エッジデバイス向けAI圧縮技術の開発

電子情報部　○奥谷悠典　田村陽一　　企画指導部　笠原竹博

１．目　的

近年，AIを製品の外観検査に応用する試みが活発になっている。外観検査用AIには重みパラメータを多数持つ深層学習が利用されるため，高価かつ高性能なコンピュータ（例えばデスクトップコンピュータ）が必要であり，工場での活用が課題となっている。

そこでコンピュータへ安価かつ小型なエッジデバイスを活用する試みが注目されている。しかしエッジデバイスは搭載メモリ量が限られるため，少量のメモリで外観検査用AIを実行する技術が求められている。本研究では，エッジデバイスで外観検査用AIを実行するために，必要な精度を維持しつつメモリ使用量を削減する圧縮技術の開発を行なった。また，外観検査用AIの圧縮率と処理速度の関係の評価も行なった。

２．内　容

2.1　概要

事前にデスクトップコンピュータ上で，外観検査用AIの圧縮率とメモリ使用量・精度・処理速度の関係を検証した。その後エッジデバイス上で，外観検査用AIの動作可否および処理速度を検証した。実験条件は表1のとおりである。本研究の題材はCNN層をベースとしたオートエンコーダを使った画像検査用AIである。圧縮手法には低ランク近似を用いた。また，後処理には再学習を実施した。評価対象のカプセル薬の外観検査用AIは，一般的にAIの評価に用いられるMVTec Anomaly Detection Datasetのものである。本実験では圧縮前の外観検査用AIの作成時と同じ学習データ・評価データを用いた。AUC（AIの精度の指標）は学習1エポック毎に算出し，最良値を該当条件における評価値とした。

2.2　メモリ使用量の検証

外観検査用AIの実行時に占有するメモリ使用量の検証を行なった。圧縮率とメモリ使用量との関係を図1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を示している。また，各圧縮率の試行回数は10回である。図1から，99.0%まではメモリ使用量が減少するが，99.0%を超えると減少が見られないことが分かる。

表１　実験条件

AI の種類	深層学習（オートエンコーダ）
用途	カプセル薬の外観検査
圧縮手法	低ランク近似(Tucker 分解)
最大エポック数	200
バッチサイズ	16
損失関数	MSE
精度の指標	AUC
学習データ数	219
評価データ数	23(OK) / 109(NG)

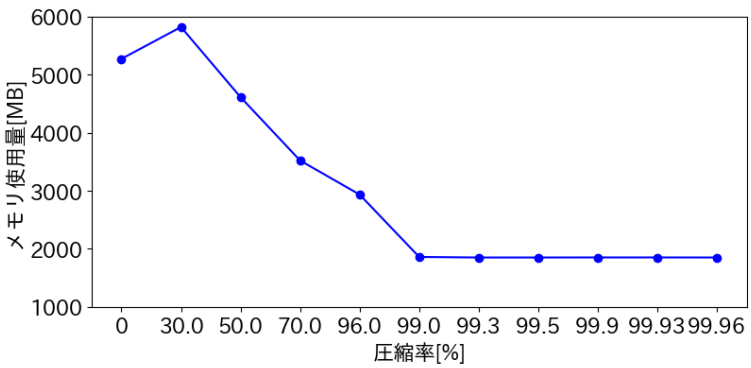

図１　圧縮率とメモリ使用量の関係

2.3　精度の検証

外観検査用AIの精度変化について検証を行なった。圧縮率とAUCとの関係を図2に示す。圧縮率の算出方法や試行回数は2.2と同様である。図2に示すように，圧縮率が99.0%を超えると急激にAUCが低下するものの，

約99.0%までは圧縮前のAIと同程度の精度を維持する結果が得られた。

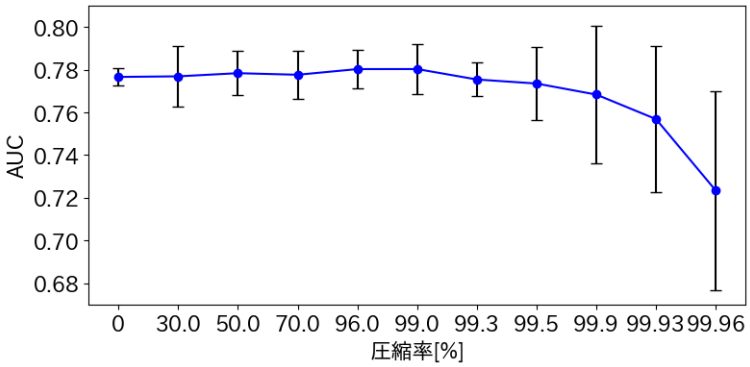

図２　圧縮率とAUCの関係

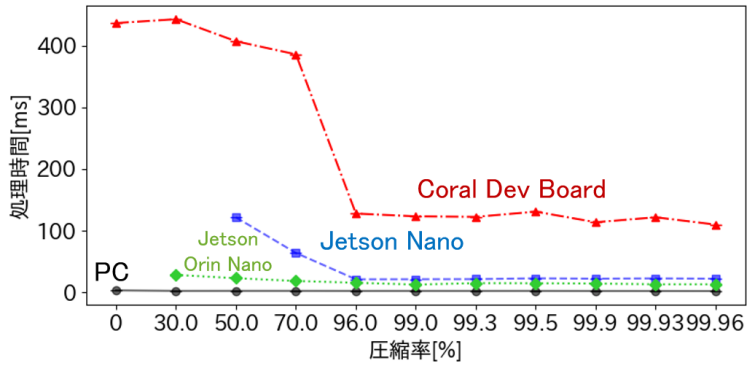

図３　圧縮率と処理時間の関係

2.4　処理速度の検証

外観検査用AIの処理速度について検証を行なった。処理時間の計測区間は，画像1枚を外観検査用AIへ入力し，外観検査用AIがデータを出力するまでである。なお，エッジデバイスにはGoogle LLC社製 Google Coral Dev Board・NVIDIA Jetson Nano開発者キットB01(Jetson Nano)およびNVIDIA Jetson Orin Nano開発者キット(Jetson Orin Nano)の3種類を使用した。各圧縮率での試行回数は1000回である。

圧縮率と処理時間の関係を図3に示す。処理時間は最初の1回を除いた 999 回分の平均値である。またグラフがない箇所は該当条件で外観検査用AIが動作しなかったことを示している。未圧縮の状態では動作しなかったエッジデバイスが圧縮によって動作可能になることが分かった。

また本技法での圧縮は処理の高速化にも効果がある結果も示された。エッジデバイス毎に見てみると，Jetson NanoおよびJetson Orin Nanoはデスクトップコンピュータと遜色ない処理速度を得られることが分かった。Coral Dev Boardは処理速度に劣るが，低率の圧縮でも動作可能なことが分かった。

３．結　果

本研究ではエッジデバイスで外観検査用 AI を実行することを目的に，外観検査用 AI の圧縮技術の開発を行なった。その結果，以下の成果が得られた。

・圧縮率を上げるに連れてメモリ使用量は減少し，99%までは削減効果があった
・精度は 99%の圧縮まで元の AI と同等の精度を維持した

以上の結果から，本技法における最適な圧縮率は 99%と結論付けた。

また，本技法での圧縮は外観検査用 AI をエッジデバイスで実行可能にすることに加え，デスクトップコンピュータ上でも処理速度の向上の目的へ応用が可能であることが明らかになった。

謝　辞

本研究の一部は公益財団法人 I-O DATA 財団様の助成を受けています。