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　４ 年　 ６ 月 １８ 日現在

機関番号：１４７０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０２０

研究課題名（和文）Neuro-Coding/Unificationを用いたCNNのコンパクト化

研究課題名（英文）CNN compaction based on Neuro-Coding/Unification

研究代表者
和田　俊和（Wada, Toshikazu）

和歌山大学・システム工学部・教授

研究者番号：００２３１０３５

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では，CNNの機能をできるだけ低下させずに，計算量を削減するCNNの圧縮問題を取り扱った．当初は、学習データに対して各チャネルがどのように反応するのかをベクトルで表現し、線形従属の関係にあるチャネルのうち一方を削除し、残されたチャネルで削除したチャネルの働きを肩代わりさせる手法を提案した．その後，削除したチャネルの働きを，残されたチャネルにかける全ての重みを変更する「再構築」によって，次層の誤差を最小化するREAPという方法に拡張した．さらに，ResNetを直列化して圧縮するSRN，各層ごとの際的な圧縮率の配分を求めるPROという方法を提案した．

研究成果の学術的意義や社会的意義
提案したREAPを普通に実装すると，チャネルを削除して最小二乗法による再構築をする，という操作を全チャネルに対して適用して，再構築後の誤差が最小になるチャネルを削除するという計算になる．本研究では，この計算を双直交基底を用いてワンショットで解けるようにしたことが学術的に価値があると考えている．また，広く用いられているResNetに含まれる分岐のある層は削除することができなかったが，この分岐を無くすことによって，広い範囲のCNNにREAPの適用が行えるようになった．さらに，REAPでどの層をどれだけ圧縮すれば最適かを求めるPROによってCNN全体の最適な圧縮が可能になった．

研究成果の概要（英文）：We address the pruning problem for CNN that reduces the number of channels while keeping the accuracy. We started from a simple idea that similar channel pair having similar behaviors for training data can be replaced by a single channel. We further extended this idea to " reconstruction" that after pruning the channel all weights are updated for keeping the activation patterns in the following layer. Our method REAP guarantees we can find the optimal pruning channel that produces minimal error after the reconstruction. Also, we proposed a method SRN that serialize ResNet for applying REAP and PRO that estimates optimal pruning ratio for each layer.

研究分野：パターン認識，コンピュータ・ビジョン

キーワード：Neuro Coding/Unification　CNN pruning　REAP　Serialized ResNet　Pruning Ratio Optimizer

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

Deep Neural Network (DNN)は，画像や音声，さらには自然言語の理解において大きな成果を収めている機械学習の手法である．中でも，Convolutional Neural Network (CNN)は，少ないパラメータで精度の高い認識や画像生成が行えることが知られている．しかし，CNN では学習時にも推論時にも多くの計算が必要であるため，学習の効率化や，実時間の推論のためには，コア数が多い Graphic Processing Unit(GPU)の利用が欠かせない．しかし，GPU は数十万円から百数十万円と，高価であるばかりでなく，数十から百ワットと消費電力も多い．このため，組み込み系デバイスで高度な CNN を高速に動作させることは困難であり，社会に幅広くこの技術を広げていく上で大きな障壁があった．

この問題を解決するために，整数化，固定小数点化，二値化，など数の量子化によって高速化を実現する手法，重みの中にゼロ要素を増やす事によって高速化を試みる手法，重みを複数の重みの積として表現する事で高速化を試みる手法，DNN のニューロンや CNN のチャネルを削減する Pruning の手法，などが検討されてきた．特に，Pruning の手法は他の手法と併用することが可能であるため，よく研究が行われてきたが，Pruning によって生じる次層のアクティベーションパターンの誤差が拡大しやすいという問題点があった．この誤差を，「再構築」と呼ばれる重みの更新によって小さくする CP と呼ばれる方法が存在していたが，再構築後に誤差が最小となるニューロンやチャネルを選ぶことはできていなかった．

