Skip & Swap: Efficient Weight Spreading for Decentralized Machine Learning with Non-IID Data

Asato Yamazaki[1,a)] Takayuki Nishio[1,b)] Yuko Hara-Azumi[1,c)]

Abstract: Although several methods have been proposed for decentralized machine learning, they faced a non-convergence problem for non-independent and identically distributed (non-IID) data. We propose a novel decentralized machine learning method, which efficiently spreads weight parameters among networked devices so that each model can be quickly trained even with non-IID data. Our evaluation quantitatively demonstrated that our method outperforms existing methods in terms of convergence and accuracy even with highly non-IID data in sparse network topologies.

Keywords: Decentralized Machine Learning, non-IID Data, Weight Spreading

1. Introduction

By the development of IoT devices equipped with various sensors, the amount of collected data has been dramatically increasing. In order to reduce communication loads on cloud and central servers as well as to address privacy concerns, it is preferable to locally process the data on the devices. Besides, considering differences in data collected from multiple devices, distributed machine learning is expected to be used for IoT devices, where an identical neural network model is deployed on the devices and only weight parameters of the trained models are communicated among them. Recently several methods have been proposed for "decentralized" machine learning, such as Decentralized Federated Learning [1] and Gossip Stochastic Gradient Descent (SGD) [2] that averages weight parameters between neighboring devices. Most of these existing methods are effective only for independent and identically distributed (IID) data in dense communication topologies.

In this work, we propose a novel decentralized machine learning method on top of Gossip SGD, named as *Skip Swap SGD*, which can deal with even highly non-IID data in sparse communication topologies. As aforementioned, Gossip SGD is known to be poor at convergence in high accuracy for all classes of non-IID data. Hence, recently Primal-Dual Method of Multipliers (PDMM) SGD [3] (hereafter simply PDMM) was developed to make Gossip SGD theoretically effective for non-IID data by introducing an additional variable to accelerate the learning. However, as our evaluation will empirically disclose later, PDMM cannot learn highly non-IID data well more than targeted. We then found that exchanging parameters quickly between connected devices in a lightweight manner is effective for improving model convergence, particularly for highly non-IID data in sparse topologies. In our work, weight skipping and swapping (that will be elaborated in Section 3) help quick parameter exchanges with neighboring devices and far devices (at least two hops away), respectively, without incurring computational overhead. We demonstrate that our method achieves faster convergence in environments where even PDMM cannot learn well.

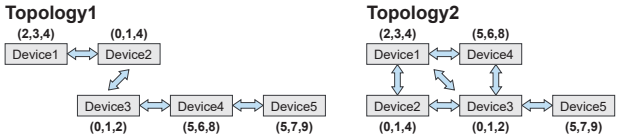

Fig. 1 Non-IID data setup in a sparser topology (Topology 1) and a denser topology (Topology 2). The local labels on each device are shown in parentheses. Please see the legend of Fig. 2 to find what image class each label represents.

2. Motivation

Here we explain a previous work that inspired our work. FedSwap [4], a "centralized" federated learning method for non-IID data, proposed an effective approach to exchange weight parameters (i.e., swap). In FedSwap, each device communicates and exchanges weight parameters with each other via a central server after updating its own weight parameters using local data. This swapping enables to train the models for non-local data and achieve fast convergence even for highly non-IID data.

Contrary to FedSwap [4], this work targets "decentralized" machine learning, for which it is difficult to exchange parameters with distant devices especially in sparse topologies (e.g., Devices 1 and 5 in Topology 1 of **Fig. 1**). Consequently, in order to achieve effective decentralized machine learning with non-IID data, an efficient method that can spread trained parameters of each device among a set of networked devices (i.e., both neighboring devices and distant devices) is required.

3. Skip Swap SGD

We propose a novel decentralized machine learning method, *Skip Swap SGD*, by which each device not only exchanges (or *swaps*) parameters with neighboring devices to train its own model, but also plays a role of "hub" to exchange (or *skip*) pa-

[1] Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[a)] yamazaki.a.ah@m.titech.ac.jp
[b)] nishio@ict.e.titech.ac.jp
[c)] hara@cad.ict.e.titech.ac.jp

Algorithm 1 Parameter exchanges in Skip Swap SGD

1: Initialization of $w_i, w_{i\to j}$
2: **for** $t \in \{0, \ldots, T-1\}$ **do**
3: $\quad w_i \leftarrow w_i - \alpha \nabla F_i(w_i, x_i^t)$
4: $\quad$ select $j \in N(i)$ at random and receive $w_{j\to i}$
5: $\quad w_{i\to j} \leftarrow w_i$
6: $\quad w_i \leftarrow w_{j\to i}$
7: $\quad$ **for** $k \in N(i)\backslash\{j\}$ **do**
8: $\quad\quad w_{i\to k} \leftarrow w_{j\to i}$
9: $\quad$ **end for**
10: **end for**

rameters of two disconnected devices via itself. For example, in Topology 1 of Fig. 1, the former corresponds to parameter exchanges between Devices 1 and 2, while the latter corresponds to that between Devices 1 and 3 via Device 2.

