AI 芯片和大模型计算技术发展与应用

Development and Application of AI Chip and Large Model Computing Technology

王迎帅 [1] 张元刚 [2] 杨宜镇 [3] 杨文娟 [2]

Yingshuai Wang[1] Yuangang Zhang[2] Yizhen Yang[3] Wenjuan Yang[2]

1. 上海侬找信息科技有限公司 中国·上海 200949
2. 上海泽阳智能科技有限公司 中国·上海 200120
3. 南京工业大学浦江学院 中国·江苏 南京 211134

1. Shanghai Nongzhao Information Technology Co., Ltd., Shanghai, 200949, China
2. Shanghai Zeyang Intelligent Technology Co., Ltd., Shanghai, 200120, China
3. Pujiang College, Nanjing Tech University, Nanjing, jiangsu, 211134, China

摘 要：A 芯片和大模型计算技术作为人工智能领域的重要组成部分，正在引起越来越多的关注。随着科技的发展和人们对智能化技术的需求增加，AI 芯片和大模型计算技术不仅在学术研究领域得到广泛应用，也在商业和工业领域展现出巨大的潜力。论文将探讨 AI 芯片和大模型计算技术的发展历程、应用案例以及它们的结合应用，同时也会讨论它们所面临的挑战和未来发展的趋势。通过深入了解 AI 芯片和大模型计算技术的发展与应用，我们可以更好地把握当前人工智能技术的前沿动态，并为未来的研究和创新提供参考依据。

Abstract: The A-chip and large model computing technology, as important components of the field of artificial intelligence, are attracting increasing attention. With the development of technology and the increasing demand for intelligent technology, AI chips and large model computing technology have not only been widely applied in academic research, but also shown enormous potential in commercial and industrial fields. This paper will explore the development history, application cases, and combined applications of AI chips and large model computing technology, as well as the challenges they face and future development trends. By deeply understanding the development and application of AI chips and large model computing technology, we can better grasp the cutting-edge dynamics of current artificial intelligence technology and provide reference basis for future research and innovation.

关键词：AI 芯片；模型，计算机

Keywords: AI chip; mode; computer

DOI: 10.12346/csai.v2i1.9261

1 AI 芯片技术的发展与应用

AI 芯片是人工智能技术的核心推动力之一，它为实现复杂的计算任务提供了高效的硬件支持。AI 芯片根据其设计架构和功能特点可以分为通用 AI 芯片和专用 AI 芯片。通用 AI 芯片具有灵活性和适应性，能够处理多种不同类型的任务，而专用 AI 芯片则更加专注于某一特定领域的优化运算。

AI 芯片的功能广义上所有面向 AI 应用的芯片都可以称为 AI 芯片。目前一般认为是针对 AI 算法做了特殊加速设计的芯片。现阶段，这些人工智能算法一般以深度学习算法为主，也可以包括其他浅层机器学习算法。

AI 芯片技术经过多年的发展，取得了显著的进步。从最初的基于规则的专家系统到如今的深度学习和神经网络模型，AI 芯片的设计和性能逐渐提升。AI 芯片在图像识别、语音识别和自然语言处理等领域的应用得到了广泛认可和应用。例如，在图像识别方面，AI 芯片可以协助实时检测与分类，使得安全监控和智能驾驶等应用更加可靠和高效。在语音识别和自然语言处理领域，AI 芯片的快速计算

【作者简介】王迎帅（1974-），中国辽宁人，硕士，工程师，从事区块链、工业互联网、工业物联网芯片及相关应用、传感器等研究。

能力可以实时解析和响应大量的语音指令，提升用户的交互体验。

近年来，随着 ChatGPT 的出现和发展，大量 AI 大模型不断涌现，并由此延伸出了一系列大模型上层应用，也促进了通用人工智能产业蓬勃发展。

近年来，随着ChatGPT的问世，大量AI大模型不断涌现，并由此延伸出了一系列大模型上层应用，也促进了通用人工智能产业蓬勃发展。

ChatGPT 等大模型训练，首先需要对海量数据进行并行处理；其次，大模型应用也带动了终端用户使用频率大幅提高，数据流量巨大，从而对服务器的数据处理能力、可靠性及安全性等要求出现显著提升；最后，在大模型新的应用场景中，数据的质和量均发生显著变化，非结构化数据占比激增，数据规模也明显增长。

