

基于图学习的空天地一体化网络合作推理方法

徐煜凯¹, 徐波¹, 张凌豪¹, 王琴¹, 赵海涛²

(1. 南京邮电大学通信与信息工程学院, 江苏 南京 210003; 2. 南京邮电大学物联网学院, 江苏 南京 210003)

摘要: 在空天地一体化网络 (SAGIN, space-air-ground integrated network) 中, 通过推理已训练的深度神经网络 (DNN, deep neural network) 能够实现多样化的人工智能 (AI, artificial intelligence) 服务。由于 DNN 的参数数量不断增大, 在有限的通信和计算资源下, 推理任务的执行过程通常需要多方的合作。本论文将认知无线电技术 (CR, cognitive radio) 融入 SAGIN 中, 以最小化合作推理时延为目标, 在卫星通信干扰阈值约束、设备服务质量约束以及发射功率约束下, 以无人机 (UAV, unmanned aerial vehicle) 为推理任务的发起者建立考虑设备关联、模型分割以及功率优化的合作推理问题。该问题是多维变量耦合的非凸问题, 本文根据优化变量将原问题分解, 提出一种融合二次无约束二值优化 (QUBO, quadratic unconstrained binary optimization) 和注意力机制的图学习法, 将目标函数和约束条件建模为 QUBO 形式, 并且通过哈密顿量松弛生成可微的损失函数, 进而利用动态多头注意力机制对图结构问题中的节点信息进行聚合, 从而完成设备关联模型分割优化。此外, 使用连续凸逼近 (SCA, successive convex approximation) 对发射功率进行优化, 最后对三部分问题交替迭代求解。仿真结果表明, 所提方法能够将合作推理任务总时延降低 11%。

关键词: 空天地一体化网络; 合作推理; 图神经网络; 资源优化

中图分类号: TN929.5

文献标志码: A

doi: 10.11959/j.issn.2096-3750.XXXX.

Graph Learning-based Cooperative Inference Method for Space-Air-Ground Integrated Networks

XU Yukai¹, XU Bo¹, ZHANG Linhao¹, WANG Qin¹, ZHAO Haitao²

1. College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

2. College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: In the space-air-ground integrated network (SAGIN), diverse artificial intelligence (AI) applications are realized through the inference of pre-trained deep neural networks (DNNs). Owing to the ever-increasing parameter scale of DNNs, the execution of inference tasks under constrained communication and computational resources typically necessitates multi-party collaboration. This paper integrates cognitive radio (CR) technology into SAGIN, aiming to minimize inference latency while considering constraints such as satellite communication interference thresholds, device quality-of-service requirements, and transmit power limitations. We formulate a cooperative inference problem involving device association, model partitioning, and power optimization, with unmanned aerial vehicle (UAV) serving as task initiators. The resulting problem is a non-convex optimization with coupled multi-dimensional variables. Accordingly, we decompose the original problem based on optimization variables and propose a graph learning method combining quadratic unconstrained binary optimization (QUBO) and an attention mechanism. Specifically, the objective function and constraints are modeled in QUBO form, and a differentiable loss function is derived through Hamiltonian relaxation. A dynamic multi-

收稿日期: XXXX-XX-XX; 修回日期: XXXX-XX-XX

通信作者: 徐波 xubo@njupt.edu.cn

基金项目: 国家自然科学基金 (U2441226)

Foundation Items: The National Natural Science Foundation of China (U2441226)

head attention mechanism is then employed to aggregate node information in the graph-structured problem, optimizing device association and model partitioning. Additionally, transmit power is optimized using successive convex approximation (SCA). Ultimately, the three subproblems are solved alternately in an iterative process. Simulation results indicate that the proposed method reduces the total latency of collaborative inference tasks by 11%.

Key words: space-air-ground integrated networks, collaborative inference, graph neural network, resource optimization

0 引言

近年来,地面通信技术取得了显著进步,特别是在城市环境中,通过先进的系统架构和密集部署的基础设施,实现了高质量的传输服务。然而,地面基站的选址受限于地理环境,且地面网络的架构较为固定,在偏远地区和应急场景下难以提供可靠的通信服务^[1]。针对该问题,集成低地球轨道(LEO, low earth orbit)卫星、无人机(UAV, unmanned aerial vehicle)和地面设备(GD, ground device)的空天地一体化网络(SAGIN, space-air-ground integrated network)被认为是第六代移动通信系统(6G, 6th-generation mobile system)的重要研究方向^[2]。在SAGIN中,LEO卫星作为陆基网络的补充,由于不受地面条件限制并具备快速响应能力,可通过无缝网络覆盖提供无处不在的服务^[3]。此外,UAV作为连接天基网络和陆基网络的枢纽,依托轻量化和高机动的优势,可以辅助SAGIN根据具体需求进行动态的精细组网^[6]。因此,SAGIN已成为当前通信网络领域的重要研究方向。

与此同时,随着深度学习技术的快速发展,基于深度神经网络(DNN, deep neural network)的人工智能(AI, artificial intelligence)应用越来越多,包括图像分类^[7]、自动驾驶^[8]和语义通信^[9]等。这类AI应用依赖于DNN模型的推理,即使用预先训练的DNN模型处理输入的数据样本,并输出分类或预测结果。值得关注的是,SAGIN凭借其独特的架构优势,为上述AI应用提供了极具潜力的部署平台。具体而言,SAGIN的分布式网络架构能够有效支持边缘计算和多源数据协同聚合,从而显著提升了DNN模型推理的实时性和准确性^[10]。然而,随着DNN的参数数量不断增大,SAGIN中资源有限的UAV和LEO卫星难以在本地高效地完成推理任务^[11]。针对该问题,设备之间或者设备与边缘服务器合作完成模型推理任务已成为可行方案^[12-14]。文献[12]基于动态分区策略分割DNN模型,将部分

DNN模型卸载到边缘服务器,以降低推理能耗。类似地,文献[13]将模型分割问题建模为一个有限域内的优化停止问题,并通过逆向归纳法推导出最优的模型分割策略,在线决定何时停止本地推理。此外,文献[14]中设计了一种基于分布式强化学习的模型分割和计算资源联合优化方法,以降低设备和边缘服务器的合作推理时延。

虽然已有工作能够提高合作推理任务的执行效率,但是没有被有效应用于SAGIN中,特别是缺乏对SAGIN中复杂干扰的考量。具体而言,随着SAGIN中设备数目和DNN的规模激增,在执行合作推理任务时,设备之间或设备与边缘服务器之间的通信会产生额外的开销,此时频段复用引起的同频干扰抑制了SAGIN的传输能力,进而增加了合作推理任务中的传输时延。针对该问题,需要通过合理调配网络资源并且设计设备关联策略以提高合作推理效率。文献[15]探讨了卫星-UAV网络的多域资源分配问题,以提高数据传输速率为目标,离线联合优化子信道分配、发射功率和无人机悬停时间。文献[16]提出了基于多智能体深度强化学习的设备关联和功率分配方法,采用云端更新机制以最大化长期全局能量效率。文献[17]提出了一种基于多智能体双深度Q网络的分布式端到端网络切片切换决策算法,用于在线优化UAV位置、设备关联和资源分配。上述方案虽然能够合理地进行用户关联并分配网络资源,但是没有考虑SAGIN中频谱资源极度稀缺的现实困境。例如,当大量设备在非授权频段运行时,需确保这些设备不对其他合法无线电业务产生干扰,并克服接入困难、传输速率受限及通信时延显著增加等问题,而传统的静态频谱管理方案难以有效应对这一挑战。

针对此问题,认知无线电(CR, cognitive radio)技术被提出。该技术通过频谱共享机制,允许次级网络(SN, secondary network)在满足干扰约束条件下动态接入主网络(PN, primary network)的频谱资源,其中,主网络和次级网络中的用户分别被称为

