人工智能芯片技术体系研究综述

施羽暇
（中国信息通信研究院，北京 100191）

摘　要：人工智能技术是当前各国关注的新焦点。人工智能技术的发展对计算芯片提出了新的需求，深度学习算法需要海量数据的训练，而传统计算架构无法支撑深度学习算法的大规模计算需求，因此新架构的人工智能芯片层出不穷。分析了人工智能芯片不同的技术路线，比较了不同路线的特点，研究了人工智能芯片产业全球及我国的发展态势，分析了我国人工智能芯片发展面临的机遇与挑战，并对未来人工智能芯片技术发展趋势进行了展望。

关键词：人工智能；芯片；技术路线
中图分类号：TP331.2
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2019063

Review on artificial intelligence chip technology system

SHI Yuxia
China Academy of Information and Communications Technology, Beijing 100191, China

Abstract: Artificial intelligence technology is the new focus of current countries. The development of artificial intelligence technology has put forward new requirements for computing chips. Deep learning algorithms require the training of massive data, while traditional computing architectures can’t support the large-scale computing requirements of deep learning algorithms. Therefore, artificial intelligence chips of new architectures emerge one after another. The different technical routes of artificial intelligence chips were analyzed, the characteristics of different routes were compared, the development trend of artificial intelligence chip industry and studied, the opportunities and challenges of artificial intelligence chip development in China were analyzed, and the future development of artificial intelligence chip technology was forecasted.

Key words: artificial intelligence, chip, technical route

1　引言

当前，互联网的快速发展为人工智能（artificial intelligence，AI）提供了丰富的大数据资源，得益于数据、算法、芯片三大因素的推动，人工智能技术快速崛起。芯片是支撑人工智能产业发展的核心因素，因此谷歌（Google）等新兴互联网巨头、英特尔（Intel）等国际传统IT 企业纷纷在人工智能芯片领域投入大量精力进行研发，形成不同的技术路线。另一方面，集

收稿日期：2018−09−05；修回日期：2019−03−21

成电路芯片也是我国的“短板”领域，具有极高的战略意义。回顾集成电路的发展，一直是依靠工艺、架构和应用 3 方面来拉动的。随着摩尔定律接近极限，工艺改进已经无法降低成本。人工智能的密集计算型需求已成为当前芯片技术的主要驱动力之一。通用处理器的架构已经无法适应人工智能算法的高需求，各种新的架构成为当前处理器芯片性能提升的重要手段。GPU（图形图像处理器）、FPGA（现场可编程门阵列）、ASIC（专用集成电路）等异构芯片纷纷抢占先机，类脑神经元结构芯片的出现颠覆传统的冯诺依曼结构，给产业发展带来新的变革。

2　人工智能芯片定义与分类

当前的人工智能芯片有 3 种含义：第一种是指能处理人工智能通用任务且本身具有核心 IP（知识产权）的处理器芯片；第二种是指运行或者嵌入人工智能算法的普通处理器芯片；第三种是指具备加速语音、图像等某一项或多项任务的计算效率及迭代能力的处理器芯片。

人工智能芯片按架构体系又可以分为 CPU（通用处理器）、GPU、DSP（数字信号处理器）、FPGA、ASIC 和类脑芯片。按使用的场景可以分为云侧和端侧芯片，每一侧又可以按任务分为训练和推理，如图 1 所示。

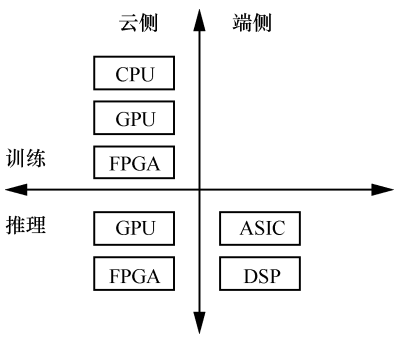

图 1　人工智能芯片按任务分类

结合上述两种分类，目前在云侧主要以训练的任务为主，芯片架构以 CPU、GPU、FPGA 为主；在端侧主要以推理的任务为主，端侧芯片无法承担巨大的运算量、性价比低，所以芯片架构主要以 ASIC、DSP 为主。

