エッジデバイス向けAI圧縮技術の研究

－精度を維持しつつ省メモリでAIを実行する技術－

奥谷悠典* 笠原竹博**

外観検査用AIを導入するには，高速な演算処理性能を備えた高価なデスクトップコンピュータが必要なことが課題である。デスクトップコンピュータに代わる安価なデバイスとしてエッジデバイスが注目されているが，エッジデバイスは現状，搭載するメモリ量が少ないために外観検査用AIを実行できない。そこで，外観検査用AIの一種であるオートエンコーダを圧縮してメモリ使用量を削減する手法を開発した。本研究では，枝刈りと低ランク近似の2種類の圧縮手法を開発した。その結果，低ランク近似で圧縮することにより，精度を維持しながらメモリ使用量を削減できることを示した。また，同圧縮手法を用いることで，エッジデバイスで外観検査AIを実行可能であることを確認した。さらに，処理速度の向上にも効果があることを明らかにした。

キーワード: エッジデバイス，AI，オートエンコーダ，枝刈り，低ランク近似

Research on AI Compression Technology for Edge Devices
-Executing AI with Reduced Memory Consumption while Preserving Accuracy-

OKUTANI Yusuke and KASAHARA Takehiro

Implementing AI for visual inspection often requires expensive desktop computers with high-speed processing capabilities, which poses a significant challenge. While edge devices are gaining attention as cost-effective alternatives to desktop computers, they frequently lack sufficient memory to execute AI for visual inspection. To address this issue, we developed a method to compress autoencoder, thereby reducing memory usage. In this study, we evaluated two compression techniques: pruning and low-rank approximation. Our results demonstrate that low-rank approximation can effectively reduce memory usage while maintaining accuracy. We confirmed that this compression method enables the execution of visual inspection AI on edge devices. Furthermore, we discovered that this approach also contributes to improved processing speed.

Keywords : edge computer, AI, autoencoder, pruning, low-rank approximation

1．緒　　言

深層学習の登場を契機に，近年様々な業界でAIが活用されている。このうち製造業においては深層学習を用いた外観検査用AIを活用し，外観検査を自動化する動きが活発である。一方で，外観検査用AIでは高い演算処理能力が求められることから，高価なデスクトップコンピュータが必要である。これに対し，デスクトップコンピュータに比べ低価格で，AI向けの高速演算プロセッサを有しているエッジデバイスが注目されている。エッジデバイスは価格面で優れているほか，小型，低消費電力といった特徴も備えている。しかし，エッジデバイスは搭載しているメモリ量が少なく，メモリ不足のために外観検査用AIを実行できないことが課題である。

これに対し，深層学習を圧縮する研究が進められている。Hubaraらは少ないメモリ量で深層学習を実行するためにパラメータの量子化を行っている[1)]。しかし，量子化は深層学習の精度低下を招きやすい[2)]。精度低下が少なくメモリ使用量の削減効果が期待される手法として，枝刈りや低ランク近似がある。これらの手法による圧縮はクラス分類用途の深層学習を中心に取り組まれてきた[3),4)]。一方で，外観検査用AI，具体的にはCNN(Convolutional Neural Network)層から成るオートエンコーダ構造を備えた深層学習への取り組みは進

*電子情報部　**企画指導部/電子情報部

んでいなかった。

本研究では，エッジデバイスで外観検査用AIを実行可能にすることを目的に，外観検査用AIを圧縮し，その際のメモリ使用量や精度の変化を検証した。また，圧縮した外観検査用AIをエッジデバイスへ実装し，外観検査用AIの実行可否および処理速度に関して検証を行った。

２．圧縮手法の検討

２．１　実験条件

対象にした外観検査用AIは21層のCNN層から成るオートエンコーダである。以下に述べる枝刈りと低ランク近似の2通りの手法でこのオートエンコーダを圧縮し，未圧縮のものと比較した。実験条件を表1に示す。

表1　実験条件

ハードウェア	デスクトップコンピュータ
最大エポック数	200
バッチサイズ	16
損失関数	MSE
精度の評価指標	AUC
学習データ数	219
評価データ数	23(OK) / 109(NG)

ハードウェアにはデスクトップコンピュータを用い，精度の評価指標にはAUC(Area Under the Curve)を用いた。学習データ数は219，評価データ数はOK，NG合わせて132である。