主用户 (PU, primary user) 和次级用户 (SU, secondary user)^[18]。特别地, 根据国际电信联盟无线电通信部门 (ITU-R, International Telecommunication Union Radiocommunication) 的频谱规划, Ka 频段中仅 19.7-20.2 GHz 和 29.5-30 GHz 分别专用于卫星通信的下行和上行链路, 其余频段均需与地面固定业务共享^[19]。在此背景下, 将 CR 技术引入 SAGIN, 可提高频谱资源的利用率^[20]。例如, 文献[21]研究了融合 CR 的卫星-高空平台网络, 在保证 PN 干扰阈值约束的情况下, 建立了 SN 和速率最大化问题, 联合优化了 SN 中的信道分配和功率分配。文献[22]考虑多个 UAV 作为空中基站通过共享卫星网络的频谱为地面用户提供服务, 并设计了一种多智能体深度强化学习算法, 以最小化长期任务周期内所有地面用户的总传输时延。文献[23]研究了一种 CR-SAGIN 框架, 其中, 卫星网络作为 PN, UAV 辅助智能车联网作为 SN, 并通过联合优化 UAV 轨迹、设备关联和发射功率最大化所有车辆的吞吐量。但是, 上述研究没有特别针对合作推理任务, 未考虑资源分配结果对模型分割策略的影响。此外, SAGIN 中大量的异构设备存在复杂的连接关系, 导致联合优化问题规模庞大且具有复杂性和非凸性, 一般算法难以兼顾求解精度和求解速度。

针对上述挑战, 本文提出了一种面向合作推理任务的 CR-SAGIN 框架。在此框架中, 卫星网络是

PN, 地面卫星接收器作为 PU 优先使用频谱, 而 UAV 网络是 SN, 即 UAV 作为 SU 与 PU 共享相同频谱并发起与 GD 之间的合作推理任务, 同时 UAV 需动态调整发射功率以确保对 PU 的干扰低于预设阈值。为了提高 SN 中的合作推理效率, 本文建立了设备关联、模型分割以及功率优化的联合优化问题。以最小化总推理时延为目标, 同时, 考虑卫星通信干扰阈值约束、GD 服务质量 (QoS, quality of service) 约束以及 UAV 发射功率约束。本文考虑到该问题是变量耦合的非凸问题, 根据优化变量将原问题分解为三个子问题。为取得求解精度和求解速度的均衡, 本文提出了一种融合二次无约束二值优化 (QUBO, quadratic unconstrained binary optimization) 和注意力机制的图学习法, 将设备关联子问题和模型分割子问题建模为 QUBO 形式, 并且通过松弛问题哈密顿量生成可微的损失函数, 进而利用动态多头注意力机制对图结构问题中的节点信息进行聚合, 最后对图神经网络 (GNN, graph neural network) 进行无监督训练, 从而完成设备关联优化和模型分割优化。此外, 使用连续凸近似 (SCA, successive convex approximation) 对发射功率进行优化, 最后通过交替迭代实现全局优化。仿真结果表明, 在多种不同场景下, 与多种基准方案相比, 该方法能够显著降低合作推理总时延, 提高合作推理任务的效率。

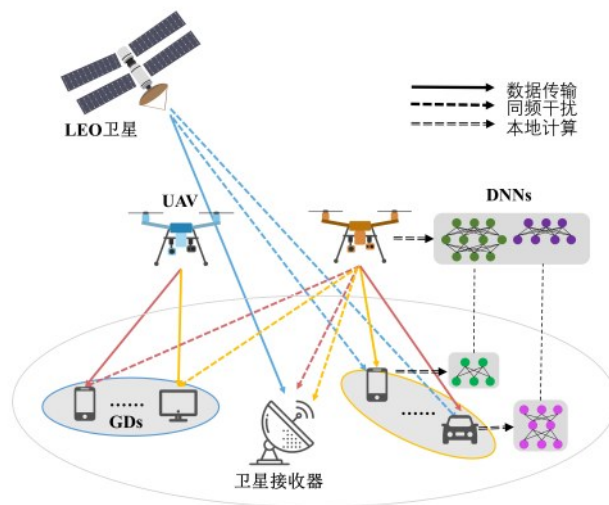


图1 面向合作推理任务的CR-SAGIN系统

1 系统模型

本文考虑一个面向合作推理任务的CR-SAGIN系统,如图1所示,其中,一个LEO卫星与一个地面卫星接收器组成的卫星网络作为PN, Q 个UAV组成的多UAV网络作为SN。多UAV网络可以与卫星网络共享相同的频谱以向它们正在服务的 K 个GD传输数据。为方便表示,将UAV的索引定义为 $q \in \{1, 2, \dots, Q\}$, 同时将受UAV服务GD的索引定义为 $k \in \{1, 2, \dots, K\}$ 。UAV的频谱资源被划分为多个子信道并与GD进行关联,子信道的索引定义为 $m \in \{1, 2, \dots, M\}$, 同时将与UAV q 关联的GD集合定义为 $\mathcal{K}_q$ 。由于所有UAV共享多个子信道,因此UAV会对其他UAV服务的GD造成干扰。

1.1 合作推理模型

由于在灾害救援、交通监管等常见AI应用场景下,DNN模型需要UAV收集的高空信息作为输入,所以本文将UAV作为合作推理任务的发起者。假设UAV发起的每个推理任务有 N 个子任务(DNN层),对于子任务 $n \in \{1, 2, \dots, N\}$, A_n 表示子任务 n 的计算工作量,单位为cycles。 B_n 表示子任务 n 的输出特征大小,即下一个子任务(第 $(n+1)$ 个子任务)的输入特征大小,单位为bits。特别地, B_0 用于表示第一个子任务的输入数据的大小,例如UAV采集的图像数据^[24]。UAV作为合作推理任务的发起者在高空采集原始信息,同时将DNN至多分割为两部分,一部分在UAV本地计算,另一部分卸载至关联的GD上计算。此外,GD可接入任一UAV的空闲信道。将UAV与GD之间的关联情况定义为二进制变量 $\mu_{q,k,m} \in \{0, 1\}$, 其中, $\mu_{q,k,m} = 1$ 表示UAV q 与GD k 通过信道 m 关联,否则 $\mu_{q,k,m} = 0$ 。此外,UAV可以将子任务卸载到与其关联的GD上进行合作推理,使用二进制变量 $x_{q,k,n} \in \{0, 1\}$ 表示每个子任务的计算位置, $x_{q,k,n} = 1$ 表示UAV q 与GD k 合作的同时,将第 n 个子任务在UAV本地处理,否则 $x_{q,k,n} = 0$ 。

1.2 计算模型

假设UAV分配固定的计算单元用于并行执行本地的推理任务,将UAV q 的本地计算速率定义为 f_q ,其单位是cycles/s。因此每架UAV的本地计算延迟为:

$$l_q^{\text{cp}} = \max \left\{ \sum_n \sum_m \frac{\mu_{q,k,m} x_{q,k,n} A_n}{f_q} \mid k \in \mathcal{K}_q \right\} \quad (1)$$

同理,将GD k 的本地计算频率定义为 f_k ,其单位是cycles/s。因此每个GD的本地计算延迟为:

$$l_k^{\text{cp}} = \sum_q \sum_n \sum_m \mu_{q,k,m} \frac{(1 - x_{q,k,n}) A_n}{f_k} \quad (2)$$

