手机终端人工智能关键技术研究

庞涛，丘海华，潘碧莹
（中国电信股份有限公司研究院，广东 广州 510630）

摘　要：随着移动终端软硬件技术的进步，移动终端的机器学习能力得以挖掘。从手机终端人工智能技术整体框架入手，研究了端侧人工智能硬件加速技术，对比了当前主流端侧机器学习框架和神经网络模型压缩等软件技术，分析了 AI 应用发展趋势，总结了手机终端人工智能技术进展和未来发展趋势。

关键词：手机终端；人工智能；机器学习；神经网络模型
中图分类号：TP334.1
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020045

Research on artificial intelligence key technologies of mobile terminal

PANG Tao, QIU Haihua, PAN Biying
Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: With the progress of software and hardware technology of mobile terminals, the machine learning ability of mobile terminals has been explored. Starting from the overall framework of mobile terminal artificial intelligence technology, the end-to-end artificial intelligence hardware acceleration technology was studied, mainstream on-device machine learning frameworks, neural network model compression and other software technologies were compared, the development trend of AI application was analyzed, and the current technological progress and future development trend of mobile terminal artificial intelligence were summarized.

Key words: mobile terminal, artificial intelligence, machine learning, neural network model

1　引言

有研究表明，通用计算正在逐步衰退[1]，摩尔定律正在逐渐失效，移动智能终端通用芯片技术发展遭遇了瓶颈。芯片尺寸从当前 7 nm 制程往下发展至 2～3 nm，已经接近硅材料芯片的物理极限，因此手机芯片发展的技术路线从提高集成度扩展到多核心、异构芯片协同等发展方向。为满足人工智能算法在端侧部署的计算需求，在手机终端上诞生了人工智能计算加速芯片（以下简称 AI 芯片）。

应用软件可以通过芯片厂商或机器学习框架提供的 SDK 和 API 调用底层硬件的 AI 加速计算能力。但是对于应用开发者而言，开发 AI 应用需

收稿日期：2019−05−31；修回日期：2020−01−20

要针对不同的手机平台对相应的SDK或API进行适配，而且各芯片厂商选择支持的深度学习框架、模型算子都有差异，从而限制了当前端侧AI应用的快速普及。

总之，以手机为代表的移动智能终端为了赋能AI，在芯片、系统和应用生态等方面实现跨越式发展的同时仍然面临诸多问题[2-3]。本文聚集研究手机终端的人工智能软硬件关键技术。

2 整体架构

当前的人工智能主要依靠诸如机器学习、深度学习等AI技术实现机器自主学习数据信息，过程大致是先将特定领域的大数据（图片、音频数据）通过深度学习技术训练得到AI算法模型，该模型可以被加载在包括手机在内的各种终端中用于机器推断（inference）。

手机行业针对AI算法模型的加速计算，特别推出人工智能手机[1-2]的概念，其定义相对宽泛，因为主流手机基本上都能通过CPU运行浅层的神经网络。是否需要对人工智能手机设定明确的指标门槛，还是仅仅针对手机的“人工智能能力”进行规范，是国内外标准组织尚在研究的内容。

机器学习涵盖了大量的高维数据以及变量，构成了高维数据计算问题，张量（tensor）方法能够十分有效地处理这种高维数据进行机器学习。因此，当前的手机终端人工智能技术主要在芯片底层提供张量加速计算硬件单元，通过AI算法模型的接口和SDK等软件技术为上层应用提供加速AI推断过程计算的技术。

图1给出了手机终端人工智能技术的整体框架，手机终端的AI能力由硬件和软件协同提供。硬件层面，需要对AI软件的运行提供足够的算力，能加速机器学习和神经网络常用的矢量和张量计算；软件层面，主要提供运行于手机的机器学习框架和各类型AI应用。

