DOI：10.19816/j.cnki.10-1594/TN.2021.04.087

AI芯片的落地场景和关键技术分析

安宝磊

（北京比特大陆科技有限公司 北京 100029）

摘　要：人工智能（AI）技术正处于黄金发展时期，AI相关的研究和应用对计算能力提出了很高的要求和挑战，因此AI芯片技术在近几年成为了研究的热点，很多公司和科研机构都发布了相关的产品，AI芯片也从技术验证进阶到工业场景落地阶段。本文将对AI芯片主要的落地场景和所需的关键技术做介绍和分析。

关键词：集成电路；人工智能；AI芯片；专用集成电路

中图分类号：TP18;TN409　　**文献标识码**：A　　**国家标准学科分类代码**：510

Analysis of the landing scenarios and key technologies of AI chip

AN Baolei

（*Beijing Bitmain Technology Co., Ltd., Beijing 100029, China*）

Abstract: Artificial intelligence technology is in the golden age of development. The research and application of AI put forward high requirements and challenges for computing power. Therefore, AI chip technology has become a research hotspot in recent years. Many companies and scientific research institutions have released AI products, and AI chip has also advanced from technical verification to landing stage. It will introduce and analyze the main landing scenarios and key technologies of AI chip in this paper.

Keywords: IC; artificial intelligence; AI chip; ASIC

0　引　言

深度学习给计算机视觉、自然语言处理、个性化推荐、语音识别等领域都带来了突破性的进展[1-2]，在计算机视觉方向，图像分类[3]、目标检测[4]的准确率已经达到甚至超过人类的水平。深度学习神经网络需要大量的矩阵乘加运算，对于硬件计算算力有很高的要求，CPU和传统计算架构无法满足对于并行计算能力的需求，需要特殊定制的人工智能（artificial intelligence, AI）芯片[5]。2016年以来，国外的英伟达、谷歌、亚马逊、特斯拉，国内华为、百度、比特大陆、寒武纪等公司都开始了AI芯片相关的研发[6-7]，其中一些公司的AI芯片已经迭代了三代以上，并在市场上进行了一段时间的推广，完成了试点项目。可以说AI芯片已经跟随着人工智能产业的发展进入到了场景化落地的阶段。

1　AI芯片功能和技术架构

1.1　AI芯片功能

深度学习由训练（training）和推理（inference）两个任务组成，因此AI芯片的主要功能就是训练和推理[8]。训练和推理对于芯片的要求存在一定的差别，一般情况下，面向训练和推理的芯片会分别进行设计。

1）训练

训练任务是对大量的数据在平台上进行学习，并形成具备特定功能的神经网络模型。训练芯片的目的是让算法研究人员可以快速的验证算法方案和

安宝磊（通信作者），人工智能资深专家，主要研究方向为嵌入式人工智能。E-mail：anbaolei1984@163.com

观测试验结果，对AI芯片有高算力、高容量、高传输速率和通用性的要求。

2）推理

推理任务是指基于已经训练好的模型，针对输入数据计算结果。需要根据实际应用场景对推理芯片的算力、功耗、工业等级、价格成本等因素做综合考量。华为训练和推理AI芯片的主要技术规格如表1所示。

表1 华为昇腾芯片主要规格

Table 1 Huawei Ascend AI chip specifications

芯片	昇腾 Ascend910	昇腾 Ascend310
功能	训练	推理
工艺/nm	7	12
算力	INT8 640TOPS FP16 320TFLOPS	INT8 22TOPS FP16 11TFLOPS
功耗/W	310	8
内存	HBM2E	2 * LPDDR4x

1.2 AI芯片技术架构

目前一般认为AI芯片是针对AI算法做了特殊加速设计的芯片，所以传统CPU架构并不属于AI芯片的范畴。从技术架构上AI芯片分为GPU、FPGA、专用集成电路、类脑芯片等[9]，如表2所示。

2 AI芯片应用和落地场景

AI芯片按着部署的位置可分为云、边缘和端侧3类[10]。

2.1 云侧AI芯片和硬件

云侧主要指部署在数据中心的服务器群，AI硬件以PCIe加速卡的形式插在服务器上，用来提供AI计算的算力，一台服务器通常可以插多张加速卡。云端训练的加速卡如英伟达A100功耗都在250 W以上，而推理卡一般都会将功耗限制在PCIe供电的峰值75 W，无需额外供电。如果单颗芯片的功耗控制的比较理想，可以在一张卡里放置多颗芯片。比如华为的Atlas300推理卡（如图1所示）里安放了4颗昇腾310芯片（单颗芯片功耗8 W），比特大陆的SC5+推理卡里放有3颗BM1684芯片（单颗芯片功耗16 W）。

