



量子人工智能中的对抗学习

沈培鑫 蒋文杰 李炜康 鲁智德 邓东灵

Adversarial learning in quantum artificial intelligence

Shen Pei-Xin Jiang Wen-Jie Li Wei-Kang Lu Zhi-De Deng Dong-Ling

引用信息 Citation: *Acta Physica Sinica*, 70, 140302 (2021) DOI: 10.7498/aps.70.20210789

在线阅读 View online: <https://doi.org/10.7498/aps.70.20210789>

当期内容 View table of contents: <http://wulixb.iphy.ac.cn>

您可能感兴趣的其他文章

Articles you may be interested in

利用超导量子电路模拟拓扑量子材料

Topological quantum material simulated with superconducting quantum circuits

物理学报. 2018, 67(22): 220302 <https://doi.org/10.7498/aps.67.20181857>

基于人工神经网络在线学习方法优化磁屏蔽特性参数

Online learning method based on artificial neural network to optimize magnetic shielding characteristic parameters

物理学报. 2019, 68(13): 130701 <https://doi.org/10.7498/aps.68.20190234>

人工带隙材料的拓扑性质

Topological properties of artificial bandgap materials

物理学报. 2017, 66(22): 224203 <https://doi.org/10.7498/aps.66.224203>

HgTe/CdTe量子阱中自旋拓扑态的退相干效应

Dephasing effect of quantum spin topological states in HgTe/CdTe quantum well

物理学报. 2019, 68(22): 227301 <https://doi.org/10.7498/aps.68.20191072>

一维扩展量子罗盘模型的拓扑序和量子相变

Topological orders and quantum phase transitions in a one-dimensional extended quantum compass model

物理学报. 2018, 67(19): 190301 <https://doi.org/10.7498/aps.67.20180855>

马约拉纳零能模的非阿贝尔统计及其在拓扑量子计算的应用

Non-abelian statistics of Majorana modes and the applications to topological quantum computation

物理学报. 2020, 69(11): 110302 <https://doi.org/10.7498/aps.69.20200812>

专题: 机器学习与物理

量子人工智能中的对抗学习*

沈培鑫¹⁾ 蒋文杰¹⁾ 李炜康¹⁾ 鲁智德¹⁾ 邓东灵^{1)2)†}¹⁾ (清华大学交叉信息研究院, 北京 100084)²⁾ (上海期智研究院, 上海 200232)

(2021 年 4 月 25 日收到; 2021 年 5 月 26 日收到修改稿)

量子人工智能是一个探究人工智能与量子物理交叉的领域: 一方面人工智能的方法和技术可以用来解决量子科学中的问题; 另一方面, 量子计算的发展也可能为人工智能, 尤其是机器学习, 提供新的范式, 极大促进人工智能的发展. 然而, 量子机器学习和经典学习系统对于对抗样本同样具有脆弱性: 在原始数据样本上添加精心制作的微小扰动将很可能导致系统做出错误的预测. 本文介绍经典与量子对抗机器学习的基本概念、原理、以及最新进展. 首先从经典和量子两个方面介绍对抗学习, 通过二维经典伊辛模型和三维手征拓扑绝缘体的对抗样本揭示出经典机器学习在识别物质相时的脆弱性, 同时利用手写字体的对抗样本直观展示出量子分类器的脆弱性. 随后从理论层面上分别阐述经典与量子的“没有免费午餐”定理, 并探讨了量子分类器的普适对抗样本. 最后, 分析并讨论了相应的防御策略. 量子人工智能中对抗学习的研究揭示了量子智能系统潜在的风险以及可能的防御策略, 将对未来量子技术与人工智能的交叉产生深刻影响.

关键词: 量子人工智能, 量子对抗学习, 量子分类器, 拓扑物质相**PACS:** 03.67.-a, 03.67.Ac, 05.30.Rt, 07.05.Mh**DOI:** 10.7498/aps.70.20210789

1 引言

在过去的十年里, 无论是从图像识别^[1]到自然语言处理^[2], 还是从医学诊断^[3]到自动驾驶^[4], 人工智能领域都取得了巨大的成功, 引发了现代社会诸多领域的技术革命^[5,6]. 特别地, 基于人工神经网络的 AlphaGo^[7,8]与 AlphaFold^[9]分别在围棋和预测蛋白质结构方面取得了里程碑式的突破. 人工智能主要有三条发展路线: 符号主义、连接主义与行为主义^[10]. 基于人工神经网络的机器学习属于连接主义, 它是实现人工智能的一个重要途径, 近年来发展非常迅速^[11]. 与此同时, 近期实验上展现的量子优越性也标志着量子计算领域^[12]取得了开拓性的进展^[13,14]. 这两个快速发展的领域的交叉融通催生了一个新的研究前沿——量子人工智能^[15-17].

一方面, 可以利用人工智能技术来解决量子科学中的难题, 例如量子多体问题^[18]、量子态层析^[19]、拓扑量子编译^[20]、无序材料中的结构转变^[21]、量子非定域性探测^[22]以及物质相分类^[23-33]等. 另一方面, 直接运行在量子计算机上的量子算法具有巨大的潜力, 可以增强、加速甚至革新^[34-41]某些经典人工智能算法. 其中具有代表性的算法包括 Harrow-Hassidim-Lloyd 算法^[34]、量子主成分分析^[35]、量子生成模型^[38-40]和量子支持向量机^[42]等. 毫无疑问, 人工智能和量子物理之间的相互融通将极大促进两个领域的发展. 当前, 量子物理与人工智能的交叉研究主要集中在与机器学习的交叉方面, 与人工智能另外两条路线 (即符号主义与行为主义) 之间的交叉研究还相对缺乏.

在经典人工智能领域, 近期有许多工作揭示出基于深度人工神经网络的分类器在对抗场景中的

* 国家自然科学基金 (批准号: 12075128)、清华大学启动经费 (批准号: 53330300320) 和上海期智研究院资助的课题.

† 通信作者. E-mail: dldeng@mail.tsinghua.edu.cn

脆弱性^[43-45]: 向原始数据添加精心制作的微量 (甚至是人眼无法察觉的) 噪声, 可能会导致分类器以非常高的置信度做出错误的预测. Szegedy 等^[46,47] 利用一个著名的例子直观地展示了深度学习的脆弱性: 在初始的熊猫图像添加了肉眼不可察觉的对抗噪声后, 分类器将其错误地标识为长臂猿, 且置信度大于 99% (图 1). 这些精心设计用来欺骗分类器的输入样本被称为对抗样本. 目前, 人们普遍认为对抗样本在经典机器学习中广泛存在——无论输入数据类型和神经网络的细节如何, 几乎所有学习模型都易遭受对抗攻击^[43-45]. 从理论计算机科学的角度来看, 经典分类器关于对抗微扰的脆弱性与“没有免费午餐”定理之间存在深刻的联系. 需要指出的是, 对抗学习中的对抗样本与生成对抗网络 (generative adversarial networks, GAN)^[48] 产生的样本存在本质区别. 前者目的在于攻击已经训练好的学习模型, 后者目的在于模拟目标数据集以生成新的数据样本.

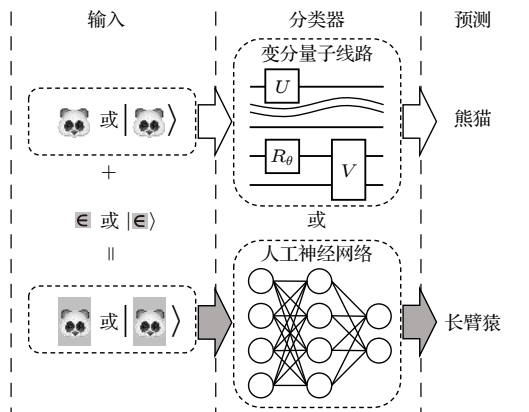


图 1 量子与经典对抗学习示意图 输入的原始熊猫图像样本可以编码为经典或量子数据, 分类器 (包含变量子线路或人工神经网络) 能够以非常高的准确率识别出熊猫; 但添加少量精心制作的噪声后, 同一分类器将以非常高的置信度把轻微修改过的熊猫图像错误分类为长臂猿

Fig. 1. A schematic illustration of quantum and classical adversarial learning. The image of a panda can be encoded as classical or quantum data. A classifier, which uses either variational quantum circuits or classical artificial neural networks, can successfully identify the image as a panda with the state-of-the-art accuracy. However, adding a small amount of carefully crafted noise will cause the same classifier to misclassify the slightly modified image into a gibbon with a notably high confidence.