Algorithm 1 describes a pseudo code of Skip Swap SGD on device i. We denote weight parameters on device i and those that are sent from devices i to j as w_i and $w_{i\to j}$, respectively. First w_i and $w_{i\to j}$ are both initialized (line 1). The total number of training iterations is defined as T, and training and communications are conducted repeatedly at every iteration (line 2). Each model is trained by its own local dataset using SGD as an optimizer with the learning rate α, loss function F_i, and a mini batch x_i^t at iteration t (line 3). After local training, device i randomly selects one device from $N(i)$, which is a set of devices connected to device i (line 4). For swapping parameters, device i sends its trained parameters to device j (line 5), and its own parameters are overwritten by parameters of device j (line 6). Then, for skipping parameters, device i sends parameters of device j to the other devices in $N(i)$ (lines 7-9). These procedures are lightweight as our method does not need averaging calculation of parameters (unlike Gossip SGD) or heavy computation for additional variables (unlike PDMM).

4. Evaluation

We quantitatively simulated PDMM and Skip Swap SGD (hereafter "Skip Swap") in NVIDIA GeForce RTX3070 for an image classification task (CIFAR-10). A neural network consisting of four convolutional layers, two pooling layers, and two fully-connected layers was assumed to be deployed on IoT devices. We used cross-entropy error for the loss function and set mini-batch and epoch sizes to 100 and 200, respectively. Each device updates its own model by communicating with its neighbors every local training. Fig. 1 describes two experimental environments with different sparsity of topology. For example, (2,3,4) for Device 1 in Topology 1 indicates that this device has three types of data (bird, cat, and deer). Due to space limitations, we show the results of Device 5, which would be hard to communicate with other devices in both topologies and hence can clearly compare the effects of PDMM and Skip Swap.

Figure 2 shows the results on transition of test accuracy, where the x-axis presents the number of epochs and the y-axis indicates accuracy for each class. As shown in Figs. 2(a) and (c), PDMM could not quickly converge and failed to achieve high accuracy for non-local data. This means that even though PDMM was developed for non-IID data, the data distribution in our evaluation was more strongly biased than targeted. On the other hand, Figs. 2(b) and (d) show that Skip Swap successfully converged to high accuracy for all classes. To quantify the training efficiency, **Table 1** shows time per epoch on Device 5 in Topology 1 and Topology 2 with a reference to the original Gossip SGD (specified as Gossip). These results well demonstrated the high efficiency of Skip Swap over PDMM as Skip Swap has less computation and communication costs. Furthermore, Skip Swap even slightly sped up than Gossip by saving the averaging calculation of exchanged parameters. In summary, Skip Swap achieves quick convergence to high accuracy even under challenging environments (i.e., highly non-IID data in sparse topologies).

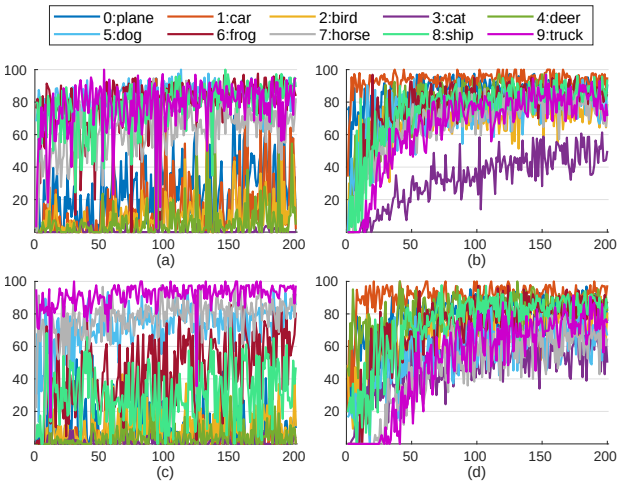

Fig. 2 Results of Device 5: (a) PDMM in Topology 1, (b) Skip Swap in Topology 1, (c) PDMM in Topology 2, and (d) Skip Swap in Topology 2.

Table 1 Time per epoch (seconds)

	Gossip	PDMM	Skip Swap
Topology 1	124.11	241.51	121.48
Topology 2	130.36	267.62	129.30

5. Conclusion

We proposed a novel decentralized distributed machine learning method, Skip Swap SGD, that enables to efficiently spread weight parameters among neural network models on networked devices. We demonstrated that our proposed method makes quick convergence even for highly non-IID data regardless of communication topological sparsity. In our future work, we will further improve our method by reducing communication frequency and/or selecting appropriate devices to exchange parameters.

Acknowledgments This work was partially supported by JSPS KAKENHI Grant Numbers JP20K21789 and JP20H04154, and JST, PRESTO Grant Number JPMJPR2035.

References

[1] Liu, W., Chen, L. and Zhang, W.: Decentralized Federated Learning: Balancing Communication and Computing Costs, *IEEE Trans. on Signal Inf. Process. Netw.*, Vol. 8, pp. 131–143 (2022).
[2] Jin, P. H. et al.: How to Scale Distributed Deep Learning?, *Proc. of ML Systems Workshop* (2016).
[3] Niwa, K. et al.: Edge-consensus Learning: Deep Learning on P2P Networks with Nonhomogeneous Data, *Proc. of Int'l Conf. on Knowledge Discovery & Data Mining*, pp. 668–678 (2020).
[4] Chiu, T.-C. et al.: Semisupervised Distributed Learning with non-IID data for AIoT Service platform, *IEEE Internet of Things Journal*, Vol. 7, No. 10, pp. 9266–9277 (2020).