大数据与人工智能：谁成就了谁？

软件学院 2019011799 罗华坤

人工智能（AI）与大数据（Big Data）作为ABC（cloud computing）的两员，是目前世界互联网行业火热的研究方向。然而，社会上仍存在着一些疑问——人工智能与大数据，究竟谁成就了谁？只有解决了这个问题，才能确定未来大数据与人工智能行业的发展方向与发展力度。在我看来，**人工智能与大数据的关系就像汽车与汽油**，谁也离不开谁。

1.人工智能与大数据“相遇”前

人工智能的概念被普遍认为于1956年在达特茅斯会议上提出。但由于理论尚浅、算力不足、数据稀缺等原因，人工智能领域经历了数次寒冬，使其在21世纪前都发展的不温不火。然而中间也诞生了许多影响深远的模型，例如多层感知机、支持向量机等。这些都被归于浅层机器学习领域，即有效但缺少数据需要较多训练经验技巧，更别说其难以忍受的训练时间了，因此难以被实际应用。[1]

而大数据于1996年被首次提出，IBM总结了其五大特点：大量、高速、多样、低价值密度、真实性。[2]从中可以看出，大数据巨量的特点曾令人十分头疼。尽管其真实来自生活且大量，但由于人手有限，在面对大量数据时只能抽样或直接抛弃，这也是许多银行早期采取的手段，因此大数据的价值密度十分低，间接导致处理这些数据的成本较为高昂。而且随着时间的增长、设备的更新，数据的复杂度增长，更加剧了数据浪费，供远过于求。当时有预测显示，数据增长的速度将很快会超过储存设备增长的速度。[3]

与数据爆炸相对的是，数据孤岛现象的广泛出现，即企业内各部门数据孤立、企业间无数据沟通，使得企业在内部运行成本增加、与外部交流难度加大。[6]无论是人工智能还是大数据，当时都面临着巨大的挑战，**人工智能更像人力车，而大数据还是未开采的原油**。

2.大数据对人工智能的影响

进入21世纪后，GPU的出现大幅度地提高了算力，各种硬件设施也都得到了一定的提升，使得从前需要数天训练的模型如今也可快速得到。此时深度学习初见雏形，由于其从零开始训练、需要抽取不同的特征以不断修正模型、容易过拟合的特性，恰好需要大量的数据来训练模型，大数据也为此提供了可能。而且当时许多研究和实验表明，在模型相似的情况下，更大的数据往往意味着更高的准确率，进一步推动了‘**big data is all you need**’的浪潮，打破了以往忽视数据重视算法的局面。[4]

随后李飞飞斥重金创建了ImageNet数据库，开始运营每年一度的ImageNet比赛。[5]在过程中人们惊奇发现，经过ImageNet数据集预训练后的模型往往表现更好，于是大数据开启了AI的一大发展浪潮，吸引了大量资本进入AI领域。在ImageNet竞赛期间还诞生了许多新型CNN网络，例如GoogLeNet、Resnet、DenseNet等。**可以这么说，大数据不仅在实际训练中对AI有着举足轻重的作用，还为人工智能吸引了大批资金，使其进一步发展。**

同时，**各种各样的数据催生了多样的人工智能**，例如大量气象数据促进了人工智能预报领域的发展，大大提高了预报准确率，而日常的交通数据为导航人工智能提供了源源不断的燃料，自动驾驶领域如今也发展得如火如荼。大数据时代充分利用了其特点，放弃了一部分因果关系，而更注重数据的相关性，**与AI的内部运行特点十分契合**，促进了AI的发展。[7]该研究方式如今也被称为第四范式，即人脑+电脑（电脑主导），区别于原来人脑主导的第三范式。

3.人工智能对大数据的影响

人工智能与大数据结合后，大数据从前的劣势反而成为其优势。可以24小时运行的电脑、人类望尘莫及的计算速度都大大提高了大数据的利用效率。不仅如此，人工智能对大数据的运用还促进了**数据处理技术的进步**，许多平台、软件在这一过程中诞生，例如spark，hadoop，各大IT公司也创立了自己的大数据查询、计算、储存平台，例如阿里巴巴、网易已有了较为成熟的大数据处理体系。

