Rethinking Edge AI Architecture

Elizabeth Nathania Witanto, Yustus Eko Oktian, Sang-Gon Lee*

Dongseo University
*Corresponding Author

Abstract

The proliferation of 5G connections and Internet-of-Things (IoT) drives to the data explosion generated by the massive number of IoT devices (e.g., sensors, cameras, etc.) and end-devices (e.g., smartphones, tablets, etc.). Rapidly increasing data volume brings more advantages and challenges to Artificial Intelligence (AI) development. Data is the heart of AI. The conventional way to process generated data is to transfer it over the internet to the data center. Sending bulks of data from the IoT devices to the cloud data center causes high financial cost, transmission delay, and privacy leakage. It is about time that the technology trend is shifting to so-called Edge AI. In this paper, we explain and evaluate about three possible types of architecture for the Edge AI training process. There are centralized, decentralized, and distributed. In addition, we present the pros and cons of each architecture.

I. Introduction

We are living in an era of rapid technological and communication development. Most recently, the proliferation of 5G connections and Internet-of-Things (IoT) drives to the data explosion generated by the massive number of IoT devices (e.g., sensors, cameras, etc.) and end devices (e.g., smartphones, tablets, etc.). Research from International Data Corporation (IDC) shows that the amount of data generated by connected IoT devices, forecast to grow to 41.6 billion by 2025, is expected to generate 79.4 zettabytes (ZB) of data [1]. This amount of data brings more advantages and challenges to Artificial Intelligence (AI) development. Data is the heart of AI. More training data will increase the accuracy results of the AI algorithm.

The conventional way to process generated data is to transfer it over the internet to the data center. However, there is a problem due to the concerns of performance, cost, and privacy. Send bulks of data from the IoT devices to the cloud data center is highly non-trivial even with fast connections. The financial cost and transmission delay can be prohibitively high, and privacy leakage is a crucial concern.

It is about time that the technology trend is shifting to so-called Edge AI. Edge AI is a system that uses machine-learning algorithms to process data generated by a hardware device at the local level [2]. Thus, it will conduct the process closer to the IoT devices and data sources. Since the data does not need to be transfered through the internet, the device can make a real-time decision in a matter of milliseconds. According to Vector ITC Group, the latency of cloud computing would be seconds; with Edge AI, the times are reduced to less than 400ms [2]. The combination of edge computing and AI comes with several benefits compared to traditional cloud computing, including low-latency, less bandwidth consumption, privacy protection,

scalability, and adaptability [3].

In this paper, we explain and evaluate about three possible types of architecture for the Edge AI training process. There are centralized, decentralized, and distributed. Besides, we present the pros and cons of each architecture.

II. Edge AI Architecture

The workflow in the AI environment is divided into two parts, training, and inference. In the training process, the algorithm gives and calculates the value of weights of the model. The output is the task result, and there is a loss function to evaluate the correctness of the result by calculating the error rate. The inference process happens after training. It tests the model by giving the input and shows the predictions. In the scope of our paper, we show possible types of architecture for Edge AI training only.

It is crucial to design and choose the suitable architecture to gain optimized advantages from Edge AI. Based on references [4] and [5], we conclude three types of architecture, centralized, decentralized, and distributed. We evaluate and present the pros and cons of each type of architecture. It is worth noting that there is some value to trade-off for each architecture. Figure 1 shows three types of architecture that we describe as follows:

1. *Centralized*: training of the model is happening in the cloud while edge devices such as surveillance cameras, traffic lights, smart watches, and smart phones send the training data to the cloud.

Pros:

Low-computation latency. The central server has enough resources (memory, storage, energy) to compute the training algorithm. Therefore, it will reduce the computation-latency compare to constraint device with constraint resources.

High-accuracy. The computation that happens in a central server with enough resources (memory, storage, power) can train more data. More data will increases the accuracy of the results.

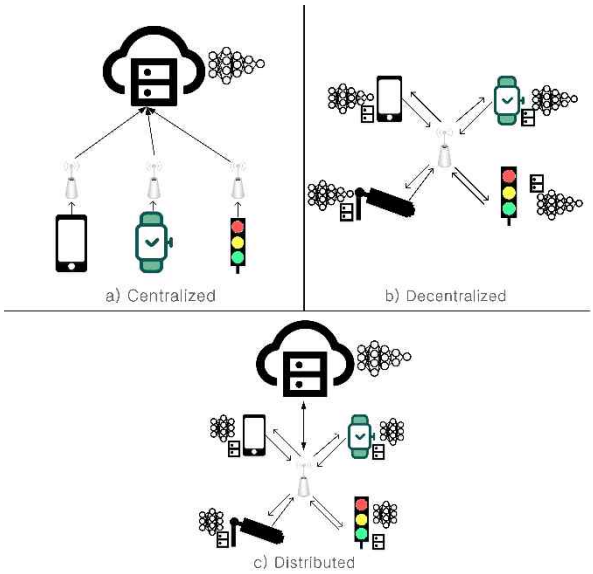

Figure 1. Edge AI Training Architecture.

