

人工智能的生态树及算力研究^{*}

周娜¹, 何铮², 何为民³

(1. 海口经济学院, 海口 571132; 2. 海南政法职业学院; 3. 海南久迪物联网科技有限公司)

摘要: 介绍了通用人工智能 AGI 的生态树, 重点分析了 AGI 生态树的决定性基础建设—算力。算力位于 AGI 科技发展金字塔的底端, 它的发展制约了人工智能的发展。此外, 还介绍了当前算力发展的进程和方向。

关键词: AGI 生态树; 算力; ChatGPT; 人工智能

中图分类号: TP3

文献标识码: A

Research on Ecological Tree and Computing Power of Artificial Intelligence

Zhou Na¹, He Zheng², He Weimin³

(1. Haikou University of Economics, Haikou 571132, China; 2. Hainan Vocational College of Political Science and Law; 3. Hainan Jiudi Internet of Things Technology Co., Ltd.)

Abstract: This article introduces the ecological tree of General Artificial Intelligence (AGI), with a focus on analyzing the decisive infrastructure of AGI ecological tree—computing power. Computational power is located at the bottom of the AGI technology development pyramid, and its development restricts the development of artificial intelligence. The article also introduces the current process and direction of computing power development.

Keywords: AGI ecological tree; computing power; ChatGPT; AI

0 引言

通用人工智能的生态树如图 1 所示, 它庞大的树根(算力)从大地中吸取营养(数据), 形成茁壮的树干(基座大模型), 上面结满了大苹果(纵向应用包括聊天机器人、智能操作系统、浏览器、第三方应用插件、办公套装软件、AI 图像、软件开发、医疗、制药、机器人等)。修剪刀剪去病枝烂果, 修正树势, 保证果实茁壮成长。

科技发展都是为了应用。在 AGI 生态树中, 纵向应用是结出的成千上万的大苹果。这些应用几乎无所不至, 国内外对此均有巨大的市场需求。技术、资金等进入的门槛不是太高, 而且收效快, 众多大中小企业、创业者都可参与, B 端、C 端均可分享。

树壮才能硕果累累。基座大模型的算法当前都不成问题, 采用 LLM(大语言模型)已成共识, 国内外基本所有的大模型都来源于 2017 年谷歌开源的 Transformer 模型。国外的大模型有微软和 OpenAI 的 GPT-4、谷歌的 PaLM 2、亚马逊的 Titan 大模型、Intel 的 Aurora genAI 模型、Meta 公司的 LLaMA 模型, 而国内则是百度“文心一



图 1 AGI 强人工智能生态树

言”等的百模争雄。近日, 智源研究院发布了“悟道 3.0”系列大模型及算法, 这是目前我国市场上为数不多的可商用开源大语言模型之一。虽然“悟道 3.0”的模型参数没有公布, 但 2021 年 6 月发布的“悟道 2.0”模型参数规模达到 1.75 万亿, 是当时我国首个万亿级模型。而且, “悟道

^{*} 基金项目: 2023 年海南省高等学校科学研究项目—基于大数据的海南特色分布式养老监护系统的研究(Hnky2023-50)。

3.0”中的天演团队将“天演”接入我国新一代百亿亿次超级计算机——天河新一代超级计算机。通过“天演—天河”的成功运行,实现鼠脑 V1 视皮层精细网络等模型仿真,计算能耗均可降低约 10 倍以上,计算速度实现 10 倍以上提升,达到全球范围内最极致的精细神经元网络仿真性能。

数据是肥料,来源于人类的自然积累,位于人脑(经验)、书本(文字)、音像作品中。ChatGPT 的训练使用了来自于 2021 年 9 月前互联网公开的文本数据(维基百科、书籍、期刊、Reddit 社交新闻站点等),共 45 TB、近 1 万亿个单词。但国内 AGI 中公益数据不多,数据的版权分散在各个机构、互联网厂商的手中,若能建立中文数据集的产业联盟,打破版权壁垒,将人工智能训练数据统一规划,则必是好事一桩。