２．研究の目的

本研究では，DNN のモデルから，内在する冗長性を削減し，同じ性能でより小規模な DNN を自動的に作成する技術を構築することを目的とする．申請者らは，すでに，Neuro-Coding/Neuro-Unification と呼ばれる方法を考案し，実験を通じて他手法に勝る有効性を確認している．但し，この手法は「全結合層」のみを対象としているため，CNN で用いられる「畳み込み層」の圧縮は行えない．そこで，本研究では，Neuro-Coding/Neuro-Unification の考え方を畳み込み層に拡張し，CNN の圧縮を行う．さらに，２つの類似したニューロン/チャネルを一つに置き換えるという Neuro-Unification の考え方を発展させ，CP と同様に再構築を行う方法に発展させ，さらに再構築後に誤差が最小になるニューロンやチャネルを Pruning する方法や，広く用いられている ResNet の Pruning, CNN 全体の最適な削減率を求める方法についても明らかにする．

３．研究の方法

まず，Neuro-Coding/Neuro-Unification の考え方を CNN に拡張する際には，ニューロンはチャネルに対応するものと考える．そして，データに対する各チャネルの挙動は Feature Map 上のランダムにサンプリングした複数の位置での値を用いてベクトル化する．これらベクトルの挙動が類似するものを統合し，重みを更新する事によって次層のアクティベーションパターンの誤差を最小化する．

再構築の方法としては，あるチャネルを Pruning 層間の重みパラメータ全てを最小二乗法で最適化する事で誤差の最小化を行うが，Pruning するチャネルは再構築後の誤差が最小化されるものを選ぶべきである．しかし，あるチャネルを消して最小二乗法を適用するということを繰り返すと，計算量が増え過ぎてしまう．この問題を解消するために双直交基底を用いたワンショットの解法を構築する．この手法を Reconstruction Error Aware Pruning(REAP)と呼ぶ．

ResNet の圧縮では，分岐を持つ層を Pruning できないため，圧縮対象となる層が限定されてしまう.このため，恒等写像を表す重みを導入することによって，ResNet を一旦直列化し，それを REAP で圧縮する．この手法を Serialized Residual Network(SRN)と呼ぶ．

REAP は，層単位の Pruning 手法であるため，CNN 全体を圧縮する場合，どの層をどれくらい圧縮すれば，全体の精度に与える影響が少なく，効率良く圧縮できるかがわからないと，ネットワーク全体に対する最適な適用が行えない．このため，各層を大まかに圧縮し，出力層の誤差に及ぼす影響を調べる．そして，計算量と最終層の誤差の関係を折れ線グラフで表現し，これを参照してどの層をどれくらい圧縮すれば良いかを推定する手法 Pruning Ratio Optimizer(PRO)を構築する．

４．研究成果

Neuro-Coding/Unification に関しては，下記の国際会議と，学術論文の発表がある．下記研究では，ImageNet データセットでトレーニングした Vgg16（精度 0.895）の全結合の最終層を 1/2, 1/3 に圧縮した場合の精度がそれぞれ 0.879, 0.864 となり，他手法（DPP）の精度（0.856, 0.763）よりも大幅に高い性能が得られた．また，CNN の圧縮については，同様にトレーニングした Vgg16 に対して計算量が 1/2, 1/3 になるように圧縮すると，それぞれ 0.845, 0.729 になり，他手法（OP, ThiNet）のそれぞれの結果（0.024, 0.006）（0.245, 0.022）と比べて大幅な改善がなされてい

ることが確認できた.

[1] Koji Kamma, Yuki Isoda, Sarimu Inoue, and Toshikazu Wada. Behavior-Based Compression for Convolutional Neural Networks. ICIAR2019 (Best Paper).

[2] Koji Kamma, Yuki Isoda, Sarimu Inoue, and Toshikazu Wada. Neural Behavior-Based Approach for Neural Network Pruning. IEICE Transactions on Information and Systems, Vol. E103-D, No. 05, pp. 1135-1143, 2020.

