AI 与边缘计算结合的双向优化

在面向物联网、大流量等场景下，为了满足更广连接、更低时延、更好控制等需求，云计算在向一种更加全局化的分布式节点组合形态进阶，边缘计算是其向边缘侧分布式拓展的新触角。

作者|华为云原生团队

以物联网场景举例，设备产生大量数据，上传到云端进行处理，会对云端造成巨大压力，为分担云端的压力，边缘计算节点可以负责自己范围内的数据计算。

同时，经过处理的数据从边缘节点汇聚到中心云，云计算做大数据分析挖掘、数据共享，同时进行算法模型的训练和升级，升级后的算法推送到边缘，使边缘设备更新和升级，完成自主学习闭环。

对于边缘AI总体来说，核心诉求是高性能、低成本、高灵活性。其技术发展趋势可总结为以下几点：

可编程性、通用性；

伸缩性，同一个架构支持不同场景

低功耗，适应更多边缘场景的环境和电力要求

软硬件深度结合

高效的分布式互联和协作计算能力

笔者分别从边缘计算AI加速、端/边/云协同以及边缘计算AI框架等三个部分继续深入剖析AI应用与边缘计算结合之后的双向优化，进一步优化AI应用的用户体验。

01

边缘计算AI加速

针对基于边缘计算场景进行AI加速，笔者参考相关论文认为大致可归结为以下四个方面：

云边协同（云端训练、边缘推理）

为弥补边缘设备计算、存储等能力的不足，满足人工智能方法训练过程中对强大计算能力、存储能力的需求，研究人员提出云计算和边缘计算协同服务架构。如下图所示，研究人员提出将训练过程部署在云端，而将训练好的模型部署在边缘设备。显然，这种服务模型能够在一定程度上弥补人工智能在边缘设备上对计算、存储等能力的需求。

模型分割（云边协同推理）

为了将人工智能方法部署在边缘设备，如下图提出了切割训练模型，它是一种边缘服务器和终端设备协同训练的方法。它将计算量大的计算任务卸载到边缘端服务器进行计算，而计算量小的计算任务则保留在终端设备本地进行计算。显然，上述终端设备与边缘服务器协同推断的方法能有效地降低深度学习模型的推断时延。然而，不同的模型切分点将导致不同的计算时间，因此需要选择最佳的模型切分点，以最大化地发挥终端与边缘协同的优势。

模型裁剪

为了减少人工智能方法对计算、存储等能力的需求，一些研究人员提出了一系列的技术，在不影响准确度的情况下裁剪训练模型，如在训练过程中丢弃非必要数据、稀疏数据表示、稀疏代价函数等。下图展示了一个裁剪的多层感知网络，网络中许多神经元的值为零，这些神经元在计算过程中不起作用，因而可以将其移除，以减少训练过程中对计算和存储的需求，尽可能使训练过程在边缘设备进行。在参考文献中，作者也提出了一些压缩、裁剪技巧，能够在几乎不影响准确度的情况下极大地减少网络神经元的个数。

设计轻量级加速体系架构

在工业界，有很多公司开始研究低功耗加速芯片。如寒武纪公司推出的思元系列及华为公司推出的昇腾系列，能够适配并兼容多样化的硬件架构，进而支撑边缘计算典型的应用场景。

在学术界，对于边缘AI硬件的设计工作主要集中在提高深度神经网络及相关算法如CNN、FCN和RNN等的计算性能和效率。研究人员利用神经网络的冗余和弹性等特性来优化计算操作和数据移动，以降低NN算法在专用硬件上的功耗并提高性能。下表总结了一些低功耗机器学习处理器的相关情况。

AI在过去几年中，为互联网应用、工业互联网、医学和生物学及自动驾驶等领域带来了突飞猛进的进展。同时，随着边缘计算的逐步成熟，业界必将更加关注边缘计算AI加速方面的研究进展。