因此，基于传统 CPU 的服务器已无法满足算力需求，转而对基于 AI 芯片的服务器需求迅猛增长。据公开数据显示，2025 年全球 AI 服务器市场规模预计将达到 288 亿美元。AI 服务器的快速增长，大力拉动了对 AI 芯片需求的增长。从大模型的训练到推理，先进的 AI 芯片扮演着重要角色。

据 ChatGPT 的公开数据显示，ChatGPT 整个训练算力消耗非常大，达到了 3640PF-days（即假如每秒计算一千万亿次，需要计算 3640 天），换算成英伟达 A100 芯片，它单卡算力相当于 0.6P 的算力，理想情况下总共需要大概 6000 张，在考虑互联损失的情况下，需要一万张 A100 作为算力基础。在 A100 芯片 10 万人民币 / 张的情况下，算力的硬件投资规模达到 10 亿人民币。而整个的数据中心还需要处理算力以及服务器等，规模应该在 100 亿人民币以上。近期，英伟达等 GPU 芯片企业股价飞涨，GPU 芯片中国国内企业超高购买包括存货则进一步侧面验证了我们对于 AI 芯片的开发研究，以及对大模型的应用发展的技术储备变得至关重要 [1]。

面对灵活多变的大模型应用场景，以 GPU、ASIC、FPGA 等为代表的 AI 芯片被广泛应用于 AI 服务器中，与 CPU 组合来满足高吞吐量、高并发和并发互联的需求；传统 CPU 也从单核心向多核心转变，满足并发处理能力和速度提升的需要。

表 1 列出了 AI 芯片的三种技术架构，其优缺点比较如下。

AI 部署模式正在发生转变，它们不仅被部署于数据中心，而且越来越多地被部署在功耗和散热要求比较严格的边缘设备上。现在，每瓦功耗所提供的性能（或称为性能 / 功耗比）通常比简单的性能指标（TOPS）更为重要。随着 AI 算法的不断演进，网络模型和数据格式也在不断演化发展。

2 大模型计算技术的发展与应用

大模型，又称为预训练模型、基础模型等，是“大算力 + 强算法”结合的产物。大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调，如在下游特定任务上的小规模标注数据进行二次训练，或者不进行微调，就可以完成多个应用场景的任务。

大模型计算技术的发展与应用已经取得了巨大的进步，成为推动人工智能领域发展的关键技术之一。在过去的几年里，随着计算硬件的不断更新和算法的改进，大模型计算技术在机器学习、数据挖掘和推荐系统等多个领域都取得了显著的成果。大模型指的是机器学习领域中具有大规模参数和架构的深度学习模型。这些模型通常包括成千上万的神经元和数百万到数十亿的参数，需要大量的计算资源来进行训练和推理。

大模型计算技术的基本原理是利用大规模的计算资源来训练和优化复杂的模型。这些计算资源可以是传统的服务器集群，也可以是云计算平台或者分布式计算系统。通过充分利用计算资源，大模型计算技术可以处理海量的数据，并从中挖掘出隐藏的模式和规律。

大模型计算技术的发展历程可以追溯到深度学习的兴起。深度学习作为一种基于神经网络的机器学习方法，需要大量的计算资源来进行训练和推断。因此，为了满足深度学习的需求，各种大规模计算平台和算法优化方法相继提出。

在机器学习领域，大模型计算技术的应用广泛而深入。例如，在图像分类任务中，研究人员利用大规模计算平台和深度神经网络，实现了高精度的图像分类和识别。在数据挖掘领域，大模型计算技术可以处理海量的结构化和非结构化数据，从中挖掘出有价值的信息。在推荐系统领域，大模型计算技术可以基于用户的行为数据和偏好，实现个性化的推荐 [2]。

表 1 AI 芯片的三种技术架构及其优缺点比较

技术架构	优点	缺点
GPU 芯片（图形处理器）	通用处理器，编程灵活性高有更强的并行计算能力成熟的开发环境	相对于 FPGA 和 ASIC 相比 CPU 芯片价格和功耗过高
ASIC 芯片（专用集成电路）	可针对专门的任务和需求进行定制化，可实现低成本、低功耗、高性能等	芯片通用性差，可编程架构设计难度高，投入更大
FPGA 芯片（现场可编程门阵列）	半定制化，可对芯片硬件层进行编程和配置。相对于 GPU 有更低的功耗等	需掌握硬件编程语言，相对于 ASIC 有一定的电子管冗余，功耗和成本有进一步压缩空间