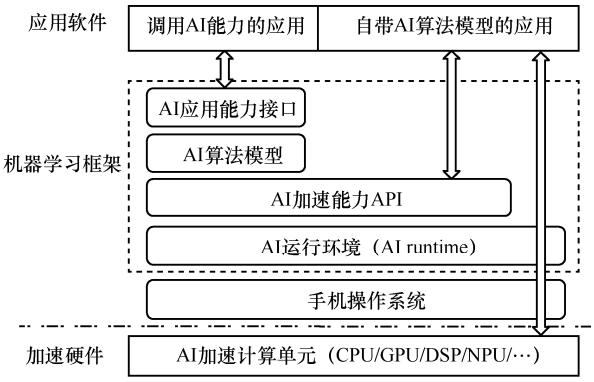

图1 手机终端人工智能技术整体框架

3 硬件技术

机器学习模型可以在各种形式的处理器上运行，从微处理器到CPU、GPU到DSP，问题的核心在于处理器运算这类模型的速度及效率。

从当前最常用的深度神经网络（DNN）来看，由于其基本操作是神经元（网络节点）和突触（节点之间连接）的处理，而x86和ARM这类传统的CPU指令集是为了进行通用计算而设计的，其基本操作为算术操作（加减乘除）和逻辑操作（与或非），往往需要众多指令才能完成一个神经元的处理，因此处理DNN模型数据的效率不高。GPU是目前云端运行DNN的主流选择，但其单位面积的成本较高、能耗较大，不适合用于移动终端上的AI加速计算。因此，业界设计了各种神经网络专用硬件加速器，例如Temam提出一种基于多层感知机的加速器，Chen等[4]提出了神经网络加速器DaDianNao等。专用NN加速芯片最大的特点就是单指令可以完成多个神经元的计算，并对神经元数据在芯片上的传输提供了一系列专门的支持。

在移动设备上，由于能耗、体积等限制，一般使用专用领域处理器架构（DSA）的DSP或神经网络处理器（NPU）进行加速，而当前阶段移动端DSA AI硬件架构基本不支持训练，只支持前向推断。

业界一般采用每秒操作数量（operations per

second，OPS）作为衡量硬件算力水平的一个性能指标。乘法累加（multiply-accumulate，MAC）操作是 DNN 运算的基础，通常一次 MAC 运算可看作两次深度学习的运算操作。

2017 年，苹果在 A11 仿生芯片上率先集成神经网络引擎。华为则在麒麟 970 芯片平台搭载了独立的 NPU。高通则通过神经网络处理引擎（NPE）软件框架对底层的 CPU、GPU、DSP 进行调度，提供异构的 NN 加速方案。联发科同样在 P60 平台集成了双核人工智能处理器（APU）。此时手机上的 AI 算力为 0.6~2 TOPS。

2018 年，苹果将 A12 芯片的 NN 加速器从 A11 的双核设计升级为八核设计，华为麒麟 980 处理器中 NPU 升级为双核架构，高通随后发布的骁龙 855 处理器同样为 AI 计算集成了新的 Hexagon 张量加速器（HTA），并扩展了矢量计算单元。2019 年年初，联发科 Helio P90 也将 APU 升级到 2.0 版本，提升了主频和计算速度。从产业发展可以看出，手机终端集成独立的 AI 加速单元在高端机型上已成为主流趋势。此时手机上的 AI 算力已经增长到 2.25~7 TOPS。

图 2 给出高通骁龙 845 芯片中各计算单元运行两个自定义神经网络的速度和功耗对比。其中，图 2(a)是运行包括 4 层节点数，分别为 128、256、512、10，核心尺寸为 3×3 像素的卷积层和 1 层 1 001 个节点的全连接网络；图 2(b)是运行 3 层节点数，分别为 512、128、1 001 的全连接网络。两个网络的输入为 100 张 224×224 像素的图片，不考虑输出，单纯用于检验 AI 芯片对神经网络最常见的卷积运算和全连接运算加速效果。由图 2 可以看到，具有向量计算加速单元的 DSP 的计算速度和功耗远远优于 CPU 和 GPU。华为的 NPU 和联发科的 APU 均表现出类似的优势。随着 AI 加速解决方案成本的降低，越来越多的中低端机型也有可能搭载独立的 AI 加速硬件单元，为用户带来更好的 AI 应用体验。

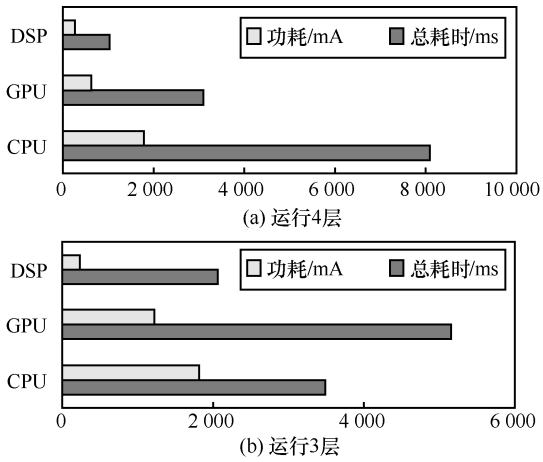

图 2　全连接网络的速度和功耗对比

如果从能效角度来看，目前手机 AI 加速芯片的处理效率在 0.5 TOPS/W 上下，理论上可称为高能量效率的 AI 芯片需要达到 10~20 TOPS/W，由此可见，手机 AI 芯片距离实现真正的高能量效率还有相当大的差距。

