西 南 交 通 大 学

Research and Implementation of Machine Learning Deployment at the Network Edge

机器学习边缘架构部署的研究与实现

年 级： 2016 级 \_\_\_\_

学 号： 2016113705

姓 名： 张羽晨

专 业： 软件工程

指导教师： 滕飞

二零二零年五月

边缘计算与深度学习的融合:一份综合调查

注：(1) **本页前面页数应为偶数**，以保证此页双面打印时为正面！

(2) 正文中文用宋体，英文与数字用Times New Roman，均为小四号。其它格式与本科毕业设计（论文）格式要求相同。

(3) 此部分为文献翻译的中文全文。

Ⅰ. 绪论

从服务器集群的云数据中心到个人电脑和智能手机,甚至包括可穿戴的物联网设备， 随着计算和存储设备技术的高速发展和快速扩散，我们现在正处于一个以信息为中心的时代。在这个时代，计算无所不在，计算服务也从云端延申到网络边缘。根据美国思科公司(Cisco)发布的白皮书[1]，2020年，将有500亿的物联网设备接入互联网；另一方面，思科还估算到2021年的时候，每年在云服务器外部产生的数据量将达到850泽字节(ZB)。然而，目前全球的数据中心流量容量只有20.6ZB，远远低于需要被传输的数据量。这表明，为大数据技术而生的数据中心也在进行着一场革新，从大体量的云数据中心迁移到越来越多的边缘设备。然而，现有的云计算架构正渐渐不再能够管理这些大规模分布式计算能力以及进行数据分析，原因如下：1) 大量的计算任务需要交付到云中进行处理[3]，这无疑对网络容量和云计算基础设施的计算能力提出了严峻的挑战；2)许多新型的应用，比如车辆的协作自主驾驶，对响应延迟有着严格的要求，云架构由于物理上的原理用户，可能难以满足应用的实时性[4]。

因此，边缘计算[5], [6]作为一种吸引人的替代方案被提出，它特别强调要将计算任务尽可能地安排在接近数据源和最终用户的地方。当然，边缘计算和云计算并不是相互排斥的[7] , [8]。相反，边缘是云的补充和扩展。与单纯的云计算相比，边缘计算与云计算相结合的主要优势有三方面: 1) **缓解骨干网流量压力**，分布式边缘计算节点无需与云交换相应数据即可处理大量计算任务，减轻了网络的流量负荷；2)**敏捷的服务响应**，托管在边缘的服务可以显著降低数据传输的延迟，提高响应速度；3)**强大的云备份**，当边缘结点资源不足时，云可以提供强大的处理能力和海量存储。

作为一种典型的广泛适用的新型应用[9]，各种基于深度学习的智能服务和应用借由在计算机视觉(CV)和自然语言处理(NLP)[10]等领域的巨大优势，已经改变了人们生活的很多方面。这些成就不仅源于深度学习(DL)的发展，而且与不断增长的数据和计算能力密不可分。

然而，对于更广泛的应用场景，如智能城市、车联网等，却没有出现很多成熟的智能服务，其因素主要有以下几点：

* 成本因素：云中的深度学习模型的训练和推断需要设备或用户向云传输大量数据，从而消耗大量的网络带宽；
* 延迟因素：访问云服务的延迟通常没有保证，可能不足以满足许多时间关键型应用程序的需求，如协作自主驾驶[11]；
* 可靠性问题：大多数云计算应用程序依赖于无线通信和骨干网将用户与服务连接起来，但在许多工业场景中，智能服务必须具有高可靠性，即使在网络连接丢失的情况下也是如此；
* 隐私问题：深度学习所需的数据可能包含大量的隐私信息，并且隐私问题会影响到智能家居和城市等领域。

