07

Distributed learning algorithm and medical application

Privacy-aware machine learning

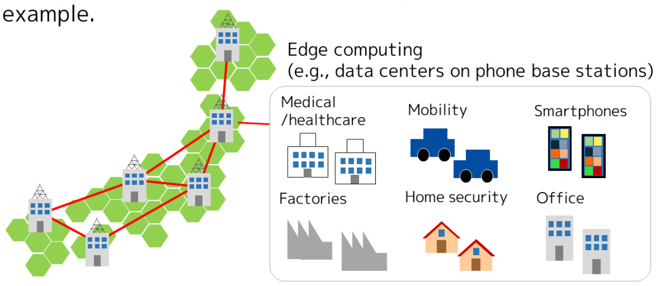
Abstract

While use of massive data benefits on training deep learning models, aggregating all data into one physical location (e.g. a central cloud) may not be possible due to data privacy concerns. For example, according to the EU GDPR, data transmission should be minimized among processing nodes. Our goal is to construct training algorithms to obtain global deep learning models that can be adapted to all data, even when individual nodes only have access to different subsets of the data. We assume that this algorithm is allowed to communicate between nodes in an asynchronous/sparse manner, exchanging such information as model variables or their update differences. However, data are prohibited from being moved from the node on which they reside. We aim to indirectly exploit the overall data across countries and provide high performance services for such industries as the medical/health-care field while protecting privacy.

Goal and application

Background: We are entering an era of distributed data processing (inference/training) due to data volume, privacy-aware issues, and legal regulations, e.g., GDPR.

Goal: To train deep learning models without aggregating data to a central cloud, where asynchronous communication among nodes are allowed to exchange latent variables, for example.



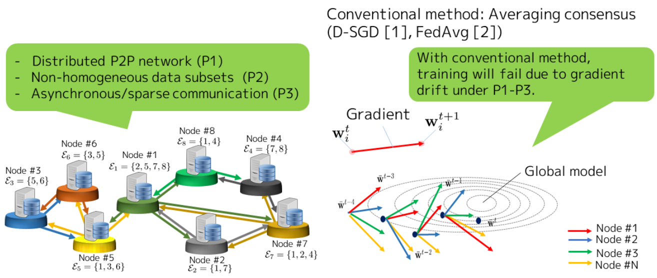
Problems

Aim: Our goal is to obtain a global model, which is equivalent to a model trained at a central cloud using all datasets, even data distributed in N(≧2) nodes.

Problem 1: Network structure is distributed in P2P manner to scale service at any scale.

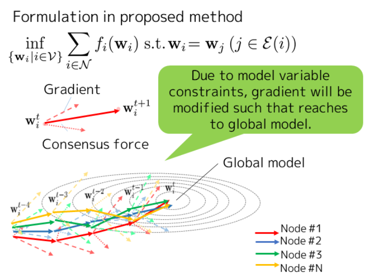
Problem 2: Non-homogeneous data subsets are placed for each node. Training procedure is unstable due to gradient drift.

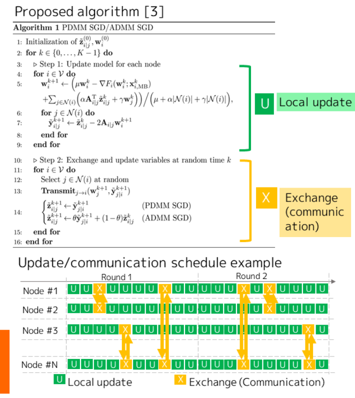
Problem 3: Communication among nodes is in asynchronous/sparse manner due to large-scale network.



Asynchronous decentralized deep learning

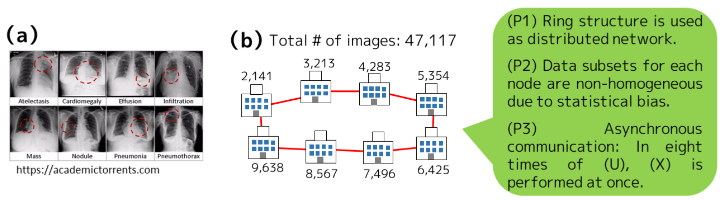
Proposed algorithm: We solved a model variable constrained minimization problem. Training is achieved by alternatingly repeating (U) local node model updates and (X) exchange auxiliary variables for progress w.r.t. making consensus between nodes. This scheme is runnable on arbitrary network structure with asynchronous communication.



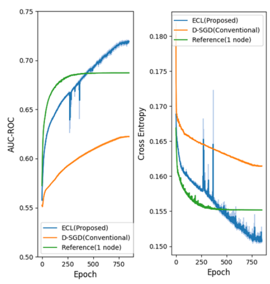


Medical image analysis

We trained 14 different disease detection models using chest X-ray datasets [4]. Image data were not transmitted from eight hospital nodes.



(Experimental results)



- Detection accuracy for some diseases was measured using Area Under curve of Receiver Operating Characteristic (AUC-ROC).
- Conventional method (orange) did not attain global model performance (green) due to gradient drift.
- Proposed method (blue) reached global model performance (green) even though P1-P3 are present.
- Detection accuracy for a part of disease was practical level: AUC-ROC of 0.75 or higher for e.g., emphysema, pneumothorax, cardiomegaly, pleural effusion.

References

[1] J. Chen, A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, Vol. 60, No. 8, pp. 4289–4305, 2012.
[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, "Communication–efficient learning of deep networks from decentralized data," in *Proc. Artificial Intelligence and Statistics (AISTATS 2017)*, pp. 1273–1282, 2017.
[3] K. Niwa, N. Harada, G. Zhang, W. B. W Kleijn, "Edge-consensus learning: deep learning on P2P networks with nonhomogeneous data," in *Proc. the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, pp. 668–678, 2020.
[4] National Institutes of Health (NIH) clinical center, ChestXray14 data set.

Contact

Kenta Niwa / Learning and Intelligent Systems Research Group, Innovative Communication Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp