分布式计算环境

边缘计算研究现状分析

院（系）名称： 计算机学院

专 业 名 称 ： 计算机科学技术

学 生 姓 名 ： 万天娇

二〇二一年八月

目 录

[1. 边缘智能 4](#_Toc79944701)

[1.1 在边缘上进行推断的深度学习模型 4](#_Toc79944702)

[1.2 在边缘上训练的深度学习模型 5](#_Toc79944703)

[1.3 用深度学习优化边缘计算 5](#_Toc79944704)

[1.4 边缘计算对深度学习服务的支持 7](#_Toc79944705)

[2.优化与设计 8](#_Toc79944706)

[2.1 物理架构 8](#_Toc79944707)

[2.2 数据和服务放置 8](#_Toc79944708)

[2.3 资源调度 9](#_Toc79944709)

[3.系统与工程 10](#_Toc79944710)

[参考文献 12](#_Toc79944711)

随着万物互联时代的到来，网络边缘设备产生的数据量快速增加，带来了更高的数据传输带宽需求，同时，新型应用也对数据处理的实时性提出了更高要求，传统云计算模型已经无法有效应对，因此，边缘计算应运而生。边缘计算的基本理念是将计算任务在接近数据源的计算资源上运行，可以有效减小计算系统的延迟，减少数据传输带宽，缓解云计算中心压力，提高可用性，并能够保护数据安全和隐私。得益于这些优势，边缘计算从2014年以来迅速发展，经历了技术储备期、快速增长期，从2018年起已进入稳定发展期。

我将边缘计算领域目前的主要研究内容分为三块进行分析，第一个是边缘智能，也是就边缘计算和人工智能的结合，是目前论文中的热点研究方向，研究内容主要围绕以下几类：在边缘上进行推断的深度学习模型，在边缘上训练的深度学习模型，用深度学习优化边缘计算，边缘计算对深度学习服务的支持。第二个是针对边缘计算的优化与设计，将研究内容自底向上地划分为物理架构、内容放置、资源调度三个板块进行分析，后两部分被研究较多；第三个是系统与工程，该部分的文章结合实际场景展开研究，SEC（ACM/IEEE Symposium on Edge Computing）会议里这部分论文出现的较多。

# 边缘智能

当下人工智能（特别是深度学习）算法的计算需求增加，但受限于智能终端设备本身体积、能耗等因素，很多大型算法不能在终端上直接运行。为了解决这个问题，一方面我们可以探索模型压缩方法以减少运算量，另一方面我们可以运用边缘计算的方法将算力在终端、边缘与云之间进行合理的分配与卸载。边缘智能则是人工智能（深度学习）与边缘计算的一个交叉领域。边缘智能主要研究如何将人工智能模型，包括统计学习、深度学习及强化学习，放在网络边缘端执行。

对深度学习与边缘计算的研究主要总结为以下几类：在边缘上进行推断的深度学习模型，在边缘上训练的深度学习模型，用深度学习优化边缘计算，边缘计算对深度学习服务的支持。

## 在边缘上进行推断的深度学习模型

1. 深度学习模型的优化

深度学习任务通常是计算密集型的，需要占用大量内存。但是在边缘，没有足够的资源来支撑大规模的深度学习模型。所以需要优化深度学习模型，并量化它们的权重来降低资源成本。其中，最重要的挑战是：如何转换或重新设计深度学习模型使它们适合边缘设备，并且尽可能减少模型性能的损失。针对不同场景，先主要有两种优化方法，一是对资源相对充足的边缘节点，采用通用的优化方法；另一种是针对资源预算紧张的终端设备，采用细粒度的优化，针对模型输入、模型结构、模型选择、模型构架分别进行优化设计的研究。

1. 模型的拆分

对深度学习模型进行水平拆分，沿末端、边缘和云进行拆分是最常见的拆分方法，问题在于如何智能地选择拆分点。确定拆分点的过程一般可以分为三个步骤[1]：1）测量和建模不同DNN层的资源成本和层间中间数据的大小； 2）通过特定的层配置和网络带宽来预测总成本； 3）根据延迟需求等从候选拆分点中选择最佳选择。另一种模型拆分是垂直拆分，特别是对于CNN [2]，与水平分割相反，以网格方式融合层并垂直拆分，将CNN层划分为独立计算任务。

