Deep Learning for the Internet of Things with Edge Computing

He Li, Kaoru Ota, and Mianxiong Dong

Research Fund for Postdoctoral Program of Muroran Institute of Technology - KDDI Foundation

Objectives

- Introduce deep learning for IoTs into the edge computing environment.
- Design a novel offloading strategy to optimize the performance of IoT deep learning applications with edge computing.
- Test the performance of executing multiple deep learning tasks in an edge computing environment with our strategy.

Introduction

We introduce deep learning for IoT into the edge computing environment to improve learning performance as well as to reduce network traffic. We formulate an elastic model that is compatible with different deep learning models. We state a scheduling problem to maximize the number of deep learning tasks with the limited network bandwidth and service capability of edge nodes. We design offline and online scheduling algorithms to solve the problem. We perform extensive simulations with multiple deep learning tasks and given edge computing settings.

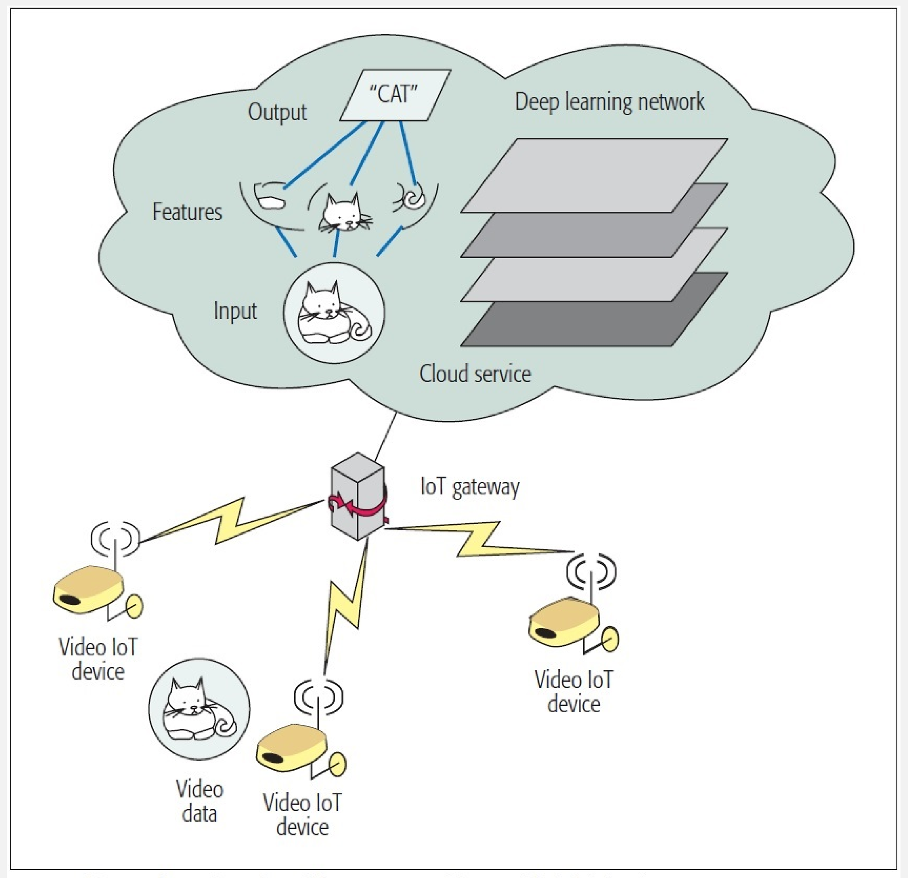

FIGURE 1. Deep learning for video recognition with IoT devices.

Deep Learning and Edge Computing

Edge computing is proposed to move computing ability from centralized cloud servers to edge nodes near the user end. Edge computing brings two major improvements to the existing cloud computing. The first one is that edge nodes can preprocess large amounts of data before transferring them to the central servers in the cloud. The other one is that the cloud resources are optimized by enabling edge nodes with computing ability.

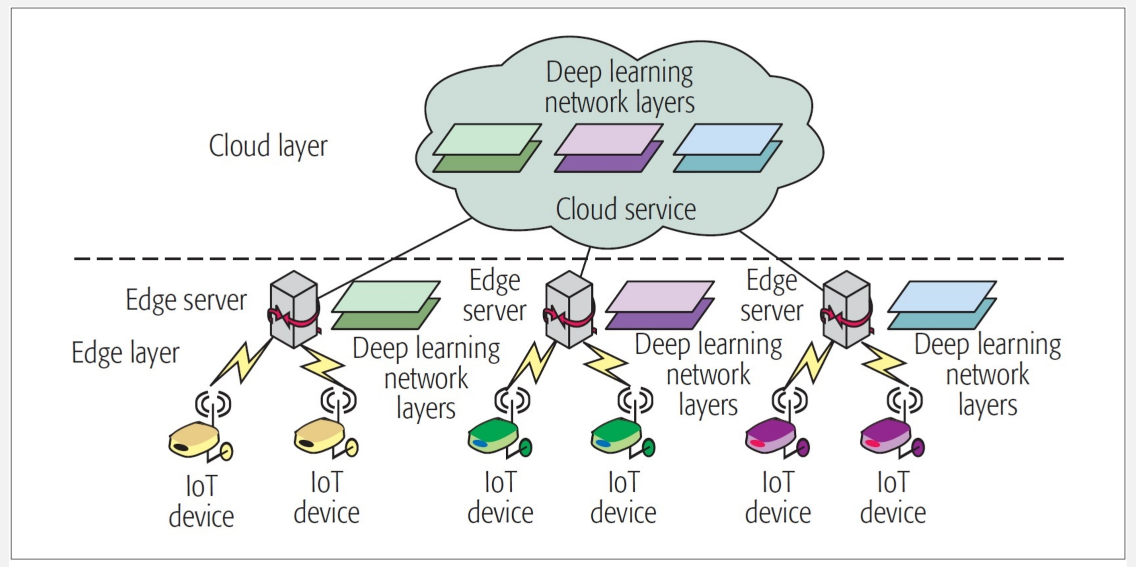

FIGURE 2. Edge computing structure for IoT deep learning.

