Upgrade of Deep Neural Network-based Optical Monitors by Communication-Efficient Federated Learning

Takahito Tanimura and Masayuki Takase
Research and Development Group, Hitachi Ltd., 1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan
takahito.tanimura.kz@hitachi.com

Abstract: We present an efficient scheme to upgrade DNN-based optical monitors collaboratively trained through multiple network operators without revealing each confidential data, applying federated learning with pre-model size reduction based on transferable lottery ticket hypothesis.

1. Introduction

A use of federated learning (FL) [1] has been proposed and demonstrated [2-4] to train machine learning (ML)- and/or deep neural network (DNN)-based optical monitors [5] that are one of the vital parts of future autonomous optical networks. In existing centralized model training manner, a data availability in optical networks is an issue to train DNN-based optical monitors, because the monitors require training data with enough diversity and quantity. FL can solve this issue by assembling many pieces of small dataset owned by many individual data owners with keeping data privacy requested by regulations that do not allow to share and/or move data across different data owners/regions.

A typical setting of the FL-based training for optical monitors is shown in Fig. 1(a). The FL system enables to collaboratively train model among multiple clients (e.g., administrative network operators) without revealing each own dataset. Fig. 1(b) shows typical FL process between a FL model aggregator and several FL clients in a single FL communication round.

Although the FL systems can solve the data availability issue, there is still room to investigate toward realistic implementations in optical networks. Such FL systems for optical networks would perform their distributed training via data control network (DCN) channels and/or other supervisory control channels of optical networks. This suggests communication cost between the FL clients and the FL model aggregator becomes a key issue in practical implementation, because these control channels usually provide limited capacity rather than main data channels such as 400 Gb/s.

In this paper, we present a scheme to tackle this issue on distributed learning of DNN-based optical monitors. The presented scheme enables communication-efficient FL-based optical monitor training by reducing a model size of initial DNN based on the lottery ticket hypothesis (LTH) [6]. As an early-stage proof of concept, we numerically demonstrated the scheme on the scenario to upgrade convolutional neural network (CNN)-based OSNR estimator from supporting single modulation type (64-GBd QPSK) to supporting multiple modulation types (64-GBd QPSK and 16QAM).

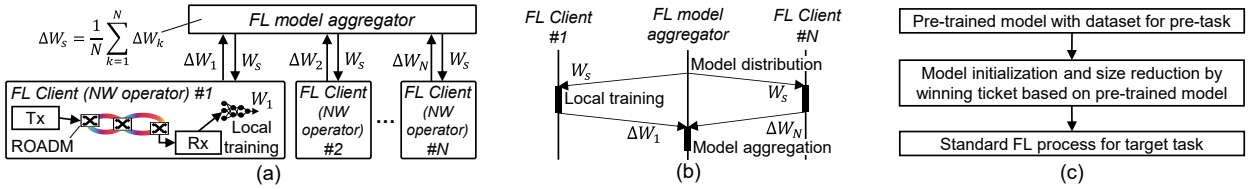

Fig. 1. Schematic diagram of (a) DNN-based optical monitor training with federated learning (FL), (b) sequence of one communication round of FL, (c) server-side model initialization based on the lottery ticket hypothesis.

2. Communication-efficient federated learning for training DNN-based optical monitors

Considering the upgrade scenario of DNN-based optical monitors, we assume the FL system holds a pre-trained model for pre-task related to monitoring at the beginning of this scenario. The FL system re-trains the model in the standard FL scheme for upgrading the model. To re-train the model with communication-efficient manner, we insert a pre-process before starting the standard FL [1] for model initialization and resizing. This method is based on Ref. [7]. As shown in Fig. 1(c), the method consists of two stages: (1) resizing and initializing model based on pre-trained model for surrogate pre-task, e.g., monitoring for QPSK signals only, and (2) re-training the model for target upgraded task, e.g., monitoring for both QPSK and 16QAM signals, through standard FL process.

In the first stage that resizes a model at the server, a pre-trained DNN model by the pre-task is used to extract a reduced sub neural network from the original DNN, based on the winning ticket of the LTH [6]. The pre-task is similar yet different from the target task; thus, the extracted sub-network (the winning ticket of the LTH) based on the pre-task should be different one extracted for the target task in general. Nonetheless, recently, observations that some winning tickets can be transferable to the different other tasks have been reported in Ref. [8]. This study is based on the observations of the transferable winning ticket.

The specific methodology to find the winning ticket, i.e., reduced neural network model, is the one-shot method [6]. First, we prepare the mask m of the original model, initializing it with one. The mask is a flag of trainable parameters of the original model. We selected p % of the smallest magnitude of the parameters in the pre-trained DNN model, and the corresponding mask parameters set to zero, i.e., untrainable. Finally, by using initial parameters of original model θ_0, the winning ticket of the LTH is generated as $m \cdot \theta_0$. Through the process, the number of the trainable parameter of the model, corresponding to model size, is reduced from N to (1 - p) N, where N is total number of parameters of the original DNN model.