2.1　不同芯片的优缺点比较

不同架构的芯片在通用/专用性、性能、功耗方面有各自的优点和缺点。CPU 通用性最强，但时延严重、散热高、效率最低；GPU 相比其他芯片，通用性稍强、速度快、效率高，但是在神经网络的执行阶段效率低；DSP 速度快、能耗低，但是任务单一，目前成熟商品仅作为处理器 IP 核使用；FPGA 具有低能耗、高性能以及可编程等特性，相对于 CPU 与 GPU 有明显的性能与能耗优势；ASIC 可以更有针对性地进行硬件层次的优化，从而获得更好的性能。当然，ASIC 芯片的设计和制造需要大量的资金、较长的时间周期和工程周期，而且深度学习算法也在快速迭代，ASIC 类芯片一旦定制无法再次进行写操作，FPGA 具有硬件可升级、可迭代的优势。所以当前阶段，GPU 配合 CPU 将是人工智能芯片的主流，而后随着视觉、语音、深度学习的算法在 FPGA 上的不断优化，之后会固化到 ASIC 上以降低成本。

2.2　不同芯片的技术路径

当前处理器芯片主要遵循两条发展路径：一种是延续传统冯诺依曼计算架构，以加速硬件计算能力为主要目的，从通用处理器（CPU）、图像处理器（GPU）到数字信号处理器（DSP），再到半定制电路（FPGA）和全定制电路（ASIC），这 5 种类型的芯片通用性依次递减，为升级方向；另一条路径是遵循非冯诺依曼计算架构，以类脑芯片为代表，采用人脑神经元的结构来提升计算能力，但是从落地情况来看，若实现真正产业化还需要搭建生态系统，包括建立起一整套编程环境、编译器等工具。当前人工智能芯片沿着从通用到专用的方向不断演进（见表 1），未来将从专用走向另一程度的通用。

表 1　国际人工智能芯片企业

芯片架构	代表公司	发布时间	简介	专用性
CPU	英特尔	2017 年	通用计算处理器，至强系列加速深度学习处理	L1
GPU	英伟达	2016 年 4 月	目前商用最广泛的 AI 芯片，可以执行深度学习和神经网络任务	L2
DSP	新思科技	2015 年 4 月	仅作为处理器 IP 核使用	L3
FPGA	微软	2016 年 9 月	可以执行 Bing 的机器学习算法	L4
TPU	谷歌	2016 年 5 月	专为深度学习算法 TensorFlow 设计的专用集成芯片	L5
TrueNorth	IBM	2015 年 10 月	模仿人脑神经元和神经突触的结构，功耗非常低	颠覆经典冯氏架构

注：专用性从 L1 到 L5 依次提高。

2.3　不同技术路线产业情况分析

从产业的角度来看，不同技术路线的企业有不同的特点。

（1）基于 CPU 技术阵营的企业

以英特尔公司为典型代表。英特尔的优势一直在制造和集成工艺上，劣势在于 CPU 的通用架构设计，运行效率受限。当前 CPU 虽然在机器学习领域的计算大大减少，但是不会被完全取代，英特尔推出针对深度学习算法的至强处理器系列产品，企业在实际使用过程中兼顾成本和性能两方面的因素，选择 CPU 产品，因此 CPU 依然发挥着不小的作用。

（2）基于 GPU 的技术阵营

以英伟达（NVIDIA）和 AMD 为典型代表。英伟达针对深度学习算法，推出全新 Volta 架构 GPU——NVIDIA TeslaV100 GPU 计算卡。GPU 主要从事大规模并行计算，比 CPU 运行速度快，并且比其他专用 AI 处理器芯片价格低。AMD 也推出了世界上第一款 7 nm 级 GPU，专注于移动设备使用的 GPU。

对比于 CPU 和 GPU，GPU 只能完成 CPU 的部分功能，但执行速度却快很多。GPU 在某些应用如加密货币挖掘方面更具成本效益，但 CPU 却拥有更广泛的消费者基础。

（3）基于 DSP 的技术阵营

以新思科技（Synopsys）和楷登（Cadence）为典型代表。基于 DSP 进行加速器设计，如 Synopsys 公司的 EV 处理器、Cadence 公司的 Tensilica Vision 等。目前基于 DSP 的设计有一定的局限性，一般都是针对图像和计算机视觉的处理器 IP 核芯片，速度较快，成本不高。

（4）基于 FPGA 的技术阵营

代表企业有赛灵思（Xilinx）和阿尔特拉（Altera）。FPGA 具有三大优点：单位能耗比低、硬件配置灵活、架构可调整。但是，FPGA 的使用有一定门槛，要求使用者具备硬件知识。目前赛灵思和阿尔特拉公司采用最新的 CMOS 节点工艺制造 FPGA 芯片，利用先进工艺提升性能。

