边缘智能：现状和展望

李肯立[1,2]，刘楚波[1]
1. 湖南大学信息科学与工程学院，湖南 长沙 410082；2. 国家超级计算长沙中心，湖南 长沙 410082

摘要
边缘智能（即将人工智能融入边缘计算，部署在边缘设备）作为更快更好地提供智能服务的一种服务模式，已逐渐渗入各行各业。然而，当前边缘智能面临着巨大挑战。首先简要介绍了边缘智能的起源与概念；然后总结了边缘智能面临的三大挑战；最后，概括了当前针对边缘智能挑战的5个研究方向。为相关读者了解边缘智能和相关人员研究边缘智能提供一定的参考。

关键词
边缘计算；人工智能；边缘智能

中图分类号：TP39　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019025

Edge intelligence: state-of-the-art and expectations

LI Kenli[1,2], LIU Chubo[1]
1. College of Information Science and Engineering, Hunan University, Changsha 410082, China
2. National Supercomputing Center in Changsha, Changsha 410082, China

Abstract
Edge intelligence (EI, which merges artificial intelligence (AI) into edge computing and deploys AI methods on edge devices) is regarded as a very efficient measure to provide faster and better intelligent services, having been successfully applied to various fields. However, current EI faces great difficulties. Firstly, a brief introduction to EI was given, and then, three challenges in EI were summarized. Finally, current five research directions for solving the EI challenges were outlined. The paper was expected to provide a better understanding for people who want to know EI, and help for researchers who study EI to have an overall direction guideline.

Key words
edge computing, artificial intelligence, edge intelligence

1 引言

近年来，随着万物互联时代的到来和无线网络的普及，网络边缘的设备数量和其产生的数据量都急剧增长[1]。根据Cisco云指数的预测[2]，到2021年，全球范围内将有超过500亿的终端设备，这些设备每年产生的数据总量将达到847 ZB，其中约有10%的数据需要进行计算处理。另外，智能终端设备已成为人们生活的一部分，人们对服务质量的要求有了进一步提升[3-5]。在这种情况下，以云计算为代表的集中式处理模式将无法高效地处理边缘设备产生的数据，无法满足人们对服务质量的需求。其劣势主要体现在以下两个方面。

● 实时性不够。在云计算服务模式下，应用需要将数据传送到云计算中心进行处理，这增大了系统的时延。以无人驾驶汽车为例，高速行驶的汽车需要在毫秒级的时间内响应，一旦由于数据传输、网络等问题导致系统响应时间增加，将会造成严重的后果。

● 带宽不足。边缘设备产生的大量数据全部传输至云计算中心，给网络带宽造成了极大的压力。例如，飞机波音787每秒产生的数据超过5 GB[6]，但飞机与卫星之间的带宽不足以支持数据的实时传输。

边缘计算模型应运而生。边缘计算是部署在边缘终端设备和云计算中心之间的一种新型计算模型[7-8]。边缘计算的资源是指从数据源到云计算中心路径上的任意计算和网络资源，是一个连续系统[1]。在该模式下，边缘设备上亦有计算。边缘计算具有几个明显的优点：首先，边缘设备处理了部分产生的临时数据，不再需要将全部数据上传至云端，只需要传输有价值的数据，这极大地减轻了网络带宽的压力，且减少了对计算存储资源的需求。其次，在靠近数据源端进行数据处理，能够大大地减少系统时延，提高服务的响应时间。

人工智能（artificial intelligence，AI）技术迅猛发展，它是研究、开发用于模拟、延伸和扩展人的智能的一门新的技术科学。所谓智能，就是人脑比较过去、预测未来的能力。人的大脑不是计算机，不会亦步亦趋、按部就班地根据输入产生输出。人的大脑是一个庞大的记忆系统，它存储着在某种程度上反映世界真实结构的经验，能够记忆事件的前后顺序及相互关系，并依据记忆做出预测。

人工智能的最大优势在于它可以通过对外部声音、图像、环境或思维方式的处理，制定出问题的解决方法，对各类信息进行智能化的分析处理，解决人们在生活中面对的各种难题[9]。例如，人工智能在疾病诊断方面的应用，不仅可以辅助医生做出诊断治疗方案，也能帮助非医学人员正确处理病情。根据对模拟得出的数据信息的处理分析，人工智能可以快速制定出相应的解决方案，并以类比人类思维的方式运作[10]。在相同情况下，人工智能技术相比于人工处理更加高效快速，极大地促进了人们生活水平的进步。人工智能技术的优势主要体现在以下两个方面。