Pruning 後に残った全ての重みを更新する再構成を行い，再構成後に最も誤差が小さくなるニューロンやチャネルを Pruning する REAP については，下記(国内査読付き会議論文，国際会議論文，学術論文)が研究業績である．ImageNet でトレーニングした Vgg16 の畳み込み層に適用した場合，計算量を半減させて速度を倍にした場合，REAP では精度は 2%しか低下せず，再学習する事によって元の 89.5%よりも 0.2%上昇するという結果が得られている．つまり，圧縮した方が精度が向上したという結果である．これは，他のいずれの手法よりも優れた結果である．

Table 1 VGG16 on ImageNet. The changes of top-5 accuracy from the baseline (89.5%) are reported (The greater, the better.). In this table, "rt" stands for "retraining". *our implementation.

Speed-up ratio	Method	Acc. before rt	Acc. after rt	Retrain epoch#
×2	REAP	**-2.0%**	**+0.2%**	10
	NU [9]	-5.0%	-	-
	CP [4]	-2.7%	0.0%	10
	*ThiNet [5]	-65.0%	-1.0%	10
	SPP [19]	-	0.0%	-
×5	REAP	**-9.4%**	**-1.3%**	10
	CP [4]	-22.0%	-1.7%	10
	*ThiNet [5]	-88.8%	-3.4%	10
	SPP [19]	-	-2.0%	-

[3] Koji Kamma and Toshikazu Wada. Accelerating the Convolutional Neural Networks by Smart Channel Pruning. MIRU2019 (Long oral, 学生奨励賞).

[4] Koji Kamma and Toshikazu Wada. Reconstruction Error Aware Pruning for Accelerating Neural Networks. ISVC2019.

[5] Koji Kamma, Toshikazu Wada, Biorthogonal System Based Channel Selection Algorithm for Neural Network Pruning. PRMU2020-2 (2020 年度研究奨励賞).

[6] Koji Kamma and Toshikazu Wada. REAP: A Method for Pruning Convolutional Neural Networks with Performance Preservation. IEICE Transactions on Information and Systems, Vol. E104-D, No. 01, pp. 194-202, 2021.

ResNet を直列化してから圧縮する SRN については，下記の研究業績がある．この手法を，ResNet-20/32/44/56 に適用し，CIFAR-10 で評価した結果を次の表に示す．この表では，元の ResNet とそれを直列化した SRN，その SRN をスクラッチから学習させた naïve，および，SRN を REAP で圧縮した pruned の４つを比較している．SRN に変換すると恒等変換に相当する重みとチャネルが増加するため，FLOPS がオリジナルに比べてほぼ倍になる．しかし，pruned では，圧縮により元の ResNet と同じ計算量になるようにしてある．これを見る限り，SRN の pruned では，オリジナルに比べて，同じ計算量で Test error がいずれも低く抑えられていることが確認できる．

[7] Koji Kamma, Toshikazu Wada, Serialized Residual Network. MIRU2020 (Short oral).

Table VI.2: The results on CIFAR-10. The SRN models consistently outperform the ResNet models. Unexpectedly, the SRN models run faster than the ResNet models in some cases, despite the doubled FLOPs. *We did not measure the inference time of SRN-x-naïve as it must be the same with SRN-X.

