様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２１ 日現在

機関番号：３２６６０
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７７６８

研究課題名（和文）アルゴリズムと計算機アーキテクチャの協調による深層学習における学習の高速化基盤

研究課題名（英文）Acceleration framework for training deep learning by cooperative with algorithms and computer architectures

研究代表者
前田　慶博（Maeda, Yoshihiro）

東京理科大学・工学部電気工学科・講師

研究者番号：８０８４３３７５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究では，幅広い分野で活用されるDeep neural network（DNN）の学習における高速化について検討を行った．具体的には，DNNモデルの軽量化手法であるプルーニングや量子化と呼ばれるアルゴリズムに対して，計算機アーキテクチャの観点も踏まえた学習時にも活用できるDNN軽量化について検討を行った．検討の結果，プルーニングや量子化を基にした計算機アーキテクチャの恩恵をうけることが可能なアルゴリズムによってDNNの学習の高速化を実現できることを示した．

研究成果の学術的意義や社会的意義
本研究は，Deep neural network（DNN）の学習の高速化について検討をするものである．DNNは，画像処理分野では物体認識や超解像など様々なコンピュータビジョンにおけるタスクの更なる高精度化を実現している．DNNの活躍は画像処理分野だけにとどまらず，様々な研究領域での活用や産業界においても商用利用されている．本研究によって学習の高速化が実現でき，更なる活用の幅が広がるものである．

研究成果の概要（英文）：In this research, we aimed to accelerate the training process of deep neural networks (DNNs) used in various fields. We focused on DNN optimization techniques such as pruning and quantization, which help simplify DNN models. Considering computer architecture viewpoints, our study explored how these techniques can be applied during training. We found that pruning and quantization-based algorithms can accelerate the training process for DNN by leveraging computer architecture.

研究分野：画像処理

キーワード：深層学習　計算機アーキテクチャ　高能率計算　高速化

１．研究開始当初の背景

近年，Convolutional Neural Network の発展である Deep neural network（DNN）技術への期待が高まっている．DNN は，画像処理分野では物体認識や超解像など様々なコンピュータビジョンにおけるタスクの更なる高精度化を実現している．

DNN における学習と推論という二つの過程について，学習過程は多大な計算リソースとそれを用いても膨大な時間を必要とするという大きな問題を抱えている．DNN を活用した研究の場合， DNN のネットワークアーキテクチャや学習率などのハイパーパラメータを変更すると，再度の学習を必要とする．そのため，微少な変更に対しても多大な時間を必要とする．特に近年のネットワークモデルは非常に多くのハイパーパラメータを有しており，学習に要する時間は増大する一方である．これは，DNN を活用した研究のトライアンドエラーの回数に大きな制約を与えてしまう．つまり，膨大な学習に要する時間が DNN の更なる発展を大きく阻害している．そのため，DNN における学習過程の高速化が強く求められている．

２．研究の目的

本研究では，DNN における学習過程の高速化を目的として，DNN を実現するアルゴリズムを計算機アーキテクチャの観点も考慮して検討を行う．具体的には，DNN モデルの軽量化手法であるプルーニングや量子化等のアルゴリズムや DNN に入力するための前処理に関して，計算機アーキテクチャによる高速化の恩恵をうけることが可能なアルゴリズムの検討を実施する．そして，アルゴリズムと計算機アーキテクチャを協調させることによる学習の高速化方法を学術的に明らかにする．

３．研究の方法

本研究では，DNN における学習過程の高速化を目的として，DNN を実現するアルゴリズムを計算機アーキテクチャの観点より検討を行う．

（１）プルーニングによる DNN モデルの軽量化

プルーニングは，DNN モデルの軽量化を実現する技術であり，DNN モデルの軽量化は，学習の高速化に繋がる．この項目では，これに焦点をあて，finetune などの再学習時における DNN モデルの軽量化や学習前にプルーニングを行うことによる DNN モデルの軽量化について検討する．

（２）計算機アーキテクチャを考慮した量子化による DNN アーキテクチャの軽量化

量子化は DNN 内で扱う数値の表現範囲，すなわち bit 長をより短いものに置き換える手法である．これでは数値の bit 長が短くなるため，DNN モデルの軽量化そして学習の高速化に繋がる．

