Research and Application of Edge Computing Based on Deep Learning

Wei Cui
Department of electronic engineering, Chongqing Aerospace Polytechnic, Chongqing, 400021, China
Email: tracy_cw@163.com

Abstract. With the exponential growth of IoT terminals, smartphones, and wearable devices, traditional centralized cloud computing models have been unable to efficiently process data generated by edge devices. In order to meet the challenges, edge computing has been proposed. The location close to the terminal device meets the high computing volume, low latency, privacy requirements of deep learning on edge devices, bandwidth advantages, efficiency and scalability. We first introduced the background and motivation for running AI at the edge of the network, reviewed the basic concepts of deep learning, and then we provided the overall architecture of edge computing based on deep learning. We discussed three computing and inference models on terminal devices, edge servers and cross-edge devices, and describes the method to improve and optimize the edge deep learning model. Finally, we discuss the application scenarios and future opportunities of edge deep learning.

1. Introduction
Traditional cloud computing enables companies to store and process data and other computing tasks outside their own physical hardware and across remote server networks. Big data is efficiently calculated, analyzed, and stored in large data centers. However, centralized cloud computing is not suitable for all applications and scenarios. The number of mobile devices, wearable devices, sensors, and other Internet of Things (IoT) devices at the edge of the Internet has grown exponentially. Massive data calculation and processing requirements, such as the detection of product quality in manufacturing, the forecast of key equipment operating conditions or video surveillance, etc. Although this has brought opportunities to all areas of science, engineering, business, etc., it has also emerged the challenge of flooding data. This type of demand usually has high real-time characteristics and needs to be analyzed by artificial intelligence technology. [1]One option is to transfer this data to a cloud data center for analysis, but it will lead to increased network traffic, high latency and privacy leaks, faster and more reliable data processing will become crucial. Another option is to analyze on the device and process the data locally by implanting artificial intelligence (AI) programs on the device, but because many AI applications require high computing power, which greatly exceeds resource and energy-constrained terminals capacity, terminal performance and energy efficiency may be affected. The edge intelligence formed by the combination of edge computing and artificial intelligence (deep learning) has the potential to solve these challenges by pushing cloud services from the core of the network to the edge of the network closer to the device and data source. The edge deep learning framework is shown in Figure 1.
2. Overview of Deep Learning
Deep learning (DL) is a subset of AI and machine learning. It uses multiple layers of cascade, each layer is composed of multiple neurons, and can generate non-linear output based on input data. The neurons in the input layer receive the data and propagate it to the intermediate layer (hidden layer). The neurons in the intermediate layer generate a weighted sum of the input data, use a specific activation function to output the weighted sum, and then propagate the output to the output layer.[2] The final result is displayed in the output layer. In recent years, DL has been widely recognized because it has the ability to learn complex models and perform representation learning, has more complex layers and abstraction layers, and can learn advanced functions to achieve high-precision inference in tasks. Layer artificial neural network has achieved great success in effectively solving various problems such as image recognition, computer vision, autonomous driving, natural language processing, and speech recognition. Multilayer Perceptron (MLP), Convolutional Neural Network (CNN) And recurrent neural networks (RNN) have been widely used in the industry, such as Google, Facebook and Nvidia. The difference between deep learning and traditional machine learning technologies is that they can automatically learn representations from data such as images, videos, or text without the need to introduce manual coding rules or artificial knowledge, but deep learning requires strong computing power (high Performance machines and GPUs), need to reduce the model training time through parallel architecture or combined with clusters and cloud computing. In addition, a large amount of labeled data is needed, for example, a large number of pictures and videos required for autonomous driving development.

3. Edge Deep Learning Model

3.1. Computing Model on Terminal Devices
The user terminal device executes a deep learning algorithm. In Figure 2 (a), the terminal device obtains the DNN model from the edge server and performs model training and inference locally. During the inference process, the mobile device does not communicate with the edge server. The performance of this mode depends on the local device itself, and a sufficient number of CPU, GPU, and RAM resources must be available on the terminal device.