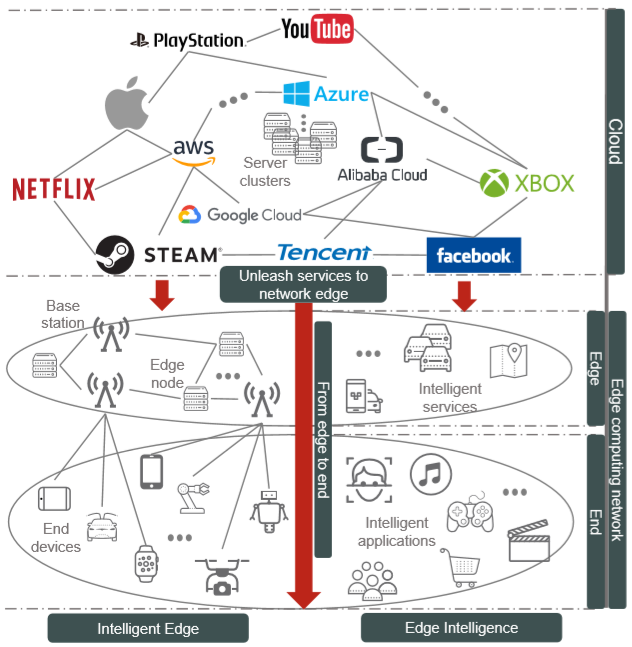


图.1. “边缘智能”和“智能边缘”

由于边缘结点比云端更接近于用户，因此边缘计算有望解决上述的问题。事实上，边缘计算正在逐渐与人工智能(AI)相结合，在实现“边缘智能”和“智能边缘”方面相互受益，如图.1.所示。边缘智能和智能边缘并不是相互独立的。边缘智能是目标，而智能边缘中的深度学习服务也是边缘智能的一部分。智能边缘可以为边缘智能提供更高的服务吞吐量和资源利用率。

具体来说，一方面，边缘智能希望将深度学习计算尽可能地从云中推到边缘，从而支持各种分布式、低延迟和可靠的智能服务。如图.2.所示，其优点包括：1) 深度学习服务部署在发出请求的用户附近，只有在需要额外处理[12]时，云才会参与，大大降低了向云发送数据进行处理的延迟和成本；2)由于深度学习服务所需的原始数据存储在本地的边缘或用户设备本身，而不是云上，因此增强了对用户隐私的保护；3)多级分层的计算架构提供了更可靠的深度学习计算；4)通过丰富的数据和应用场景，边缘计算可以**促进深度学习的普及应用**，实现“**随时随地为每个人和每个组织提供AI”**的前景[13]；5)多样化和有价值的DL服务可以扩大边缘计算的商业价值，加速其部署和增长。

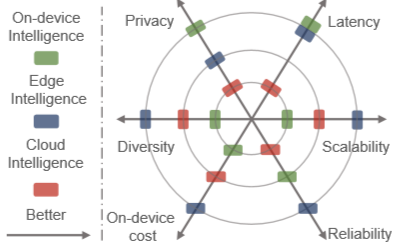


图.2. 云端,用户设备端和边缘智能的多方面能力比较

另一方面，智能边缘的目标是将**深度学习集成到边缘架构**中，实现动态、自适应的边缘架构维护管理。随着通信技术的发展，网络接入方式也越来越多样化。同时，边缘计算基础设施作为中间媒介，使得无处不在的终端设备和云之间的连接更加可靠和持久。因此，终端设备、边缘结点和云端逐渐合并成一个共享资源的社区。然而，维护和管理这样一个涉及无线通信、网络、计算、存储等的庞大而复杂的整体架构(community)是[15]面临的主要挑战。典型的网络优化方法依赖于固定的数学模型;然而，快速变化的边缘网络环境和系统很难精确建模。深度学习可以解决这个问题:当面对复杂繁琐的网络信息时，深度学习可以依靠其强大的学习和推理能力，从数据中提取有价值的信息，做出自适应的决策，实现相应的智能维护和管理。

