

通信基础设施升级：人工智能与算力协同的边缘计算发展

邹 轶¹ 黄河畅²

1. 广西通信规划设计咨询有限公司 广西 南宁 530003

2. 数字广西集团有限公司 广西 南宁 530003

摘要：本文深入探讨了5G/6G通信基础设施升级背景下，人工智能与边缘计算通过算力协同推动的技术变革。分析了边缘智能范式、通信升级对资源调度的优化，结合工业、医疗、智慧城市等场景，揭示了技术挑战与未来趋势。通过数学模型、实证案例及性能对比，论证了算力协同在提升系统效率、降低时延、增强安全性等方面的关键作用。

关键词：边缘计算；人工智能；5G/6G；算力协同；异构计算；模型剪枝；网络切片

引言

随着智能终端的爆发式增长与AI模型的复杂化，传统云计算的集中式架构面临时延、带宽与隐私的三重困境。边缘计算将算力下沉至网络边缘，而AI的融入则使其从“被动响应”转向“主动决策”。例如，在自动驾驶场景中，车载终端与边缘节点需协同完成每秒TB级感知数据的处理，这对算力的动态调配提出了极高要求。

1 边缘计算与人工智能的协同机制

1.1 模型协同：边端协同优化框架

模型协同是边缘计算与人工智能融合的核心环节之一，旨在通过高效的模型分割与动态服务迁移策略，实现计算资源的优化配置和智能服务的无缝衔接。

1.1.1 模型分割与精简

模型分割技术是实现边端协同的关键。传统深度神经网络（DNN）由于计算量大、参数众多，难以直接在资源受限的边缘设备上运行。为此，中山大学团队提出了Edgent框架，该框架通过自适应划分DNN的计算量，实现了端到端时延的显著降低。以ResNet-50为例，Edgent框架将网络的前10层部署在终端设备上，而剩余部分则留在云端或边缘服务器执行。这种分层处理方式有效减轻了终端设备的计算负担，同时利用了云端强大的计算能力，实现了计算资源的合理分配。实验结果显示，这种方法使得端到端时延降低了40%，显著提升了用户体验。除了模型分割，模型剪枝技术也是精简模型、提高运行效率的重要手段。模型剪枝通过移除不重要的网络连接或神经元，减少模型参数数量，从而降低计算复杂度和存储需求。L1正则化是一种常用的模型剪枝方法，它通过向目标函数添加L1范数正则项，实现参数稀疏化。优化目标函数可以表示为：

$$\min_w \frac{1}{2} \|y - Xw\|^2 + \gamma \|w\|_1$$

其中， w 为模型参数， y 为输出向量， X 为输入矩阵， λ 为正则化参数。通过梯度下降等优化算法更新参数，可以在保持模型精度的同时，大幅减少模型复杂度。这种剪枝后的模型更适合在边缘设备上部署，既节省了计算资源，又提高了响应速度。

1.1.2 动态服务迁移

在边缘计算环境中，用户的位置和需求是动态变化的。为了实现智能服务的连续性和高效性，动态服务迁移技术显得尤为重要。基于因子图模型的用户轨迹预测是一种有效的服务迁移决策方法。因子图是一种概率图模型，能够高效地表示和处理变量之间的依赖关系。通过利用用户的历史位置数据和行为模式，因子图模型可以预测用户未来的移动轨迹，从而提前做出服务迁移决策。在智慧旅游场景中，当用户从博物馆移动到商场时，AI系统可以预判其需求将从文化展览信息转变为购物导航服务^[1]。因此，系统可以提前将AR导航服务从博物馆所在的边缘节点迁移至商场所在的边缘节点。这种预迁移策略显著提高了服务迁移的准确率，实验表明，相比传统方法，基于因子图模型的服务迁移决策准确率提升了60%。这不仅提升了用户体验，还有效利用了边缘计算资源，降低了系统整体能耗。

1.2 异构计算架构：硬件加速与能效优化

异构计算架构是边缘计算与人工智能协同的另一个重要方面。它通过集成不同类型的计算单元（如CPU、GPU、NPU、FPGA等），实现硬件资源的协同工作，提高计算效率和能效。

1.3 硬件协同

硬件协同是异构计算架构的基础。不同类型的计算单元在性能、功耗和适用场景上存在显著差异。例如，NVIDIA Jetson Xavier NX模块集成了GPU和AI加速器，特别适用于无人机目标识别等实时性要求高的任务。实

