第 30 卷第 6 期　　功能材料与器件学报　　Vol. 30, No. 6
2024 年 12 月　　JOURNAL OF FUNCTIONAL MATERIALS AND DEVICES　　Dec., 2024

文章编号: 1007－4252(2024)06－0300－10　　DOI: 10.20027/j.gncq.2024.0042

低功耗人工智能计算系统研究进展综述

陈　华[1,2], 曲益明[1], 吴文豪[1,2], 赵　毅[1,2,3,*]

（1. 华东师范大学集成电路科学与工程学院, 上海 200241; 2. 中国电子科技南湖研究院, 嘉兴 314001; 3. 浙江大学信息与电子工程学院, 杭州 310027）

摘要: 最近, 随着大数据和硬件能力的快速增长, 人工智能取得了显著发展, 人工神经网络(Artificial Neural Network, ANN)已被成功应用于解决学术界和工业界的许多问题。然而, 在边缘设备上部署人工神经网络仍具有挑战性。这些场景一般对功率、体积有严格的限制, 同时对系统延迟和实时性有较高要求, 因此构建低功耗人工智能计算系统需要在性能、功率、体积之间进行权衡。本文综述了目前低功耗人工智能计算系统的研究现状, 介绍和分析了低功耗人工智能计算硬件和软件工具, 阐述了存在的技术挑战, 讨论了系统的评估方法和指标, 并展望了未来发展趋势。

关键词: 嵌入式人工智能; 边缘人工智能; 低功耗人工智能; 深度学习; 神经形态

中图分类号: TP368　　**文献标志码**: A

A review of progresses in low-power artificial intelligence computing systems

CHEN Hua[1,2], QU Yiming[1], WU Wenhao[1,2], ZHAO Yi[1,2,3,*]

(1. School of Integrated Circuits, East China Normal University, Shanghai 200241, China; 2. China Nanhu Academy of Electronics and Information Technology, Jiaxing 314001, China; 3. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: Recently, with the rapid growth of big data and hardware capabilities, artificial intelligence (AI) has achieved significant development. Artificial neural networks (ANNs) have been successfully applied to solve numerous problems in academia and industry. However, deploying AI networks on edge devices remains challenging. These scenarios generally have strict limitations on power and size, while also have high requirements for system latency and real-time performance. Thus, building low-power AI computing systems involves making trade-offs among performance, power, and size. This article reviews the current state of low-power AI computing systems, introduces low-power AI computing hardware and

收稿日期: 2024-07-24;　　**修订日期**: 2024-08-12

基金项目: 科技创新 2030-"新一代人工智能"重大项目 (No.2020AAA0109001); 国家自然科学基金资助项目 (No.U23B2040).

作者简介: 陈　华（1983— ）, 男, 学士, 工程师, 主要研究方向为嵌入式系统架构、硬件设计（E-mail: chenhua@cnaeit.com）

通信作者: 赵　毅 (1977—), 男, 博士, 教授, 主要研究方向为面向存储融合架构的器件与芯片 (E-mail: yizhao@zju.edu.cn)

software tools, elaborates on existing technical challenges, discusses system evaluation methods and metrics, and looks to future development trends.

Key words: embedded AI; edge AI; low-power AI; deep learning; neuromorphic

0 引言

近年来,人工智能技术取得了快速发展。尤其是随着深度学习的普及,人工智能应用日益丰富,在安防 [1]、机器人 [2]、医疗 [3]、自动驾驶 [4] 等多个领域实现了应用落地。同时,人工智能算法对算力、能效、数据传输带宽等提出更高的要求。另外,由于摩尔定律放缓,只依靠通用的中央处理器(Central Processing Unit, CPU)、图形处理器(Graphics Processing Unit, GPU)的计算系统已无法满足需要,专用的人工智能加速硬件成为研究热点。为满足不同的应用需求,通常将人工智能系统分为云计算、边缘计算、端侧计算三大类。这三类系统具备不同的硬件资源以完成不同的任务,云计算主要用于大规模数据的集中处理、模型训练、大模型推理等算力要求高的任务,边缘和端侧设备则更适合小规模的智能分析和本地推理任务。

在边端设备上部署人工智能算法具有显著的优势,因其不需要将数据从传感器传输到云端处理,节省了传输带宽,降低了系统延迟,还具有更好的隐私性。但同时,边端智能应用也会带来新的设计挑战,在大多数情况下系统需要在特定的延迟内完成实时处理,功耗和面积也有严格的限制。典型的应用如微型无人机、机器人等,对智能处理系统功率的要求一般在瓦级,待机时功耗甚至在几十毫瓦内。因此,边端人工智能计算系统需要在性能、功率、成本和面积之间进行权衡。

本文介绍了当前用于边缘侧和端侧的低功耗人工智能计算系统。与之前关于人工智能加速器的综述不同 [5,6,7],本文重点关注功耗≤2 W 的低功耗人工智能计算系统,包括硬件平台、软件工具等。本文剩余章节安排如下:第 1 节主要介绍低功耗人工智能计算硬件;第 2 节介绍用于低功耗人工智能计算的软件工具;第 3 节讨论系统评估指标;第 4 节总结全文并展望未来的发展趋势。