随着嘈杂中型量子 (noisy intermediate-scale quantum, NISQ) 时代的到来, 量子计算有望在近期的科研应用领域中大放异彩^[49]. 然而, 尽管量子

计算机能够在某些方面展示出超越经典计算机的性能, 但在对抗学习中也会展现出脆弱性. 事实上, 量子分类器的脆弱性最近开始受到广泛关注^[50-52], 也因此衍生了一个全新研究方向——量子对抗机器学习. 文献^[53] 在理论上表明, 添加一个随着量子分类器比特数目指数减小的对抗微扰 (adversarial perturbation) 就足以影响学习模型的输出. 这表明量子分类器的稳健性 (robustness) 与高维希尔伯特 (Hilbert) 空间潜在的量子优势之间存在竞争关系: 展示量子优势需要高维的希尔伯特空间, 但量子分类器的稳健性随空间维度的增加而减弱. 最近文献^[50] 在量子机器学习的框架下探究了不同的对抗场景, 其诸多数值实例表明量子分类器同样很容易受到精心制作的对抗样本的攻击. 目前这个新兴的研究方向正在迅速增长, 吸引了越来越多来自不同领域研究者的关注, 但它仍处于起步阶段, 许多重要问题仍待探讨.

本文首先介绍近期经典对抗机器学习的研究进展和技术方法, 探讨其在识别物质相任务中得到的对抗样本实例. 随后介绍经典机器学习中的“没有免费午餐”定理及其在量子情况下的推广, 还回顾了对量子分类器脆弱性的相关研究. 有工作表明, 研究者可以生成一个通用的对抗样本来迷惑一组不同量子分类器, 同时也能够利用一个普适的对抗微扰将不同的初始样本全部变成对抗样本^[54]. 基于这些研究结果, 人们可以针对性地开发出实用的防御策略以对抗相应的攻击. 通过对抗训练, 可以使得分类器针对特定类型的对抗扰动的稳健性有显著提高. 最后, 对量子人工智能中的对抗学习做出总结与展望, 希望能给 NISQ 时代的量子机器学习提供有价值的指导.

2 经典对抗机器学习

经典对抗机器学习的早期探索可以追溯到垃圾邮件过滤 (spam filtering) 问题, 即垃圾邮件的发送方与抵制方的博弈. 一般来说, 用户的邮箱地址为外界得知后, 一些恶意的群体便可能为了商业利益等向这个邮箱发送广告邮件甚至电脑病毒. 为了抵御这种侵权行为, 人们开发了邮件过滤器以区分正常邮件与恶意邮件并对后者加以拦截. 这些早期的邮件过滤器可以看作是线性的分类器, 通过对邮件中的词汇与已采集到的恶意邮件的特征词汇

相对比来做出分类的判断. 而作为恶意邮件的发送方, 为了躲过邮件过滤器的检测, 便会采取一系列的手段, 如修改恶意邮件的特征词汇、增加正常词汇的比例等. 以上便是对抗机器学习中防御与攻击的例子. 需要说明的是, 在实际的对抗机器学习的运行中, 防御与攻击的发展往往是一个迭代的过程, 即攻击方与防御方都会根据对方技术的演变而做出相应的调整与改进, 以提高己方成功的概率.

在上述邮件过滤的对抗学习问题中, Dalvi 等^[55]以及 Lowd 和 Meek^[56]在此背景下研究了线性分类器对于对抗样本的脆弱性. 他们发现, 在不影响邮件包含的目的信息下只对邮件内容做小幅度精心设计的修改, 就可以顺利地通过过滤器的筛选. Barreno 等^[57]在 2006 年首先针对对抗机器学习做出了广泛的讨论, 并给出了不同的攻击分类. 此后对抗机器学习得到了全方位的发展, 研究内容包括攻击策略与攻击背景、对抗攻击的防御手段、以及机器学习的安全性评估等^[44,45,58,59]. 以下将对对抗机器学习做一个宏观的介绍, 以及从不同角度对对抗机器学习做出分类.

在对抗过程中, 攻击者能够获得的攻击目标的模型信息 (如训练数据、模型结构等) 在不同的情形下有不同等级. 根据对目标的信息由完全了解到完全不了解, 可以将攻击划分为白盒攻击 (white-box attack)、灰盒攻击 (gray-box attack) 和黑盒攻击 (black-box attack). 对于攻击者的能力来说, 我们根据其可以同时操纵训练集和测试集或只能操纵测试集将其分为毒药攻击 (poisoning attack) 和躲避攻击 (evasion attack). 前文提到的绕过垃圾邮件过滤器便是躲避攻击的一个例子, 即通过技术手段修改测试集并让所攻击的模型给出攻击者所期望的结果. 而通过与用户进行交互对话来改善表现的语音模型则属于毒药攻击中的一类, 比如某些恶意用户带有歧视性的言论将会使这个模型有潜在的风险. 此外, 可以将攻击的目标分为无差别攻击 (untargeted attack) 和针对性攻击 (targeted attack). 顾名思义, 无差别攻击是指攻击的目的为使分类器分类错误, 而不关注将样本判错成哪个错误标签. 针对性攻击则希望样本被错判成某个特定标签.

对于防御者而言, 一个直接的防御方式是根据攻击方已有的攻击手段来重新设计和训练机器学习模型, 从而使其具有抵御这一类攻击的能力. 对

于躲避攻击中提到的在某一类数据微扰得到的对抗样本, 可以将其连同原始的标签加入到训练集进行训练, 以此来增加模型面对攻击时的抵抗能力——这便是所谓的对抗训练 (adversarial training)^[60]. 除此之外, 一个更加理论化的策略是稳健优化 (robust optimization)^[61]. 这个方法的思想在于把防御的过程看成一个极小化极大问题 (minimax problem), 即通过扰动训练集以最大化损失函数, 并通过训练模型来最小化损失函数, 以此让重新训练的模型获得较好的防御能力 (详见第 5 节的讨论).

近年来深度学习得到了长足的发展, 其在人脸识别、自动驾驶等领域的应用受到了广泛的关注^[1,4]. 然而, 研究者们发现深度学习同样也存在着被对抗样本攻击的威胁^[46,48]. 在一个已经训练好的可以正确识别熊猫的深度学习模型中, 即使添加一个肉眼难以察觉的扰动, 也很可能会使这个模型给出的预测结果变为长臂猿 (图 1). 如果这类攻击没有得到解决而被恶意者利用, 将会使深度学习的实际应用中存在严重的安全隐患. 例如在自动驾驶汽车上, 如果前方一个停车告示牌被贴上一层精心设计的扰动薄膜, 被汽车的识别程序识别为常速行驶, 便可能引发安全事故^[44]. 因此深度学习中的对抗攻击和防御策略也受到了广泛的关注. 目前已经被提出的较为著名的攻击算法有快速梯度符号法^[62]、基本迭代法^[63]、动量迭代法^[64]、投影梯度下降法^[62]等, 这些算法通过精心设计的扰动来使模型无法给出正确的预测. 为了防御这些攻击, 同样有很多的策略被提出. 除了上文提到的稳健优化方法, 在深度学习中, 随机化和去噪往往也有着不错的效果. 这是由于深度学习有着较强的表达能力以及对随机噪声、去噪操作有着较好的稳健性, 而这些操作能够有效地消除对抗样本的扰动. 在对抗攻击与防御策略两者的快速发展和博弈中, 这些成果必将为机器学习的安全性和稳定性提供有益的指导.

3 机器学习物质相中的对抗样本

在凝聚态物理中, 识别不同的物质相并分辨出他们之间的相变点是一个重要且有趣的问题. 近年来, 机器学习在处理这类问题上取得了一系列令人瞩目的进展^[24-33,65-70], 其中包括监督学习和无监督

学习等方法. 与此同时, 一个亟待回答的问题是, 这些通过机器学习方法所获得的结论是否可靠? 它们能否抵御得住来自对抗样本的攻击? 针对这个问题, 文献 [71] 进行了一系列的探索, 其研究的内容包含经典二维伊辛 (Ising) 模型和三维手征拓扑绝缘体不同物质相的分类.