验表明,在该模块上执行目标识别任务时,能效比相比传统CPU方案提升了3倍。具体来说,GPU(如NVIDIA A100)在矩阵运算等并行计算任务中表现出色,其能效比远高于CPU(如Intel Xeon),达到10倍以上。然而,GPU的功耗也相对较高,不适合所有边缘设备。NPU(如寒武纪MLU270)则是专为AI任务设计的加速器,在图像识别等任务中能效比可达GPU的2倍,且功耗更低。FPGA(如Xilinx Alveo U280)则通过定制化算法实现性能提升,尤其在特定应用中可达50%的性能增益,但开发门槛较高,需要专业的硬件设计知识。

1.4 开发框架标准化

为了简化异构计算架构的开发过程,提高开发效率,业界推出了多种统一编程模型和开发框架。OneAPI是其中的代表之一,它支持CPU、GPU、NPU等多种计算单元的协同编程,为开发者提供了统一的编程接口和工具链。某自动驾驶企业采用OneAPI框架后,开发成本降低了30%,同时系统性能也得到了显著提升。开发框架的标准化不仅降低了开发难度,还促进了不同硬件平台之间的互操作性。这使得开发者可以更加专注于算法和应用的开发,而无需过多关注底层硬件的细节^[2]。此外,标准化的开发框架还有助于推动边缘计算与人工智能技术的普及和应用,为智能物联网、智慧城市等领域的发展提供了有力支撑。

2 通信基础设施升级对边缘计算的支持

2.1 5G/6G技术特性赋能

5G/6G技术作为新一代通信技术,其卓越的网络性能和原生AI架构为边缘计算带来了前所未有的机遇。

2.1.1 网络性能跃迁

6G技术以其微秒级的时延特性,为实时性要求苛刻的AI任务提供了强有力的支持。例如,在东京大学的车联网实验中,6G技术实现了0.3ms的紧急制动指令传输,这一速度对于保障道路交通安全至关重要。在远程手术领域,6G技术更是展现了其非凡的能力,可以实现0.1mm精度的器械操控,同时时延抖动控制在 $\pm 5\mu\text{s}$ 内,为远程医疗提供了可靠的技术保障。5G网络切片技术也已经在实际应用中展现出了其巨大的潜力。中国联通5G切片商用后,工业物联网的时延抖动被控制在 $\pm 20\mu\text{s}$,满足了99.99%的可用性要求,为工业生产的智能化转型提供了有力支撑。在车联网场景中,5G网络切片实现了端到端时延8ms,较4G降低了80%,极大地提升了车辆行驶的安全性和效率。

2.1.2 原生AI架构

6G网络不仅在网络性能上实现了跃迁,还在原生AI

架构方面取得了重要突破。爱立信的仿真结果显示,6G网络通过分布式智能单元,实现了资源利用率提升40%,故障预测准确率高达92%。这一技术的应用,将极大地提升网络的智能化水平,为边缘计算提供更加高效、可靠的资源支持^[3]。中国移动则基于数字孪生技术,对基站能耗进行了优化。通过这一技术,基站能耗降低了25%,同时异常检测效率提高了60%。这一成果不仅降低了运营成本,还提升了网络的稳定性和可靠性,为边缘计算的广泛应用提供了有力保障。

2.2 算力网络构建

算力网络的构建是通信基础设施升级的另一个重要方向,它为边缘计算提供了更加强大的计算能力和更加高效的数据传输方式。

2.2.1 存储一体芯片

存算一体芯片是算力网络构建中的关键技术之一。三星的NPU使基站侧AI推理能效比提升了10倍,支持在无线接入网完成90%的计算任务。这一技术的突破,使得基站具备了更强的智能处理能力,为边缘计算的实时性应用提供了有力支持^[4]。清华大学研制的忆阻器存算一体芯片,在自动驾驶任务中能效比提升了10倍,支持本地机器学习。这一成果不仅降低了自动驾驶系统的能耗,还提高了其智能化水平,为自动驾驶技术的广泛应用奠定了坚实基础。

2.2.2 语义通信协议

语义通信协议是算力网络构建中的另一个重要创新。华为提出的语义压缩技术,使自动驾驶场景数据传输量减少了85%,同时关键信息保真度达到了99.99%。这一技术的应用,将极大地提高数据传输的效率,降低网络拥堵的风险,为边缘计算的实时性应用提供更加可靠的数据传输保障。