Model	Test error (%)	FLOPs	Inf. time (msec) at batch size		
			1	4	16
ResNet-20	8.46	**40.5M**	191.1	55.0	23.5
SRN-20	**7.19**	80.6M	**173.7**	50.9	25.3
SRN-20-naïve	7.76	80.6M	-*	-*	-*
SRN-20-pruned	8.17	**40.5M**	174.7	**49.8**	**20.4**
ResNet-32	7.40	**68.8M**	268.7	73.2	29.2
SRN-32	**6.32**	137.2M	247.0	**67.9**	36.5
SRN-32-naïve	8.66	137.2M	-*	-*	-*
SRN-32-pruned	7.12	**68.8M**	**245.6**	68.0	**28.4**
ResNet-44	7.18	**97.1M**	347.6	91.0	36.6
SRN-44	**6.03**	193.9M	321.8	85.1	46.6
SRN-44-naïve	9.58	193.9M	-*	-*	-*
SRN-44-pruned	6.98	**97.1M**	**310.5**	**84.1**	**35.0**
ResNet-56	6.63	**125.4M**	432.6	111.9	44.9
SRN-56	**5.62**	250.5M	397.9	102.2	56.9
SRN-56-naïve	11.52	250.5M	-*	-*	-*
SRN-56-pruned	6.57	**125.4M**	**388.7**	**102.1**	**42.1**

PRO に関しては，以下の学術論文が研究成果である．これは，CP を提案した MIT のグループが提案している先行手法 AMC と競合する手法である．AMC が強化学習による各層の圧縮率を決定するのに対して，PRO では各層を大まかに削減して，最終層の誤差を次の図のように折れ線で見積もる(a)．これを，FLOPs と最終層での誤差のグラフ(b)に変換し，これを利用して圧縮する層を決定している．

[8] Koji Kamma, Sarimu Inoue, and Toshikazu Wada. Pruning Ratio Optimization with Layer-Wise Pruning Method for Accelerating Convolutional Neural Networks. IEICE Transactions on Information and Systems, Vol.E105-D, No.1, pp.161-169. 2022.

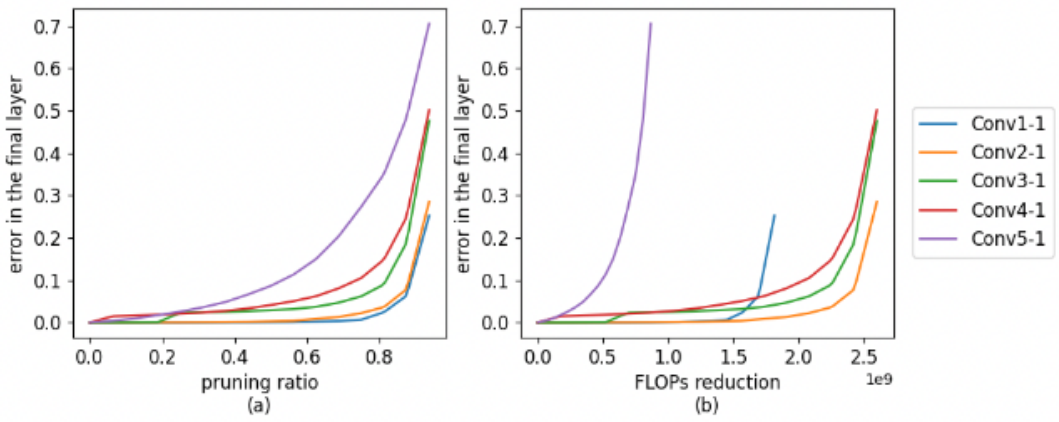

Figure V.3: (a) Relationship of the error in the final layer and pruning ratio in each layer. (b) Relationship of the error in the final layer and FLOPs reduction in each layer, which we actually use for selecting the layer to be pruned.

この方法で各層の圧縮率を決定し，REAP もしくは CP で圧縮した場合と，AMC で圧縮率を決定し，CP で圧縮した場合をまとめたのが次の表である．この表は，ResNet-56 を CIFAR-10 でトレーニングしたネットワーク（精度 92.8%）の計算量を半分にした結果である．この表から，PRO と REAP の組み合わせの精度が最も良く，92.1%となっている．次に精度が高いのが PRO と CP を組み合わ

せた手法である．

Table V.2: The results with the ResNet-56 model on the CIFAR-10 dataset. The top-1 accuracy are reported (The greater, the better.). The baseline accuracy is 92.8%.

