戦略的創造研究推進事業　CREST
研究領域「イノベーション創発に資する人工知能基盤技術の創出と統合化」
研究課題「リアルタイム性と全データ性を両立するエッジ学習基盤」

研究終了報告書

研究期間　2017年10月～2020年3月

研究代表者：松谷　宏紀
（慶應義塾大学理工学部、准教授）

§1 研究実施の概要

(1)実施概要

一般的なニューラルネットワークの学習では誤差逆伝搬法と確率的勾配降下法などによって重みパラメータを反復的に最適化する。学習に要する計算コストは非常に高く、GPU(Graphics Processing Unit)による並列計算が有効である。このため、工場やプラント、データセンタなどの実環境で異常検知等の推論処理を実行する場合、1) 学習に用いる教師データを現場で収集し、2) 集めた教師データを高性能計算機に移したうえで学習処理を行い、3) 学習結果である推論器を現場(エッジ環境)にデプロイするという手順を踏む。このアプローチの問題点は、現場の状況が変わるたびに上記の 1) ～ 3) をやり直す必要がある点である。例えば、製造ラインにおいては、他の装置の稼働状況や工具の摩耗、工場のレイアウト変更によってセンサの値の出方、つまり、正常パターンやノイズパターンが変動する。材料の材質や形状の違いによってもセンサの値の出方は異なってくる。今まで正常扱いしていたパターンを途中から「異常」として検出したくなることもある。このような状況変化のたびに上記の 1) ～ 3) を再実行するのはコストや手間が大きい。

このような問題を解決するために、本研究ではニューラルネットワークを用いたオンデバイス学習技術を確立した。オンデバイス学習では、学習と推論を同じ計算機上、とりわけ計算資源の限られたエッジデバイス上で行う。オンデバイス学習を可能にするにはニューラルネットワークの学習コストを大幅に削減する必要があった。本提案ではオンライン逐次学習アルゴリズムをもとに、行列演算のボトルネック部分を簡易計算化したうえで、これにともなう学習の不安定さを改善している。このような「エッジでのニューラルネットワークの学習」は教師無し異常検知、もしくは半教師有り異常検知との相性が良い。本研究でも次元圧縮アルゴリズムと組み合わせて異常検知を実現している。この場合、上述の製造ラインの例であれば、状況が変化するたびにオンデバイス学習器の「学習ボタン」を押すことで、現場だけで追加学習もしくは再学習ができる。追加学習に要する時間は、パラメータ依存ではあるが、かなり短く抑えることができている。例えば、CEATEC 2019 にて披露したサーモグラフィを用いた電源の異常検知では、入力次元数を 4800 としたき再学習に要する時間は数十秒から一分弱であった。

しかし、これだけでは個々のエッジデバイスに集まる学習データの質と量には限界があり、また、ラベル付き教師データを前提としにくいため、異常が検知された後の高度な要因推定に課題が残る。そこで、エッジ AI とクラウド AI を組み合わせようというのも本研究の提案である。具体的には、個々のエッジデバイスで異常と判断されたデータのみをサーバに転送し、集まった異常データに対して異常の要因等をラベルとして付与し、深層学習フレームワークを用いて学習するフローを構築した。エッジデバイスで異常が検出されると、そのデータをサーバに転送し、上記の深層学習フレームワークを用いて推論することで異常の要因を推定できる。すでに回転機械を対象とした異常の要因推定で効果を発揮している。

オンデバイス学習のアルゴリズム部分は慶應義塾大学グループと東京大学グループが共同で開発した。具体的には、オンライン逐次学習アルゴリズムの軽量化、次元圧縮技術との組み合わせ、逐次学習の安定化手法、コンセプトドリフトに追従するための忘却手法、複数定常点に対処するためのアンサンブル手法などを提案している。当初は固定長の入力データを前提としていたが、カメ

ラ映像による人の異常行動検出の際に、可変長の入力データに対処する手法を開発した。

実装については慶應義塾大学グループが中心となって、安価に入手できる Raspberry Pi Zero や小規模 FPGA で動作するプロトタイプを開発した。クラウド側については東京大学グループが中心となって開発した。これには深層学習フレームワークと組み合わせた要因推定機能、クラウド上で動作するオンデバイス学習器のインスタンス等が含まれる。

フィックスターズグループはオンデバイス学習を用いた異常検知のためのソフトウェアフレームワークの構築、フィールドテストに使用する実験機の開発を行った。また、スマートインダストリーを対象にオンデバイス学習の実証実験を行った。実証実験では、予め現場で収集しておいた実データを用いた予備評価から、現場に実験機を設置して稼働させるところまで実施している。本研究の成果は JST 新技術説明会や CEATEC 2019 などでも積極的に展開し、高い評価を得ている。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1. 概要:

「エッジ AI」と呼ばれる技術の多くはエッジデバイスによる推論処理をターゲットにしており、学習は行わない。学習まで行う場合であっても対象がスマートフォン等の比較的高性能なデバイス、もしくは、組込み GPU の利用を想定しているケースが多い。一方、本研究では数千次元にも及ぶ入力データを前提に、ニューラルネットワークの学習を低コストな組込み CPU で実現している。小規模 FPGA(Field-Programmable Gate Array)を用いた試作にも成功している。

2. 概要:

