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「社会インフラ映像処理のための
高速・省資源深層学習アルゴリズム基盤」

研究終了報告書

研究期間　2016年12月～2019年　3月

研究代表者：篠田　浩一
（東京工業大学情報理工学院　教授）

§1 研究実施の概要

(1)実施概要

安心・安全なスマート社会の実現のために、ドライブレコーダーの映像や監視カメラの映像を用いて事故や犯罪を防止するシステムの実現が望まれている。そのためには近年人工知能分野で急速に進展している深層学習技術を用いるアプローチが有望である。しかし、大量の高精細映像における小さい物体やその些細な動きを精度よく検出・解析するためにはまだ課題が多い。また、通信量の削減のために端末側で処理を行う必要があるが現状では深層ニューラルネットワークのサイズが大きく難しい。

この問題意識を受け、本申請課題では、大量の高精細(HD)映像から高性能かつ実時間の物体検出・異常検知を端末側で行うことを可能にする、高速・省メモリの深層学習・解析アルゴリズム基盤の構築を目標とした。機械学習と高性能計算の研究者が協力して、東工大の誇る大型計算機 TSUBAME を活用することにより、システムからアプリケーションまでの階層を一気通貫して同時に最適化する「Co-Design」のフレームワークのもと研究開発を行うことを大きな特色とする。スモールフェーズにおいては深層学習処理を4階層に区分し、4つの研究グループの各々が 1 つの階層を担当し、高速化・省コスト化を行った。

まず、横田 G は、「個々の計算ノードにおける計算量を削減するための行列構造化アルゴリズム」の研究を行った。特に、深層学習の二次最適化手法の一つである自然勾配法において、Fisher 情報行列の逆行列を Kronecker 因子分解を用いて高速に求める手法を提案した。ImageNet-1K, ResNet-50 を用いた学習において、従来の確率的最急降下法を用いた一次最適化に比べ 3 倍程度早く収束することが確認できた。

次に、松岡 G は、「ノード間の通信処理を削減するための高並列アルゴリズムと資源スケジューリングによる全体最適化」の研究を行った。特に、深層学習にしばしば用いられる畳み込み演算の分散並列処理を、メモリ・ノード数などの条件に応じて自動的に最適化するライブラリを開発した。1.2-1.6 倍の高速化を実現した。

続いて、篠田 G は、「知識の構造を活用した高速な深層学習アルゴリズム」の研究を行った。特に、映像における誕生日パーティーなどの「イベント」の検出において、ケーキや帽子などの各種オブジェクトの検出器とそれらの間の意味関係を用いることで、イベントの学習データがごく少量の場合でも頑健に検出する手法を提案した。NIST TRECVID ワークショップで世界 2 位の性能を達成した。

最後に、村田 G は、「リアルタイム認識・解析のための Deep Net 構造のコンパクト化アルゴリズム」の研究を行った。深層ニューラルネットワークにおいて重み係数の枝刈り、量子化、符号化を行うことにより、そのサイズを 90 分の 1 にまで小さくすることに成功した。

これらの成果を統合することにより、当初の目標の一つである 1000 倍の高速化を実現した。さらに、深層学習高速化の標準的なベンチマークである ImageNet の学習高速化において、現時点で、1024GPU のみを用いる条件下ではあるものの、世界 2 位を達成している。2018 年度末には産業総合研究所の大規模計算機 ABCI を用いてノード数を増やし、留保条件なしの世界最速を達成する予定である。

(2)顕著な成果

＜優れた基礎研究としての成果＞

1.

概要:

映像からイベント検出において、イベントを構成する概念間の関係を利用することで、学習データが少なくても安定して深層学習を行う手法を構築した。概念間の意味関係を用いてイベントの学習データが不要なゼロショット識別器を構築する。そして、それをごく少量の学習データを用いて学習した識別器と組み合わせる。NIST TRECVID ワークショップのマルチメディアイベント検出タスクで世界 2 位の性能を達成した。

2.

概要:

機械学習処理に頻出する畳み込み処理には計算量とメモリ要求量が異なる複数の実装があるが、1 つの畳み込み処理を適切なサイズに分割することで限られたメモリ量での計算速度の最大化を行うことができることを示し、cuDNN のラッパライブラリである μ-cuDNN を実装した。これにより DeepBench および Caffe の畳み込み演算を平均 1.60 倍高速化した。

3.

概要:

深層学習の大規模並列化において問題となる、バッチサイズの増大による汎化性能の低下を低減するため、2次の最適化手法を用いる方法を提案した。従来の2次の手法でボトルネックとなっていた Fisher 行列の計算を Kronecker 因子分解による近似法を用いて高速化した。また、様々な正則化手法を取り入れることで Fisher 行列の特異性を取り除き、過学習を抑えることに成功した。学習の中盤から Fisher 行列の更新頻度を下げる手法を採用し、ステップあたりの計算時間を1次の最適化手法とほぼ同等にまで低減した。

＜科学技術イノベーションに大きく寄与する成果＞

1.