1 低功耗人工智能计算硬件

目前有多种硬件可用来构建人工智能计算系统,如 CPU、GPU、现场可编程门阵列(Field Programmable Gate Array, FPGA)、专用集成电路(Application Specific Integrated Circuit, ASIC)和片上系统(System On Chip, SOC)等。CPU 和 GPU 是通用处理器架构的代表,有着非常广泛的应用,软件工具成熟,易于使用,但硬件没有针对人工智能应用进行优化,存在一定局限。如 CPU 计算资源欠缺;GPU 有大量的并行计算资源,但能效不高。ASIC 是专为人工智能应用设计的定制硬件,一般也称为神经网络处理单元(Neural Network Unit, NPU),其性能优异,运行效率高,但通用性不强,开发周期长,成本高;FPGA 介于两者之间,可以在一定程度上定制加速硬件,无需流片,开发周期和成本适中;另外还有一类可重构计算硬件,其可编程性介于 FPGA 和 ASIC 之间,兼具两者的特点。

由于人工智能应用具有计算密集和存储密集的特点,学术界和工业界涌现出了多种新架构和新技术,如近存计算 [8]、存内计算(Processing in Memory, PIM)[9] 等,进一步提升了计算效率。此外,以脉冲神经网络(Spiking Neural Network, SNN)为代表的第三代神经网络模型兴起,已有实际应用 [10]。SNN 建立在脉冲神经元的基础上,模拟了神经元和突触状态,更接近于生物神经元,具备更低的功耗,有较大发展潜力。

具体到本文研究的低功耗人工智能计算平台,在实际应用场景中,系统往往要处理多种复杂任务,单一的硬件架构难以满足要求,因此 SOC 是发展趋势,专用的人工智能加速 ASIC 一般和 CPU 一起构成 SOC 来使用。为方便讨论,本文中我们将 SOC 和 ASIC 归为一类。

1.1 基于 CPU 架构的硬件

基于 CPU 的低功耗人工智能计算硬件主要包括微处理器(Micro Processor Unit, MPU)和微控制器(Micro Controller Unit, MCU),作为通用处理器,有大量的商用产品。

典型的 MPU 如 ARM Cortex-A7 处理器,在 28 nm 制造工艺下,运行在 1.2-1.6 GHz 频率下,典型功耗约 100 mW[11]。商用产品一般以 SOC 形

式提供，组成系统时还需要外扩存储器，系统功耗一般小于 2 W。

MCU 因其功耗较低，更适合应用在低功耗场景；同时其架构多样，如 Cortex-M4、Cortex-M55、Cortex-M7、RISC-V。目前有大量商用产品，如 ambiq 公司的 Apollo3[12]、Apollo4[13]，ST 公司的 STM32L4[14]、STM32H7[15]。这些微控制器内置存储器，可提供大约几 MOPS 至几 GOPS 的算力，系统峰值功耗可控制在 1 W 以内，典型运行功耗在几十 mW 至几百 mW。

1.2 基于 GPU 架构的硬件

GPU 包括大量算术逻辑单元（Arithmetic Logic Unit, ALU），这种并行架构十分适合人工智能这种计算密集型应用，但同时也会消耗大量的能耗，在低功耗场景下的使用存在一些限制。

目前 GPU 的主流厂商是 NVIDIA，其拥有极高的市场占有率。NVIDIA 先后发布了多个 GPU 架构，如 G80、Volta、Turing 和 Ampere。截至目前，NVIDIA 针对嵌入式领域推出的功耗最低的产品是 Jetson Nano 模组 [16]，该产品由 128 核 GPU 和四核 ARM Cortex-A57 CPU 组成，AI 算力达到 472GFLOPS，结合 NVIDIA Jetpack SDK 可运行大多数 NN 后端和框架，最小功耗约为 5W。

1.3 专用集成电路和片上系统

相比于 CPU 和 GPU 这些通用处理器，ASIC 具有更好的性能和更高的能效。表 1 列出了功耗≤2W 的具有代表性的 ASIC 和 SOC 产品。

表 1 典型的功耗≤2 W 的 SOC 和 ASIC 产品

Table1 List of typical SOC and ASIC products with power consumption ≤2 W

型号	公司/机构	硬件架构	峰值算力（TOPS）	功耗（mW）	能效（TOPS/W）	支持的网络类型
NDP200	Syntiant	MCU+DSP+PIM	0.006	1	6.40	CNN、RNN
Max 78000	Analog Devices	MCU+NPU	0.056	28	2.00	CNN、RNN
Ergo	Perceive	CPU+NPU	4.000	72.7	55.02	CNN、RNN
RV1126	Rockchip	CPU+NPU	2.000	201	9.95	CNN、RNN
TrueNorth	IBM	neuromorphic	1.890	500	3.78	SNN
PuDianNao	Institute for Computing Technology	NPU	1.060	596	1.78	k-means, k-nearest neighbors, naive bayes, SVM, linear regression, classification tree, DNN
Tianjic	Tsinghua	neuromorphic	1.210	950	1.27	ANN、SNN
MyriadX	Intel Movidius	NPU	4.000	1 500	2.67	CNN、RNN
Hi3516DV500	Hisilicon	CPU+NPU	2.000	2 000	1.00	CNN、RNN
TPU Edge	Google	TPU	4.000	2 000	2.00	CNN、RNN
Journey2	Horizon Robotics	NPU	4.000	2 000	2.00	CNN、RNN