● 人工智能技术可以在很大程度上解决现实生活中无法处理的难题。当人们对模型的运算情况不清楚时，人工智能可以将其很好地解决，实现由结果现象回溯本质方法的处理。例如，针对大量杂乱、看似无关的原生数据，人们可以采用人工智能方法对数据进行处理、分类，提取出有用的数据。

● 人工智能可以实现对数据的加速处理。它主要是通过对外部声音、图像、环境或思维模式行为的模仿，结合计算机技术，加快对信息和数据的处理速度，从而给用

户带来更好的体验。

人工智能技术已成功应用于目标识别、智能搜索、语言处理、智能交通等领域。然而，由于人工智能方法包含大量的计算，当前人工智能大部分计算任务部署在云计算中心等大规模计算资源集中的平台上，这极大地限制了人工智能带给人们的便利。

为此，边缘智能应声而出。边缘智能是指终端智能，它是融合网络、计算、存储、应用核心能力的开放平台，并提供边缘智能服务，满足行业数字化在敏捷连接、实时业务、数据优化、应用智能、安全与隐私保护等方面的关键需求。将智能部署在边缘设备上，可以使智能更贴近用户，更快、更好为用户地提供智能服务。

随着网络技术的发展和移动设备的高度普及，边缘智能自提出以来，已引起了国内外政府、学术界与工业界的高度关注。然而，边缘智能仍处于发展初期，面临着巨大的挑战。

2 边缘智能面临的挑战

人工智能方法主要包括训练和推断两部分。训练主要是利用已知结果的大量数据来训练模型，即根据已有的大量数据来拟合模型及参数。推断主要是用拟合好的模型对未知结果的数据结果进行预测。训练过程需要大量的计算资源和存储资源。将人工智能部署在边缘设备上，主要面临以下几个挑战。

● 计算、存储、能耗等资源受限。相比于云计算中心等大规模服务器集群，边缘计算的资源是指从数据源到云计算中心路径上的任意计算和网络资源（如摄像头、网关等）[1]。因而，边缘设备的计算、存储能力往往远远小于专用服务器的计算、存储能力，无法满足人工智能训练所需的大量计算和存储资源。除此之外，部分边缘设备采用蓄电池等小型供电设备，无法满足计算所需的能耗。

● 边缘网络资源不足。根据Cisco云指数的预测[2]，到2021年，全球范围将有超过500亿的终端设备，每年产生的数据总量将达到847 ZB。相比而言，全球数据中心的存储能力预计仅能达到2.6 ZB，而网络流量为19.5 ZB。

● 人工智能在“边缘”并行困难。人工智能是研究使计算机模拟人的某些思维过程和智能行为（如学习、推理、思考、规划等）的学科。人工智能模型模拟人脑神经元的互联关系，模型复杂，各组成部分之间依赖性强，在分布式环境下并行困难。

3 边缘智能的主要研究方向

计算模型的创新带来的是技术的进步，而边缘智能的巨大优势也促使人们直面挑战、解决问题，推动相关技术的发展。针对第2节中分析的边缘智能面临的挑战，研究人员针对性地设计了相应的解决方案。本节总结了当前针对边缘智能难题的5个研究方向，包括边云协同、模型分割、模型压缩、减少冗余数据传输以及设计轻量级加速体系结构。其中，边云协同、模型分割、模型压缩主要是减少边缘智能在计算、存储需求方面对边缘设备的依赖；减少冗余数据传输主要用于提高边缘网络资源的利用效率；设计轻量级加速体系结构主要针对边缘特定应用提升智能计算效率。

(1) 边云协同

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，有研究文献[1-12]提出云计算和边计算协同服务架构。如**图1**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。