４．研究成果

（１）再学習における学習済み重みをプルーニングによる DNN 学習の高速化

PCA による最適構造の探索とスパース性に基づくプルーニングを組み合わせた CNN モデルの最適化手法を提案した．図 1 に提案手法の概要を示す．提案手法では，まず，主成分分析（Principal component analysis: PCA）によって最適なチャネル数を決定する．その後，学習済みの重みを用いて，チャネルごとに L1 ノルムの大きさに基づいた重要度を計算しプルーニングを行う．これにより，少ない再学習で精度を保持した軽量なモデル構築を実現した．

2 種類のデータセットを用いた物体認識タスクにより提案手法の有効性を検証した．対象モデルは vgg16_bn を使用した．比較手法として，学習後にプルーニングを行う手法を用いた．再学習のエポック数は，提案手法では 20 エポック，比較手法では 200 エポックと 20 エポックとした．また，累積寄与率αは 99.0%と設定した．

表 1 に提案手法と比較手法の識別精度と再学習時間を示す．top1 と top5 精度は，それぞれ，出力が正解だと推定された上位 1 個と上位 5 個のラベルの中に正解ラベルが含まれている割合を表す．提案手法は，20 エポックと少ない学習時間で，従来手法と同程度の精度を達成している．これらの結果より，提案手法は従来手法に対して，同程度もしくはわずかな精度の低下で大幅な再学習時間の短縮に成功したといえる．

（２）学習前/学習後のプルーニングの組み合わせによる DNN 学習の高速化

学習後にプルーニングを行う Pruning after training （PaT）と学習前にプルーニングを行う Pruning at Initialization（PaI）を組み合わせた段階的な CNN モデルの軽量化手法を提案した．提案手法では，学習前に主成分分析（PCA）によって軽量化モデルの最適構造を決定し，短時間の学習後に学習済みの重みを活用したプルーニングにより更なる軽量化を行う．これにより，短い学習時間と高精度を両立した軽量化モデルを実現する．図 2 に提案手法の概要を示す．提案手法では，学習前のプルーニングにおいては，正規分布に従う乱数を順伝搬することで，特徴量マップを取得し，それを基にプルーニングを行う．そして，学習後のプルーニングでは，前項（１）の手法を適用することでプルーニングを行う．

データセット CIFAR100 を用いた物体認識タスクにより提案手法の有効性を検証した．対象モデルは vgg16_bn とした．比較手法には，PaI の代表的な手法を用いた．エポック数は比較手法