3.2. Computing Model on Edge Servers
This mode is suitable for situations where the resources of the terminal device, such as power, computing, and memory are limited, and it is not suitable for deploying large and powerful deep learning neural networks with real-time execution requirements. This method considers the transfer of
deep learning computing from terminal devices to more powerful edge servers or clouds. The cloud center is not suitable for edge applications that require short response times. Edge servers are close to users and can quickly respond to user requests. In Figure 2 (b), the terminal device sends the input data to the edge server. When the DL model training and inference are completed on the edge server, the prediction result will be returned to the device. In this reasoning mode, since the DNN model is located on the edge server, it is easy to implement the application on different mobile platforms. However, the main disadvantage is that inference performance depends on the network bandwidth between the device and the edge server. In Figure 2 (c), a small DL model is deployed on the terminal and a large DL model is deployed on the edge server. The terminal can determine whether to push data to the edge service training or the local device according to network conditions such as device resources, data volume, and delay.

3.3. Computing Model on Cross-edge Devices

The computing model on cross-edge devices refers to the joint computing between terminal devices, edge servers, and cloud data centers. This mode is suitable for situations where device resources are severely limited. In Figure 2 (d), the terminal device transmits data to the edge server through dimensionality reduction or feature extraction. The edge server executes some layers or dimensions of the model, the cloud data center executes the remaining layers or dimensions, and feeds back the final results to the end device. This model depends heavily on the quality of the network connection.

4. Methods for Improving Edge Deep Learning Models

4.1. Hardware

From a hardware perspective, custom application-specific integrated circuits (ASICs) and field-programmable gate array (FPGA)-based DNN accelerators are usually more energy efficient than traditional CPUs and GPUs.[3][5] This method requires software tools to take advantage of the acceleration provided by the hardware.

Figure 2. Edge deep learning model
4.2. Software
From a software perspective, currently the following methods are often used:

Model compression: Model compression refers to optimization of latency, energy consumption, privacy, and memory footprint issues. [4] Weight pruning (removing low-weight neurons), parameter quantization (floating point numbers are changed to low-bit width numbers), compact architecture design (training smaller DNNs by using output predictions generated by larger DNNs) are popular model compression methods. Compared with the original model, the compressed model may lose some accuracy, and the compressed model does not necessarily save a lot of energy, but a comprehensive trade-off is acceptable.

Computing offloading: Computing offloading technology refers to the resource-constrained device that completely or partially offloads compute-intensive tasks to a cloud environment with sufficient resources. It mainly addresses the shortcomings of mobile devices in resource storage, computing performance, and energy efficiency. [5] The computing offloading technology not only reduces the pressure on the core network, but also reduces the delay caused by transmission. This method can reduce latency, reduce energy consumption, and ensure privacy.

Model exits early: In order to ensure the accuracy of calculation, deep learning models usually have a deep hierarchy. This method is based on the sensitivity of the application to accuracy and delay, and a trade-off is made between the two. Depending on where the model exits early, program exit points can be set at edge terminal devices, edge servers, and cloud data centers to obtain different accuracy and delay times. For example: the edge device collects data, the edge server serves as the first exit point, and the cloud data center serves as the second exit point. [6]

Edge caching: Optimize latency by caching DNN inference results. Cache and reuse task results at the edge of the network, reducing application query latency.

Input data preprocessing: When data is sent to the edge server, the data preprocessing avoids redundant calculations of DNN model inference, thereby improving inference accuracy, reducing inference delay, and reducing energy costs.

5. Typical Applications of Edge Deep Learning

5.1. 5G
With the large-scale deployment of high-bandwidth, low-latency, high-capacity 5G networks around the world, in addition to providing traditional communication services, telecom operators will integrate edge learning capabilities in 5G base stations to provide customers with edge computing capabilities.

