大数据与计算模型

李国杰
中国科学院计算技术研究所，北京 100190

摘要
当前，人工智能持续升温，大语言模型吸引了众多人士的关注，并在全球范围内掀起了一股热潮。人工智能的成功本质上不是大算力“出奇迹”，而是改变了计算模型。首先，肯定了数据对于人工智能的基础性作用，指出合成数据将是未来数据的主要来源。然后，回顾了计算模型的发展历程，重点介绍了神经网络模型与图灵模型的历史性竞争；指出了大模型的重要标志是机器涌现智能，强调大模型的本质是“压缩”；分析了大模型产生“幻觉”的原因。最后，呼吁科技界在智能化科研中要重视大科学模型。

关键词
人工智能；大数据；计算模型；神经网络模型；合成数据；涌现

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024017

Big data and computing models

LI Guojie
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract
At present, artificial intelligence continues to heat up. Large language models have attracted much attention and set off a wave of enthusiasm around the world. The success of artificial intelligence is not essentially a "miracle" of large computing power, but a change in computing models. Firstly, this paper affirms the fundamental role of data in AI, and points out that synthetic data will be the main source of data in the future. Then, this paper reviews the development of computing models, highlights the historic competition between neural network models and Turing models. We points out that the important hallmark of large language models is the emergence of intelligence in machines, emphasizes that the essence of large language models is "compression", and analyzes the reasons for the "illusion" of large language models. Finally, we call on the scientific community to attach importance to large scientific models in "AI for research(AI4R)".

Key words
artificial intelligence, big data, computing model, neural network model, synthetic data, emergence

0 引言

如今，大数据在降温，而大模型还在持续升温，AI for Science的惊艳成果正吸引着人们的眼球。如今，科技界对AI的认识和如何选择AI的技术发展路线还存在不少争议，以下是争议较多的几个问题。

- 以数据为中心，还是以模型为中心；
- 大语言模型（large language model，LLM）能否成为通用智能模型；
- 以模拟人类智能为目标，还是发展可能与人类不同的机器智能；
- 连接主义AI与符号主义AI的结合是否必要和可行；
- 机器有没有理解能力；
- 神经元计算模型是否不同于图灵计算模型。

人工智能还处在“伽利略时代”，或者说“牛顿时代的前夜”，面临着诸多未知和挑战。我们要看到现有技术路径的局限性，不能只追求增量式改进；要提倡百家争鸣，过早地锁定技术路线会阻碍人工智能的发展。本文对与大数据和计算模型有关的一些有争议的话题发表一些看法，旨在抛砖引玉，引起大家讨论。

1 数据的基础作用和发展趋势

1.1 数据是人类认识复杂世界的基本途径

大量的科学和工程实践表明：只要找到足够多具有代表性的样本（数据），就可以运用数据找到一个模型或者一组模型的组合，使得它和真实情况非常接近。在更高质量数据集上训练的模型，可能只需要较少的训练或较少的模型参数。近几年，人工智能的重大突破已经凸显出数据的巨大作用。

近年来，人工智能取得重大突破，得益于大数据、大模型和大算力，三者缺一不可。大数据和大算力是大模型发挥作用的前提条件。20世纪，算力和数据都是短板，无法释放大模型的潜力。如今，GPU弥补了算力短板，互联网提供了海量数据，如此才凸显出大模型的威力。目前，最受关注的是算力。NVIDIA公司的市值超过万亿美元，这显示出投资界对算力的高度重视。但从实际应用来看，数据可能要排在第一位。现阶段人工智能的主流是数据智能，从某种意义上讲，没有数据就没有智能，没有数据就没有解释。单纯提高算力不是万能药，野蛮地提高算力对于扩大复杂问题的求解规模没有实际意义。例如，若围棋棋盘扩大到20×20（只增加一行一列），野蛮搜索的算力需要提高10^{18}倍。

