Edge-to-Cloud Intelligence: Enhancing IoT Devices with Machine Learning and Cloud Computing

SATYANARAYAN KANUNGO
Independent Researcher, Principal Data Engineer, USA

Abstract- ***The rapid growth of the Internet of Things (IoT) has exponentially increased the number of connected devices that generate large amounts of data. Leveraging advanced technologies such as machine learning and cloud computing is key to generating meaningful insights and enabling intelligent decision-making. This article describes the concept of edge-to-cloud intelligence, which combines edge and cloud computing paradigms to improve the capabilities of IoT devices. Explore the benefits, challenges, and considerations of edge computing, machine learning, and cloud computing in the context of IoT. Additionally, we are exploring the integration of these technologies to create a smart and efficient IoT ecosystem. This paper highlights the critical role of edge computing in enabling real-time analysis and decision-making at the edge while leveraging the power of cloud resources for advanced analytics, scalability, and storage. We discuss various use cases and examples of edge-to-cloud intelligence implementation and address challenges related to scalability, security, and privacy. Finally, we consider new trends and future directions in this field. By leveraging intelligence from the edge to the cloud, IoT devices can realize their full potential, enabling innovative applications in areas such as healthcare, manufacturing, transportation, and smart cities, resulting in increased efficiency. , reliability, and user experience.***

Indexed Terms- ***Edge-to-Cloud Intelligence, Internet of Things (IoT), Machine Learning, Cloud Computing, Edge Computing, IoT.***

I. INTRODUCTION

Since its inception, cloud computing has become widespread and has greatly changed people's lifestyles. Many large companies, including Google, Amazon, and Microsoft, have launched their cloud computing services (Google Cloud, Amazon Web Services, and Microsoft Azure, respectively). Cloud computing with many remote servers can intelligently provide real-time computing, storage, and network services to users according to their requirements regarding resource types, amounts, etc. In this case, users can easily utilize these cloud services for a small fee or completely free of charge.

1. Edge Computing

The development of the Internet of Things (IoT) is driving the production and application of numerous hardware devices and sensors worldwide. These hardware devices and sensors are capable of sensing the physical environment around them and converting environmental information into data. After these large amounts of data are transferred to the cloud for computation or storage, data consumers can access the cloud data and extract the information they need according to their individual needs.

However, with the continued development and widespread application of IoT, more and more issues are emerging in cloud computing. For example, when data generated by global endpoints is computed and stored in a centralized cloud, there are problems with low throughput, high latency, bandwidth bottlenecks, data protection, centralized vulnerabilities, and additional costs, e.g. Indeed, many application scenarios in the IoT, especially in the Internet of Vehicles (IoV), require high speed and low latency for data processing, analysis, and the return of results.

To address these challenges of cloud computing, a new computing paradigm called edge computing (EC) has attracted great attention. Simply put, the core idea of the EC model is to move data processing, storage, and computing operations originally required in the cloud to the edge of the network, closer to the end devices. This helps reduce data transmission time and device response time, reduces pressure on network

bandwidth, reduces data transmission costs, and also achieves decentralization.

1.1 Why We Need Edge Computing

We will explain the importance of EC from three perspectives: the "big data era" brought about by IoT, stricter requirements for high network stability and responsiveness, and consideration for privacy and security.

1.2 The Big Data Era Caused by the Internet of Things.

Although the concept of IoT was proposed for supply chain management in 1999, it currently covers a much wider area. His integration of IoT into traditional industries has led to the emergence of many new application areas, such as smart homes, smart grids, smart traffic, and intelligent manufacturing. The idea of IoT is that things connected to the Internet form a huge network, allowing things to connect anytime and anywhere. According to the International Data Corporation (IDC), with the continued development of the Internet of Things, the number of different sensors, smartphones, health applications, and online social platforms is rapidly increasing, and the resulting global data will be It is expected to increase to 175 zettabytes (ZB) in 2019. Prediction. This vast amount of data has aided the world of big data.

In the era of big data, the most direct and easiest way to work with this data is to transfer it to the cloud for processing. As reported by Cisco in 2018, global annual cloud IP traffic was 6.0 ZB in 2016 and is expected to reach 19.5 ZB in 2021. However, the computing power of the cloud increases linearly. This is significantly slower than current data growth rates. Considering the rapid growth of data, cloud computing is no longer completely reliable.

