深度学习的10年回顾与展望

Deep Learning: Past Decade and Future

韩炳涛/HAN Bingtao[1,2]，刘涛/LIU Tao[1,2]，唐波/TANG Bo[1,2]
（1. 中兴通讯股份有限公司，中国 深圳 518057；
2. 移动网络和移动多媒体技术国家重点实验室，中国 深圳 518055）
(1. ZTE Corporation, Shenzhen 518057, China;
2. The State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China)

DOI：10.12142/ZTETJ.202206013
网络出版地址：https://kns.cnki.net/kcms/detail//34.1228.TN.20221221.1442.003.html
网络出版日期：2022-12-23
收稿日期：2022-10-15

摘要：过去10年深度学习在算法、算力、数据方面获得了长足发展，使人工智能（AI）技术突破商用限制，行业应用场景日益广泛，产业规模持续扩大。在基础模型方面出现了卷积、注意力机制等关键突破；在学习方法方面，强化学习、自监督学习、大模型并行训练等使模型学习能力大大加强。新型AI计算芯片不断涌现，使计算能效提升百倍。未来10年，深度学习若要保持可持续的指数增长态势，绿色、高效、安全将成为新的核心要素。空间计算、近似计算等技术有望使AI芯片效能继续获得百倍提升。一系列生态融合工具的出现将解决目前日趋严峻的生态碎片化问题。AI安全、可信将成为AI技术应用的基本要求。

关键词：深度学习；AI芯片；推理加速；可信AI；开源

Abstract: In the past ten years, deep learning has made great progress in algorithm, computing power, and data, which has enabled artificial intelligence (AI) technology to meet commercial requirements, and has an increasingly wide range of application in various kinds of business, and the scale of the industry has continued to expand. In terms of basic models, there have been key breakthroughs such as convolution and attention mechanisms; in terms of learning methods, technologies such as reinforcement learning, self-supervised learning, and parallel training of large-scale model have greatly enhanced performance. New AI chips continue to emerge, and computing energy efficiency has increased by a hundredfold. In the next ten years, deep learning will maintain a sustainable exponential growth trend, and green, efficient, and safe will become the new core elements. Spatial computing, approximate computing and other technologies are expected to continue to improve the performance of AI chips by a hundredfold. Some integration tools will appear to solve the increasingly severe ecological fragmentation problem. AI security and trustworthiness will become the basic requirements for the application of AI technology.

Keywords: deep learning; AI chip; inference accelerating; trusted AI; open source

1 第1个10年回顾

2012年AlexNet[1]横空出世，掀起第3次人工智能（AI）浪潮。从此AI进入深度学习时代。在深度学习的第1个10年，数据、算法、算力三大要素得到迅速发展。与前两次浪潮不同的是，在第3次浪潮中AI技术一举突破商用限制，拥有日益广泛的行业应用场景，产业规模持续扩大，打消了人们对于第3次浪潮何时终结的疑虑。

1.1 算法长足发展

深度学习的特点是可以将基础算子以层层叠加的方式组成复杂的神经网络，并使用反向传播算法统一实现神经网络的训练。使用如此的简单方法即可构建任意复杂模型。这种能力使深度学习成为一种适用于多种任务的通用算法。

在过去10年中，基础模型经历了两次跨越式发展。第1次跨越是以AlexNet为代表的卷积神经网络。2015年ResNet[2]的出现使得这一阶段的发展达到高峰。在这一阶段人们普遍认为，更深的神经网络将具备更强的表征能力。因此，研究者主要思考如何增加神经网络的深度。ResNet通过引入跨层shortcut连接，成功将网络深度提升至150层以上。之后的研究虽将网络深度提升至1 000层以上，但模型性能提升幅度越来越小，因此百层左右的网络成为应用的主流选择。此外，这一阶段发展了大量基于卷积计算的算子，在提取空间和时间局部特征方面取得了很好的效果，使得图像、语音模式识别准确率大幅提升，产生了诸如语音输入、人脸识别等第一批可商业化的技术，为第三次AI浪潮创造了一个良好的开端。

第2次跨越是以2016年出现的以Transformer[3]为代表的注意力机制神经网络。注意力机制此前在神经网络中仅是辅助性算子，但Transformer创造性地将其作为网络核心算子，引发了一系列重大创新。Transformer最初解决了长短期记忆网络（LSTM）[4]等循环神经网络计算效率低、训练容易过拟合等问题。2017年基于Transformer的预训练语言模型

BERT[5]利用海量样本的无监督预训练大幅提升下游任务表现能力，使大规模样本预训练和少量样本精调成为模型训练新范式。此后，自监督预训练[6]更是将这一范式推向高潮。研究人员很快发现语言模型规模越大，表现就越好。模型规模在短短的两年内迅速突破了千亿参数级别。2019年，拥有1 700亿参数的GTP-3[7]模型在对话、知识问答、吟诗作赋等多项任务中展示出的能力令人印象深刻。深度学习从此迈入大模型时代。现如今相关模型规模已经达到百万亿级别[8]。此外，研究人员发现Transformer具备跨模态通用性。2018年，ViT[9]模型证明Transformer除了适用于处理自然语言相关任务外，在处理图像任务方面也不输于卷积神经网络。最新的DALL.E、紫东太初、M6[10-12]等多模态模型更是可以同时处理文本、语音、图像多模态数据。自监督预训练、大模型、多模态等创新Transformer成为当今最重要的深度学习模型，为深度学习的发展带来无限可能。

除了基础模型，过去10年在学习方法上也取得了重大进展。学习方法主要包括监督学习和无监督学习两大类。监督学习比较容易，但需要对数据进行标注，这个过程通常需要耗费大量人力。强化学习是一种特殊的监督学习，它只需要一个回报信号而无须对每条数据进行标注。在强化学习过程中，算法以一种“试错”的方式对问题空间进行探索，从而找到一种最优（获取最大回报）的策略。深度学习模型和强化学习方法相结合，产生多项重要成果，大幅拓展了深度学习的应用边界。2016年AlphaGo[13]战胜九段专业棋手，AI进入大众视野，第3次AI浪潮开始井喷。2017年AlphaGo Zero[14]完全不依赖人类的围棋知识，仅从最基本的围棋规则开始，经过72 h的训练，棋力就可远超AlphaGo。2018年AlphaZero[15]使用同一个模型和算法，同时掌握国际象棋、将棋、围棋，显示出强化学习有实现通用AI的潜力。强化学习在德州扑克、DOTA、星际争霸等视频游戏[16-18]中也达到顶尖人类玩家的水平。在真实环境中使用强化学习的研究也取得很大进展。使用强化学习算法不仅可以对机械臂实现适应性控制，可以完成诸如网线插拔等灵巧型任务[19]，甚至还可以操作复杂的可控核聚变托卡马克装置，实现对装置中高温等离子体形状、位置的跟踪和控制[20]。最近，强化学习在科学领域也取得不小进展。例如，AplhaTensor[21]可以发现各种大小的矩阵乘法的速算方法，而人类科学家还没能发现任何一种大于3×3规模矩阵的速算方法。强化学习在多种任务中体现出通用性，使其成为实现“通用AI”一条重要技术路线，不断吸引更多学者参与到研究中来。