概要:

画像認識のベンチマークである ImageNet の学習を ABCI スパコンのグランドチャレンジ制度を利用し 4096GPU を用いて行った。2次の最適化手法を様々な正則化手法と組み合わせることでバッチサイズの増大の問題を低減し、Kronecker 因子分解による近似法を用いて高速化を行った。集団通信アルゴリズムに ring と recursive halfing/doubling のハイブリッドな手法を考案した。その結果、学習精度を維持しながらも ImageNet の収束までの最短反復数の世界記録を達成することができた。

2.

概要:

モバイル端末等の限られたストーレッジおよび消費電力の環境下での使用に適したサイズに Deep Net を圧縮する研究を行った。従来手法をベースに、それを改良することで、Deep Net の

サイズ圧縮のベンチマークとして用いられる AlexNet を 1/90 のサイズに圧縮し、かつ圧縮後の学習精度も悪化していなかった。また特別なハードウェアを必要とすることなく圧縮を実現した。

3.
概要:
アプリケーションに特化した構造をもつ深層ニューラルネットワークを設計する方法論を開発た。より具体的には、本申請課題の応用として重要な人間の動作を認識する動作認識のための高性能な深層ニューラルネットワークを構築した。このモデルは、人間の四肢、胴体やその関係性をグラフで表現し、その関係性を保ったネットワークを構築することで従来よりも小さいサイズでより高い認識瀬能を実現している。

＜代表的な論文＞
[1] T. M. Le, N. Inoue, K. Shinoda, A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition, Proc. British Machine Vision Conference (BMVC), Sep. 3, 2018.
[2] Yosuke Oyama, Tal Ben-Nun, Torsten Hoefler, Satoshi Matsuoka, “Accelerating Deep Learning Frameworks with Micro-batches”, Cluster 2018, Belfast, UK, September 2018
[3] Arie Wahyu Wijayanto, Jun Jin Choong, Kaushalya Madhawa, Tsuyoshi Murata, ″Towards Robust Compressed Convolutional Neural Networks″, The 6th IEEE International Conference on Big Data and Smart Computing (IEEE BigComp 2019), pp.168-175, 2019.

§2 研究実施体制

(1)研究チームの体制について

① 篠田グループ
研究代表者:篠田 浩一(東京工業大学情報理工学院 教授)
研究項目 ・知識の構造を活用した高速な深層学習アルゴリズム

② 松岡グループ
主たる共同研究者:松岡 聡(東京工業大学情報理工学院 教授)
研究項目 ・ノード間の通信処理を削減するための高並列アルゴリズムと資源スケジューリングによる全体最適化

③ 村田グループ
主たる共同研究者:村田 剛志(東京工業大学情報理工学院 准教授)
研究項目 ・リアルタイム認識・解析のための Deep Net 構造のコンパクト化アルゴリズム

④ 横田グループ
主たる共同研究者:横田 理央(東京工業大学情報理工学院学術国際情報センター 准教授)
研究項目 ・個々の計算ノードにおける計算量を削減するための行列構造化アルゴリズム

(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

- シンガポールのNTUと映像検索において共同研究を行っている。米国国立標準技術局(NIST)主催の国際映像検索評価ワークショップの人物動作検索タスクにおいて共同チームを作り研究している。11月に米国メリーランド州で開催されたワークショップでその成果を発表した。
- 2018年度後半から同じくシンガポールのInstitute for Infocomm Research (I2R)における深層学習チーム(リーダー: Dr. Vijay Chandrasekhar氏)と共同研究を開始した。2018年10月より、学生の交換(インターンシップ)を開始した。主に大規模敵対的生成学習(Generative Adversarial Network, GAN)の分野で研究を進めている。
- JST CREST AI領域の別課題「自然言語処理による心の病の理解:未病で精神疾患を防ぐ」の研究代表者である岸本泰士郎講師と、AMEDの医療のデジタル革命実現プロジェクトの委託研究の形で、音響情報からの認知症診断のテーマで共同研究をしている。これは、人間の発声における音響情報のみを用いて認知症かどうか、あるいは、どの程度重症か、に関し、深層学習を用いて自動診断するものである。2018年9月に国際会議Interspeechで発表した成果[1]を応用した。音声を書き起こしたテキストを用いる場合よりも若干性能は低いものの90%近い判別性能を得ている。今度も共同研究を続けていく予定である。
- 2018年度後半から、同じJST CREST AI領域の別課題「脳波の機械判読によるてんかん診断・治療支援AIの研究代表者である田中聡久教授と、AIPネットワークラボ特別経費の支援を頂き、「てんかん患者の頭蓋内脳波を用いた言語活動のAIとデコーディングとマッピング」のテーマで共同研究を行っている。まだ開始したばかりであるが、主に音声モデルから脳波モデルへの転移学習や生成モデルの深層学習などの分野で貢献する予定である。
- 横田Gと松岡Gが共同で行った国内最大のAIスパコンであるABCIを占有するグランドチャレンジにおいて、2次の最適化手法を用いることでImageNetベンチマークの学習にかかる反復数を1/3に低減できることを明らかにした。これと並行して行われていたソニーによる同等のベンチマークにおける大規模実行では、学習時間の世界最短記録である3.7分を達成した。横田Gで開発している2次の最適化手法とソニーの開発しているバッチサイズを徐々に増大させる手法は大規模並列深層学習において相補的な技術であるため、現在共同研究の協議中である。