A survey of Deep learning at the Edge computing

SUBMISSION DATE / POSTED DATE

22-04-2022 / 27-04-2022

CITATION

Mayumu, Nicanor (2022): A survey of Deep learning at the Edge computing. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.19640265.v1

DOI

10.36227/techrxiv.19640265.v1

A survey of Deep learning at the Edge computing

Nicanor Mayumu
School of Computer Science and Engineering
Central South University
932 Lushan S Rd, Yuelu District, Changsha, Hunan
Email: 214708038@csu.edu.cn

Abstract—**Deep learning is revolutionizing machine learning application domains such as big data analysis, computer vision, and natural language processes. Deep learning is a subset of machine learning that make use of Neural Network(NN) with many processing layers and parameters. This consumes a lot of computing power and memory during the training and testing phases. A common approach to overcoming this computational limitation of deep learning is to use cloud resources. To do this, data to be processed need to be transferred from the source to the cloud. Cloud-based big data processing presents many challenges as the amount of data continues to grow and the data processing requirements differ. Edge computing has emerged to solve these challenges. This article explores how deep learning is used in edge computing and discusses some of the common challenges that have arisen so far.**

1. Introduction

Deep learning is revolutionizing machine learning applications domains such as Big data analysis, computer vision and natural language process. For example in agriculture, convolutional neural network were used for detecting disease in plant [1]. The complexity in deep architectures leads to a high resources consumption in computational power and memory for training and testing phase.

To overcome this constraint in computational power of deep learning, one way is to process algorithm and data on the cloud resources. For that, the data to be processed must be transferred from its sources to a the cloud. However, cloud-based solution has many challenges:

Latency: Some real-time applications such as computer vision in autonomous vehicle, traffic, monitoring, etc. need a real-time inference. For example, in self-driving car, camera frames require real-time feedback to detect and avoid obstacles [2]. However, transmission of large amount of data from different sources (IOT objects) to the cloud for processing may result in queuing and large load of network transmission bandwidth which demonstrated the cloud-based solution limitation.

Scalability: Sending large amounts of data to the cloud requires cloud flexibility. To be able to increase or decrease functionality by processing resource requests taking workload capacity into account while maintaining performance. For example, not all data sent to the cloud can help you draw conclusions using deep learning.

Security and privacy: Some private data are sent to the cloud, such as users face, voice, location. The big issues is who is using their data, for what purpose. So, security and privacy is one of most critical functionality of cloud computing. For instance, the user location can be sending to the cloud in real-time using IOT objects, this can be a high risk when terrorist capture this data.

With increasing data volumes and computing power, edge computing is the best solution to the challenges of cloud computing in terms of latency, scalability, security, and data protection.

Edge computing technology provides artificial intelligence services to rapidly growing end devices and data, making them more stable. Edge computing is closer to the source of the data, store, secure and process data at the edge devices for real-time inference.

In this paper, we present some application of deep learning at the edge. We will investigate the intersection of deep learning [3] and edge computing [2] from the architecture perspective to the challenges. We selected five papers among twenty best one to investigate our purpose. In the following section, we present the applications of deep learning at the edge and then give a survey of five selected articles.

2. Deep learning at the edge

Deep in deep learning comes from the number of hidden layers in the architecture (Fig 1). While extra layers allow a deep neural network to reach best performance, it come up with some limitation:

Data: while efficient to deal with unstructured data, deep learning architectures required a huge amount of data to be trained for better performance.

Computational power: Even with all necessary data, training deep neural network requires GPUs which contain a large numbers of cores compared to the CPUs. Not all machines meet the computational requirements for deep learning.

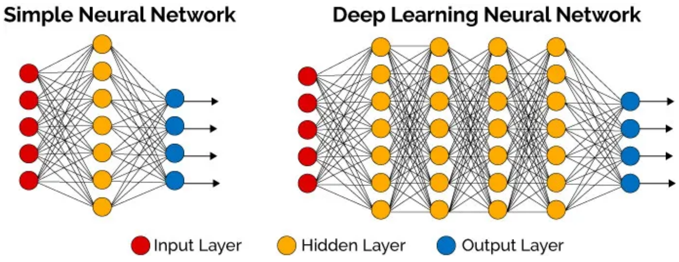

Figure 1. From neural network to deep neural network.
Source: "https://thedatascientist.com/what-deep-learning-is-and-isnt/"

Training time: Training deep neural network take more than hour. Training time increase with the amount of data and parameters to be train.

To address these challenges on edges devices, several researches have been made. For example, Huyn et al proposed a Mobile GPU-based Deep Learning framework for addressing the computational power challenge [4]. Sujith Ravi (Google, 2018) proposed Learn2Compress technology developed for the purpose of reducing model size without loosing accuracy. In definitive, it is possible to run deep learning on edge devices despite computational challenges. In the following sections, we will present some works that applied deep learning on edge devices.

2.1. Computer vision

Computer vision is a sub-field AI that allows computers to derive insights from images and videos inputs. Several researches is going on computer vision, from object detection to object recognition applied in real industrial cases such in autonomous vehicles, transportation, mining industry, healthcare, etc. Deep learning techniques has improved the computer vision accuracy for tasks such in image, videos detection or recognition.

Tan Zhang et al designed and implemented a wireless camera architectures that process video analysis at the edge computing nodes (ECN) while the relevant part of the video feed is uploaded to the controller on the internet [5]. This method allow author's to reduce bandwidth consumption while providing accurate real-time feedback.