5.2. Smart Transportation
Autonomous vehicles are embedded with a large number of sensors such as cameras, lidars, etc. Uploading sensor data to the cloud computing center will increase the difficulty of real-time processing. Therefore, performing edge computing on the car can speed up processing and enhance the real-time feature of road environment decisions. The power supply of the drone itself is limited. If the data is transmitted to the cloud center, the energy consumption is high, and the delay is high. Drones generate a large amount of high-definition video data, making it difficult to achieve real-time transmission of wireless networks and commands from receiving centers. [7] Edge computing solves these problems well. The edge end processes the data sensed by the drone, reducing the energy consumption of data transmission and ensuring real-time performance.

5.3. Smart Home
With the popularization of the Internet of Everything application, home life has become more and more intelligent and convenient. However, the practice of connecting smart lighting systems, smart TVs, smart robots and other devices to the cloud computing center through the WiFi module is far from complete. Meet the needs of smart homes. In the smart home environment, a large number of wireless sensors and controllers are deployed in the rooms, pipes, floors and walls. For the consideration of data transmission load and data privacy, the processing of these sensitive data should
be completed within the home. Traditional cloud computing models are no longer fully applicable to smart home applications, and the edge computing model makes it easier to connect and manage smart home devices within the home, and processes the data generated by these devices locally, reducing the load of data transmission bandwidth.[8]

5.4. VR / AR

Both VR and AR technologies need to collect real-time information related to user status including user position and orientation, then perform calculations and process them based on the calculation results. [9] The server can provide it with rich computing and storage resources, cache the audio and video content that needs to be pushed, and based on the one-to-one correspondence between position technology and geographic location information, determine the pushed content based on the location information, and send it to the user or quickly simulate a three-dimensional dynamic visual and interact with users.

6. Conclusion

Edge computing is a distributed computing architecture that brings computing and data storage closer to where it is needed, thereby reducing response time and saving bandwidth, and solving connectivity, latency, scalability, and security challenges. With the massive access of mobile terminals and IoT terminals, edge computing based on AI will help us reconstruct the Internet, thereby better supporting new application scenarios such as 5G, autonomous driving, cloud gaming, VR / AR .

7. References

[1] Xiaoyi Fan, Yutao Huang, Xiaqiang Ma, Jiangchuan Liu and Victor C. M. Leung 2019 Exploiting the edge power an edge deep learning framework CCF Transactions on Networking, vol 2 (1) pp 4-11
[2] Junyang Zhang, Yang Guo, Xiao Hu and Rongzhen Li 2018 Design and Implementation of Deep Neural Network for Edge computing IEICE Transactions on Information and Systems vol E101.D, Issue 8 pp 1982-1996
[3] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo and Junshan Zhang 2019 Edge Intelligence: Paving the last mile of AI with edge computing Proceedings of the IEEE vol 107,Issue 8 pp 1738-1762
[4] Surat Teerapittayanon, Bradley McDanel, H.T. Kung 2017 Distributed Deep Neural Networks over the cloud the edge and end devices 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS) Atlanta, GA, USA pp 328-339
[5] Xiaofei Wang, Yiwen Han, Victor C.M. Leung, Dusit Niyato, Xueqiang Yan and Xu Chen 2020 Convergence of Edge Computing and Deep Learning: A Comprehensive Survey IEEE Communications Surveys & Tutorials, vol 8 pp 47118-47132
[6] Jiasi Chen and Xukan Ran 2019 Deep Learning With Edge Computing:A Review Proceedings of the IEEE vol107,Issue 8 pp 1655-1674
[7] Weisong Shi, Hui Sun, Jie Cao, Quan Zhang and Wei Liu 2017 Edge Computing_A New Computing Model in the Era of IOT Computer Research and Development 54(5) pp 907-924
[8] Zishu Li, Renchao Xie, Li Sun and Tao Huang 2018 Mobile Edge Computing Survey Telecommunications Science pp 87-101
[9] Fengbin Zheng, Dongwei Zhu, Wenqian Zang, Jinlin Yang and Guanghui Zhu 2020 Edge Computing-A New Computing Paradigm Review and Application Research Frontiers of Computer Science and Technology pp 208-225