Syntiant 推出的 NDP200[17] 可以在 1 mW 的功耗下提供高度精确的推理，其核心架构由 Syntiant 加速器和嵌入式 ARM Cortex-M0 处理器组成，芯片在 100 MHz 的工作频率下运行。Syntiant 加速器是一个超低功耗、高度灵活的深度学习推理引擎，同时内置 HiFi DSP，可以较低的功耗同时运行多个应用，包括人工智能视觉功能、多传感器融合、语音命令识别、声音事件检测，以及其它音频、运动和压力感应等应用。

Analog Devices Inc 推出的 MAX78000[18]，由

两个低功耗内核 ARM Cortex-M4 和 RISC-V、基于浮点处理单元的微控制器和卷积神经网络加速器组成，可以最大限度地减少卷积神经网络(Convolutional Neural Network, CNN)的功耗和延迟。其后，又陆续推出了 MAX78002 等产品，具有更高的算力。

Perceive 推出的 ERGO 产品 [19]，内置 ARC EM5D RISC/DSP 处理器和多个神经网络加速器。该芯片的计算性能相当于 4TOPS，每瓦功耗可达 55 TOPS/W，同时支持多种类型的神经网络，包括卷积神经网络和递归神经网络(Recurrent Neural Network, RNN)。该产品具有广泛的应用，适用于摄像头、智能门铃和锁、家用机器人、运动相机、无人机以及其它消费类和企业级产品。

RV1126[20] 是 Rockchip 推出的智能视觉 SOC 芯片，采用 14 nm 制程工艺制造，基于四核 ARM Cortex-A7 内核，内置 2T 算力 NPU，支持 4K30FPS H.264/H.265 视频编解码。其自研图像处理单元可实现多级降噪、HDR 等功能。同时，内置 HDAEC 算法，支持麦克语音阵列，可有效增强声音采集及拾音距离。

TrueNorth[21] 是 IBM 参与 DARPA 的研究项目 SyNapse 的成果，单颗芯片集成了 100 万个“神经元”，256 个“突触”，4 096 个并行分布的神经内核，是专注于 SNN 的硬件加速器。

PuDianNao[22] 是一款定制 ASIC，是 DianNao[23] 系列中针对嵌入设备的加速器，具有改进的内存访问功能，可实现低延迟和低功耗。其内部可实现 7 种常用的机器学习算法：k-means、k-nearest neighbors、naive bayes、support vector machine、linear regression、DNN。

Tijic[24] 是同时支持 SNN 和 CNN、RNN 的加速器，采用了众核架构，每个核可以配置成 SNN 单元或 CNN / RNN 单元。另外，单个核还可以配置为兼容模式，即接受 SNN 的输入，并在计算后转化为 CNN / RNN 的输出；或者反之，将 CNN、RNN 网络的输入转化为 SNN 输出。

Movidius Myriad X[25] 是一款视觉处理 SOC，支持 8 位和 16 位整数以及 16 位浮点量化，可以接入多达 8 路高清传感器输入。通过并行处理和最小化数据移动，它可以在图像、视觉智能处理任务中实现 4TOPS 的峰值性能。

Hi3516DV500[26] 是 Hisilicon 推出的面向视觉应用的智能 SOC，其内置双核 A55，集成了高效的神经网络推理引擎，算力最高可达 2TOPS。该芯片最高支持 2 路 sensor 输入，支持最高 5M@30fps 的 ISP 图像处理能力，支持宽动态、多级降噪、六轴防抖、多光谱融合等多种传统图像增强和处理算法。

谷歌开发的 Coral edge TPU[27] 是谷歌为加速基于边缘设备的神经网络推理而研发的专用芯片，峰值算力为 4TOPS，能效可达 2TOPS/W。Edge TPU 主要用于推理加速和轻量级迁移学习，并支持常用的深度学习框架(如 TF-Lite)和神经网络模型。

Journey 2[28] 是 Horizon Robotics 面向汽车和边缘智能推出的一款高性能 SOC 芯片，CPU 为双核 ARM Cortex-A53，最高运行频率 1 GHz，同时采用 Horizon Robotics 专有的深度学习计算“BPU”核心，可以在 2 W 的功耗下提供了 4TOPS 算力，芯片通过了 AEC-Q100 Grade 2 车规级认证。

1.4　现场可编程门阵列和可重构计算硬件

可编程架构因具有灵活性的优点而成为快速原型开发的主要选择，从可编程粒度的角度又可分为细粒度和粗粒度两类。FPGA[29] 是一种细粒度可编程的体系结构，通过硬件描述语言(Hardware Description Language, HDL)进行编程，可实现逻辑门的编程重构。可重构计算 [30](Coarse-grained Reconfigurable Architecture, CGRA)则是粗粒度可编程的代表，通常对其处理单元(Processing Element, PE)进行编程重构，一般能在更高的时钟频率下运行。