互联网上的文本数据有限,且有很多虚假、违反法律/道德和意识形态以及侵犯隐私的信息。下一代大模型的参数可能达到万亿级以上,为了避免数据短缺问题成为训练的瓶颈,人们推出了“合成数据”。合成数据(synthetic data)是通过计算机技术人工生成的数据,虽然不是由真实事件产生的数据,但能够在数学上或统计学上反映原始数据的属性,因此可以作为原始数据的替代品来训练、测试并验证大模型。合成数据具有更高的效率(较短的时间内大量生成)、更低的成本、更高的质量(通过深度学习算法合成原始数据中没有的罕见样本,规避用户隐私问题)。据称,到 2024 年,人工智能和数据分析项目中的数据预计将有 60% 来自合成数据。

对人工智能的监管(法律、道德、伦理)就是对果树的修剪。为了促进生成式人工智能技术健康发展和规范应用,国家网信办在 2023 年 4 月 11 日发布了关于《生成式人工智能服务管理办法(征求意见稿)》。2023 年 2 月 16 日,中美等 60 多国签署声明“军事领域负责任使用人工智能”。5 月 22 日,OpenAI 的创始人奥特曼、总裁和首席科学家联合撰文称:“我们最终可能需要类似于 IAEA(国际原子能机构)的东西来进行超级智能方面的努力;任何超过一定能力(计算等资源)门槛的努力都需要接受国际权威机构的检查,要求进行审计,测试是否符合安全标准,对部署程度和安全级别进行限制等。”

庞大的根系是果树生长结果的根本。算力作为根系,一直都是掣肘人工智能发展的基石。

在人工智能(AI)研究过程中,有两种流派:一种流派是严格的逻辑推理,另一种流派是研究模拟生物学的人类及大脑。后者在 20 世纪 80 年代推出了神经网络学说,把人脑看成是一台碳基计算机,用我们的硅基计算机来模拟生物的进化。随着神经网络模型不断变大,神经元不断增加,不知在突破某个值后是否会发生突变,变得更为智能。

有专家按照现代神经网络的架构分析人类神经网络参数最少为 100 万亿(100~1000 万亿),仅同时激活 10%,其算力最少需 0.79 EFLOPS,功耗仅几十瓦。计算机所需要的数据空间大概为 400~4 000 TB^[1]。

神经网络学派推出了循环卷积等各种算法,但在 20 世纪没有真正奏效的重要原因是当时计算机运行速度不够快、数据集不够大。随着“神经元”的膨胀,大模型(LLM)计算复杂度以指数级增加,当时计算机有限的内存和处理速度不足以解决任何实际的 AI 问题,而每次算力的突破都带来人工智能的爆发。IBM 研制的深蓝(算力为 11.38 GFLOPS)、谷歌的 AlphaGo(算力是深蓝的 30 万倍)带动了深度学习的大发展,而这次 NVIDIA 的 A100、H100 及 CUDA 则助推了 GPT 生成式大模型的问世。

人类通常不断有各种科学幻想和预言(期待的成果),而一些优秀的科学家为实现这些幻想和预言研究出各种理论(牛顿的力学及万有引力、霍金的黑洞及宇宙起源、爱因斯坦的相对论等)。理论往往都超前于当时的科技水平,超前于算法,算法则要求有算力做支撑。有分析认为,当前的 AGI 算力和能耗都远比不上人脑的碳基计算机!当下,千亿级、万亿级参数的生成式大模型比人脑百万亿级参数还差很多,但能耗却高得可怕。2023 年 4 月,OpenAI 就因需求量过大而停止了 ChatGPT Plus 的销售。据数据预测,到 2030 年全球超算算力将达到 0.2 ZFLOPS,平均年增速超过 34%,未来 10 年人工智能算力需求将会增长 500 倍以上。为此,除 NVIDIA 外,几乎所有的芯片大厂都在布局 AI 算力芯片。

AMD 近期就推出一款新一代超级 AI 芯片——将 CPU 和 GPU 融合在一起的 MI300X。该芯片是针对 LLM 的优化版,拥有 192 GB 的 HBM 内存,与 NVIDIA H100 相比,有 2.4 倍的 HBM 密度和 1.6 倍的 HBM 带宽优势,这可以在芯片上容纳更大的模型,以获得更高的吞吐量。单个 MI300X 可以运行一个参数多达 800 亿的模型,这是全球首次在单个 GPU 上运行这么大参数量的模型。这表明在无人驾驶 L5 等应用项目中需要强大边缘计算算力的地方,MI300X 将大放异彩。

