

研究終了報告書

「機械学習する IoT 通信ネットワーク基盤」

研究期間:2020 年 11 月～2024 年 3 月

研究者:西尾 理志

1. 研究のねらい

IoT では、パソコンやスマートフォンのみならず、ロボットや自動車、無数のセンサ端末などあらゆるものが通信ネットワーク(以下、NW)で接続され、見守り、健康管理、生産自動化など我々の生活を豊かにする様々なアプリケーションが登場し始めている。さらに将来的には、自動運転車やドローンに搭載された多種多様なセンサからリアルタイムに情報収集し、人工知能技術(AI)を駆使し解析・活用することで、これまでにない新たな価値が想像されることが期待される。

AI による IoT データ処理は、クラウドコンピューティングにより、データを集約し集中的に処理する方式が主流である。一方で、IoT 端末の台数は 2030 年には 500 億台にも達するとされ、IoT 端末数増加に伴う爆発的な IoT データトラフィック増加への対応が急務である。これまでに圧縮によるデータ量削減や、クラウドではなく通信 NW の端(エッジ)で計算処理を行うエッジコンピューティングが検討されているが、圧縮・解凍処理に膨大な計算を要しリアルタイム性が損なわれたり、エッジサーバでの計算負荷や通信負荷が増加しボトルネックになったりと、従来技術だけでは、AI による次世代の IoT 実現は困難である。加えて、IoT 端末が人々の生活や企業の生産活動のあらゆる場面に導入されるに従い、データの機密性に配慮する必要性が益々増加する。データの機密性を保護する方法として暗号化や匿名化があるが、暗号化/匿名化処理によりリアルタイム性が損なわれたり、また処理を行っても収集・集約されたデータから個人情報や機密情報が推定されたりする可能性がある。

本研究では、この IoT-AI トラフィックの爆発的増加とデータプライバシーの 2 つの課題を解決するため、IoT データの地産地消を実現する通信と AI データ処理が一体化した IoT 基盤構築を目指す。IoT 端末により生成されるデータは、主にその端末が稼働する場所に関連したアプリケーションに用いられることが想定される。例えば高齢者や子供の見守りでは、その場に設置された多数の IoT 端末が対象者をセンシングし、データを解析し、その結果によって自治体や近隣の人々が行動する。このように、データの生産地と活用される場所は密接している。そこで、クラウドを介さずに IoT データを収集するローカル NW 内で分散的にデータ処理することで、コア NW のトラフィック削減、リアルタイム性向上および、機密情報流出の危険性を大きく低減することを目指す。

2. 研究成果

(1)概要

本研究は AI モデルの分散的学習と、学習した AI モデルを用いた分散推論処理の大きく 2 つのテーマに分けられる。両テーマに共通した目標として、省通信トラフィック、低遅延、データプライバシー保護をかね、目標達成のための様々な技術開発を実施した。

分散学習においては、Federated Learning (FL)に着目し、その性能を改善することで目標を達成した。FL はデータをローカルストレージに保持したまま共有等を行わずともそのデータを用いた学習を可能にする訓練フレームワークであり、データプライバシー保護可能な機械学

習として近年非常に大きな注目を集めている。本研究では、FL の通信トラフィックを大幅に削減する新たな学習手法を確立し、FL の通信トラフィックを 99%程度削減できることを示した[代表的論文 1]。また、本技術はデータに偏りがある場合 (non-IID) やラベルに誤りがある (noisy label) においても推論精度の低下を抑制できる画期的な手法である。遅延削減に向けては、端末選択手法 (Client Selection) 技術を改良し、訓練における計算及び通信遅延を削減できることを示した。また、学習処理の完全分散化技術についても研究し、中央サーバがなくとも、ノード間通信のみを介して中央サーバのアシストがある場合と同程度の性能をもつモデル訓練できる手法を確立し、従来の分散 FL と比べて通信量を 20%から条件によっては 99%以上削減可能であることを示した。加えて、画像生成モデルや IoT マルウェア検出タスクへの応用可能性を検証し、タスクに応じてより効率的にモデル訓練する手法を確立した。