经典二维伊辛模型的哈密顿量可以表示为

$$H_{\text{Ising}} = -J \sum_{\langle ij \rangle} \sigma_i^z \sigma_j^z, \quad (1)$$

其中在 i 点的 z 方向自旋 $\sigma_i^z = \pm 1$, 最近邻自旋的耦合强度设为单位能量 ($J = 1$). 为了测试机器学习关于对抗样本的脆弱性, 采用一个全连接的前馈神经网络来训练由蒙特卡罗模拟生成的经典自旋构型数据, 其在测试集的准确率能够达到 97%. 在对抗攻击的阶段, 随机选取了一个铁磁相样本, 运用对抗攻击的算法对其迭代地增加扰动, 最终该模

型以 88% 的高置信度将这个受到扰动的样本错误地分类为顺磁相. 图 2 展示了一个更为极端的例子: 尽管修改前的未扰动数据 (图 2(a)) 与对抗样本 (图 2(b)) 只相差了一个自旋, 分类模型却给出了截然不同的预测结果.

对于拓扑相数据, 考虑一个三维手征拓扑绝缘体 [72,73]:

$$H_{\text{TI}} = \sum_{\mathbf{k} \in \text{BZ}} \Psi_{\mathbf{k}}^\dagger H_{\mathbf{k}} \Psi_{\mathbf{k}}, \quad (2)$$

其中 $\Psi_{\mathbf{k}}^\dagger = (c_{\mathbf{k},1}^\dagger, c_{\mathbf{k},0}^\dagger, c_{\mathbf{k},-1}^\dagger)$, $c_{\mathbf{k},\mu}^\dagger$ 表示在动量 $\mathbf{k} = (k_x, k_y, k_z)$ 处在 $\mu = -1, 0, 1$ 态上的费米产生算符, 求和遍历整个布里渊区 (Brillouin zone, BZ). 单体的动量空间的哈密顿量 $H_{\mathbf{k}} = \lambda_4 \sin k_x + \lambda_5 \sin k_y + \lambda_6 \sin k_z - \lambda_7 (\cos k_x + \cos k_y + \cos k_z + h)$, 其中 $\lambda_{4,5,6,7}$ 代表四个无迹的 Gell-Mann 矩阵, $0 < |h| < 1$ 、 $1 < |h| < 3$ 、 $|h| > 3$ 分别对应着三种不同的拓扑物

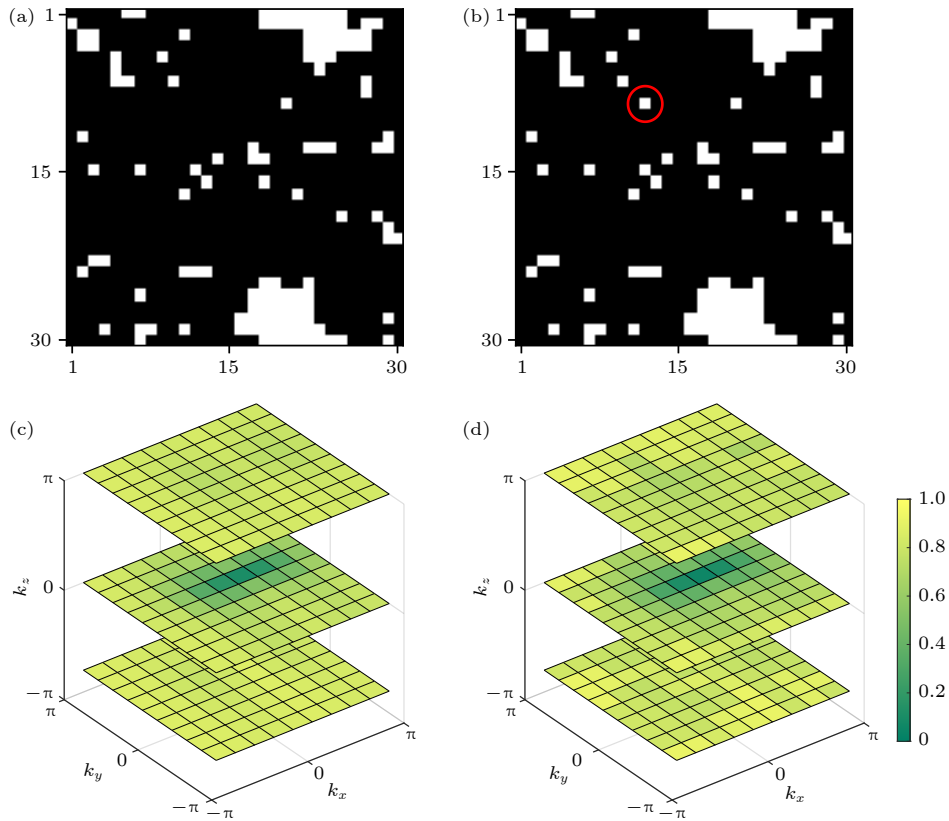


图 2 机器学习物质相中的对抗样本 (a) 一个原始的经典二维伊辛模型铁磁相的自旋构型; (b) 被分类器错误识别成顺磁相的对抗样本, 其相对于 (a) 只改变了一个自旋; (c) 一个原始的三维手征拓扑绝缘体的拓扑相样本; (d) 被分类器错误识别成其他相的对抗样本, 其相对于 (c) 只有肉眼难以识别的细微差别

Fig. 2. Adversarial examples in machine learning phases of matter: (a) A legitimate sample of the spin configuration in the ferromagnetic phase of the two-dimensional (2D) classical Ising model; (b) an adversarial example misclassified as the paramagnetic phase, which only differs from the original legitimate one shown in (a) by a single pixel; (c) a legitimate sample of the topological phase of three-dimensional (3D) chiral topological insulators; (d) an adversarial example misclassified as the other phase, which only differs from the original legitimate one shown in (c) by a tiny amount of noises that are imperceptible to human eyes.

质相^[72]. 首先使用一个三维卷积神经网络来进行初步的训练和分类, 随后采用动量迭代法生成对抗微扰. 与二维伊辛模型类似, 对抗样本在这个卷积神经网络模型以极高的概率被误分类成其他相. 为了在视觉上更加直观, 对处于拓扑相的图 2(c) 运用快速梯度符号法生成其相应的对抗样本 (图 2(d)). 尽管二者通过肉眼观察难以分辨, 但神经网络却会把添加扰动的对抗样本识别成其他相.

通过这些数值实验, 可以观察到在识别物质相及其相变的问题上, 机器学习在对抗攻击面前有其脆弱性, 为此需要制定相应的防御策略. 特别地, 本文以三维卷积神经网络训练拓扑相数据的例子来演示对抗训练方法. 在通过快速梯度符号法和投影梯度下降法等方式获得对抗样本后, 将其与原始数据一起作为训练集重新训练神经网络, 所获得的新模型在对抗样本上也具有很高的识别准确率. 需要指出的是, 上述的防御手段具有一定的局限性: 对抗训练后的新模型一般只能抵御特定的攻击方法 (训练用的对抗样本由此攻击方法产生), 当面对黑盒攻击或其他类型的对抗攻击时, 需要根据具体的情形再次重新训练网络. 以上工作揭示了机器学习应用于物质相识别时潜在的脆弱性, 同时对机器学习在其他物理学问题上的应用也提供了有益参考.

4 量子对抗机器学习

4.1 “没有免费午餐”定理

基于人工智能的技术早已应用在各个领域, 从日常生活到科学探究, 从休闲娱乐到工业生产, 逐渐深入人类社会的各个方面^[10]. 然而在诸如金融、国防、医疗等对安全性有着严格要求的领域, 对抗机器学习的发展使得人们难以忽略其背后潜藏的风险^[74]. 正如前文提到, 在经典人工智能领域中有许多攻击算法可以用于构造对抗样本, 这使得精心训练的人工智能模型面临重大的安全挑战^[75]. 从理论上阐明对抗学习和对抗样本是一件充满挑战但却意义重大的事情. 对这一问题的研究不仅可以帮助我们明确经典人工智能算法面临的安全性问题, 更可以指引我们理解量子人工智能算法的相关性质^[15–17]. 我们已经见证了中等规模量子信息处理技术的突破: 谷歌公司 53 个超导量子比特的“悬铃木”量子处理器被成功应用于随机线路取样^[13];

我国 76 个光子的量子计算原型机“九章”在 200 s 内成功获得 5000 万个样本的高斯玻色取样^[14], 均标志着量子技术取得突破性进展. 将量子计算技术应用于人工智能领域是一个前沿交叉领域, 近年来正蓬勃发展^[17]. 从理论上剖析量子人工智能的原理不仅可以帮助研究人员挖掘其潜力, 更能明确其限制所在, 这是当下此领域最为核心的研究内容之一. 量子对抗学习就是将对抗学习的思想和技术应用在量子人工智能领域, 这可以帮助研究人员更好地理解其内在的性质及可能存在的限制.