1. 提前退出推断（EEol）

通过利用 EEoI，可以在边缘设备上使用深度学习模型的浅层部分进行快速局部推理。通过这种方式，边缘设备上的浅层模型可以快速进行初始特征提取，如果有较高置信度，可以直接给出推理结果。否则，部署在云中的额外大型深度学习模型将执行进一步的处理和最终推理。与直接将深度学习计算卸载到云端相比，这种方法具有更低的通信成本，并且比边缘设备上的模型修剪或量化具有更高的推理精度[3]。

## 在边缘上训练的深度学习模型

1. 分布式训练

两种方案[4]：一种解决方案是，每个终端设备基于本地数据训练一个模型，然后在边缘节点聚合这些模型的更新。另一种是每个边缘节点训练自己的局部模型，并交换和细化它们的模型更新以构建全局模型。边缘的大规模分布式训练虽然避免了将大量原始数据集传输到云端，但不可避免地引入了边缘设备间的梯度交换通信成本。此外，在实际应用中，边缘设备可能会遭受更高的延迟、更低的传输速率，因此进一步阻碍了不同边缘设备间的深度学习模型梯度交换。

1. 联邦学习

联合学习是一种新兴的但有前途的方法，主要步骤：1) 从中央服务器下载全局深度学习模型，2）使用自己的数据在下载的全局模型下训练其本地模型， 3) 仅将更新后的模型上传到服务器进行模型平均。通过将训练数据限制在设备端，可以保护数据隐私和降低安全风险，从而避免将训练数据上传到云端所引起的隐私泄露问题。

联邦学习可以解决以下问题：1)训练数据可以不满足不是独立同分布的，2）有限的通信资源，3)各个节点训练数据和训练资源分布不均匀，4）数据隐私和安全问题。

目前对于联邦学习和边缘计算的研究主要围绕通信有效性、资源优化、安全提升等方向展开。

## 用深度学习优化边缘计算

1. 用深度学习优化缓存策略

深度学习的应用：传统的缓存方法通常计算复杂度高，因为它们需要大量的在线优化迭代来确定用户和内容的特征，以及内容放置和交付的策略。使用DNN网络来处理从用户的移动设备收集的原始数据，从而提取用户和内容的特征形成基于特征的内容热度矩阵[5]。对热度矩阵用基于特征的协同过滤算法来估计核心网络的流行内容。当使用DNN网络优化边缘缓存策略时，可以通过离线训练避免在线大量计算迭代。一个 DNN网络由一个用于数据正则化的encoder层和隐藏层构成，可以用最优启发式算法训练来确定缓存策略，从而避免在线优化迭代[6]。

强化学习的应用：与基于DNN的边缘缓存不同，强化学习可以利用用户和网络的上下文，并采用自适应策略来最大化长期缓存的性能[7]。传统的强化学习算法受限于难以手工提取特征，以及难以处理高维度数据和动作的缺陷[8]。相比于与深度学习无关的传统强化学习，例如，将Q-learning和Multi-Armed Bandit (MAB) learning[9]结合，深度强化学习的优势在于 DNN 网络可以从原始观察数据中学习关键特征。结合了强化学习和深度学习的智能体，可以直接从高维观测数据优化其在边缘计算网络中缓存管理的策略。

在[10]中，用DDPG训练DRL智能体以最大化长期缓存命中率，以做出适当的缓存替换决策。这项工作考虑了单个BS的场景，其中DRL智能体决定是否替换缓存的内容，在训练中，奖励被设计为缓存的命中率。

1. 用深度学习优化边缘任务卸载

边缘计算允许边缘设备在能量​​、延迟、计算能力等条件约束下将部分计算任务卸载到边缘节点，这些约束引发了一系列问题：1) 哪些边缘节点应该接收任务，2）边缘设备应该卸载的任务比例是多少，3）应该为这些任务分配多少资源。任务卸载是NP难的问题，因为要涉及通信和计算资源的组合优化以及边缘设备的竞争。特别是优化需要同时考虑随时间变化的无线环境（例如变化的信道质量）和任务卸载的请求，因此更倾向选择基于学习的优化方法。其中，基于深度学习的方法比其他方法具有更大优势，所以许多研究用深度学习或强化学习的方法优化边缘任务卸载。