难度最大同时也是最具发展潜力的无监督学习方法，特别是在生成式模型领域，在过去10年产生了两个重大的方法创新。一个是在2014年，Goodfellow提出的生成对抗网络（GAN）模型及其创新的对抗训练方法[22]，被LeCun认为是过去10年中机器学习领域中最有趣的想法。对抗训练方法通过同步优化生成器、判别器，使两者达到纳什均衡。这种方法可以生成更加清晰的图片，但训练过程不稳定。此后对GAN的改进成为研究热点，特别是在2019年BigGAN[23]改进了大规模网络下对抗训练不稳定的问题，使batch size增大至2 048，模型参数达到1.7亿，在生成图像的真实性和多样性上取得巨大进步，生成了可以假乱真的图像。另一个是自监督学习，以变分自动编码器（VAE）[24]为代表的自动编码器通过将模型分割为编码器、解码器两个部分，先将数据编码到隐变量空间，再从隐变量空间解码恢复数据。这种方法使数据自身成为标签，在不使用任何人工标注的情况下从大规模无标签数据中学习数据特征。除了以原始数据作为标签外，其他多种建立自标签的方法也陆续被发现。2021年Diffusion Model[25]则是将向原始图像添加的高斯噪音作为标签，让模型从加噪的图像中预测噪音，从而学习得到降噪编码器。将这样多个降噪编码器层层叠加，就可以从噪音中得到图像。这种方法可以使深度学习模型生成前所未有的高清、逼真图像。2022年潜在扩散模型（LDM）[26]大幅提升了高分辨率图像的效率，使AI内容生成技术更加实用化。AI在未来音乐、视频、游戏、元宇宙内容生成中有广阔的应用前景。

正是由于过去10年中深度学习算法的长足发展，如今AI已在千行百业中拥有广泛的应用场景[27-36]，产业规模持续扩大，成为数字经济下不可或缺通用基础技术，对经济增长意义重大。

1.2 算力需求驱动芯片迅速发展

J. SEVILLA等[37]对AI主要算法所需要的算力进行了汇总。过去10年在模型训练方面，模型所需的算力增长超过了100万倍。深度学习对算力的巨大需求推动了AI芯片快速发展。在这10年中，主流AI芯片架构经历了3代进化。

第1代（2012—2016年）AI芯片架构是通用图形处理器（GPGPU）。这一时期深度学习刚刚起步，网络规模并不大。这一代芯片架构没有针对神经网络计算进行加速的特殊设计，而是利用GPGPU已有的单指令多线程（SIMT）计算核心来提升向量、矩阵并行计算效率。SIMT架构特点是硬件根据数据自动分支，既可以像单指令多数据（SIMD）一样高效，又可以像多指令多数据（MIMD）一样灵活。但SIMT是为通用计算设计的，依赖共享内存交换中间数据，功耗大，算力并不高。

第2代AI芯片架构从2016年开始出现，时至今日仍是主流。这一时期卷积神经网络成为最主流的算法。AI芯片以加速卷积神网络为首要目标。ResNet成为AI芯片性能测试标准。这一代架构以谷歌张量处理器（TPU）[38]为代表，其主要特点是将AI计算抽象为标量、向量、矩阵3类计算。计算核心包含对应的3种专用计算单元，可以提供很高的峰值算力。同时核心内置容量较大的静态随机存取存储器（SRAM）作为本地存储。因此，第2代AI芯片架构在算力和功耗上相对于第1代架构有了巨大的提升。但是这一代架构在通用性上较差，在应对各种尺寸的神经网络时难以表现出很好的计算效率，同时在可编程、灵活性上不如第1代架构，面对不断涌现的新算法和新场景，日益显示出应用场景的局限性和软件开发的高成本弊端。

同一时期，GPGPU在计算核心中增加专门的矩阵计算单元Tensor Core[39]，这样既拥有高性能，又拥有强大的可编程性和灵活性，依靠完备的工具链和成熟的生态，具有突出的市场竞争力。因此，对于第2代AI芯片架构，从数量上看是百花齐放，从市场占有率上看却是一枝独秀。拥有Tensor Core的GPGPU，例如NVIDIA Volta、Ampare系列，成为这一代AI芯片的最终赢家。

第3代AI芯片架构产生于2019年，这一时期出现了Transformer模型。该模型迅速发展，并与卷积神经网络形成了分庭抗礼的局面。特别是随着大规模预训练模型和多模态的进展，Transformer很可能会最终取代卷积神经网络（CNN）。摆在芯片架构设计面前的挑战有两个：（1）需要对Transformer进行优化设计。相对于CNN，同等算力的Transformer模型对带宽的要求更高，这增加了芯片设计的难度。（2）系统需要具备优秀的水平扩展能力，以满足急速增长的大模型训练算力需求。这一代架构以GraphCore[40]、Tenstorrent[41]为代表，其特点是在单一芯片拥有上百甚至上千个计算核心。同时芯片间具备良好的水平扩展能力，可以实现从单核到百万核的无缝扩展。为保证如此大规模并行计算高效运行，需要采用软硬件协同设计，特别是需要图编译器对多核上的计算任务派发和数据路由做出优化调度，以便隐藏数据传输等额外开销，实现一加一等于二的并行计算效果。然而，这一代架构大幅增加了编译器的开发难度，芯片可编程性和灵活性相对上一代架构并未得到明显的提升，工具链和生态建设难度大。与此同时，GPGPU的TensorCore已具备专用的Tensformer加速引擎。第3代AI芯片架构中谁是最终胜利者，仍需要时间来给出答案。

算法的不断发展对AI芯片架构提出越来越高的要求。我们认为未来AI芯片架构必须要具备如下综合能力：在性能方面，对Transformer模型有优秀的加速能力；在功耗方面，8位整数（INT8）等效算力达到10 TOPS/W以上；在通用性方面，对各种规模的模型都可以达到较高的硬件利用率；在可编程性方面，可以通过编程支持新的算法且容易开发，具备完整的工具链，能够快速完成模型的开发和部署。

