一场学习革命

(北京大学　赵鹏巍　编译自 Marric Stephens. *Physics World*, 2019, (3): 45)

机器学习的基础工作早在上世纪中叶已经奠定。然而，正如Marric Stephens所发现的，越来越强大的计算机以及过去十年间不断改进的算法，正推动着机器学习在从医学物理到材料科学等诸多方面应用的爆发。

当银行打电话提醒你的信用卡在陌生时间发生可疑大额消费时，这不大可能是经由一位亲自梳理过你个人帐户的好心银行员工发现的，相反，这更像是机器学会了判断与犯罪活动相关联的行为，并在你的账单中发现了一些意外情况。银行的计算机系统利用算法可以静静地、高效地监控你帐户上的盗刷迹象。

监控信用卡只是“机器学习”的一个例子。机器学习是指计算机系统通过在给定的示例集上训练，从而发展出灵活自主地执行任务的能力的过程。作为更广义的人工智能领域的一个子集，机器学习技术能够被应用于任意可挖掘输入输出之间关联的大型复杂数据集中。对于上面提到的银行，算法将分析大量合法以及非法交易数据，以便从给定输入(“凌晨3点产生的大额消费”)产生输出(“盗刷嫌疑”)。

机器学习不仅仅应用于金融领域，也正在被应用于医疗保健、交通运输、刑事司法以及科学研究等许多其他领域。

量子问题

机器学习对量子物理，特别是“量子多体问题”的求解可能产生深远的意义。这类问题产生于一组相互作用着的物体，且只能通过考虑其量子性质才能理解整个体系。美国纽约西蒙斯基金会Flatiron研究所的物理学家 Giuseppe Carleo 指出：“量子多体问题的共性在于研究体系的性质原则上需要完全了解体系的多体波函数这一事实”，而多体波函数，用Carleo的话说，“是一个怪物，其复杂程度随着体系组分数量的增加而呈现出指数级的增长”。例如，考虑一组由若干粒子组成的系统，每个粒子均有顺时针和逆时针两个自旋方向。对于两粒子系统，共有四种可能的状态，三粒子系统则有八种可能状态，这时体系仍然是容易处理的。但随着粒子数增多，复杂程度很快就会超出可控范围。

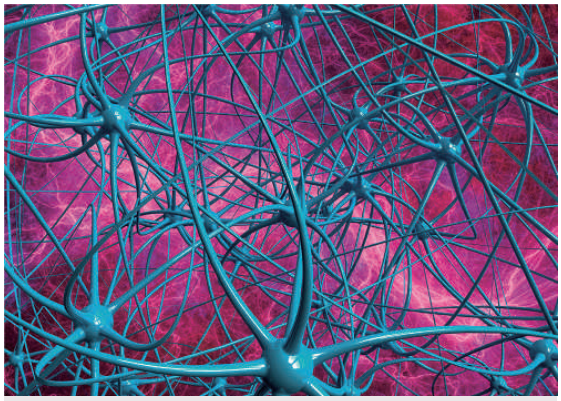

图1　新型学习。人工神经网络近似模拟真实生物大脑的工作模式，每个输入经由一层或多层隐藏的人工神经元处理

传统方法无法有效地处理具有一定组分数量的量子多体问题，所以Carleo及其瑞士苏黎世联邦理工学院的同事 Matthias Troyer 使用了机器学习方法。他们发现一个相对较“浅”的神经网络——仅使用单个隐藏层，已经可以有效地“学习”表征体系的波函数了，一维或二维晶格上的自旋问题便是一例。

与求解量子多体问题相同的困难也出现在“量子态层析成像”中。正如层析成像从外部测量来重建物体的内部结构一样，量子态层析成像通过对量子态易于测量的部分进行少量次数的测量来确定一个系统的量子状态。与量子多体问题一样，编码在波函数中的信息量随着系统中组分数量的增加而呈指数增长。

量子比特在量子计算机中的纠缠方式是一个值得描述的量子态，这使得量子态层析成像对于理解量子计算机应该如何应对噪声和退相干性至关重要。问题在于任何实用的量子计算机都将包含数十或数百个量子比特，故直接去确定其量子态的方法是不合适的。这正是人工神经网络可以发挥作用的地方，Carleo发现神经网络可以高效地重建含100个量子比特的量子计算机的量子态，而标准方法仅限于大约

8个量子比特。

机器学习的应用还远没有结束。机器学习方法仅是最近才被应用于量子物理领域，这意味着研究仍处于原理论证阶段。实际上，Carleo及其同事展示的方法通常涉及仅含一或两个隐藏层的神经网络，而更成熟的商业应用——如Google和Facebook等，能够利用更深层的架构，并在针对特定任务优化的专用硬件上运行。

