Distributed learning algorithm and medical application

07 革新的な非同期分散型学習アルゴリズムと医療画像への応用

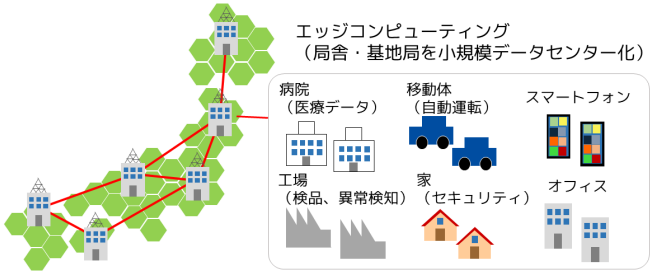
データを漏洩させない機械学習

どんな研究	１か所に集約したデータを使ってモデルを学習することが一般的です。しかし、データ量の激増やプライバシー保護の観点からデータ蓄積や学習／推論処理は分散化されるでしょう。データを各ノード（例：基地局）から外に出すことなく、機械学習モデルを学習する手法を提案します。
どこが凄い	分散蓄積されたデータは、統計的に偏っていると仮定することが自然です（例：一部クラスのデータが存在しない）。その状況で、ノード同士がモデル等の変数を非同期に交換（通信）しながら、全データを使って学習したかのようなグローバルモデルを得るアルゴリズムを開発しました。
めざす未来	地域／国／世界中のデータ全体を間接的に取り扱えるようにすることで、プライバシーを保護しながらも、高度な知を形成したり、高性能なサービス（例：医療）を提供できるようにしたいです。

目的・アプリケーション

背景：データ量の増大、プライバシー保護、法的規制（e.g. GDPR）の観点で、データの処理（学習／推論）が分散化される時代になる。

目的：分散蓄積されたデータを各ノード（e.g. 基地局）から外に出すことなく、深層学習モデルを学習したい。（ただし、モデルの更新差分等の補助情報をノード間で交換することを許容。）



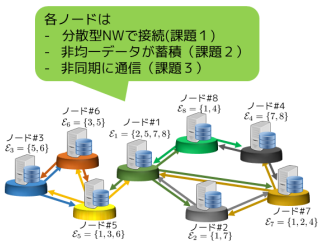
研究の課題（難しいポイント）

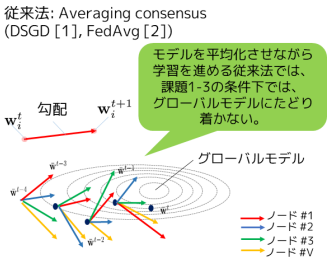
目的: 全部のデータを一か所に集約し、一か所で学習したモデルと同等のもの（グローバルモデル）を得たい。

課題１（分散、P2P型）サービス規模を任意のスケールで拡張するために、任意のネットワーク（NW）構造を利用できるようにしたい。

課題２（非均一データ）統計的に偏ったデータが各ノードに蓄積されても、安定して学習したい。

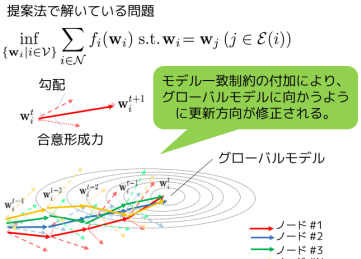
課題３（非同期通信）ＮＷ全体が同期して通信しなくても、安定して学習したい。

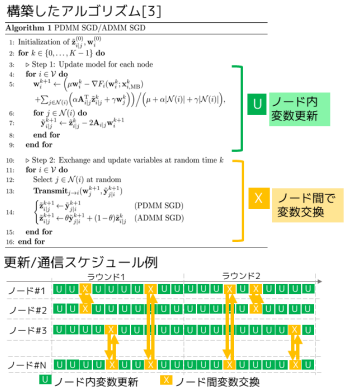




非同期分散型深層学習アルゴリズム

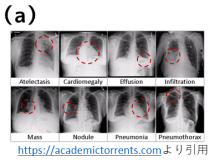
提案方式：制約付最小化問題を変形して再帰的な学習則を導出。その学習則では、(U) ノード内で変数（モデルとその補助変数）を更新するステップと（X）ノード間で非同期に通信（補助変数を交換）を交換するステップを交互に繰り返す。（任意のネットワーク構造で非同期通信で動作）

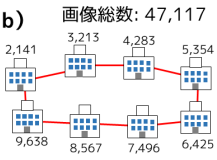




医療画像解析応用

胸部Ｘ線画像[4]から１４種類の疾患（下図**a**）を診断するモデルを生成。ただし、８箇所の病院（ノード）からデータを外に出さない。



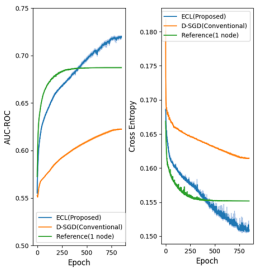


（課題１）分散ＮＷの一つとして、リングＮＷを利用

（課題２）各ノードに配分したデータ数が大きく異なるため、統計的性質に偏りが存在（左図**b**）

（課題３）更新（U）8回に対して通信（X）1回の割合

(実験結果)



- 従来法（橙）は各ノードがバラバラな方向に学習を進めようとするため、グローバルモデルにたどり着かなかった。
- 提案法（青）は、課題１－３の状況でも学習が進み、１か所で学習したグローバルモデル（緑）に近い性能を得た。
- 一部疾患に対する検知レベルが高い。実用性あり（肺気腫、気胸、心肥大、胸水等のAUC-ROCが0.75以上）。

関連文献

[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.

連絡先

丹羽 健太 （Kenta Niwa）協創情報研究部 知能創発環境研究グループ

Email: cs-openhouse-ml@hco.ntt.co.jp