ML-Based Task-Oriented Offloading: A Survey

Dongwook Won
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
dwwon@uclab.re.kr

Thanh Phung Truong
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
tptruong@uclab.re.kr

Chihyun Song
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
chsong@uclab.re.kr

Jaemin Kim
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
jmkim@uclab.re.kr

Tung Son Do
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
dwwon@uclab.re.kr

Sungrae Cho
School of Computer Science and Engineering
Chung-Ang University
Seoul 06974,
South Korea
srcho@cau.ac.kr

Abstract—**In the Industrial Internet of Things (IIoT), machine learning (ML) plays a crucial role in driving intelligent, data-driven processes, especially in applications that require computation-intensive tasks, such as real-time defect detection, predictive maintenance, and quality inspection. To meet the demands of these ML tasks, effective offloading strategies are essential for reducing latency and ensuring high inference accuracy. This paper surveys recent advancements in ML-based task-oriented offloading strategies, focusing on two primary directions: (1) enhancing timeliness and quality in computation-intensive tasks and (2) ensuring accuracy-aware offloading for high-reliability applications in IIoT. The first direction addresses the need for rapid processing by minimizing task delays and optimizing real-time scheduling, while the second focuses on maintaining inference accuracy across resource-constrained environments, essential for industrial applications. By categorizing and analyzing the latest strategies, this paper highlights the evolution of ML-based task offloading in IoT and IIoT systems, emphasizing the balance between timeliness, accuracy, and resource management.**

Keywords—Semantic communication, Offloading

I. Introduction

In the era of Industry 4.0, the Industrial Internet of Things (IIoT) has emerged as a transformative technology, integrating machine learning (ML) capabilities across industrial processes to enable intelligent, data-driven decisions [1]. With advancements in computation-intensive ML applications, such as real-time defect detection, predictive maintenance, and quality inspection, the demand for effective task offloading strategies has intensified. In these scenarios, ensuring rapid processing and high inference accuracy is critical, as it directly affects operational outcomes and can lead to costly disruptions if mishandled.

Traditional cloud computing solutions support these computation-heavy tasks by providing extensive resources for ML model inference. However, the inherent latency associated with cloud-based offloading, due to long data transmission distances, poses significant challenges for time-sensitive applications. As a response, edge computing has gained prominence, allowing tasks to be processed closer to data sources, thereby reducing latency and enabling more immediate responses [2]. The combination of edge, cloud, and in some cases, end-device processing, has led to the adoption of end-edge-cloud collaborative architectures, which offer a versatile solution for meeting diverse task demands in IIoT.

The unique challenges in ML-based task-oriented offloading have given rise to two primary research directions: (1) enhancing the timeliness and quality of computation-intensive tasks and (2) ensuring accuracy-aware offloading for applications with high reliability requirements. The former emphasizes reducing task latency while maintaining data quality, a vital factor in real-time processing applications. The latter, on the other hand, focuses on maintaining the inference accuracy of ML tasks across resource-constrained environments, particularly within IIoT, where data precision is essential for industrial efficiency and safety. This paper surveys recent advancements in these two areas, highlighting the evolution of task-specific, ML-based offloading strategies that address the unique needs of modern IoT and IIoT systems.

II. ML-based Task-Oriented Offloading Strategies

A. *Timeliness and Quality in Computation-Intensive Task Offloading*

The timeliness and quality of information are crucial in computation-intensive IoT and IIoT applications, where real-time processing and rapid decision-making are prioritized. Qin et al. [3] studied timeliness in task-oriented communications, particularly in status update applications. They derived expressions to quantify the timeliness of updates and developed an iterative optimization process to minimize delays, ensuring tasks are offloaded and processed within acceptable timeframes. This focus on timeliness forms the foundation for advancing real-time scheduling strategies in latency-sensitive networks.

979-8-3315-0694-0/25/$31.00 ©2025 IEEE

Fan et al. [4] extended this focus to ML-based edge-assisted systems, where they proposed a quality-aware framework for task inference. Their approach integrates joint optimization for task offloading, resource allocation, and data quality management to enhance task accuracy while maintaining low latency. By applying Lyapunov optimization, they reduced the complexity of the problem, optimizing resource allocation and achieving efficient task processing within quality and timeliness constraints.