（5）基于 ASIC 的技术阵营

以谷歌公司为典型代表。谷歌推出张量处理单元 TPU3.0，是专用的逻辑电路，配合 TensorFlow 框架使用，当前为谷歌公司专用，还不是市场化产品。此外，当前算法架构还未完全稳定，若主流深度学习算法发生较大变化，ASIC 芯片不能像 FPGA 很快改变架构，适应变化，对企业而言成本较昂贵。

（6）颠覆经典的冯诺依曼架构路线

以 IBM 公司为典型代表。IBM 在 2016 年公布了 Truenorth（真北）芯片的详细发展计划，Truenorth 是一款基于人脑神经形态混合信号的计算机芯片。IBM 在 2016 年描述了 Truenorth 芯片的架构、评估板系列、参考系统和软件生态系统，

把计算单元作为神经元，把存储单元作为神经突触，把传输单元作为轴突。真北芯片采用的工艺是三星 28 nm 低功耗的技术，具有 4 096 个神经突触核心，每一个神经突触核心含有 256 个神经元和 64 KB 内存突触，共有 100 万个神经元和 256 万个突触，实时作业功耗仅为 70 MW。但该芯片有可能实现人工智能领域的通用化路径，但从短期来看，离大规模商业生产还有很远的距离。

3　国外企业最新发展情况

当前，由于人工智能应用场景化定制的特点以及知识产权保护等，互联网巨头、传统 IT 巨头等纷纷加大自研芯片投资力度。国际互联网四巨头——谷歌、苹果（Apple）、脸书（Facebook）、亚马逊（Amazon），国内的互联网企业如百度、阿里也纷纷开展芯片业务。如谷歌发布 TPU3.0，性能比之前提升 8 倍多，配合开源框架 TensorFlow，打造闭环生态；微软发布基于 FPGA 的低时延深度学习云平台 Project Brainwave；亚马逊定制人工智能芯片，用于未来的 Echo 设备。百度发布了面向 AI 应用的昆仑芯片，阿里巴巴公司成立平头哥半导体有限公司，注重阿里业务场景的定制化 AI 芯片，华为发布了端侧芯片麒麟 980 和云侧芯片昇腾系列。

目前，美国企业在 AI 芯片领域优势地位明显。各个企业也搭建了各自的硬件和开源平台，抢夺生态竞争权，具体见表 2。

未来主导芯片的产业生态系统有可能出现转型升级，类似谷歌、亚马逊这样的 AI 巨头，重整生态，用云服务来挤压底层硬件供应商的战略布局已经很明显。微软的 Brainwave 平台以及脸书的 PyTorch 1.0 软件和硬件都与谷歌形成了竞争关系，都希望与谷歌的 TensorFlow+TPU 进行抗衡。此外，ARM 发布了第一代面向 AI 和机器学习的处理器“Trillium”；英伟达发布了新的图灵架构。人工智能芯片已经成为国际产业竞争的新焦点。

4　我国人工智能芯片领域产业发展情况

从我国发展来看，我国在集成电路领域的技术基础较薄弱，但是在人工智能芯片学术研究上起步较早，如中国科学院寒武纪芯片 2014—2016 年间在深度学习处理器指令集上获得创新进展，在 2016 年国际计算机体系结构年会中，约 1/6 的论文引用寒武纪开展神经网络处理器研究。不仅初创企业投入足够的资金研究，华为、百度等公司也参与建设布局，积极抢位。另一方面，面对垂直细分领域的 AI 芯片市场前景广阔。随着人工智能应用场景的细分市场越来越多，专门为某些应用场景定制的芯片性能优于通用芯片，终端芯片呈现碎片化、多样化的特点，并且目前尚未形成市场垄断，我国公司仍然有较多的机会。我国人工智能芯片企业布局见表 3。

表 2　国际典型企业人工智能生态建设情况

代表企业	开源框架	硬件	代表产品	应用
谷歌	TensorFlow	TPU	AlphaGo	移动应用
IBM	SystemML	TrueNorth	Waston	医疗服务
微软	ML.NET	FPGA	小冰、小娜	私人助手
脸书	PyTorch	在研	Chatbot	聊天机器人
亚马逊	MXNET	Inferentia	Echo	智能音箱
苹果	Turi Create	Apple Neural Engine	Siri	移动应用