上式表示对于所有GD k ,若其与UAV q 关联,则将两者间所有本地子任务的计算时延进行累加,从而得到本地计算总时延。

1.3 通信模型

本文采用了一种广泛使用的空地(A2G, air-to-ground)自由空间路径损失模型^[25]。由于UAV在高空飞行且以视距(LoS, line-of-sight)传播为主,该自由空间路径损耗模型未考虑小尺度快衰落效应。此外,UAV移动性引起的多普勒效应假设已在卫星接收器和GD端得到完全补偿^[26]。因此,从UAV q 到卫星接收器 r 的信道增益为:

$$g_{q,r} = \frac{\beta_o}{H_u^2 + \|w_q - w_r\|^2} \quad (3)$$

其中, H_u 的为UAV的固定高度, w_q 和 w_r 分别为UAV q 和卫星接收机 r 的水平坐标, β_o 为距离1m处的参考信道增益。

类似地,UAV q 到GD k 的信道增益可以表示为:

$$g_{q,k} = \frac{\beta_o}{H_u^2 + \|w_q - w_k\|^2} \quad (4)$$

其中, w_k 为GD k 的水平坐标。

此外,与空地信道模型的假设类似,LEO与GD之间的星地链路同样以LoS传播为主^[27]。因此,从卫星 s 到GD k 的信道增益可以表示为:

$$g_{s,k} = \frac{\beta_o}{D^2 + \|w_s - w_k\|^2} \quad (5)$$

其中, D 是卫星离地面的高度, w_s 是卫星的水平坐标。针对大规模LEO卫星星座的实际应用场景,考虑到单颗卫星服务时长有限,期间信道增益波动范围较小,为降低算法复杂度,本文使用单颗卫星服务周期内的平均信道增益替代时变的增益。

考虑到设备关联过程的时间尺度远大于信道变化周期,本文假设快衰落效应已被平均化,且可获得理想CSI。因此,GD k 与UAV q 关联并占用信道 m 时,传输速率为:

$$R_{q,k,m} = B_m \cdot \log_2(1 + \gamma_{q,k,m}) \quad (6)$$

式中, B_m 为信道带宽, $\gamma_{q,k,m}$ 为GD k 的信干噪比

(SINR, signal-to-interference-plus-noise ratio) 可表示为:

$$\gamma_{q,k,m} = \frac{p_{q,m}g_{q,k}}{\sum_{j \neq q} p_{j,m}g_{j,k} + P_{SS}g_{s,k} + \sigma_k^2} \quad (7)$$

其中, P_{SS} 为卫星发射功率, σ_k^2 为GD端加性高斯白噪声 (AWGN, additive white Gaussian noise), $\sum_{j \neq q} p_{j,m}g_{j,k}$ 表示除 UAV q 以外其他所有 UAV 传输造成的同频干扰, $p_{q,m}$ 表示 UAV q 在信道 m 上的发射功率。

此外, 卫星接收器 r 收到的来自 UAV 的干扰为:

$$\zeta_r = \sum_{q=1}^Q \sum_{m=1}^M p_{q,m}g_{q,r} \quad (8)$$

值得注意的是, UAV 作为 SU 需控制发射功率以确保对卫星接收器的干扰 ζ_r 低于预设阈值 $\zeta_r^{\max}$ 。

此外, 二进制变量 $y_{q,k,n-1}^d$ 被用于判断 GD k 是否需要下载 UAV q 发起的第 $(n-1)$ 个子任务的输出中间数据, 即

$$y_{q,k,n-1}^d = x_{q,k,n-1} - x_{q,k,n} \quad (9)$$

则每架 UAV 与每个 GD 合作推理任务的中间传输延时为:

$$l_{q,k,n}^d = \sum_{m=1}^M \mu_{q,k,m} \left[x_{q,k,N} \frac{B_N}{R_{q,k,m}} + (1 - x_{q,k,1}) \frac{B_0}{R_{q,k,m}} + \sum_{n=2}^N y_{q,k,n-1}^d \frac{B_{n-1}}{R_{q,k,m}} \right] \quad (10)$$

上式括号中 $x_{q,k,N} \frac{B_N}{R_{q,k,m}}$ 表示 UAV 本地完成全部推理任务后, 将最终结果传输至 GD 所产生的时延, $(1 - x_{q,k,1}) \frac{B_0}{R_{q,k,m}}$ 表示 UAV 直接将原始数据传输至 GD 所产生的时延, $\sum_{n=2}^N y_{q,k,n-1}^d \frac{B_{n-1}}{R_{q,k,m}}$ 表示 UAV 将推理过程中的特征结果发送给 GD 所产生的时延。

综上所述, 合作推理总时延可以表示如下:

$$l_{\text{total}} = \sum_{q=1}^Q l_q^{\text{cp}} + \sum_{k=1}^K l_k^{\text{cp}} + \sum_{q=1}^Q \sum_{k=1}^K l_{q,k,n}^d \quad (11)$$

即总时延 l_{total} 包括所有 UAV 和 GD 的本地计算时延以及两者间的数据传输时延。

1.4 问题建立

以 SAGIN 中合理推理时延最小化为优化目标, 该网络的总时延由 UAV 的本地计算时延、UAV 与 GD 的中间数据传输时延以及 GD 的本地计算时延

三部分构成。将卫星通信的服务质量, 设备 QoS 以及发射功率满足需求作为约束条件, 联合优化合作推理计算卸载、设备关联以及发射功率分配。所建立的优化问题为:

$$\begin{aligned} \text{P1: } \min_{x, \mu, p} l_{\text{total}} \\ \text{s.t. } \text{C1: } \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} \leq N_q, \quad \forall q \\ \text{C2: } \sum_{q=1}^Q \sum_{m=1}^M \mu_{q,k,m} = 1, \quad \forall k \\ \text{C3: } \sum_{q=1}^Q \sum_{m=1}^M \mu_{q,k,m} \gamma_{q,k,m} \geq \gamma_k^{\min}, \quad \forall k \\ \text{C4: } \mu_{q,k,m} \in \{0, 1\}, \quad \forall q, k, m \\ \text{C5: } \zeta_r \leq \zeta_r^{\max}, \quad \forall r \\ \text{C6: } 0 \leq p_{q,m} \leq p_k^{\max}, \quad \forall q, m \\ \text{C7: } y_{q,k,n-1}^d \geq 0, \quad \forall q, k, n \geq 2 \\ \text{C8: } \sum_{m=1}^M \mu_{q,k,m} - x_{q,k,n} \geq 0, \quad \forall q, k, n \\ \text{C9: } x_{q,k,n} \in \{0, 1\}, \quad \forall q, k, n \end{aligned} \quad (12)$$

其中, C1 是指一个 UAV 可以关联的最大 GD 数约束, C2 表示一个 GD 只能与一架 UAV 关联并占用一个子信道, 由于本文只考虑被 UAV 服务的 GD, 故每个 GD 必须与 UAV 关联, C3 表示为了满足 GD 的 QoS 需求, GD 接收到的下行链路 SINR 不能小于其最小 SINR 要求 $\gamma_k^{\min}$, C5 表示卫星接收器 r 作为 PU 收到来自 SU 的干扰需低于预设阈值, C6 表示 UAV 的传输功率限制, C7 表示 UAV 和 GD 之间只进行下行数据传输, C8 表示 UAV q 与 GD k 关联时其对应的计算卸载二元变量 x 为 1, 否则为 0。

2 基于图学习的合作推理时延优化算法

问题 P1 具有非凸特性且存在耦合的多维变量, 难以直接获得全局最优解。为求解该问题, 首先, 基于变量解耦思想, 将原问题分解为设备关联优化、模型分割优化以及发射功率优化三个独立的子问题, 并且通过交替迭代求解三个子问题来实现全局性能的渐进优化。

2.1 设备关联优化

在模型分割策略和 UAV 发射功率配置确定的条件下, 本文构建了设备关联优化子问题 P2:

$$\begin{aligned}
& \text{P2: } \min_{\mu} l_{\text{total}} \\
& \text{s.t. C1: } \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} \leq N_q, \quad \forall q \\
& \text{C2: } \sum_{q=1}^Q \sum_{m=1}^M \mu_{q,k,m} = 1, \quad \forall k \\
& \text{C3: } \sum_{q=1}^Q \sum_{m=1}^M \mu_{q,k,m} \gamma_{q,k,m} \geq \gamma_k^{\min}, \quad \forall k \\
& \text{C4: } \mu_{q,k,m} \in \{0,1\}, \quad \forall q,k,m
\end{aligned} \quad (13)$$