1 提高算力的路径

(1) 提高制程及 Chiplet

根据摩尔定律,集成电路上可以容纳的晶体管数目大约每经过 18 个月便会增加一倍,因此微处理器的性能每隔 18 个月提高一倍,价格下降一半。而光刻“制程”的进步是每 18 个月单位面积硅片上可容纳的晶体管数目翻倍的主要原因。经历了二十多年的正增长后,新世纪摩尔生长曲线在变缓,尤其是近年来芯片“制程”的提升越来越难,从 16 nm 到 7 nm 节点,芯片制造成本也在大幅提升,

出现“摩尔定律失效”的议论。在将芯片制程继续从 7 nm 降至 5 nm、3 nm、2 nm 的同时,芯片业也在从各方面应对 AGI 算力提升和功耗降低面临的挑战。

Chiplet 是一种芯片设计和集成的方法,它将一个大型 AI 芯片分解为多核异构、多个独立的功能模块片段(称为 IP 核或 Chiplet)。每个 Chiplet 指已经过验证的、可以重复使用的具有某种确切功能的集成电路设计模块,如图形处理器(GPU)IP、神经网络处理器(NPU)IP、视频处理器(VPU)IP、数字信号处理器(DSP)IP 等。Chiplet 设计可以使芯片设计更加模块化,具有更强的灵活性。不同的芯片片段可以独立设计和优化,然后通过集成技术组合成一个完整的芯片。这种模块化的设计使芯片开发更加容易,而且每个芯片片段可以在独立的制造工艺下生产,而不是整个芯片都使用同一种复杂的制造工艺,这样可以降低制造成本,提高芯片的产量和良品率。

当前计算机均为冯·诺依曼架构,计算单元和存储单元分离,存储带宽制约了计算系统的有效带宽,造成时延长、功耗高等问题。采用多维集成芯片将多个芯片堆叠在一起,使存储与计算完全融合,以新的高效运算架构进行二维和三维矩阵计算,具有更大算力(1000 TFLOPS 以上)、更高能效(超过 10~100 TFLOPS/W)、降本增效三大优势,有效克服了冯·诺依曼架构瓶颈,实现计算能效的数量级提升。

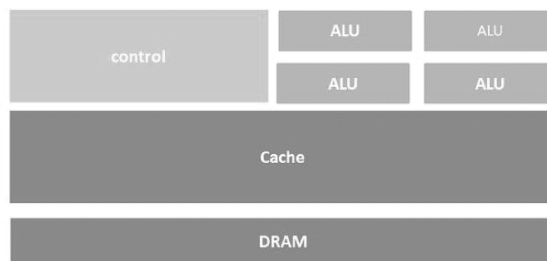
AMD 近期推出的 MI300X 就采用了 3D 堆叠技术和 Chiplet 设计,配备了 9 个基于 5 nm 制程的芯片组,置于 4 个基于 6 nm 制程的芯片组之上。

(2) CPU 到 GPU、TPU、NPU 和多核异构

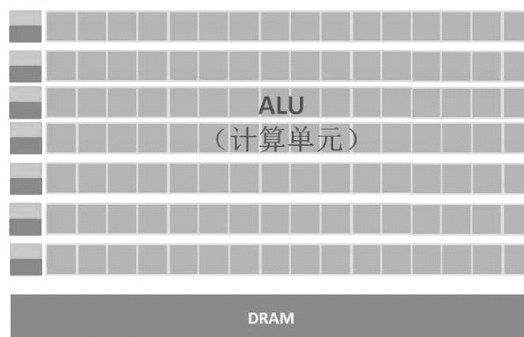
最初的处理器是 CPU,采用冯·诺依曼架构,存储程序,顺序执行。在图 2(a)CPU 架构图中,负责计算的 ALU 区占的面积太小,而 Cache 和 Control 单元占据了大量空间。为提升算力,出现了多线程的流水线作业,继而出现新结构的各类处理器。

