Adversarial attacks and defenses in explainable artificial intelligence: A survey

摘要

可解释的人工智能 （XAI） 方法被描述为调试和信任统计和深度学习模型以及解释其预测的补救措施。然而，对抗性机器学习 （AdvML） 的最新进展凸显了最先进的解释方法的局限性和脆弱性，使其安全性和可信度受到质疑。当应用于高风险决策和知识发现时，纵、愚弄或公平清洗模型推理证据的可能性会产生有害后果。本调查全面概述了有关对抗性攻击的研究，包括机器学习模型的解释以及公平性指标。我们引入了统一的方法符号和分类法，为来自 AdvML 和 XAI 交叉研究领域的研究人员和从业者提供了共同的基础。我们将讨论如何防御攻击并设计强大的解释方法。我们列出了 XAI 中现有的不安全因素，并概述了对抗性 XAI （AdvXAI） 的新兴研究方向。未来的工作应解决改进解释方法和评估方案的问题，以考虑报告的安全问题。

**介绍**

可解释的人工智能 （XAI） 方法（简要概述见 [1]，综合调查见 [2]），例如 PDP [3]、SG [4]、LIME [5]、IG [6]、SHAP [7]、TCAV [8]、Grad-CAM [9] 等事后解释，提供了各种机制来解释机器学习模型的预测。对 XAI 的普遍批评是它无法忠实地解释黑盒预测功能 [10]。尽管如此，解释在自动驾驶 [11] 或药物发现 [12] 等应用中取得了成功，并且可以用来更好地理解 AlphaZero [13] 等大型模型的推理。

最近，与对抗性机器学习相关的研究 [AdvML， 14–16] 在 XAI 研究中变得更加普遍，该研究揭示了解释方法的漏洞，这些漏洞引发了对其可信度和安全性的担忧 [17]。数据中毒 [18-21]、模型操纵 [22-25] 和后门 [26-27] 等官方攻击成为 XAI 方法的突出失败模式。同时，提出了诸如数据集中采样 [28， 29] 和模型正则化 [30-33] 等防御模式来抵消这些攻击。图 1 说明了由于数据和模型影响解释方法的输出而存在的常见故障模式。攻击者的主要目标是通过更改数据或模型来纵模型的解释，从而欺骗解释的接收者 [34]。它在敏感应用中成为真正的威胁，例如解释帮助法官做出保释决定的决策系统 [35]。自然，对解释的攻击根据所考虑的机器学习任务、模型架构和 XAI 方法的类别而有所不同，而 XAI 方法没有很好地系统化。

图 2 显示了最常见的此类攻击机制，称为对抗示例 [36-40]，即对输入图像施加的轻微扰动会极大地改变模型预测的“狗”类的解释。事实证明，通过访问模型，可以使用微分来优化输入扰动，这些扰动会广泛影响解释，而不会影响模型的预测性能。在更高级的设置中，可以毒害用于模型训练的训练数据，以便仅纵一组预定义的输入，例如来自特定类。随着时间的推移，人们提出了不需要访问模型的替代优化方法来执行此类作 [41， 42]。

在图 2 中，我们还可以观察到一种简单的防御模式，它可以抵消对抗性示例。使用不同算法获得的解释的聚合表明不太容易受到这种输入作的影响 [43]。这是因为攻击者以单一解释方法为目标，并且他们的融合更加健壮。最有效的防御机制是修补解释方法 [28， 44] 还是改进模型算法，例如在训练中使用正则化 [30， 45] 存在公开争论。

相关调查。当然，存在广泛的文献调查解释方法 [2， 46]、对模型预测的对抗性攻击 [16， 47]，特别是 XAI 在 AdvML 中的应用 [48]。然而，社区缺乏对解释和公平性指标的对抗性攻击的研究的全面概述，这将允许为未来的工作奠定共同基础。值得注意的是，Mishra等[49]简要回顾了特征重要性和反事实解释的稳健性。持续跟踪 XAI 出现的故障模式并修补解释框架仍然很重要。

贡献。在这项调查中，我们旨在突出我们所谓的对抗性可解释人工智能 （AdvXAI） 中迅速兴起的跨领域研究。评估 XAI 中安全威胁的范围是在实际应用中成功采用解释方法的关键。我们贡献了关于模型解释（第 3 节）和针对这些攻击的防御机制（第 4 节）的对抗性攻击的当前知识状态的系统化。此外，我们还面临着与攻击机器学习公平性指标密切相关的工作（第 5 节）。对 50 多篇论文的全面概述使我们能够在 AdvXAI 中指定机遇、挑战和未来的研究方向（第 6 节）。