2 第2个10年展望

在深度学习的第2个10年，数据、算法、算力三大要素依旧占据核心地位。但随着AI的应用越来越广泛和深入，绿色、生态、可信将成为AI可持续发展新的核心要素。

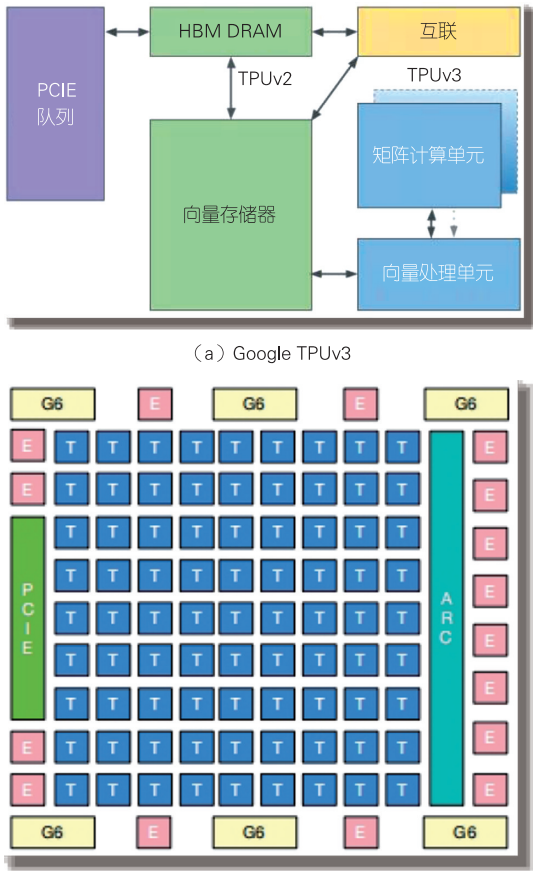
2.1 AI芯片创新实现绿色节能

2019年一项研究表明，完成一次Trasformer（Big）模型训练所排放的二氧化碳高达282吨，相当于5辆汽车整个生命周期的CO_2排放量[42]。目前，全世界1%的发电量被用于AI计算。全球AI计算能耗年增长率为37%。据此估算，下一个10年，AI计算将消耗全世界发电量的15%左右，将为环境带来沉重的负担。为了实现绿色可持续发展，必须不断研究更有效率的AI芯片。

提升AI芯片效率的一个方向是空间计算。众所周知，AI芯片功耗与数据在芯片内搬运的距离正相关。借助创新的芯片架构设计，减少完成每次操作数据在芯片内需要移动的距离，就可以大幅降低芯片的能耗。

这里我们对Google TPUv3和Tenstorrent Wormhole两个AI芯片进行对比。如图1（a）所示，TPU计算核心设计是采用一个较大的向量和矩阵计算单元同本地SRAM相连接，完成一个神经网络算子的计算，需要将数据从Vector Memory搬移到Matrix Multiply Unit完成矩阵乘计算，然后再搬移到Vector Unit完成Element-wise计算。这种大计算、大存储单元的设计导致每次计算数据平均移动距离达到毫米级别，因此芯片功耗高，以至于必须采用水冷才能使TPU集群系统正常运行。在图1（b）中，每颗Wormhole芯片包含80个Tensix计算核心。每个计算核心拥有约5 TOPS的算力以及1.5 MB的本地存储。由于大多数计算能够在单核心内完成，因此更小的核心能够缩短数据移动距离。只有少数的跨核心计算才需要将数据搬运到更远的地方。据估算，Wormhole芯片每次操作数据的平均移动距离只有TPUv3的1/10左右。因此，Wormhole芯片功效比要高得多，达到3 TOPS/W@INT8，而TPUv3的为0.6 TFOPS/W@BF16。

将一个包含大计算、大存储单元的计算核心拆分为多个包含小计算、小存储单元的计算核心，可以有效降低每次计算数据移动的平均距离，从而降低芯片能耗。这也成为新一



▲图1 Google TPUv3和Tenstorrent Wormhole架构示意图

代AI芯片的设计趋势。然而，这种多核并行计算会引入额外的开销，导致计算效率降低。相应的解决方案是通过软硬件架构协同设计，将一个计算任务拆分为多个子任务，然后将子任务指派到不同的计算核心上，并规划任务之间数据传输路径，最优匹配芯片的算力、存储、数据传输带宽、互联拓扑结构，减少数据移动距离，从而实现性能最优、功耗最低。这种将多个计算任务在空间（多核）上进行调度的计算方式被称为“空间计算”。

实现多核空间计算需要软硬件协同设计。在硬件方面，为提升并行计算效率，计算核心可以增加对AI并行计算常用通信模式的硬件支持，如Scatter、Gather、Broadcast等，对数据包进行封装、压缩等，在核间互联上优化片上网络拓扑结构和动态路由能力。在软件方面，由于空间计算的优化非常复杂，非开发人员所能负担，需要编译器自动实现任务的拆分、指派、路由规划，在运行时需要完成计算过程控制，特别是对空间计算过程中产生的各种异常（如丢包、乱序、拥塞）进行处理。

未来空间计算的一条演进路线是在存计算（At-Memory）。在存计算可以把一个大的计算核心拆分为上万个微型计算核心，而不仅仅是上百个小核心。在这种架构下，每个计算数据平均移动距离将进一步降低至微米级，功效比可以超过10 TOPS/W@INT8。例如Untether AI公司的Bo-queria[43]芯片拥有上万个处理引擎（PE）。每个PE配置6 kB本地内存，整个芯片的内存带宽高达PB/s级。PE与本地内存之间的数据移动距离仅有几微米，功效比高达30 TFOPS/W@FP8。然而，由于存在面积限制，每个PE功能简单、灵活性差，只适用于一些特定算法，目前只能进行推理，无法进行训练。此外，将计算任务部署在上万个PE上，对编译器的优化能力提出了更高的要求。

空间计算技术的另一条演进路线是确定性设计。编译器优化能力对空间计算的性能至关重要，但只能利用静态信息对计算进行调度。因此，重新设计系统的软件-硬件界面、静态-动态界面，使编译器能够利用更多的静态信息，成为一个新的技术演进方向。例如，Groq公司的张量流处理器（TSP）[44]芯片采用确定性硬件设计，芯片中没有Arbiter、Crossbar、Cache等“响应型”组件，允许编译器进行时钟级的调度。编译器可以精确地调度每个核上的计算、内存访问和数据传输，使得指令流在运行期内完全避免共享资源的访问冲突，因此可以实现无锁，系统极为高效。但是，这种确定性设计需要编译器接管到硬件状态机级别，复杂度很高。实现系统级硬件确定性非常复杂，需要实现全局时钟、链路延迟补偿、时钟漂移补偿等机制，引入硬件对齐计数器、软件对齐计数器、指令集。

随着3D封装技术的日趋成熟，空间计算还可以向3D的方向发展。将一颗大计算核心拆分为多个小核心，并在3D方向堆叠起来，可以进一步缩短数据移动的距离，从而进一步降低芯片功耗，提升能效比。此外，相对于传统2D芯片，经由3D封装技术，3D Mesh、3D torus等片上网络（NOC）拓扑更有效率，从而给编译器留下更大的调度优化空间，进一步提升空间计算性能。

提升AI芯片效率的第2个方向是近似计算。深度学习模型的一个特征是对精度要求不高。计算过程中出现的误差并不会显著影响模型的最终判定结果。近似算法可以减少内存使用和计算复杂度，使计算更加高效。

低精度计算是深度学习近似计算一个重要的技术方向。使用低精度的数据类型，可以有效减少芯片面积和功耗。例

如，INT8的乘法和加法运算所消耗的能量仅为32位浮点数（FP32）的1/30和1/15[45]。目前混合精度训练技术可以使用FP16位半精度浮点数和FP32单精度浮点数配合完成模型训练。Transformer模型的训练则可以使用更低的精度浮点数。例如，NVIDA在其最新的Hopper架构中实现了FP16和FP8混合精度训练Transformer模型[46]。未来仍有可能出现更低精度的训练算法。