1) FPGA: FPGA 包含可编程逻辑块和可配置互连，使用 HDL 语言描述和定义逻辑电路和互连，可以灵活更改设计。由于人工智能的快速发展，FPGA 被广泛用于人工智能硬件加速的快速定制和性能评估 [31]。例如，Lite-CNN[32] 采用 INT8 量化方法在 AMD ZYNQ XC7Z020 FPGA 中实现了卷积神经网络的加速，算力达到 410GOPS。同时，使用 FPGA 构建人工智能应用也存在诸多挑战：a)通常 FPGA 片上存储容量较小，不可避免地需要外部存储器，数据传输带宽和功耗会限制系统性能；b)乘法器资源偏少，限制了系统算力提升；c)可编程的互连会消耗大量的芯片面积和功耗，同时实现

复杂电路时其关键路径长,限制了工作频率。当前主流的 FPGA 厂商(如 Xilinx)已在片上集成专用的人工智能加速硬件,往异构 SOC 的方向发展。

2) 可重构计算硬件:可重构计算硬件一般由一系列网状互连的 PE、寄存器和共享存储器组成。PE 一般由标准单元电路构建,因此相比于 FPGA,牺牲了可重配置性,但效率大大提升。Thinker[33]是 CGRA 的典型代表,其架构如图 1 所示。通过对 PE 的指令级动态配置,可实现 PE 阵列不同部分执行不同运算的异构并行计算能力;同时通过优化的数据重排模块,以较小的带宽实现了数据的传输;还引入了稀疏化处理,可在输入数据为 0 的情况下关闭时钟以节省功耗。Thinker 芯片使用 65 nm 工艺制造,能够在 386 mW 的功耗下实现 409.6 GOPS 的算力。

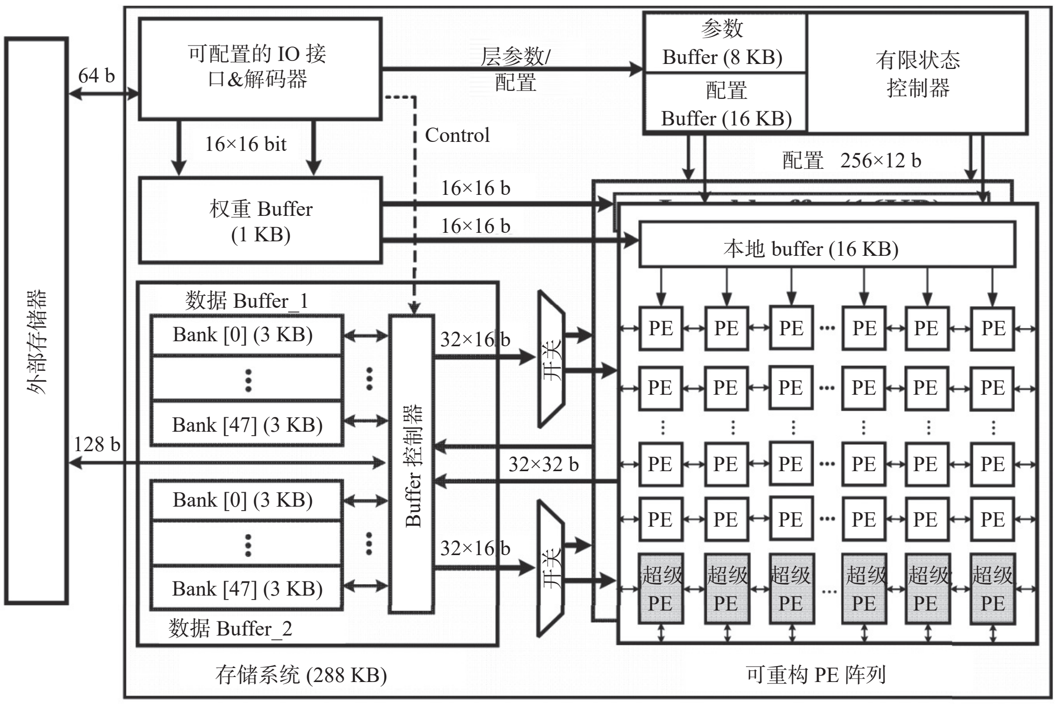

图 1 Thinker 芯片整体架构图

Fig. 1 Architecture of Thinker chip

2 低功耗人工智能计算软件工具

在端侧设备上部署人工神经网络的典型流程如图 2 所示,主要包括训练、模型转换和优化以及部署运行等阶段。训练的主流工具是 TensorFlow[34]和 PyTorch[35]。当训练出的模型部署到 MCU、SOC 等低功耗硬件平台时,需要部署

