大数据与决策研究

（政策与技术跟踪专题）

2021 年第 16 期（总第 59 期）

广西壮族自治区信息中心
广西壮族自治区大数据研究院　　2021 年 5 月 17 日

编者按：目前人工智能技术成为全球瞩目的新焦点，已全面渗透到医疗、教育、自动驾驶、安防等领域。人工智能技术的崛起离不开数据、算法、算力三大要素，而人工智能芯片解决了传统芯片架构算力不足的问题，成为当前人工智能技术的核心硬件基础。我国“十四五”规划中针对新一代人工智能，提出要加强专用芯片研发攻关。本期主要介绍人工智能芯片（AI 芯片）相关情况。

本期要目

人工智能芯片的功能及类型

一、AI 芯片定义

广义上所有面向人工智能（AI）应用的芯片都可以称为 AI 芯片，一般认为是针对 AI 算法做了特殊加速设计的芯片。按设计思路主要分为三大类：专用于机器学习尤其是深度神经网络算法的训练和推理的加速芯片、受生物脑启发设计的类脑仿生芯片、可高效计算各类人工智能算法的通用 AI 芯片。

二、AI 芯片功能

1. 训练。在平台上，通过对大量的数据进行学习，形成具备特定功能的神经网络模型。该功能对 AI 芯片有高算力、高容量和访问速率、高传输速率、通用性的要求。

2. 推理。利用已经训练好的模型，通过计算，对输入的数据得到各种结论。该功能的芯片主要注重算力功耗比、时延、价格成本的综合能力。实验证明低精度运算（如 float16，int8）可达到几乎和 float32 同等的推理效果，所以 AI 推理芯片也有低精度算力的要求。

三、AI 芯片类型

目前 AI 芯片主要包括图形处理器（GPU）、现场可编程门阵列（FPGA）、数字信号处理（DSP）、专用集成电路（ASIC）、众核处理器、类脑仿生芯片和通用 AI 芯片等。

（一）图形处理器（GPU）

GPU 是一种由大量核心组成的大规模并行计算架构，专为同时处理多重任务而设计，原本的功能是帮助 CPU 处理图形显示的任务，尤其是 3D 图形显示。为了执行复杂的并行计算，快速进行图形渲染，GPU 的核数远超 CPU，但每个核拥有的缓存相对较小，数字逻辑运算单元也更简单，适合计算密集型的任务。

深度神经网络的训练过程中计算量极大，而且数据和运算是可以高度并行的。由于 GPU 与深度学习的需求不谋而合，最先被引入运行深度学习算法，成为高性能计算领域的主力芯片之一。但由于 GPU 不能支持复杂程序逻辑控制，仍然需要使用高性能 CPU 配合来构成完整的计算系统。

（二）现场可编程门阵列（FPGA）

FPGA 作为专用集成电路领域中的一种半定制电路出现，既解决了定制电路灵活性上的不足，又克服了原有可编程器件门电路数量有限的缺点。FPGA 利用门电路直接运算，速度快，用户可以自由定义这些门电路和存储器之间的布线，改变执行方案，以期得到最佳效果。FPGA 可以集成重要的控制功能，整合系统模块，提高了应用的灵活性。与 GPU 相比，FPGA 具备更强的计算能力和更低的功耗。

目前，FPGA 的主要厂商 Xilinx 和 Altera 推出了专门针对 AI 加速的 FPGA 硬件和软件工具。各个主要的云服务厂商，比如亚马逊、微软、阿里云、华为云等推出了专门的云端 FPGA 实例来支持 AI 应用。

（三）数字信号处理（DSP）

DSP 是一种由大规模集成电路芯片组成的用来完成某种信号处理任务的处理器，善于测量、计算、过滤或压缩连续的真实模拟信号。针对滤波、矩阵运算、FFT（快速傅里叶变换）等需要大量乘加法运算的特点，DSP 内部配有独立的乘法器和加法器，从而大大提高了运算速率。

目前应用于 AI 领域的 DSP 主要用于处理图像、视频等视觉系统方面的任务。这些 DSP 中加入了专为深度神经网络定制的加速部件，如矩阵乘和累加器、全连接的激活层和池化层等。

（四）专用集成电路（ASIC）

ASIC 是一种为专用目的设计的，面向特定用户需求的定制芯片，在大规模量产的情况下具备性能更强、体积更小、功耗更低、成本更低、可靠性更高等优点。ASIC 分为全定制和半定制。全定制一般比半定制的 ASIC 芯片运行速度更快，但开发效率低。半定制使用库中标准逻辑单元，设计时可以从标准逻辑单元库中选择门电路、加法器、比较器、数据通路、存储器甚至系统级模块和 IP 核，从而提高系统设计效率。