由于推理对精度的要求更低，因此在完成模型训练之后，我们可以将模型转化为更低精度的数据类型表示，这个技术称之为模型量化。目前，INT8量化技术已经相当成熟，INT4量化技术仍然面临一些困难。特别是在模型中使用了非线性激活函数时，模型准确率下降很多。对此，一种思路是使用INT8和INT4自适应混合精度量化，另一种思路是将模型量化为FP8。FP8的面积和功耗仅有INT8的一半，但模型判定准确率没有明显下降。

近似计算的另一个演进路线是稀疏计算。研究发现，深度学习模型的权重存在一定的稀疏性，即部分权重值为零或者非常接近于零，特别是Transformer模型的稀疏度更大。利用模型的稀疏性可以省略不必要的计算，从而提升模型计算的效率。例如，NVIDIA A100 GPGPU中的4选2稀疏加速可以将芯片等效算力提升一倍[47]，同时功耗保持不变。Tenstorrent Wormhole芯片更是可以在模型稀疏度90%的情况下，将芯片等效算力提升100倍。未来软硬件协同下稀疏计算仍然会是一个非常有前景的技术方向。新模型的稀疏化算法、稀疏加速计算核心仍然是研究的热点。

未来10年，依靠制程提升能效比的难度越来越大，而空间计算、近似计算在提升芯片能效比方面存在巨大潜力。相对于目前的主流AI芯片，未来的芯片效能将有数十倍的提升，是AI产业实现双碳目标的有力保障。

2.2 生态融合实现降本增效

深度学习模型的研发和应用可以分为两个阶段，一是模型的训练，二是模型的应用服务。完成训练并达到业务性能要求的模型，最终形成各种形式的模型应用服务，产生商业价值。当前，从模型训练完成到部署的过程，还存在诸多痛点，无法很好的满足规模化部署的要求。

首先，目标硬件多种多样，如X86/ARM中央处理器（CPU）、GPGPU、现场可编程门阵列（FPGA）、专用集成电路（ASIC）芯片等。随着新的AI芯片层出不穷，各厂商芯片之间架构、指令集、软件工具链互不兼容，缺乏统一标准，容易引起生态碎片化问题。上层算法和应用与底层硬件紧耦合。跨硬件部署同一模型需要大量移植工作，这大幅增加了深度学习模型的研发成本和应用难度。其次，部署阶段的场景主要分为云侧、边缘侧、端侧，有基于容器化部署场景，也有基于嵌入式硬件部署的场景。不同部署场景对模型推理的性能需求、计算资源、App调用方式等要求不同。因此不同部署方案需要具备不同的技术。再次，模型开发使用的训练框架各不相同，如TensorFlow、PyTorch、Paddle - Paddle、Caffe、Keras、OneFlow。不同框架训练后保存的模型格式均不相同，在部署时需要做针对性处理，即需要一一转换到目标硬件支持的模型格式。但转换路径较为繁杂，用户需要付出较多的学习成本。

性能优化也是深度学习模型在落地时经常遇到的问题，例如计算时延高、吞吐量低、内存占用大等。在不同的应用场景和部署环境下，模型的优化目标不完全相同。例如，在端侧部署中，内存和存储空间均非常有限，模型的优化目标是减小模型的大小；在自动驾驶场景下，由于计算平台算力有限，对模型的优化侧重于在有限的算力下，尽可能提升吞吐量，降低时延。模型优化技术包括模型压缩和硬件执行优化，涉及模型剪枝、量化、稀疏化、模型中间表示（IR）、可执行文件的编译器，以及基于硬件架构的高性能计算等多项关键技术点。

为应对上述挑战，中兴通讯主导了Adlik开源项目[48]。Adlik是将深度学习模型部署至特定硬件并提供模型应用服务的端到端工具链，能够与多种推理引擎协作，提供灵活的模型加速、部署、推理方案，助力用户构建高性能AI应用。

Adlik的整体架构包括模型优化器、编译器和引擎模块。它支持各类模型在云、边、端侧多种硬件上的灵活部署和高效执行。

Adlik模型优化器支持多种结构化剪枝方法，能够有效降低模型参数量和算力需求，支持多节点、多GPU并行剪枝以提升系统效率，同时支持自动剪枝方法。用户只需要指定神经网络类型（如ResNet-50）和限制条件（如算力、延迟），模型优化器会自动决定模型每一层的通道数，得到在限制条件下最优的模型结构[49-50]。在模型量化方面，Adlik模型优化器支持8 bit量化，可以利用少量校准数据快速实现8 bit训练后量化（PTQ）；也支持量化感知训练（QAT）算法，提升量化模型精度。Adlik模型优化器提供不同的蒸馏方法，能够应用于各种深度学习任务（如图像分类、目标检测等）。如表1所示，针对ResNet-50模型优化研究，在执行剪枝、蒸馏和INT8量化后，Adlik模型推理吞吐量提升13.82倍，同时模型准确率没有降低[51]。

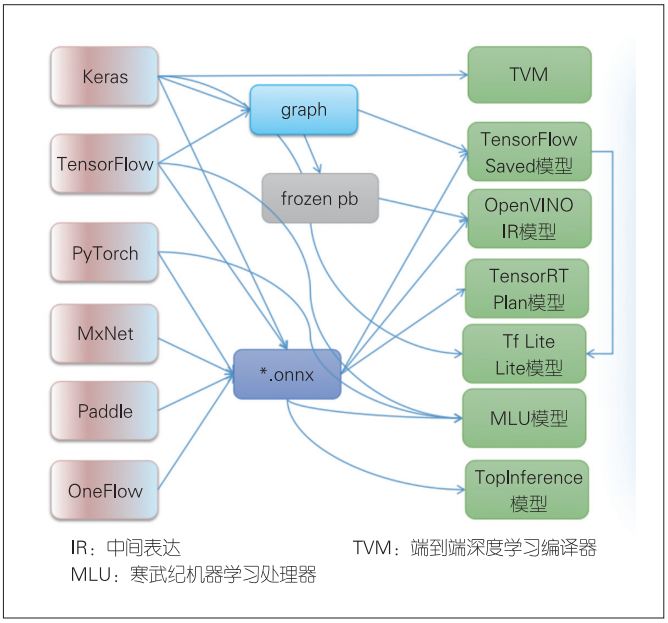
Adlik模型编译器支持不同的训练框架模型格式和推理框架模型格式之间的转换，并易于扩展，如图2所示。因

▼表1 Adlik模型优化器性能测试结果

模型优化方法	吞吐量（OpenVINO）	精度/%
基线	432	76.80
自动剪枝	1 615	73.30
自动剪枝+蒸馏	1 615	77.50
自动剪枝+蒸馏+INT8量化	6 401	77.00

INT8：8位整数　　OpenVINO：开放视觉推理及神经网络优化

此，在设计上Adlik模型编译器采用自动构建有向无环图（DAG）的方式生成源模型格式和目标模型格式的转换路线。用户只需要给出源和目标模型格式，Adlik模型编译器就可以使用最优转换路线，端到端地完成模型格式的转换[52]。目前，除了业界常用的TensorFlow和PyTorch之外，Adlik还引入了国产训练框架PaddlePaddle和OneFlow，并支持国产推理芯片厂商（寒武纪、燧原等）的推理模型格式。



