智能边缘计算：计算模式的再次轮回

原创：刘云新 微软研究院AI头条 前天

编者按：人工智能的蓬勃发展离不开云计算所带来的强大算力，然而随着物联网以及硬件的快速发展，边缘计算正受到越来越多的关注。未来，智能边缘计算将与智能云计算互为补充，创造一个崭新的智能新世界。本文中，微软亚洲研究院系统与网络研究组首席研究员刘云新将为大家介绍智能边缘计算的发展与最新研究方向。

智能边缘计算的兴起

近年来，边缘计算（Edge Computing）在学术界和工业界都成为了一个热门话题。事实上，边缘计算是相对于云计算（Cloud Computing）而言的。在云计算中，所有的计算和存储资源都集中在云上，也就是数据中心（Datacenter）里；在终端设备上产生的数据通过网络传输到云上，计算任务和数据处理都在云上进行。而**在边缘计算中，计算和存储资源被部署到边缘上（边缘服务器或者终端设备），可以就近对本地的数据进行处理，无需把数据传输到远端的云上，从而避免网络传输带来的延迟。**

虽然边缘计算成为广受关注的热门话题的时间并不久，但边缘计算的概念并不新。早在2008年，微软研究院的 Victor Bahl 博士邀请了学术界和工业界的知名学者，包括卡内基·梅隆大学的 Mahadev Satyanarayanan 教授、AT&T 实验室的 Ramón Cáceres博士、兰卡斯特大学（Lancaster University, U.K.）的Nigel Davies教授、英特尔研究院（Intel Research）的 Roy Want 博士等，一起探讨云计算的未来时 [1]，就提出了基于 Cloudlet 的边缘计算的概念；并于次年在 IEEE Pervasive Computing 期刊上发表了广为人知的名为 “The Case for VM-based Cloudlets in Mobile Computing”的文章 [2]。

此后，越来越多的研究人员开始关注边缘计算。值得一提的是，2016年，首届专注于边缘计算的学术会议 The First IEEE/ACM Symposium on Edge Computing 在美国华盛顿特区召开 [3]。目前，边缘计算已成为相关顶级学术会议（比如MobiCom）的重要专题之一。在工业界，2017年微软公司 CEO 萨提亚·纳德拉就将边缘计算和云计算并列成为全公司的战略之一。之后，各大云计算公司和运营商都纷纷推出了自己的边缘计算服务；边缘计算相关的创业公司更是不断涌现。

在人工智能时代，边缘计算不仅仅只是计算，更是智能+计算，我们称之为**智能边缘计算（Intelligent Edge Computing）**。

计算模式的轮回：在集中式和分布式之间的摇摆

唯物辩证法指出，事物的发展总是曲折、循环往复，并在波浪中不断前进的。计算模式（Computing Paradigm）也不例外。如图1所示，如果我们回顾计算模式的发展历史，就会发现一个简单的规律：计算模式是在集中式计算和分布式计算之间不断摇摆，往复式发展前进的。

