PhD Forum Abstract: Diffusion-based Task Scheduling for Efficient AI-Generated Content in Edge Networks

Changfu Xu
Hong Kong Baptist University, Hong Kong and BNU-HKBU United International College, Zhuhai, China
changfuxu@uic.edu.cn

ABSTRACT

The Artificial Intelligence-Generated Content (AIGC) technique has gained significant popularity in creating diverse content. However, the current deployment of AIGC services is a centralized framework, thus leading to high response times. To address this issue, we propose a diffusion-based task scheduling method that considers the integration of the diffusion model, Deep Reinforcement Learning (DRL), and Mobile Edge Computing (MEC) technique to improve the AIGC efficiency. This challenges efficient server selection without prior information in dynamic MEC systems. We formulate our problem as an online integer linear programming problem aiming to minimize task offloading delay. Furthermore, we propose a novel AIGC Task Scheduling (DDRL-ATS) algorithm based on Diffusion DRL (DDRL) that effectively addresses this problem. The DDRL-ATS algorithm achieves efficient AIGC tailored for heterogeneous MEC environments. Additionally, an online Adaptive Multi-server Selection and Allocation (DDRL-AMSA) algorithm based on DDRL is proposed to further enhance the AIGC efficiency. Moreover, our DDRL-AMSA algorithm achieves near-optimal solutions within approximate linear time complexity bounds. Finally, experimental results validate the effectiveness of our method by showcasing at least a reduction of 13.54% in task offloading delay compared to state-of-the-art methods.

KEYWORDS

AIGC, Collaborative MEC, Diffusion model, Task scheduling, Deep reinforcement learning

1 PROBLEM

The technique of Artificial Intelligence-Generated Content (AIGC) has garnered substantial attention and demonstrated considerable success in various industrial applications, such as ChatGPT, Sora, and QWen [1]. These models autonomously generate satisfactory textual, graphical, and video content in response to provided prompts. However, owing to the intricate architecture of AIGC models, they often necessitate a substantial allocation of computing resources for effective task processing. Consequently, current AIGC models predominantly operate within centralized frameworks deployed on Cloud Server (CS), leading to prolonged service response times [6]. Notably, users may experience wait times of 40-60 seconds for image generation on platforms such as Hugging Face (https://huggingface.co/spaces). Therefore, there is an urgent need to design methods aimed at enhancing the Quality of user Experience (QoE) for AIGC applications.

2 RELATED WORKS

The Mobile Edge Computing (MEC) technique has been recognized as a promising solution to provide low service delay for many computation-intensive Internet of Things applications [4]. According to this advantage, the MEC technique was applied to improve the QoE of AIGC services [8]. For example, Xu *et al.* [9] investigate edge-based pre-trained foundation models serving problems for mobile AIGC services. In this work, a joint AIGC model caching and inference framework in MEC systems is presented, which manages AIGC models and allocates resources to satisfy users' requests efficiently. Wang *et al.* [6] explore a distributed inference method for next-word prediction in edge networks. This method efficiently utilizes the computing resources of Edge Servers (ESs), thus reducing task offloading latency. However, these methods are designed by leveraging traditional machine learning algorithms, which limits the performance of the MEC system.

Recently, since diffusion models have achieved state-of-the-art results in many AIGC applications, a diffusion-based Q-learning method is proposed to seek optimal actions in terms of behavior cloning and policy improvement [7]. This method utilizes a conditional diffusion model to represent the policy in a Deep Q-Network (DQN), improving action reward [5]. Furthermore, Du *et al.* [2] propose an edge-based AIGC approach to maximize the human-aware content quality. This approach performs discrete action decisions for AIGC tasks based on the diffusion model and Soft Actor-Critic (SAC) model [3], thus improving the user's subjective experience. However, we note that these existing methods do not consider the response time of AIGC services, which diminishes the QoE of the entire system.

3 DIFFUSION-BASED TASK SCHEDULING FOR EFFICIENT AIGC IN EDGE NETWORKS

Based on the above analyses, we observe that the feature of iterative refinement in diffusion models can guide the improvement of the actor policy at each reverse diffusion step. Therefore, we propose a diffusion-based task scheduling method that integrates the diffusion model, Deep Reinforcement Learning (DRL), and MEC technique to improve the AIGC efficiency.

The system architecture of our method is shown in Figure 1. The network consists of several Base Stations (BSs) and each BS is equipped with one ES. Each ES deploys an AIGC model that has been trained. Different ESs have different computing capacities. All BSs are connected via a wired core network. Each ES has a scheduler with a processing queue. The processing queue stores the new arrival task workloads waiting for processing by CPU or GPU if GPU is available. Based on this MEC system, when the AIGC task requests arrive at a BS at a time slot, the BS scheduler will decide which task workloads are allocated to which ESs for parallel processing in a distributed way.

Our method considers the distributed deployment of AIGC services on ESs. Then, we formulate our problem as an online integer

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00063

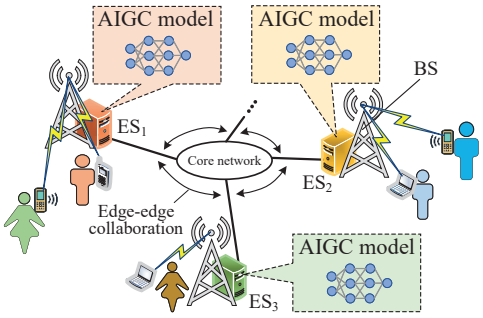

Figure 1: System architecture of our method.

linear programming problem to minimize the sum of all task offloading delays in the entire system. Furthermore, we propose an AIGC Task Scheduling (DDRL-ATS) algorithm to solve the problem based on a Diffusion DRL (DDRL) model. By taking our method, the AIGC tasks are parallelly processed in multiple ESs, thus reducing the response time of AIGC services.