P2 是一个具有复杂约束条件的 NP-hard 问题，其求解空间随问题规模呈指数级增长，导致传统优化算法难以在计算效率和求解精度之间取得良好平衡。针对此类复杂组合优化问题的求解，近期研究提出了将 QUBO 框架与 GNN 相结合的创新方法^[28]。QUBO 框架能够将复杂的组合优化问题转化为适合 GNN 处理的图结构表示，并且生成的问题哈密顿量松弛后可作为用于训练 GNN 的可微损失函数，而 GNN 凭借其对图结构数据强大的处理能力，可以显著提升优化过程的收敛速度。然而，在本文研究的 CR-SAGIN 场景下，该方法的直接应用面临两个主要挑战：首先，系统复杂的目标函数和约束条件需要经过特殊处理才能转化为标准的 QUBO 形式；其次，传统的 GNN 难以有效捕捉密集图结构中的关键节点特征，从而导致优化精度下降。为解决上述问题，本节提出一种基于 QUBO 框架和注意力机制的新型图学习方法。该方法设计专门的 QUBO 转化策略以处理复杂的目标函数和约束。同时，引入注意力机制增强 GNN 对关键节点的识别能力。

首先，将子问题 P2 的目标函数和约束建模为 QUBO 形式并给出问题哈密顿量。目标函数的哈密顿量 $H_{\text{P2}}^{\text{obj}}$ 可以表示为

$$\begin{aligned}
H_{\text{P2}}^{\text{obj}} = & \sum_{q=1}^Q \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} \left[\sum_{n=1}^N A_n \left(\frac{z_{q,n}}{f_q} + \right. \right. \\
& \left. \left. \frac{(1 - x_{q,k,n})}{f_k} \right) + \right. \\
& \left. \frac{B_N x_{q,k,N} + (1 - x_{q,k,1}) B_0 + \sum_{n=2}^N B_{n-1} y_{q,k,n-1}^d}{R_{q,k,m}} \right] \quad (14)
\end{aligned}$$

其中， $z_{q,n}$ 为判断 UAV q 是否需要本地计算子任务 n 的二元变量，此处为常数，具体定义见下一节，

$\sum_{n=1}^N A_n \left(\frac{z_{q,n}}{f_q} + \frac{(1 - x_{q,k,n})}{f_k} \right)$ 表示合作推理的计算时延，

$\frac{B_N x_{q,k,N} + (1 - x_{q,k,1}) B_0 + \sum_{n=2}^N B_{n-1} y_{q,k,n-1}^d}{R_{q,k,m}}$ 表示合作

推理的传输时延。

此外，C1 为线性不等式约束，难以直接转化成 QUBO 形式。为此，额外引入二进制变量 $a_{q,\beta} \in \{0,1\}, 0 \leq \beta \leq N_q$ 表示 UAV 关联的 GD 数。如果 UAV q 关联的 GD 数目为 β ，则 $a_{q,\beta}$ 为 1，否则为 0。

因此，C1 可以转化为 $\sum_{\beta=1}^{N_q} \beta a_{q,\beta} = \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} = \beta, \forall q$ ，进而其哈密顿量 $H_{\text{P2}}^{\text{C1}}$ 可表示如下：

$$\begin{aligned}
H_{\text{P2}}^{\text{C1}} = & \sum_{q=1}^Q \left[\left(1 - \sum_{\beta=0}^{N_q} a_{q,\beta} \right)^2 + \left(\sum_{\beta=0}^{N_q} \beta a_{q,\beta} - \right. \right. \\
& \left. \left. \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} \right)^2 \right] \quad (15)
\end{aligned}$$

其中，第一项确保每个 UAV q 关联的 GD 数目只有一种取值情况。第二项限制 UAV q 的实际关联数等于 β ，两者与约束 C1 逻辑等效。

同时，为了满足 C2，加入哈密顿量 $H_{\text{P2}}^{\text{C2}}$ 如下：

$$H_{\text{P2}}^{\text{C2}} = \sum_{k=1}^K \left(\sum_{q=1}^Q \sum_{m=1}^M \mu_{q,k,m} - 1 \right)^2 \quad (16)$$

上式规定每个 GD k 与所有无人机 q 和信道 m 的关联变量 $\mu_{q,k,m}$ 之和等于 1，即保障 GD k 只能与一架 UAV 关联并占用一个子信道。

类似地，C3 是一个复杂的非线性不等式约束，难以直接转化为 QUBO 形式，为此，额外引入二进制变量 $b_{q,k,m} \in \{0,1\}$ 用于判断 GD 的 QoS 要求是否被满足， $b_{q,k,m} = 1$ 表示 GD k 与 UAV q 关联并占用信道 m 时，下行链路 SINR γ_{qkm} 大于其最小要求 SINR，否则 $b_{q,k,m} = 0$ 。从而 C3 被转化为线性不等式约束 $\mu_{q,k,m} \leq b_{q,k,m}$ ，进而可以用哈密顿量 $H_{\text{P2}}^{\text{C3}}$ 表示：

$$H_{\text{P2}}^{\text{C3}} = \sum_{q=1}^Q \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} (1 - b_{q,k,m}) \quad (17)$$

综合考虑上述因素，P2 的问题哈密顿量 H_{P2} 可以由目标函数的哈密顿量和约束的哈密顿量的加权和表示^[29]：

$$H_{\text{P2}} = H_{\text{P2}}^{\text{obj}} + w_{\text{P2}}^{\text{C1}} H_{\text{P2}}^{\text{C1}} + w_{\text{P2}}^{\text{C2}} H_{\text{P2}}^{\text{C2}} + w_{\text{P2}}^{\text{C3}} H_{\text{P2}}^{\text{C3}} \quad (18)$$

其中， $w_{\text{P2}}^{\text{C1}}, w_{\text{P2}}^{\text{C2}}, w_{\text{P2}}^{\text{C3}}$ 是相应约束的权重因子，确定不同约束在总哈密顿量中的重要性。

其次，根据上述问题哈密顿函数 H_{P2} ，将变量

之间的耦合关系抽象为包括点集 $\mathcal{V}_\mu$ 和边集 $\mathcal{E}_\mu$ 的无向图 $\mathcal{G}_\mu = (\mathcal{V}_\mu, \mathcal{E}_\mu)$, 将每个二进制决策变量 $\mu_{q,k,m}$ 与对应顶点 $v_{q,k,m}^\mu \in \mathcal{V}_\mu$ 关联起来。若决策变量之间存在耦合 (H_{p2} 中对应的二次项系数不为0), 则在对应的两点间创建一条边, 所有边组成边集 $\mathcal{E}_\mu$, 该图的平均连接数 $R_\mu = |\mathcal{E}_\mu|/|\mathcal{V}_\mu|$ 。

然后, 结合问题哈密顿量 H_{p2} 和对应的无向图 $\mathcal{G}_\mu$, 使用 GNN 求解 P2。考虑到本文研究的图结构具有连接密集特性, 为增强模型对关键连接 (如严格约束惩罚项导致的变量耦合) 的识别能力, 使用将多头注意力机制与 GNN 相结合的图注意力网络 (GAT, graph attention network)^[30]。此外, 为解决深层 GAT 面临的过平滑问题, 引入跳跃知识 (JK, jumping knowledge) 模块来实现多尺度特征融合^[31]。具体而言, 本文的神经网络模型由 L 层具有 H 头注意力机制的 GAT 和一层 sigmoid 输出层组成, 并由 JK 模块进行层与层之间的特征融合。进而将离散的二进制决策变量 $\mu_{q,k,m}$ 松弛为连续概率参数 $\varphi_{q,k,m}^{q,k,m}(\theta)$, 并且构建可微的损失函数 $\mathcal{L}_{p2}(\theta)$:

$$\mu_{q,k,m} \rightarrow \varphi_{q,k,m}^{q,k,m}(\theta) \in [0,1] \quad (19)$$

$$H_{p2} \rightarrow \mathcal{L}_{p2}(\theta) \quad (20)$$

然后, 使用该损失函数对 GNN 的节点表示进行无监督训练。具体而言, 首先随机初始化 D_μ 维的节点嵌入向量, 然后依次进行前向传播计算节点表示、基于松弛哈密顿量计算损失函数、反向传播更新网络参数, 迭代多次直至满足预先设定的迭代停止条件, 最终得到最佳概率输出 φ_μ^* 。训练完成后, 采用基于概率阈值的贪婪算法将 φ_μ^* 映射回二元决策变量, 从而得到最优设备关联策略 μ^* 。