表2 AI芯片技术架构

Table 2 Technical architecture

技术架构	优点	缺点	代表产品
图形处理器(GPU)	编程灵活性高，相比CPU，有更高的并行计算能力。有最成熟的软件生态。	相对于FPGA和ASIC，价格和功耗过高。	训练：英伟达A100、V100 推理：英伟达T4、Xavier NX
现场可编程门阵列(FPGA)	半定制，可对芯片硬件层进行编程和配置。相对于GPU有更低的功耗。	硬件编程语言难以掌握，功耗和成本有进一步压缩空间。	赛灵思（Xilinx）
专用集成电路(ASIC)	针对专门的任务进行定制，可实现低成本、低功耗、高性能。	芯片通用性差，可编程架构设计难度高、研发投入大。	训练：华为昇腾910 推理：华为昇腾310、比特大陆BM1684、寒武纪MLU270等
类脑芯片	突破冯·诺伊曼架构瓶颈，用芯片去模拟大脑神经网络的结构来达到最优的性能和功耗	尚未成熟，还处在实验室阶段	IBM TrueNorth、斯坦福Nurogrid

图1 华为推理板卡Atlas300

Fig.1 Huawei Atlas300

芯片在云侧应用最多的场景就是在各大互联网公司的数据中心，用来做视频内容审查、个性化推荐、语音识别等业务。另外，政府的智慧城市等项目，都会建设AI计算中心作为算力支撑，来处理海量视频数据的结构化等业务。

2.2 边缘侧AI芯片和硬件

相对于云计算，边缘计算在延时、可靠性、成本、部署便利等方面有明显的优势[11]。边缘计算使用的AI硬件以盒子和模组的形态为主，AI芯片工作在片上系统（SoC）模式，除了AI计算以外还可以在芯片的CPU上实现业务应用。因为需要处理多路视频

流,边缘侧AI芯片都会配备视频编解码和图像处理硬件加速单元,以比特大陆的边缘计算盒SE5为例,如图2所示,可以支持30路以上1080 P高清视频的硬解码,17.6 Tops的算力足够支持16路人脸识别业务。

图2　比特大陆AI计算盒SE5

Fig.2　Bitmain SE5

边缘计算模组面向的是有硬件定制开发需求的场景,如在无人机、机器人中的AI计算模块,模组设计需要具备接口丰富、体积小、低功耗、可宽温工作、高集成度等特点。

随着AI应用在各行各业的赋能,边缘计算的落地场景也日益增多,总结了当前一些常见的功能需求,很多边缘AI硬件产品里也集成了这些功能的算法和应用,从而可以作为完整的解决方案提供给用户使用。

1)工地和工厂安全生产:包括安全帽佩戴监测、工服穿着监测、明烟明火监测(如图3所示)等功能。

图3　明烟明火监测

Fig.3　Fire detection

2)工业质检:生产缺陷和瑕疵检测、OCR检测、尺寸测量等。

3)智能加油站:使用人脸、车辆和号牌识别等技术实现人员离岗检测、占用车位检测、入口拥堵检测、打手机和抽烟检测等。

4)明厨亮灶:厨师帽佩戴监测、厨师服穿着监测、口罩识别等。

5)电力巡检:刀闸状态识别、仪表智能分析、操作规范监测等。

2.3　端侧AI芯片和硬件

在手机、摄像头、汽车等终端设备中使用的AI芯片。端侧AI芯片对于功耗的要求非常严格,通常不超过5 W。对于芯片的算力需求相对于云端和边缘有所降低,一般都在10 Tops以内[12]。面向安防应用的终端芯片需要支持高清视频编码和图像处理(ISP)功能。可穿戴设备、智能音箱等市场也对端侧AI芯片有不断增长的需求。

3　AI芯片关键技术

3.1　面向应用场景设计芯片规格

通过前文的分析可以看到,不同的应用场景对于芯片的算力、功耗、成本、视频图像处理能力、工作温度等要求存在很大的差异。在最初设计芯片规格时就要明确芯片设计所面向的应用场景和场景对于芯片具体的需求,能够让芯片设计的制程工艺、MAC阵列、存储器、IP核、接口的选型和配比更加准确和合理。

3.2　工具链易用性和编译器优化

英伟达GPU之所以占据领先的市场份额的主要原因之一就是具备成熟的软件生态,支持所有主流深度学习框架(PyTorch\TensorFlow\PaddlePaddle\MxNet)。研发一套成熟易用的工具链软件对于AI芯片公司来说是非常重要的,在深度学习框架支持、算子完备性、自定义算子支持、性能剖析、错误信息提示和调试等方面做的越好,越有助于产品被市场认可和推广[13]。

在评估一颗AI芯片的时候,不能简单的只看芯片规格中的算力峰值Tops,推荐的评估方法是用经典网络模型(Resnet50,Mobilenet v2最为常用)在芯片上实测,主要测试指标包括:延时(ms)、吞吐量(image/s)、算力利用率(image/s/T)、算力功耗比(image/s/w),以评测出芯片真实的表现。其中算力利用率考察的是编译器优化能力,算力利用率越高,表示编译器优化的越好。编译器主要通过图优化、芯片