牛顿力学、相对论、量子力学的成功，使很多科学家相信靠人类的抽象思维能力就可以了解宇宙中任何事物的奥秘，几个简单的公式就可以解释万事万物。但是，客观世界本质上具有不确定性。从根本上讲，解决很多复杂问题的关键在于消除不确定性，只能用数据来消除不确定性。例如，大气物理学已经有很多理论，有很多精确的偏微分方程，但天气预报的效果始终不尽人意。华为的盘古和DeepMind的Graph Cast，通过理解海量气候数据中的复杂模式来进行预测，不是通过物理方程的计算来预报天气，其预报精度和速度都超过传统的数值天气预报。这一事实表明，数据是人类认识复杂世界的基本途径。

1.2 以数据为中心，还是以模型为中心

在人工智能发展历程中，数据和模型哪

一个更重要，这个问题一直存在争议。这两年大语言模型发展势头很猛，争议也更加激烈。谷歌大脑的创始人吴恩达认为，在过去十年中，人工智能最大的转变是向深度学习的转变，神经网络架构问题已经基本解决，今后十年会转向以数据为中心。现在更有效的方法是固定神经网络架构，寻找改善数据的方法。以数据为中心的AI是一门对成功构建AI系统所需的数据进行系统工程的学科。对于许多实际应用来说，拥有50个精心设计的样本就足以向神经网络解释你想让它学习什么，比如缺陷检查系统。在许多根本不存在巨大数据集的行业中，必须将重点从大数据转移到好数据。

图灵奖得主杨立昆（leCun）的看法则完全不同，他认为大多数人类知识与语言无关，基于自监督的语言模型无法获得关于真实世界的知识，这些模型在本质上是不可控的。我们需要学习一个具备常识推理与预测能力的世界模型，而世界模型是AI大模型未来最理想的道路。很显然，杨立昆的观点是重点发展新的模型，而不是在数据上下功夫。

对技术发展方向的选择不完全是一种学术判断，甚至代表了一种信仰。当神经网络模型受到学术界普遍排斥的时候，Hilton等一批学者毫不动摇，坚信神经网络模型与人脑有一定的相似性。正是因为这一份信仰，神经网络才得以大翻身。在神经网络模型取得巨大成功的今天，科技界对大模型的前途仍然存在两种针锋相对的预判。以OpenAI公司为代表的一方认为，只要扩大模型和数据的规模、增加算力，未来的大模型很可能会涌现出目前没有的新功能，呈现更好的通用性。另一种观点认为，目前的大语言模型不可能实现通用人工智能，必须研究新的智能模型和通用AI技术。

许多学者对ChatGPT等大模型的规模与性能的关系做了研究。一方面，要提高LLM的性能，数据量、模型参数和计算量都要数量级地同步提升[1]。也就是说，目前LLM的代价很高。另一方面，究竟多大的模型是必需的，对于这一问题还没有结论。有学者在研究少样本学习理论，降低模型成本还有很大的研究空间。大语言模型可能不是实现通用人工智能的最佳道路，只是人工智能发展过程中的一个阶段性成果，但它比前两波人工智能采用的技术具有更大的使用价值。大语言模型也不是万能的，尽管其泛化能力超出人们的预期，但本质上还是封闭范围的AI，学到的是互联网上的书本知识，还没有学到与人类实践活动有关的大量隐性知识。目前，关于神经元计算模型的计算复杂性的研究成果还不多，这个课题值得深入研究。在人工智能界独占鳌头的技术一般红火十年左右，如Transformer已经流行6年多了，可能几年后会被新的模型和算法取代。

1.3 合成数据将是未来数据的主要来源

数据并不是取之不尽的资源，数据消费的速度远远高于数据自然产生的速度。有预测认为，到2026年文本数据将被训练完，图像数据将在2040年左右用完。自然语言处理将实现从对已有数据的消费（自然语言理解）到全新数据的生产（自然语言生成）的跃迁，这将是一次巨大变革。合成数据将是未来数据的主要来源。以自动驾驶为例，自动驾驶汽车在道路上可能遇到的每个场景，是无法通过收集真实世界的驾驶数据来覆盖的。Waymo公司（Google母公司旗下研发自动驾驶汽车的子公司）从现实世界收集的驾驶数据仅为300万英里(约483万千米)的数据，而到2019年，自动生成的模拟驾驶数据已达到100亿英里（约161亿 km）的数据。合成数