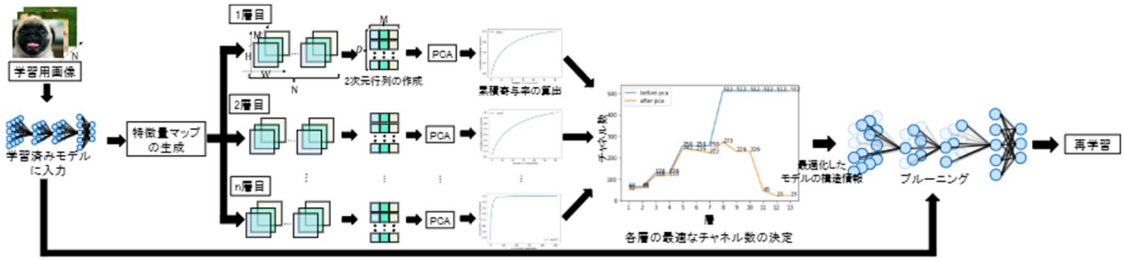

図 1：学習済み重みを活用したプルーニングの流れ

表 1：学習済み重みを活用したプルーニングにおける識別精度と再学習時間

データセット	圧縮率	top1[%]			top5[%]			再学習時間 [s]	
		手法 [1] 200epoch	手法 [1] 20epoch	提案手法 20epoch	手法 [1] 200epoch	手法 [1] 20epoch	提案手法 20epoch	手法 [1] 200epoch	提案手法 20epoch
CIFAR10	43%	93.13	89.37	**93.25**	**99.56**	99.49	99.46	2442.1	**243.8**
	47%	93.04	88.75	**93.30**	99.52	99.51	**99.53**	2382.8	**243.0**
	56%	92.99	88.88	**93.17**	**99.52**	99.37	99.45	2454.0	**240.9**
	65%	**93.14**	89.12	92.83	**99.51**	99.49	99.41	2380.6	**241.8**
CIFAR100	24%	**72.10**	61.38	70.92	**89.72**	86.97	88.97	2440.5	**247.5**
	33%	**71.78**	60.38	70.66	**89.32**	86.24	88.63	2444.7	**243.2**
	51%	**71.20**	60.12	69.53	**89.26**	86.50	88.44	2444.1	**244.2**
	64%	**70.14**	57.13	67.20	**89.12**	84.88	87.95	2396.1	**242.9**

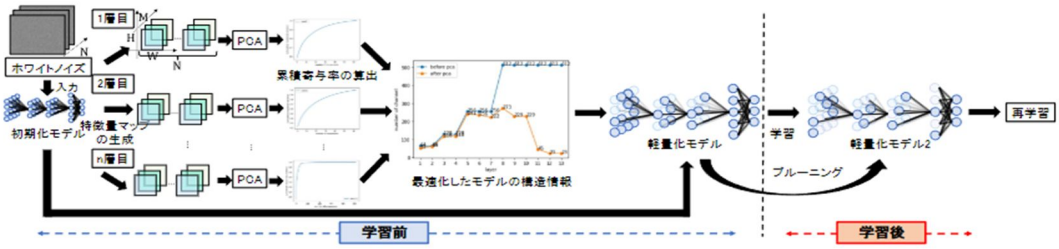

図 2：学習前/学習後のプルーニングを組みあせた手法

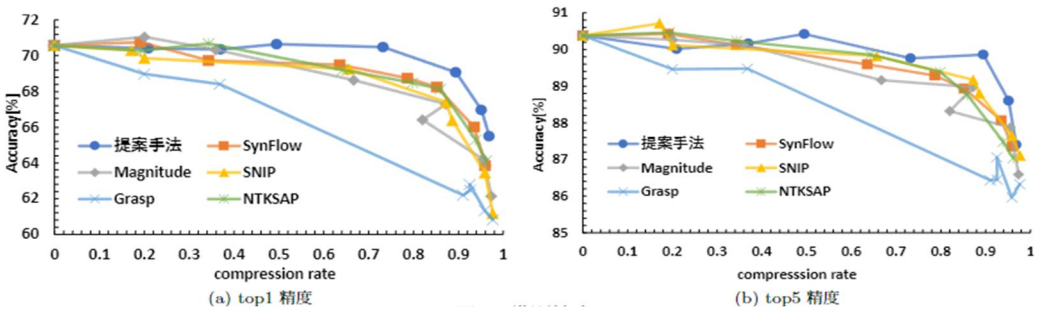

図 3：学習前/学習後のプルーニングにおける識別精度

が 200，提案手法は事前学習 30，再学習 170 の計 200 とした．また，累積寄与率αは 99.0%と設定した．

図 3 に top1 及び top5 精度を示す．図における compression rate は元モデルと軽量化モデルを比較した際のパラメータ削減率を表す．図 3 より，提案手法は低圧縮率では比較手法と同程度の精度であるが，高圧縮率では高い精度を示していることが分かる．以上より，提案手法は比較手法と比べて，精度を維持した軽量化モデルを PaI によって短い学習時間で構築できていることが確認できる．

（3）計算機アーキテクチャを考慮した量子化による DNN アーキテクチャの軽量化

DNN の軽量化手法である量子化において，計算機アーキテクチャに親和性が高い数値表現に基づいた上で，実際に観測した重み分布に従ったクリップ処理とビット割り当てを行う量子化手法を提案した．提案手法では，指数部と仮数部のビット配分の決定に実際の重み分布を用い，チャネル単位で量子化を行う．また，重み分布の特徴に合わせて外れ値のクリップ処理を行う．これにより，元モデルの精度をなるべく維持したモデル量子化を実現する．

提案手法では，従来手法で仮定されていた量子化誤差の評価関数に関して，実際の重み分布を基にした重みの生起確率分布を作成し，それに基づいて量子化を行う．評価実験では，従来の量子化手法と比較して，高い精度を実現した．