▲图2 Adlik模型编译依赖图

Adlik模型应用服务存在Serving和Embedding两种方式。Adlik Serving以独立的微服务部署，支持多个客户端的推理请求服务。支持表述性状态转移（REST）和远程过程调用（RPC）接口，相关模型的版本控制和管理，可以在保持业务不中断的情况下完成模型的滚动升级。Adlik Serving的特色是以插件的方式部署和隔离各种运行时的环境，如TensorFlow、OpenVINO、Tf Lite、TensorRT、Paddle Inference等，使应用可按需加载。Serving SDK提供模型推理开发的基础类库。用户可扩展实现推理运行时的自定义开发，如实现多模型在进程内协作的推理服务、低时延嵌入式设备的推理服务等。Serving SDK提供模型上载、模型升级、模型调度、模型推理、模型监控、运行时隔离等基础模型管理功能，以及用户定制与开发推理服务的C++应用程序编程接口（API）。应用根据自身的需求，定制开发自己的模型和运行时。Serving SDK提供标准的扩展点，方便用户高效地定制新的模型和运行时环境。

Adlik支持云、边缘、端3种部署场景并提供相应的特性支持[53]：（1）在云侧，支持原生容器化部署方案、优化和编译完成的模型，可以和Adlik Serving Engine镜像一起打包，发布为应用服务镜像，并在指定硬件的容器云上运行；（2）在边缘侧，支持在启动的Adlik Serving Engine服务上加载优化和编译完成的模型，支持多模型实例调度功能，减少边缘侧计算资源的占用；（3）在端侧，支持用户优化和编译完成的模型，结合特定的计算引擎依赖库和交叉编译工具链，可编译为运行在指定硬件上的可执行文件。同时Adlik可以提供C/C++的API接口，用来提供模型编排能力，为用户提供低延时、小体积并可在指定硬件上运行的模型应用。

Adlik是对生态融合的一次尝试，用一套统一的工具链打通不同框架和硬件供应商相互割裂的生态，从而实现深度学习部署应用降本增效，为下一个10年更大规模的深度学习应用打下良好基础。未来Adlik将进一步围绕深度学习端到端性能优化、AI应用在异构平台上的部署与运行、高性能计算、模型运维等技术方向发展，持续构建社区生态，推动产业推动数字化变革，为用户打通深度学习应用的全流程，真正实现高效率、低成本的AI应用落地，助力不同行业实现智慧化转型，为数字经济发展提供强劲动力。

2.3 安全可信任实现深度应用

随着AI广泛应用于金融、交通、医疗等诸多领域，AI自身的脆弱性、黑盒等导致的安全问题和可信危机逐渐突显。例如，以色列科研人员生成的9张万能人脸可以冒充超40%的人[54]，微软聊天机器人Tay发表歧视女性相关言论[55]，没有任何犯罪记录的黑人被AI判定为更具危险性，自动驾驶汽车引发多起交通事故等。

在此背景下，世界主要国家和组织，纷纷出台AI安全和可信的法律法规、道德伦理规范和标准，用于规范和引导AI的安全生产和应用，并将AI的安全使用上升到国家战略高度。例如，中国将“促进公平、公正、和谐、安全，避免偏见、歧视、隐私和信息泄露等问题”写入《新一代AI伦理规范》[56]总则。

综合AI安全、可靠、可解释、可问责等方面的需求，

可信AI的概念被提出[57]。可信AI被业界归结为4个方面。（1）可靠性：AI系统在面临恶意攻击和干扰的情况下，能够提供正确决策和正常服务的能力；（2）隐私安全性：AI的开发和应用不能造成个人或者群体隐私信息的泄露；（3）可解释性（透明性）：AI系统的决策能够被人类用户理解，并能提供相应的解释；（4）公平性（包含个体公平性和群体公平性）：AI系统不因个体或群体差异而给出不公正的输出。因此，我们应该规范、安全地开发和使用AI，在享受技术发展带来红利的同时避免技术自身缺陷带来的负面影响。

发展可信AI意义重大，其价值主要体现在以下两个方面：

（1）有助于打破数据孤岛，充分释放数据要素价值，决定AI未来发展应用的广度和深度。一方面数据要素作为重要的战略资源，需要充分流通和共享才能释放巨大的价值，加速社会的数字化转型；另一方面数据使用过程中的隐私保护已经成为法律、法规的基本要求，例如一般数据保护条例（GDPR）[58]、《中华人民共和国网络安全法》[59]等。在此背景下发展以联邦学习[60]为代表的隐私安全机器学习方法、隐私安全计算[61]就显得尤为重要。这对打破因隐私安全造成的数据孤岛、挖掘各行各业的数据价值具有重大意义。

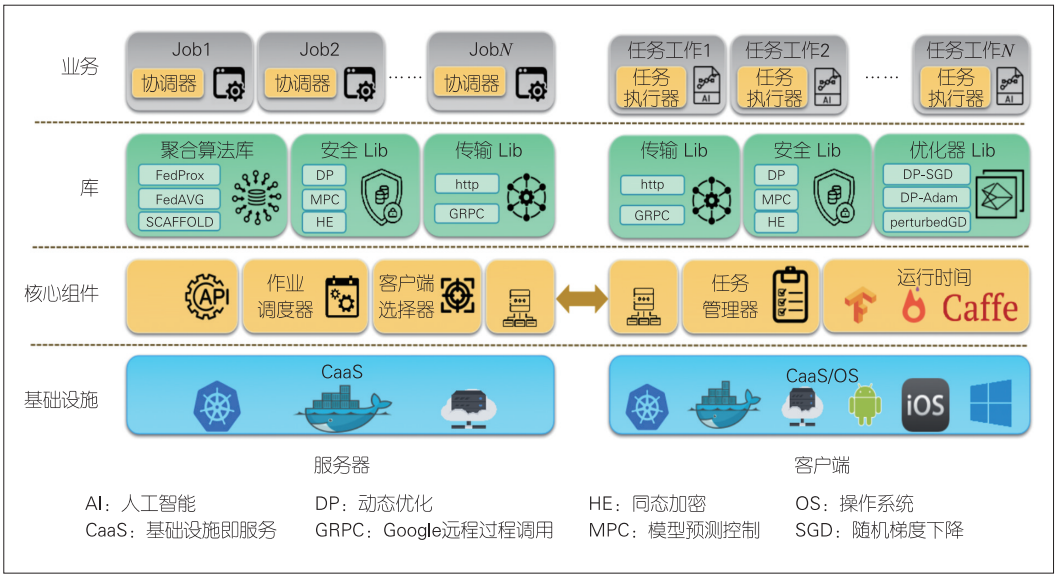
（2）安全、可靠、透明、合乎伦理规范的AI能消除人们对AI的疑虑，从而释放产业价值。AI的内生安全[62]已经引发人们的担忧，具体表现在：（a）贯穿AI生命周期、种类繁多的攻击会引起人们对可靠性的担忧，相关攻击包括对抗样本攻击、投毒攻击、后门攻击、模型窃取等[63-66]；（b）AI的黑盒特点使系统难以给出决策依据，导致在安全关键领域的决策难以被采纳；（c）AI在某些应用场景中表现出来的公平性缺失[67]，引发人们对其道德伦理的担忧。解决上述问题，构建公众对AI的信心，才能让AI被广泛接纳和使用，从而进一步扩大产业规模和价值。