Some IoT application scenarios require very fast response speeds. For example, in intelligent driving scenarios, sensor devices such as cameras are installed in self-driving cars. These sensor devices can continuously collect data from the environment during autonomous driving mode. In a cloud computing model, this data is uploaded to the cloud for calculations and the results are returned to the vehicle's control chip. Considering the complex driving environment of the vehicle, this method is very time consuming and may even prevent the intelligent vehicle from making appropriate decisions in time, which may have serious consequences.

1.3 More Stringent Requirements of Network Stability and Response Speed.

In the field of augmented reality (AR) and virtual reality (VR), mobile AR/VR applications require continuous transmission of high-resolution video, which places high demands on data processing power, network stability, and response speed. With today's data growth, cloud computing power is no longer able to meet these demands. However, uploading all data to the cloud will cause severe network congestion. Due to limited network bandwidth, the data generated by a large number of IoT devices places a heavy burden on the network bandwidth, making cloud computing unable to meet the latency and responsiveness requirements of these scenarios. Additionally, this data can contain large amounts of noise and errors. Some studies have shown that only one-third of the data collected by most sensors is accurate. Storing this worthless data in the cloud wastes many cloud server resources and network bandwidth.

1.4 Privacy and Security.

Cloud computing has outsourcing capabilities. Cloud computing necessitates that users host local data in the cloud. This leads to many data security and privacy issues. Long-distance transmission between devices and the cloud can cause data loss, which can compromise data integrity and accuracy. Additionally, highly centralized data processing and storage can also be a serious problem. If a central system device fails due to a benign error or malicious attack, other devices will be adversely affected. Privacy issues refer to theft or use by other unauthorized persons, companies, or organizations. Data owners lose control over the data uploaded to the cloud, making it difficult to ensure data protection.

II. THE DEFINITION OF EDGE COMPUTING

The origins of EC date back to 1999, when Akamai proposed a content delivery network (CDN) that caches web pages close to the client in order to improve web page loading efficiency. The concept of EC was borrowed from cloud computing infrastructure to extend the concept of CDN.

The EC currently has various definitions. OpenStack, for example, defines EC as a model for providing cloud and IT environment services to application developers and service providers at the edge of the network. In Reference, the authors believe that the "edge" of EC refers to all computing and network resources between the data source and the cloud. Examples: smartphones, gateways, micro data centers, and cloud networks. It's also understandable that EC is moving some cloud resources and tasks to the edge, closer to users and data sources.

It should be noted that EC cannot replace the role and benefits of cloud computing, as the computing power and storage capacity of the cloud are essential. The emergence of EC aims to compensate for cloud computing's limitations, and the relationship between EC and cloud computing should be complementary. Therefore, how to adjust the relationship between cloud and edge so that they can work together more efficiently and securely is a question that needs to be considered.

2.1 Data Generated at the Network Edge Need AI to Fully Unlock Their Potential:

As the number and variety of mobile and IoT devices are rapidly increasing, large amounts of multimodal data (audio, images, video, etc.) from the physical environment are continuously collected on the device side. In this context, there is a functional need for AI that can quickly analyze these vast amounts of data and derive insights from it for quality decision-making. Deep learning, one of the most popular AI technologies, automatically recognizes patterns in data collected by edge devices, such as population distribution, traffic flow, humidity, temperature, pressure, and air quality, and identifies anomalies. Provides functionality to detect. Insights from the collected data feed into real-time predictive decision-making (e.g., public transportation planning, traffic control, driver warnings, etc.) in response to rapidly changing environments, thereby increasing operational efficiency. Improve. As predicted by Gartner, by 2022, more than 80% of enterprises' IoT projects will include an AI component (currently only 10%).