GPU(图形处理器)。在图像处理中,每个像素点都需要处理,而且处理的过程和方式十分相似,都需要相同的精度、不高的运算,为此推出了 GPU(Graphics Processing Unit)。GPU 架构示意图如图 2(b)所示,它有数量众多的计算单元和超长的流水线,特别适合处理大量的类型统一的数据。它将求解的问题分解成若干子任务,各子任务并行计算。虽然 GPU 为了图像处理而生,但它在结构上并没有专门为图像服务的部件,只是对 CPU 的结构进行了优化与调整,所以 GPU 不仅可以在图像处理领域大显身手,还被用来进行科学计算、密码破解、数值分析、海量 AGI 数据处理。

TPU(Tensor Processing Unit)即张量处理器。张量是机器学习中常见的多维数组结构,谷歌就专门为此开发



(a) CPU 架构



(b) GPU 架构

图 2 CPU 及 GPU 的架构图

了一款 TPU 芯片。TPU 专注于高效地执行大规模张量计算,具有高带宽和低延迟的内存访问以及快速的张量操作指令,用于进行机器学习和人工智能任务中的张量计算,提供高效、快速且能耗低的计算性能。据称,TPU 有高性能和能效效率,与同期的 CPU 和 GPU 相比,可以提供 15~30 倍的性能提升,以及 30~80 倍的效率(性能/瓦特)提升。2021 年谷歌发布的第四代 TPU v4,每个芯片拥有 2.5 PFLOPS 的算力。

NPU(Neural Network Processing Unit)是用电路模仿人类的神经元,深度学习的基本操作是神经网络中神经元和突触的处理,其中存储和处理是一体化的。NPU 通常不单独存在,而是存在于多核异构的处理器中。典型代表有国内的寒武纪芯片和 IBM 的 TrueNorth。Dian-NaoYu 指令直接面对大规模神经元和突触的处理,一条指令即可完成一组神经元的处理,并对神经元和突触数据在芯片上的传输提供了一系列专门的支持。

(3) 人工智能专用计算机

芯片能力直接影响算力训练效果和速度,当前最好的 AI 芯片是 NVIDIA 的 A100 和 H100,其最高浮点算力分别实现 19.5 TFLOPS 和 67 TFLOPS,但 AGI 所需算力和单个处理器所有的算力差距太大,当前的 AGI 大模型的计算都需要成千上万个 GPU 堆叠起来应用。业内公认,做好 AI 大模型的算力门槛是 1 万枚 A100 芯片,GPT3.5 大模型需要 2 万块 A100 来处理训练数据。近期,谷歌推出了拥有 26 000 片 H100 的超级 AGI 计算机 A3。

这样大量 GPU 堆起来使用就需要采用专门针对深度学习 and 神经网络的互联硬件、架构和软件工具包,从而实现更高的计算速度和更低的功耗。通过优化和协调来协同作用,再去组合模拟优化的路径。NVIDIA 在 2006 年发布了一个名为 CUDA(Compute Unified Device Architecture)的软件工具包。他们搭建的 CUDA 开发者平台以良好的易用性和通用性让 GPU 可以用于通用超级计算,使用 CUDA 和不使用 CUDA,两者在计算速度上往往有数倍到数十倍的差距。26 000 块 H100D 的 A3 超级计算机单独叠起用的算力是 $26\,000 \times 67\text{ TFLOPS} = 1.724\text{ EFLOPS}$,而采用 CUDA 后可提供高达 26 EFLOPS 的算力。Meta 的超级计算机 RSC(Research Super Cluster) 花费了数十亿美元,配备了由 16 000 个 NVIDIA A100 GPU 组装成的 2 000 个 NVIDIA DGX A100 系统(8 个 A100+CUDA),集联起来,在其巅峰时期算力可以达到近 5 EFLOPS,所以 A100、H100、CUDA 三者确定了 NVIDIA 成为当前全球 AGI 算力霸主的地位。

在谷歌的数据中心中,大规模的 TPU 芯片则是通过光开关互连网络连接在一起的。光开关是一种高带宽、低延迟的互连技术,它使用光纤作为传输介质,实现快速的数据传输和通信。通过光开关,TPU 芯片之间可以进行高速、并行的通信,以实现大规模的并行计算。这种互连方式能够提供低延迟和高带宽的通信性能,确保在分布式计算环境中各个 TPU 芯片能够有效地协同工作,共同完成复杂的机器学习任务。谷歌的光开关互连网络架构为其算力集群提供了高效的通信基础设施,使得 TPU 能够实现高性能、大规模的机器学习计算。这种光开关连接的设计能够支持谷歌处理海量数据和进行复杂计算,提供出色的计算能力和可扩展性。