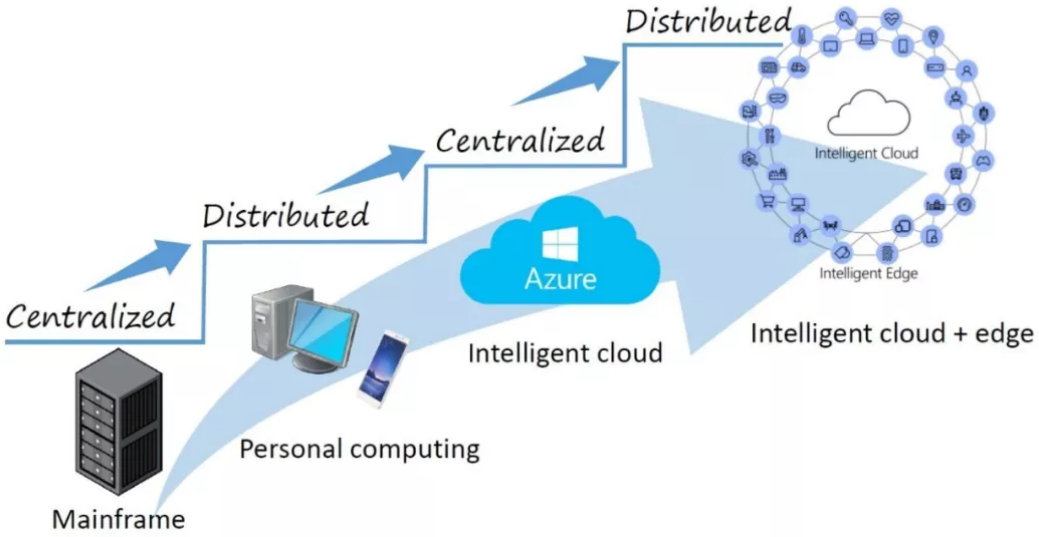

图1：计算模式的发展历史

在大型机（Mainframe）时代，计算资源稀缺，很多人共享一台主机，计算是集中式的；到了个人计算（Personal Computing）时代，硬件变得小型化，价格低廉，人们可以拥有自己的个人设备，计算成为了分布式的；在云计算时代，通过高速网络，人们可以共享云上的海量的计算和存储资源，计算模式又回到集中式的。此时，人工智能蓬勃发展，云上提供的众多智能服务带来了智能云计算。而随着边缘计算的出现，计算模式再一次成为分布式的。现在，我们不仅有智能云，还有智能边缘。

智能边缘计算的出现当然不仅仅是满足表面上的简单规律，背后有其必然性和强大的驱动力，是计算机软硬件和新应用新需求不断发展的必然结果。

首先，随着物联网特别是智能物联网（AIoT）的发展，各种新型智能设备不断涌现，产生了海量的数据。比如，监控摄像头已经无处不在（据统计，在伦敦每14个人就有一个监控摄像头 [4]），每天产生大量的视频数据。而每辆自动驾驶汽车每天更是会产生多达5TB的数据。把所有这些数据都传输到云上进行处理是今天的云和网络无法承受的。

其次，新的场景和应用需要对数据在本地进行处理。比如，自动驾驶和工业自动化对数据处理的实时性有很高的要求。数据传输带来的网络延迟往往无法满足实时性的要求，如果网络发生故障可能带来灾难性后果。再如，人们对个人隐私越来越关注，而很多数据（视频、图片、音频等）都包含大量的个人隐私。保护个人隐私的最好的方法就是在本地进行数据处理，不把个人数据传到网络上去。

另外，同样重要的是，**硬件的快速发展使得智能边缘计算成为可能。**随着 AI 算法的日益成熟，人们开始设计制造专用的 AI 芯片，特别是专门用于深度学习模型推理的 AI 芯片，这些 AI 芯片不仅数据处理能力强大，而且尺寸小、功耗低、价格便宜，可以应用到各种边缘设备上，为智能边缘计算提供了坚实的硬件基础。

需要指出的是，**智能边缘计算并不是要取代云计算，而是和云计算互为补充**，一起更好地为用户提供服务。云计算和边缘计算会不断融合；智能计算分布在不同的地方，但又相互连接，协同合作。

智能边缘计算中的关键问题研究

在微软亚洲研究院，我们致力于研究智能边缘计算中的关键问题，更好地将 AI 赋能于边缘设备（包括终端设备和边缘服务器）和应用，提高智能边缘计算的系统性能和用户体验。具体来说，目前我们主要关注以下几个研究方向：

