Deep Learning-Based Offload and Load Balancing in IoT Fog

Abdellahi Krama and Sofiane Ouni

September 13, 2024

Deep learning-based offload and load balancing in IoT Fog

Abdellahi KRAMA
CRISTAL Lab
University of Manouba
Tunis, Tunisia
abdellahi.krama@etudiant-fst.utm.tn

Sofiane Ouni
CRISTAL Lab
University of Manouba
Tunis, Tunisia
sofiane.ouni@insat.rnu.tn

***Abstract*— In this paper, we addressed the problem of resource management in IoT-Fog platforms. Several approaches are proposed, including task offloading, resource allocation, etc. Our approach will be based on resource prediction. We hope that this approach will help IoT devices improve their behavior in finding available resources either locally or in the Fog tier. Collecting a dataset of the processing history of jobs in the IoT-Fog environment will allow us to predict the load level of the devices and then the resources needed to process these workloads using one of the deep learning approaches. The combination of a prediction tool with one of the already existing approaches such as Deep Reinforcement Learning will allow us to automate the management of resources in the IoT-Fog environment in a more suitable way.**

Keywords—IoT, Fog, Resources Prediction, Deep Learning

I. PROBLEM STATEMENT AND MOTIVATION

With the notable growth in the use of the Internet of Things devices in different areas of our daily lives and given the limited capacity of these devices in terms of computing [14] [15], storage and energy, it becomes essential to think about resources management in this environment. To overcome resource limitation in IoT devices, researchers introduced the notion of task offloading [1]. Several approaches were used in task offloading [2], Deep Reinforcement Learning [3] [16], game theory [4] and others. Despite all the research that has focused on task offloading, there are still some challenges facing it. The authors in [1] and [2] have cited some examples of these challenges like Distributed Deep Learning, Dynamic Quality of Service, Workload Prediction and Run-time Estimation and other challenges. I chose to focus in this thesis on the prediction of conditions and resources in order to automate resource management in IoT, fog and cloud platforms. Predicting conditions and resources can optimize task offloading at both IoT level and fog level. In fog level, if we know for example that such a fog node is usually not accessible at such a time, the available resources on this node, or its distance from IoT devices; all these information allows us to make the appropriate decision about this fog node. At the IoT level, the prediction of workload makes it much easier for us to decide whether to process the task locally or offload it to Fog or cloud. The challenges we face here are the heterogeneity of IoT devices and therefore, the difficulty of collecting a dataset from the history of workload that allows us to train models. The diversity of IoT application areas in our daily life, as well as the novelty of the topic of resource management in IoT Fog and cloud platforms, encouraged me to choose this topic in the hope of achieving a qualitative addition by designing a deep learning-based framework that predicts conditions and resources to automate resource management in IoT, fog and cloud platforms.

II. BACKGROUND AND RELATED WORK

There are several approaches to offload tasks for the Internet of Things (IoT). Some approaches of offloading are based on the game theory for intelligent resource allocation and task offline in Fog networks [5]. More recent approaches based on Deep Reinforcement Learning (DRL) are presented to optimize the offload process for IoT. These approaches often formulated the problem of optimization to reduce latency time and energy consumption [6]. A multi-objective optimization to minimize processing time and energy consumption has been proposed in [7], where the concurrency of IoT devices with fog resources was considered as a game. Authors in this designed a Distributed Task Offloading (DTO) algorithm and ε-DTO to achieve load balancing between fog nodes using the Nash equilibrium point.

A Deep Meta Reinforcement learning based Offloading (DMRO) framework that combines multiple parallel Deep Neural Networks (DNNs) and deep Q-learning algorithms to make offloading decisions was proposed in [8]. It minimizes the delay and energy consumption of IoT devices in Mobile Edge Computing (MEC) environment.

A task partition and scheduling algorithm (TPSA) was proposed in [9] to minimize the weighted sum of the overall computing service delay for vehicle users and service failure penalty in Vehicular Networks. This algorithm utilizes DRL to conduct the complex decision-making problem in a dynamic environment by re-formulating the problem into a Markov Decision Process, where states represent the network state includes the computing data amount in zones and the delay for edge servers to finish the tasks the set of system states. The action space includes three elements: the index of receiver RSU, helper RSU, and deliver RSU. The cost model represents the Q-value. Reinforcement Learning was also proposed in [10] with Particle Swarm Optimization (PSO), where PSO is used at IoT level to select optimal node to offload its workload. Whereas RL is used at the fog level to select suitable cloud. In this work, authors also used a Neuro-fuzzy model to isolate IoT nodes that try to congest the network by sending invalid data. If the IoT device is sensed to have sent invalid data for offloading, the data is dropped.

Authors in [11] proposed a middleware solution to manage the resources across the cloud, fog and edge spectrum, the purpose of which is to meet the service level objectives (SLOs) of the IoT application. SLO according to this paper refers to minimizing the deployment costs and ensure longevity of scarce edge resources, such as battery. This approach that can intelligently switch between fog and edge resources as the user moves is needed to meet the SLO, but it does not allow the task to be downloaded from fog to process on IoT (reload action).

III. APPROACH AND UNIQUENESS

In this section, I introduce our approach for resources and conditions predictions, this is a new approach that will facilitate the task offloading mission. We can achieve this goal by collecting a dataset extracted from the behavior history of an IoT Fog system. We start with the implementation of a virtual system, then we apply the results to a real system and update them if necessary.