2.2 Edge Computing Can Prosper AI With Richer Data and Application Scenarios:

It is widely accepted that the recent deep learning boom is driven by four factors: algorithms, hardware, data, and application scenarios. While the influence of algorithms and hardware in deep learning development is intuitive, the role of data and application scenarios has been largely overlooked. To improve the performance of deep learning algorithms, the most commonly used approach is to improve the DNN with more neuron layers. In this way, the DNN needs to learn more parameters, which also increases the data required for training. This demonstrates the importance of data in AI development. Once you realize the importance of data, the next question is: where does it come from? Traditionally, data was typically generated and stored in mega data centers. However, with the rapid development of the IoT, that trend is now reversing. According to a Cisco report, a huge amount of IoT data will be generated at the edge shortly. When this data is processed by AI algorithms in cloud data centers, it consumes large amounts of bandwidth resources and places a significant load on cloud data centers. To address these challenges, edge computing has been proposed to achieve low-latency data processing by transferring computing power from cloud data centers to the edge, or edge side. H. High-performance AI processing becomes possible.

Edge computing and AI complement each other from a technological perspective, but their application and proliferation are also mutually beneficial.

2.3 AI Democratization Requires Edge Computing as a Key Infrastructure:

AI technology has achieved great success in many digital products and services in daily life, including B. Online shopping, service recommendations, video monitoring, smart home devices, etc. AI is also a key driver behind innovations such as self-driving cars, smart finance, cancer diagnostics, and drug discovery. Beyond the examples above, leading IT companies are using AI democratization or declaring ubiquitous AI. To achieve this, AI needs to get "closer" to people, data, and devices. Edge computing is superior to cloud computing in achieving this goal. First, edge servers are located closer to people, data sources, and devices compared to cloud data centers. Second, edge computing is also more affordable and accessible

compared to cloud computing. Finally, edge computing has the potential to offer a wider variety of AI application scenarios than cloud computing. These benefits inevitably make edge computing a key enabler of ubiquitous AI.

2.4 Edge Computing Can Be Popularized With AI Applications:

In the early days of edge computing's development, the cloud computing community was wondering what high-demand applications could take edge computing to the next level that cloud computing cannot reach and what edge computing's killer applications are. There was always concern about what it was. To address these questions, since 2009, Microsoft has focused on what types of things should be moved from the cloud to the edge, from voice command recognition, AR/VR, and interactive cloud gaming to real-time video analytics. We have continued to consider the importance of safety. In comparison, real-time video analytics is considered a killer application for edge computing, and. Real-time video analytics, a new application built on computer vision, continuously acquires high-resolution video from surveillance cameras and uses high computing, high bandwidth, advanced data protection, and low latency. Edge computing is the only viable approach to meeting these stringent requirements. Looking back at the developments in edge computing mentioned above, we can predict that new AI applications in areas such as IoT, intelligent robots, smart cities, and smart homes will play a key role in the proliferation of edge computing. This is primarily because many mobiles and IoT-related AI applications are a family of practical applications that are compute-intensive, energy-intensive, privacy- and latency-sensitive, and therefore naturally mesh well with edge computing. This is to be done.

Edge AI has received a lot of attention recently due to the advantages and necessity of running AI applications at the edge. In December 2017, the white paper “A Berkeley View of Systems Challenges for AI” published by the University of California, Berkeley, presented cloud-edge AI systems as an important research direction to achieve mission goals. Ta: critical and personalized AI. In August 2018, Edge AI first appeared on his Gartner Hype Cycle. Gartner predicts that edge AI is still in the innovation trigger stage and will reach a productivity plateau in the next five to 10 years. The industry is also conducting numerous pilot projects aimed at cutting-edge AI. Specifically, traditional cloud providers such as Google, Amazon, and Microsoft are working to bring intelligence to the edge by allowing end devices to perform ML inference locally using pre-trained models. We introduced a service platform on the Edge AI service platform. For edge AI chips, there are a variety of high-end chips on the market that power ML models, including Google Edge TPU, Intel Nervana NNP, Huawei Ascend 910, and Ascend 310.

III. SCOPE AND RATING OF EDGE INTELLIGENCE

Although the term edge AI or EI is quite new, exploration and practice in this direction started early. As mentioned earlier, in 2009, Microsoft developed an edge-based prototype to support mobile voice command recognition, an AI application, to demonstrate the benefits of edge computing. Although research is still in its early stages, there is still no formal definition of EI.