2022 年 11 月 30 日，OpenAI 发布 ChatGPT（GPT，Generative Pre-trained Transformer），一款人工智能技术驱动的自然语言处理工具，能够通过学习和理解人类的语言来进行对话和互动，甚至能完成撰写邮件、视频脚本、文案、翻译、代码等任务。ChatGPT 是基于 GPT 模型构建的基于 Web 端的“聊天机器人”，对于每一个对话提问，由后端已训练好的 GPT3.5 或 GPT4 模型进行预测，并实时返回文字预测的结果，从而实现对话任务。总的来说，ChatGPT 是一个能够生成文本，回答问题和进行自然语言对话的 AI 模型。它可以帮助我们完成各种任务，如聊天、写作、信息检索和问题回答等。

ChatGPT 中用到的我们一般称之为大型语言模型，也称大语言模型、大模型（Large Language Model，LLM；Large Language Models，LLMs）。

大语言模型是一种深度学习模型，特别是属于自然语言处理（NLP）的领域，一般是指包含数千亿（或更多）参数的语言模型，这些参数是在大量文本数据上训练的，例如模型 GPT-3，PaLM，LLaMA 等，大语言模型的目的是理解和生成自然语言，通过学习大量的文本数据来预测下一个词或生成与给定文本相关的内容。从实际应用表现来看，大语言模型具备回答各种问题、编写文章、编程、翻译等能力，如果深究其原理，LLM 建立在 Transformers 架构之上，并在很大程度上扩展了模型的大小、预训练数据和总计算量。

自此以 ChatGPT 为首的应用通过人工智能生成内容模式（AIGC）将人们的视线再一次聚焦在了人工智能应用领域，AIGC 全称为 AI-Generated content，是指基于大型预训练模型，通过已有数据寻找规律，并通过适当的泛化能力生成相关内容的技术。人工智能 AIGC 的应用基础就是基于大模型训练的一种产物，通俗来说，大模型通常是计算机在大规模无标注数据上进行训练，学习出一种特征和规则，基于大模型进行应用开发时，将大模型进行微调，或者不进行微调，就可以完成多个应用场景的任务；并且，大模型具有自监督学习能力，不需要或很少需要通过人工标注的数据进行训练，降低训练成本，因而能够加速 AI 产业化进程，降低 AI 应用的门槛。另外，随着大模型不断地迭代，大模型能够达到更强的通用性以及智能程度，从而使得 AI 能够更广泛地赋能各行业应用。

3 AI 芯片和大模型计算技术的未来发展趋势和应用前景

3.1 AI 芯片和大模型计算技术的未来发展趋势

定制化 AI 芯片的兴起：随着各行业对 AI 技术的需求多样化，定制化 AI 芯片将成为趋势，以满足特定领域的需求。

小尺寸、低功耗的 AI 芯片：随着技术的进步，未来 AI 芯片将越来越小型化、低功耗，使其能够广泛应用于移动设备和物联网领域。

多模态集成的大模型计算技术：未来的大模型计算技术将更加注重多模态融合，实现图像、语音、文本等多种信息的联合推理和分析能力。

智能边缘计算兴起和分布式智能并行。为了满足对实时性和隐私保护的需求，将 AI 计算推向边缘设备成为未来的发展方向，边缘计算将进一步提升 AI 技术的应用前景。考虑到算力算法资源和延迟等因素，一方面，未来 AI 大模型更多应用于边缘计算和分布式智能系统，提供更高效、低能耗的智能服务。另一方面，甚至可能出现去中心化、中心数据中心与边缘数据中心多路径互联的全分布式云。边缘计算将为未来的百亿级终端提供人工智能运算能力，形成万物感知、万物互联、万物智能的智能世界。

AI 芯片将伴随着下列 AI 技术的发展而不断进化。

① AI 技术将逐步实现更多能力的智能化。随着算法模型、芯片设计和传感器技术的进一步提高，AI 技术的可靠性和性能将得到大幅提高。在未来的几年里，AI 技术将逐渐实现对更加复杂的自然语言、图像和视频数据的智能化处理，使得 AI 技术在更多领域得到广泛应用。

② AI 技术将逐步实现更加个性化的服务。在过去，大多数的 AI 算法和系统都是采用相同的方法，针对同样的数据进行处理，这种统一的处理方式往往无法满足不同用户的需求。未来，更多基于 AI 技术的个性化应用，使得 AI 系统能够针对不同用户进行知识的挖掘和服务的提供，从而更好地满足用户需求。

③ AI 技术将逐步实现更强的商业化价值。AI 技术已经成为现代技术进步的重要推手，各种企业和机构都在向 AI 技术转型。未来，将会有更多基于 AI 技术的成功商业模式，使得 AI 技术在商业价值上得到进一步的提升。