一般的なエッジ AI では学習する環境とデプロイする環境が異なり、再学習も容易ではないため汎化性能が高くないと使い物にならず、高度なニューラルネットワークが使われる傾向にあった。一方、我々のオンデバイス学習では、学習する環境とデプロイする環境が同一である。しかも、逐次学習によってコンセプトドリフトにも追従でき、複数インスタンス化によるアンサンブル手法によって表現能力を高めている。これによってシンプルなニューラルネットワークを使っている割に応用範囲が広い。

3. 概要:

近年、分散深層学習の枠組みにスマートフォン等のモバイル端末を参加させるフェデレーションが注目を浴びている。とくにオンデバイス学習の場合、エッジデバイスに集まる学習データの量に限界があるため、ドメイン内の他所の学習結果の利活用が重要である。本研究では各エッジデバイスが計算した中間重み(生データではない)を一箇所に集め、必要に応じて分配する仕組みの理論的な検討まで行った。また、複数エージェントによる分散強化学習の効率的実装も検討した。これらが実現できればエッジ AI 全般の欠点を克服できる。

＜科学技術イノベーションに大きく寄与する成果＞

1．概要：

エッジ環境で AI を利用する場合、現場の状況が変わるたびに 1) 教師データを現場で収集し、2) 集めた教師データを高性能計算機に移したうえで学習を行い、3) 学習結果である推論器を現場にデプロイするという手順を踏む必要があり、これが現場に AI を導入する際の課題になっている。本研究では User-in-the-loop のアプローチでこの問題を解決している。必要に応じてオンデバイス学習器の「学習ボタン」を押すことで、エッジによる追加学習もしくは再学習を短時間で実現できる。

2．概要：

研究成果は 2019 年 9 月の JST 新技術説明会、10 月の CEATEC 2019 での展示など積極的に発信している。反響は我々の予想を遥かに上回るものであった。同 10 月にはデータセンタでの異常検知に関するプレスリリース、工場等での異常検知に関する新聞掲載があり、多くの事後問い合わせをいただいた。ご期待に応えるため、オンデバイス学習のお試し版（クラウド版、Raspberry Pi 版）を準備し、順次公開している。

3．概要：

本プロジェクトでは、当初より商用化を前提とした活動をし、工場等を対象としたオンデバイス学習による異常検知の実証実験の機会にも恵まれた。具体的には、予め現場で収集しておいた実データを用いた予備評価から、実際に実験機を現場に設置して稼働させるところまで実施できている。このような連携の枠組みがあったからこそ、実データを用いた、もしくは現場での実証実験を経験でき、提案手法の改善につなげることができた。

＜代表的な論文＞

[1] Mineto Tsukada, Masaaki Kondo, Hiroki Matsutani, ″A Neural Network Based On-device Learning Anomaly Detector for Edge Devices″, arXiv:1907.10147, July 23, 2019.

[2] Shaswot Shresthamali, Masaaki Kondo, Hiroshi Nakamura, ″Power Management of Wireless Sensor Nodes with Coordinated Distributed Reinforcement Learning″, Proc. of the 37th IEEE International Conference on Computer Design (ICCD'19), Nov 2019.

[3] Tomoya Itsubo, Mineto Tsukada, Hiroki Matsutani, ″Performance and Cost Evaluations of Online Sequential Learning and Unsupervised Anomaly Detection Core″, Proc. of the 22nd IEEE Symposium on Low-Power and High-Speed Chips and Systems (COOL Chips 22), Apr 2019.

§2 研究実施体制

(1)研究チームの体制について

① 「慶應義塾大学」グループ

研究代表者:松谷 宏紀 (慶應義塾大学理工学部 准教授)

【研究項目 A】(1) FPGA を利用したエッジ上での高速リアルタイム学習基盤

【研究項目 A】(3) オンライン学習を効率化するエッジ上での転移学習およびマルチタスク学習基盤

【研究項目 B】(6) 高速並列分散深層学習基盤

【研究項目 C】(7) 応用を意識した要件検討

② 「東京大学」グループ

主たる共同研究者:近藤 正章 (東京大学大学院情報理工学系研究科 准教授)

【研究項目 A】(2) Realtime Knowledge と Deep Knowledge を活用する高効率ニューラルネットワーク認識基盤

【研究項目 B】(4) エッジの学習の監視と学習結果のリカバリ

【研究項目 B】(5) エッジ上での転移学習およびマルチタスク学習のサポート

【研究項目 C】(7) 応用を意識した要件検討

③ 「フィックスターズ」グループ

主たる共同研究者:塩田 靖彦 ((株)フィックスターズ 執行役員)

【研究項目 C】(7) 応用を意識した要件検討

【研究項目 C】(8) エッジコンピューティングサーバ Olive をベースにした統合環境の構築準備

(2)国内外の研究者や産業界等との連携によるネットワーク形成の状況について

産業界と密に連携しながら実証実験を進めた。詳細は§4(非公開)に記す。本研究では合計14件の基調講演もしくは招待講演を行った。そのうち7件は国際会議での講演である。詳細は別紙の業績一覧に記す。さらに、JSTおよびDFKIの主催でドイツ人工知能研究センタにて開催されたInternational Workshop on Intelligence Augmentation and Amplification (IAA Workshop 2019)にも参加し、欧州系の研究者とのネットワークを構築した。