架构相关优化等技术，驱动模型在芯片上的高效执行。

3.3 低精度量化技术

模型低精度量化可以有效减少模型大小，加速模型的推理速度，在学术和工业界已经被广泛研究和应用，目前主要使用float16和int8两种计算精度，int8精度需要使用量化技术[14]，为了使量化模型的精度相对于浮点模型的损失尽可能小，研发人员要在量化方法和实现方式上做很多的工作。当前通过模拟训练量化、非对称量化、per-channel量化、混合精度计算等技术都让精度有了一定的提升。用户在芯片选型时，也要验证量化后的模型在芯片上实测的精度表现。

4 结 论

通过近几年的不断发展，AI芯片技术已经进入到了工业落地阶段，根据应用场景来设计芯片已经成为趋势，用户在选型时也要选取适合自己产品场景需求的芯片。对于芯片算力不能只看厂商标称的算力峰值，而要做实际测试。本文对AI芯片的落地场景和关键技术做了调研和分析，希望可以帮助读者更加深入的了解AI芯片的设计和研发过程。AI芯片的发展日新月异，谷歌甚至已经开始让AI自己来设计芯片，一些公司也开始研究基于RISC-V的芯片解决方案[15]，以适应AI时代的需要。

参考文献

[1] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998 , 86(11): 2278-2324.

[2] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]. Proc of the 25th International Conference on Machine Learning. ACM Press, 2008: 1096-1103.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, ACM, 2012: 1097-1105.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

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

[6] 清华大学 . 人工智能芯片技术白皮书 (2018)[R/OL]. (2018-12-11)[2010-01-20]. https://www.tsinghua.edu.cn/publish/thunews/9659/2018/20181217102627644168087/20181217102627644168087_.html.
Tsinghua University. 2018 white paper on AI chip technologies[R/OL]. (2018-12-11)[2010-01-20]. https:// www.tsinghua. edu. cn/publish/thunews/9659/2018/20181217102627644168087/20181217102627644168087_.html.

[7] 安宝磊 . AI芯片发展现状及前景分析[J]. 微纳电子与智能制造, 2020, 2(1): 91-94.
AN B L. Analysis of the development and prospect of AI chips[J].Micro/nano Electronics and Intelligent Manufacturing, 2020, 2(1): 91-94.

[8] 尹首一, 郭珩, 魏少军 . 人工智能芯片发展的现状及趋势[J]. 科技导报, 2018, 17: 45-51.
YIN SH Y, GUO H, WEI SH J. Present situation and future trend of artificial intelligence chips[J]. Science & Technology Review, 2018, 17: 45-51.

[9] 尹首一 . 人工智能芯片概述[J]. 微纳电子与智能制造, 2019, 2: 7-11.
YIN SH Y. Overview of artificial intelligence chip[J]. Micro/nano Electronics and Intelligent Manufacturing, 2019, 2: 7-11.

[10] 赵春昊 . 基于应用场景的人工智能芯片技术分类方法研究[J]. 智能计算机与应用, 2020, 10(9): 225-228.
ZHAO CH H. A classification method of AI chip based on application scenarios[J]. Intelligent Computer and Applications , 2020, 10(9): 225-228.

[11] 张展, 张宪琦, 左德承, 等 . 面向边缘计算的目标追踪应用部署策略研究[J]. 软件学报, 2020, 31(9): 71-88.
ZHANG ZH, ZHANG X Q, ZUO D CH, et al. Target tracking application deployment strategy for edge computing[J]. Journal of Software, 2020, 31(9): 71-88.

[12] 谭洪贺, 余凯. 端侧AI芯片的挑战和展望[J]. 人工智能, 2018, 2: 113-121.
TAN H H, YU K. The challenge and prospect of edge AI chip[J]. Artificial Intelligence, 2018, 2: 113-121.

[13] LIN W F, TSAI D Y, TANG L, et al. ONNC: A compilation framework connecting ONNX to proprietary deep learning accelerators[C]. 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), IEEE, 2019.

[14] 黄钲喆, 杜慧敏, 常立博. 卷积神经网络混合截断量化[J]. 计算机辅助设计与图形学学报, 2021, 33(4): 553-559.
HUANG ZH ZH, DU H M, CHANG L B. Mixed-clipping quantization for convolutional neural networks[J]. Journal of Computer-Aided Design & Computer Graphics , 2021, 33(4):553-559.

[15] COCOCCIONI M, ROSSI F, RUFFALDI E, et al. Vectorizing posit operations on RISC-V for faster deep neural networks: Experiments and comparison with ARM SVE[J]. Neural Computing and Applications, 2021(33): 10575-10585.