因此，考虑到边缘智能和智能边缘，也就是边缘深度学习，在多个方面都面临着一些相同的挑战和实际问题，我们确定了以下五种技术对边缘深度学习至关重要：

1)应用于边缘架构上的深度学习应用：作为一种技术框架系统地组织起来边缘计算和深度学习，并提供智能服务；

2) 边缘架构中的深度学习推断：侧重于深度学习在边缘计算架构中的实际部署和推理，满足精度、延时等不同需求；

3) 为深度学习设计的边缘架构：在网络架构、硬件和软件方面支持深度计算的边缘计算平台；

4) 边缘本身的深度学习建模训练：对分布式边缘设备的边缘智能进行深度学习的建模和训练，以满足其资源和隐私方面所受约束。

5) 用于边缘架构优化的深度学习：将深度学习应用于维护和管理边缘计算网络(系统)的不同功能，如边缘缓存[16]、计算卸载[17]。

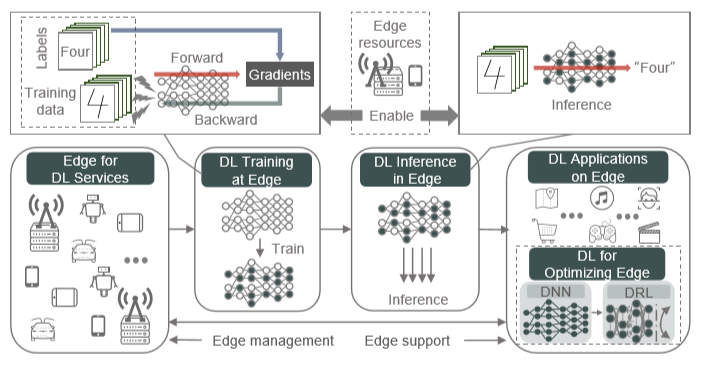


图.3. 核心边缘深度学习技术概览

如图.3.所示，“边缘上的深度学习应用”和“优化边缘的深度学习”分别对应于边缘智能和智能边缘的理论目标。为了支持它们，首先需要对各种深度学习模型进行密集的计算训练。在这种情况下，对于利用边缘计算资源来训练各种深度学习模型的相关工作，我们将它们归类为“边缘的深度学习训练”。其次，为了支持和加速边缘深度学习服务，我们将重点放在支持在边缘计算框架和网络中深度学习模型的有效推理的各种技术上，称为“边缘中的深度学习推理”。最后，我们将所有适应边缘计算框架和网络以更好地服务于边缘深度学习的技术分类为“面向深度学习的边缘计算”。

据我们所知，与我们的工作最相关的现有文章包括[18]-[21]。与我们对边缘深度学习更广泛的讨论不同，[18]关注于无线通信领域边缘智能中机器学习(而不是深度学习)的使用，即：训练机器学习模型在网络的边缘来改善无线通信。而[19]-[21]的主要贡献是对有关深度学习推断和训练的讨论。与这些工作不同的是，本研究集中在以下几个方面：1)通过边缘计算、跨网络、通信和计算全面考虑深度学习的部署问题;2)研究深度学习和边缘计算在这5个因素下的收敛的整体技术谱;3)指出深度学习和边缘计算是相互有利的，只考虑在边缘上部署深度学习是不完全的。

本文组织如图.4.所示，在本章节中，我们已经给出了本次调查的背景和动机。接下来，我们分别在第二节和第三节中介绍与边缘计算和深度学习相关的一些基础知识。以下各节介绍五种提出的核心支持技术，也即边缘上的深度学习应用(第四节)、优化边缘的深度学习(第五节)、边缘的深度学习训练(第六节)、边缘中的深度学习推理(第七节)、面向深度学习的边缘计算(第八节)。最后,在第九节我们讨论了目前的经验教训和开放挑战。在第十节，我们给出了本论文的总结。文中提及的所有缩写含义可在表Ⅰ.中查询。

Ⅱ. 边缘计算基础

三级节标题

三级节标题

三级节标题

二级节标题

二级节标题

一级节标题

二级节标题

二级节标题

一级节标题

英文文献原文标题

注：① 此部分为英文文献全文。

② 如文献原文可编辑，则须遵守此部分的模板格式。

③ 如文献原文为PDF格式，且文献格式遵照英文期刊或会议论文规范，则英文原文部分勿须遵守此处模板（包括页眉），直接用A4纸张打印后置于此处与翻译一起装订。

一级节标题

二级节标题

三级节标题

三级节标题

三级节标题

二级节标题

二级节标题

一级节标题

二级节标题

二级节标题

一级节标题