很多人工智能算法的实质是通过有限的数据拟合未知的函数, 以期对没有学习过的数据也能做出精准的预测^[76]. 在实际应用中, 人工智能算法可以用于处理多种多样的问题, 比如在机器翻译问题中利用机器学习寻找两种语言之间的映射, 从而使得计算机可以自动地将一种语言翻译成为另一种语言^[77]. 未知的函数可能是非常复杂的, 比如上面提到的机器翻译问题中, 语言所包含的信息量非常庞大, 语言之间的映射是非常复杂且一般难以直接计算的^[78], 因此需要使用复杂的学习模型从大量数据中寻找可能的规律. Wolpert 与 Macready^[79]早在 1997 年就提出, 在没有对未知函数做出限制的情况下, 所有的学习算法都是无法区分优劣的, 这是“没有免费午餐”定理的体现. 因此, 对于待解决问题的先验知识决定了对学习算法的选择. 另一方面, 在训练学习模型的时候, 还需要大量的数据以提高学习模型预测新数据的准确率. 一般情况下, 更大的数据规模意味着更准确的模型, 而数据规模的限制往往会使得学习模型预测准确率降低. 事实上, 数据规模与模型准确率的关系也可以通过“没有免费午餐”定理来定量刻画^[80]. 假设未知函数 f 是从有限集合 \mathcal{X} 到有限集合 \mathcal{Y} 的映射, S 是包含 N 个数据的训练集, h_S 是学习算法基于已知的数据集 S 学习得到的模型, $R_f(h_S)$ 为模型预测错误的概率, 则有如下关系:

$$\mathbb{E}_f [\mathbb{E}_S [R_f(h_S)]] \geq \left(1 - \frac{1}{|\mathcal{Y}|}\right) \left(1 - \frac{N}{|\mathcal{X}|}\right), \quad (3)$$

其中 $|\mathcal{X}|$, $|\mathcal{Y}|$ 分别代表集合元素的个数. (3) 式的直观理解为: 训练数据集包含有关未知函数的知识, 学习算法的核心是从中发现并学习其中的知识, 从而应用到新的数据预测. 因此, 如果训练数据过于稀疏, 其中包含的信息非常稀少, 学习算法一般是无法学习到足够的知识的. 需要指出的是, 上述结

论是针对最一般情况的平均结果. 具体到某些特殊的问题, 基于先验知识仍可能通过较少的数据量做出较为准确的预测.

量子机器学习使用量子数据作为学习算法的输入与输出. 量子数据一般表示为量子比特的物理状态, 因此全部的量子数据构成一个维度指数增长的希尔伯特空间. 量子学习算法的目标是从已知的数据集中学习从输入空间到输出空间的未知么正映射. 经典人工智能的学习效果用学习模型预测错误的概率来表示, 在量子机器学习中, 可以使用学习得到的么正映射与未知的目标映射之间的距离 $R_U(V_S)$ 来表征学习效果. 与经典“没有免费午餐”定理相对应的量子“没有免费午餐”定理有如下表述 [81]:

$$\mathbb{E}_U[\mathbb{E}_S[R_U(V_S)]] \geq 1 - \frac{1}{d(d+1)}(N^2 + d + 1), \quad (4)$$

其中 U 是未知的目标映射, S 是包含 N 个量子数据的训练集, V_S 是基于数据集 S 得到的学习结果, d 是输入空间的维度. 与经典结论相比, 可以看到有限数据集的大小被希尔伯特空间的维度所替代. 在训练集规模与空间指数维度相当的时候, 一般可以获得比较好的学习效果. 同样需要指出的是, 这个结论是针对全部可能么正映射而言, 对于某些特殊的么正映射, 通过较少的数据量做出较为准确的预测是可能实现的.

“没有免费午餐”定理给出了训练集规模与学习效果之间的关系. 在应用学习算法处理实际问题的时候, 由于问题本身的复杂性, 往往只能获得一部分的数据. 另一方面, 即使获得了大量的数据, 由于现行计算能力的限制, 学习模型的参数规模往往受到一定限制. 因此在大部分情况下, 只能获得对于目标问题的近似解, 这就意味着预测一般难以做到完全准确, 这是对抗攻击能够成功的重要原因.

4.2 量子分类器的脆弱性

分类任务是一类常见的人工智能任务 [10], 日常生活的人脸识别 [82]、自动驾驶 [4] 中的信号检测、物理研究中的物质相识别 [24-33, 65-70] 等问题, 其内核都是分类任务. 在量子人工智能领域, 分类任务同样广泛存在 [17]. 不幸的是, 无论是经典分类器还是量子分类器, 在面对精心设计的对抗样本时, 其预测准确率都会显著下降 [50, 60]. 即使对抗样本与原

来的样本相差无几, 甚至仅仅改变图片的一个像素值, 都会使得分类器分类错误——这就是分类器的脆弱性.

在前文中已经看到, 在使用学习模型识别物质相的任务中 [71], 可以使用多种攻击算法, 让原本分类正确的样本在经过微小的扰动之后被错误识别. 并且随着扰动强度的增加, 能够被正确识别的样本数会迅速下降. 实际上, 这一现象是广泛存在的, 这是“没有免费午餐”定理与高维数据独特的统计性质相结合的结果. 学习任务的全部数据的集合构成一个概率空间, 任意数据出现概率即是其在这个空间中的体积. 因此, 全部数据组成的集合是以概率为测度的测度空间. 同时还需要衡量数据受扰动的强度, 因此我们引入距离函数来衡量扰动前后数据之间的差别. 装备了距离函数的高维测度空间有一个非常重要的现象叫做测度集中 [83]. 物理上有许多这样的例子, 比如对于平衡状态下的自由粒子集合. 经典统计力学告诉我们, 这些系统的状态总可以使用一些简单的宏观量来描述, 这些宏观量就是系统微观状态的平均. 进一步的计算可以证明, 随着系统粒子数的增加, 平衡状态下系统偏离这些统计平均值的概率迅速下降, 即是在相空间中, 系统微观状态的概率展现了集中的性质 [84].

在经典人工智能领域, “没有免费午餐”定理告诉我们, 在实际情况下难以得到预测完全正确的学习结果. 因此在数据组成的测度空间中, 总是能够找到被分类错误的数据集合 B . 假设数据集所在的空间具备概率集中性质, 大部分数据点存在于错误分类的集合 B 附近, 从而微小的扰动就可以使得原本分类正确的样本得到错误的分类结果. 使用 k 代表分类的类别, h 代表学习算法给出的映射, acc_ϵ 代表在强度 ϵ 的扰动下分类器的准确率, 那么可以得到如下关系 [85]:

$$acc_\epsilon(h|k) \leq \min \left(acc(h|k), e^{-\frac{1}{2\sigma_k^2}(\epsilon - \epsilon_0)^2} \right), \quad (5)$$

其中 σ_k 是与数据空间几何性质相关的常数, ϵ_0 是与分类器准确率相关的常数. 可以看到, 在存在扰动的情况下, 分类器的准确率随扰动强度的增加而指数下降. 这一结论在使用卷积神经网络识别 MNIST 数据集的实验中得到了验证 [85].

在量子人工智能领域, 可以通过在 $SU(d)$ 中选择么正变换作用在固定初始态以制备量子数据,

因此可以将对量子态的分类转换为对么正变换的分类, 其中 d 是对应希尔伯特空间的维度. 装备了 Haar 测度和 Hilbert-Schmidt 距离的 $SU(d)$ 是一个具有测度集中性质的空间, 因此有类似的关系 [53]:

$$acc_{\epsilon} \leq \frac{2}{acc} e^{-\frac{d}{4}\epsilon^2}, \quad (6)$$

其中 ϵ 为对数据扰动的强度, d 为希尔伯特空间维度, acc 为分类器的准确率, acc_{ϵ} 为允许强度为 ϵ 的扰动下分类器的准确率. 可以发现, 量子分类器的准确率同样随扰动强度的增加而指数下降.