本综述作为一个平易近人的前景，以识别潜在的研究差距并确定未来的方向。我们首先纳入了自 Ghorbani 等人 [38] 以来发表在主要机器学习会议（ICML、ICLR、NeurIPS、AAAI）和期刊（AIj、NMI）上的可见论文。然后，我们广泛搜索了他们的引文网络，以查找在其他场所发表的与 AdvXAI 相关的论文。我们故意排除了大量主要关注解释评估而与对抗性情景无关的论文 [参见 50，51]。

**背景**

在这里，我们简要介绍了对机器学习预测的对抗性攻击方法（第 2.2 节），以及用于解释人工智能模型的方法（第 2.3 节）。熟悉这些基本概念的读者可以跳到第 3 节。

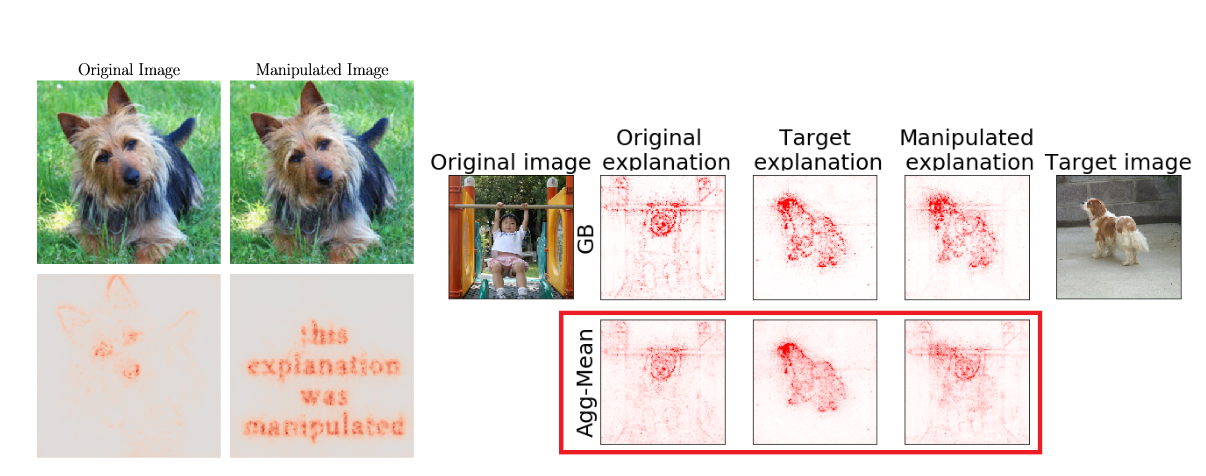
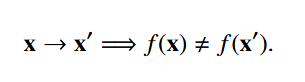


图 2：对抗性示例是对图像分类器预测的局部解释的最常见攻击。左 [改编自 40]：原始图像被归类为“狗”，其解释指出了影响此决定的特征。图像可以通过人眼不可见的扰动进行对抗性更改。纵的图像仍然被归类为“狗”，但它的解释发生了巨大变化。右 [改编自 43]：可以通过集成多种算法产生的解释来防御对抗性示例。在此示例中，攻击者旨在更改原始图像的引导式反向传播解释，使其看起来与目标图像的解释相似;而不会影响模型的预测。虽然成功纵了引导式反向传播解释，但三种替代解释方法的输出的聚合平均值仍然更能抵抗对抗性攻击。

2.1符号

2.2对抗性机器学习

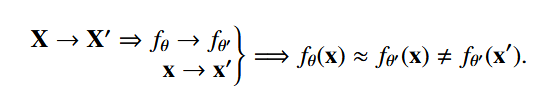
自 2004 年以来，对抗性方法在（深度）机器学习中取得了巨大的发展 [52]。通常，这些算法旨在攻击模型的行为，这与 AI 系统的安全性、安全性和稳健性有关。机器学习中探索最多的一类攻击是对抗性示例 [36， 37， 53]，尤其是在计算机视觉任务中，其目标是对输入数据进行最低限度的修改，以便欺骗模型进行错误分类。我们使用引入的符号来描述这种攻击策略，方式如下：



对抗性示例是与愚弄示例（随机输入）的正交概念，例如人类无法识别的图像，模型以非常高的置信度预测这些示例[55]。Brown等[56]提出了一种方法来创建对抗性补丁，当添加到输入中时，会触发错误分类。Athalye等[57]连续证明了物理世界中存在3D对抗对象，当从多个拍摄时，这些对象会被错误分类。Su et al. [58] 在他们的开创性论文中表明，通过仅扰动图像中的单个像素，可以欺骗模型进行错误分类。虽然成像数据可能是最具说明性的，但对抗性示例最初是针对预测表格和文本数据的机器学习模型进行研究的[36,59]。文本输入的对抗性扰动包括拼写错误、标点符号错误以及交换单词或字符 [60]。