In the following FL stage, standard FL based on FedAvg [1] is employed to train the winning ticket-based model for the target task. By utilizing the winning ticket model instead of the original model, the required communication cost in FL to upgrade the DNN-based optical monitors can be reduced by (1 - p) in terms of relative communication cost between FL clients and a FL model aggregator.

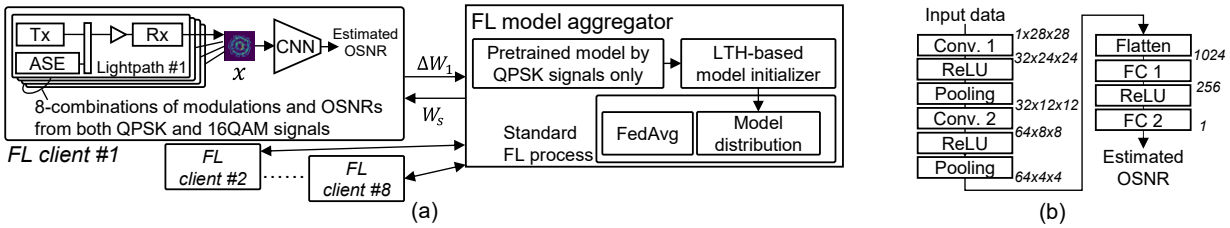

Fig. 2. Simulation setup of (a) federated learning for CNN-based OSNR estimation, and (b) CNN used in this study. Conv. 1: 2-dimentional convolution layer with 32 channels 5 × 5 kernel, ReLU: rectified linear unit, Pooling: 2 × 2 max pooling, Conv. 2: 2-dimentional convolution layer with 64 channels 5 × 5 kernel, FC 1 and 2: fully-connected layers, Flatten: reshaping from 64 × 4 × 4 to 1,024, CNN: convolutional neural network, FL: federated learning, ASE: amplified spontaneous emission noise source, Tx: optical transmitter, Rx: digital coherent optical receiver.

3. Simulation setup

Figure 2(a) shows simulation setup for the proof of the concept. Standard FL part in the simulation was implemented based on [9]. CNN model to estimate OSNR was deployed as an original model for eight FL clients connected with a single FL model aggregator. Detailed of the CNN are described in Fig. 2(b). Dataset to train the CNN is a set of 2D constellation histograms, which are 28 × 28 pixels single channel images generated from 512 IQ data points sampled after adaptive equalizer of digital coherent receiver (1 sample/symbol).

The original model was pre-trained before starting FL process by supervisory manner with a dataset consists of QPSK signals only with OSNRs ranging from 10 to 25 dB (total 10,080 histograms). An initial model for FL was generated from the pre-trained model by its winning ticket of the LTH with p = 0.9.

In following FL stage, a local dataset of each client contained 5,040 constellation histograms composed of randomly selected 8 different combinations of modulation formats (64-GBd 16QAM or QPSK) and OSNRs ranging from 10 to 25 dB with 1-dB step. Note that the dataset for FL was different from the dataset used for pre-training. For the simplest proof of concept, we added only ASE to the signals and investigation including fiber nonlinearity was left for future work. Each local CNN at FL client was trained with only their local dataset with SGD optimizer on learning rate of 10^{-3} and batch size of 128. An update information of model weights ΔW_k extracted from local model at k-th FL client was uploaded to the FL server. The FL server aggregates all local update information ΔW_k (k = 1, 2, … 8) to generate weights of a shared CNN model by FedAvg (local epoch = 3). The model update of the shared model was sent to all FL clients for updating their local CNN models.

After the FL training, the trained model was evaluated by test dataset contained 2,048 constellation histograms generated by another 16QAM/QPSK modulation pattern generated by Mersenne twister with different seed.

4. Results and discussion

First, we evaluated functionality of CNN-based OSNR monitors on both full-model (no model resizing before FL) and LTH-based model (selected 90 % of the parameters of the full model was set to zero). Figures 3 show mean estimated OSNR values by the CNN-based OSNR monitors as a function of actual OSNR values. Fig. 3(a) shows

the pre-trained model had no functionality to estimate OSNR for 16QAM signals, since the model was trained with QPSK data only. In Fig. 3(b), we evaluated the full model trained with FL manner. Note that the full model was initialized with random values before FL. The full model was successfully trained to estimate correct OSNR values for both QPSK and 16QAM signals without revealing raw dataset of each client. Next, we investigated functionality of the reduced model, i.e., the LTH-based model with $p = 0.9$. Figure 3(c) shows the LTH-based model can correctly estimate OSNR values for both QPSK and 16QAM, even with the 90 % model size reduction.