分类器的脆弱性是一个内禀的性质, 不依赖于学习模型或算法的具体细节. “没有免费午餐”定理表明, 在很多情况下会得到问题的近似解, 总是存在一部分数据被分类错误, 这是其脆弱性的根本原因. 人工智能算法一般用于处理复杂的高维数据, 而高维数据所表现出的测度集中的性质是分类器准确率急剧下降的重要原因. 大部分数据分布在错误分类的数据附近, 一些微小的扰动就可以使其分类错误. 这两个原因使得分类器的脆弱性成为一个难以避免的问题.

4.3 量子分类器的对抗样本

接下来介绍如何在不同模式下生成攻击量子分类器的对抗样本. 根据攻击者对量子分类器和学习算法的掌握情况, 分为白盒攻击和黑盒攻击两个模式. 白盒攻击模式假设攻击者掌握量子分类器和学习算法的全部信息. 在白盒攻击模式下, 探讨针对性攻击和无差别攻击两种情况. 在黑盒攻击模式下, 攻击者掌握极少 (甚至没有) 关于量子分类器和训练算法的信息.

4.3.1 白盒攻击模式: 无差别攻击

无差别攻击以分类器识错样本为目的, 在多分类问题中, 攻击者并不在意将样本判错成哪个错误标签. 在经典对抗学习领域有一个著名的例子: 攻击者戴上一个精心设计的眼镜便可以迷惑面部识别系统从而假装成其他人 [86]. 考虑到在白盒攻击模式下, 攻击者掌握量子分类器的结构和学习算法, 他们可以先用某个数据集先训练量子分类器, 使得该量子分类器达到很高的准确率, 然后固定分类器里已训练好的量子门参数, 将施加在样本数据上的微扰视为优化参数, 希望求得最优的扰动函

数, 使得量子分类器最大概率识错该数据集上的对抗样本, 即最大化如下损失函数:

$$U_{\delta} \equiv \operatorname{argmax}_{U_{\delta} \in \Delta} L(h(U_{\delta}|\psi)_{\text{in}}; \Theta^*), a), \quad (7)$$

其中 Θ^* 是量子分类器中已训练好的参数, $|\psi\rangle_{\text{in}}$ 是输入数据样本, U_{δ} 是施加在 $|\psi\rangle_{\text{in}}$ 上的、限制在 Δ 以内的对抗微扰, a 是样本对应的正确标签, h 是分类器基于已知的数据集学习得到的模型, L 是常用的交叉熵损失函数. 图 3(a) 展示了在手写字体 MNIST 数据集上的无差别攻击结果. 可以发现, 在所有原始样本上添加很小的、肉眼几乎不可分辨的微扰, 就能以很大概率获得被分类器识别错的对抗样本. 并且当对抗样本数据和原始数据之间相似度降低到 73% 时, 所有的对抗样本都会被识别错, 这说明了量子分类器的脆弱性 [50].

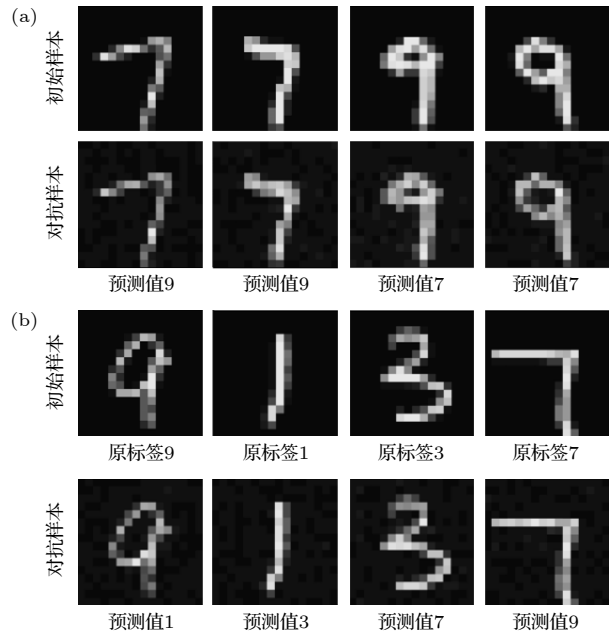


图 3 量子分类器在识别 MNIST 中手写字体图片时的对抗样本 (a) 经过无差别攻击, 量子分类器以极高置信度将数字 7, 9 分别识别成 9, 7, 即使对抗样本和初始样本的差别非常微小; (b) 通过针对性攻击, 量子分类器将把对抗样本预测为给定错误标签, 尽管对抗样本和初始样本相差无几

Fig. 3. Adversarial examples in quantum learning of MNIST hand-written images: (a) After untargeted attacks, the quantum classifier will misclassify the images of digit 7 (9) as digit 9 (7) with notably high confidence, although the differences between the adversarial and legitimate images are tiny; (b) after targeted attack, the quantum classifier will misclassify the adversarial examples into the category with the targeted label, even though the adversarial and legitimate images only differ slightly from each other.

4.3.2 白盒攻击模式: 针对性攻击

针对性攻击目的是让分类器将带有某种标签的样本数据错判成攻击者期望的标签. 比如攻击者可以试图迷惑带有面部识别系统的电子产品并将他识别成该产品的实际拥有者, 从而取得该电子产品的控制权^[86]. 由于要针对性地让对抗样本被预测为某一个特定标签, 考虑最小化如下损失函数:

$$U_{\delta}^{(t)} \equiv \operatorname{argmin}_{U_{\delta}^{(t)} \in \Delta} L\left(h\left(U_{\delta}^{(t)}|\psi\right)_{\text{in}}; \Theta^*\right), a^{(t)}, \quad (8)$$

其中 $a^{(t)}$ 是攻击者期望的标签 (与初始正确的标签 a 不同), $U_{\delta}^{(t)}$ 是施加在 $|\psi\rangle_{\text{in}}$ 上的对抗微扰. 在图 3(b) 中, 从 MNIST 数据集中随机选取数字 1, 3, 7, 9 的原始样本各一个, 采用基本迭代法^[63] 最小化 (8) 式以得到相应的对抗样本. 尽管原始样本和相应的对抗样本之间的区别肉眼几乎无法察觉, 但是添加微扰之后, 数字 9, 1, 3, 7 将会被量子分类器分别识别成 1, 3, 7, 9. 这表明在针对性模式下, 量子分类器也是易受攻击的^[50].

4.3.3 黑盒攻击模式

在黑盒攻击模式下, 攻击者掌握极少 (甚至没有) 关于量子分类器和训练算法的信息, 这种情况在实际情况中更为普遍, 所以攻击者难以直接根据分类器结构特征来生成对抗样本. 但是由于对抗样本的可传递性 (能够欺骗某一个机器学习模型的对抗样本, 也会有一定的概率可以成功欺骗其他模型), 攻击者仍然能够在不掌握量子分类器的具体信息, 甚至没有量子资源的情况下, 产生能够欺骗量子分类器的对抗样本^[46,47,87].

文献 [50] 探究了对抗样本在经典机器学习模型之间, 以及经典机器学习模型和量子机器学习模型之间的可传递性. 作者先用 MNIST 数据集训练好两个经典分类器 (卷积神经网络和前馈神经网络), 再在白盒无差别攻击模式下, 采用不同的迭代方法分别生成它们的对抗样本, 最后检验训练好的量子分类器识别这些对抗样本的准确率. 结果表明, 尽管量子分类器和经典分类器的结构差异巨大, 但是基于经典机器学习模型所生成的对抗样本也可以用来有效攻击量子机器学习系统.