Besides, to further enhance the AIGC efficiency, we also propose an Adaptive Multi-server Selection and Allocation (DDRL-AMSA) algorithm based on DDRL. The DDRL-AMSA algorithm simultaneously enables multiple ESs' idle resources to accelerate task processing. Moreover, it achieves a near-optimal solution to our problem within an approximate linear time complexity [8].

4 PRELIMINARY RESULTS

We have developed an edge computing prototype consisting of 3 Jetson-Xavier-NX devices, 1 computer with an Intel Core i7 2.2 GHz Processor, and 1 computer with 12 Intel Core i5 2.7 GHz Processors and an NVIDIA Quadro RTX 4000 graphics card. Our experiments are executed in this prototype environment. The special results are given as follows.

DSAC Algorithm. To validate the effectiveness of the diffusion model in DRL, we implement the Diffusion-based SAC (DSAC) algorithm and compare it with two well-known DRL baselines of DQN [5] and SAC [3]. The well-known DRL environment of CartPole-v1 is used in our experiments. The experimental results are shown in Fig. 2. Fig. 2 shows that our DASC algorithm significantly outperforms the DQN and SAC baselines in reward. In particular, compared to the two baselines, it improves the average rewards of 9.57% and 18.36%, respectively.

DDRL-AMSA Algorithm. We compare the proposed DDRL-AMSA algorithm with two state-of-the-art baselines of DRLCoEdge [4] and SMCoEdge [8]. The experiments are executed in a simulation environment with random task generation. The results illustrate the superiority of our DDRL-AMSA algorithm over the DRLCoEdge and SMCoEdge methods, achieving a minimum 18.22% and 13.54% reduction in task offloading delay, respectively.

The above results strongly demonstrate that our method significantly improves the AIGC's efficiency and reliability in MEC systems, and has substantial improvement compared to the state-of-the-art methods.

5 CHALLENGES AND FUTURE WORKS

Challenges: Different AIGC tasks usually present varying quality requisites, such as precision, response time, and model magnitude,

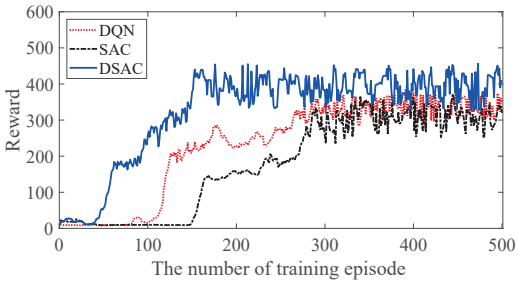

Figure 2: Performance comparison of DQN, SAC, and DSAC.

thereby necessitating diverse allocations of computing and networking resources. However, the finite nature of ES resources challenges the resource allocation for efficient AIGC. Furthermore, the availability of computing resources in ESs remains uncertain in dynamic MEC systems, posing a substantial challenge in achieving minimal task offloading delay absent prior information. Moreover, the process of ES selection and workload allocation inherent to AIGC tasks is decided online, which renders the attainment of an optimal solution within polynomial time computationally infeasible.

Future Works: In this work, we only tested the superiority of our method over the state-of-the-art methods on some simple DRL applications (e.g., CartPole). In future works, we plan to do experimental tests on complex AIGC applications (e.g., image inpainting) by our method. Also, we intend to extend our method to other types of neural networks, aiming to seek better results.

ACKNOWLEDGMENTS

I started my PhD study in September 2021. My expected graduation date is September 2025. I am many thanks to Prof. Tian Wang, Wentao Fan, and Jianxiong Guo's valuable guidance in this work.

REFERENCES

[1] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023), 1–44.

[2] H. Du, Z. Li, D. Niyato, J. Kang, Z. Xiong, and D. I. Kim. 2023. Enabling AI-generated content (AIGC) services in wireless edge networks. *arXiv preprint arXiv:2301.03220* (2023).

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning (PMLR)*, Vol. 80. PMLR, 1861–1870.

[4] Mushu Li, Jie Gao, Lian Zhao, and Xuemin Shen. 2020. Deep reinforcement learning for collaborative edge computing in vehicular networks. *IEEE Transactions on Cognitive Communications and Networking* 6, 4 (2020), 1122–1135.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[6] Shangshang Wang, Ziyu Shao, and John C.S. Lui. 2023. Next-Word Prediction: A Perspective of Energy-Aware Distributed Inference. *IEEE Transactions on Mobile Computing* (2023), 1–14. https://doi.org/10.1109/TMC.2023.3310536

[7] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. 2022. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193* (2022).

[8] Changfu Xu, Jianxiong Guo, Yupeng Li, Haodong Zou, Weijia Jia, and Tian Wang. 2024. Dynamic parallel multi-server selection and allocation in collaborative edge computing. *IEEE Transactions on Mobile Computing* (2024), 1–15. https://doi.org/10.1109/TMC.2024.3376550

[9] Minrui Xu, Dusit Niyato, Hongliang Zhang, Jiawen Kang, Zehui Xiong, Shiwen Mao, and Zhu Han. 2023. Sparks of Generative Pretrained Transformers in Edge Intelligence for the Metaverse: Caching and Inference for Mobile Artificial Intelligence-Generated Content Services. *IEEE Vehicular Technology Magazine* 18, 4 (2023), 35–44.