2.2 模型分割优化

在设备关联策略和 UAV 发射功率配置确定的条件下, 本文构建了模型分割优化子问题 P3:

$$\begin{aligned} \text{P3: } \min_x & l_{\text{total}} \\ \text{s.t. } & \text{C7: } y_{q,k,n}^d - 1 \geq 0, \forall q,k,n \geq 2 \\ & \text{C8: } \sum_{m=1}^M \mu_{q,k,m} - x_{q,k,n} \geq 0, \forall q,k,n \\ & \text{C9: } x_{q,k,n} \in \{0,1\}, \forall q,k,n \end{aligned} \quad (21)$$

子问题 P3 是一个具有复杂目标函数形式的 NP-hard 问题, 性质与 P2 相似, 所以采用相同的算法求解。首先, 将子问题 P3 的目标函数和约束建模为 QUBO 形式并给出问题哈密顿量。目标函数中的

l_q^{cp} 项为非线性项, 难以直接转化为 QUBO 形式, 为此, 额外引入二元变量 $z_{q,n} \in \{0,1\}$ 用于判断 UAV 是否需要本地计算子任务 n , 如果 UAV q 本地计算子任务 n 则 $z_{q,n} = 1$, 否则为 0。则 l_q^{cp} 可以转化为带有约束 C10 的线性项:

$$l_q^{\text{cp}} = \sum_{n=1}^N \frac{z_{q,n} \cdot A_n}{f_q} \quad (22)$$

$$\text{C10: } z_{q,n} = \bigvee_{k=1}^K x_{q,k,n}, \forall q,n$$

进一步, 舍弃无关的常数项后目标函数的哈密顿量 H_{p3}^{obj} 可以表示为:

$$\begin{aligned} H_{p3}^{\text{obj}} = & \sum_{q=1}^Q \sum_{k=1}^K \sum_{m=1}^M \mu_{q,k,m} \left[\sum_n \left(\frac{z_{q,n}}{f_q} - \frac{x_{q,k,n}}{f_k} \right) + \right. \\ & \left. \frac{B_N x_{q,k,N} - B_0 x_{q,k,1} + \sum_{n=2}^N B_{n-1} y_{q,k,n-1}^d}{R_{q,k,m}} \right] \end{aligned} \quad (23)$$

此外, 额外引入的 C10 是非线性约束, 难以直接转化为 QUBO 形式, 特别地, 在最小化问题中, $z_{q,n} \geq x_{q,k,n}, \forall q,k,n$ 与 C10 逻辑等价, 故 C10 可以使用哈密顿量 H_{p3}^{C10} 表示:

$$H_{p3}^{\text{C10}} = \sum_{q=1}^Q \sum_{k=1}^K \sum_{n=1}^N (1 - z_{q,n}) x_{q,k,n} \quad (24)$$

同时, 为了满足约束 C7, 加入哈密顿量 H_{p3}^{C7} 如下:

$$H_{p3}^{\text{C7}} = \sum_{q=1}^Q \sum_{k=1}^K \sum_{n=2}^N (1 - x_{q,k,n-1}) x_{q,k,n} \quad (25)$$

上式规定对于任意 UAV q 与 $\text{GD}k$ 之间的合作推理任务, 若第 $(n-1)$ 个子任务已经卸载至 $\text{GD}k$ 上计算, 则第 n 个子任务必然不在 UAV q 本地计算, 即 UAV 与 GD 之间只进行下行数据传输。

类似地, 为了满足约束 C8, 加入哈密顿量 H_{p3}^{C8} 如下:

$$H_{p3}^{\text{C8}} = \sum_{q=1}^Q \sum_{k=1}^K \sum_{n=1}^N \left(1 - \sum_{m=1}^M \mu_{q,k,m} \right) x_{q,k,n} \quad (26)$$

上式在 UAV q 和 $\text{GD}k$ 不关联时, 限制两者间所有合作推理子任务 n 对应的计算位置决策变量 $x_{q,k,n} = 0$ 。

综合考虑上述因素, P3 的哈密顿量可以表示如下, 其中 $w_{p3}^{\text{C7}}, w_{p3}^{\text{C8}}, w_{p3}^{\text{C10}}$ 分别为惩罚项权重:

$$H_{p3} = H_{p3}^{\text{obj}} + w_{p3}^{\text{C7}} H_{p3}^{\text{C7}} + w_{p3}^{\text{C8}} H_{p3}^{\text{C8}} + w_{p3}^{\text{C10}} H_{p3}^{\text{C10}} \quad (27)$$

其次, 根据上述问题哈密顿函数 H_{p3} , 将变量之间的耦合关系抽象为包括点集 $\mathcal{V}_x$ 和边集 $\mathcal{E}_x$ 的无

向图 $\mathcal{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$, 将每个二进制决策变量 $x_{q,k,n}$ 与每个顶点 $v_{q,k,n}^x \in \mathcal{V}_x$ 关联起来。若决策变量之间存在耦合 (H_{P3} 中对应的二次项系数不为 0), 则在对应的两点间创建一条边, 所有边组成边集 $\mathcal{E}_x$, 该图的平均连接数 $R_x = |\mathcal{E}_x|/|\mathcal{V}_x|$ 。最后, 随机初始化 D_x 维的节点嵌入向量并使用上一节提出的 GNN 对 P3 进行

$$\begin{aligned} \text{P4: } \min_p \quad & \sum_{q=1}^Q \sum_{k=1}^K \sum_{m=1}^M \frac{\mu_{q,k,m}}{B_m \log_2(1 + \gamma_{q,k,m})} \left[x_{q,k,N} B_N + (1 - x_{q,k,1}) B_0 + \sum_{n=2}^N y_{q,k,n-1}^d B_{n-1} \right] \\ \text{s.t. } \quad & \text{C5: } \zeta_r \leq \zeta_r^{\max}, \forall r \\ & \text{C6: } 0 \leq p_{q,m} \leq p_q^{\max}, \forall q, m \end{aligned} \quad (28)$$

问题 P4 是一个变量耦合的非凸问题, 难以直接求解。根据 SCA 基本原理, 对 P4 的求解可以转换为近似凸问题的迭代求解。为此, 引入辅助变量 $t_{q,k,m} = \ln(1 + \gamma_{q,k,m})$, 并添加约束 $t_{q,k,m} > 0$, 使目标函数保持凸性。此外, 在第 i 次迭代后的可行点 $(\mathbf{p}^{(i)}, \mathbf{t}^{(i)})$ 对 $e^{t_{q,k,m}}$ 一阶泰勒展开得到其下界, 将辅助变量引入的非凸等式约束松弛为凸不等式约束:

$$\begin{aligned} \text{P5: } \min_{\mathbf{p}, \mathbf{t}} \quad & \sum_{q=1}^Q \sum_{k=1}^K \sum_{m=1}^M \frac{\mu_{q,k,m} \ln 2}{B_m t_{q,k,m}} \left[x_{q,k,N} B_N + (1 - x_{q,k,1}) B_0 + \sum_{n=2}^N y_{q,k,n-1}^d B_{n-1} \right] \\ \text{s.t. } \quad & \text{C5: } \zeta_r \leq \zeta_r^{\max}, \forall r \\ & \text{C6: } 0 \leq p_{q,m} \leq p_q^{\max}, \forall q, m \\ & \text{C11: } t_{q,k,m} > 0, \forall q, k, m \\ & \text{C12: } p_{q,m} g_{q,k} \leq \left[e^{t_{q,k,m}^{(i)}} (1 + t_{q,k,m} - t_{q,k,m}^{(i)}) - 1 \right] I_{q,k}^{(i)}, \forall q, k, m \end{aligned} \quad (31)$$