综上所述，可信AI决定和影响着AI发展的可持续性和未来产业规模，而规范、法律、标准的出台更让其成为发展AI的必选项和基本准入门槛。

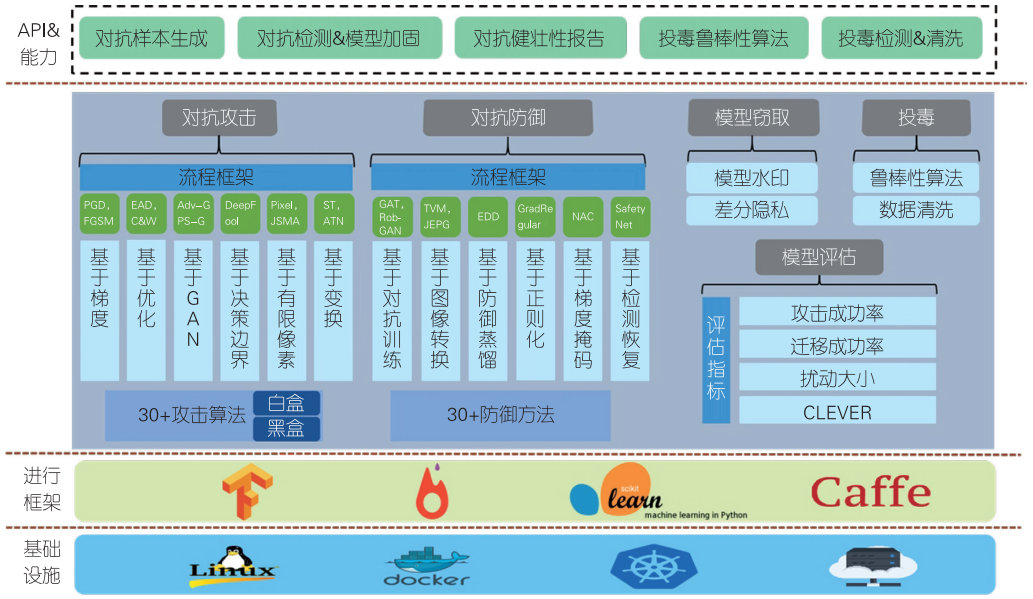
中兴通讯在可信AI方面积极投入，创建了Nuersafe开源社区（https://github.com/neursafe）。Nuersafe开源社区包含联邦学习、AI安全、AI公平和AI可解释4个平台，覆盖可信AI的4个要素。目前我们重点研究了联邦学习（Neursafe-FL[68]）和AI安全工具（Neursafe-Security）。下面我们将针对这两个方面做详细介绍。

（1）Neursafe联邦学习。该平台的目标是在隐私安全的前提下，打造可靠、高效、易用的联邦学习解决方案，如图3所示。为了实现这一目标，在设计和实现中我们应做出如下几个方面的考虑：（a）进行微服务架构设计，以满足系统灵活部署的需求，提供单机、Cross-Silo、Cross-Device 3种部署模式，并可满足科研验证、跨企业数据孤岛、海量设备联合训练等多种场景的需求。（b）拥有完备的框架能力，可提供分布式资源管理和作业调度能力，通过调度算法最大化联邦学习性能。（c）通过核心组件的高可用设计和作业级的容错处理机制，保证系统持续可服务性。（d）支持Tensorflow和Pytorch两种主流底层机器学习框架，并支持框架扩展；通过用户层极简的联邦API设计，最大程度上保留底层框架编程习惯，降低原机器学习算法向联邦学习迁移的成本。（e）封装基于差分隐私和安全多方计算等隐私算法库，并标准化算法接口，支持算法扩展。（f）提供多种算法（FedAvg、Scaffold、FedProx、FedDC等[69-72]）构成的聚合和优化算法库，满足不同数据异构场景下的收敛效率需求。

（2）Neursafe AI安全。如图4所示，该平台以工具化的方式，提供AI对抗攻击、模型鲁棒性检测以及模型加固和对



▲图3 Neursafe联邦学习架构



Adv-G：对抗生成对抗网络
API：应用程序接口
ATN：对抗变换网络
C&W：基于优化的攻击
EAD：弹性网络攻击
EDD：扩展防御蒸馏
FGSM：快速梯度符号下降法
GAN：生成对抗网络
GAT：生成对抗训练
JPEG：JPEG Compression，JPEG压缩
JSMA：基于雅克比的显著图攻击
NAC：噪声增强分类器防御
PGD：投影梯度下降法
PS-G：感知敏感生成对抗网络
Rob-GAN：一种基于GAN的防御方法
ST：空间变换攻击
TVM：总体方差最小防御

▲图4 Neursafe 人工智能安全架构

抗样本检测等能力。该平台可以实现：（a）统一服务入口，屏蔽底层算法实现，支持命令行、API和SDK 3种接口形式，一键完成对模型的对抗攻击、鲁棒性检测、安全防御加固等功能使用。（b）支持30+的黑、白盒攻击算法，其中30+的防御算法涵盖了当前主流且经典的攻防算法。（c）对当前主流的攻击和防御算法进行分类，如基于梯度的攻击、基于遗传算法的攻击、基于对抗训练的防御等，提取同类算法共性，在算法基类中实现框架代码，简化后续算法创新开发工作量。（d）支持Auto Attack，自学习攻击参数；支持多种攻防算法的正交组合，增强综合攻防能力。（e）攻防算法一次编码，兼容Tensorflow和Pytorch，解决了主要当前攻防工具支持底层框架单一问题。（f）支持模型鲁棒性检测功能，能进行模型鲁棒性的综合评估，生成界面优化的评估报告。（g）支持模型加固、对抗样本检测、对抗样本恢复3种防御手段，满足不同场景下的安全防御需求，在模型已经上线运行的情况下可以通过增加前置检测网络来实现安全防御。

可信AI的研究进展对AI的可持续发展至关重要。中兴通讯将继续关注可信AI，对未来可信AI的研究工作有如下规划：（1）将坚持开源运作，和业界一起共筑可信AI未来；（2）补齐当前在公平和可解释方面的缺失，构建可信AI的全方位能力；（3）针对可信AI中的问题，如联邦学习中的性能问题、AI安全中的碎片化问题，跟踪业界最新进展，对算法进行创新研究，逐步扫除解决方案的落地障碍；（4）坚持产品化的思维，站在用户角度，提供简单、用户友好的解决方案。