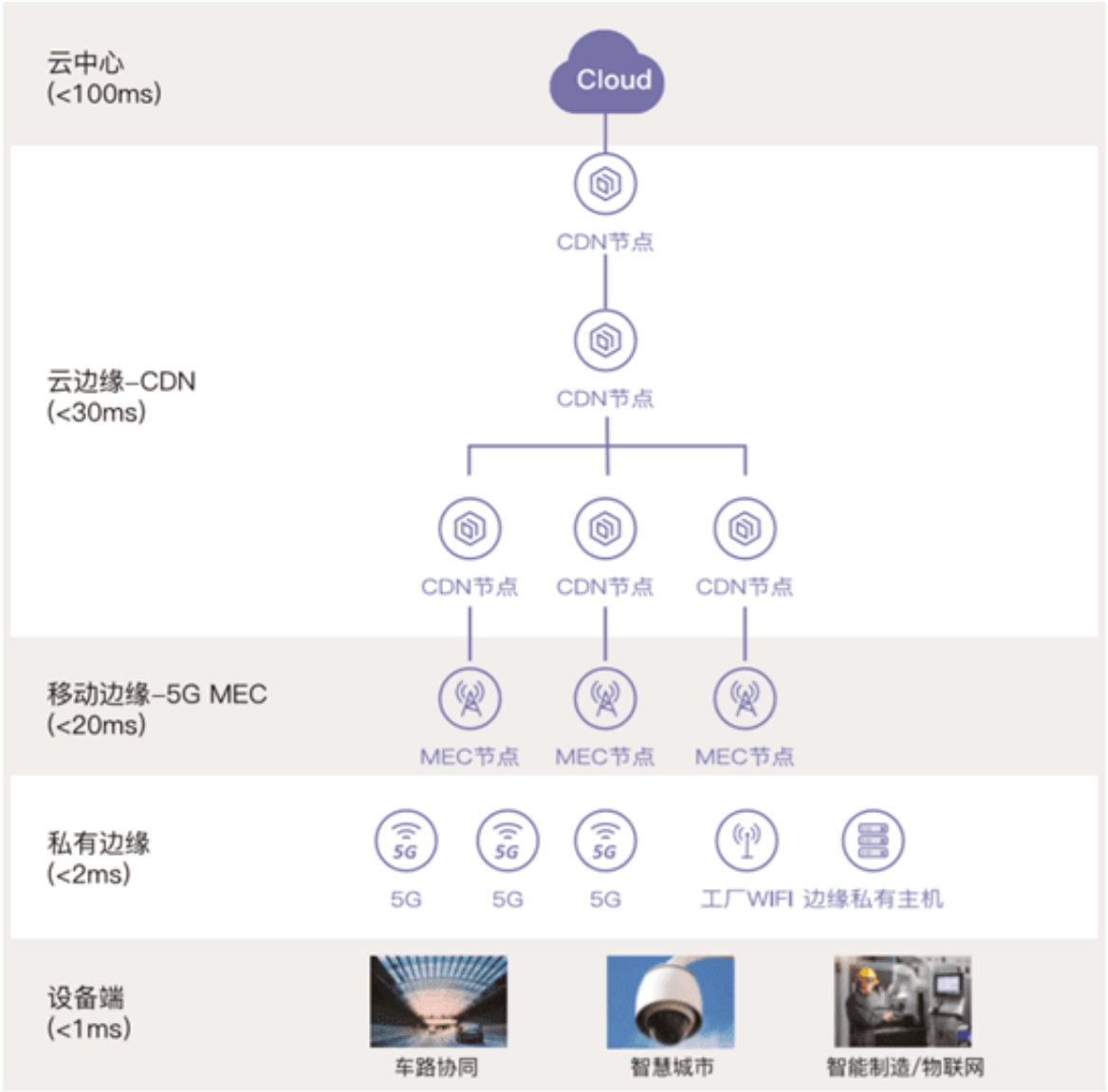
由于边缘计算场景的特点，其硬件的异构化程度会显着高于传统数据中心，对现有计算框架也会有非常大的挑战。如何快速支持异构的计算芯片并保证计算的高效，也非常值得产业内的研发力量持续投入。

02

端/边/云协同

资源协同

对于边缘计算，需要对计算资源和网络资源有全局的判断，比如边缘设备、边缘节点及中心云资源的使用情况，站在全局角度，进行资源的合理分配，确保性能、成本、服务最优。



数据协同

边缘AI会处理用户的数据，可以从两个维度来考虑。

一方面， 横向考虑，边缘的网络环境多种多样，终端用户设备具有移动性，可能会从一个服务节点移动到另一个服务节点，从一个边缘移动到另一个边缘，从WiFi切换为5G移动网络，甚至从一个运营商切换到另一个运营商，那么用户在旧的环境中产生的数据如何与新环境中的AI程序进行同步会成为一个问题。这里的数据协同不仅需要技术上的支持，更需要商业模式上的支持。

另一方面 ，纵向考虑，如下图所示，用户在边缘侧产生的数据按照隐私级别可以分为不同类型，如User-Private、Edge-Private、Public等，这些数据可以自下而上分层储存在云边协同系统中的不同层级的数据库中，同时也可以对应不同算力支持的边缘AI的访问权限，例如可以允许云上运行的AI程序读取Public数据来训练一个通用的模型，在边缘侧的AI可以读取Edge-Private数据来在通用模型的基础上训练边缘模型等等。