分散推論においては、Model chaining と名付けた学習済みモデルを分割・共有・分散処理する手法(現在では Split computing という名称で同様の技術が広まりつつある)を用いて、分散化とデータプライバシー保護の実現を目指した。本プロジェクトでは通信効率の改善を重視し、通信路特性に適した特性をモデルに持たせる COMTune という手法を提案した。特に無線ネットワークにおけるパケット損失に着目し、モデルのパケット損失耐性を向上させる訓練方法を確立し、パケット損失によりデータの一部が欠損した状況においても正しい推論ができる確率を向上した。また、横国大 白川先生との共同研究により、端末の計算性能や通信性能、データ分布に応じて、モデルの構造やパラメータをチューニングするフレームワークを確立し、推論精度と遅延のトレードオフを改善した。

これらの成果は、計 18 件(投稿準備中・投稿中を含む)の国際論文誌や国際学会予稿で発表され、合計 240 件以上の被引用を達成するなど、注目を集めつつある。加えて、実用化に向けた取り組みとして、Split computing framework を実装し、物体検出や生成系 AI への適用可能性を示すとともに、オープンソースとして公開し、実験や PoC に利用可能とした。また、学部発表や CEATEC への出展を通して知り合った、国内外の研究者や企業と連携し、基盤 A や JST ASPIRE など大型競争的資金を獲得し共同研究を実施し、学術および産業界での連携を拡大している。

(2) 詳細

A) 分散学習

A-1) 省トラフィック連合機械学習手法 DS-FL

本研究では、従来のモデル更新情報の交換ではなく、ロジットと呼ばれるモデルの出力情報のみ交換し学習する Distillation-based Semi-supervised Federated Learning (DS-FL) を提案し、従来の 100 分の1程度の通信量での分散学習を実現した。ロジットは更新情報と比べて 100~1000 分の 1 程度のデータ容量となるため、狭帯域な IoT ネットワークでも利用可能となる。また、ラベルのないデータセットとロジットを活用することで、データ分布に偏り

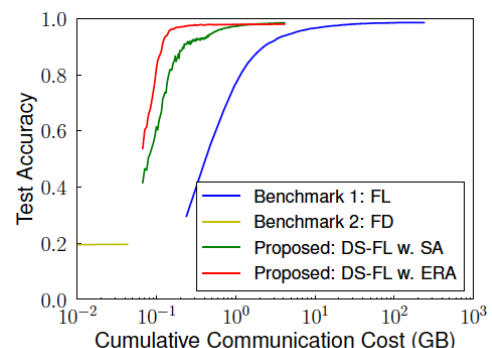


図 A-1: 通信量に対する推論精度

がある場合においても、高性能なモデルの学習を可能とした。本方式について、Fashion MNIST や IMDb など画像や文章分類タスクを用いて、学習に必要なトラフィック量とテスト性能を評価した。図 A-1 に MNIST データセットでの評価結果を示す。提案手法は従来の 100 分の 1 程度のトラフィック量で、既存手法と同程度の精度で学習できることを示した。本成果は難関論文誌である IEEE Trans. Mobile Computing に掲載され、掲載から現在に至るまで月間 Popular Article 上位 10 件にランクインし続けている他、200 件以上の被引用など高い注目を集めている。

A-2) Client Selection による遅延削減手法

本研究は研究者の代表的研究である Client Selection 手法を改良したものであり、多腕バンディットの応用により計算資源や通信資源が変動する環境においても学習にかかる計算と通信の遅延を低減する手法である。多腕バンディットの Reward 設計を工夫することで収束性を高め、総遅延時間を改善している。

A-3) CMFD: 通信効率の良い完全分散型 Federated Learning

本研究成果は東大 田谷昭仁氏および京都工繊大 山本高至氏との共同研究成果であり DS-FL を分散型に拡張し、分散合意制御の観点から理論的解釈をあたえたものである。本研究は、ネットワーク上で隣接する端末同士が協調し並列分散的にモデルを更新するためのアルゴリズムを提案し、任意のトポロジ、任意のモデル、偏りのあるデータなど悪条件においても、従来手法より高い精度のモデルの学習を実現した。本研究について、様々なタスクやネットワークトポロジについて追加実験を実施し、提案手法の優位性の一般性を示した。本成果は IEEE Transactions on Signal and Information Processing over Networks に掲載された。