深度学习的应用：在[11]中，计算卸载问题被表述为一个多标签分类问题。通过离线穷举求解，得到的最优解可用于训练DNN网络，其中输出为卸载决策。通过这种方式，可能最优解不需要在线求解，避免了延迟的卸载决策，并且计算复杂度可以转移到训练上。

强化学习的应用：虽然将计算任务卸载到边缘节点可以提高计算任务的处理效率，但是卸载的可靠性会受到无线环境质量的影响。在[12]中，为了最大化卸载效用，作者首先量化了各种通信模式对任务卸载性能的影响，并提出用DQL在线选择最佳卸载边缘节点和传输模式。为了优化总卸载成本，修改了 Dueling-DQL和Double-DQL网络，智能体可以为终端设备分配边缘计算和带宽资源。

1. 深度学习用于边缘节点的管理和维护

其他研究围绕深度学习对边缘通信的优化、边缘节点安全性的提高，以及联合节点优化等展开。

## 边缘计算对深度学习服务的支持

深度学习服务的广泛部署，尤其是移动深度学习，需要边缘计算的支持。这种支持不仅仅是在网络架构层面，边缘硬件和软件的设计、适配和优化同样重要。具体来说，1）定制的边缘硬件和相应优化的软件框架和库可以帮助深度学习执行更高效；2）边缘计算架构可以实现深度学习计算的卸载；3）设计良好的边缘计算框架可以更好地维护在边缘运行的深度学习服务；4) 用于评估边缘深度学习性能的平台有助于进一步改进上述实现。

这一部分的研究主要围绕硬件对深度学习的支撑，通信和计算模型的支撑，定制的深度学习边缘框架，以及边缘深度学习的性能评估等工作展开。

# 优化与设计

这一方向的论文主要针对三部分进行优化设计，我把它自底向上地划分为：物理架构、内容放置、资源调度。物理架构这部分论文研究的是边缘计算的架构，这涉及边缘站点的放置及部署、无线网络规划和路由等。第二部分，内容放置主要围绕服务选择、服务部署等问题提出解决方案，它是建立在物理架构基础上的。既然搭建好了边缘网络，接下来要做的就是数据和服务（内容）的部署和放置。这会涉及到许多问题要解决，比如，服务架构的选择，以微服务架构为例，需要为各类服务选择合适的边缘站点部署实例，这就得面临服务器选择、服务放置，服务部署等问题。如果待部署的服务有复杂的组合结构，那么也会涉及到服务组合的问题等，这一块的文章就主要解决这些问题。第三部分的资源调度是建立在物理架构和内容放置的基础上的，这类论文研究的是如何合理调度资源以提供更加优质的服务质量和用户体验。包括计算卸载方案的设计、用户数据在不同边缘站点之间的同步与迁移、移动性管理等。

## 物理架构

以IEEE期刊（2021年）上的文章Design and Simulation of a Hybrid Architecture for Edge Computing in 5G and Beyond[13]为例，该文章针对5G 及更高版本的边缘计算（如自动驾驶汽车、增强现实和远程手术）超低延迟的要求，提出了一种混合架构，该架构利用可持续的技术（例如，D2D 通信、MIMO、SDN 和 NFV），并具有可扩展性、可靠性和超低延迟支持等主要特性。

## 数据和服务放置

以IEEE期刊（2021年）上的文章An Online Framework for Joint Network Selection and Service Placement in Mobile Edge Computing[14]为例。以前的相关工作主要集中在通过动态服务放置来优化通信延迟问题，而忽略了接入网络选择对接入延迟的关键影响。在这篇文章中，研究了联合优化接入网络选择和服务放置的问题, 通过平衡接入延迟、通信延迟和服务切换的成本来提高 QoS。