采用 ASIC 芯片进行深度学习算法加速，表现最为突出的是 Google 的 TPU。TPU 比同时期的 GPU 或 CPU 平均提速 15~30 倍，能效比提升 30~80 倍。

（五）众核处理器

众核处理器采用将多个处理核心整合在一起的处理器架构，主要面向高性能计算领域，作为 CPU 的协处理器存在。众核处理器适合处理并行程度高的计算密集型任务，如基因测序、气象模拟等。比起 GPU，众核处理器支持的计算任务的控制逻辑和数据类型要更加复杂。Intel 的至强融核处理器（Xeon Phi）是典型的众核处理器。众核处理器的结构能有效地利用现代网络和服务器等应用中较高的线程并行度，适用于数据中心部署的各类 AI 训练和推理任务。

（六）类脑仿生芯片

类脑仿生芯片的主流理念是采用神经拟态工程设计的神经拟态芯片。神经拟态芯片采用电子技术模拟已经被证明的生物脑的运作规则，从而构建类似于生物脑的电子芯片，即“仿生电子脑”。神经拟态主要指用包括模拟、数字或模数混合超大规模集成电路 VLSI（也包括神经元或者神经突触模型的新型材料或者电子元器件）和软件系统实现神经网络模型，并在此之上构建智能系统的研究。受到脑结构研究的成果启发，复杂神经网络在计算上具有低功耗、低延迟、高速处理、时空联合等特点。

（七）通用 AI 芯片

AI 芯片的最终成果将是通用 AI 芯片，并且最好是淡化人工干预的自学习、自适应芯片。目前尚没有真正意义上的通用 AI 芯片诞生，而基于可重构计算架构的软件定义芯片

或许是通用 AI 芯片的出路。软件定义芯片是将软件通过不同的管道输送到硬件中来执行功能，使芯片能够实时地根据软件、产品、应用场景的需求改变架构和功能，实现更加灵活的芯片设计。清华大学微电子学研究所设计的 AI 芯片 Thinker，采用可重构计算架构，支持多种 AI 算法。（科技导报《人工智能芯片发展的现状及趋势》）

人工智能芯片的发展及应用

一、发展现状

人工智能（AI）芯片技术领域的国外代表性企业包括Google、NVIDIA、Intel 和 SAMSUNG 等公司，国内则呈现出百花齐放的态势，主要包括中科寒武纪、地平线机器人、深鉴科技、百度、华为、平头哥等。其中寒武纪在 2016 年发布的 AI 处理器是世界首款商用深度学习专用处理器。2019 年 9 月 25 日，阿里巴巴旗下的芯片研发公司平头哥发布了首颗云端超大型 AI 推理芯片——含光 800，在业务测试中，一颗该芯片的计算能力相当于十颗 GPU（通用 AI 处理器）。该芯片目前已经部署在阿里云平台，供阿里内部的多个视觉业务场景大规模使用，未来还将应用到医疗影像、自动驾驶等场景。

全球各大芯片公司和科技公司都在积极进行 AI 芯片的布局。就目前来看，AI 芯片领域逐渐呈现出三股势力：第一股势力是致力于通用 AI 芯片的专业芯片厂商，比如 NVIDIA、Intel、AMD 等；第二股势力是致力于定制化 AI 芯片的新兴 AI 独角兽企业，比如商汤科技、旷视科技、依图科技；第三股势力是致力于云端 AI 芯片的互联网公司，比如谷歌、百度、阿里巴巴等。

二、应用场景

目前人工智能芯片主要应用于云端训练、云端推理、终端推理领域，在云计算、ADAS、智能终端、智能安防等领域已实现了较为广泛的应用。

（一）云计算

用于云端训练和推理，大多数的训练工作都在云端完成。移动互联网的视频内容审核、个性化推荐等都是典型的云端推理应用。目前 GPU 因其通用性好、性能强、编程环境优良、生态成熟等因素在云端训练市场占据主流，但 GPU 存在投资研发成本较高、生态构建较难等问题，谷歌、微软、华为、百度等国内外科技公司开始尝试布局云端专用芯片以提高效率、抢占市场。云端主要的代表芯片有谷歌 TPU、Nvidia TESLA V100、华为昇腾 910、Nvidia TESLA T4、寒武纪 MLU270 等。

（二）ADAS

随着人工智能技术的不断发展，智能汽车也得到高速发展。无人驾驶技术也是当下发展的热点，其中 ADAS（Advanced Driving Assistant System，高级驾驶辅助系统）做出了决定性的贡献。ADAS 需要处理大量的激光雷达、摄像头等传感器实时采集的数据，并在极短时间内处理完数据并及时反馈。ADAS 的优越性体现在对控制模型优化和综合信息处理的算法上，主要包括神经网络控制和深度学习算法等。AI 芯片的飞速发展，可以满足 ADAS 中的图像分析、