4.3.4 对抗微扰并不是随机噪声

以上揭示了量子机器学习系统在面对对抗微扰时普遍存在的脆弱性. 值得强调的是, 对抗微扰

并不是随机噪声, 而是精心设计的扰动. 文献 [50] 分别将随机噪声和对抗微扰施加在原始数据上, 对比它们对分类器识别准确率的影响. 结果表明, 当施加非关联退相干噪声的时候, 识别准确率将随着相似度 (原始样本和对抗样本之间的保真度) 的降低而线性减小^[50]. 这种准确率和稳健性之间的折衷关系在考虑完全未知的噪声时也有所体现^[51]. 但是在施加对抗微扰的情况下, 当相似度偏离 100% 时, 识别准确率将显著减少, 甚至当准确率为 0 的时候, 相似度还可以维持在较高的水平^[50]. 这表明了对抗微扰并不是随机噪声, 同时也体现了量子分类器针对随机噪声具有很好的稳健性.

4.4 对抗样本的普适性

上述讨论了对于某一个特定分类器, 在不同攻击模式下生成对抗样本的过程. 目前已经知道, 添加一个随着量子分类器比特数目 n 指数减小的对抗微扰就足以得到对抗样本, 即对抗微扰强度 ϵ 的下限随着量子线路规模的增大而指数降低^[53]. 文献 [54] 在此基础上将以上结论推广到 k 个分类器: $\epsilon \sim O(\ln k / 2^n)$ 的对抗微扰就足以生成一个以较大概率同时欺骗 k 个分类器的对抗样本. 这个结论根本源于高维希尔伯特空间中的测度集中现象^[88], 与量子分类器的具体结构、训练算法以及数据集无关.

此外, 文献 [54] 还证明了对于一个给定的量子分类器, 将一个普适的对抗微扰施加在 m 个不同的初始样本上, 则量子分类器的判错率至少以 $1 - \delta$ ($0 < \delta < 1$) 的概率限制在内禀错误率附近^[89]:

$$|R_E - \mu(\mathcal{E})| \leq \sqrt{\frac{1}{2m} \ln\left(\frac{2}{\delta}\right)}, \quad (9)$$

其中, R_E 是量子分类器的判错率, $\mu(\mathcal{E})$ 是给定量子分类器在 Haar 测度下无对抗攻击时的内禀错误率, \mathcal{E} 是误分类的数据集. (9) 式说明当 m 越大时, 量子分类器的判错率越接近于内禀错误率. 当 m 非常大时, 可以用内禀错误率来估算量子分类器的判错率.

进一步地, 文献 [54] 证明内禀错误率的期望值满足如下不等式^[80,81,90]:

$$\mathbb{E}_U [\mathbb{E}_S [\mu(\mathcal{E})]] \geq 1 - \frac{d'}{d(d+1)} (N^2 + d + 1), \quad (10)$$

其中 U 是未知的真实目标映射, S 是规模为 N 的

训练集, d 是输入空间的维度, d' 是输出标签的数目. 从 (10) 式可以看出, 内禀错误率的期望值随着输出标签数或训练集数目的增加而减小, 随着样本空间的维度增大而增大. 当 d 趋于无穷时, 内禀错误率会逼近 100%, 且与量子分类器的结构和训练算法无关.

综上所述, 尽管在所有可能的数据样本中添加单一的对抗微扰平均而言并不会使对抗风险增大. 但是对仅包含 m 个初始样本的有限集合而言, 单一的对抗微扰仍然可能会增加量子分类器的判错率. 文献 [54] 通过详实的数值实验证实这种普适对抗微扰的存在: 将其添加到一组不同的初始样本中, 便能以极大概率生成一组对抗样本同时欺骗某个特定的分类器.

5 防御

通过以上的讨论可以清楚地得到, 经典和量子的分类器都很容易受到对抗微扰的影响——对抗样本普遍存在. 这可能引发人们对量子学习系统可靠性的关切, 尤其在那些安全性至关重要的领域, 例如自动驾驶 [4] 和医疗诊断 [74] 等. 因此, 研究可能的防御策略以提高量子分类器的稳健性具有重要的理论与实际意义 [91].

实际上, 完全消除对抗攻击带来的风险是非常困难的. 首先, 很难为对抗学习过程创建一个精确的理论模型. 这是一个高度非线性且复杂的非凸优化过程, 目前缺乏适当的理论工具来分析它 [61]. 因此, 要从理论上分析并阐明一个特定的防御策略能否防御对抗攻击是极其困难的. 此外, 要达到防御对抗攻击的目的, 我们希望分类器能够对每种可能的输入产生正确的输出, 其数量通常随着问题的变大而指数增长. 原则上这要求学习模型的复杂度指数增长. 在大多数情况下, 机器学习模型只能在所有可能输入的一小部分样本中具有良好表现 [10,11,76].

尽管如此, 近年来在经典对抗机器学习领域, 人们提出了多种防御策略来减小对抗样本带来的影响, 其中包括对抗训练 [60]、梯度隐藏 [92]、量子噪声添加 [91]、防御性蒸馏 [93] 和防御-生成对抗网络 (defense-generative adversarial network, defense-GAN) [94] 等. 每种策略都有其自身的优点和缺点, 然而并没有一个策略能够应对所有类型的对抗攻

击. 本节探讨如何提高量子分类器关于对抗攻击的稳健性. 数值实验表明, 前文提到的对抗训练方法可以显著提高量子分类器在防御特定对抗攻击时的表现.

对抗训练的基本思想是通过将对抗样本注入训练集中来增强模型的稳健性. 这是一种非常简单且直接的方法: 给定多种攻击策略生成的对抗样本, 可以把这些对抗样本与初始样本结合成新的训练集, 并重新训练分类器. 特别地, 如果采用稳健优化 [61] 方法来训练量子分类器, 那么可以将任务归结于一个典型的极小化极大的优化问题:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^N \max_{U_{\delta} \in \Delta} L(h(U_{\delta}|\psi)_{\text{in}}^{(i)}; \Theta), y^{(i)}), \quad (11)$$

其中 $|\psi\rangle_{\text{in}}^{(i)}$ 是受到攻击的第 i 个样本, $y(i)$ 表示其原始对应标签. 重新训练量子分类器可以降低对抗攻击的风险, 该风险由输入样本最坏情况的平均损失函数来描述. 这种极小化极大问题在稳健优化领域已经得到广泛的研究, 存在诸多解决此类问题的方法. 其中一个卓有成效的方法是把上述表达式拆分为两部分: 外部最小化和内部最大化. 内部最大化与生成对抗样本的过程完全相同, 已经在上述章节进行了讨论. 外部最小化则归结为最小化对抗样本的损失函数, 可以运用前文提到的迭代算法实现.

值得一提的是, 虽然经过对抗训练的量子分类器针对某类对抗扰动的稳健性确实会得到显著提升, 但是一般只能对由相同攻击方法生成的对抗样本表现良好. 当攻击者采用其他攻击策略时, 分类器的防御性能可能急剧下降. 此外由于黑盒攻击的梯度掩蔽 (gradient masking), 对抗训练倾向于使量子分类器在防御白盒攻击上相比于黑盒攻击更加稳健 [87,92]. 实际上, 尽管一种防御策略可以抵抗针对量子分类器的一种攻击方式, 但是其将不可避免地会给出并了解其防御机制的攻击者开放另一种漏洞.

在经典对抗学习领域, 另一种新的防御机制最近受到了广泛关注, 它可以有效地应对白盒和黑盒攻击. 这便是上文提到的 defense-GAN [94]. 该机制不直接把输入样本导入分类器, 而是首先把样本导入 GAN 的生成器中重新生成输入数据, 然后再将其输入分类器. 该防御策略的核心在于利用 GAN 强大的表达能力减弱甚至过滤掉对抗微扰带来的影响. 最近 GAN 的量子版本 (QGAN) 已经在理

论上被提出^[39,95],同时 QGAN 的实验原型机也在超导量子电路上得以实现^[40,96].研发一个 defense-QGAN 来增强量子分类器关于对抗微扰的稳健性也是至关重要的,这将是未来一个有趣且充满前景的研究方向.

6 总结与展望

本文系统地回顾了经典和量子机器学习在不同情况下关于对抗样本的脆弱性.由于测度集中现象^[88],对抗样本存在的普遍性是高维空间中量子机器学习应用的基本特征.无论是经典还是量子的分类器,在原始数据中添加精心设计的细微扰动都可能导致分类器以非常高的置信度做出错误的预测.本文总结了针对多种不同的任务和分类器生成对抗样本的通用方法,并回顾了在不同对抗环境中的具体示例.通过对抗训练,研究人员发现分类器对于特定类型的对抗扰动的脆弱性可以得到显著抑制.