Method	FLOPs	Acc. before rt	Acc. after rt	Time for optim.
PRO & REAP	×0.500	**90.6**%	**92.1**%	4,237 sec
PRO & CP	×0.498	90.0%	92.0%	3,800 sec
AMC & CP	×0.501	79.0%	91.4%	6,885 sec
uniform & REAP	×0.510	86.3%	91.2%	-

以上の４つの内容をまとめたものは，研究代表者が指導する菅間幸司氏の博士論文としてまとめられ，要約された内容が下記の研究会論文として発表されている．この論文は，2020 年度情報処理学会研究会推薦博士論文（各研究会 1 件）に選ばれている．

[9] Koji Kamma, Toshikazu Wada, Behavior-based DNN Compression: Pruning and Facilitation Methods. 2020-CVIM-226.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 KAMMA Koji、INOUE Sarimu、WADA Toshikazu	４．巻 E105.D
２．論文標題 Pruning Ratio Optimization with Layer-Wise Pruning Method for Accelerating Convolutional Neural Networks	５．発行年 2022年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 161～169
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2021EDP7096	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Koji Kamma, Toshikazu Wada	４．巻 E104.D
２．論文標題 REAP: A Method for Pruning Convolutional Neural Networks with Performance Preservation	５．発行年 2021年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 194-202
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2020EDP7049	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 KAMMA Koji、ISODA Yuki、INOUE Sarimu、WADA Toshikazu	４．巻 E103.D
２．論文標題 Neural Behavior-Based Approach for Neural Network Pruning	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 1135～1143
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2019EDP7177	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Koji Kamma, Toshikazu Wada
２．発表標題 Serialized Residual Network
３．学会等名 MIRU 2020 (Short Oral Session)
４．発表年 2020年

１．発表者名
磯田雄基，和田俊和

２．発表標題
Feature Sharing: 特徴マップの共有による複数DNNの統合

３．学会等名
研究報告コンピュータビジョンとイメージメディア（CVIM）

４．発表年
2021年

１．発表者名
Kamma K., Wada T.

２．発表標題
Reconstruction Error Aware Pruning for Accelerating Neural Networks.

３．学会等名
In: Bebis G. et al. (eds) Advances in Visual Computing. ISVC 2019. Lecture Notes in Computer Science, vol 11844. Springer, Cham（国際学会）

４．発表年
2019年

１．発表者名
Kamma K., Isoda Y., Inoue S., Wada T.

２．発表標題
(2019) Behavior-Based Compression for Convolutional Neural Networks.

３．学会等名
In: Karray F., Campilho A., Yu A. (eds) Image Analysis and Recognition. ICIAR 2019. Lecture Notes in Computer Science, vol 11662. Springer, Cham（国際学会）

４．発表年
2019年

１．発表者名
Koji Kamma, Toshikazu Wada

２．発表標題
Accelerating the Convolutional Neural Network by Smart Channel Pruning.

３．学会等名
(MIRU2019 Long Oral)

４．発表年
2019年

１．発表者名 磯田 雄基，菅間 幸司，和田 俊和
２．発表標題 Neuro Coding/Unificationを用いたDNNの効率的なパラメータ数削減法
３．学会等名 (MIRU2019 Poster presentation)
４．発表年 2019年

〔図書〕　計0件

〔出願〕　計2件

産業財産権の名称 ニューラルネットワーク処理装置、ニューラルネットワーク処理方法、及びコンピュータプログラム	発明者 和田俊和、菅間幸司	権利者 和歌山大学
産業財産権の種類、番号 特許、特願2020-123973	出願年 2020年	国内・外国の別 国内

産業財産権の名称 ニューラルネットワークの圧縮方法、ニューラルネットワーク圧縮装置、コンピュータプログラム、及び圧縮されたニューラルネットワークデータの製造方法	発明者 和田俊和　菅間幸司	権利者 国立大学法人 和歌山大学
産業財産権の種類、番号 特許、特願2019-137019	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔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