A-4) Tram-FL: 省エネな完全分散学習手法

本研究は、連合機械学習における通信や計算により消費電力を削減可能な新たな訓練手法である。通常の分散連合機械学習では、M 台の端末すべてがモデルを更新し、更新したモデルを隣接端末と交換して統合するが、システム全体で端末数 M に比例した電力を消

モデル巡回型分散FL



費する。システム全体で端末数 M に比例した電力を消費する。本研究では、 $N(<< M)$ 個のモデルを端末間に流通させ、モデルを受け取った端末のみが更新や統合を行う。特に $N=1$ の場合、更新が直列的になるため処理時間は増大するが、ある時刻において計算・通信する端末が 1 台となり、消費電力を $1/M$ に削減できる。本成果の一部が国際会議 IEEE CCNC に採択されたほか、IEEE 論文誌に投稿している。

B) 分散推論

B-1) Split computing の通信路特性への適合手法 COMTune

本研究は、Split computing における推論性能低下を抑制しつつ、パケット損失耐性の向上と通信トラヒック削減を実現する。具体的には、学習済みモデルに対し、Dropout によるパケット損失のエミュレーション、および、非可逆圧縮によるデータ量削減を同時に施した上でモデルの再訓練を行い、実際の通信路上でパ

ケット損失やデータの非可逆圧縮を行なった場合でも高い推論精度を維持するようにモデルのチューニングを行う。本手法により通信量を 1/16 に削減しつつも 60%もパケットする環境においても数ポイントの推論精度に抑制することができた。本研究成果は国際会議 1 件、論文誌採録 1 件の成果をあげた。

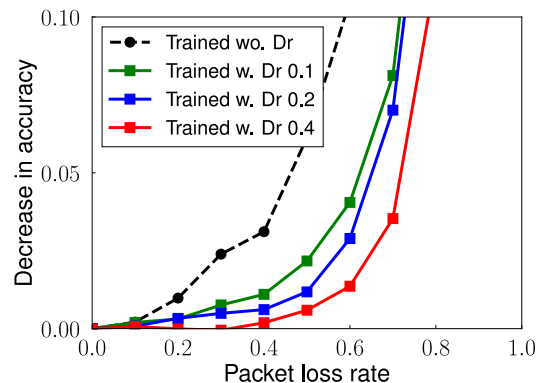


図 B-1: パケット損失率に対する推論性能

B-2) 通信と機械学習の同時制御

本研究は“通信性能と機械学習性能のトレードオフ”に着目し、通信側もあわせて制御することでこのトレードオフを改善する。具体的には、Dropout rate や非可逆圧縮率など機械学習側のパラメータと、伝送レートや再送制限など無線 LAN パラメータを同時制御する MWJC (ML-WLAN Joint Control)を提案した。この同時制御問題では、通信と機械学習の間で生じる複雑な相互作用から、定式化し陽に最適解を求めることが困難である。そのため、強化学習、特に多腕バンディットアルゴリズムにより、学習的に最適な制御パラメータを探索する。提案手法により、片方だけ制御する場合と比べて遅延時間を削減できることを NS3 によるネットワークシミュレーションにより示した。

B-3) Split computing に向けたモデル構造探索 NSAC

本研究も“通信性能と機械学習性能のトレードオフ”に着目し、Model chaining あらため Split computing により適したモデル構造を探索する問題に取り組んだ。横国大白川先生との共同研究で実施している。Neural Architecture Search の応用により、高い推論精度を達成しつつ、低遅延・少ない通信量・損失耐性のあるモデルを探索する。提案手法では、従来は考慮されない通信遅延やデバイスでの計算遅延を考慮した目的関数に従いモデルを探索することで、同程度の推論精度だが計算処理と通信の合計遅延を一定以下にする構造を探索可能であることを示した。本成果は国際会議 IEEE Globecom の Workshop で発表している。

B-4) Split computing に向けたモデルチューニングの分散学習手法

本研究は、Split computing における端末の計算性能・通信性能・データ分布にあわせた専用モデルを効率的に訓練するための手法である。Split computing において、モデルのカスタ