4　软件技术

4.1　手机终端机器学习框架

要释放终端上 AI 加速硬件的计算能力，还需要软件层面的紧密配合。尽管如图 1 中所示，某些与手机终端深度整合的对速度要求苛刻的 AI 应用，可以通过芯片厂商提供的底层固件接口直接调用 AI 加速单元，但此类底层接口并不开放给普通开发者使用。为了能让应用开发者方便、安全地使用 AI 硬件加速资源，各芯片平台都提供了手机终端机器学习框架，包含 SDK 和神经网络运行环境（NN runtime）。与云端的机器学习框架不同[10]，终端侧的机器学习框架以推断为主，裁减了训练部分。另外，为了在终端上加速运行神经网络模型，还需要使用各自框架中提供的模型优化和转换工具，支持将其他机器学习框架训练的 NN 模型转换为可在手机本地支持的模型格式。

当前，除了 Android（TensorFlowLite）和苹果（CoreML）两大操作系统提供的端侧机器学习框架，主要移动芯片厂商和部分整机生产商也各

自推出了面向自身平台的机器学习框架，例如，高通的 SNPE、华为的 HiAI 引擎、联发科的 NeuralPilot、小米的 MACE 等。由于各自所在产业链中的位置不同，各框架的开放策略和提供的 AI 能力并不一致。

主流手机终端机器学习框架的对比见表 1。

由此可见，由于硬件解决方案的差异，产业链上各家纷纷推出手机终端机器学习框架的局面导致了端侧机器学习 SDK“碎片化”的问题，给移动终端应用开发者带来了困扰。尽管在 Android 系统生态中 TensorFlow Lite 和 AndroidNN 试图在操作系统层面统一 AI 加速计算能力的调用，但是芯片和终端设备厂商出于自身产品生态和客户需求的考虑，对 Android 原生 NN API 生态的支持和完善并不积极。

除了表 1 所列的框架，在移动端还有腾讯推出的手机端前向推理框架 NCNN、百度开源框架 Paddle-mobile（MDL）、阿里巴巴开源的轻量级移动端推理引擎（MNN）等，由于这些框架没有专用硬件适配支撑，大多数 AI 运算在 ARM CPU/GPU 上进行，效率较低。

4.2 模型压缩

在云端完成训练的机器学习模型，尤其是 DNN 模型动辄数百兆的尺寸并不适合直接用于手机终端，通常的做法是将 DNN 模型进行压缩，减少体积，提升推理速度。模型的压缩方法通常分为网络量化、网络剪枝、低秩分解、知识迁移网络、紧凑网络设计 6 类[5]。

网络模型的量化方法相对易于实现。由于用于推断的网络权值已经固定下来，无后向传播过程，相比训练过程对精度的要求低很多，可以用低精度，如半长 16 位的浮点型（FP16）或者用 8 位的整型（INT8）来替换高精度模型中的浮点型参数做推断，这种数据替换过程称为定点量化。研究结果表明，定点量化后的模型相比浮点模型在采用各种恢复精度措施后没有特别大的推断精度损失[8-9]。目前手机端的 FP16 和 INT8 量化较为常见，另外，网络模型的二值量化（−1，+1）和三值量化（−1，0，+1）等方法也受到了越来越多的关注。虽然苹果 CoreML 等框架宣称支持二值化网络，但目前还没有见到可实际应用的二值化网络模型。

除了量化外，利用模型剪枝[6]、权值共享[7]等手段，减少模型参数实现模型压缩也是比较常见的方法。基于模型裁剪的方法很多，基本思路是挑选出模型中不重要的参数，将其剔除而不会

表 1 主流手机终端机器学习框架对比

对比项	高通 SNPE	华为 HiAI	联发科 NeuralPilot	小米 MACE	Android TensorFlow Lite	苹果 CoreML
开放程度	面向大众开放 SDK，HTA 能力需授权调用	面向大众开放 SDK	面向 OEM 合作方开放 SDK	开源	开源	面向大众开放 SDK
支持转换的模型格式	TensorFlow、Caffe、PyTorch（Caffe2）等	TensorFlow（支持版本较低）、Caffe	TensorFlow、TensorFlowLite	TensorFlow、Caffe、ONNX	TensorFlow、TensorFlowLite	Caffe、Keras、XGBoost、ONNX 等
模型量化	在线量化/离线量化	离线量化	离线量化	离线量化	离线量化	在线量化/离线量化
AI 接口	只提供底层混合异构硬件的加速能力 API	底层提供硬件加速 API 和上层的人脸识别等应用能力接口	底层硬件加速接口，支持 Android NN API	底层硬件加速接口	支持 Android NN API 和更上层面的应用能力接口	支持底层硬件加速 API 和上层语音识别、游戏加速等能力接口