3 结束语

经过3次发展浪潮，AI已经快速走出低谷期。在第2个10年学术研究、产业落地的双轮驱动下，研究者数量、论文数量、数据量、算力、产业规模等维度将保持指数增长态势。绿色、高效、安全是下一个10年深度学习维持可持续指数增长的3个新的核心要素，是实现中国新一代AI发展规划三步走[73]、2030年AI核心产业突破十万亿元的关键。

参考文献

[1] 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-Volume. ACM, 2012: 1097-1105. DOI: 10.5555/2999134.2999257
[2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90
[3] VASWANI A, SHAZEER N, PARMAR N, et al. 2017. Attention is all you need [EB/OL]. [2022-10-12]. https://arxiv.org/abs/1706.03762
[4] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735
[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-10-12]. https://aclanthology.org/N19-1423/
[6] LIU X, ZHANG F J, HOU Z Y, et al. Self-supervised learning: generative or contrastive [J]. IEEE transactions on knowledge and data engineering, 2022, 35(1): 857-876. DOI: 10.1109/TKDE.2021.3090866
[7] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-10-12]. https://arxiv.org/abs/2005.14165
[8] 马子轩, 翟季, 韩文弢, 等. 高效训练百万亿参数预训练模型的系统挑战和对策[J]. 中兴通讯技术, 2022, 28(2): 51-58
[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. [2022-10-12]. https://arxiv.org/abs/2010.11929
[10] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation [EB/OL]. [2022-10-12]. https://arxiv.org/abs/2102.12092
[11] LIU J, ZHU X, LIU F, et al. OPT: omni-perception pre-trainer for cross-

modal understanding and generation [EB/OL]. [2022−10−12]. https://arxiv.org/abs/2107.00249
[12] LIN J, MEN R , YANG A, et al. M6: a Chinese multimodal pretrainer [EB/OL]. [2022−10−12]. https://arxiv.org/abs/2103.00823
[13] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484−489. DOI: 10.1038/nature16961
[14] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676): 354−359. DOI: 10.1038/nature24270
[15] SILVER D, HUBERT T, SCHRITTWIESER J, et al. Mastering chess and shogi by self−play with a general reinforcement learning algorithm [EB/OL]. [2022−10−12]. https://arxiv.org/abs/1712.01815
[16] BROWN N, SANDHOLM T. Safe and nested subgame solving for imperfect−information games [EB/OL]. [2022−10−12]. https://arxiv. org/abs/1705.02955
[17] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning [EB/OL]. [2022−10−12]. https://arxiv. org/abs/1912.06680
[18] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi−agent reinforcement learning [J]. Nature, 2019, 575(7782): 350−354. DOI: 10.1038/s41586−019−1724−z
[19] WU Z, LIAN W Z, UNHELKAR V, et al. Learning dense rewards for contact−rich manipulation tasks [EB/OL]. [2022−10−12]. https://arxiv. org/abs/2011.08458
[20] DEGRAVE J, FELICI F, BUCHLI J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning [J]. Nature, 2022, 602(7897): 414−419. DOI: 10.1038/s41586−021−04301−9
[21] FAWZI A, BALOG M, HUANG A, et al. Discovering faster matrix multiplication algorithms with reinforcement learning [J]. Nature, 2022, 610 (7930): 47−53. DOI: 10.1038/s41586−022−05172−4
[22] GOODFELLOW I, POUGET−ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139−144. DOI: 10.1145/3422622
[23] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis [EB/OL]. [2022−10−12]. https://arxiv. org/abs/1809.11096
[24] KINGMA D P, WELLING M. Auto−encoding variational bayes [EB/OL]. [2022−10−12]. https://arxiv.org/abs/1312.6114
[25] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [EB/OL]. [2022−10−12]. https://arxiv.org/abs/2006.11239
[26] ROMBACH R, BLATTMANN A, LORENZ D, et al. High−resolution image synthesis with latent diffusion models [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10674−10685. DOI: 10.1109/CVPR52688.2022.01042
[27] 熊先奎, 袁进辉, 宋庆春. 面向分布式AI的智能网卡低延迟Fabric技术 [J]. 中兴通讯技术, 2020, 26(5): 23−28
[28] SHI W Q, SUN Y X, HUANG X F, et al. Scheduling policies for federated learning in wireless networks networks: an overview [J]. ZTE communications, 2020, 18(2): 11−19
[29] 曹晓雯, 莫小鹏, 许杰. 面向边缘智能的空中计算 [J]. 中兴通讯技术, 2020, 26 (4): 31−37. DOI: 10.12142/ZTETJ.202004007
[30] YANG H H, ZHAO Z Y, QUEK T Q S. Enabling intelligence at network edge: an overview of federated learning [J]. ZTE communications, 2020, 18 (2): 2−10. DOI: 10.12142/ZTECOM.202002002
[31] 朱近康. 知识+数据驱动学习: 未来网络智能的基础 [J]. 中兴通讯技术, 2020, 26(4): 46−49. DOI: 10.12142/ZTETJ.202004011
[32] LIU W C, SHEN M Q, ZHANG A D, et al. Artificial intelligence rehabilitation evaluation and training system for degeneration of joint disease [J]. ZTE communications, 2021, 19(3): 46−55. DOI: 10.12142/ZTECOM.202103006
[33] 程强, 刘姿杉. 数据驱动的智能电信网络 [J]. 中兴通讯技术, 2020, 26(5): 53−56. DOI: 10.12142/ZTETJ.202005010
[34] ZHANG C C, ZHANG N, CAO W, et al. AI−based optimization of handover strategy in non−terrestrial networks [J]. ZTE Communications, 19(4): 98−104, DOI: 10.12142/ZTECOM.202104011
[35] YANG K, ZHOU Y, YANG Z P, et al. Communication−efficient edge AI inference over wireless networks [J]. ZTE communications, 2020, 18(2): 31−39
[36] 李高, 王威, 吴启晖. 面向低轨卫星的频谱认知智能管控 [J]. 中兴通讯技术, 2021, 27(5): 7−11. DOI: 10.12142/ZTETJ.202105003
[37] SEVILLA J, HEIM L, HO A, et al. Compute trends across three eras of machine learning [C]//Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1−8. DOI: 10.1109/IJCNN55064.2022.9891914
[38] JOUPPI N P, YOUNG C, PATIL N, et al. In−datacenter performance analysis of a tensor processing unit [C]//Proceedings of 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2017: 1−12
[39] NOVID. Nvida tesla V100 GPU architecture [EB/OL]. [2022−10−12]. https://images. nvidia. com/content/volta−architecture/pdf/volta−architecture−whitepaper.pdf
[40] JIA Z, TILLMAN B, MAGGIONI M, et al. Dissecting the graphcore IPU architecture via microbenchmarking [EB/OL]. [2022−10−12]. https://arxiv.org/abs/1912.03413
[41] PATEL, D. Tenstorrent wormhole analysis: a scale out architecture for machine learning that could put Nvidia on their back foot [EB/OL]. [2022−10−12]. https://www. semianalysis. com/p/tenstorrent−wormhole−analysis−a−scale
[42] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP [EB/OL]. https://arxiv. org/abs/1906.02243
[43] BEACHLER R, SNELGROVE M. Untether ai: boqueria [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS). IEEE, 2022: 1−19. DOI: 10.1109/HCS55958.2022.9895618
[44] ABTS D, KIM J, KIMMELL G, et al. The Groq Software−defined Scale−out Tensor Streaming Multiprocessor: from chips−to−systems architectural overview [C]//Proceedings of 2022 IEEE Hot Chips 34 Symposium (HCS). IEEE, 2022: 1−69. DOI: 10.1109/HCS55958.2022.9895630
[45] HOROWITZ M. 1.1 Computing's energy problem (and what we can do about it) [C]//Proceedings of 2014 IEEE International Solid−State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014: 10−14. DOI: 10.1109/ISSCC.2014.6757323
[46] ANDERSCH M. Nvidia Hopper architecture in−depth [EB/OL]. [2022−10−12]. https://developer. nvidia. com/blog/nvidia−hopper−architecture−in−depth
[47] POOL J. Accelerating inference with sparsity using the Nvidia ampere architecture and NVIDIA TENSORRT [EB/OL]. [2022−10−12]. https://developer. nvidia. com/blog/accelerating−inference−with−sparsity−using−ampere−and−tensorrt
[48] GITHUB. Lfai−landscape [EB/OL]. [2022−10−12]. https://github. com/lfai/lfai−landscape
[49] 王成灿. Adlik在模型剪枝量化上的实践 [EB/OL]. [2022−10−12]. https://zhuanlan.zhihu.com/p/197837122
[50] 潘佳懿. 模型优化算法 [EB/OL]. [2022−10−12].https://zhuanlan.zhihu.com/p/543576964
[51] 中兴通讯, 英特尔. 英特尔联手中兴优化深度学习模型推理 实现降本增效[EB/OL]. [2022−10−12]. https://wiki. lfaidata. foundation/display/ADLIK/Discussion+materials
[52] 韩雪微. Adlik 深度学习模型编译器介绍[EB/OL]. [2022−10−12]. https://zhuanlan.zhihu.com/p/368595113
[53] 刘涛. 异构计算系列(三): Adlik在深度学习异构计算上的实践 [EB/OL]. [2022−10−12]. https://www.infoq.cn/article/eg4kwzd1uofwjssuzfgt
[54] SHMELKIN R, FRIEDLANDER T, WOLF L. Generating master faces for dictionary attacks with a network−assisted latent space evolution [C]// Proceedings of 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE, 2022: 1−8. DOI: 10.1109/FG52635.2021.9666968
[55] Wikipedia. Tay [EB/OL]. [2022−10−12]. https://en.wikipedia.org/wiki/Tay
[56] 中华人民共和国科学技术部. 新一代AI治理原则 [EB/OL]. [2022−10−12]. https://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html
[57] 中国信息通信研究院, 京东探索研究院. 可信AI白皮书 [R]. 2021
[58] VOIGT P, BUSSCHE A V D. The EU general data protection regulation (GDPR): a practical guide [M]. Cham: Springer International Publishing, 2017
[59] 中华人民共和国网络安全法 [EB/OL]. [2022−10−12]. http://www.cac.gov.cn/2016−11/07/c_1119867116.htm
[60] YANG Q, LIU Y, CHEN T J, et al. Federated machine learning: concept and applications [J]. ACM transactions on intelligent systems and technology, 2019, 10(2): 1−19. DOI: 10.1145/3298981
[61] LI F H, LI H, NIU B, et al. Privacy computing: concept, computing framework, and future development trends [J]. Engineering, 2019, 5(6): 1179−1192. DOI: 10.1016/j.eng.2019.09.002
[62] 方滨兴, 崔翔, 顾钊铨. 人工智能安全论述[EB/OL]. [2022−10−12]. https://tx.pcl.ac.cn/CN/article/openArticlePDF.jsp?id=15