环境感知等环节对计算速度的要求。传统芯片的计算延时无法满足无人驾驶的应用场景，只有 AI 芯片才能实时处理随时变化的交通信息及各类传感器的反馈信息。英伟达、英特尔等公司近年先后针对自动驾驶推出高算力（100TOPS 以上）主控芯片，我国的华为 MDC600、黑芝麻科技华山 2 号等芯片的算力达到 100TOPS 以上，可以满足 L3 级别以上自动驾驶需求。

（三）智能终端

华为在 2017 年 9 月推出包含专用 AI 模块的麒麟 970 芯片，并成功应用在随后推出的 Mate10 系列的智能手机上。该款芯片搭载了寒武纪的 NPU（Neural-network Processing Unit，嵌入式神经网络处理器），成为“全球首款智能手机终端 AI 芯片”。紧跟其后，Apple 公司发布了 A11 Bionic 芯片并应用在 iPhone X 系列的手机终端。A11 Bionic 芯片中应用了双核架构神经网络处理引擎，提升了用户在拍照等方面的使用体验。在智能手机终端上应用 AI 芯片可以让其具备更强的深度学习和推断能力，让各类基于深度神经网络图像处理技术的应用能够为用户提供更完美的使用体验。

除了在智能手机上，AI 芯片在其他智能终端上也有广泛的应用，比如智能家居、无人机领域。在 AI 芯片的帮助下，扫地机器人可以敏捷地躲避障碍物，较脏的地方重点清扫，无人机的图像处理功能也变得更加强劲。

（四）智能安防

安防领域现在正处于“智能化”升级阶段，边缘推理芯

片使得安防摄像头具有了推理、筛选功能，而不是只具备之前单纯的影像记录功能。比如安装在家里的 AI 智能监控系统，可以对陌生面孔进行识别，并且通过 WiFi 将安装在家里的视频、音频、监控门窗损坏的传感器等连接起来，使用 AI 技术进行筛选，将有用信息发送到手机、笔记本电脑等终端，紧急时刻还可以主动向出门在外的主人进行报警提醒。智慧安防监控系统将使用边缘推理芯片的安防摄像头所采集到的信息传递到云端，并且在 IT 系统所在的云端引入人工智能、大数据等技术，有效降低了传统安防领域过度依赖人力的问题。除此之外，AI 技术可以用在商店客流量监控器、入侵者检测器等设备中，采用图像识别技术，过滤掉监控视频中的无用信息，自动识别不同物体，使得统计结果更为可靠。

（五）VR 设备

VR（Virtual Reality，虚拟现实）是指利用计算机图形系统和多种控制的接口设备，在计算机上生成的可交互的三维环境，并给用户提供沉浸感觉的一种技术。现在 VR 设备主要用在游戏体验和电影电视方面。微软为自身 VR 设备研制的 HPU 芯片，可同时处理来自多个摄像头和多种运动传感器的数据，并具有 CNN（卷积神经网络）运算的加速功能，能够充分满足 VR 设备实现功能的计算需求。

三、AI 芯片的关键技术

AI 芯片是基于新工艺、新器件，从工具到架构都有所优化的新型芯片，加上不断迭代的算法和超前、多样化的应用，

才形成了 AI 芯片的核心竞争力。AI 芯片涉及的关键技术众多，其中主要分为五大类：工艺、器件、芯片架构、算法、应用这五大类，以下重点对其中几点进行介绍。

（一）新型工艺与器件的突破

基于 CMOS 工艺节点（16nm，7nm，5nm）的不断突破，器件的集成度越来越高。一些新兴器件，比如 3D NAND、Flash Tunning FETs、FeFET、FinFET，应用的越来越广泛。3D NAND 通过堆叠内存颗粒来扩展 2D NAND 闪存的存储容量，可满足移动消费端 AI 芯片的存储要求。FeFET（Ferroelectric Gate Field-effect Transistors）虽然与现有逻辑晶体管的结构相同，但是具有可拓展、非易失、低功率等优点。高带宽片外存储器技术，如 HBM、高速 GDDR、LPDDR、STT-MRAM 的应用也可以大大提升芯片的存储和运算能力，STT-MRAM 为自旋转移矩磁阻内存，其第一代产品是基于 40nm 制造工艺，容量只有 32MB。第二代产品基于 28nm 制造工艺，容量已增加到了 128MB。LPDDR（Low Power Double Data Rate SDRAM）作为第二代低功耗内存技术，由于功耗低、体积小的特点，可以用于移动终端的 AI 芯片中。