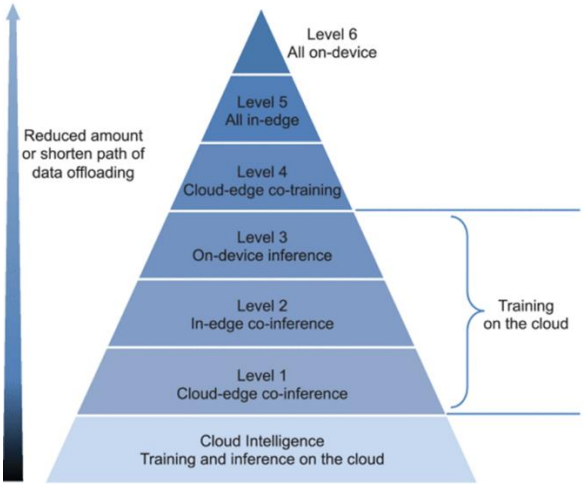
Currently, most organizations and news organizations are using the paradigm of running AI algorithms locally on the device using data created on the device (sensor data or signals). EI is referred to as although this currently represents the most common approach to EI in the real world (e.g. high-end AI chips), it is important to note that this definition significantly limits the solution range of his EI. Local execution of computationally intensive algorithms, such as DNN models, is extremely resource-intensive and requires high-end processors on the device. Such strict requirements not only increase the cost of EI but are also incompatible and unfriendly with existing legacy devices with limited computing power.

In this paper, we argue that the scope of EI should not be limited to running AI models only on edge servers or devices. As more than a dozen recent studies have shown, leveraging edge-cloud synergies to run his DNN models reduces both end-to-end latency and energy consumption compared to local execution approaches. It can be reduced. Because of these practical benefits, we believe that such collaborative hierarchies should be integrated into the design of

efficient EI solutions. Furthermore, existing considerations on EI mainly focus on the inference phase (i.e., AI model execution) and assume that the AI model training is performed on a power cloud data center. Among the training phases, the inference phase dominates. However, this means that large amounts of training data must be transferred from the device or edge to the cloud, leading to prohibitive communication overhead and privacy concerns.

Instead, EI takes full advantage of available data and resources within a hierarchy of end devices, edge nodes, and cloud data centers to optimize the overall performance of DNN model training and inference. I believe that the paradigm should be this indicates that while EI does not necessarily mean that the DNN model is fully trained or derived at the edge, it can operate in a coordinated manner on cloud edge devices through data offloading. Specifically, we classify EI into six levels according to the amount of data offload and path length, as shown in the figure. In detail, the definition of different EI levels is given as follows:

1. Cloud Intelligence: DNN model training and inference occur entirely in the cloud.
2. Level 1 - Cloud-Edge Collaborative Conference and Cloud Training: Train the DNN model in the cloud, but infer the DNN model in an edge-cloud collaborative manner. Edge-cloud collaboration means that some data is outsourced to the cloud.
3. Level 2 - In-edge conference and cloud training: Train the DNN model in the cloud, but the DNN model infers in an in-edge manner. In-edge here means that model inference is performed within the network edge. This can be achieved by completely or partially offloading data (via D2D communication) to edge nodes or nearby devices.
4. Level 3 - On-device inference and cloud training: Train the DNN model in the cloud, but the DNN model is fully inferred locally on the device. "On-device" here means the data is not outsourced.
5. Level 4 - Cloud Edge Collaborative Training and Inference: Training and inference of DNN models in both edge and cloud collaboration.
6. Level 5 – All in-edge: DNN model training and inference, both in-edge style.
7. Level 6 – All on the device: Both training and inference of the DNN model is done on the device.



CONCLUSION

The convergence of edge computing, machine learning, and cloud computing has paved the way for edge-to-cloud intelligence, revolutionizing the capabilities and possibilities of IoT devices. This integration enables IoT devices to collect, process, and analyze data in real-time at the edge while leveraging the power of cloud resources for advanced analytics and scalability.

The benefits of edge-to-cloud intelligence are numerous. Processing critical data locally enables faster decision-making, reduces latency, and increases reliability. Additionally, it optimizes resource utilization and increases efficiency by offloading non-time-critical tasks to the cloud. Additionally, the combination of machine learning algorithms and cloud computing enables IoT devices to extract valuable insights from large amounts of data, enabling predictive and prescriptive analytics to drive intelligent automation and informed decision-making. It will be possible.

