基于模型强化学习的边缘计算无线大语言模型推理自适应层切分方法

陈宇轩[1]，李荣鹏[1]，于小雪[1]，赵志峰[2]，张宏纲[1]
[1]浙江大学信息与电子工程学院，中国杭州市，310027
[2]之江实验室，中国杭州市，310012

摘要：在边缘计算环境中优化大型语言模型（LLMs）的部署对提升隐私保护和计算效率至关重要。为实现高效的无线LLM推理，本文全面分析了主流开源LLMs中不同分割点的影响。本文引入一个基于模型的强化学习（MBRL）框架，以确定边缘和用户设备（UE）之间的最佳分割点。通过引入奖励替代模型，该方法显著减少了频繁的性能评估的计算成本。广泛的仿真结果表明，该方法在不同网络条件下有效地平衡了推理性能和计算负载，为去中心化环境中LLM的部署提供稳健的解决方案。

关键词：大型语言模型；边缘计算；基于模型的强化学习；分裂推理；Transformer模型
https://doi.org/10.1631/FITEE.2400468