针对不同设备的模型压缩和优化。高精度的深度学习模型通常都十分庞大，由数百万甚至以亿计的参数构成。运行这些模型需要耗费大量的计算和内存资源。虽然智能边缘设备的处理和存储能力大幅增长，但仍远远比不上云计算设备。因此，如何把深度学习模型在资源受限的边缘设备上运行起来是一个巨大的挑战。传统的模型压缩和优化（比如剪枝、量化等）主要关注的是在如何把模型变小的同时尽量少损失模型精度。然而，边缘设备的特点是类型多、差异性大，处理器类型性能和内存大小千差万别。我们认为，没有一个统一的模型能够适用于所有的边缘设备，而是应该结合硬件的特性，为不同的设备提供最适合的模型，不仅考虑模型大小和精度损失，更要考虑模型在设备上的执行性能，比如延迟和功耗等。

基于异构硬件资源的系统优化。即使有了一个可以运行的模型，如何提高模型的运行效率仍是一个值得深入研究的课题。我们需要一个高效的模型推理引擎，把系统性能提高到极致。这不仅需要软件层面的系统优化，更要有软件和硬件的协同设计，能够充分利用底层硬件的能力。边缘设备往往有着各种异构的硬件资源，比如智能手机拥有大小不同的 CPU 核（ARM big.Little）、DSP、GPU、甚至 NPU。而现有的系统往往只能利用其中一种计算资源（比如 CPU 或者 GPU），还不能充分发挥硬件的性能。我们的工作致力于研究如何充分利用同一设备上的异构硬件资源，深度优化系统性能，大大降低模型执行的延迟和能耗。

隐私保护和模型安全。如前所述，用户隐私数据保护是一个重要的课题。在边缘设备无法运行高精度模型的情况下（比如在低端的监控摄像头上），利用云计算或者边缘服务器来执行深度学习模型就不可避免。在这种情况下，我们就需要研究如何利用远程的计算资源的同时还能不泄露用户的隐私数据。另外，在边缘设备上运行模型还带来了一个新的问题——模型的安全。训练一个好的模型需要花费巨大的人力、物力。因此，模型是重要的数字资产。在云计算模式下，模型的存储和运行都在云上，终端用户无法直接接触模型数据。而在边缘计算中，模型是部署到本地设备上的，恶意用户可以破解终端系统，复制模型数据。所以，如何在智能边缘计算中保护模型的安全就是一个新的重要研究课题。

持续学习和合作学习。智能边缘计算还带来了新的改善模型的机会。目前的模型训练和模型使用通常是割裂的。一个模型在事先收集好的数据集上进行训练，然后被部署到设备上进行使用。然而，模型使用中的数据通常是和训练时的数据集不一样的。比如，每个智能摄像头由于其位置和光线的不同，它们看到的图像内容和特征都不尽相同，从而导致模型精度下降。我们认为，模型被部署到设备上以后，应该根据设备上的输入数据进行适配和优化，而且随着设备处理越来越多的新数据，它应该从中学习到新的知识，持续不断地提高它的模型，这就是持续学习（Continuous Learning）。此外，多个设备还应该把它们学习到的不同的新知识合并到起来，一起合作来改进完和善全局的模型，我们称之为合作学习（Collaborative Learning）。与主要关注如何利用多方数据集进行模型训练而不相互泄露数据的联邦学习（Federated Learning）不同，持续学习和合作学习的重点是如何在模型部署后从新获取的数据中学习新的知识。

此外，我们还关注**智能边缘计算中的各种新场景和新应用**，比如视频分析、VR/AR、自动驾驶、AIoT 等，特别是随着 5G 的到来，如何构建更好的智能边缘+智能云的系统，为这些场景和应用提供更好的支撑。

在过去两年，我们和国内外的高校紧密合作，在这些研究方向上取得了一系列的进展，也在相关学术会议上发表了多篇论文。其中，我们和北京大学和美国普渡大学关于如何利用缓存技术（Cache）提高卷积神经网络（CNN）执行效率的工作发表在 MobiCom 2018上 [5]；和哈尔滨工业大学等学校合作的关于如何利用模型稀疏性（Sparsity）加速模型执行的工作发表在 FPGA 2019和 CVPR 2019上 [6] [7]；和韩国 KAIST 等学校合作的关于如何利用 SGX 保护用户隐私的工作发表在 MobiCom 2019上 [8]；和美国纽约大学和清华大学合作的关于合作学习的工作发表在 SEC 2019上 [9]。