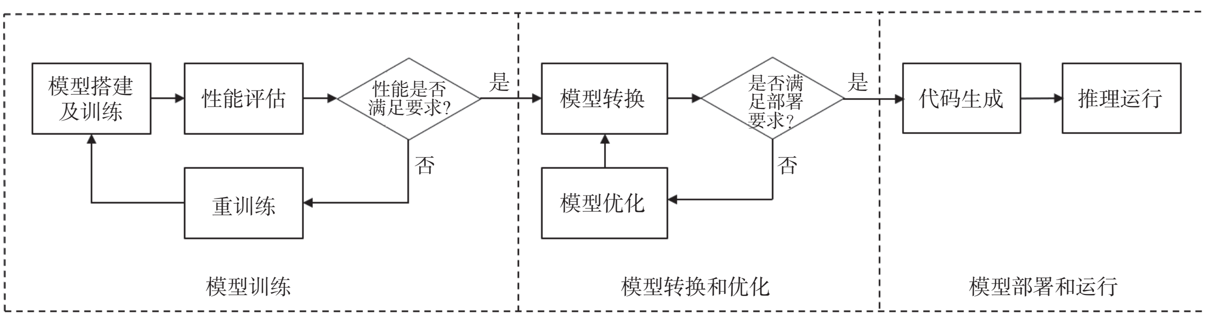

图 2 神经网络在端侧部署的典型流程

Fig. 2 Typical flowchart of neural network deployment on the edge devices

工具支持。通常情况下, 部署工具指 AI 编译器 [36], 或者 AI 推理框架, 一般需要具备模型的转换、优化、加载和执行功能。相较于云计算应用, 在低功耗人工智能硬件上部署神经网络模型存在着诸多挑战:

1) 硬件资源受限: 低功耗人工智能的计算能力和存储空间通常比较有限。以 MCU 为例, 其核心时钟频率一般在几十兆赫兹至两三百赫兹, 片内静态存储空间一般小于 1MB; 即使是 SOC 硬件, 受限于功耗等因素, 其算力一般也小于 4TOPS。

2) 硬件异构特性: 低功耗人工智能计算硬件覆盖了从 MCU 和 FPGA 到特定的 AISC 和 SOC, 以及可重构计算硬件。以 MPU 和 MCU 为例, 主流的架构包括 ARM Cortex 和 RISC-V, 它们的指令集不同, SOC 则更为多样, 这使得应用程序难以跨平台移植。

3) 软件碎片化: 边端人工智能应用的发展时间相对较短, 软件生态还没有达到稳定的状态。通常, 此类应用程序会包括一个操作系统(如 Linux)或者实时 OS(如 FreeRTOS)、一个推理框架(如 NCNN[37]、TensorFlow Lite[38])、或一个 C 语言的裸机实现。上述多样性限制了通用性, 导致软件工具的碎片化。

表 2 列出了典型的适用于低功耗人工智能计算平台的软件工具。不难看出, TVM[39] 具备一定的通用性, 可以支持多种架构的硬件, 其它软件则主要针对特定硬件。

表 2　典型的适用于低功耗人工智能计算平台的软件工具

Table2 List of typical software tools suitable for low-power AI computing platforms

软 件	支持的神经网络类型	支持的硬件平台	是否开源
TVM	ANN	CPU、GPU、MCU、FPGA等	Y
mircoTVM	ANN	ARM Cortex-M、nRF 5340	Y
TensorFlow Lite	ANN	支持Android 、iOS、Linux系统的硬件、ARM Cortex-M、ESP32、Himax WE-I Plus	Y
uTensor	ANN	ARM Cortex-M	Y
STM32 Cube.AI	ANN	STM32 ARM Cortex	N
NanoEdge AI Studio	ANN	STM32 ARM Cortex	N
Edge Impulse	ANN	ARM Cortex-M、ARM Cortex-A、ESP32	N
FANN-on-MCU	ANN	ARM Cortex-M、PULP	Y
CMSIS-NN	ANN	ARM Cortex-M	Y
MCUNET	ANN	STM32 ARM Cortex	Y
ARM-NN	ANN	ARM Ethos Processor、ARM Mali GPU、ARM Cortex-A	Y
ELL	ANN	ARM Cortex-A、ARM Cortex-M、Arduino、micro:bit	Y
NCNN	ANN	Intel、AMD、ARM、Apple等厂家的CPU、GPU	Y
RKNN	ANN	Rockchip SOC	Y
Vitis AI	ANN	AMD SOC、FPGA	Y
SpikingJelly	SNN	CPU、GPU、Lynxi SOC	Y

microTVM[40] 是 TVM 的扩展, 主要针对微控制器进行模型优化和推理部署, 支持裸机设备上的模型推理。目前支持 Zephyr RTOS 的 Cortex-M 微控制器, 同时也可移植到 RISC-V 等其它处理器。

TensorFlow Lite 是一个开源的深度学习框架, 支持边缘感知学习推理, 兼容 iOS、嵌入式 Linux、Android 和一系列微控制器, 其支持 ARM Cortex-M 系列、ESP32 架构硬件等, 即使是只有 16 kB 内

存的 ARM Cortex M3 也可运行一些神经网络模型。

uTensor[41] 是一个基于 Tensorflow 的轻量级机器学习推理框架，主要针对 ARM Cortex-M 处理器。

STM32Cube.AI[42] 和 NanoEdge AI Studio[43] 都是 ST 公司针对 STM32 ARM Cortex-M 系列 MCU 开发的软件工具。STM32Cube.AI 通过图形界面导入预训练的神经网络模型，并转换为适配 STM32 MCU 运行的代码，支持常用的模型文件格式，配合 STM32Cube IDE 可以在不同型号的 STM32 MCU 之间移植。NanoEdge AI Studio 集成了异常检测、分类等机器学习库，软件会根据处理器、内存、传感器等参数，搜索生成一个合适的 NanoEdge AI 库，并直接集成到嵌入式应用程序中，同时还可以在线采集现场数据进行学习和调试，极大提升了开发效率。