类似上述理念，2018年7月，谷歌推出两款大规模开发和部署智能连接设备的产品：Edge TPU和Cloud IoT Edge。Edge TPU是一种小型的专用集成电路（application specific integrated circuit，ASIC）芯片，用于在边缘设备上运行TensorFlow Lite机器学习模型。Cloud IoT Edge是一个软件系统，它可以将谷歌云的数据处理和机器学习功能扩展到网关、摄像头和终端设备上。用户可以在Edge TPU或者基于GPU/CPU的加速器上运行在谷歌云上训练好的机器学习模型。Cloud IoT Edge可以在Android或Linux设备上运行，关键组件包括一个运行时（runtime）。Cloud IoT Edge 运行在至少有一个CPU的网关类设备上，可以在边缘设备本地存储、转换、处理数据，同时，还能与物联网（Internet of things，IoT）平台的其他部分进行无缝互操作。

（2）模型分割

为了将人工智能方法部署在边缘设备，参考文献[13]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。如**图2**所示，参考文献[13]提出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。

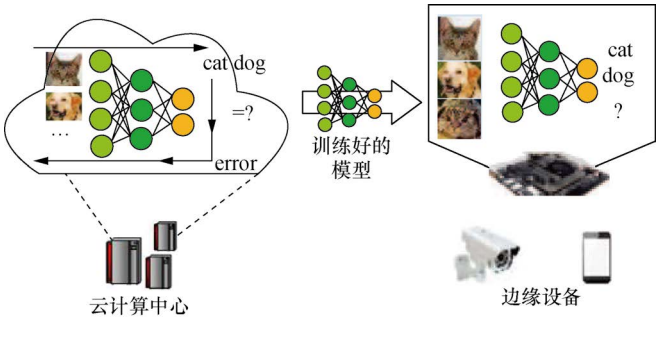

图1 云计算中心协同边计算服务器服务的过程

（3）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[14]、稀疏数据表示[15-16]、稀疏代价函数[15-17]等。**图3**展示了一个裁剪的多层感知网络。如**图3**所示，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[18]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。

（4）减少冗余数据传输

为了节省带宽资源，部分研究人员在不同的环境中提出了各式各样减少数据传输的方法，主要表现在边云协同和模型压缩中。例如，在参考文献[11]中，作者提出只将在边缘设备推断有误的数据传输到云端再次训练，以减少数据传输。在参考文献[19]中，作者提出在不影响准确度的情况下移除冗余数据，以减少数据的传输。

（5）设计轻量级加速体系结构

虽然NVIDIA公司的GPU芯片在数据中心人工智能的训练阶段占据了主导地位，但是依靠电池供电的边缘设备需要低功耗、小面积的加速芯片才能进行有效的推理。为此，相关研究人员从不同的角度进行了考虑，设计了许多针对边缘设备的加速体系结构。例如，在参考文献[19]中，作者针对压

缩、裁剪环境下的网络模型设计了一个加速器。在参考文献[20]中，与传统对不同种类应用一一对应设计加速硬件方法不同，作者提出采用可编程硬件加速，即对不同种类的应用使用同一硬件重写编程加速，以提高资源利用率，减小加速硬件的面积。在参考文献[21]中，作者提出利用可重构硬件特性，针对多种应用设计加速体系结构，在维持硬件面积的同时扩大应用范围。在工业界，亦有公司研究低功耗加速芯片。例如，与Altera公司和Xilinx公司研究数据中心功耗达到数十瓦的大型现场可编程门阵列（field programmable gate array，FPGA）芯片不同，Efinix公司的目标是研究功耗只有一瓦但可以运行神经网络加速器的设备。

4 结束语

经过近几年的发展，人工智能已在许多场景下发挥了重要作用，其巨大的优势有望在更多方面推动科技和社会生活的进步。然而，由于人工智能对计算能力的高需求，目前为止，人工智能的训练过程多集中在云计算中心等大规模计算集群中，端-云间的传输时延已成为智能服务的一道门墙。为使智能更贴近用户，更好地为人们服务，边缘智能应运而生，即结合人工智能和边缘计算技术，将智能服务由云计算中心推送到边缘设备，以改善智能服务质量。

本文总结了边缘智能面临的3个挑战，并总结了当前大部分研究人员针对这3个挑战提出的5个方向的解决方案，包括边云协同、模型分割、模型压缩、减少冗余数据传输以及设计轻量级加速体系结构。