显然, P5 为凸问题, 可以使用 CVX 工具箱求解, 进而得到最佳的功率分配 $\mathbf{p}^*$ 。

2.4 联合优化算法

基于上述对三个子问题的研究, 本文提出了一种联合优化算法。该算法采用迭代优化框架, 依次求解设备关联优化子问题 P2、模型分割优化子问题 P3 和发射功率优化子问题 P4, 优化其中某组变量时固定其余两组变量。特别地, 每次迭代获得的局部最优解将作为下一次迭代的初始值, 从而保证算法的单调收敛性。设置最大迭代次数 $\eta_{\max}$ 和收敛精度 ϵ 。第 η 次迭代结束前计算当前轮次的总推理时延 $l_{\text{total}}^{(\eta)}$, 并与上一轮次 ($\eta - 1$) 的总推理时延 $l_{\text{total}}^{(\eta-1)}$ 比较, 若 $|l_{\text{total}}^{(\eta)} - l_{\text{total}}^{(\eta-1)}| \leq \epsilon$, 则停止迭代, 并输出结果。算法 1 完整描述了该优化算法的具体实现流程。在算法 1 中, 每个迭代轮次包含三个优化步骤, 首先, 使用图学习法求解 P2 的过程主要涉及 GAT 模型的训练阶段, 故其时间复杂度可近似为

求解, 得到最优的模型分割策略 $\mathbf{x}^*$ 。

2.3 发射功率优化

在设备关联策略和模型分割策略确定的条件下, 舍弃无关常数, 本文构建了发射功率优化子问题 P4:

$$p_{q,m} g_{q,k} \leq \left(e^{t_{q,k,m}^{(i)}} (1 + t_{q,k,m} - t_{q,k,m}^{(i)}) - 1 \right) I_{q,k}^{(i)} \quad (29)$$

其中, $I_{q,k}^{(i)}$ 为第 i 次迭代后与 UAV q 关联的 GD k 收到的干扰, 可表示为:

$$I_{q,k}^{(i)} = \sum_{j \neq q} p_{j,m}^{(i)} g_{j,k} + P_{SS} g_{s,k} + \sigma_k^2 \quad (30)$$

进而, 在迭代点 $(\mathbf{p}^{(i)}, \mathbf{t}^{(i)})$ 处可构建问题 P4 的近似凸问题:

$\mathcal{O}(E_\mu LH | \mathcal{V}_\mu | ((3R_\mu D_\mu + R_\mu^2 + D_\mu^2)))$, 其中 E_μ 为求解 P2 时模型收敛所需的迭代次数。其次, 使用图学习法求解 P3 的时间复杂度为 $\mathcal{O}(E_x LH | \mathcal{V}_x | ((3R_x D_x + R_x^2 + D_x^2)))$, 其中 E_x 为求解 P3 时模型收敛所需的迭代次数^[32]。最后, 使用 SCA 算法求解 P4, 对于每个近似凸子问题使用内点法求解, 其时间复杂度为 $\mathcal{O}(E_p (QM)^{3.5})$, 其中 E_p 为近似求解次数^[33]。假设算法 1 运行 E 个轮次后达到收敛条件, 则整体时间复杂度为 $\mathcal{O}(E(E_\mu LH | \mathcal{V}_\mu | ((3R_\mu D_\mu + R_\mu^2 + D_\mu^2)) + E_x LH | \mathcal{V}_x | ((3R_x D_x + R_x^2 + D_x^2)) + E_p (QM)^{3.5}))$ 。

3 仿真测试

本文考虑到 5 架 UAV 被均匀部署在 $2000\text{m} \times 2000\text{m}$ 的区域内, 同时 50 个 GD 随机分布在整个区域, 模拟的 Walker Star 星座模型中 80 颗 LEO 卫星

算法1 联合优化算法**输入:** 系统参数、GNN模型参数、迭代参数

- 1: 初始化设备关联策略、模型分割策略、发射功率分配
- 2: 初始化 $l_{\text{total}}^{(\eta)} \rightarrow \infty$
- 3: **for** $\eta = 1$ to η_{max} **do**
- 4: 步骤1: 设备关联优化(图学习法)
- 5: 根据式(18)构建问题哈密顿量 H_{p_2} , 并构造无向图 $\mathcal{G}_\mu$
- 6: 将 H_{p_2} 松弛为 $\mathcal{L}_{p_2}(\theta)$, 以训练 GNN 输出 μ^*
- 7: 步骤2: 模型分割优化(图学习法)
- 8: 根据式(27)构建问题哈密顿量 H_{p_3} , 并构造无向图 $\mathcal{G}_x$
- 9: 将 H_{p_3} 松弛为 $\mathcal{L}_{p_3}(\theta)$, 训练 GNN 输出 x^*
- 10: 步骤3: 功率分配优化(SCA 算法)
- 11: 根据式(31)构建凸近似问题
- 12: 使用 CVX 工具箱迭代求解, 输出 p^*
- 13: 步骤4: 收敛判断
- 14: 计算当前轮次合作推理总时延 $l_{\text{total}}^{(\eta)}$
- 15: **if** $|l_{\text{total}}^{(\eta)} - l_{\text{total}}^{(\eta-1)}| \leq \epsilon$ **then**
- 16: 输出结果, 并退出循环
- 17: **end if**
- 18: **end for**

输出: 最优的设备关联策略 μ^* 、模型卸载策略 x^* 、功率分配 p^*

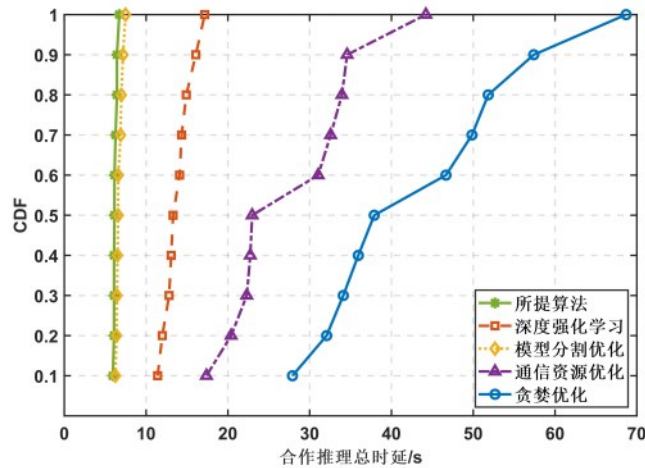
均匀分布在5个不同的轨道上, 高度为800km, 倾角为85°, 最小仰角为15°^[34], 目标区域的纬度和经度分别为32°N和120°E。具体的仿真环境参数如表1所示。星座规模主要影响单颗卫星对目标区域的服务时长, 然而, 通过测算分析发现, 单颗卫星服务时间远大于DNN模型合作推理时间, 合作推理期间卫星对UAV和GD产生的通信干扰波动幅度可忽略不计。因此, 在本文的仿真系统中, 可合理忽略星座规模变化对仿真结果的影响。特别地, 考虑

到灾害救援、交通监管等常见应用场景下无人机均需执行高空图像识别任务, 本文选取VGG16深度卷积神经网络作为基准推理模型^[35]。

表1 仿真环境参数

参数	数值
GD数 K	50
GU最低要求 $\text{SINR}_{\gamma_k}^{\min}$	0dB
AWGN功率 σ_k^2	-169 dBm/Hz
UAV数 Q	5
UAV高度 H_u	100m
UAV最大传输功率 $p_q^{\max}$	100W
卫星高度 D	800km
卫星发射功率 P_{ss}	1000W
卫星接收器干扰阈值 $\zeta_r^{\max}$	-50dBm
子信道数 M	10
GD计算速率 f_k	$2 \times 10^{13} \sim 1 \times 10^{14}$ cycles/s
UAV计算速率 f_q	3×10^{13} cycles/s