ムは通信トラフィック削減や推論性能向上には必須であるが、現在は従来の集中的な学習手法のみが検討されている。本研究では、これまで検討してきた分散学習を応用することで、データを収集・共有することなく学習も可能とする。特に、学習済みモデルを活用することで端末の有するラベルなしデータのみを用いてモデル訓練する手法を提案し、データ集約した場合と同程度の制度のモデルを訓練可能であることを示した。

C) 実用化

C-1) Split computing framework の実装と公開

Split computing framework を実装し、物体検出や生成系 AI への適用可能性を示すとともに、オープンソースとして公開し、実験や PoC に利用可能とした。通信エミュレーションによりパケット損失率や遅延を任意に設定できる機能を搭載しているほか、UI を備え、使用するモデルや実験設定を調整しながら性能テストが可能である。また、送信する中間層出力の可視化機能も有し、どのようなデータがネットワークを流れていくかを直接観測できる。

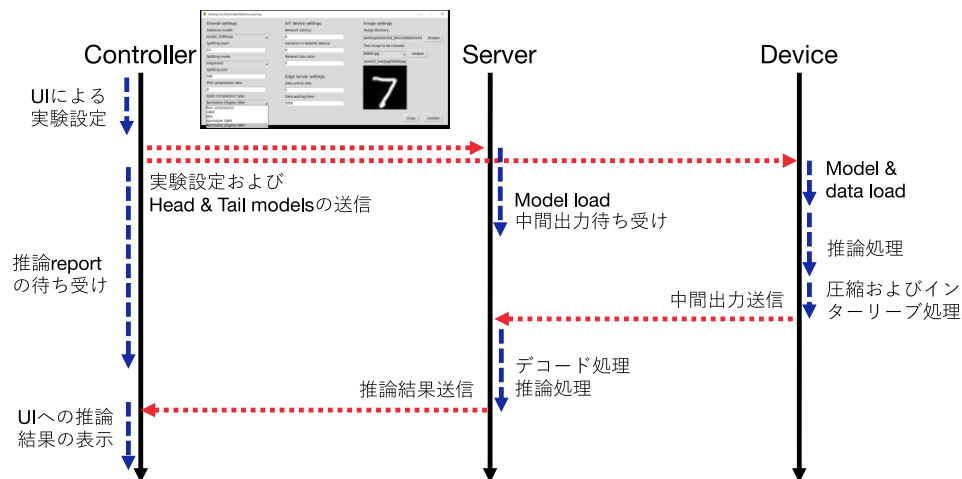


図 C-1: Split Computing Framework

C-2) Λ -Split: 生成系 AI の分散推論の実証

本研究では、 Λ -Split という 3 分割による入出力を秘匿する推論手法の提案と生成系 AI 特に LLM を用いたテキスト生成と Diffusion model による画像生成において、3 分割分散推論の実現可能性を示している。LLM においては埋込ベクトル、Diffusion model においては推定ノイズのみがネットワークを介して送信されるため、生成モデルへの入力と生成モデルの出力を秘匿することができる。また、トラフィック削減手法を提案し、単に分割した場合と比べて、90%以上の通信量削減が可能であることを実験的に示した。