据不是完全随机产生的，首先要收集真实的数据子集，分析和识别其中的规律和趋势，然后使用这些数据来生成新的数据，这些新数据有可能弥补没有收集到的数据。合成数据不可避免地依赖一部分真实数据用于自身的创建。因此，合成数据永远不会完全取代它所依赖的初始数据。合成数据可以用来验证或扩展已知规律，加速科研的进程，但不大可能直接揭示在原始数据集中不存在的全新规律。

算力网现在很红火，但算力网传送的是数据、模型和任务，算力本身是不能被传送的。在互联网服务中，音视频数据是用户消费的产品，按流量计费，传送越多，公司收入越多。而在算力网中，数据传送是计算成本的一部分，应尽可能减少数据传送。算力网首先要考虑数据在地理分布上的合理性，让同行业的数据相对集中并靠近计算资源。如果大部分数据是合成数据，有算力的地方可以在当地产生需要的大部分数据，从而大大减轻数据传送的压力。到那时，算力网的主要作用转变为广域分布式计算的任务分配和模型的传送。此外，构建算力网应当考虑合成数据的发展趋势。

2 关于计算模型的历史回顾和思考

2.1 人工智能的突破源于计算模型的改变

计算模型有不同的层次，图灵机模型是一种通用的计算模型，理论上可以实现所有的计算。而目前流行的卷积神经网络等智能计算模型是专用的计算模型。通用计算模型有很多种，包括图灵计算（离散变量计算，即递归计算）、神经网络计算（数据驱动的图灵计算）、模拟计算（连续变量计算）和量子计算等新计算模型。计算模型是分析可计算性和计算复杂性的基础。图灵机模型只是若干个计算模型的一种。通用计算模型在可计算性上都是等价的，但对于某些计算问题，不同模型的计算效率有天壤之别。例如，在量子计算模型上用Shor算法做大数分解是多项式复杂性，而在图灵机模型上做大数分解是指数复杂性。

经典的图灵计算模型的递归形式是$f(x,0)=g(x)$；$f(x,y+1)=h(x,y,f(x,y))$。机器学习的递归形式是$f_0(x,0)=g(x)$；$f_2(x,y+1)=h(x,y,f_1(x,y))$；$f\approx L(g)$。在图灵计算模型中，递归迭代函数和输入都是确定的。而在机器学习模型中，每次迭代的结果都产生一组新的迭代方程，f随输入数据而变化，f_0、f_1、f_2是递归函数f的水平分裂。机器学习是不同于经典图灵计算的数据驱动型递归计算[2]。

问题的复杂性随计算模型的改变而改变。人们常说的NP困难问题是对确定性图灵计算模型而言的。自然语言理解、模式识别等NP困难问题，在大语言模型上能被有效解决，这说明大语言模型对这类问题的求解效率远远高于图灵计算模型。人工智能的成功本质上不是大算力“出奇迹”，而是改变了计算模型。从理论上讲，现在还没有明确的证据表明，神经网络模型能够为NP完全问题提供多项式时间的解法（只是针对某些问题实例有多项式复杂性的近似解）。AI研究的新近发展体现了一种趋势，放弃绝对性，拥抱不确定性，即只求近似解或满足一定精度的解，这或许是这次AI“意外”取得成功的深层原因。

图灵机模型和神经网络模型各有优缺点，适合于不同的计算问题。若对一个领域已经有较透彻的理解，要求完全正确或非常精确的解，选择图灵机模型一般更合适。若对一个领域了解不深入，问题很复杂，只求近似解，选择神经网络模型可能

更合适。需要注意，理论上有些NP问题求近似解仍然是NP困难问题。LLM求解NP问题是针对某些问题实例，而不是针对整个问题类。

2.2 两种计算模型的历史性竞争

人工智能经历了60多年的发展，曾经两次跌入低谷，目前是兴起的第三次浪潮。波浪式的发展始终围绕符号主义和连接主义的竞争，而背后实际上是图灵机模型和神经网络模型的竞争。从源头上理清神经网络模型的发展脉络，有助于我们了解它的潜力和局限性。

1936年，图灵在《论可计算数及其在判定问题上的应用》中提出图灵机模型，这个模型成为80多年来计算机和人工智能发展的基本模型。1943年麦卡洛克（McCulloch）和皮茨（Pitts）提出了神经元计算模型，这个模型在可计算性上与图灵模型是等价的（理论上无限容量的神经网络模型被认为是图灵完备的，即可以模拟任何图灵机的计算过程）。对自动机理论而言，神经网络模型可能比图灵模型更有价值。