值得强调的是,本文所讨论的量子对抗学习与前面提到的 QGAN 存在本质区别^[39,40,43,95,97].QGAN 包含两个主要部分:生成器和判别器.它们通过对抗博弈的方式进行交替训练:在每个学习回合中,判别器都会优化其策略,以识别生成器产生的虚假数据,而生成器会进一步更新其伪造数据的机制来欺骗判别器.最终,这种动态博弈的训练过程将达到纳什(Nash)均衡.理想情况下,生成器在达到纳什均衡后生成的数据将与原始训练集中的真实数据满足相同的统计规律,而判别器将不能以大于一半的概率分辨出伪造的数据.因此,QGAN 的主要目标是生成匹配原始训练数据的统计信息的新数据(无论是经典的或量子的).与之不同的是,本文介绍的对抗学习主要侧重如何生成对抗样本以及如何防御对抗攻击.

本文仅揭示了对抗学习的冰山一角,该领域还有许多重要问题值得进一步研究.本文的总结主要集中在基于人工神经网络和变分量子线路的监督学习上,但是无监督学习和强化学习也可能遭受脆弱性问题的困扰^[59].因此,将对抗学习推广到其他机器学习类型或许是可行的.另外,我们猜想深度学习中的对抗微扰的普遍存在性与量子多体物理中的正交灾难现象(即添加任意弱的局部扰动后,量子系统的基态在热力学极限下与原始基态正

交)^[98,99]之间可能存在深远的联系.最后,设计并进行一个实验来展示对抗样本的普遍性和普适微扰的存在性也亟待研究.这将是未来量子技术在人工智能中实际应用的重要一步,尤其是在安全性至关重要的领域,例如无人驾驶汽车、恶意软件检测、生物识别和医疗诊断^[74]等.

感谢清华大学交叉信息研究院龚维元、蒋飏和马克斯-普朗克研究所陆思锐的有益讨论.

参考文献

- [1] Krizhevsky A, Sutskever I, Hinton G E 2012 *Proceedings of the 25th International Conference on Neural Information Processing Systems* (Volume 1) New York, USA, December 3–8, 2012 p1097
- [2] Hinton G, Deng L, Yu D, Dahl G E, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury B 2012 *IEEE Signal Process. Mag.* **29** 82
- [3] Kononenko I 2001 *Artif. Intell. Med.* **23** 89
- [4] Grigorescu S, Trasnea B, Cocias T, Macesanu G 2020 *J. Field Robot.* **37** 362
- [5] LeCun Y, Bengio Y, Hinton G 2015 *Nature* **521** 436
- [6] Jordan M, Mitchell T 2015 *Science* **349** 255
- [7] Silver D, Huang A, Maddison C J, et al. 2016 *Nature* **529** 484
- [8] Silver D, Schrittwieser J, Simonyan K, et al. 2017 *Nature* **550** 354
- [9] Senior A W, Evans R, Jumper J, et al. 2020 *Nature* **577** 706
- [10] Russell S, Norvig P 2020 *Artificial Intelligence: A Modern Approach* (Hoboken: Pearson) pp1–61
- [11] Bishop C 2006 *Pattern Recognition and Machine Learning* (New York: Springer-Verlag) pp225–284
- [12] Nielsen M A, Chuang I L 2010 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press) pp171–352
- [13] Arute F, Arya K, Babbush R, et al. 2019 *Nature* **574** 505
- [14] Zhong H S, Wang H, Deng Y H, Chen M C, Peng L C, Luo Y H, Qin J, Wu D, Ding X, Hu Y, Hu P, Yang X Y, Zhang W J, Li H, Li Y, Jiang X, Gan L, Yang G, You L, Wang Z, Li L, Liu N L, Lu C Y, Pan J W 2020 *Science* **370** 1460
- [15] Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S 2017 *Nature* **549** 195
- [16] Dunjko V, Briegel H J 2018 *Rep. Prog. Phys.* **81** 074001
- [17] Das Sarma S, Deng D L, Duan L M 2019 *Phys. Today* **72** 48
- [18] Carleo G, Troyer M 2017 *Science* **355** 602
- [19] Torlai G, Mazzola G, Carrasquilla J, Troyer M, Melko R, Carleo G 2018 *Nat. Phys.* **14** 447
- [20] Zhang Y H, Zheng P L, Zhang Y, Deng D L 2020 *Phys. Rev. Lett.* **125** 170501
- [21] Deringer V L, Bernstein N, Csányi G, Ben Mahmoud C, Ceriotti M, Wilson M, Drabold D A, Elliott S R 2021 *Nature* **589** 59
- [22] Deng D L 2018 *Phys. Rev. Lett.* **120** 240402
- [23] Deng D L, Li X, Das Sarma S 2017 *Phys. Rev. B* **96** 195145
- [24] Zhang Y, Kim E A 2017 *Phys. Rev. Lett.* **118** 216401
- [25] Carrasquilla J, Melko R G 2017 *Nat. Phys.* **13** 431
- [26] van Nieuwenburg E P L, Liu Y H, Huber S D 2017 *Nat. Phys.* **13** 435