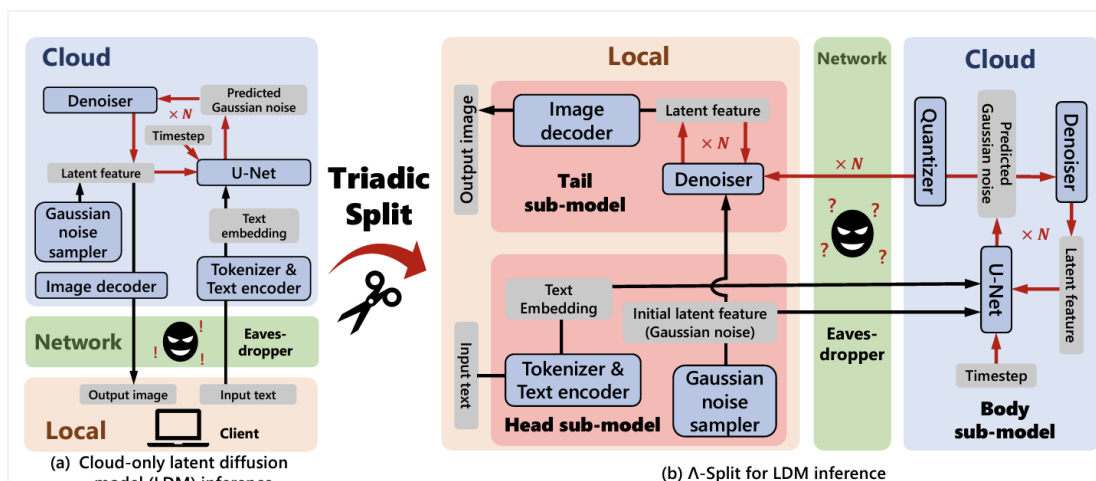


図 C-2: Λ -Split. Diffusion model の場合。

C-3) Federated Learning による IoT マルウェア検出向上

本研究では、分散学習 Federated Learning の応用先の一つとして、教師なし異常検知に取り組んだ。異常検知は IoT の最も一般的なタスクの一つである。例えば、工場の製造機械に取り付けた IoT センサによる加速度や音の観測値から製造機械の異常動作を検出したり、IoT 機器に感染したマルウェアをトラヒックパターンから検出したりといった用途が想定される。特に教師なし異常検知は、教師信号が不要であることから、訓練データを端末で準備する必要がある FL とタスクとしても相性が良い。本研究では特にマルウェア検出を対象に、「通常、FL はデータを集約して学習する場合よりもモデルの性能が悪くなるが、特定の条件の異常検出タスクでは FL のほうが性能がよくなる」という仮説を検証し実証した。本研究は KDDI 研究所との共同研究で実施し、本成果は国際会議 IEEE Globecom で発表し、今後論文投稿予定である。

D) 連携

本さがけプロジェクトを通して、国内大学の研究者 13 名、海外研究者 4 名、3 企業と連携し、共同研究や共著論文執筆、競争的資金申請、Tutorial 実施、Workshop 開催申請を行った。未だ実っていないものもあるが、今後も継続していく。

特に代表的な例をあげる。

- 横国大 白川先生と連携し、Split computing に特化した NAS 技術や Dynamic Split Computing の研究を実施している
- ロボティクス・自動運転等の分野において埼玉大 安積先生を筆頭に、東工大 原先生、芝浦工大 新熊先生と基盤 A を獲得し、アプリケーションを意識した分散機械学習の研究を強化していく
- 電通大 須藤先生(代表)、佐藤先生と共同応募した JST ASPIRE が採択された。海外研究者として Mehdi Bennis 教授と連携し、通信分野における機械学習人材の育成と知の循環を目指す。
- Deakin University Jihong Park 先生とは 2021 年に共同チュートリアルを実施した他、2023 年には東工大に来訪しセミナーを実施した。

3. 今後の展開

機械学習の発展およびIoTの展開とともに、分散学習・分散推論技術の重要性は増大する一方であり、本さがけ研究をベースに、国内外の研究者や企業との共同研究を通してさらに発展させていく。技術的に Federated Learning や Split Computing はすでに社会実装も可能な技術であり、いくつもの実証実験が報告されている。一方で、アプリケーションやサービスごとにそれぞれの制約を考慮したアルゴリズム改良は必要であり、学術的に解決していく必要がある。これは成果 D に記載した大学・起業との共同研究により解決していく。また、データの秘匿性に関しては技術課題が多く残る。特に Split Computing においては、入力データが比較的容易に復元可能であることが報じられており、この Data leakage 問題の解決は不可欠である。