未来展望

智能边缘计算之后是什么？计算模式会沿着既有历史路线继续轮回吗？未来会是怎样的？

我们无法准确预测未来，但我们相信世界一定会变得越来越数字化、智能化，一定会变得更加美好。在微软看来，**整个世界正在成为一台巨大的计算机** [10]。不管你是在家里、在办公室、还是在路上，不管是在工厂、在商场、还是在各行各业，借助分布在各处的强大计算能力，我们可以利用人工智能处理由无处不在的传感器采集到的数据，创造出丰富多彩的工作和生活体验。**未来的计算一定是以用户为中心的，智能环境和设备随时随地感知用户的状态和需求，将用户所需的数据和信息准确推送给用户，为人们提供更好的服务。**

这是一个技术创新的黄金时代，有无数令人兴奋的问题等待我们去解决。希望有志于计算机系统研究的同仁能够加入我们，一起为建设更加美好的未来贡献自己的一份力量。简历请投递至邮箱：msra-srg-hire@microsoft.com

来源：沈向洋博士在2018微软人工智能大会上的演讲 [10]

参考文献

[1] V. Bahl, "10 years is an eternity in the tech world, but we are just getting started," 19 10 2018. [Online].
https://www.microsoft.com/en-us/research/blog/10-years-is-an-eternity-in-the-tech-world-but-we-are-just-getting-started/

[2] M. Satyanarayanan, P. Bahl, R. Cáceres and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," IEEE Pervasive Computing, vol. 8, no. 4, pp. 14-23, 2009.

[3] "The First IEEE/ACM Symposium on Edge Computing," 27-28 10 2016. [Online].
http://acm-ieee-sec.org/2016/

[4] J. Ratcliffe, "How many CCTV Cameras are there in London 2019?," 29 5 2019. [Online].
https://www.cctv.co.uk/how-many-cctv-cameras-are-there-in-london/

[5] M. Xu, M. Zhu, Y. Liu, F. X. Lin and X. Liu, "DeepCache: Principled Cache for Mobile Deep Vision," in Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, 2018.

[6] S. Cao, C. Zhang, Z. Yao, W. Xiao, L. Nie, D. Zhan, Y. Liu, M. Wu and L. Zhang, "Efficient and Effective Sparse LSTM on FPGA with Bank-Balanced Sparsity," in Proceedings of 27th ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2019.

[7] S. Cao, L. Ma, W. Xiao, C. Zhang, Y. Liu, L. Zhang, L. Nie and Z. Yang, "SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity through Low-Bit Quantization," in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019.

[8] T. Lee, Z. Lin, S. Pushp, C. Li, Y. Liu, Y. Lee, F. Xu, C. Xu and L. Zhang, "Occlumency: Privacy-preserving Remote Deep-learning Inference Using SGX," in Proceedings of the 25th Annual International Conference on Mobile Computing and Networking, 2019.

[9] Y. Lu, Y. Shu, X. Tan, Y. Liu, M. Zhou, Q. Chen and D. Pei, "Collaborative Learning between Cloud and End Devices: An Empirical Study on Location Prediction," in Proceedings of the Fourth ACM/IEEE Symposium on Edge Computing, 2019.

[10] 沈向洋, "让云计算和人工智能帮助每一个人," 2018 微软人工智能大会. [Online].
https://www.microsoft.com/china/events/ArtificialIntelligence2018.aspx

你也许还想看：

感谢你关注“微软研究院AI头条”，我们期待你的留言和投稿，共建交流平台。来稿请寄：msraai@microsoft.com。