Aishwarya D and Minu R.I deployed a convolutional neural network (CNN) on a Raspberry device to classify and identify various actions such as pulling, pushing, and other hand movements in real-time. [6]

Guanxiong et al. designed a Smart Traffic Monitoring System, an hybrid edge-cloud based system for monitoring traffic using convolutional neural network techniques.

2.2. Virtual reality

Virtual reality (VR) refers to a digital representation of real world or experience. Currently, VR is applied on video gaming, training and business. Unfortunately, several applications of VR are commercial.

Smit et al. described a qualitative study of children shopping in a virtual reality supermarket. The introduction provides a firm justification for a dual focus on environmental and health behaviors through a focus on food-consuming behaviors prediction [12].

2.3. Internet of Things

IoT networks are systems of related heterogeneous devices such as vehicles and homes appliances. These connected objects equipped with communication and data transfer functions improve quality of life by [9]. For some reason discussed in introduction, data generated from IOT Objects such WIFI, RFID, Camera, wearable devices, etc. have to be processed at the edge instead of sending to the data center of cloud computing for performing process.

Azimi et al. designed the hierarchical computer architecture HiCH for real-time health monitoring. These edge computing-based architectures focus on detecting cardiac arrhythmias in patients with cardiovascular disease (CVD) using deep neural network approaches [10]. Sufian et al., proposed a learning-based edge computing approach to home care monitoring. In particular, models based on pretrained convolutional neural network models was integrated to leverage edge devices with small amounts of ground data and fine-tuning methods to train the model [11].

3. Challenges

Although we have shown that it is possible to deploy deep learning on edge devices, there are several challenges that remain to be investigated such as:

- Service integration : The fact that edge computing uses multiple platforms such as edge servers, network protocol, end devices and having different architectures makes edge computing difficult to allocate resources from a development point of view which makes integration difficult.
- Energy consumption: Running deep learning on edge devices is energy consuming yet most end devices like our smartphones have a limited capacity in energy. Hence it is important to take this factor into account when compressing a model.
- Storage: Ends device and edge devices have very limited capacity, yet some applications such as video analysis require a large storage capacity. Hence the importance of combining edge computing with the cloud that can take care of storage and in-depth analysis. But the challenge remains what portion of data will have to keep and analyze in the edge and which portion to send to the cloud.

4. Conclusion

In this survey, we examined the current state of deep learning at the edge computing devices. Computer vision, virtual reality and internet of things were discussed as exemplary application drivers. Many unresolved issues remain such improved performance, resource allocation, energy consumption. It is important to address these challenges for a good performance of our models and a long life of our end devices and edge devices.

References

[1] Konstantinos P. Ferentinos, *Deep learning models for plant disease detection and diagnosis, Computers and Electronics in Agriculture*, Volume 145, 2018, Pages 311-318, ISSN 0168-1699.

[2] K. Cao, Y. Liu, G. Meng and Q. Sun, *"An Overview on Edge Computing Research,"* in IEEE Access, vol. 8, pp. 85714-85728, 2020, doi: 10.1109/ACCESS.2020.2991734.

[3] Ian Goodfellow, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1. MIT press Cambridge, 2016.

[4] Loc N Huynh, Youngki Lee, and Rajesh Krishna Balan. Deepmon: Mobile gpu-based deep learning framework for continuous vision applications. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, pages 82–95. ACM, 2017.

[5] Tan Zhang, Aakanksha Chowdhery, Paramvir Victor Bahl, Kyle Jamieson, and Suman Banerjee. The design and implementation of a wireless video surveillance system. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, pages 426–438. ACM, 2015.

[6] Aishwarya D., Minu R.I., Edge computing based surveillance framework for real time activity recognition, ICT Express, Volume 7, Issue 2, 2021, Pages 182-186, ISSN 2405-9595.

[7] "Get Ready to Hear a Lot More About 'XR'". Wired. 1 May 2019. ISSN 1059-1028. Retrieved 29 August 2020.

[8] Nathan Matsuda, Brian Wheelwright, Joel Hegland, and Douglas Lanman. 2021. VR Social Copresence with Light Field Displays. ACM Trans. Graph. 40, 6, Article 1 (December 2021), 13 pages. https://doi.org/10.1145/3478513. 3480481.

[9] Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of things: A survey on enabling technologies, protocols, and applications. IEEE Commun. Surv. Tutor. 2015, 17, 2347–2376.

[10] Azimi, I.; Anzanpour, A.; Rahmani, A.M.; Pahikkala, T.; Levorato, M.; Liljeberg, P.; Dutt, N. Hich: Hierarchical fog-assisted computing architecture for healthcare iot. ACM Trans. Embed. Comput. Syst. (TECS) 2017, 16, 174.

[11] Sufian, Abu and You, Changsheng and Dong, Mianxiong: A Deep Transfer Learning-based Edge Computing Method for Home Health Monitoring, 2021 55th Annual Conference on Information Sciences and Systems (CISS), 2021, 10.1109/ciss50987.2021.9400321.

[12] Smit, E.S.; Meijers, M.H.C.; van der Laan, L.N. Using Virtual Reality to Stimulate Healthy and Environmentally Friendly Food Consumption among Children: An Interview Study. Int. J. Environ. Res. Public Health 2021, 18, 1088. https://doi.org/10.3390/ijerph18031088