乐辉,王耀鹏,等.深度学习中的对抗攻击与防御[J].网络与信息安全学报,2020,6(5):36.

[9]Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu. Adversarial Attacks and Defenses in Deep Learning. Engineering, https://doi.org/10.1016/j.eng.2019.12.012.