算力协同

通过合理的模型拆解，将不同的服务模型根据资源、成本、质量、时延等要求部署在合适的位置。通过完成的协同计算框架，确保各子模型之间的协同处理。比如结合产品设计，我们可以将简单的识别推理全部置于端侧设备，如需要判断视频中的物体属于动物还是植物等。

但是进一步的识别功能，我们可以结合边缘侧的推理能力，识别动物为猫科动物或犬科动物等。如果用户需要更加精细的识别，我们可以将边缘侧的识别结果及处理之后得到的特征数据发送至云端，结合云端完善的数据模型和知识体系，将该猫科动物判定为是东北虎还是华南虎。这样通过端、边、云三者的协同，能够在极大保证用户体验的同时，合理的使用各类资源。

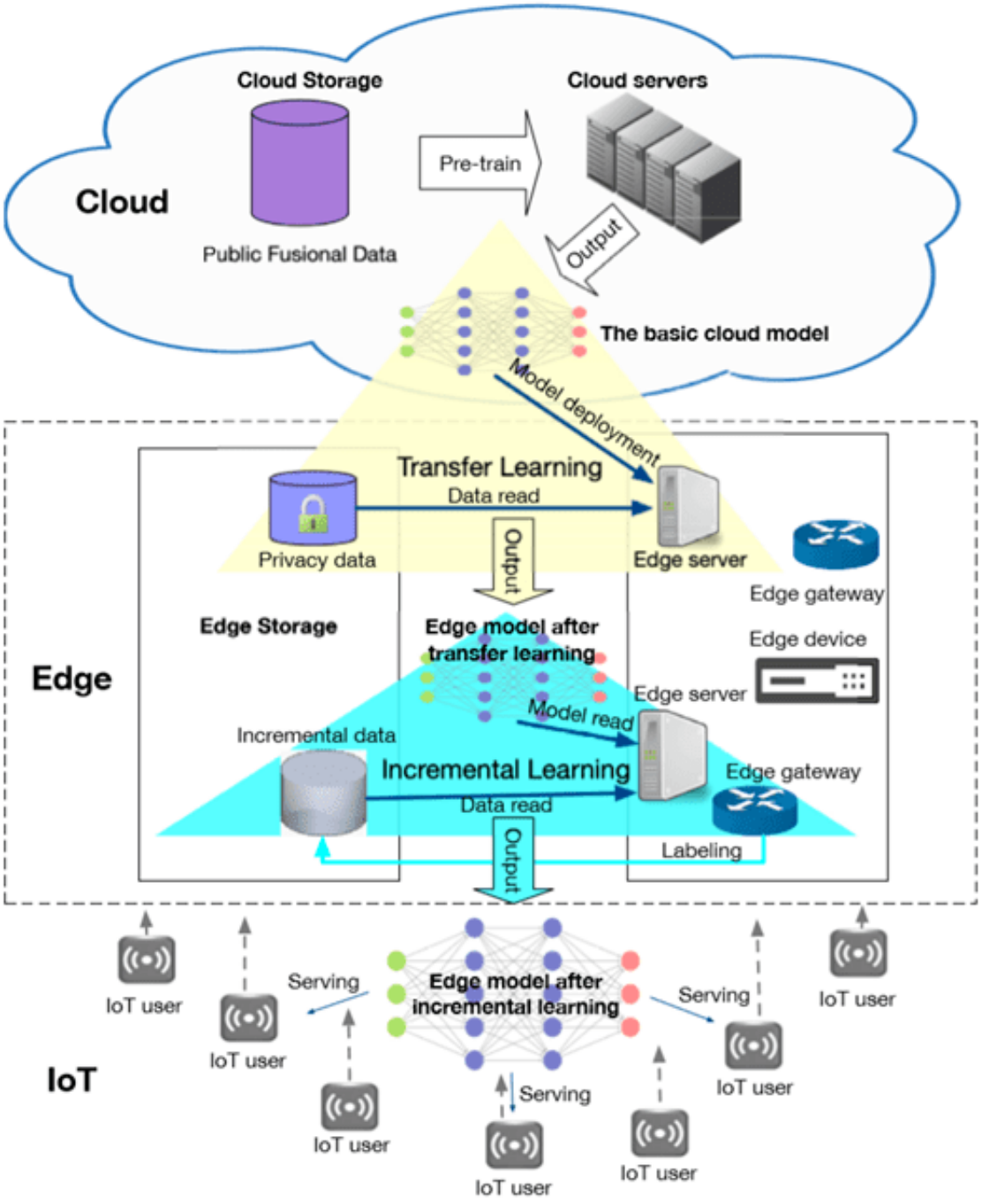
合理利用算力协同，也能够做到在边缘侧进行训练。目前工业界还没有成熟的模式，但学术界有相关的研究。如下图所示的ICE智能协同计算框架，将边缘AI的训练分为三个阶段：