IV. CURRENT STATUS

We implemented a DRL environment using OpenAI gym python library. This environment consists of *n* IoT devices and *m* fog nodes. The state is a vector is a vector s of size m+1 where s[0] represents the load level of IoT device n, s[i] (0<i<=m) is the load level of fog node i. compared with previous works, we add a new action that allows us to cancel an offload to a fog node and (reload) the task on IoT if it becomes less expensive. In this implementation, we use LSTM [12] to predict workload of IoT devices. We compared our approach to related works by implementing 3 others approaches as follow:

1) Local Computing Only: (LCO). In this approach there is no offload. i.e. the IoT device processes all tasks locally.

2) Deep Reinforcement Learning Offloading: (DRLO) which represents the existing Deep Reinforcement Learning approaches. used in comparison with our approach that has the newest action reload.

3) Deep Reinforcement Learning Offloading Reloading: (DRLOR), it allows us to reload the task on IoT devices if its load level is low enough.

4) Persistence Deep Reinforcement Learning Offloading Reloading: (PDRLOR). in such step we use LSTM to predict workload of IoT devices.

Figure Fig. 1. shows some results. (These results are the subject of a journal article currently being submitted)

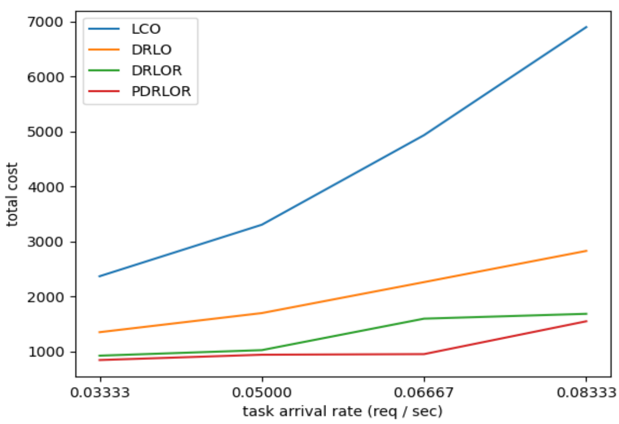

Fig. 1. Sum weighted cost variation according to λ

V. EXPECTED RESULTS AND CONTRIBUTIONS

The major contribution of this thesis is in the design of an intelligent management system for Fog and cloud resources in IoT platforms with offloading and load balancing techniques. That may fill a research gap and economize the utilization of IoT devices.

REFERENCES

[1] F. Saeik et al., “Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions”, Computer Networks 195 (2021) 108177, 18 May 2021.

[2] A. Hazraa et al., “Fog Computing for Next-Generation Internet of Things: Fundamental, State-of-the-Art and Research Challenges”, Computer Science Review · March 2023

[3] M. Tang, and V. Wong “Deep Reinforcement Learning for Task Offloading in Mobile Edge Computing Systems”, IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. PP, NO. 99, MONTH 2020

[4] S. YU-JIE et al., “Balanced Computing Offloading for Selfish IoT Devices in Fog Computing”, IEEE 10.1109/ACCESS.2022.3160198

[5] A. Mebrek and A. Yassine, "Intelligent Resource Allocation and Task Offloading Model for IoT Applications in Fog Networks: A Game-Theoretic Approach," IEEE Transactions on Emerging Topics in Computational Intelligence, doi: 10.1109/TETCI.2021.3102214.

[6] M. A. Ebrahim, G. A. Ebrahim, H. K. Mohamed and S. O. Abdellatif, "A Deep Learning Approach for Task Offloading in Multi-UAV Aided Mobile Edge Computing," IEEE Access, vol. 10, pp. 101716-101731, 2022, doi: 10.1109/ACCESS.2022.3208584

[7] G. Qu, H. Wu, R. Li and Pengfei Jiao, “DMRO: A Deep Meta Reinforcement Learning-based Task Offloading Framework for Edge-Cloud Computing”, IEEE Transactions on Network and Service Management, 2021, doi: 10.1109/TNSM.2021.3087258.

[8] C. Lin, G. Han, X. Qi, M. Guizani, “A Distributed Mobile Fog Computing Scheme for Mobile Delay-Sensitive Applications in SDN-Enabled Vehicular Networks”, IEEE Transactions on Vehicular Technology, 2020.

[9] R. Somula and Sasikala R, “A Load and Distance Aware Cloudlet Selection Strategy in Multi-Cloudlet Environment”, International Journal of Grid and High-Performance Computing, 2019

[10] M. Aazam et al., “Offloading in fog computing for IoT: Review, enabling technologies, and research opportunities” Future Generation Computer Systems, 19, 04, 2018

[11] S. Shekhar et al., “URMILA: A Performance and Mobility-Aware Fog/Edge Resource Management Middleware”, Conference: 2019 IEEE 22nd International Symposium on Real-Time Distributed Computing (ISORC), May 2019 DOI:10.1109/ISORC.2019.00033

[12] S. HOCHREITER, AND J. SCHMIDHUBER, “Long Short-Term Memory”, Neural Computation · December 1997.

[13] https://www.gymlibrary.dev/content/environment_creation

[14] Fatma Hmissi, Sofiane Ouni. An MQTT Brokers Distribution Based on Mist Computing for Real-Time IoT Communications, *Wireless Personal Communications*, 13 July 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-695717/v1]

[15] Karim Kamoun, Fatma Hmissi, Sofiane Ouni, et al., Semantic Distributed MQTT Broker for Optimized Data Flow in IoT Platforms, Transactions on Emerging Telecommunications Technologies, Wiley edition, 2024; 35(2):e4945. doi: 10.1002/ett.4945.

[16] Mohamed Ali Zormati, Hicham Lakhlef, Sofiane Ouni, Review and analysis of recent advances in intelligent network softwarization for the Internet of Things, Computer Networks (Impact Factor 5.493), Elsevier, vol. 241, 2024, p. 110215, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2024.110215.