8. 人工智能中的神经网络被广泛用于生活中的各个领域中，但它的“黑匣子”特性却让人诟病。神经网络的可解释性在某种程度上对它的应用落地造成了阻碍。你对神经网络在某些领域效果出色但缺乏可解释性这一现象有什么看法？

打开神经网络黑箱·量变引起质变

20373843 秦铭悦

1. 深度学习与神经网络

深度学习是一类新兴的多层神经网络学习算法，因其缓解了传统训练算法的局部最小性，引起机器学习领域的广泛关注。它克服了过去人工智能中被认为难以解决的一些问题,且随着训练数据数量的显著增长以及芯片处理能力的剧增，在目标检测和计算机视觉、自然语言处理、语音识别和语义分析等领域成效卓然，因此也促进了人工智能的发展。

深度学习是包含多级非线性变换的层级机器学习方法，深层神经网络是目前的主要形式，其神经元间的连接模式受启发于动物视觉皮层组织。

1. 应用举例

１ 语音识别

微软研究院的语音识别专家Ｌｉ和 Ｄｏｎｇ 从２００９年开始和深度学习专家 Ｈｉｎｔｏｎ合作。２０１１年微软基于深度神经网络的语音识别研究取得成果，彻彻底底改变了语音识别原有的技术框架。深度神经网络采用模拟人脑的多层结果，逐级地进行信息特征抽取，最终形成适合模式分类的较理想特征。这种多层结构和人脑处理语音图像信息的时候，是有很大的相似性的。

２ 图像识别

e.g. 拍照/截图识别

内容审核与监管

视频内容分析

纸质文档电子化

３ 自然语言处理

计算机处理或“理解”自然语言的学科

４ 搜索广告 ＣＴＲ预估

搜索广告是搜索引擎的主要变现方式，而按点击付费ＣＰＣ又是其中被最广泛应用的计费模式。在ＣＰＣ模式下，预估的 ＣＴＲ （ｐＣＴＲ）越准确，点击率就会越高，收益就越大。

1. 黑匣子特性及问题
2. 特性

神经网络就像“炼丹炉”一样，投喂大量数据，或许能获得神奇的效果。炼丹功后，神经网络也能对没见过的数据进行预测了。然而，这种情况下，神经网络其实成了黑匣子——神经网络是个黑匣子，大致有两层意思。一个是说，我们不能刻画网络具体是在做什么（比如在每一层提取什么样的特征）。另一个是说，我们不知道它为什么在做这些，为什么会有效。

事实上，神经网络之所以起作用，最直观的原因就是，它由大量非线性函数组成。这些非线性函数，使得网络可以学习原始数据中各种抽象级特征。然而，也正是因为神经网络中的这些非线性函数，使得人类往往难以理解，它们是如何起作用的。

2、问题

①错误预判：深度学习的神经网络如果用在医学方向，对疾病进行预测，那么神经网络下的“判断”就不可轻信。医生和法律相关的研究者往往更乐意采用可解释模型，例如线性回归、决策树，因为神经网络在疾病预测中的确出过问题。人们利用神经网络预测肺炎患者的病情发展，其中一项患者特征为是否有哮喘病史。神经网络经过训练后预测，有哮喘病史的患者死于肺炎的可能性较低。但其实结果恰好相反，哮喘本身会给肺炎带来雪上加霜的效果。之所以数据表明哮喘患者较少死于肺炎，往往是因为哮喘能被及早发现，所以患者得肺炎后能被及早治疗。

如果这种神经网络被应用于实践中，将会带来非常危险的结果。

②无法控制：

e.g.被恶意引导学习性别歧视理论、种族歧视理论的AI

③AI带来的潜在危险：Google的机器学习研究员Maya Gupta说，硅谷的研究人员们也在试图打开AI的“黑匣子”。除了运行后操作的准确性，所有人心中还有一个非常大的顾忌：因为不知道它在做什么，所以不确定能不能相信它。

④理论混乱：

围棋的胜利，是机器数据和计算机算力的胜利。媒体神话AI，为学术界带来投资，这是好事；但是研究者仍需宠辱不惊，脚踏实地，不骄不躁，为大众还原真理。

科学的发展一直就是一个迭代的过程：在以往的结论上提出新的问题和假设，设计实验验证，用形式化的理论进行解释；发现实验中不可解释的意外性收获，再做出假设或提出新的问题，设计实验，循环迭代，万不可造神。

1. 我的观点

深度学习目前就好比实验物理，有很多实验观察，但没有理论能够解释。也跟物理学的发展一样，实验总是走在前面。可是没有Michelson／Morley 光速不变的实验观察，也就没有爱因斯坦的狭义相对论。

对于神经网络，打开“黑匣子”与技术应用两者都很重要，切不可矫枉过正。如果人类目前的认识水平无法探求神经网络原理，那么不妨继续应用该技术，并在运用过程中积累数据，量变引起质变，充分利用随着时代发展不断迭代升级的硬件，直到黑匣子被打开。

以下是Ali Rahimi在文章Lessons from Optics, The Other Deep Learning中将光学和深度学习进行的类比：

1. 从结构来看，无论是深度学习还是镜头，都是一层层叠起来的。
2. 从设计流程来看，当要设计一个镜头的时候，往往以一个已知的镜头组合作为基础，然后你跑个仿真，看看这个基础镜头组合的表现和你需要达到的要求存在哪些差距，在合适的地方插入合适的组件来磨平差距。接着，你用一个数值优化器来调参数，以发挥上述镜头组合最大的功效。这个类似于深度学习里面的优化和调超参过程。
3. 从系统组成部件来看，光学组件有的起到反射作用，有的起到衍射作用，有的起折射作用，有的起到散射作用，有的起到相位校正作用等等。深度学习组件（conv、pool、relu等等）有的起到学习空间相关作用，有的起到防止过拟合作用，有的起到增加非线性作用等等。
4. 从发展历史来看，伽利略时代虽然没有光学理论，类似于现在的深度学习，但是同样造出来了人类历史上第一架天文望远镜，推动了天文学的发展。现在深度学习也理论不足，但是极大地推动了人工智能的发展。
5. 从结果看，在几百年里，经过科学家的不懈努力，光学终于形成了一整套比较完备的体系，使得现在的光学工程师在设计镜头的时候有迹可循，而不是像几百年前的伽利略一样靠经验设计。深度学习里面的很多现象还得不到较好的解释。深度学习现象与对应的解释光学也经历过黑箱时代，深度学习现在就是黑箱时代，理论的前进一般会晚于实践，但是假以时日，理论肯定会完备起来，形成类似于当前光学那样的层层抽象的学术体系，只不过这个过程需要所有从业者一起的努力，和一定时间的积淀，也许几十年，也许上百年。

参考文献

[1]孙志军,薛磊,许阳明,王正.深度学习研究综述[J].计算机应用研究,2012,29(08):2806-2810.

[2]余凯，贾磊，陈雨强，徐伟. 深度学习的昨天、今天和明天

[3]周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.

[4] 微信公众号：人工智能与大数据技术. 神经网络“炼丹炉”内部构造长啥样？牛津大学博士小姐姐用论文解读.原文作者Oana-Maria Camburu

[5] Ali Rahimi.Lessons from Optics, The Other Deep Learning.