(4) 下一代计算机

当前,硅基计算机的摩尔定律似乎快走到尽头,于是人们寄希望于“类人脑生物计算机”的非硅基计算机。未来的计算机可能会采用一种或多种可能的技术和架构技术提高算力计算性能和效率。

① 量子计算机:量子计算机利用量子力学原理来执行计算。这种计算机可以大大提高处理大量数据和解决复杂问题的速度(例如加密和化学模拟)。近期量子计算机好消息不断:在国际超算大会(ISC)上,NVIDIA 公布了一个搭载 384 颗 Grace CPU 超级芯片的超级计算机 Isambard 3。这台超级计算机 FP64 峰值性能达到约 2.7 petaFLOPS,功耗低于 270 kW。当前全球大量的量子计算研究都在 NVIDIA GPU 上运行。欧洲量子计算设施于利希超算中心(JSC)计划与 NVIDIA 共同建立一座量子计算实验室。在 NVIDIA 量子计算平台的基础上开发一台量子超级计算机,作为于利希量子计算用户基础设施

(JUNIQ)的一部分,运行高性能、低延迟的量子一经典计算工作负载。近期,IBM 科学家在《自然》杂志上发表了论文《容错前的量子计算实用性证据》。容错量子计算指的是有量子纠错保护的量子计算。IBM 科学家宣布,已经设计出一种方法来管理量子计算的不可靠性,从而得出可靠、有用的答案。英特尔公布了名为 Tunnel Falls 的硅自旋量子芯片,它拥有 12 个硅自旋量子比特,是英特尔迄今为止研发的最先进的硅自旋量子比特芯片。

② 光子计算机:光子计算机使用光子而不是传统的电子来执行计算,这种计算机可以实现更高的速度和更低的功耗,并且更容易进行并行计算。

③ DNA 计算机:DNA 计算机使用 DNA 分子来执行计算,这种计算机可以处理大量数据,并且具有很好的并行性。

2 合作共赢

算力是 AGI 的决定性的基础建设,是 AGI 金字塔的底座,需要人力和资金的巨大投入。ChatGPT 全面放导致用户激增、算力不足。除大量购入 GPU 外,近期又和甲骨文(Oracle)讨论了一项协议,如果任何一家公司为使用大规模 AI 的云客户所提供的计算能力不足,那么双方就将相互租用对方服务器。

而日本经产省 2022 年 11 月牵头组建了“技术研究组合最先端半导体技术中心(LSTC)”,随后,丰田汽车、索尼、NTT、NEC、软银等 8 家大型公司成立合资企业 Rapidus 公司,计划最早 2025 年在日量产 2 nm 芯片,希望 2025 到 2030 年的几年中为其他企业提供代工服务。

国内重心放在了大模型上,近 3 年大模型发布数量为世界之首(达 79 个),已超过美国,但实际上国内 AGI 的短板在于算力,国内 AGI 都受制于算力,仅面向 B 端用户或内测,其使用效果、影响力、评价有目共睹。正如百度 CEO 李彦宏所说:“文心一言跟 ChatGPT 的差距大约是两个半月,这两个月的差距我们要用多长时间才能赶上,也许很快,也许永远也赶不上。”

面对芯片封控,算力是买不来的,也不是几年时间可以一蹴而就的,也不是短期能解决的。在面对封锁的重大问题上(如两弹一星、北斗、航天等),国家都采用举国体制,赶上了世界先进水平。AGI 重要性不逊于核能,其核心算力牵涉到国民经济的各个方面,需要在国家层面上,组织大厂合作共赢。当前国内百模竞争只会分散人力、财力,从而拉大差距。

参考文献

- [1] _pinnacle_. 模拟人类大脑每秒计算数量级及参数容量的估计 [EB/OL]. [2023-06]. <https://blog.csdn.net/u010165147/article/details/50481303/>.

通信作者:何为民,254716268@qq.com。

(责任编辑:薛士然 收稿日期:2023-06-12)