Edge Impulse[44] 是一种开发嵌入式人工智能模型的云服务，提供了端到端的解决方案，包括一整套开发流程，覆盖数据采集和分析、预处理、模型训练和评估、模型优化和压缩、模型转换和部署、设备管理等。同时，通过图形界面和自动化机器学习技术，可大大简化开发过程。

此外，还有一些针对 MCU 平台的软件工具。FANN-on-MCU[45] 基于快速人工神经网络库，可生成基于 PULP[46]、ARM Cortex-M MCU 上运行的代码。CMSIS-NN[47] 是一个针对 ARM Cortex-M 微控制器的神经网络库，其提供了一套高效、轻量的神经网络 API，使开发者可以在资源受限的微控制器上运行深度学习模型。MCUNET[48] 提出了一种高效网络架构搜索与轻量推理引擎联合设计的方案。

ARM NN[49] 是面向 ARM Cortex-A CPU 和 Arm Mali GPU 的高效推理框架，可运行于 Android 和 Linux OS 之上，同时提供了对 ARM Ethos-N NPU 的支持。ELL[50] 是微软发布的用于嵌入式机器学习的软件库，支持 micro:bit、树莓派和 Arduino 硬件。

NCNN 是腾讯优图实验室推出的面向手机平台的 AI 推理框架，其主要考虑手机端的部署和使用。NCNN 不依赖任何第三方库，代码使用 C/C++，实现了较好的运行效率。

RNKK[51] 是 Rockchip 针对自家芯片推出的软件开发套件，提供模型转换、推理和性能评估功能，支持 CPU 和 Rockchip NPU 硬件。

Vitis AI[52] 是 AMD 公司针对自家产品推出的 AI 开发解决方案，它包括一系列 AI 模型、优化的深度学习处理器单元内核、工具、库与示例设计，支持在 AMD SOC、FPGA 等平台快速高效部署 AI 应用。

SpikingJelly[53] 则提供了全栈式的脉冲深度学习解决方案，提供神经形态数据处理、深度 SNN 的构建、替代梯度训练、ANN 转换 SNN、权重量化和神经形态芯片部署等功能。

3 系统评估指标

低功耗人工智能计算系统具有硬件异构、软件碎片化、场景多样等特点，在构建低功耗系统时，不可避免地需要对各类低功耗人工智能计算系统进行评估，以做出合适的选择。通常在描述人工智能计算系统时使用以下性能指标:

1)算力: 它通常表示为系统执行推理任务时每秒可以执行的操作数，数值越高，性能越好。量化算力的指标是每秒千兆运算(GOPS)或者每秒 tera 运算(TOPS)。

2)存储容量: 存储器需求主要取决于模型权重和输入输出数据的大小，一个精度较高的分类和检测模型往往需要数兆或者数百兆字节的权重存储空间。在低功耗应用中，降低存储需求对提升系统能效、降低功耗起着重要作用。

3)推理准确性: 推理准确性描述了系统正确地完成任务的能力。一些人工智能应用，如人脸识别，要求非常高的准确率。

4)处理延迟: 处理延迟通常以时间单位(微秒、毫秒和秒)为单位来表示，用于测量输入信号到达与结果生成之间的时间差。对于边端场景的实时系统而言，降低延迟至关重要。

5)功耗: 功耗通常以瓦(W)、毫瓦(mW)或微瓦(μW)为单位来表示，是系统的总功耗。人工智能处理系统的功耗主要包括存储器访问功耗、计算功耗和外设的功耗等。通常，边缘设备是由电池供电，降低功耗可以延长系统工作时间，有着积极意义。

6)能效: 能效用于描述计算系统每消耗一瓦功率所能提供的操作数，通常表示为每瓦千兆运算

(GOPS/W)或每瓦 tera 运算(TOPS/W)。

7)成本: 成本用于描述构建整个系统所需的费用, 是各个部件的成本之和。具体到单个芯片时, 成本取决于芯片面积和采用的制造工艺。

正如嵌入式行业中使用 Coremark[54] 作为评价处理器性能的标准, 学术界和工业界也在推动人工智能推理应用的评测标准。AI Benchmark[55] 是苏黎世联邦理工学院推出的 AI 性能评测套件, 它主要面向智能手机和移动平台 SOC, 包含 78 项 AI 和计算机视觉任务测试, 指标涵盖了速度、精度、初始化时间、能效等多个维度。EEMBC MLMark[56] 则用于评测嵌入式推理设备的性能和精度, 包括图像分类和对象检测任务。