B. *Accuracy-Aware Offloading in IIoT Applications*

Accuracy-aware offloading is essential in IIoT, where the reliability of ML-driven inference directly impacts industrial processes. Fan et al. [5] proposed an accuracy-aware task offloading and resource allocation scheme for ML-based IIoT applications. Their approach employs Lyapunov optimization to balance task accuracy with resource utilization, considering the computational needs of various ML models. This framework ensures that inference accuracy requirements are met across edge and cloud layers, which is critical in industrial settings where errors can lead to costly operational disruptions.

Khoramnejad et al. [6] introduced a stability and accuracy-focused offloading scheme in UAV-assisted smart farming, integrating multi-agent reinforcement learning to balance task accuracy and resource stability. Their system optimizes offloading decisions for UAVs, ensuring accurate data collection and analysis while managing UAV energy consumption effectively. This study exemplifies the importance of accuracy-focused offloading in IoT applications where data precision is integral to successful operations. Concurrently, Peng et al. [7] tackled multi-objective optimization within end-edge-cloud systems, focusing on balancing accuracy with energy and latency. Their NSGA-III-based framework is designed to handle diverse workflows in IIoT, where both accuracy and resource efficiency are crucial. By integrating accuracy considerations into the offloading strategy, this approach highlights the value of accuracy-aware offloading in environments with complex, computation-intensive tasks.

These two primary themes underline the evolution of ML-based, task-oriented offloading strategies, demonstrating the necessity of balancing timeliness, accuracy, and resource management to meet the demands of modern IoT and IIoT systems.

III. Lesson Learned

The study of ML-based task-oriented offloading strategies for IoT and IIoT applications highlights several critical insights:

A. *Importance of Real-Time and Accuracy Trade-Offs*

The findings underscore the necessity of balancing timeliness and accuracy for computation-intensive tasks. For applications where real-time responsiveness is paramount, strategies that prioritize low latency are critical. However, applications demanding high precision, such as quality inspection or predictive maintenance, benefit from accuracy-aware offloading frameworks.

B. *Challenges of Resource-Constrained Environments*

Many IIoT applications operate under tight resource constraints, particularly at the edge. Effective task offloading requires not only efficient resource allocation but also adaptive scheduling mechanisms that can mitigate bottlenecks and reduce energy consumption. This lesson emphasizes the importance of lightweight ML models and resource-aware algorithms to improve task processing in constrained environments.

IV. Future Directions and Open Challenges

Despite significant progress, several open challenges remain, presenting future research directions in ML-based task-oriented offloading strategies for IoT and IIoT:

A. *Model Update Data Offloading for Continuous Learning*

ML-based tasks in IoT and IIoT require continuous learning to adapt to evolving environments. To keep models up-to-date, data for model updates (e.g., new training data or model parameters) can be periodically offloaded to the edge or cloud for further learning. This approach is especially valuable in applications like predictive maintenance or quality control, where conditions change and require ongoing adjustments to the model. However, the challenge lies in managing resource consumption and latency while ensuring timely updates to maintain optimal model performance.

B. *Selective Data Offloading for Efficient Learning*

Instead of offloading all data, a selective approach can prioritize only the most valuable data for ML model training. For instance, in a defect detection system, rare instances of defects are more critical than routine data. By selectively offloading this high-value data for model updates, the system can quickly improve model accuracy while conserving resources. This approach requires lightweight methods for real-time data evaluation, enabling edge devices to prioritize data offloading based on its significance to model performance.

C. *Enhancing Reliability with Explainable Inference Offloading*

In IIoT environments, explainability is crucial for ML-based inference, especially in fields like automated inspection, where real-time decision-making is required. An effective approach is to use explainable models at the edge for initial analysis and offload the inference results to the cloud for more complex follow-up analysis if needed. For example, if suspicious data is detected in real-time, a preliminary assessment can be conducted at the edge, and the data can be further examined in the cloud to confirm the results. This layered approach enhances trust and ensures high reliability in industrial applications, particularly where safety and compliance are critical.