对模型的效果造成太大的影响。权值共享的基本思想是多个网络连接的权重共用一个权值。基于低秩分解的神经网络加速与压缩方法，其出发点是找到与张量 W 近似，但计算量更小的张量 W'，因为预训练的卷积模型卷积核存在着低秩特性，所以常采用低秩分解（SVD）的方法进行压缩。

以上方法主要针对已经训练过的大型网络模型进行压缩，而知识迁移（或知识蒸馏）和紧凑（轻量化）网络设计则直接从网络结构上进行轻量化的设计，再通过训练生成轻量化的模型。知识迁移网络方法会使用两种类型的网络，一种是教师网络；另一种是学生网络。该方法首先训练教师网络，达到非常高的性能，然后通过使用学生网络对教师网络进行拟合，从而把教师网络学到的知识迁移到学生网络中。一般来说，教师网络是一个大型神经网络或神经网络集合，而学生网络则是一个紧凑而高效的神经网络。而 MobileNet、SqueezeNet、ShuffleNet 等网络则是以轻量化为设计目标的典型代表[11]，其设计的出发点是寻找更高效的“网络计算方式”（主要针对卷积方式），从而使网络参数减少的同时，不损失网络性能。通过引入 Depth-wise 卷积、1×1 卷积核减少特征图数量等方式降低卷积层的参数量。

5　AI 应用

5.1　端云结合的 AI 应用

人脸识别、拍照识物、文字翻译、语音助手等诸多 AI 应用在手机上出现。早期受限于端侧的算力及模型的尺寸，AI 应用主要在云端实现，云端拥有强大的算力和数据汇聚能力，是深度学习训练和建模的最佳场景。但由于能耗、成本、时延性等问题，在终端设备上远程运行 AI 算法对于终端用户来说体验并不完美。

云服务提供商通过后端即服务（backend as a service，BaaS）提供 API，开发者只需几行简单的代码就可以整合一些后端功能，如数据存储、应用分析等。Google Firebase 是 BaaS 平台的典型代表，Firebase 提供了机器学习套件（MLkit），其中文字识别、面部监测等功能既可以在终端上实现，也可以在云端实现。

随着端侧算力的增加和隐私、安全、时延等方面的考虑，使用端侧 AI 能力的应用越来越普及。以语音助手为例，Google Assistant 已经宣称将云端 100 GB 大小的人工智能模型缩小到了 500 MB，让手机实现不依赖于云端的语音助手成为可能。

随着 5G 时代的到来，终端、边缘、云都被赋予了 AI 能力，手机终端作为个人用户的直接入口，适用于各种应用场景。Google 提出了联邦机器学习（federated learning）技术[12]，云端通过整合海量端侧训练的 AI 模型对输入法中预测字词的准确率进行优化和提升。

5.2　AI 应用能力开放

由于 AI 产业链上各厂商的定位不同，并不是每家厂商都会为应用开发者提供高层次的 AI 应用能力接口，例如芯片解决方案提供商的定位决定了高通 SNPE 只会为其客户提供基于底层硬件的加速能力，DNN 模型由客户自己提供。

而华为、苹果、谷歌这些整机解决方案提供商则会为用户提供应用层面的 AI 能力接口，例如华为的 HiAI Engine 和 HiAI Service，将端侧和云侧的视觉、语音识别、语义理解、推荐等能力封装成开放的 API，为应用开发者提供融合开放的 AI 能力。苹果则通过 CoreML 为应用开发者提供了计算机视觉、自然语言处理和游戏开发套件等 AI 能力开放接口。

由于手机终端底层软硬件 AI 加速方案的差异，使得开发者需要面对 AI 应用多平台适配的问题，而通过标准化 AI 应用能力开放接口，为开发者提供一套跨平台的统一 AI 应用能力调用接口，是看起来最有希望解决 AI 应用开发“碎片化”困

局的方案。

6 结束语

当前的智能手机纷纷以加载人工智能技术为卖点，不同于以往游戏手机、AR 手机等局部功能创新与优化，人工智能技术正深入融合进智能手机的各个层面，智能手机演进接下来的形态是 AI 手机。