However, deploying edge-to-cloud intelligence to IoT devices is not without its challenges. Scalability and resource limitations, security vulnerabilities, and privacy issues at the edge require careful consideration and robust solutions. As technology advances, it becomes increasingly important to address these challenges to ensure widespread adoption and success of edge-to-cloud intelligence in the IoT ecosystem.

Looking ahead, the future of edge-to-cloud intelligence is bright. New technologies such as 5G networks, edge AI chips, and federated learning will further advance the capabilities of IoT devices and enable more sophisticated and autonomous edge computing. Potential applications span fields as diverse as smart cities, healthcare, manufacturing, and transportation, where edge-to-cloud intelligence can drive transformative change and improve efficiency, reliability, and user experience. Masu.

In summary, edge-to-cloud intelligence represents a powerful paradigm for improving his IoT devices through machine learning and cloud computing. By leveraging the strengths of edge computing and cloud resources, IoT devices can realize their full potential, enable innovative applications, and drive innovation. As researchers, industry experts, and policymakers continue to explore and address challenges, edge-to-cloud intelligence will shape the future IoT landscape and usher in a new era of smart, connected devices. There is no doubt about it.

APPENDIX

Edge Computing:
Edge computing refers to the practice of processing data near the source or "edge" of a network, that is, near where the data is generated. This approach reduces latency, improves privacy and security, and minimizes the need for extensive bandwidth. Through data processing and analysis at the edge, IoT devices can effectively filter and aggregate data before forwarding it to the cloud for further analysis and storage.

Machine learning at the edge:
Deploying machine learning capabilities to edge devices enables real-time decision-making and predictive analytics without relying solely on the cloud. This is especially useful in scenarios where low latency is important or where connectivity to the cloud can be intermittent. Machine learning models deployed at the edge can perform tasks such as anomaly detection, predictive maintenance, and intelligent data filtering.

Cloud Computing:
Cloud computing provides scalable, on-demand access to a shared pool of computing resources. By leveraging the cloud, IoT devices can offload compute-intensive tasks, store and access large amounts of data, and benefit from advanced analytics and machine learning algorithms. Cloud platforms also enable centralized management and monitoring of IoT deployments.

Edge-to-cloud architecture:
edge-to-cloud architecture integrates edge and cloud computing to deliver a comprehensive IoT solution. In this architecture, edge devices collect and process data locally using machine learning capabilities at the edge. Processed data is transferred to the cloud for further analysis, long-term storage, and integration with other data sources. The cloud also provides backend services for device management, over-the-air updates, and remote monitoring.

ACKNOWLEDGMENT

I am deeply grateful to my writer, Johnson Dare, for their guidance, expertise, and unwavering support throughout the entire research process. Their insightful feedback and valuable suggestions have significantly shaped the direction and quality of this work.

REFERENCES

[1] Khan, A. U. R., Othman, M., Madani, S. A., & Khan, S. U. (2014). A survey of mobile cloud computing application models. IEEE Communications Surveys & Tutorials, 16(1), 393–413.

[2] Durao, F., Carvalho, F., Fonseka, A., & Garcia, V. C. (2014). A systematic review on cloud computing. The Journal of Supercomputing, 68(3), 1321–1346.

[3] Shi, W., & Dustdar, S. (2016). The promise of edge computing. Computer, 49(5), 78–81.

[4] Qin, M., Chen, L., Zhao, N., Chen, Y., Yu, F. R., & Wei, G. (2018). Power-constrained edge computing with maximum processing capacity for IoT networks. IEEE Internet of Things Journal, 6(3), 4330–4343.

[5] Ghosh, A. M., & Grolinger, K. (2021). Edge-cloud computing for internet of things data analytics: Embedding intelligence in the edge with deep learning. IEEE Transactions on Industrial Informatics, 17(3), 2191–2200.

[6] Zhou, P., Chen, W., Ji, S., Jiang, H., Yu, L., & Wu, D. (2019). Privacy-preserving online task allocation in edge-computing-enabled massive crowdsensing. IEEE Internet of Things Journal, 6, 7773–7787.