使用对抗性输入纵模型预测的威胁增加了在现实世界中部署和应用 AI 系统的风险。旨在对抗攻击的算法成为防御，包括数据增强 [16]、模型正则化 [61] 和蒸馏 [62]。他们的最终目标是使攻击者更难欺骗模型的行为，即最大限度地降低潜在风险。通常，检测攻击是成功防御的初步（且充分）步骤 [63]。

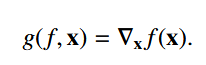
机器学习模型的另一种被广泛研究的失败模式是后门攻击 [64， 65]，其中（通常）人们假设对手可以访问模型，例如在外包训练的情况下。其目标是使模型对具有对手已知的特定触发模式的新输入进行错误预测。创建后门的最常见方法是以某种方式毒害用于训练的数据，以便对抗模型与所需模型无法区分：



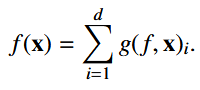
更广泛地说，已经提出了针对不同对抗目标的各种针对机器学习模型的中毒攻击 [66]，例如降低分类准确性 [67]。Cina' et al. [47] 介绍了与模型预测相关的中毒攻击（和防御）的全面系统化。攻击根据攻击者的目标进行分类，例如违反可用性或完整性;知识，例如拥有对系统的完全访问权限或将其视为黑匣子;能力，例如扰动数据的可能方式;最后是攻击者的策略，例如扰乱训练或测试数据。防御措施根据它们是清理数据还是修改学习算法/模型进行分类。这种对抗性攻击的系统化有助于进一步研究 AI 的不安全性。

2.3 可解释的人工智能

在研究 AdvML 的同时，人们对使我们能够理解和解释机器学习模型的方法越来越感兴趣 [68， 69]。XAI（或解释）方法在各种应用中都很有用，如监管审计 [70]、模型调试 [71] 或科学发现 [12]。可解释性的起源与人工智能系统有关，这些系统应该为其决策提供推理 [72， 73]，通常人类可以理解。目前，关于开发的 XAI 方法是否足以获得可信的 AI 系统，存在公开讨论 [74， 75]。例如，考虑一个（可微）模型输出相对于输入的简单梯度，作为特征影响的直观解释 [4， 76]：

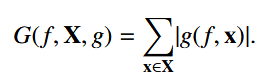


机器学习中探索最多的一类解释是局部事后特征归因 [5–7， 77]，它表明每个输入特征对模型结果预测的贡献程度：



获取向量 g（ f，x） 的最先进的方法是估计 Shapley 值 [7， 77， 78] 的算法，这是一种起源于博弈论的解决方案。如今，考虑到不同的数据模态、模型算法、解释质量指标或所需的计算效率，存在大量变化[79]。特别是对于表格数据，基于扰动的解释方法 [5， 7] 允许以与模型无关的方式解释单个预测。相反，基于梯度的 [6， 80] 和基于传播的 [81] 局部事后解释更特定于深度神经网络，深度神经网络是在非结构化数据（例如图像和文本）上训练的最先进的预测模型。

与局部解释相辅相成的是全局解释，这些解释总结了模型预测中的模式，在数据分布中保持一致。全局特征重要性 [82–84] 指标量化了模型对特定特征的依赖。例如，输入子集的特征归因聚合成为一个自然的重要性指标 [85]：



特征效应 [3， 84， 86] 解释，例如部分依赖图 [3]，可视化了预期模型的预测与特定特征的值之间的全局关系。特定于深度神经网络的全局解释包括基于概念的解释 [8]，它将人类可理解的概念与预测的类别联系起来，例如，对 “斑马 ”的预测对图像中条纹的敏感程度。

请注意，还有其他各种方法可以解释机器学习模型 [参见例如 1、2、46]，我们提到了本文调查的对抗性攻击最针对的解释方法的非详尽子集。此外，与 AdvXAI 密切相关的是评估解释的稳健性 [49] 和在对抗场景中应用解释方法 [48] 的研究。

1. **对模型解释的对抗性攻击**

据我们所知，Ghorbani et al. [38] 是 mention2 的第一个贡献，并提出了一种针对解释方法的对抗性攻击，特别是（卷积）神经网络中基于梯度的特征归因 [4， 6]。攻击背后的直觉是在解释变化的梯度方向上迭代扰动输入：x\_i+1 = x\_i + α ·sign( ∇\_xg（ f， x\_i） 。给定多次迭代和步长 α，该算法会产生相似的输入 x′，但可能具有截然不同的解释 g（ f， x′）。此优化过程受以下约束：扰动保持较小 ∥x − x′∥∞ ≤ ε并且两个输入具有相同的预测类 f （x） = f （x′）。类似的策略被用来攻击另一类解释方法——影响函数[87]，它返回训练输入对解释输入损失的重要性。尽管对图像数据的实验凸显了解释方法的脆弱性，但这从根本上说是因为神经网络模型本身容易受到此类攻击[38]。