随着5G无线技术的快速发展和部署，移动边缘计算（MEC）已成为一种新的技术来满足移动应用程序的低延迟要求，但由于分散的无线通信环境和容量受限的移动边缘计算节点，这种新技术的一个问题就是：如何为移动用户保持令人满意的服务质量 (QoS)。这种QoS通常在端到端延迟方面，极易受到接入网络瓶颈和通信延迟的影响。为了解决这个问题，首先，这篇文章同时考虑了MEC中的接入网络选择和服务放置问题，考虑访问延迟、切换成本和通信延迟来提高MEC应用程序的QoS。具体来说，当用户访问间接连接的边缘云上的服务时，会产生通信延迟。接入延迟和切换成本分别由AP的排队延迟和业务的动态迁移引起。由于公式化的长期优化问题本质上是随机的，即包含未来的不确定性，如用户移动性，本篇文章设计了一种高效的在线框架，将长期时变优化问题分解为一系列一次性子问题。为了解决一次性问题的NP难度，本篇文章设计了一种基于匹配和博弈论的计算效率高的两阶段算法，近乎实现了最优的解决方案。

## 资源调度

以2021年发表在《IEEE Transitions on Mobile Computing》上的Robust Task Offloading in Dynamic Edge Computing[15]为例。

移动边缘计算通过把任务从终端设备卸载到附近的边缘服务器，来实现更好的应用程序响应能力。然而由于移动性和功率限制，边缘服务器集会变得不确定， 当某些服务器出现故障时，在它们上运行的任务也将失败。为了克服边缘服务器的不稳定性，本篇文章提出了一种鲁棒的任务卸载方案，可以抵抗h（≥1）个边缘服务器故障。对于 h=1 的情况，本文设计了在线原始对偶算法，可以在任务到达时卸载它们，并从理论上推导出其竞争率。本文还将原始对偶算法扩展到h=2的情况，并进一步扩展到h ≥ 3的一般情况。通过仿真实验，证明本文的方案可以很好地处理边缘服务器的故障，并实现接近最佳的 DEC 吞吐量。

# 系统与工程

这类文章研究的场景包括视频监控的实时处理、目标检测、嫌疑行为识别等，车联网和自动驾驶中的数据实时感知与处理，无人机的实时路径规划与农业灌溉等。即使此类文章研究的是资源调度和优化的工作，也与从优化角度入手的论文完全不同。这些论文的工作通常借助一些单片机作为边缘设备，借助Docker和Kubernetes搭建benchmarks，并开发相应的中间件来解决一些实际问题。和优化类的文章相比，这类文章所需的研究周期更长，可以看出，它们需要涉及较到要解决的问题所在领域的专业知识。

例如，2019年边缘计算顶会SEC上的文章:Managing Edge Resources for Fully Autonomous Aerial Systems[16]。完全自主的空中系统可以通过软件完全执行复杂的任务。如果用户可以很好选择软件、计算硬件和飞机，则与人类相比，FAAS驾驶的无人机系统能更快，更安全地完成任务。另一方面，管理不善的边缘资源会降低任务的执行速度，浪费能源并增加成本。本文提出了一种模型驱动的FAAS管理方法。研究人员进行实际的FAAS任务飞行，配置文件计算和飞机资源使用，并为预期需求建模。通过创建FAAS验证了模型的有效性，并在许多系统设置下测量了任务吞吐量。通过开源FAAS套件SoftwarePilot发布的FAAS基准，执行现实的任务：自主摄影、搜索和救援以及农业侦察。该模型错误预测为4％，对比方法产生10％到24％的错误。作者团队研究了：（1）GPU加速、扩展和横向扩展，（2）板载、边缘和云计算，（3）能源和资金预算，（4）软件驱动的GPU管理。研究发现，模型驱动的管理可以使任务量提高10倍，并将成本降低87％。

再例如，另一篇文章I-SAFE:Instant Suspicious Activity identiFicationat the Edge using Fuzzy Decision Making[17]。城市图像通常用作取证分析，通过设计可用于减轻事故。收集更多图像后，人们很难从成千上万个视频剪辑中确定具体事件的某些帧。研究人员希望有一个实时，主动的监视系统，该系统可以立即检测可疑人员，识别可疑活动，或提高警惕。本文通过采用模糊决策在边缘（I-SAFE）上引入即时可疑活动识别来提出一种法务监视策略。针对安全人员的决策制定了模糊控制系统。根据轻量级深度机器学习（DML）模型提取的视频功能做出决策。根据一线执法人员的要求，选择并模糊化了一些功能，以应对人员决策过程中存在的不确定性状态。在边缘层次结构中使用功能可以最大程度地减少即时警报的通信延迟。此外，利用微服务体系结构，I-SAFE方案可在网络边缘增加复杂性，并具有良好的可扩展性。经测试，I-SAFE方案识别可疑活动的平均时间为0.002秒。