不幸的是，量子物理的奇异特性使得这些更复杂的神经网络不能简单地直接移植到量子问题上，Carleo等人不得不几乎从头开始设计算法，目前还未达到与机器学习应用前沿相当的复杂度，而赶上那些成熟的算法，将允许人工神经网络解决更复杂的量子问题。Carleo指出："我认为未来几年会看到原理方法与技术实现之间的鸿沟越来越小，并带来我们现在甚至无法想象的应用。"

寻找新材料

虽然人工神经网络在给出有用的结果前通常必须有大量数据集作为输入，但美国弗吉尼亚大学的Prasanna Balachandran使用一些数据需求并不是很大的方法。他的研究目标是从巨大的多维可能性空间中找出相对较小的能制造有益材料的配方空间。通过试错法来寻找这样的空间将会花费太长时间，而且对应属性已知材料的区域也仅仅是全空间的极小一部分。

Balachandra用来解决这一问题的方法是一种特殊类型的机器学习，称为统计学习。这种方法通过假设数据特征遵循严格的统计规律来绕过机器学习对大量训练集的需求。他解释道："我们训练机器学习模型去掌握我们已经知道的东西，进而用这些模型来预测我们不知道的东西"。

在新材料研究中，我们知道某些材料组合的行为，而想要预测的是每个其他可能组合的性质。然而，预测给定材料性质的可信度依赖于对其邻近材料的了解情况，因此，对于每个预测，Balachandran还量化了与每个预测值相应的误差。

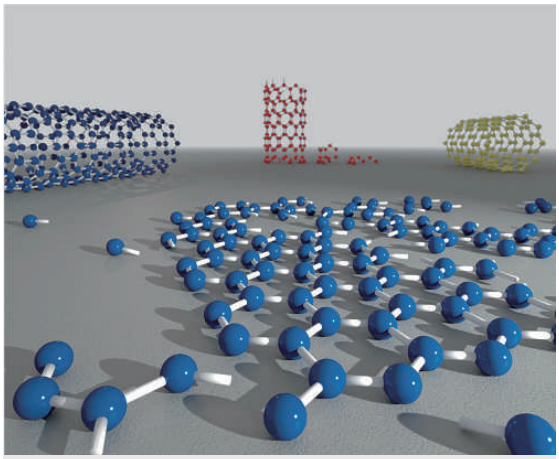

图2 大浪淘沙，沉者为金。统计学习可用于从大量可能的材料结构中筛选出可控的数量以进行实验

统计，统计，统计

尽管机器学习技术已经在医学、量子和材料物理学方面形成了具体的成果和无可替代的前景，但在统计物理学中的进展却没有那么清晰。在法国巴黎萨克莱大学研究机器学习理论的Lenka Zdeborová承认道，"我们仍在等待一个被科学同行认可的，没有机器学习的帮助就无法完成的重要例子。"

当然，机器学习技术在统计物理学中有一些有希望的进展，但Zdeborová说这些技术迄今尚未处于该领域的前沿。她指出，有数十篇论文使用神经网络来研究一些统计模型，例如二维伊辛模型——该模型描述了二维晶格上自旋粒子之间的相互作用，但迄今还没有得到任何全新的发现。

机器学习尚未推动统计物理学的进步也许会令人感到失望，但相关的知识与前景必定走向另一方向。试想识别图像所需的神经网络，每个图像都会包含大量数据(像素)并且会伴有噪音(因为任何给定的图像都会被大量不相关的特征掩蔽)，而且网络中的不同权重之间也存在相关性。

令人高兴的是，多维、有噪声和有关联的问题正是统计物理学家自上世纪中叶以来一直在学习如何应对的问题。研究一种被称为"自旋玻璃"材料的Zdeborová说道："试想针对无序系统中发展出来的物理学理论"。这样的系统具有许多粒子(即许多的维度)，具有有限的温度(即具有热噪声)，而且具有许多粒子间的相互作用(即许多的关联)。事实上，在某些情况下，描述机器学习模型的方程与用于处理统计物理系统的方程完全相同。

这一发现可能是发展一个全面理论的关键，这一理论可以解释为什么这些方法会如此有效。虽然目前机器学习的发展也许比几十年前通常预测的还要进步一些，但其成功仍然主要来自经验性的试错法。Zdeborová总结道："我们希望能够预测最优的机器学习架构，知道应该如何设置参数以及采用何种算法，目前我们还没有线索知道如何在不投入大量人力的情况下获得这些知识。"