第一阶段为预训练阶段（pre-train），云上的应用可以通过读取云端存储的公共数据，来训练一个通用的模型，这是边缘AI的底座，如果没有云端强大算力的帮助，只靠边缘侧算力是难以得到比云端训练更优秀的AI模型。

第二阶段将通用模型下发至边缘侧，读取边缘私有数据，通过转移学习（Transfer Learning），来得到边缘模型。

第三阶段读取增量数据，利用增量学习（Incremental Learning）生成最终边缘模型，这个最终边缘模型就可以用于用户侧的推理了。

这种三步学习的算力协同的模式，可以更好地满足边缘智能的个性化服务需求。



通过资源协同、算力协同以及数据协同，边缘智能能够高效合理的利用端、边、云的各类资源，极大地优化AI应用在边缘计算场景下的用户体验，进一步放大AI应用的商业价值。

03

边缘计算AI计算框架

在云数据中心中，算法执行框架更多地执行模型训练任务，在训练过程中需要接收大规模、批量化的信息数据，比较关注训练的迭代速度和收敛率等。而在边缘设备上更多的是执行预测任务，输入一般是实时的小规模数据，大多数场景下执行预测任务，因此更加关注于预测算法执行速度及端侧或边缘侧的资源开销。

目前业界针对边缘计算场景，也提出了针对性的设计方案，例如用于移动设备或嵌入式设备的轻量级解决方案TensorFlow Lite，Caffe2和Pytorch等。

TensorFlow Lite

TensorFlow Lite 提供了转换 TensorFlow
模型，并在移动端（mobile）、嵌入式（embeded）和物联网（IoT）设备上运行 TensorFlow
模型所需的所有工具。

特点：

只含推理（inference）功能，使用的模型文件需要通过离线的方式训练得到。

最终生成的模型文件较小，均小于500kB。

为了提升执行速度，都使用了ARM NEON指令进行加速。

支持跨平台，包括Linux、Android和iOs。

Caffe2

Caffe2 是一个兼具表现力、速度和模块性的深度学习框架，是 Caffe 的实验性重构，能以更灵活的方式组织计算。Caffe2可帮助开发人员和研究人员训练大规模机器学习模型，并在移动应用中提供 AI 驱动的用户体验。现在，开发人员可以获取许多相同的工具，能够在大规模分布式场景训练模型，并为移动设备创建机器学习应用。

特点：

可以在iOS系统、Android系统和树莓派（Raspberry Pi）上训练和部署模型；

使用比较简单，只需要运行几行代码即可调用Caffe2中预先训练好的Model Zoo模型；

NVIDIA（英伟达），Qualcomm（高通），Intel（英特尔），Amazon（亚马逊）和Microsoft（微软）等公司的云平台都已支持Caffe2；

PyTorch

PyTorch 是最新的深度学习框架之一，由 Facebook 的团队开发，并于 2017 年在 GitHub 上开源。PyTorch 很简洁、易于使用、支持动态计算图而且内存使用很高效，因此越来越受欢迎。

特点：

改进现有的神经网络，提供了更快速的方法——不需要从头重新构建整个网络，这是由于 PyTorch 采用了动态计算图（dynamic computational graph）结构，而不是大多数开源框架（TensorFlow、Caffe、CNTK、Theano 等）采用的静态计算图；

强大的社区支持，facebook的FAIR强力支持，FAIR是全球TOP3的AI研究机构。PyTorch论坛，文档，tutorial，一应俱全。FAIR的几位工程师更是全职维护开发，github上PyTorch每天都有许多pull request和讨论。

支持iOS系统、Android系统运行

这些边缘AI执行框架通过优化移动应用程序内核、预先激活和量化内核等方法来减少执行预测任务的延迟和内存占用量。

此外，边缘计算在AI训练提速、安全信息预处理、边云一体的AI算法上仍处于起步阶段。设计面向轻量级、高效和可扩展的边缘计算AI框架是实现边缘智能，极大拓展更多边缘AI场景落地的重要步骤。

04

结语

AI和边缘计算已获得国内外学术界和工业界的广泛关注和认可，并且已经在很多商业场景下发挥作用。将AI应用部署至边缘已成为提升智能服务的有效途径。

尽管目前边缘智能仍处于发展的初期，然而，笔者相信，边缘智能够产生极大的促进效果，并成为各行各业的黏合剂和智能产业发展的催化剂，促进多个行业的升级转型。