为验证本文所提算法的性能优势, 本文在多种场景下将其与四种基准方案进行对比。“贪婪优化”方案指关联的UAV和GD之间的推理任务由计算能力较强的一方本地计算, 同时, UAV平均分配子信道发射功率, GD优先接入下行SINR最优的信道。“深度强化学习”方案指文献[36]中提出的基于深度强化学习的在线联合优化卸载策略和通信资源分配的优化方案。“模型分割优化”方案指仅采用本文所提方法对模型分割策略进行优化。“通信资源优化”方案指仅采用本文所提方法对设备关联策略和发射功率分配进行优化。

**图2** 十种不同场景下合作推理总推理时延 CDF

如图2所示, 本文随机生成十种不同的GD位置分布场景, 并测试不同方案在这十种场景下的合作推理总时延累计分布函数(CDF, cumulative distribution function), 从图中可以明显看出, 所提算法能够更高效地进行合作推理, 尤其是在处理SAGIN中异构设备间的复杂连接关系时, 通过

GNN对复杂图结构的高效表征学习可以更加合理地进行设备关联和模型分割, 从而显著降低合作推理时延。此外, SCA高效收敛于局部最优的功率优化策略, 有效抑制了同频干扰, 降低了传输时延, 进一步提高了合作推理效率。

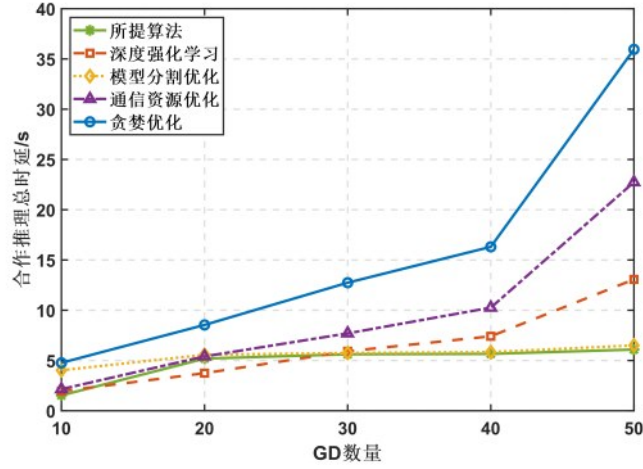


图3 不同GD数量合作推理总时延

如图3所示, 在区域内部署5架UAV的情况下, 比较了不同GD数量下各方案的合作推理总时延。从图中可以看出随着GD数量的增加, 总推理时延显著上升, 这是由于推理任务数量增加且同频干扰加重, 进而导致计算时延和传输时延增大。显然, 所提算法在存在大量GD的复杂网络环境下具有显著优势。这一现象验证了GNN在建模复杂网络连接关系方面的独特优势。

不同卫星接收器干扰阈值下合作推理总时延如图4所示, 只对涉及发射功率优化的算法进行比较

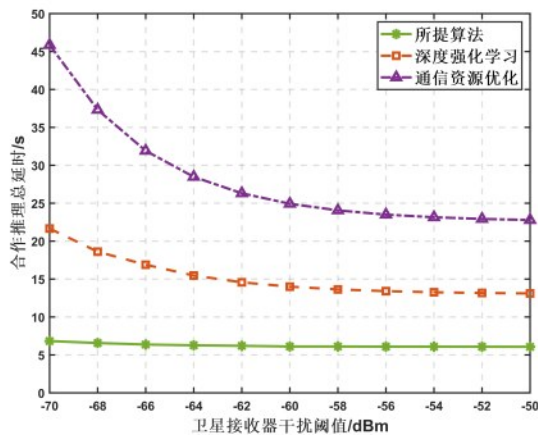


图4 不同卫星接收器干扰阈值合作推理总时延

较, 即本文所提算法, “深度强化学习”方案以及“通信资源优化”方案。仿真结果显示所提算法在各对比方案中始终保持最低的合作推理总时延。特别是在严格的干扰阈值约束条件下仍能保持稳定的性能表现, 展现出优异的鲁棒性。这是由于联合优化框架通过多维参数协同优化, 显著拓了解空间搜索范围, 并且基于GNN的智能搜索机制有效提升了复杂约束条件下的解空间探索能力。

不同宽带下合作推理总时延如图5所示, 从图中可以看出, 随着带宽增加, 各算法性能均有所提升, 但所提算法始终保持着稳定的性能优势, 相较于“贪婪优化”方案, 本算法平均降低时延93%; 相比“深度强化学习”方案, 性能提升达65%; 与“通信资源优化”和“模型分割优化”方案相比, 合作总推理时延分别降低了86%和11%。

4 结束语

本文研究了融合CR的SAGIN中高效合作推理方法, 联合优化设备关联、模型分割以及功率分配以降低总推理时延。考虑卫星通信干扰阈值约束、设备服务质量约束以及发射功率约束, 以最小化推理时延为目标, 建立了设备关联、模型分割以及功

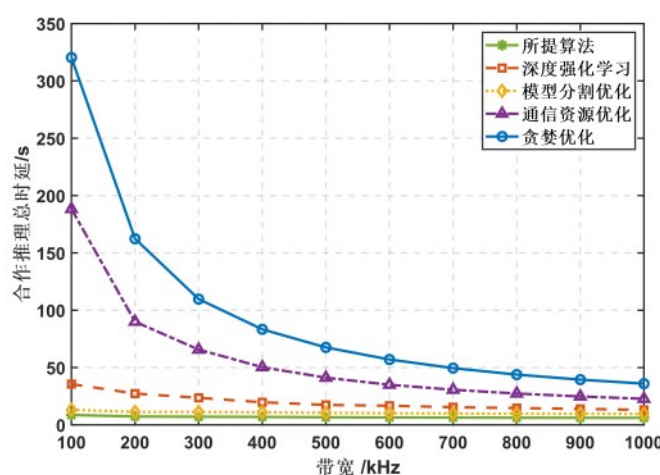


图5 不同带宽合作推理总时延

率优化的联合优化问题。该问题具有非凸性和复杂性，为了解决该问题，本文将其根据优化变量分解为三个子问题，并进行交替迭代求解。为取得求解精度和求解速度的均衡，本文提出了一种融合 QUBO 框架和注意力机制的图学习方法用于优化设备关联和模型分割，同时，使用 SCA 优化功率分配。仿真结果表明，在面向常见的 DNN 时，所提方法能够提高合作推理任务的执行效率。在实际系统部署中，所提算法的应用会引入一定的计算与资源开销。为降低算法对系统性能的影响，在未来的研究中，可考虑将 UAV 与地面云服务器进行协同部署，利用云端强大的计算资源分担本地计算负载，或采用网络剪枝、量化压缩等模型轻量化技术，在保证 GNN 模型性能的同时有效降低计算复杂度。