其中就包括了 ChatGPT。从 2022 年开始，虽然随着以 OPENAI 为代表的 ChatGPT 推出，加快了人工智能场景的落地应用，但是行业的商业落地仍处于早期，主要面临着场景需求碎片化、人力研发和应用计算成本高、长尾场景数据较少导致模型训练精度不够、模型算法从实验室场景到真实场景效果差距大等行业痛点。而大模型的出现能够在提高模型的通用性、降低训练研发成本方面等方面降低 AI 落地应用的门槛。

基础大模型和垂直行业模型协同并进发展。通用大模型沉淀的知识与认知推理能力向垂直行业模型输出，通过预训练和专用预训练实现业务场景的感知、认知、决策、执行能力，再将执行与学习的结果反馈给大模型，通过监督微调及强化学习等手段优化后的大模型，能够学习到该领域或行业的特定知识和规律，形成一套有机循环的智能系统，增加大模型产业的参与者与应用方，加速模型进化。

3.2 AI 芯片和大模型计算技术的应用前景

过去十年中，通过“深度学习 + 大算力”从而获得训练模型是实现人工智能的主流技术途径。由于深度学习、数据

和算力这三个要素都已具备，全世界掀起了“大炼模型”的热潮，也催生了大批人工智能企业。但是，在深度学习技术兴起的近 10 年间，AI 模型基本上是针对特定应用场景需求进行训练的，即小模型，属于传统的定制化、作坊式的模型开发方式。传统的 AI 模型从研发到投入应用需要完成包括确定需求、数据收集、模型算法设计、训练调优、应用部署和运营维护等阶段组成的整套流程。这意味着除了需要优秀的产品经理准确确定需求之外，还需要 AI 研发人员扎实的专业知识和协同合作能力完成大量复杂的工作。大模型能够实现 AI 从“手工作坊”到“工厂模式”的转变。

大模型通过从海量的、多类型的场景数据中学习，并总结不同场景、不同业务下的通用能力，学习出一种特征和规则，成为具有泛化能力的模型底座。基于大模型进行应用开发或面对新的业务场景时，将大模型进行微调，例如在下游特定任务上的小规模有标注数据进行二次训练，或不进行微调，就可以完成多个应用场景的任务，实现通用的智能能力。由此利用大模型的通用能力可以有效地应对多样化、碎片化的 AI 应用需求，为实现规模推广 AI 落地应用提供可能。

从大模型发展历程中能够看出，多模态大模型是发展趋势之一。由于具有在无监督情况下自动学习不同任务、并快速迁移到不同领域数据的强大能力，多模态大模型被广泛认为是从限定领域的弱人工智能迈向强人工智能的路径探索。OpenAI 联合创始人、首席科学家 Ilya Sutskever 也曾表示，“人工智能的长期目标是构建多模态神经网络，即 AI 能够学习不同模态之间的概念，从而更好地理解世界”。将文本、语音、图像、视频等多模态内容联合起来进行学习，大模型由单模态向多模态方向发展，能够对更广泛、更多样的下游任务提供模型基础支撑，从而实现更加通用的人工智能模型。

更具体来看，大模型带来的更强大的智能能力，能够推动人工智能向更高级智能应用领域迈进，如 AIGC、更智能的对话客服等领域。GPT-3 等大模型在新闻文本生成、商业文本分析、法律文本分析等领域具有较高的产业应用价值[3]。

2022 年，大模型正在成为 AIGC 领域发展的算法引擎。在大模型的能力加持下，包括以文生图以及虚拟数字人等 AIGC 类应用将快速进入商业化阶段，并为元宇宙内容生产带来巨大的变革。大模型正在让人工智能技术从五年前的“能听会看”，走到今天的“能思考、会创作”，未来有望实现“会推理、能决策”的重大进步。

总之，大模型计算技术在机器学习、数据挖掘和推荐系统等领域的应用已经成为科学研究和商业创新的重要支撑。通过不断优化算法和提升计算硬件性能，我们可以预见大模型计算技术在未来的发展前景将更加广阔。

参考文献

[1] 算法小陈.一文搞懂大模型、GPT、ChatGPT等AI概念[OL].

[2] 付二局,鲁钟情,张超.AI技术在未来的发展趋势及应用[J].中国信息化,2023(6).

[3] 刘小虎.智能化控制系统中的数据采集与处理技术应用[J].电子技术,2023(3).