4. 自己評価

主たる研究目標である「データ秘匿した分散学習・推論において通信トラフィック 1/100 の達成」は、学習においては DS-FL 単独でも達成しており、さらに研究成果を組み合わせることでさらなる削減も可能である。また、推論においても、Head network のボトルネック挿入と圧縮の組み合わせにより達成可能である。一方で、当初想定していなかった課題として、層数の多い CNN モデルを浅い層で分割した時に入力データと酷似した中間層出力が送信されプライバシー保護が困難であるという課題が見つかり、これらの課題を解決する研究を追加で実施したものの、初期検討にとどまっている。また、テストベッドを構築するという目標は、ソフトウェアに関してはある程度の実装ができ、ハードウェアは多数のセンサと計算・通信デバイスを配備したテスト環境を整備したものの研究室の一室程度の規模に留まっている。一方で、小規模ながら実装システムを展示会出展するなど社会実装に向けた取り組みの他、さらなる発展を目指し産学連携を進めている。

また、科学技術及び社会・経済への波及効果としては、特に知識蒸留の応用や新たな学習フレームワークの確立において、国内外の機械学習コミュニティへ一定の貢献があるほか、国内の特に通信分野において、Federated Learning や Split Computing の技術普及において、多数のチュートリアル講演や解説記事を執筆するなど大きく貢献した。これらの研究領域内外での活動を通して、多数の共同研究や連携の機会を得ることができた。

よって、研究目標に関して概ね当初予定通りである一方、研究領域内および研究領域間で異分野の研究者ネットワークの形成に関しては期待以上の成果をあげていると自己評価する。

5. 主な研究成果リスト

(1) 代表的な論文(原著論文)発表

研究期間累積件数： 18件

1. S. Itahara, T. Nishio, Y. Koda, M. Morikura, and K. Yamamoto, "Distillation-Based Semi-Supervised Federated Learning for Communication-Efficient Collaborative Training with Non-IID Private Data," IEEE Transactions on Mobile Computing, vol. 22, no. 1, pp. 191-205, Jan. 2023.

分散的機械学習フレームワークである Federated Learning について、従来のモデル交換に基づく訓練手法に代わる、知識蒸留を応用した新たなモデル訓練方法を提案し、通信トラフィックを 98~99%削減できることを示した。加えて、データに偏りがある場合 (non-IID) やラベルに誤り

がある (noisy label) においても推論精度の低下を抑制できることを示した。
2. S. Itahara, T. Nishio, Y. Koda, and K. Yamamoto, "Communication-oriented Model Fine-tuning for Packet-loss Resilient Distributed Inference under Highly Lossy IoT Networks," IEEE Access, vol. 10, pp. 14969-14979, Feb. 2022.
パケット損失が頻発する無線 IoT ネットワークに向けた分散機械学習推論のパケット損失耐性の向上手法を提案した。提案手法により、パケット損失によるデータ欠損が発生した場合でも、正しい推論結果が得られる確率を高めることができ、70%の欠損が発生する場合でも 2 ポイント程度の精度低下で分散推論できることを示した。
3. S. Ohta and T. Nishio, "Λ-Split: A Privacy-Preserving Split Computing Framework for Cloud-Powered Generative AI," submitted to IEEE Transactions, June 2024.
複雑な構造を持つ生成系 AI モデルについて、分割可能性およびプライバシー保護性を実験的に示した。特に、Large Language Model (LLM)によるテキスト生成と Diffusion Model による画像生成モデルについて、具体的な分割方策を示し、処理の途中結果(埋込ベクトル、推定ノイズ)のみを通信することで、入出力の両方を秘匿してコンテンツ生成できることを示した。

(2)特許出願

研究期間全出願件数： 0 件(特許公開前のものも含む)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

本研究プロジェクトに関わる研究成果や発表に対し、

- 電子情報通信学会 2022 年 末松安晴賞授賞
を受賞した。また、本プロジェクトの研究補助学生の成果発表に対し
 - 2021 年度 電気通信普及財団賞 テレコムシステム技術学生賞, 板原壮平
 - 2021 年度 IEEE 関西支部学生研究奨励賞受賞, 板原壮平
 - 2022 年度 電子情報通信学会 SeMI 研究会 優秀発表賞, 依田光仁
- をそれぞれ受賞した。

加えて、Elsevier が毎年発表する top-cited scientists database において、Single Year Top2% Scientist に 2021 年から 2 年連続ランクインしている。

Elsevier Data Repository, <https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/6>