MLPerf Inference[57] 是由工业界和学术界共同推动而推出的机器学习推理基准套件, 可以用于评测机器学习硬件、软件和系统的推理性能。其主要包括单流、多流、服务器和离线四种评估场景, 并针对不同应用推出了不同的套件。当前共 4 个套件, 分别是: 面向数据中心的 MLPerf Inference: Datacenter[58]、面向边缘计算的 MLPerf Inference: Edge[59]、面向移动处理器的 MLPerf mobile inference benchmark[60] 和面向资源更为紧凑的低功耗计算系统的 MLPerf Tiny Benchmark[61]。

4　总结与展望

随着人工智能技术的快速发展, 在边端设备上部署人工智能应用已成为趋势。本文回顾了目前低功耗人工智能计算系统的现状, 阐述了存在的技术挑战, 并讨论了系统的评估方法和指标。

目前有多种硬件可用来构建低功耗人工智能计算系统, 如 CPU、GPU、FPGA、ASIC 和 SOC 等, 它们各有特点。在实际应用中, 系统处理的任务多样, 单一的硬件架构难以满足要求, 因此将不同计算架构、不同功能的模块进行集成, 发挥各自优势是未来的发展趋势。

由于硬件资源受限、异构等特点, 各商业公司均针对自家硬件开发了相应的软件工具, 软件总体呈碎片化的趋势。同时, 工业界已出现集成数据采集、分析、训练和代码生成等全流程的软件工具, 这种端到端的解决方案简化了开发过程, 大幅提升了开发效率, 是需要重点关注的方向。

学术界和工业界也出现一些新架构和新技术, 如近存计算、存内计算等, 可进一步提升系统效率。特别是基于忆阻器的存算系统具备高能效比优势, 具有较大发展潜力。同时, 为了进一步提升效率、实现极致低功耗, 将感知与存储、计算融合构建感存算一体化系统也是重要的研究方向。

此外, 以 SNN 为代表的第三代神经网络模型兴起, 结合事件相机等已有实际应用, 具备更低的功耗, 同样是值得关注的方向。

参考文献:

[1] Myagmar-Ochir Y, Kim W. A Survey of Video Surveillance Systems in Smart City[J]. Electronics, 2023, 12(17): 3567.

[2] Kunze L, Hawes N, Duckett T, et al. Artificial Intelligence for Long-Term Robot Autonomy: A Survey[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 4023-4030.

[3] Panayides A S, Amini A, Filipovic N D, et al. AI in Medical Imaging Informatics: Current Challenges and Future Directions[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(7): 1837-1857.

[4] Yurtsever E, Lambert J, Carballo A, et al. A Survey of Autonomous Driving: Common Practices and Emerging Technologies[J]. IEEE Access, 2020, 8: 58443-58469.

[5] Reuther A, Michaleas P, Jones M, et al. Survey and Benchmarking of Machine Learning Accelerators [C]. 2019 IEEE High Performance Extreme Computing Conference (HPEC). Waltham, MA, USA: IEEE, 2019: 1-9.

[6] Reuther A, Michaleas P, Jones M, et al. AI Accelerator Survey and Trends [C]. 2021 IEEE High Performance Extreme Computing Conference (HPEC). Waltham, MA, USA: IEEE, 2021: 1-9.

[7] Reuther A, Michaleas P, Jones M, et al. Survey of Machine Learning Accelerators [C]. 2020 IEEE High Performance Extreme Computing Conference (HPEC). Waltham, MA, USA: IEEE, 2020: 1-12.

[8] Hassanpour M, Riera M, González A. A Survey of Near-Data Processing Architectures for Neural Networks[J]. Machine Learning and Knowledge Extraction, 2022, 4(1): 66-102.

[9] Asifuzzaman K, Miniskar N R, Young A R, et al. A survey on processing-in-memory techniques: Advances and challenges[J]. Memories-Materials, Devices, Circuits and Systems, 2023, 4: 100022.

[10] Yamazaki K, Vo-Ho V-K, Bulsara D, et al. Spiking Neural Networks and Their Applications: A Review[J]. Brain