評価対象画像にはMVTec Anomaly Detection Dataset[5]のcupsule画像を使用した。学習データはcupsuleの学習用のgoodデータ，テストデータはテスト用のgoodデータ及び全パターンのNGデータである。メモリ使用量はオートエンコーダが画像を生成する際の高速演算プロセッサのメモリ使用量で評価した。AUCは学習1エポック毎に算出し，最良値で評価した。

２．２　枝刈り

２．２．１　手法の概要

枝刈り手法のひとつにFranklеらによって提唱された宝くじ仮説がある[6]。宝くじ仮説では精度に大きく関与する一部の重みパラメータを当たりくじと呼んでおり，本手法は当たりくじを残すように枝刈りを行う。

本実験では以下の手順を実施した。

(1) 学習前の初期値パラメータを保存する

(2) AIを学習させる

(3) 残っている重みパラメータのうち，絶対値の小さな20%を削除する

(4) 削除しなかった重みパラメータを初期値に戻して，再学習させる

(5) 手順(3)～(4)を10回繰り返す

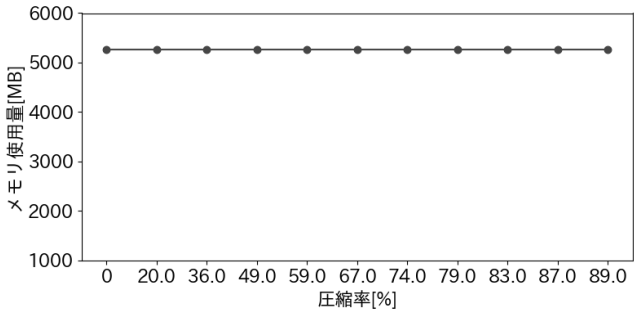
２．２．２　結果

枝刈りにより外観検査用AIを圧縮した際のメモリ使用量の変化を図1(a)に，精度の変化を図1(b)に示す。

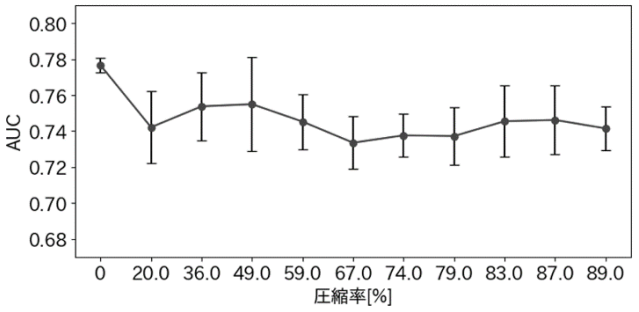
圧縮率は未圧縮時の重みパラメータ数と削減したパラメータ数の割合である。また，各圧縮率の試行回数は10回である。

図1(a)はAIを圧縮した場合でもメモリ使用量の削減効果が得られないことを示している。これは，各層やカーネルの重みパラメータの一部を削除しているものの，内部的には重みパラメータの値が0に置き換わるだけであることが原因である。

図1(b)は外観検査用AIを圧縮した場合でも精度をほぼ維持していることを示している。メモリ消費量に変化は見られず，期待した圧縮効果は得られなかった。



(a)メモリ使用量の変化



(b) AUCの変化

図1　枝刈りでの解析結果

２．３　低ランク近似

２．３．１　手法の概要

オートエンコーダの重みは高階テンソルで表現される。そこで，低ランク近似の具体的手法として，高階テンソルを分解可能なTucker分解を使用した。なお，Tucker分解にはHOOI(Higher Order Orthogonal Iteration of Tensors)を使用した。HOOIで各層毎に重みテンソルを分解し，最後に再学習を行った。

オートエンコーダにおける各層の重みテンソルはカーネルサイズの2階と入出力チャネル数の2階の計4階テンソルで構成される。そのうちカーネルサイズに関する要素は3もしくは8と小さいため，本実験では入出力チャネルの2階分のみ分解した[7)]。

２．３．２　結果

低ランク近似におけるメモリ使用量の変化を図2(a)に示す。圧縮率の算出方法や試行回数は枝刈りと同様である。図2(a)から，圧縮率が30.0%ではメモリ使用量が増加するものの，30.0%を超えるとメモリ使用量を削減できることが示された。

低ランク近似による圧縮は，圧縮前に比べて重みパラメータ数は減少するもののニューロン数が増加する。圧縮後のニューロン数は圧縮率を高めるにともない減少していく。ニューロン数と重みパラメータ数を表2に示す。