Cons:

High-communication latency and cost. To send bulks of data from edge-devices to the central server will take time, not to mention the data's size. As a result, it will increase the communication latency alongside the cost. Another consequence is increased energy consumption.

Data privacy problem. To train the data in the central server, the data need to be transfered across the network. It is unavoidable the privacy issues.

2. *Decentralized:* The model's training is happening directly in each edge-device locally without sending data to the cloud. Between

edge-devices will communicate to exchange their local model.

Pros:

Since the computation happens locally, *the communication latency and cost will be reduced.* It also *decreases energy consumption.* However, it depends on the specification of each edge device. Another advantage is decentralized type will preserve data privacy.

Cons:

High-computation latency. The edge-devices are not designed with many resources (e.g., memory, storage, power). Due to these limited resources, it takes more time until the training result converged.

Less accuracy. The resource's limitation also limits the amount of training data. The edge-devices can not train as much as the central server can. Therefore, it affects the accuracy of the training results.

Data redundancy. The training data gathered by edge-devices might be the same or partially the same. It leads to data redundancy. The further effect is the waste of computation power. The devices might train the same data repeatedly.

Distributed: each edge-device trains the model locally, and periodically the central server will aggregate the local update from each device.

Pros:

Since the training happens locally as a decentralized type, distributed type inherits the advantages from it, such as less-energy consumption, less-communication latency, and cost. Besides, the accuracy will be higher since there is a central server that aggregates the local model update from edge-devices in the network.

Cons:

The distributed type also inherits the cons from decentralized type such as data redundancy and high-computation latency. For the data redundancy problem, reference [6] suggested a solution called edge-caching. The training data will be stored in the cache. Later, when the edge-device captures the same data as that in the cache, it will not store the data to avoid redundancy and save computation power.

III. Future Research Directions

As Edge AI proliferating, it comes with challenges and future research directions for fellow developers and researchers. We describe as follows:

- **Training data curation problem and reliability**. In Edge AI, we get the training data directly from many edge-devices distributed across the edge-network. It is different from the cloud-based AI technique that uses an available dataset that is already being curated and labeled. Therefore, it raises the problem of reliability of the training data. Besides, the devices will have a different environment and high device heterogeneity.
- **Training data completeness.** The training data distributed across the edge-devices. In some cases, such as distributed architecture, the data will be sent to the central server. There is a chance that there are stagger devices and some data not arrived in the central server. Therefore, it needed to do further research to ensure all the training data from edge-devices were not missing.

IV. Conclusions

Architecture is the foundation of a system. It is vital to design and choose suitable architecture according to each system's needs for gaining maximized advantages from technology. In this paper, we evaluate three types of architecture, centralized, decentralized, and distributed. Besides, we present the pros and cons of each architecture. In the last section, we present Edge AI's future research directions that will be the next mission to accomplish for developers and researchers.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant Number: 2018R1D1A1B07047601).

[References]

[1] Eden Estopace, "IDC forecasts connected IoT devices to generate 79.4ZB of data in 2025," Jun. 22, 2019. https://futureiot.tech/idc-forecasts-connected-iot-devices-to-generate-79-4zb-of-data-in-2025/ (accessed Feb. 01, 2021).

[2] Vector ITC, "Edge AI: The Future of Artificial Intelligence," Aug. 12, 2020. https://www.vectoritcgroup.com/en/tech-magazine-en/artificial-intelligence-en/edge-ai-el-futuro-de-la-inteligencia-artificial/ (accessed Feb. 01, 2021).

[3] Maximilian Bischoff, Johannes M. Scheuermann, Christoph Kiesl, and Julian Hatzky, "The Edge is Near: An Introduction to Edge Computing!," Jun. 03, 2019. https://www.inovex.de/blog/edge-computing-introduction/ (accessed Feb. 01, 2021).

[4] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," Proc. IEEE, vol. 107, no. 8, pp. 1738 - 1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.

[5] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-Efficient Edge AI: Algorithms and Systems," IEEE Commun. Surv. Tutorials, vol. 22, no. 4, pp. 2167 - 2191, 2020, doi: 10.1109/COMST.2020.3007787.

[6] D. Xu, T. Li, Y. Li, T. Jiang, J. Crowcroft, and P. Hui, "Edge Intelligence: Architectures, Challenges, and Applications," arXiv:2003.12172, p. 53.