Distributed Learning Architecture based on Edge-cloud Collaboration

COINRG, IETF 114

draft-li-coinrg-distributed-learning-architecture-00

Authors:
Chao Li (lc96@bupt.edu.cn)
Hui Yang (yanghui@bupt.edu.cn)
Zhengjie Sun (sunzhengjie@bupt.edu.cn)
Sheng Liu (liushengwl@chinamobile.com)
Haomian Zheng (zhenghaomian@huawei.com)
Use case: IoT based on edge intelligence

The whole network architecture is divided into cloud layer, edge layer and device layer.

The device layer is composed of various IoT devices as the service object of the entire network.

The edge side consists of network terminal devices such as base stations and wifi, providing lightweight computing and storage services, as well as lightweight model training.

The cloud layer is deployed in the data center to provide massive computing services, storage services and large-scale model training.
Motivation

The development of 5G technology and the popularization of the Internet of Things, massive edge data is generated by geographically distributed mobile terminals and IoT devices. The ability of a single edge node to process data cannot meet the needs of Internet of Things services, especially the training of artificial intelligence models. Meanwhile traditional cloud computing has two shortcomings when dealing with these network edge data: ① Insufficient bandwidth; ② High energy consumption. ③ Distributed model training architecture based on edge cloud collaboration has become a feasible scheme for artificial intelligence model training.

The service delay of edge cloud collaboration training includes edge computing delay, service transmission delay and cloud processing delay. The edge computing delay and service transmission delay will affect the synchronization of model training, and the asynchronous delay will lead to the reduction of model accuracy and the extension of training time.

The architecture combines data parallelism with model parallelism, and ensures uniform edge training delay and model transmission delay through deterministic optical network technology.
Distributed Learning Architecture based on Edge-cloud Collaboration

(a) Schematic diagram of model splitting, Different splitting schemes require different bandwidth and computing resources.
(b) System architecture
(c) Schematic diagram of edge-cloud collaborative training, the edge device trains the first half of the model, and the cloud server trains the second half of the model.
Simulation Result

- **Accuracy under different training methods**
- **Edge server CPU utilization of different training method**
- **Network service QoE of different training methods**

- Our proposed architecture improves the training accuracy of edge AI models. This is because in the training process, the framework dynamically assigns training tasks, so that the model is fully trained.
- It also relieves the computational pressure in the model training process, ensures that other network services can get sufficient computational support, and its QoE is not affected by the model training.