Sciences, 2022, 12(7): 863.
[11] https://developer.arm.com/Processors/Cortex-A7
[12] https://ambiq.com/apollo3/
[13] https://ambiq.com/apollo4/
[14] https://www.st.com/content/st_com/en/search.html#q=STM32L4-t=products-page=1
[15] https://www.st.com/content/st_com/en/search.html#q=STM32H7-t=products-page=1
[16] https://www.nvidia.cn/autonomous-machines/embedded-systems/jetson-nano/
[17] Garrett D, Park Y S, Seongjong K, et al. A 1mW Always-on Computer Vision Deep Learning Neural Decision Processor [C]. 2023 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA: IEEE, 2023: 8-10.
[18] https://www.analog.com/cn/products/max78000.html
[19] https://perceive.io/product/ergo/
[20] https://www.rock-chips.com/a/cn/product/RV11xilie/2020/0427/1075.html
[21] Akopyan F, Sawada J, Cassidy A, et al. TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34(10): 1537-1557.
[22] Liu D F, Chen T S, Liu S L, et al. PuDianNao: A Polyvalent Machine Learning Accelerator[J]. ACM SIGPLAN Notices, 2015, 50(4): 369-381.
[23] Chen T S, Du Z D, Sun N H, et al. DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning[J]. ACM SIGARCH Computer Architecture News, 2014, 42(1): 269-284.
[24] Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572: 106-111.
[25] https://www.intel.cn/content/www/cn/zh/products/docs/processors/movidius-vpu/myriad-x-product-brief.html
[26] https://www.hisilicon.com/cn/products/smart-vision/pro-camera/Hi3516DV500
[27] https://coral.ai/products/
[28] https://cn.horizon.cc/journey2.html
[29] Kuon I, Tessier R, Rose J. FPGA Architecture: Survey and Challenges [M]. Now Foundations and Trends, 2008.
[30] Liu L B, Zhu J F, Li Z S, et al. A Survey of Coarse-Grained Reconfigurable Architecture and Design: Taxonomy, Challenges, and Applications[J]. ACM Computing Surveys, 2020, 52(118): 1-39.
[31] Guo K Y, Zeng S L, Yu J C, et al. A Survey of FPGA-based Neural Network Inference Accelerators[J]. ACM Trans. Reconfigurable Technol. Syst, 2019, 12(2): 1-26.
[32] Véstias M, Duarte R P, Sousa J T, et al. Lite-CNN: A High-Performance Architecture to Execute CNNs in Low Density FPGAs [C]. 2018 28th International Conference on Field Programmable Logic and Applications (FPL). Dublin, Ireland: IEEE, 2018: 399-3993.
[33] Yin S Y, Ouyang P, Tang S B, et al. A High Energy Efficient Reconfigurable Hybrid Neural Network Processor for Deep Learning Applications[J]. IEEE Journal of Solid-State Circuits, 2018, 53(4): 968-982.
[34] Abadi M, Barham P, Chen J M, et al. TensorFlow: a system for large-scale machine learning [C]. Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16). Berkeley, CA, USA: USENIX Association, 2016: 265–283.
[35] Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library [C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2019(721): 8026-8037.
[36] Li M Z, Liu Y, Liu X Y, et al. The Deep Learning Compiler: A Comprehensive Survey. IEEE Transactions on Parallel and Distributed Systems [J], 2021, 32(3): 708-727.
[37] https://github.com/Tencent/ncnn
[38] https://tensorflow.google.cn/lite?hl=zh-cn
[39] Chen T Q, Moreau T, Jiang Z H, et al. TVM: an automated end-to-end optimizing compiler for deep learning [C]. Proceedings of the 13th USENIX conference on Operating Systems Design and Implementation (OSDI'18). Berkeley, CA, USA: USENIX Association, 2018: 579–594.
[40] https://tvm.hyper.ai/docs/topic/microtvm/
[41] https://github.com/uTensor/uTensor
[42] https://www.st.com/content/st_com/zh/campaigns/stm32cube-ai.html
[43] https://www.st.com/zh/development-tools/nanoedgeaistudio.html
[44] Reddi V J, Elium A, Hymel S, et al. Edge impulse: An mlops platform for tiny machine learning [J]. Proceedings of Machine Learning and Systems, 2023, 5.
[45] Wang X, Magno M, Cavigelli L, et al. FANN-on-MCU: An Open-Source Toolkit for Energy-Efficient Neural Network

Inference at the Edge of the Internet of Things[J]. IEEE Internet of Things Journal, 2020, 7(5): 4403-4417.

[46] Rossi D, Conti F, Marongiu A, et al. PULP: A parallel ultra low power platform for next generation IoT applications[C]. 2015 IEEE Hot Chips 27 Symposium (HCS). Cupertino, CA, USA: IEEE, 2015: 1-39.

[47] Lai L, Suda N, Chandra V. CMSIS-NN: Efficient neural network kernels for Arm Cortex-M CPUs[J]. arxiv preprint arxiv, 2018: 1801.06601.

[48] Lin J, Chen W M, Lin Y, et al. MCUNET: Tiny deep learning on IoT devices[J]. Advances in neural information processing systems, 2020, 33: 11711-11722.

[49] https://github.com/ARM-software/armnn

[50] https://github.com/Microsoft/ELL

[51] https://github.com/airockchip/rknn-toolkit

[52] https://china.xilinx.com/products/design-tools/vitis/vitis-ai.html

[53] Fang W, Chen Y Q, Ding J H, et al. SpikingJelly: An open-source machine learning infrastructure platform for spike-based intelligence[J]. Science Advances, 2023, 9(40): 1480.

[54] https://www.eembc.org/coremark/

[55] https://ai-benchmark.com/

[56] https://www.eembc.org/mlmark/

[57] Reddi V J, Cheng C, Kanter D, et al. MLPerf Inference Benchmark [C]. 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain: IEEE 2020: 446-459.

[58] https://mlcommons.org/benchmarks/inference-datacenter/

[59] https://mlcommons.org/benchmarks/inference-edge/

[60] Reddi V J, Kanter D, Mattson P, et al. MLPerf Mobile Inference Benchmark: An Industry-Standard Open-Source Machine Learning Benchmark for On-Device AI[J]. Proceedings of Machine Learning and Systems, 2022, 4: 352-369.

[61] Banbury C, Reddi V J, Torelli P, et al. MLPERF Tiny Benchmark[J]. arxiv preprint arxiv, 2021: 2106.07597.