また，この他，DNN の学習における前処理の高速化として，フィルタリング処理に関する高速化も実現した．

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Yuto Sumiya, Tomoki Otsuka, Yoshihiro Maeda, and Norishige Fukushima	４．巻 29
２．論文標題 Gaussian Fourier Pyramid for Local Laplacian Filter	５．発行年 2022年
３．雑誌名 IEEE Signal Processing Letters	６．最初と最後の頁 11-15
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LSP.2021.3121198	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Nogami Haruki、Kanetaka Yamato、Naganawa Yuki、Maeda Yoshihiro、Fukushima Norishige	４．巻 24
２．論文標題 Decomposed Multilateral Filtering for Accelerating Filtering with Multiple Guidance Images	５．発行年 2024年
３．雑誌名 Sensors	６．最初と最後の頁 633～633
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/s24020633	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kanetaka Yamato、Takagi Hiroyasu、Maeda Yoshihiro、Fukushima Norishige	４．巻 12
２．論文標題 SlidingConv: Domain-Specific Description of Sliding Discrete Cosine Transform Convolution for Halide	５．発行年 2024年
３．雑誌名 IEEE Access	６．最初と最後の頁 7563～7583
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2023.3345660	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Naganawa Yuki、Kamei Hirokazu、Kanetaka Yamato、Nogami Haruki、Maeda Yoshihiro、Fukushima Norishige	４．巻 12
２．論文標題 SIMD-Constrained Lookup Table for Accelerating Variable-Weighted Convolution on x86/64 CPUs	５．発行年 2024年
３．雑誌名 IEEE Access	６．最初と最後の頁 15800～15819
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2024.3354720	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tsubokawa Teppei、Tajima Hiroshi、Maeda Yoshihiro、Fukushima Norishige	４．巻 83
２．論文標題 Local look-up table upsampling for accelerating image processing	５．発行年 2023年
３．雑誌名 Multimedia Tools and Applications	６．最初と最後の頁 26131～26158
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11042-023-16405-7	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計27件（うち招待講演　0件／うち国際学会　7件）

１．発表者名 Arashi Hirose, Yoshihiro Maeda, Takayuki Hamamoto
２．発表標題 Optimization of CNN Structure based on Principal Component Analysis and Sparsification
３．学会等名 International Workshop on Advanced Image Technology (IWAIT)（国際学会）
４．発表年 2024年

１．発表者名 峰尾光治，前田慶博，浜本隆之
２．発表標題 重み分布に基づいたチャネルごとのクリップとビット割り当てによるCNNの量子化
３．学会等名 画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)
４．発表年 2023年

１．発表者名 広瀬嵐，前田慶博，浜本隆之
２．発表標題 主成分分析とスパース化によるCNNモデルの構造決定とプルーニング
３．学会等名 画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)
４．発表年 2023年

１．発表者名
林晃平，前田慶博，福嶋慶繁

２．発表標題
値域フーリエ級数展開画像群のウェーブレット変換による局所コントラスト操作

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2023年

１．発表者名
野上遥貴，前田慶博，福嶋慶繁

２．発表標題
Tensor Coreを用いたガウシアンフィルタの高効率実装

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2023年

１．発表者名
Kohei Hayashi, Yoshihiro Maeda, and Norishige Fukushima

２．発表標題
Local Contrast Enhancement with Multiscale Filtering

３．学会等名
Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）（国際学会）

４．発表年
2023年

１．発表者名
峰尾光治，前田慶博，浜本隆之

２．発表標題
重みの分布に応じた適応的なビット割り当てによるニューラルネットワークの量子化

３．学会等名
映像情報メディア学会年次大会

４．発表年
2023年

１．発表者名
Soichiro Honda, Yoshihiro Maeda, and Norishige Fukushima

２．発表標題
Dataset of Subjective Assessment for Visually Near-Lossless Image Coding based on Just Noticeable Difference

３．学会等名
International Conference on Quality of Multimedia Experience (QoMEX)（国際学会）

