LLM-Based Task Offloading for MEC Systems

Muhammad Sohaib, Sang-Woon Jeon

Hanyang University

sangwoonjeon@hanyang.ac.kr

LLM 기반 MEC 시스템을 위한 테스크 오프로딩

소화이브무하마드, 전상운
한양대학교

Abstract

The proliferation of Internet of Things (IoT) devices has fueled the demand for mobile edge computing (MEC) to support low-latency, high-performance applications. However, uneven load distribution among MEC servers can cause increased delays and degraded quality of service. This paper proposes a large language model (LLM)-based task offloading policy designed to achieve efficient load balancing by minimizing the average maximum server delay. The proposed approach leverages the reasoning capabilities of LLMs to make dynamic offloading decisions based on real time system states, heterogeneous resource capacities, and varying workloads.

Ⅰ. Introduction

The widespread adoption of IoT devices require high data rates, computational and storage resources, for which MEC [1] has emerged as promising technology. MEC offers low latency and localized processing, but fully leveraging its potential requires optimized use of channels, computation, and power. Existing heuristic and learning-based offloading methods struggle with network dynamics or demand extensive training. Recent advances show that LLMs [2] can provide strong reasoning and adaptive decision-making, making them promising tools for MEC resource management.

In this work, we propose an LLM-based online task offloading policy for MEC systems, aimed at achieving efficient load balancing in real time. The proposed approach applies the LLM repeatedly to newly arrived tasks, enabling dynamic offloading decisions that minimize the maximum server delay across multiple edge servers at each time slot, thereby improving delay performance and enhancing overall system utilization.

Ⅱ. Simulation Results

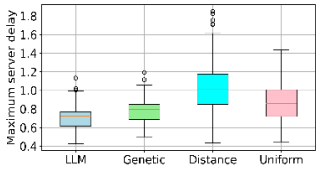

Fig. 1. Distribution of maximum server delays over time for different task offloading schemes.

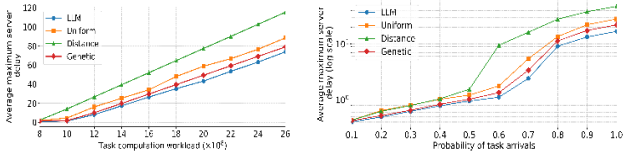

Fig. 2. Average maximum server delay versus task arrival probabilities and task computation workloads.

We compare the performance of proposed method with three baseline schemes: genetic, distance, and uniform scheme. Fig. 1 illustrates the distribution of maximum server delays over time for each of the previously mentioned schemes for L = 30 and λ = 0.5. Fig. 2 depicts how the average maximum server delay changes as the task computation workload and generation probability increase. As illustrated in figures, LLM-based scheme outperforms the benchmark schemes.

Ⅲ. Conclusion

In this work, we proposed an LLM-based task offloading policy for MEC networks, aimed at achieving effective load balancing by minimizing the average maximum server delay. The proposed scheme makes efficient task offloading decisions in real time without a training stage required for machine learning or reinforcement learning.

ACKNOWLEDGMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant RS-2024-00405128. S.-W. Jeon is the corresponding author.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322-2358, 2017.

[2] K. S. Kalyan, "A survey of GPT-3 family large language models including chatgpt and GPT-4," *Natural Language Processing Journal*, vol. 6, p. 100048, 2024.