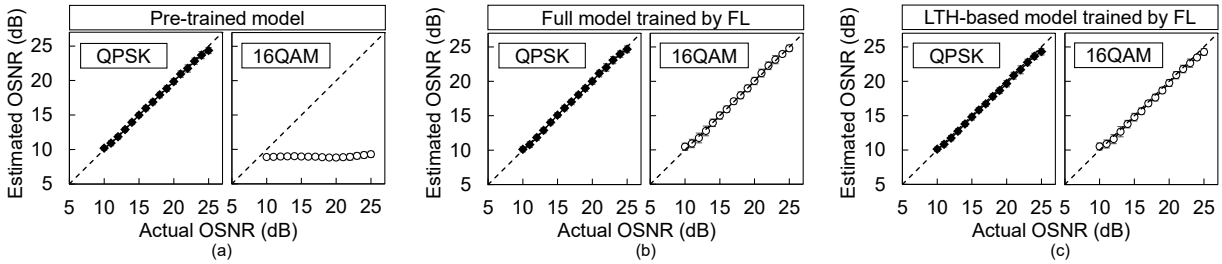

Fig. 3. CNN-estimated OSNR values with (a) pre-trained model, (b) full model, and (c) LTH-based resized model trained by FL.

For investigation on communication cost for FL, we plotted mean OSNR estimation errors calculated from test dataset in Figs. 4. We compared the LTH-based model with both the full model and a random reduction model that sets randomly selected parameters of the original DNN model to zero and untrainable (model reduction rate of 0.9). Fig. 4(a) shows the mean errors as a function of communication round of FL. Although the LTH-based FL outperformed the random model reduction, the required communication round at mean error of 0.5 dB with the LTH-bases FL was almost same but slightly larger than that with the full-model FL. Figure 4(b) shows mean OSNR estimation error as a function of relative communication cost calculated by the number of communication round and parameters of CNN model used for FL. The result shows the LTH-based FL successfully reduced the required communication cost to achieve test error of 0.5 dB by around 80 %, comparing with the full-model FL case. Additionally, the LTH-based FL also outperformed the random model reduction. This result suggests the LTH-based FL obtained a "better" sub-network in the pre-processing stage than randomly chosen ones.

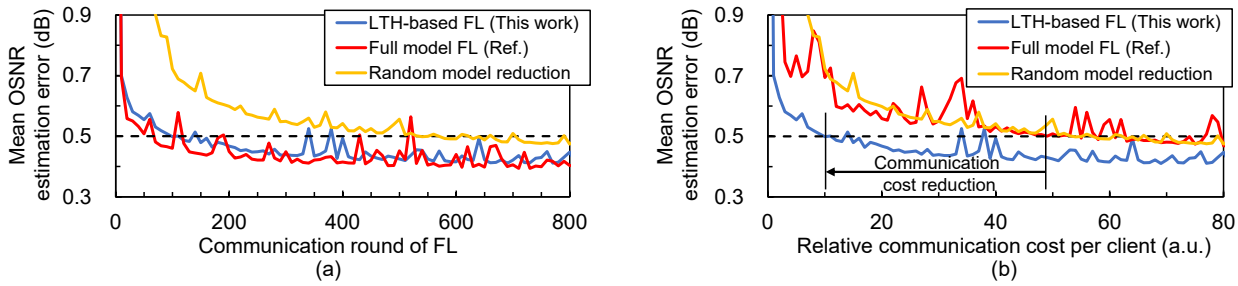

Fig. 4. Mean estimation error of test dataset as a function of (a) communication round, and (b) relative communication cost for FL.

5. Summary

We presented communication efficient scheme to upgrade DNN-based optical monitors managed and trained by multiple network operators in FL. The simulation demonstrated the scheme reduced required communication cost of FL by around 80 % in the specific scenario for upgrading the CNN-based OSNR monitor.

References

[1] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," AISTATS 2017, pp 1273-1282.
[2] B. Shariati *et al.*, "Demonstration of federated learning over edge-computing enabled metro optical networks," ECOC 2020, Tu5A.2.
[3] N. Hashemi *et al.*, "Vertical federated learning for privacy-preserving ML model development in partially disaggregated networks," ECOC 2021, We3E.3.
[4] T. Tanimura *et al.*, "Optical status representation by collaborative and unsupervised learning," OECC 2022, WB3.4.
[5] T. Tanimura and T. Hoshida, Deep learning techniques for optical monitoring. in Alan Lau (Eds.) and Faisal Khan (Eds.), Machine learning for future fiber-optic communication systems (1st edition), Elsevier (2022).
[6] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," ICLR 2019.
[7] T. Tanimura *et al.*, "Compressing model before federated learning by transferrable surrogate lottery ticket," accepted for IEEE CCNC 2023.
[8] A. Morcos *et al.*, "One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers," NeurIPS 2019. pp. 4933-4943.
[9] F. Sattler *et al.*, "Robust and communication-efficient federated learning from non-IID data," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 9, pp. 3400-3413 (2020).