从硬件角度来看，通用性与效率的权衡仍是现阶段手机 AI 芯片设计重点考虑的因素，当前 AI 芯片与 NN（尤其是 DNN）模型深度绑定，移动终端搭载的 AI 芯片仅可支持 NN 正向运算，无法支持 NN 的训练。现阶段，AI 硬件的通用性和速度、功耗不可兼得，想要更好的性能就得牺牲通用性，就需要针对特定领域进行深度定制。另外，当前手机终端的 AI 算力还存在瓶颈，体现在需要连续实时计算的、有多种网络混合的 AI 应用场景，如视频类、AR/VR 类应用场景下，仅仅依靠端侧计算，手机终端会发热、运算速度下降。因此可以预见，手机芯片仍将大幅提升 AI 算力的同时努力降低单位能耗。而且，随着端侧训练需求的出现，手机产业链也开始布局端侧训练的加速技术研究。

从软件角度来看，人工智能技术中软件发挥的作用非常大[13]。一旦各家硬件方案定型，接下来要比拼的就是软件生态，软件生态决定了 AI 应用的丰富度、易用性，从而决定了市场走向。推出移动终端机器学习框架也是产业链上各家切入和把控移动终端人工智能产业链的重要举措。当前，手机终端的机器学习软件框架众多，就算是谷歌和苹果也难以脱颖而出。在这种“碎片化”的局面下，通过将上层 AI 应用能力开放接口进行标准化定义，也许是壮大 AI 应用生态相对可行的方案。

未来，随着移动终端 AI 能力的不断增强，用户可以感受到手机更人性化，能正确理解用户意图和情绪。在 AI 能力的加持下，手机越来越个性化的背后是用户隐私数据的使用和分析。学术界也已经开始在这方面研究如何利用 AI 技术保护好用户的隐私，确保在保障隐私和安全的前提下提升用户的个性化体验。

参考文献：

[1] NEIL C. THOMPSON, SVENJA SPANUTH. The decline of computers as a general purpose technology: why deep learning and the end of Moore’s law are fragmenting computing[EB]. 2018.

[2] 艾瑞咨询. 中国人工智能手机行业研究报告[R]. 2018.
IResearch. China artificial intelligence mobile phone industry research report[R]. 2018.

[3] 中国电信. 中国电信全网通 AI 手机白皮书 2.0[R]. 2018.
China Telecom. China Telecom full netcom AI mobile white paper 2.0[R]. 2018.

[4] CHEN Y J, CHEN T S, XU Z W, et al. DianNao family: energy-efficient hardware accelerators for machine learning[J]. Communications of the ACM,2016,59(11):105-112.

[5] CHENGJ, WANG PS, LIG, et al. Recent advances in efficient computation of deep convolutional neural networks[J]. arXiv, 2018: 1802.00939.

[6] LI H, ASIM KADAV, IGOR DURDANOVIC, et al. Pruning filters for efficient convnets[J]. arXiv, ICLR 2017:1608.08710.

[7] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding[J]. arXiv Preprint arXiv,2015:1510.00149.

[8] BENOIT J, SKIRMANTAS K, BO C, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[J]. arXiv,2017:1712.05877.

[9] SHENG T, FENG C, ZHUO S J, et al. A quantization-friendly separable convolution for mobilenets[J]. arXiv, 2019: 1803.08607.

[10] 庞涛. 开源深度学习框架发展现状与趋势研究[J]. 互联网天地, 2018(4): 46-54.
PANG T. Research on development status and trend of open source deep learning framework[J].Internet World, 2018(4): 46-54.

[11] 余霆嵩. 纵览轻量化卷积神经网络: SqueezeNet、MobileNet、ShuffleNet、Xception[Z]. 2018.
YU T S. Overview of lightweight convolutional neural networks: SqueezeNet、MobileNet、ShuffleNet、Xception[Z]. 2018.

[12] HARD A, RAO K, MATHEWS R, et al. Federated learning for mobile keyboard prediction[J]. arXiv, 2018: 1811.03604v1.

[13] 中国人工智能产业发展联盟（AIIA）. 手机人工智能技术与应用白皮书[R]. 2019.
China Artificial Intelligence Industry Development Alliance (AIIA). Mobile phone artificial intelligence technology and application white paper[R]. 2019.

[作者简介]

庞涛（1977− ），男，中国电信股份有限公司研究院工程师，主要研究方向为大数据应用、人工智能、智能终端技术等。

丘海华（1991− ），男，中国电信股份有限公司研究院工程师，主要研究方向为人工智能、AR 技术、边缘计算等。

潘碧莹（1994− ），女，中国电信股份有限公司研究院工程师，主要研究方向为人工智能、智能终端技术等。