人工智能、边缘计算已获得国内外政府、学术界和工业界的广泛关注和认可，已在许多应用场景下发挥作用。将人工智能部署在边缘设备已成为提升智能服务的有效途径。尽管边缘智能仍处于发展初期，然而，笔者相信，边缘智能将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，促进整个工商业体系的升级转型。

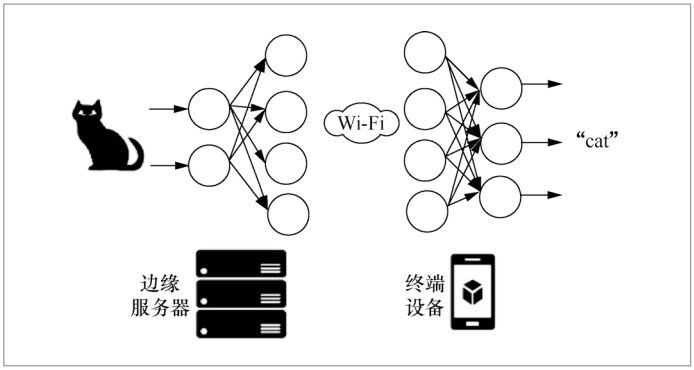

图 2 边缘服务器与终端设备协同推理

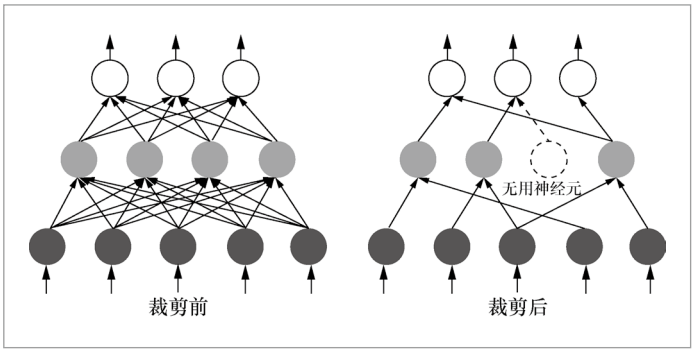

图 3 裁剪的训练模型

参考文献：

[1] 施巍松，张星洲，王一帆，等. 边缘计算：现状与展望[J]. 计算机研究与发展，2019，56(1)：69-89. SHI W S，ZHANG X Z，WANG Y F，et al. Edge computing：state-of-the-art and future directions[J]. Journal of Computer Research and Development，2019，56(1)：69-89.

[2] Cisco. Cisco global cloud index：forecast and methodology，2016-2021[R]. San Jose：Cisco，2018.

[3] WANG F，XU J，WANG X，et al. Joint

offloading and computing optimization in wireless powered mobile-edge computing systems[J]. IEEE Transactions on Wireless Communications, 2018, 17(6): 4177-4190.

[4] MAO Y Y, ZHANG J, SONG S H, et al. Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems[J]. IEEE Transactions on Wireless Communications, 2017, 16(9): 5994-6009.

[5] CHIANG M, ZHANG T. Fog and IoT: an overview of research opportunities[J]. IEEE Internet of Things Journal, 2017, 3(6): 854-864.

[6] TURNER V, GANTZ J F, REINSEl D, et al. The digital universe of opportunities: rich data and the increasing value of the Internet of things[R]. Hopkinton: EMC Corporation, 2018.

[7] BELLAVISTA P, FOSCHINI L, SCOTECE D. Converging mobile edge computing, fog computing, and IoT quality requirements[C]//2017 IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud), August 21-23, 2017, Prague, Czech Republic. Piscataway: IEEE Press, 2017: 313-320.

[8] 施巍松, 孙辉, 曹杰, 等. 边缘计算: 万物互联时代新型计算模型[J]. 计算机研究与发展, 2017, 54(5): 907-924.
SHI W S, SUN H, CAO J, et al. Edge computing-an emerging computing model for the internet of everything era[J]. Journal of Computer Research and Development, 2017, 54(5): 907-924.

[9] 李瑞驰. 人工智能的特征与应用分析[J]. 集成电路应用, 2019, 36(2): 105-106.
LI R C. Characteristics and application analysis of artificial intelligence[J]. Applications of IC, 2019, 36(2): 105-106.

[10] 谷守军, 王海永. 大数据时代人工智能在计算机网络技术中的应用[J]. 电子制作, 2017(6): 30-37.
GU S J, WANG H Y. The application of artificial intelligence in computer network technology in the age of big data[J]. Practical Electronics, 2017(6): 30-37.