除了器件工艺的突破，封装技术的提升也能大大提高微系统的性能，3D 堆叠技术是将不同功能的芯片通过层间孔互联工艺堆叠起来的系统级封装工艺技术，可以减小微系统外形尺寸，降低功耗，提高芯片速度。

（二）算法的迭代和创新

算法是 AI 技术的灵魂，目前 AI 云端芯片主要以训练、

学习为主，AI 终端芯片以推理、应用为主。但无论是类脑芯片的自我学习功能，还是终端芯片的推理功能，都是基于神经网络算法和机器学习算法。神经网络算法又是基于神经网络互联结构（比如卷积神经网络 CNN、循环神经网络 RNN、长短时记忆 LSTM 等）以及深度神经网络系统结构（如 AlexNet、GoogleNet、VGGNet 等）的一类算法。可见 AI 算法是一门很大的学问，AI 芯片必须与 AI 算法相互迭代升级，才能得到更广泛的应用。

深度学习算法是人工神经网络算法的拓展，多基于半监督式学习算法，通过输入数据的分类和回归，进而进行预测。常见的深度学习算法有：受限玻尔兹曼机 RBN、堆栈式自动编码器等。深度学习算法相比以前的神经网络算法相比，拥有更多的“神经元”、更复杂的连接层、更强大的计算能力来训练，所以基于深度学习的 AI 技术有更强的推理能力和迁移学习的能力。AI 算法的突破和创新，能让 AI 技术更上一层楼。

（三）芯片系统级结构的优化

为了满足云计算和边缘计算 AI 芯片的性能和功耗的要求，需要优化人工智能芯片的系统架构。比如，神经网络芯片的卷积神经网络吞吐量和功耗之间的平衡，就需要架构师在芯片系统级架构上给出优化方案。AI 芯片多采用多核、众核的系统级架构，以突破在提高芯片性能时遇到的三个限制（互联时延、设计复杂度、功耗）。设计这些系统级架构是

将 AI 芯片中各个处理器分别设计和优化，从而降低整体设计的复杂度；单个处理器中的互联减小了传输的距离，降低了互联时延；多核结构在满足性能提高的同时，减小晶体管整体翻转的频率，从而降低功耗，解决了单核结构高频率、高功耗的老问题。AI 芯片的结构优化方法还有很多，比如 SIMD（Single Instruction Multiple Data，单指令多数据流）技术，用于数据密集型的运算上，能让多媒体应用芯片如虎添翼。片上网络 NoC 是片上集成系统 SoC 发展来的新的通信方法，常用于多核架构中。除此之外，存储器结构、内存接口结构的优化也是 AI 芯片的关键技术。

四、AI 芯片产业的发展趋势

当今 AI 芯片的研发方向基本还是基于冯·诺依曼架构，但随着深度学习加速器的不断增加，原有的架构并不能有效地解决带宽的问题，架构的创新是 AI 芯片产业发展面临的一个不可回避的问题。目前，模仿人类脑部神经元结构设计的“脑类芯片”和“神经形态芯片”已经在研发的初期阶段。

AI 算法与 AI 芯片开发的一体化发展也将是 AI 芯片产业的趋势。首先算法公司和芯片公司分别专注各自的领域，硬件性能的提升很快会被软件消化掉。而且目前算法公司和芯片公司不能实现倾囊相授的合作，传统的合作模式无法解决芯片迭代速度慢和 AI 算法更新速度快之间的矛盾。所以现在有一些公司开始将算法和芯片作为一个整体进行开发、迭代，最终构建一个完整的人工智能平台，这种开发方式将具有明显优势。

除此之外，广阔的移动终端消费市场决定了应用在移动终端上的 AI 芯片将会是开发重点。击败世界围棋冠军的谷歌智能机器人 AlphaGo 其实是基于谷歌的云端 AI 处理器，所以击败世界冠军是一个很大的云端人工智能平台，而不是一个小小的机器人。阿里巴巴的含光 800 也是基于云平台。应用在移动终端的 AI 芯片要求体积小，集成度高、功能定制化程度高，需要芯片公司的职能集中化。现在许多公司都成立的自己专门的芯片部门，比如中兴微电子、华为海思、阿里平头哥等等，都体现了各大公司对核心芯片自主可控的追求，以及芯片研发的部门化和战略化。（电子技术与软件工程《AI 芯片的发展及应用》）

编辑部地址：南宁市体强路 18 号广西信息中心 1412 号房
联系电话：0771-6113592
电子邮箱：dsjyjs@gxi.gov.cn
网　　址：http://gxxxzx.gxzf.gov.cn/

扫描二维码获取
更多决策参考信息