1945年，冯·诺伊曼发表了一篇长达101页的报告《EDVAC报告书的第一份草案》，为计算机的发展奠定了坚实的基础。麦卡洛克和皮茨的《神经活动中内在思想的逻辑演算》是这份报告唯一的参考文献。冯·诺伊曼在给维纳的信中也提到，麦卡洛克和皮茨的大胆尝试，与图灵博士的非神经观点同样重要。后来，冯·诺伊曼在他的遗作《自复制自动机理论》中指出，图灵机和神经网络模型分别代表了一种重要的研究方式——组合方法和整体方法。麦卡洛克和皮茨对底层的零件进行了公理化定义，可以得到非常复杂的组合结构；图灵定义了自动机的功能，并没有涉及具体的零件[3]。这说明神经网络计算模型对计算机概念的形成产生了重大影响。

由于当时计算机的性能太低，数据也缺乏，基于神经网络模型构建计算机的想法无法实现。1946年11月，冯·诺伊曼给维纳写信时指出，为了理解自动机的功能及背后的一般原理，我们选择了太阳底下最复杂的一个对象……在整合了图灵、皮茨和麦卡洛克的伟大贡献后，情况不仅没有好转，反而日益恶化……这些人向世人展示了一种绝对的且无望的通用性。从此，冯·诺伊曼放弃了用神经网络模型构建计算机，转向研究自复制自动机。

值得指出的是，早在1948年，图灵也写了一篇论文《智能机器》（Intelligent machinery），提出了与图灵机不同的计算模型——“无组织机器”，它模拟婴儿的大脑皮层，通过适当的干扰训练来实现组织化。实际上，这篇论文介绍的是早期的随机连接神经网络模型，描述了目前人工智能连结主义的基本原理，包括遗传算法和强化学习等。由于没有得到他老板的认可，这篇论文一直没有被发表，直到2004年才被发现[4]。这一被历史淹没的重要论文，说明图灵同样看好神经网络模型。如果学术界早看到这篇论文，今天的计算机世界可能是另一幅模样。

冯·诺伊曼早就预言，信息理论包括两大块，即严格的信息论和概率的信息论。以概率统计为基础的信息理论对于现代计算机的设计更加重要。统计意义的正确性与确定性、计算程序的严格正确性是解决复杂问题的不同思路。图灵机模型和神经网络模型的竞争，实际上是科学技术发展史上常见的功能主义和结构主义的竞争，蒸汽机、飞机等重大发明都是先实现功能后来才研究发现其结构原理的，计算机和人工智能走的路也一样。几十年来，神经网络模型一直比不过图灵模型，在学术界受到

排挤。但有一批学者坚持不懈，终于让结构主义取得了一次初步胜利，神经网络模型开始显示它的威力。

2.3 大模型的重要标志是机器涌现智能

在AlphaFold2实现蛋白质结构预测和GPT4令人惊奇的功能中，机器猜想都发挥了关键作用，这说明大规模的机器学习神经网络已涌现出某种程度的认知智能，大模型的核心特征是“涌现”功能。神经科学家Terrence Sejnowski这样描述LLM：“达到了一个阈值，就好像一个外星人突然出现，可以用一种奇怪的方式与我们交流。”也有人比喻大模型就像毛毛虫变成蝴蝶，幼虫代表训练模型的数据，蝴蝶代表着从数据中创造的AI。

大模型是否具有涌现和理解能力，对这个问题学术界还没有形成共识。2022年，在一项针对自然语言处理的调查中，受访者被询问是否同意以下说法：根据文本训练的生成模型，在给定足够的数据和计算资源的情况下，能够在某种非琐碎的意义上理解自然语言。在480名受访者中，51%同意，49%不同意。有些学者认为，LLM所谓的“涌现”行为是度量标准引起的“海市蜃楼”，一旦改变指标进行测试，所谓的“涌现”特性就会消失。不同学者对涌现的理解可能不同，涌现未必要看性能测试曲线上是否有突变的拐点。过去的人工智能做不到的事情，今天的大模型可以做到，从宏观上看就是涌现了一些意想不到的新功能，如机器翻译、计算机生成文艺作品、新材料的发现、全自动设计CPU芯片等。可以说，大模型已经具有一定的理解和创造能力。