- [27] Wang L 2016 *Phys. Rev. B* **94** 195105
- [28] Broecker P, Carrasquilla J, Melko R G, Trebst S 2017 *Sci. Rep.* **7** 8823
- [29] Ch'ng K, Carrasquilla J, Melko R G, Khatami E 2017 *Phys. Rev. X* **7** 031038
- [30] Wetzel S J 2017 *Phys. Rev. E* **96** 022140
- [31] Hu W, Singh R R P, Scalettar R T 2017 *Phys. Rev. E* **95** 062122
- [32] Zhang Y, Mesaros A, Fujita K, Edkins S D, Hamidian M H, Ch'ng K, Eisaki H, Uchida S, Davis J C S, Khatami E, Kim E-A 2019 *Nature* **570** 484
- [33] Lian W, Wang S T, Lu S, Huang Y, Wang F, Yuan X, Zhang W, Ouyang X, Wang X, Huang X, He L, Chang X, Deng D L, Duan L 2019 *Phys. Rev. Lett.* **122** 210503
- [34] Harrow A W, Hassidim A, Lloyd S 2009 *Phys. Rev. Lett.* **103** 150502
- [35] Lloyd S, Mohseni M, Rebentrost P 2014 *Nat. Phys.* **10** 631
- [36] Dunjko V, Taylor J M, Briegel H J 2016 *Phys. Rev. Lett.* **117** 130501
- [37] Amin M H, Andriyash E, Rolfe J, Kulchytskyy B, Melko R 2018 *Phys. Rev. X* **8** 021050
- [38] Gao X, Zhang Z Y, Duan L M 2018 *Sci. Adv.* **4** eaat9004
- [39] Lloyd S, Weedbrook C 2018 *Phys. Rev. Lett.* **121** 040502
- [40] Hu L, Wu S H, Cai W, Ma Y, Mu X, Xu Y, Wang H, Song Y, Deng D L, Zou C L, Sun L 2019 *Sci. Adv.* **5** eaav2761
- [41] Schuld M, Killoran N 2019 *Phys. Rev. Lett.* **122** 040504
- [42] Rebentrost P, Mohseni M, Lloyd S 2014 *Phys. Rev. Lett.* **113** 130503
- [43] Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D 2018 arXiv: 1810.00069 [cs, stat]
- [44] Biggio B, Roli F 2018 *Pattern Recognit.* **84** 317
- [45] Miller D J, Xiang Z, Kesidis G 2019 arXiv: 1904.06292 [cs, stat]
- [46] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R 2014 arXiv: 1312.6199 [cs]
- [47] Goodfellow I J, Shlens J, Szegedy C 2015 arXiv: 1412.6572 [cs, stat]
- [48] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y 2014 *Adv. Neural Inf. Process. Syst.* **27** 2672
- [49] Preskill J 2018 *Quantum* **2** 79
- [50] Lu S, Duan L M, Deng D L 2020 *Phys. Rev. Res.* **2** 033212
- [51] Guan J, Fang W, Ying M 2020 arXiv: 2008.07230 [quant-ph]
- [52] Wiebe N, Kumar R S S 2018 *New J. Phys.* **20** 123019
- [53] Liu N, Wittek P 2020 *Phys. Rev. A* **101** 062331
- [54] Gong W, Deng D L 2021 arXiv: 2102.07788 [cond-mat, physics: quant-ph]
- [55] Dalvi N, Domingos P, Mausam, Sanghai S, Verma D 2004 *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Seattle, USA, August 22–25, 2004 p99
- [56] Lowd D, Meek C 2005 *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* Chicago, USA, August 21–24, 2005 p641
- [57] Barreno M, Nelson B, Sears R, Joseph A D, Tygar J D 2006 *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security* Taipei, China, March 21–24, 2006 p16
- [58] Huang L, Joseph A D, Nelson B, Rubinstein B I P, Tygar J D 2011 *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence* Chicago, USA, October 21, 2011 p43
- [59] Vorobeychik Y, Kantarcioglu M 2018 *Synth. Lect. Artif. Intell. Mach. Learn.* **12** 1
- [60] Kurakin A, Goodfellow I, Bengio S 2017 arXiv: 1611.01236 [cs, stat]
- [61] Ben-Tal A, Ghaoui L E, Nemirovski A 2009 *Robust Optimization* (Princeton: Princeton University Press) pp1–146
- [62] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A 2018 *International Conference on Learning Representations* Vancouver, Canada, April 30–May 3, 2018 p1
- [63] Kurakin A, Goodfellow I, Bengio S 2017 arXiv: 1607.02533 [cs, stat]
- [64] Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, Li J 2018 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Salt Lake City, USA, June 18–23, 2018 p9185
- [65] Hsu Y T, Li X, Deng D L, Das Sarma S 2018 *Phys. Rev. Lett.* **121** 245701
- [66] Rodriguez-Nieva J F, Scheurer M S 2019 *Nat. Phys.* **15** 790
- [67] Zhang P, Shen H, Zhai H 2018 *Phys. Rev. Lett.* **120** 066401
- [68] Huembeli P, Dauphin A, Wittek P 2018 *Phys. Rev. B* **97** 134109
- [69] Rem B S, Käming N, Tarnowski M, Asteria L, Fläschner N, Becker C, Sengstock K, Weitenberg C 2019 *Nat. Phys.* **15** 917
- [70] Bohrdt A, Chiu C S, Ji G, Xu M, Greif D, Greiner M, Demler E, Grusdt F, Knap M 2019 *Nat. Phys.* **15** 921
- [71] Jiang S, Lu S, Deng D L 2019 arXiv: 1910.13453 [cond-mat, physics: quant-ph]
- [72] Neupert T, Santos L, Ryu S, Chamon C, Mudry C 2012 *Phys. Rev. B* **86** 035125
- [73] Deng D L, Wang S T, Duan L M 2014 *Phys. Rev. A* **90** 041601
- [74] Finlayson S G, Bowers J D, Ito J, Zittrain J L, Beam A L, Kohane I S 2019 *Science* **363** 1287
- [75] Ren K, Zheng T, Qin Z, Liu X 2020 *Engineering* **6** 346
- [76] Goodfellow I, Bengio Y, Courville A 2016 *Deep Learning* (Cambridge: The MIT Press) pp98–165
- [77] Sutskever I, Vinyals O, Le Q V 2014 *Proceedings of the 27th International Conference on Neural Information Processing Systems* (Volume 2) Montréal, Canada, December 8–13, 2014 p3104
- [78] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I 2017 arXiv: 1706.03762 [cs]
- [79] Wolpert D H, Macready W G 1997 *IEEE Trans. Evol. Comput.* **1** 67
- [80] Shalev-Shwartz S, Ben-David S 2014 *Understanding Machine Learning: From Theory to Algorithms* (New York: Cambridge University Press) pp36–41
- [81] Poland K, Beer K, Osborne T J 2020 arXiv: 2003.14103 [quant-ph]
- [82] Schroff F, Kalenichenko D, Philbin J 2015 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* Boston, USA, June 7–12, 2015 p815
- [83] Walters M 2015 arXiv: 1508.05448 [math-ph]
- [84] Schwabl F 2006 *Statistical Mechanics* (Berlin Heidelberg: Springer-Verlag) pp1–20
- [85] Dohmatob E 2019 arXiv: 1810.04065 [cs, stat]
- [86] Sharif M, Bhagavatula S, Bauer L, Reiter M K 2016 *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* Vienna, Austria, October 24–28, 2016 p1528
- [87] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z B, Swami A 2017 *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* Abu

- Dhabi, UAE, April 2–6 2017 p506
- [88] Ledoux M 2001 *The Concentration of Measure Phenomenon* (Providence: American Mathematical Society) pp1–21
- [89] Hoeffding W 1963 *J. Am. Stat. Assoc.* **58** 13
- [90] Sharma K, Cerezo M, Holmes Z, Cincio L, Sornborger A, Coles P J 2020 arXiv: 2007.04900 [quant-ph]
- [91] Du Y, Hsieh M H, Liu T, Tao D, Liu N 2020 arXiv: 2003.09416 [quant-ph]
- [92] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P 2017 arXiv: 1705.07204 [cs, stat]
- [93] Papernot N, McDaniel P, Wu X, Jha S, Swami A 2016 *37th IEEE Symposium on Security and Privacy San Jose, USA, May 23–25, 2016* p582
- [94] Samangouei P, Kabkab M, Chellappa R 2018 arXiv: 1805.06605 [cs, stat]
- [95] Dallaire-Demers P L, Killoran N 2018 *Phys. Rev. A* **98** 012324
- [96] Huang K, Wang Z A, Song C, Xu K, Li H, Wang Z, Guo Q, Song Z, Liu Z-B, Zheng D, Deng D L, Wang H, Tian J G, Fan H 2020 arXiv: 2009.12827 [quant-ph]
- [97] Zeng J, Wu Y, Liu J G, Wang L, Hu J 2019 *Phys. Rev. A* **99** 052306
- [98] Anderson P W 1967 *Phys. Rev. Lett.* **18** 1049
- [99] Deng D L, Pixley J H, Li X, Das Sarma S 2015 *Phys. Rev. B* **92** 220201

SPECIAL TOPIC—Machine learning and physics

Adversarial learning in quantum artificial intelligence^{*}

Shen Pei-Xin¹⁾ Jiang Wen-Jie¹⁾ Li Wei-Kang¹⁾

Lu Zhi-De¹⁾ Deng Dong-Ling^{1)2)†}

¹⁾ (Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China)

²⁾ (Shanghai Qi Zhi Institute, Shanghai 200232, China)

(Received 25 April 2021; revised manuscript received 26 May 2021)

Abstract

Quantum artificial intelligence exploits the interplay between artificial intelligence and quantum physics: on the one hand, a plethora of tools and ideas from artificial intelligence can be adopted to tackle intricate quantum problems; on the other hand, quantum computing could also bring unprecedented opportunities to enhance, speed up, or innovate artificial intelligence. Yet, quantum learning systems, similar to classical ones, may also suffer adversarial attacks: adding a tiny carefully-crafted perturbation to the legitimate input data would cause the systems to make incorrect predictions at a notably high confidence level. In this paper, we introduce the basic concepts and ideas of classical and quantum adversarial learning, as well as some recent advances along this line. First, we introduce the basics of both classical and quantum adversarial learning. Through concrete examples, involving classifications of phases of two-dimensional Ising model and three-dimensional chiral topological insulators, we reveal the vulnerability of classical machine learning phases of matter. In addition, we demonstrate the vulnerability of quantum classifiers with the example of classifying hand-written digit images. We theoretically elucidate the celebrated no free lunch theorem from the classical and quantum perspectives, and discuss the universality properties of adversarial attacks in quantum classifiers. Finally, we discuss the possible defense strategies. The study of adversarial learning in quantum artificial intelligence uncovers notable potential risks for quantum intelligence systems, which would have far-reaching consequences for the future interactions between the two areas.

Keywords: quantum artificial intelligence, quantum adversarial learning, quantum classifiers, topological phases of matter

PACS: 03.67.-a, 03.67.Ac, 05.30.Rt, 07.05.Mh

DOI: 10.7498/aps.70.20210789

^{*} Project supported by the National Natural Science Foundation of China (Grant No. 12075128), the Start-up Fund from Tsinghua University, China (Grant No. 53330300320), and the Shanghai Qi Zhi Institute, China.

[†] Corresponding author. E-mail: dldeng@mail.tsinghua.edu.cn