[7] Gaura, E. I., et al. (2013). Edge mining the internet of things. IEEE Sensors Journal, 13(10), 3816–3825.

[8] Xu, Z., et al. (2020). Artificial intelligence for securing IoT services in edge computing: A survey. Security and Communication Networks, 2020, 1–13.

[9] Savaglio, C., & Fortino, G. (n.d.). A simulation-driven methodology for IoT data mining based on edge computing. ACM Transactions on Internet Technology, 1–22.

[10] Lei, L., Xu, H., Xiong, X., Zheng, K., Xiang, W., & Wang, X. (2019). Multiuser resource control with deep reinforcement learning in IoT edge computing. IEEE Internet of Things Journal, 6(6), 10119–10133.

[11] Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762.

[12] Miranda, M., Cristina, C., & Sebastiá, S. (2020). Deep learning at the mobile edge: Opportunities for 5G networks. Applied Sciences, 10(14), 4735.

[13] Wang, F., Zhang, M., Wang, X., Ma, X., & Liu, J. (n.d.). Deep learning for edge computing applications: A state-of-the-art survey. IEEE Access, 58322–58336.

[14] Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674.

[15] Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(2), 869–904.

[16] Shi, Y., Yang, K., Jiang, T., Zhang, J., & Letaief, K. B. (2019). Communication-efficient edge AI: Algorithms and systems. IEEE Communications Surveys & Tutorials, 21(2), 2167–2191.

[17] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645–1660.

[18] Reinsel, D., Gantz, J., & Rydning, J. (2018). Data age 2025: The digitization of the world: From edge to core. IDC White Paper, # US44413318.

[19] Marjani, M., Nasaruddin, F., Gani, A., Abaker, I., Hashem, T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: Architecture, opportunities, and open research challenges. IEEE Access, 5, 5247–5261.

[20] Cisco Global Cloud Index. (2018). Forecast and Methodology, 2016–2021 White Paper. Updated: February 1, 2018.

[21] Zhang, J., Chen, B., Zhao, Y., Cheng, X., & Hu, F. (2018). Data security and privacy-preserving in edge computing paradigm: Survey and open issues. IEEE Access, 6, 18209–18237.

[22] Sukhmani, S., Sadeghi, M., Erol-Kantarci, M., & Saddik, A. E. (2019). Edge caching and computing in 5G for mobile AR/VR and tactile internet. IEEE MultiMedia, 26(1), 21–30.

[23] Cai, H., Xu, B., Jiang, L., & Vasilakos, A. V. (2017). IoT-based big data storage systems in cloud computing: Perspectives and challenges. IEEE Internet of Things Journal, 4(1), 75–87.

[24] Mollah, M. B., Azad, M. A. K., & Vasilakos, A. (2017). Security and privacy challenges in mobile cloud computing: Survey and way ahead. Journal of Network and Computer Applications, 84, 38–54.

[25] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

[26] Whitaker, B. E. (2019). Cloud edge computing: Beyond the data center. Retrieved from OpenStack website.

[27] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646.

[28] Yang, R., Yu, F., Si, P., Yang, Z., & Zhang, Y. (2019). Integrated blockchain and edge computing systems: A survey, some research issues and challenges. IEEE Communications Surveys & Tutorials, 21(2), 1508–1532.

[29] Huang, L., Feng, X., Feng, A., Huang, Y., & Qian, L. (2018). Distributed deep learning-based offloading for mobile edge computing networks. Mobile Networks and Applications, 1–1. DOI:10.1007/s11036-018-1177-x.

[30] Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the Internet of Things. In Proceedings of the first edition of the MCC workshop on Mobile cloud computing (pp. 13-16). ACM.

[31] Li, P., Mao, Y., Chen, W., Zhang, Y., & Hwang, K. (2018). Edge computing for the Internet of Things: A case study. IEEE Internet of Things Journal, 6(1), 161–172.

[32] Mukherjee, M., Magno, M., & Alioto, M. (2019). Machine learning at the edge for intelligent IoT: A review. IEEE Internet of Things Journal, 7(7), 5572–5585.

[33] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.