参考文献：

- [1] YU S, GONG S W, SHI Q, et al. EC-SAGINs: Edge-Computing-Enhanced Space - Air - Ground-Integrated Networks for Internet of Vehicles[J]. IEEE Internet of Things Journal, 2022, 9(8): 5742-5754.
- [2] TAN J J, TANG F X, ZHAO M, et al. Outage Probability, Performance, and Fairness Analysis of Space - Air - Ground Integrated Network (SAGIN): UAV Altitude and Position Angle[J]. IEEE Transactions on Wireless Communications, 2025, 24(2): 940-954.
- [3] 武艳, 潘广川, 姚明许, 等. 基于深度强化学习的空天地一体化网络信息物理系统垂直切换策略[J]. 通信学报, 2024, 45(08): 180-191.
- [4] FU S, GAO J, ZHAO L. Collaborative Multi-Resource Allocation in Terrestrial-Satellite Network Towards 6G[J]. IEEE Transactions on Wireless Communications, 2021, 20(11): 7057-7071.
- [5] LEE J-H, PARK J, BENNIS M, et al. Integrating LEO Satellites and Multi-UAV Reinforcement Learning for Hybrid FSO/RF Non-Terrestrial Networks[J]. IEEE Transactions on Vehicular Technology, 2023, 72(3): 3647-3662.
- [6] CHEN R R, SUN Y J, LIANG L P, et al. Joint Power Allocation and Placement Scheme for UAV-Assisted IoT With QoS Guarantee [J]. IEEE Transactions on Vehicular Technology, 2022, 71(1): 1066-1071.
- [7] QU K G, ZHUANG W H, WU W, et al. Stochastic Cumulative DNN Inference With RL-Aided Adaptive IoT Device-Edge Collaboration[J]. IEEE Internet of Things Journal, 2023, 10(20): 18000-18015.
- [8] XU X D, YAN K W, HAN S J, et al. Learning-Based Edge-Device Collaborative DNN Inference in IoVT Networks[J]. IEEE Internet of Things Journal, 2024, 11(5): 7989-8004.
- [9] OTTER D W, MEDINA J R, KALITA J K. A Survey of the Usages of Deep Learning for Natural Language Processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 604-624.
- [10] Bakambekova A, Kouzayha N, Al-Naffouri T. On the Interplay of Artificial Intelligence and Space-Air-Ground Integrated Networks: A Survey[J]. IEEE Open Journal of the Communications Society, 2024, 5: 4613-4673.
- [11] 王俊华, 罗菲, 高广鑫, 等. 动态车辆网络场景中的协同空地计算卸载和资源优化[J]. 电子与信息学报, 2025, 47(01): 102-115.
- [12] WANG J H, Luo F, GAO G X, et al. Collaborative Air-Ground Computation Offloading and Resource Optimization in Dynamic Vehicular Network Scenarios[J]. Journal of Electronics & Information Technology, 2025, 47(01): 102-115.
- [12] HE W C, GUO S Y, GUO S, et al. Joint DNN Partition Deployment and Resource Allocation for Delay-Sensitive Deep Learning

- Inference in IoT[J]. IEEE Internet of Things Journal, 2020, 7(10): 9241-9254.
- [13] YAN J, BI S Z, Y. ZHANG Y-J A. Optimal Model Placement and Online Model Splitting for Device-Edge Co-Inference[J]. IEEE Transactions on Wireless Communications, 2022, 21(10): 8354-8367.
- [14] HAN S J, ZHANG W Z, XU X D, et al. S2E-DECI: Secrecy and Energy-Efficient Dual-Aware Device-Edge Co-Inference for AIoT [J]. IEEE Internet of Things Journal, 2024, 11(24): 39142-39157.
- [15] LIU C X, FENG W, CHEN Y F, et al. Cell-free satellite-UAV networks for 6G wide-area Internet of Things[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(4): 1116-1131.
- [16] WANG Z C, ZHANG L, FENG D Q, et al. Intelligent Cloud-Edge Collaborations for Energy-Efficient User Association and Power Allocation in Space-Air-Ground Integrated Networks[J]. IEEE Journal on Selected Areas in Communications, 2024, 42(12): 3659-3673.
- [17] WANG Y F, ZHAO L Q, CHU X L, et al. E2E Network Slicing Optimization for Control- and User-Plane Separation-Based SAGIN With DRL[J]. IEEE Transactions on Vehicular Technology, 2024, 73(11): 16680-16696.
- [18] ZHOU Z Z, ZHANG Q Q, GE J G, et al. Hierarchical Cognitive Spectrum Sharing in Space-Air-Ground Integrated Networks[J]. IEEE Transactions on Wireless Communications, 2024, 24(2): 1430-1447.
- [19] NASERI P, MATOS S A, COSTA J R, et al. Dual-Band Dual-Linear-to-Circular Polarization Converter in Transmission Mode Application to K/Ka -Band Satellite Communications[J]. IEEE Transactions on Antennas and Propagation, 2018, 66(12): 7128-7137.
- [20] 逢小玮, 蒋旭, 卢华兵, 等. 面向6G多维扩展的新型多址接入技术综述[J]. 电子与信息学报, 2024, 46(06): 2323-2334.
- PANG X W, JIANG X, LU H B, et al. An Overview of Novel Multi-access Techniques for Multi-dimensional Expanded 6G[J]. Journal of Electronics & Information Technology, 2024, 46(06): 2323-2334.
- [21] LIU R, GUO K F, AN K, et al. Resource Allocation for Cognitive Satellite-HAP-Terrestrial Networks With Non-Orthogonal Multiple Access[J]. IEEE Transactions on Vehicular Technology, 2023, 72(7): 9659-9663.
- [22] GUO S A, ZHAO X H. Multi-Agent Deep Reinforcement Learning Based Transmission Latency Minimization for Delay-Sensitive Cognitive Satellite-UAV Networks[J]. IEEE Transactions on Communications, 2023, 71(1): 131-144.
- [23] PERVEZ F, ZHAO L, YANG C G. Joint User Association, Power Optimization and Trajectory Control in an Integrated Satellite-Aerial-Terrestrial Network[J]. IEEE Transactions on Wireless Communications, 2022, 21(5): 3279-3290.
- [24] SHI W Q, ZHOU S, NIU Z S, et al. Multiuser Co-Inference With Batch Processing Capable Edge Server[J]. IEEE Transactions on Wireless Communications, 2023, 22(1): 286-300.
- [25] WU Q Q, ZENG Y, ZHANG R. Joint Trajectory and Communication Design for Multi-UAV Enabled Wireless Networks[J]. IEEE Transactions on Wireless Communications, 2018, 17(3): 2109-2121.
- [26] ABBASI O, YANIKOMEROGLU H, EBRAHIMI A, et al. Trajectory Design and Power Allocation for Drone-Assisted NR-V2X Network With Dynamic NOMA/OMA[J]. IEEE Transactions on Wireless Communications, 2020, 19(11): 7153-7168.
- [27] GUO K F, AN K, ZHANG B N, et al. Physical Layer Security for Multiuser Satellite Communication Systems With Threshold-Based Scheduling Scheme[J]. IEEE Transactions on Vehicular Technology, 2020, 69(5): 5129-5141.
- [28] SCHUETZ M J A, BRUBAKER J K, KATZGRABER H G. Combinatorial optimization with physics-inspired graph neural networks[J]. Nature Machine Intelligence, 2022, 4(4): 367-377.
- [29] ATOBE Y, TAWADA M, TOGAWA N. Hybrid Annealing Method Based on subQUBO Model Extraction With Multiple Solution Instances[J]. IEEE Transactions on Computers, 2022, 71(10): 2606-2619.
- [30] HU B, GUO K H, WANG X K, et al. RRL-GAT: Graph Attention Network-Driven Multilabel Image Robust Representation Learning[J]. IEEE Internet of Things Journal, 2022, 9(12): 9167-9178.
- [31] LI Z W, WANG Q K, YUAN C A, et al. Predicting MiRNA-Disease Associations by Graph Representation Learning Based on Jumping Knowledge Networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(5): 2629-2638.
- [32] He J C, Cheng N, Yin Z S, et al. Load-Aware Network Resource Orchestration in LEO Satellite Network: A GAT-Based Approach [J]. IEEE Internet of Things Journal, 2024, 11(9): 15969-15984.
- [33] Zeng Y P, Chen S S, Li J D, et al. Online Optimization in UAV-Enabled MEC System: Minimizing Long-Term Energy Consumption Under Adapting to Heterogeneous Demands[J]. IEEE Internet of Things Journal, 2024, 11(9): 32143-32159.
- [34] Han D-J, Fang W Z, Hosseinalipour S, et al. Orchestrating Federated Learning in Space-Air- Ground Integrated Networks: Adaptive Data Offloading and Seamless Handover[J]. IEEE Journal on Selected Areas in Communications, 2024, 42(12): 3505-3520.
- [35] ZHANG X Y, ZOU J H, HE K M, et al. Accelerating Very Deep Convolutional Networks for Classification and Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 1943-1955.
- [36] WANG Y T, QIAN Z, HE L J, et al. Intelligent Online Computation Offloading for Wireless-Powered Mobile-Edge Computing[J]. IEEE Internet of Things Journal, 2024, 11(17): 28960-28974.