与早期的纯数字不同，人工智能强大的数据处理能力与应用范围**拓宽了数据的范围**，从前基本只做通讯交流作用的图片、音频、视频现也可以用于分析中。人工智能还将许多人类看似无用的数据利用起来，提取抽象底层特征做出自己的判断。美国一研究团队就收集众多化学实验室的失败数据，从前这些数据往往被直接丢弃，而且不同的失败之间没有直接的联系，人类很难建立起直观的联系。而该团队用AI学习到了中间的特征，从而服务于未来实验预测。预测成功率达到89%，超越人类专家预测的78%。[8]

不仅如此，相较人类观测数据做出浅层笼统的判断，人工智能**将大数据运用到生活方方面面且能够针对个体做出判断**。麦当劳利用用户的购买记录，包括地点、各页面的浏览时间、乘坐的交通工具等，来向顾客定点推送食物推荐，最后让顾客的购买频率提高了35%；Spotify同样监测用户日常音乐播放情况，进而推送更符合用户品味的音乐清单；甚至google利用2009年流感疫情数据成功预测了后序感染人数的发展，精确到了州县单位。[9]

最后，随着人工智能的发展，数据的需求量越来越大，大公司越来越难以获得足量的数据，因此大数据的共享成了一个必要条件，这也**促进了数据的流通**。2009年奥巴马政府就曾提出政府数据开放计划，在github上开源了许多软件工具，使人们可以随时查看政府的数据资源，[10]自此开启了越来越多数据的开放进程，也正因如此大数据行业也得到了蓬勃的发展。

4.对现今大数据、人工智能的一点思考

尽管如今这种大数据与人工智能的结合仍十分高效，但仍可以看到其中的一些不足之处。

**大数据始终不是无穷无尽的**，无法遍历生活中的种种情况。例如在自动驾驶领域，公路上可能发生的事情是无法预料的。也许没有一个数据是婴儿坐在马路中间，未经该数据训练过的自动驾驶是否会避让仍是一个问题。况且在某些领域的数据十分紧缺，包括医学的疑难杂症的诊断上。如此依赖数据的模型显然是无法做好诊断工作的。

此外，**大规模训练消耗的资源（电力人力）如今也都达到了较高的水平**。有数据显示，训练多次Bert模型的碳排放量就相当于一次从美国到欧洲的飞行排放。在如今全球节能减排浪潮下，势必会受到更多的阻力。[11]

而幸运的是，目前科学界也对小样本学习有了一定的研究，相信未来能够正确改进如今利用模式的不足。最后，对于现今社会，人工智能与大数据的关系仍是不变的，是相互成就、互相依赖的。

参考文献

1. 余凯 贾磊 陈雨强 徐伟 深度学习的昨天、今天和明天[J]. Ji suan ji yan jiu yu fa zhan = Computer research and development, 2013, 50(9): 1799–1804.
2. Mashey B. Big Data and the Next Wave of InfraStress[J].
3. SHULL F. Data, Data Everywhere[J]. IEEE Software, 2014, 31(5): 4–7. DOI:10.1109/MS.2014.110.
4. HALEVY A, NORVIG P, PEREIRA F. The Unreasonable Effectiveness of Data[J]. IEEE Intelligent Systems, 2009, 24(2): 8–12. DOI:10.1109/MIS.2009.36.
5. RUSSAKOVSKY O, DENG J, SU H等. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211–252. DOI:10.1007/s11263-015-0816-y.
6. 刘洋 企业内部的数据孤岛现象的内在成因和解决建议[J]. 信息系统工程, 2018(4): 93–95.
7. MAYER-SCHÖNBERGER V, CUKIER K. Big Data a Revolution That Will Transform How We Live, Work, and Think[M]. Boston: Houghton Mifflin Harcourt, 2013.
8. RACCUGLIA P, ELBERT K C, ADLER P D F等. Machine-Learning-Assisted Materials Discovery Using Failed Experiments[J]. Nature (London), 2016, 533(7601): 73–76. DOI:10.1038/nature17439.
9. SAMANTHA BOMKAMP. Large Fries? Extra Sauce? Why McDonald's Wants To Track All Your Dining Habits[J]. ECN, 2017.
10. ZUIDERWIJK A M ., GASCó M, PARYCEK P等. Special Issue on Transparency and Open Data Policies: Guest Editors’ Introduction[J]. Journal of Theoretical and Applied Electronic Commerce Research, 2014, 9(3): i–ix. DOI:10.4067/S0718-18762014000300001.
11. STRUBELL E, GANESH A, MCCALLUM A. Energy and Policy Considerations for Deep Learning in NLP[J]. 2019.