表 3　我国人工智能芯片企业布局

代表企业	发布时间	AI 芯片架构	简介
寒武纪	2016 年	ASIC	专门为深度学习设计的核心处理器芯片
地平线	2016 年 4 月	FPGA/ASIC	基于深度神经网络的人工智能“大脑”平台的芯片
中星微	2016 年 6 月	DSP	嵌入式神经网络处理器 NPU
深鉴科技	2016 年	FPGA	深度学习处理器 DPU
百度	2018 年 7 月	昆仑芯片	云端全功能 AI 芯片
华为	2018 年 10 月	昇腾	全栈解决方案

但是我国在面对机遇的同时也面临诸多挑战，首先，云端市场龙头企业分布在国外，我国云端芯片与国外技术差距巨大，国外云端市场技术及生态构建成熟、优势大。我国专注云端芯片的企业较少，且尚未形成生态影响力。另一方面，我国不同企业呈现整体追逐热点快、基础不牢、后续乏力的情况。如我国从事人工智能开发处理器的初创企业有 45 家，但是基本都从事语音、视觉芯片的集成研发，定位重叠较多，并且我国目前尚未形成有影响力的芯片–平台–应用的生态。

5　结束语

未来10年是人工智能产业发展和突破的关键时期，也是人工智能芯片技术发展的重要时期。

现阶段人工智能应用对算力的需求体现在两方面，一是深度学习算法包括大量的卷积、残差网络、全连接等计算需求，在摩尔定律接近物理极限、工艺性能提升对计算能力升级性价比日益降低的前提下，仅基于工艺节点的演进已经无法满足算力快速增长的需求；二是深度学习需要对海量数据样本进行处理，强调芯片的高并行计算能力，同时大量数据搬运操作意味着对内存存取带宽的高要求，而对内存进行读写操作尤其是对片外内存进行读写访问的消耗的功耗要远大于计算的功耗，因而高能效的内存读写架构设计对芯片至关重要。

因此，一方面，从技术角度来看，对芯片架构的改进将成为提升芯片性能的主要手段，从各个企业的产品来看，也是以不同架构的升级来迭代芯片的性能为主要手段。另一方面，改善计算单元和存储单元高速的通信需求也将成为提升性能的重要趋势。

从应用的角度来看，终端芯片形态多样，如安防摄像头、智能音箱、智能机器人、智能手机等，该类任务计算量小，但是实时性要求高，注重芯片的能耗、散热、单位能耗比等指标。

从生态角度来看，软硬件协同优化已成为企业提升技术能力的主要方式，单纯的数据与算法优化已经不能满足企业的需求，需要企业采用芯片结合算法模型的方法进行优化迭代。

参考文献：

[1] JOUPPI N P. In-datacenter performance analysis of a tensor processing unit[Z]. 2017.

[2] CHEN Y H, KRISHNA T, EMER J S, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks[J]. IEEE Journal of Solid-State Circuits, 2017, 52(1): 127-138.

[3] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015(5): 436-444.

[4] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems[J]. arXiv:1603.04467.

[5] SUNDARARAJAN N, SARATCHANDRAN P. Parallel architectures for artificial networks: paradigms and implementations[M]. New York: ACM Press, 1998: 23-27.

[6] NVIDIA DGX-1. Deep learning[Z]. 2018.

[7] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[8] BENJAMIN B V, GAO P, MCQUINN E, et al. Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations[J]. Proceedings of the IEEE, 2014, 102(5): 699-716.

[9] DIKOV G, FIROUZI M, ROHRBEIN F, et al. Spiking cooperative stereo-matching at 2 ms latency with neuromorphic hardware[M]//Biomimetic and Biohybrid Systems. Berlin: Springer, 2017: 119-137.

[10] 施羽暇. 人工智能芯片技术研究[J]. 电信网技术, 2016(12): 11-13.
SHI Y X. Research on artificial intelligence process chip technology[J]. Network Technology, 2016(12): 11-13.

[11] 施羽暇. 促进人工智能芯片发展, 助力产业基础实力提升[J]. 信息通信技术与政策, 2018(11): 76-79.
SHI Y X. Promoting the development of artificial intelligence chips, enhancing the industrial foundation[J]. Information and Communications Technology and Policy, 2018(11): 76-79.

[作者简介]

施羽暇（1984– ），女，博士，中国信息通信研究院高级工程师，主要研究方向为 ICT 产业、人工智能、集成电路等，参与了“互联网+”、人工智能一系列国家政策的制定。