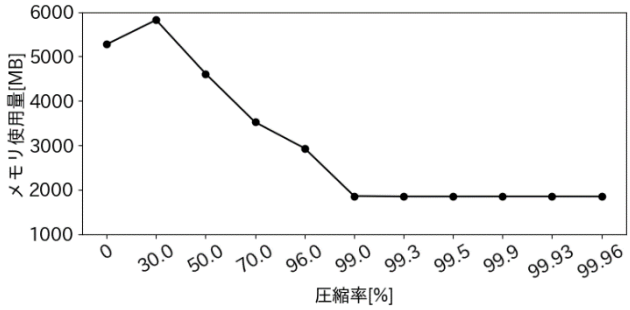
圧縮率が30.0%でメモリ使用量が増加したのは，メモリ使用量に関わる要因に重みパラメータ数の他にニューロン数が関係する。30.0%では削減した重みパラメータに関するメモリ使用量よりも増加したニューロン数に関するメモリ使用量の方が多くなるためにこのような結果になったと考えられる。

次に，精度の変化を図2(b)に示す。圧縮率が99.0%を超えると急激にAUCが低下するものの，30.0%から99.0%までは未圧縮のAIと同程度のAUCを維持する結果が得られた。

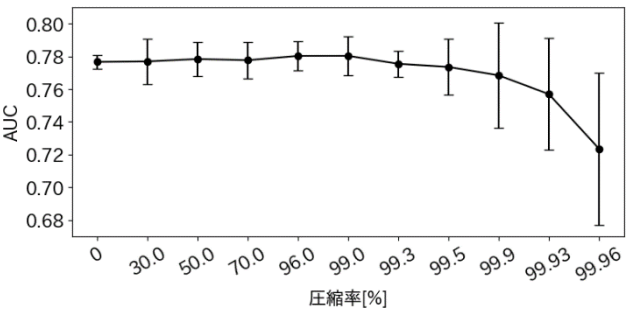
このように，低ランク近似による圧縮は枝刈りでは得られなかったメモリ使用量の削減効果が得られ，かつAUCを維持することが明らかになった。

表2　圧縮したオートエンコーダのパラメータ数

圧縮率	ニューロン数	重みパラメータ数
0%	2,393,088	34,732,835
30.0%	4,819,224	24,317,275
50.0%	4,818,826	17,388,493
70.0%	4,818,432	10,529,347
96.0%	2,777,600	1,371,567
99.96%	2,469,122	13,153



(a)メモリ使用量の変化



(b) AUCの変化

図2　低ランク近似での解析結果

３．エッジデバイスへの実装

３．１　実験条件

低ランク近似で圧縮した外観検査用AIをエッジデバイスへ実装し，動作可否の検証を行った。対象にした外観検査用AIは，前章と同じく21層のCNN層から成るオートエンコーダである。本実験では，4種類のエッジデバイスとデスクトップコンピュータで評価を行った。評価に用いたエッジデバイスとその搭載メモリ量は表3のとおりである。

実験では動作可否に加えて，処理時間も評価した。処理時間は，画像を外観検査用AI，すなわちオートエンコーダへ入力した時から，オートエンコーダが処理結果を出力した時までとし，ファイル読み込みや前処理，後処理に要する時間は含まない。各圧縮率での試行回数は1000回である。

表3　エッジデバイスの種類

エッジデバイス	搭載メモリ量
Google Coral Dev Board	1 GB
NVIDIA Jetson Nano	4 GB
NVIDIA Jetson Orin Nano	8 GB 4 GB

表4　エッジデバイスでの動作可能条件および処理時間

エッジデバイス	動作可能圧縮率	処理時間
Google Coral Dev Board	0%	436 ms
NVIDIA Jetson Nano	50%	122 ms
NVIDIA Jetson Orin Nano (4GB)	96%	13 ms
NVIDIA Jetson Orin Nano (8GB)	30%	28 ms

３．２　結果

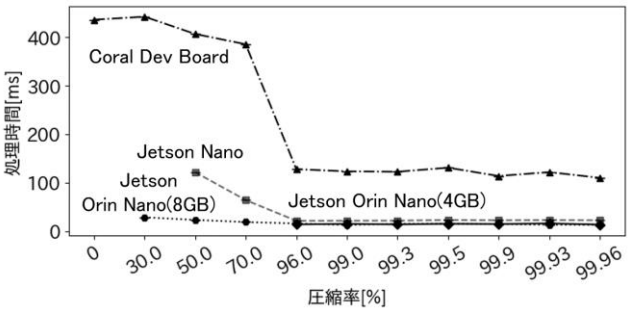
低ランク近似による圧縮後は，いずれのエッジデバイスでも動作可能になることを確認した。表4に動作可能な最低圧縮率および処理時間を示す。