４．発表年
2023年

１．発表者名
金高倭士,前田慶博, 福嶋慶繁

２．発表標題
Sliding DFTを用いた畳み込みフィルタの効率的な記述

３．学会等名
電子情報通信学会画像工学研究会（IE）

４．発表年
2023年

１．発表者名
Haruki Nogami, Sou Oishi, Tomohiro Sasaki,Yoshihiro Maeda, Norishige Fukushima

２．発表標題
Performance Evaluation of Halide Auto-Scheduler with Directional Cubic Convolution Interpolation

３．学会等名
International Workshop on Advanced Image Technology (IWAIT)（国際学会）

４．発表年
2023年

１．発表者名
林晃平, 福嶋慶繁,前田慶博

２．発表標題
局所コントラスト変換によるエッジ保存型ウェーブレット変換

３．学会等名
信号処理シンポジウム

４．発表年
2022年

１．発表者名
Yamato Kanetaka,Yoshihiro Maeda, Norishige Fukushima

２．発表標題
Basic Study on Domain Specific Description of Convolution with Sliding DFT

３．学会等名
International Workshop on Image Sensors and Imaging Systems (IWISS)（国際学会）

４．発表年
2022年

１．発表者名
広瀬嵐,前田慶博, 浜本隆之

２．発表標題
主成分分析とスパース化によるCNNネットワーク構造の最適化

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2022年

１．発表者名
大石創,前田慶博, 福嶋慶繁

２．発表標題
近似バイラテラルフィルタのJust-Noticeable Differenceに基づく画質評価指標

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2022年

１．発表者名
近藤拓海,前田慶博, 福嶋慶繁

２．発表標題
CPUマイクロアーキテクチャに応じた整数2次元畳み込みの高効率ベクトル化

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2022年

１．発表者名
小島史也,前田慶博，福嶋慶繁

２．発表標題
サブ/スーパーガウシアン分布関数の近似高速化計算

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2022年

１．発表者名
金高倭士,前田慶博，福嶋慶繁

２．発表標題
スライディング周波数変換を用いた畳み込みフィルタのドメイン固有言語

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2022年

１．発表者名
大石創，前田慶博，福嶋慶繁

２．発表標題
バイラテラルフィルタの近似誤差のJust-Noticeable Difference

３．学会等名
電子情報通信学会画像工学研究会（IE）

４．発表年
2022年

１．発表者名
小島史也，前田慶博，福嶋慶繁

２．発表標題
ガウス分布重み計算の近似によるバイラテラルフィルタの高速化

３．学会等名
電子情報通信学会画像工学研究会（IE）

４．発表年
2022年

１．発表者名
Yoshihiro Maeda, Norishige Fukushima, and Takayuki Hamamoto

２．発表標題
Color Transformation for Compressive Computing in Image Filtering

３．学会等名
Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2021年

１．発表者名
Takumi Kondo, Yoshihiro Maeda, and Norishige Fukushima

２．発表標題
Accelerating Finite Impulse Response Filtering Using Tensor Cores

３．学会等名
Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)（国際学会）

４．発表年
2021年

１．発表者名
角谷勇仁，大塚友貴，前田慶博，福嶋慶繁

２．発表標題
ガウシアンフーリエピラミッド～ローカルラプラシアンフィルタの高速化～

３．学会等名
電子情報通信学会画像工学研究会（IE）

４．発表年
2021年

１．発表者名
近藤拓海，前田慶博，福嶋慶繁

２．発表標題
Tensor CoreによるFIRフィルタの高速化

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2021年

１．発表者名
福嶋慶繁，前田慶博

２．発表標題
ローカルパッチ特徴の高速次元圧縮

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2021年

１．発表者名
八木伸二，前田慶博，浜本隆之

２．発表標題
行列演算機構を用いた積分画像生成の高速化の検討

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2021年

１．発表者名
渡邉丈裕，前田慶博，杉村大輔，浜本隆之

２．発表標題
人物移動軌跡推定のためのグラフニューラルネットワークに基づくリンク予測

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2021年

１．発表者名
渡邉丈裕，前田慶博，杉村大輔，浜本隆之

２．発表標題
人物移動軌跡推定のためのグラフニューラルネットワークに基づくリンク予測

３．学会等名
画像符号化シンポジウム(PCSJ)/映像メディア処理シンポジウム(IMPS)

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関