D. *Meta-Learning Data Offloading for Adapting to New Tasks*

In environments where tasks frequently change, meta-learning allows models to adapt quickly to new tasks. Edge devices can offload basic training data or initial model parameters to the cloud to facilitate meta-learning, enabling models to respond more rapidly to new tasks. For instance, in a manufacturing plant where product specifications and quality standards may change, meta-learning allows the model to quickly adjust to new inspection criteria. This approach supports efficient offloading and model updating, even in dynamic work environments.

These future directions emphasize the need for continuous learning, efficient data management, and enhanced reliability in ML-based task offloading. They address the unique requirements of IoT and IIoT applications, moving beyond simple inference optimization to create robust, adaptable offloading frameworks that meet the demands of various industrial settings.

V. Conclusion

This survey explored recent advancements in ML-based task-oriented offloading strategies tailored to the unique requirements of modern IoT and IIoT applications. As computation-intensive ML tasks become integral to industrial processes, efficient and reliable task offloading has proven essential to support real-time decision-making and high-precision inference outcomes. Two primary research directions—enhancing timeliness and quality for computation-intensive tasks and ensuring accuracy-aware offloading in IIoT applications—demonstrate the critical need for adaptable offloading frameworks that prioritize both latency reduction and inference accuracy.

Timeliness and quality-centered approaches address the demand for rapid processing by minimizing task delays, utilizing edge computing, and incorporating adaptive scheduling mechanisms to maintain data freshness. In contrast, accuracy-aware offloading frameworks are designed to preserve the integrity and precision of ML inference, balancing resource constraints with task-specific accuracy requirements. Both directions emphasize the importance of multi-layered (end-edge-cloud) architectures, where strategic task distribution enhances overall system efficiency.

Looking ahead, future research must focus on further refining these task-oriented offloading strategies by incorporating advanced ML techniques, such as reinforcement learning and multi-objective optimization, which dynamically adapt to changing network states and task demands. These enhancements will be key in meeting the increasingly complex requirements of IIoT systems, laying the groundwork for robust, intelligent, and resilient industrial environments.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00453301)

References

[1] T. Qiu, J. Chi, X. Zhou, Z. Ning and D. O. Wu, "Edge computing in Industrial Internet of Things: Architecture advances and challenges", IEEE Commun. Surveys Tuts., vol. 22, no. 4, pp. 2462-2488, 4th Quart. 2020.

[2] W. Zhang et al., "A survey on decision making for task migration in mobile cloud environments", Personal Ubiquitous Comput., vol. 20, no. 3, pp. 295-309, 2016.

[3] X. Qin, Y. Li, X. Song, N. Ma, C. Huang and P. Zhang, "Timeliness of Information for Computation-Intensive Status Updates in Task-Oriented Communications," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 3, pp. 623-638, March 2023, doi: 10.1109/JSAC.2022. 3229431.

[4] W. Fan, S. Li, J. Liu, Y. Su, F. Wu and Y. Liu, "Joint Task Offloading and Resource Allocation for Accuracy-Aware Machine-Learning-Based IIoT Applications," in IEEE Internet of Things Journal, vol. 10, no. 4, pp. 3305-3321, 15 Feb.15, 2023, doi: 10.1109/JIOT.2022.3181990.

[5] W. Fan, Z. Chen, Z. Hao, F. Wu and Y. Liu, "Joint Task Offloading and Resource Allocation for Quality-Aware Edge-Assisted Machine Learning Task Inference," in IEEE Transactions on Vehicular Technology, vol. 72, no. 5, pp. 6739-6752, May 2023, doi: 10.1109/TVT.2023.3235520.

[6] F. Khoramnejad, A. Syed, W. S. Kennedy and M. Erol-Kantarci, "Stability and Accuracy-Aware Learning for Task Offloading in UAV-MEC-Assisted Smart Farms," in IEEE Transactions on Network and Service Management, vol. 21, no. 5, pp. 5647-5661, Oct. 2024, doi: 10.1109/TNSM.2024.3375839.

[7] K. Peng, H. Huang, B. Zhao, A. Jolfaei, X. Xu and M. Bilal, "Intelligent Computation Offloading and Resource Allocation in IIoT With End-Edge-Cloud Computing Using NSGA-III," in IEEE Transactions on Network Science and Engineering, vol. 10, no. 5, pp. 3032-3046, 1 Sept.-Oct. 2023, doi: 10.1109/TNSE.2022.3155490.