Overview of Distributed Federated Learning: Research Issues, Challenges, and Biomedical Applications

Joohyung Jeon[1], Joongheon Kim[2], Jeongwoo Huh[1], Hyunbum Kim[3], and Sungrae Cho[1]
[1]*School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea*
[2]*School of Electrical Engineering, Korea University, Seoul, Korea*
[3]*Department of Computer Science, University of North Carolina at Wilmington, NC, USA*
joongheon@korea.ac.kr

***Abstract*—This paper discusses about distributed federated learning research issues and challenges. The federated learning is actively studied nowadays in many applications. In this paper, we will figure out the foundations of federated learning and its related research issues/challenges. Lastly, we will review the major two models (centralized and distributed) in distributed federated learning.**

I. Introduction

Artificial intelligence (A.I.) and deep learning computations are widely used in many areas. For the A.I. and deep learning procedure, we need to prepare all training data before starting the training learning computation due to supervised machine learning nature.

However, gathering all data in a single storage is not always easy. For example, gathering all patients data in a single hospital is strictly limited by law due to privacy conservation issues. Therefore, it is essential to study about the technique which is called federated learning that enables A.I. and deep learning computation over distributed data without sharing them [1]–[7].

In this paper, we study about the potential research issues and challenges in distributed federated learning (refer to Sec. II) and then present emerging up-to-date deep learning models for that (refer to Sec. III). Finally, this paper concludes in Sec. IV.

II. Research Issues and Challenges

Suppose that we have K number of distributed computing platforms ($\mathcal{C} = \{c_1, c_2, \cdots, c_K\}$). In the computing platforms, many data which cannot be shared are located. Based on the distributed data, deep learning computation should be conducted. Note that the amounts of data in distributed computing platforms are denoted as $\mathcal{N} = \{n_1, n_2, \cdots, n_K\}$.

A. Overfitting

If the value of $\mathcal{N} = \{n_1, n_2, \cdots, n_K\}$ is extremely low, it will introduce overfitting issue because A.I. and deep learning computation occurs with extremely low number of data in local computing platforms. This situation is more serious in medical deep learning applications because medical applications usually require high accuracy.

B. Data Aging

In distributed medical federated learning platforms, each platform (i.e., each hospital) has its own patients data. In some hospitals, specific medical disease information is for the last period patients. In the other hospitals, specific medical disease information is for the early period patients. Therefore, their tendencies are quite different even though the data is for the same disease. This kind of problems is called the *data aging* problem.

C. Data Distribution

Similar to overfitting issue in previous subsection, there exist data distribution issues in distributed medical federated learning applications. Suppose that one hospital is very famous in diabetes. Then the hospital might have a lot of data related to diabetes whereas the amounts of the other disease data will be less than the amount of diabetes data. This issue can introduce A.I. and deep learning training performance degradation; therefore, federated learning algorithms should take care of this issue as well.

III. Models

In distributed medical federated learning, there are two kinds of models, i.e., centralized and distributed.

- *Centralized Model [8]:* In this model, one centralized high performance computing (HPC) platform exist and all medical computing platforms are connected to the HPC platform (i.e., star topology). Each medical computing platform has its own data and the first hidden layer in the given deep learning computation model. All the other layers in the deep learning computation model are in the HPC platform. This is the model inspired by Split Learning [1]. Even though the first layer is distributed, the other all layers are in one HPC platform. Thus, this architecture is able to achieve pseudo-optimal performance. More detailed descriptions about this model are in [8].
- *Distributed Model [9]:* In this case, each local platform will be connected in a circular way. Suppose that all platforms have same model at first (i.e., they will share

the information about neural network architecture, hyper-parameters, etc). The first platform c_1 will train its own model with its own local data; and then the c_1 will pass its own up-to-data training results to the next hop platform c_2. Then the c_2 will train its own model with its own local data; and so forth. More detailed descriptions about this model are in [9].

IV. Concluding Remarks

This paper discusses about the distributed federated learning which is actively and widely discussed nowadays in deep learning platform research societies. Especially, medical deep learning is one of the main applications in federated learning. Furthermore, this paper presents research issues and challenges and then finally this paper shows major two models in the research domain.

Acknowledgment

This research was supported by National Research Foundation of Korea (2019R1A2C4070663) and also by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation)" (20170001000031001). J. Kim and S. Cho are the corresponding authors of this paper (e-mails: joongheon@korea.ac.kr, srcho@uclab.re.kr).

References

[1] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split Learning for Health: Distributed Deep Learning without Sharing Raw Patient Data," *arXiv:1812.00564*, 2018.

[2] J. Kim and W. Lee, "Stochastic Decision Making for Adaptive Crowdsourcing in Medical Big-Data Platforms," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 11, pp. 1471–1476, November 2015.

[3] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, Denver, Colorado, USA, October 2015.

[4] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A.y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, Florida, USA, 20-22 April 2017.

[5] J. Chen, X. Pan, R. Monga, and S. Bengio, "Revisiting Distributed Synchronous SGD," *arXiv preprint arXiv:1604.00981*, 2016.

[6] O. Gupta and R. Raskar, "Distributed Learning of Deep Neural Network over Multiple Agents," *Journal of Network Computing and Applications*, vol. 116, pp. 1–8, August 2018.

[7] J. So, B. Guler, A.S. Avestimehr, and P. Mohassel, "CodedPrivateML: A Fast and Privacy-Preserving Framework for Distributed Machine Learning," *arXiv preprint arXiv:1902.00641*, 2019.

[8] J. Jeon, J. Kim, J. Kim, K. Kim, A. Mohaisen, and J.-K. Kim, "Privacy-Preserving Deep Learning Computation for Geo-Distributed Medical Big-Data Platforms," in *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) (Supplemental Volume)*, Portland, Oregon, USA, June 2019.

[9] J. Jeon, D. Kim, and J. Kim, "Cyclic Parameter Sharing for Privacy-Preserving Distributed Deep Learning Platforms," in *Proceedings of the IEEE International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Okinawa, Japan, February 2019.