Google Coral Dev Boardでは未圧縮の状態でも動作し，NVIDIA Jetson Nanoは50%以上の圧縮で動作した。NVIDIA Jetson Orin Nano (4GB)は96%以上，Jetson Orin Nano (8GB)は30%以上の圧縮で動作した。処理速度はJetson Orin Nano (4GB)，Jetson Orin Nano (8GB)，Jetson Nano，Coral Dev Boardの順で速い結果となった。

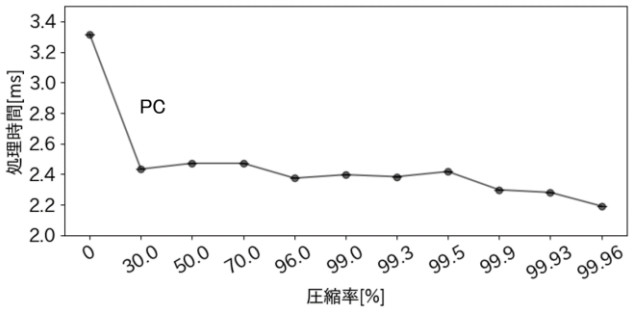
エッジデバイスでの処理時間の変化を図3(a)に示す。図3(a)中のプロットが無い箇所(例えば圧縮率0%のJetson Nanoのデータ)は，該当条件で外観検査用AIが実行不可能であったことを示している。図3(a)ではJetson Nanoは50.0%から，Jetson Orin Nano (8GB)は30.0%から，Jetson Orin Nano (4GB)は96.0%からグラフが始まっている。これは未圧縮(圧縮率0%)では実行できなかった外観検査用AIが，Jetson Nanoでは50.0%以上，Jetson Orin Nano (8GB)では30.0%以上，Jetson Orin Nano (4GB)では96.0%以上圧縮すると実行可能になることを示している。また，いずれのエッジデバイスでも96%までは圧縮率を高めるにつれて処理時間が短くなり，96%以降は一定になる結果が得られた。Coral Dev Boardは特に70%と96%の間を境に処理時間が短くなる結果が得られた。

比較のため，デスクトップコンピュータでの処理時間を図3(b)に示す。デスクトップコンピュータでもエッジコンピュータと同様に圧縮率を高めるにともない処理時間が短くなっていることが確認された。

このように低ランク近似による圧縮は，本来の目的であったメモリ使用量の削減に加え，処理の高速化にも効果があることが明らかになった。



(a)　エッジデバイス



(b)　デスクトップコンピュータ

図3　各デバイスでの処理時間の変化

４．結　　言

本研究では，外観検査用AIをエッジデバイスで実行することを目的に，外観検査用AIの圧縮を行った。その結果，低ランク近似を用いて圧縮することにより明らかになった知見を以下に示す。

(1) 精度を維持しつつ，メモリ使用量を削減できる
(2) エッジデバイスで実行可能になる
(3) 処理を高速化できる

謝　　辞

本研究の一部は公益財団法人I-O DATA財団の助成を受けて実施したものである。

参考文献

1) Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran Elyaniv, Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research. 2018. vol. 18, no.187, p. 1-30.
2) 山本康平，橘素子，前野蔵人．ディープラーニングのモデル軽量化技術. OKIテクニカルレビュー第233号. 2019. vol. 86, no.1, p. 24-27.
3) Sunil Vadera, Salem Ameen. Methods for pruning deep neural networks. IEEE Access. 2022. vol. 10, p. 63280-63300.

4) 本山義史, 遠藤敏夫, 松岡聡, 横田理央, 福田圭祐, 佐藤育郎. 低ランク近似行列による CNN における畳み込み演算の最適化. 情報処理学会研究報告. 2017. vol. 2017, no. 25, p. 1-7.

5) MVTec Academy. “The MVTec anomaly detection dataset (MVTec AD)”. MVTec Anomaly Detection Dataset: MVTec Software. https://www.mvtec.com/company/research/datasets/mvtec-ad, (参照 2024-08-20).

6) Jonathan Frankle, Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In International Conference on Learning Representations (ICLR). 2019.

7) Dwango Media Village. “深層学習のモデル学習”. https://dmv.nico/ja/casestudy/model_compression/, (参照 2024-08-20).