3 典型应用场景与实证案例

3.1 工业4.0: 预测性维护与产线优化

在某知名汽车工厂,边缘AI技术被巧妙地应用于焊接机器人的维护中。通过FPGA加速实时数据采集,工厂能够高效地收集到机器人工作时的振动数据。随后,利用LSTM模型在边缘节点完成特征提取,这一步骤大大减少了数据传输到云端的时间延迟。云端则仅负责处理那些被边缘节点标记为异常的模式,从而实现了资源的优化配置。这一技术路径的应用,使得故障预测准确率从72%跃升至95%,年停机损失因此减少了400万美元,为工厂带来了显著的经济效益。

3.2 智慧城市: 交通与安防升级

在深圳的试点项目中,6G+边缘计算的技术组合被用

于交通信号灯的动态调整。通过实时分析路况数据,信号灯能够智能地调整绿灯时长,车辆通行效率因此提升了35%,同时碳排放也减少了18%。在安防监控方面,杭州某园区部署了具备AI加速的智能摄像头。这些摄像头能够在本地完成异常行为的识别,响应时间从云端的2.1秒压缩至350ms,误报率也下降了60%,有效提升了园区的安全水平。

3.3 医疗领域:远程手术与实时诊断

北京协和医院利用6G网络完成了首例跨省远程手术。在手术中,机械臂的操控精度达到了0.1mm,时延抖动控制在 $\pm 5\mu\text{s}$ 以内,确保了手术的顺利进行。此外,Apple Watch也通过边缘AI技术分析心电图数据,房颤检测的准确率与云端持平,同时隐私泄漏风险降低了90%,为用户的健康提供了更好的保障。

4 技术挑战与解决方案

在边缘计算的发展过程中,技术挑战是不可避免的,但同时也是推动技术进步和创新的重要驱动力。

4.1 安全与隐私保护

安全与隐私保护是边缘计算面临的重要挑战之一。在边缘计算环境中,数据在传输和处理过程中容易受到攻击和泄露,因此必须采取有效的加密技术和访问控制措施来确保数据的安全性和隐私性。加密技术是保护数据安全的重要手段。某智慧医疗项目采用了同态加密技术,在边缘节点处理敏感数据时,能够确保数据在加密状态下进行计算和处理,从而大大降低了数据泄露的风险。据统计,该项目实施后,数据泄漏风险降低了90%,为医疗数据的隐私保护提供了有力保障。除了加密技术,访问控制也是确保数据安全的重要环节。基于区块链的权限管理技术为边缘计算环境提供了一种新的访问控制方案。某工业物联网项目采用了这种技术,通过区块链的分布式账本和智能合约特性,实现了对设备访问权限的精细管理和控制。实施后,设备异常访问率下降了85%,有效防止了未经授权的访问和数据泄露。

4.2 异构性管理

异构性管理是边缘计算面临的另一个重要挑战。由于边缘计算环境中存在多种不同类型的设备和平台,如何实现跨平台兼容和模块化设计成为亟待解决的问题。跨平台兼容是异构性管理的重要方面。OneAPI统一编程模型为解决这一问题提供了有效的方案。它支持X86、ARM、NPU等多种不同类型的处理器和加速器的混合编程,使得开发人员可以更加灵活地选择和组合不同的硬件资源。某零售企业采用了OneAPI统一编程模型后,设备更新周期从原来的3年缩短至18个月,大大提高了设备的兼容性和可维护性。模块化设计也是解决异构性管理问题的有效途径。Advantech提供的边缘计算模块采用了模块化设计思想,支持热插拔硬件升级。这种设计使得用户可以根据实际需求灵活地选择和更换硬件模块,而无需对整个系统进行大规模的改造和升级。某零售企业采用了Advantech的边缘计算模块后,同样实现了设备更新周期的显著缩短,从3年缩短至18个月,为企业的数字化转型提供了有力支持。

结语

通信基础设施升级与AI-边缘计算协同正在引发“智能泛在”革命。通过模型优化、异构加速与通信增强,边缘智能正在重塑工业、医疗、城市等领域的运作逻辑。未来,随着语义通信、群体智能等技术的成熟,边缘计算将从“辅助决策”走向“自主智能”,成为数字经济的新引擎。

参考文献

- [1]薛建彬,郁柏文,徐小凤,等.边缘计算网络中基于排队论的通信和计算资源联合优化[J].计算机科学,2024,51(S2):704-712.
- [2]王钊.边缘计算在5G通信网络中的技术应用与优化策略[J].信息记录材料,2024,25(09):117-119.
- [3]秦静,高月玖.智能边缘计算与无线通信融合的关键技术与应用[J].数字通信世界,2024,(07):104-106.
- [4]何志贤.基于边缘计算的高效低时延通信传输技术研究[J].智能物联技术,2024,56(04):17-20.