冯·诺伊曼的遗作《自复制自动机理论》指出，自动机理论的核心概念在于复杂性，超复杂的系统会涌现出新的原理。他提出了一个重要的概念——复杂度阈值。突破了复杂度阈值的系统，因在数据层的扩散和变异作用而不断进化，从而可以做很困难的事情。现在的神经网络模型有成千上万亿个参数，可能已接近冯·诺伊曼讲的复杂度阈值。复杂度阈值并不等于模型的规模，智能也不等同于复杂性。需要深入研究如何准确定义和测量复杂性、智能与复杂性是什么关系以及如何理解和预测涌现行为。

机器理解不同于人的理解。机器翻译可以不懂语义，AI天气预报可以不懂气象理论，这可能是一种新颖的“理解”形式，一种能够实现预测的理解形式。我们需要开发新的基准和探索方法，以深入了解不同类型的智力和理解的机制。理解、智能和意识有3个不同层次的内涵，有理解能力未必有自我意识。所谓“对齐”和“微调”是人类认知和机器“认知”的接口。即使机器有意识，源头还是人类，应当能找到人类影响机器的接口。因此，我们对机器的认知不必过于恐慌。

2.4 大模型的本质是“压缩”

20世纪90年代，Hinton就提出，深度学习的本质可能就是压缩。OpenAI首席科学家Ilya Sutskever提出，压缩可能就是学习的本质！马毅团队提出“白盒”。Transformer也指出，智能的本质就是压缩。著名计算机科学家李明教授采用第一性原理和Kolmogorov复杂性理论证明了“理解就是压缩，大模型就是压缩”。

大语言模型的本质是一个性能强大的近似无损的数据压缩器，即将输入文件的知识“压缩”后，以权重矩阵的形式存储在神经网络模型中。ChatGPT原始训练数据集的大小是900 TB，训练完成后，模型参数文件大约是64 TB，整体的压缩比约为

1:14，而传统语言模型（如Bert、RNN）的压缩比大约是1:10~1:8。

组合搜索的关键是压缩搜索空间。AlphaGo只搜索了一个很小比例的空间（约$1/10^{150}$），就能找到相当准确的满意解。中国科学院计算技术研究所（以下简称中科院计算所）做的“启蒙1号”也是将几乎无穷大的搜索空间压缩到10^6。为什么搜索空间可以被大幅度地压缩？因为许多理论上的解空间对实际求解没有意义，解分布也不是随机的。必须有效地识别和利用数据中的关键模式和结构，在巨大搜索空间中快速找到最有价值的区域。

2.5 大模型的“幻觉”

LLM的主要功能是预测（猜），不是搜索正确答案。搜索是没有创造性的，猜测可能有创造性，这种猜测可以看成人类智能的补充而不是替代。向LLM提问其实不是人类在测试计算机的智能，而是LLM在测试提问者对机器智能的了解程度，因此，这可以看成一种反向的图灵测试。OpenAI科学家Andrej Karpathy指出：从某种意义上说，大语言模型的全部工作恰恰就是制造“幻觉”，大模型就是“造梦机”。提问者是否能够让“幻觉”和自己的现实一致，很大程度取决于提问者对产生内容的检查能力。

图灵停机问题的不可判定性说明复杂系统具有不可预测性，不存在一个通用的程序能够预测所有复杂系统的运行结果。也就是说，不确定性是复杂系统的本质特征，要想弄清楚某个复杂系统的运行结果，唯一的办法就是让这个系统实际运行。长期的计算思维教育使我们习惯了用执行固定程序的观念来看待复杂系统，实际上大模型像人脑一样是个复杂系统，其信息处理过程并不是执行固定的程序。即使是在推理阶段，由于采用概率性推理，执行过程也存在随机性。相同的问题也可能生成不同的答案，结果必然有不确定性。