Deep Learning IoT in Edge Computing

The structure for IoT deep learning tasks, consists of two layers as well as a typical edge computing structure. In the edge layer, edge servers are deployed in IoT gateways for processing collected data. We first train the networks in the cloud server. After the training, we divide the learning networks. One part includes the lower layers near the input data, while another part includes the higher layers near the output data. We deploy lower layers into edge servers and higher layers into the cloud for offloading processing. The collected data are input into the first layer in the edge servers. The edge servers load the intermediate data from the lower layers and then transferred data to the cloud server as the input data for the higher layers.

Important Result

The most important benefit of deep learning over machine learning is the better performance with large data scale since many IoT applications generate a large amount of data for processing. Another benefit is that deep learning can extract new features automatically for different problems.

Scheduling Problem and Solution

To assign maximum tasks in the edge computing structure by deploying deep learning layers in IoT edge servers such that the required transferring latency of each task can be guaranteed, denoted by

$$\begin{aligned} \max \quad & \sum_{i=1}^{|E|}\sum_{j=1}^{|T|} X_{ij} \\ s.t. \quad & \sum_{i=1}^{|E|} b_{ij} \le b_i \cdot V \\ & X_{ij}\cdot d_{ij}\cdot r_{kj}/b_{ij} \le Q_j \\ & \sum_{j=1}^{|T|} l_{kj}\cdot d_{ij}\cdot X_{ij} \le c_i \end{aligned} \quad (1)$$

where Xij = 1 if task tj is deployed in edge server ei; otherwise, Xij = 0.

Performance Evaluation

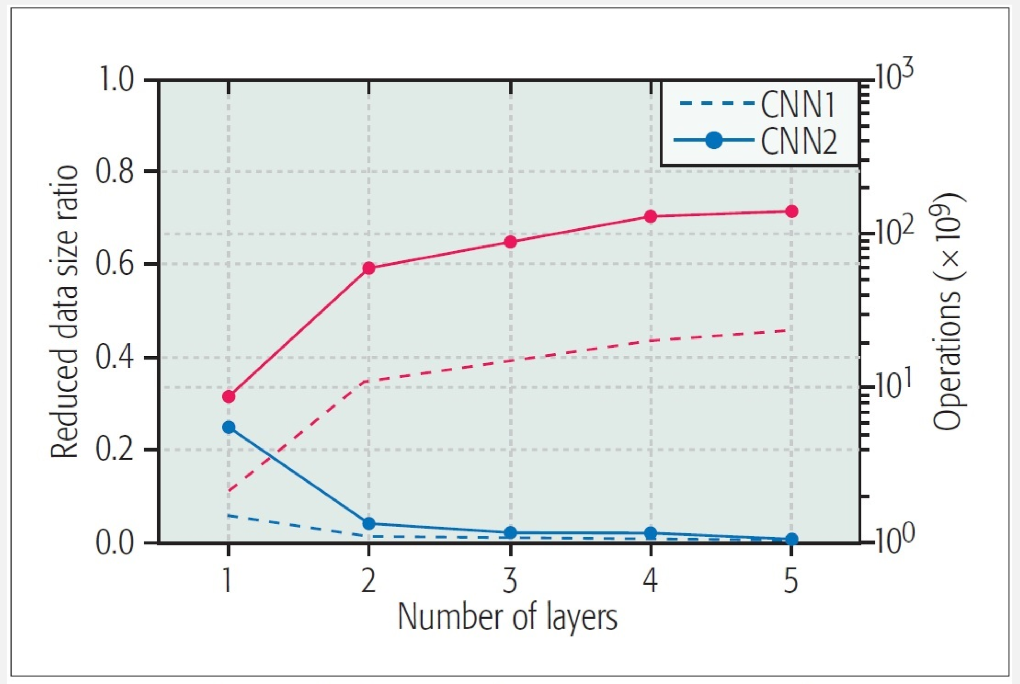

FIGURE 3. Reduced data and operations in deep learning networks.

Conclusion and Future Work

We introduce deep learning for IoT into the edge computing environment to optimize network performance and protect user privacy in uploading data. The edge computing structure reduces the network traffic from IoT devices to cloud servers since edge nodes upload reduced intermediate data instead of input data. We propose algorithms to maximize the number of tasks in the edge computing environment. In the experiments, we choose 10 different CNN models as the deep learning networks and collect the intermediate data size and computational overhead from practical deep learning applications. The results of the performance evaluation show that our solutions can increase the number of tasks deployed in edge servers with guaranteed QoS requirements. As future work, we plan to deploy deep learning applications in a real-world edge computing environment with our algorithms.

References

[1] N. Kato.
The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and perspective.
IEEE Wireless, 24(3):146–53, June 2017.

[2] K. Ota L. Li and M. Dong.
When weather matters: Iotbased electrical load forecasting for smart grid.
IEEE Commun, 55(10):46–51, Oct 2017.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Numbers JP16K00117, JP15K15976, and JP17K12669, the KDDI Foundation, and the Research Fund for Postdoctoral Program of Muroran Institute of Technology. Mianxiong Dong is the corresponding author.