[11] SONG M, ZHONG K, ZHANG J, et al. In-situ AI: towards autonomous and incremental deep learning for IoT systems[C]//2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), February 24-28, 2018, Vienna, Austria. Piscataway: IEEE Press, 2018: 92-103.

[12] TONG L, LI Y, GAO W. A hierarchical edge cloud architecture for mobile computing[C]// The 35th Annual IEEE International Conference on Computer Communications, April 10-14, 2016, San Francisco, USA. Piscataway: IEEE Press, 2016: 1-9.

[13] LI E, ZHOU Z, CHEN X. Edge intelligence: on-demand deep learning model co-inference with device-edge synergy[C]//ACM SIGCOMM Workshop on Mobile Edge Communications, August 21-23, 2018, Budapest, Hungary. New York: ACM Press, 2018: 1-10.

[14] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[15] RANZATO M A, POULTNEY C, CHOPRA S, et al. Efficient learning of sparse representations with an energy-based model[C]//The 19th International Conference on Neural Information Processing Systems, December 4-7, 2006, Vancouver, Canada. Cambridge: MIT Press, 2006: 1137-1144.

[16] RANZATO M A, BOUREAU Y L, LECUN Y. Sparse feature learning for deep belief networks[C]//The 20th International Conference on Neural Information Processing Systems, December 3-6, 2007, Vancouver, Canada. North Miami Beach: Curran Associates Inc., 2007: 1185-1192.

[17] LEE H, BATTLE A, RAINA R, et al. Efficient Sparse coding algorithms[C]//The 19th International Conference on Neural Information Processing Systems, December 4-7, 2006, Vancouver, Canada. Cambridge: MIT Press, 2006: 801-808.

[18] HAN S, POOL J, TRAN J. Learning both weights and connections for efficient neural network[C]//The 28th International Conference on Neural Information Processing

Systems, December 7-12, 2015, Montreal, Canada. Cambridge: MIT Press, 2015: 1135-1143.

[19] ZHANG S, DU Z, ZHANG L, et al. Cambricon-X: an accelerator for sparse neural networks[C]//The 49th Annual IEEE/ACM International Symposium on Microarchitecture, October 15-19, 2016, Taipei, China. Piscataway: IEEE Press, 2016: 1-12.

[20] NOWATZKI T, GANGADHAR V, SANKARALINGAM K, et al. Pushing the limits of accelerator efficiency while retaining programmability[C]//IEEE International Symposium on High Performance Computer Architecture (HPCA), March 12-16, 2016, Barcelona, Spain. Piscataway: IEEE Press, 2016: 27-39.

[21] LI Z, LIU L, DENG Y, et al. Aggressive pipelining of irregular applications on reconfigurable hardware[C]//ACM/IEEE International Symposium on Computer Architecture (ISCA), June 24-28, 2017, Toronto, Canada. Piscataway: IEEE Press, 2017: 575-586.

作者简介

李肯立 (1971-), 男, 博士, 湖南大学信息科学与工程学院教授、院长, 国家超级计算长沙中心主任, 主要研究方向为并行计算、高性能计算、网格和云计算。在国际顶级期刊、会议 (如IEEE Transactions on Computers、IEEE Transactions on Parallel and Distributed Systems、ICPP、ICDCS、CCGrid等) 上发表论文160余篇。担任IEEE Transactions on Computers等期刊编委, IEEE高级会员。

刘楚波 (1988-), 男, 博士, 湖南大学信息科学与工程学院副教授, 主要研究方向为调度和分布式系统建模、近似算法、随机算法、博弈论、云计算和边缘计算。在国际顶级期刊 (如IEEE Transactions on Parallel and Distributed Systems、IEEE Transactions on Cloud Computing、ACM Transactions on Modeling and Performance Evaluation of Computing Systems等) 上发表论文10篇。

收稿日期: 2019-02-18
基金项目: 国家重点研发计划基金资助项目 (No. SQ2018YFB020061) ; 国家自然科学基金资助项目 (No. 61702170)
Foundation Items: National Key Research and Development Program of China(No. SQ2018YFB020061), The National Natural Science Foundation of China(No. 61702170)