哥德尔不完备性定理表明，完备性和一致性不能同时满足。LLM的“幻觉”是系统不一致性的表现，泛化能力是完备性的表现。从这个角度看，“幻觉”是由于追求泛化能力造成的。LLM的泛化能力和“幻觉”是一个硬币的两面，我们需要在与“幻觉”共存的环境下发展人工智能。

3 基于大科学模型的智能化科研

3.1 大语言模型的局限

国内流行“大模型”的说法，国外并不流行large model或big model的说法，流行的是large language model（LLM）。这是OpenAI公司带领的方向，主要受ChatGPT的影响。DeepMind团队也用LLM的方法学习必要知识，但主要采用强化学习方法，机器本身也产生了很多数据。在科研工作中，我们应更加关注DeepMind团队的工作。为了区别于大语言模型，笔者建议发展大科学模型（large science model，LSM）。科研领域对模型的正确性和精度要求较高，模型具有识别自身能力不足的“自知之明”与提高模型准确性同等重要，科研大模型必须找到对付AI“幻觉”的办法。

神经网络模型的哲学基础是经验主义，实际上采用的是不完全归纳推理，存在或然性，得出的结论可能存在偏差或错误。经验主义也无法完全解释人类的创造性思维和创新能力。人工神经网络是一个有高表达能力的通用函数类，其理论逼近能力优于经典的数值函数表示。但普通算法通常无法获得理论近似率，基于点样本的有效算法能多大程度地利用这些优越的近似

性质，仍然是深度学习领域中的开放问题。

基于神经网络模型的深度学习方法难以保证高精度。由于神经网络模型满足不了13个“9”的高精确性要求，中科院计算所在全自动设计的CPU芯片“启蒙1号”的研发中，发明了一种新的机器学习模型——二进制推测图（BSD），用来表示电路逻辑。BSD不但能保证精度，而且与大语言模型一样，也具有“涌现”功能。

3.2 智能化科研与传统科研的区别

人工智能不仅应用于基础研究（AI for science，AI4S），还应用于技术研究和工程实施（AI for technology，AI4T）。因此，笔者建议将“第五科研范式”称为“智能化科研”（AI for research，AI4R）。早期的AI研究采用的数学基础是基于符号的数理逻辑，很多数学工具用不上，这也是早期AI研究跌入低谷的重要原因。这次AI之“火”复燃，明显的改进是采用了统计学的方法，处理的对象变为大量的数据，数学工具可以大显身手，这也是AI技术突飞猛进的重要原因。

传统科研的主要方式是求解函数$y=f(x)$，即通过实验和理论研究先找到反映客观规律的函数f（一般用微分方程的形式表示），或者根据已知的知识编写求解f的程序，再通过输入x求得结果y。但对于复杂或者较为通用的问题，人类还没有获得函数f的确切表达，只能通过已知的输入x和输出y来拟合函数f，这是求函数值的反问题。智能化科研（AI4R）大多是在解决“反问题”。

为特定应用编写计算机程序是可行的，但为一般智能编写计算机程序会引发组合学爆炸。经过几十年努力，现在有了另一种求解途径，即一个基于学习而不是编写计算机程序的替代方案。大模型相当于一种可能具备通用智能的应用程序，这些算法有时会失败，但足以在现实世界中处理一些复杂的问题。

参考文献:

[1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB]. arXiv preprint, 2020, arXiv: 2001.08361.

[2] 张寅生. 新一代人工智能计算模型的创新及其哲学意义[J]. 学术界, 2021(5): 59-69.
ZHANG Y S. Innovation and philosophical significance of the new generation of artificial intelligence computing models[J]. Academics in China, 2021(5): 59-69.

[3] WADA E. John von Neumann: theory of self-reproducing automata[J]. IPSJ Magazine, 2002, 43.

[4] JACK C. The essential turing[M]. Oxford: Clarendon Press, 2004: 411-432.

作者简介

李国杰（1943- ），男，中国工程院院士，第三世界科学院院士，中国科学院计算技术研究所首席科学家，中国计算机学会名誉理事长。主要从事计算机体系结构、并行算法、人工智能、大数据，计算机网络、信息技术发展战略等方面的研究，发表科学论文150多篇，出版了三本《创新求索录》文集，长期致力于发展曙光高性能计算机产业和CPU等核心技术的自主可控。

收稿日期： 2023-12-25