# 参考文献

1. Z.Zhao, Z.Jiang, N.Ling et al.,“ECRT: An Edge Computing System for Real-Time Image-based Object Tracking,” in Proc.the 16th ACM Conference on Embedded Networked Sensor Systems (SenSys 2018), 2018, pp. 394–395.
2. Z.Zhao, K.M.Barijough, and A.Gerstlauer,“DeepThings: Distributed Adaptive Deep Learning Inference on Resource-Constrained IoT Edge Clusters,” IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., vol. 37, no. 11, pp. 2348–2359, Nov. 2018.
3. S.Teerapittayanon, B. McDanel, and H.T.Kung,“Distributed Deep Neural Networks over the Cloud, the Edge and End Devices,” in IEEE 37th International Conference on Distributed Computing Systems (ICDCS 2017), 2017, pp. 328–339.
4. L.Valerio, A. Passarella, and M. Conti, “A communication efficient distributed learning framework for smart environments,” Pervasive Mob. Comput., vol. 41, pp. 46–68, Oct. 2017.
5. S.Rathore, J. H. Ryu, P. K. Sharma, and J. H. Park, “DeepCachNet: A Proactive Caching Framework Based on Deep Learning in Cellular Networks,” IEEE Netw., vol. 33, no. 3, pp. 130–138, May 2019.
6. Z. Chang, L. Lei, Z. Zhou et al., “Learn to Cache: Machine Learning for Network Edge Caching in the Big Data Era,” IEEE Wireless Commun., vol. 25, no. 3, pp. 28–35, Jun. 2018.
7. D. Adelman and A. J. Mersereau, “Relaxations of weakly coupled stochastic dynamic programs,” Operations Research, vol. 56, no. 3, pp. 712–727, 2008.
8. H. Zhu, Y. Cao, W. Wang et al., “Deep Reinforcement Learning for Mobile Edge Caching: Review, New Features, and Open Issues,” IEEE Netw., vol. 32, no. 6, pp. 50–57, Nov. 2018.
9. K. Guo, C. Yang, and T. Liu, “Caching in Base Station with Recom- mendation via Q-Learning,” in 2017 IEEE Wireless Communications and Networking Conference (WCNC 2017), 2017, pp. 1–6.
10. C. Zhong, M. C. Gursoy et al., “A deep reinforcement learning- based framework for content caching,” in 52nd Annual Conference on Information Sciences and Systems (CISS 2018), 2018, pp.
11. S. Yu, X. Wang, and R. Langar, “Computation offloading for mobile edge computing: A deep learning approach,” in IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Com- munications (PIMRC 2017), 2017, pp. 1–6.
12. K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, and Y. Zhang, “Deep Learning Empowered Task Offloading for Mobile Edge Computing in Urban Informatics,” IEEE Internet Things J., vol. 6, no. 5, pp. 7635– 7647, Oct. 2019.
13. H. Rahimi ,Y. Picaud, K. Deep Singh, G. Madhusudan, S. Costanzo , and O. Boissier, “Design and Simulation of a Hybrid Architecture for Edge Computing in 5G and Beyond” IEEE Trans. Comput, VOL. 70, NO. 8, AUGUST. 2021.
14. B. Gao, Z. Zhou, F. Liu, F. Xu, and B. Li, “An Online Framework for Joint Network Selection and Service Placement in Mobile Edge Computing” IEEE Trans. Comput. DOI 10.1109/TMC.2021.
15. H. Wang, H. Xu, H. Huang, M. Chen, S. Chen, “Robust Task Offloading in Dynamic Edge Computing”in Transactions On Mobile Computing, VOL. NO. OCT. 2020.
16. J. G, C. Stewart, J. Chumley, and S. Zhang, “Managing Edge Resources for Fully Autonomous Aerial Systems” in EC'19: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, 2019, pp. 101–112.
17. S.Y. Nikouei, Y. Chen, A. Aved, E. Blasch, and T.R. Faughnan, “I-SAFE: instant suspicious activity identification at the edge using fuzzy decision making”, in EC '19: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, 2019, pp. 101–112.