[63] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. [2022-10-12]. https://arxiv.org/abs/1312.6199
[64] JAGIELSKI M, OPREA A, BIGGIO B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning [C]// Proceedings of 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018: 19-35. DOI: 10.1109/SP.2018.00057
[65] GAO Y S, DOAN B G, ZHANG Z, et al. Backdoor attacks and countermeasures on deep learning: a comprehensive review [EB/OL]. [2022-10-12]. https://arxiv.org/abs/2007.10760
[66] TRAMÈR F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction APIs [C]//Proceedings of the 25th USENIX Conference on Security Symposium. ACM, 2016: 601-618. DOI: 10.5555/3241094.3241142
[67] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning [J]. ACM computing surveys, 2022, 54(6): 1-35. DOI: 10.1145/3457607
[68] TANG B, ZHANG C M, WANG K W, et al. Neursafe-FL: a reliable, efficient, easy-to-use federated learning framework [EB/OL]. [2022-10-12]. http://kns.cnki.net/kcms/detail/34.1294.TN.20220826.1322.001.html
[69] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [EB/OL]. [2022-10-12]. https://arxiv.org/abs/1602.05629
[70] KARIMIREDDY S P, KALE S, MOHRI M, et al. Scaffold: stochastic controlled averaging for federated learning [EB/OL]. [2022-10-12]. https://arxiv.org/abs/1910.06378v3
[71] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks [EB/OL]. [2022-10-12]. https://arxiv. org/abs/1812.06127
[72] GAO L, FU H Z, LI L, et al. FedDC: federated learning with non-IID data via local drift decoupling and correction [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10102-10111. DOI: 10.1109/CVPR52688.2022.00987
[73] 国务院. 国务院关于印发新一代AI发展规划的通知 [EB/OL]. [2022-10-12]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

作 者 简 介

韩炳涛，中兴通讯股份有限公司数据智能平台中心总工程师、移动网络和移动多媒体技术国家重点实验室多媒体研究中心副主任、Linux深度学习基金会Adlik项目负责人；研究方向为机器学习平台技术和网络智能化，以及相关核心系统架构和AI算法；拥有发明专利多项，出版专著多部。

刘涛，中兴通讯股份有限公司资深算法专家、Adlik开源项目首席架构师、AI预研项目经理；主要研究领域为AI模型并行训练、模型推理优化、高性能计算、异构硬件模型部署等；拥有多项发明专利。

唐波，中兴通讯股份有限公司资深系统架构师；主要研究领域为深度学习技术、异构资源调度、隐私安全机器学习、AI安全攻防等；主导中兴通讯公司AI平台、联邦学习、AI安全工具的设计和研发工作；拥有多项发明专利，发表论文多篇。

综合信息

《中兴通讯技术》2023年热点专题预告

期 次	专题名称	策划人
1	云网安全新挑战及智能防护技术	中国电信研究院教授级高工 解冲锋 北京邮电大学教授 杨义先
2	语义通信	清华大学教授 陶晓明 中国科学院院士 陆建华
3	数字孪生	重庆邮电大学教授、副校长 陈前斌
4	算力网络和东数西算	工业和信息化部通信科技委专职常委 赵慧玲
5	6G网络技术	北京邮电大学教授 王文东
6	面向双碳的新一代无线通信网络	华中科技大学教授 葛晓虎 西安电子科技大学教授 李建东