現場で使える低価格なAI実装モデルの構築

生産技術部　〇谷山清吾，上薗　剛*
（現 *地域資源部）

１．　はじめに

AI技術を活用する際，クラウドが利用されることが多い。しかし，クラウドAIではインターネット通信の遅延によってリアルタイム性の確保が困難であるという問題がある。更に，セキュリティの面での不安や，インターネット環境を用意できない場所では利用ができない等の問題もある。これらの問題に対して，エッジデバイスでAI処理を行うことで解決が可能になる。

そこで本研究では，AI活用の促進を目的に，低価格なエッジAIデバイスの検討および処理が軽量，かつ現場で扱える高精度なAIモデルの開発を行った。

２．　研究概要

２．１　AI圧縮手法の検討

処理能力の低いエッジデバイスでAI処理を行うために，AIの処理を軽量化する圧縮手法の検討を行った。本研究では以下の２つの手法で圧縮を検討した。また，圧縮をより強くかけると軽量化の効果は大きくなるが，AIの推論精度が落ちる可能性がある。本研究では，圧縮対象となるAIの精度が落ちないことを最優先で圧縮パラメーターの検証を行った。

①量子化：ニューラルネットワークの演算やニューロンの重みなどAIのパラメータを通常の32ビット浮動小数点から桁の小さな固定小数点に置き換えることでメモリの使用量削減や計算効率の向上を図る。

②枝刈り：ニューラルネットワーク内で重みの小さな接続や，影響の少ないノードを削除することで，計算量を削減しメモリの使用量を抑える。

２．２　AI実装デバイスの検討

AI実装を行うエッジデバイスの検討を行った。本研究では，低価格なAI実装モデルを構築するため，低価格で入手可能かつ一定の処理能力を有するエッジデバイスとしてRaspberry Pi3B+，Raspberry Pi4Bを検討した。各デバイスの概要を図１に示す。

２．３　AI実装モデルの性能評価

２．３．１　実装する対象AI

今回，AI実装モデルを構築する上で実装を行うAIを２つ作成した。各AIの概要を表１に示す。また，実装AI②については，イメージを図２に示す。実装AI①は広く，画像分類の機械学習で使用されているCIFAR-10のデータセットを用いた。実装AI②は，県内企業で製造しているねじ頭部に発生する傷の検出を目的とし，模擬的にねじに傷をつけ画像を撮影することでデータセットを作成した。

SoC	Broadcom BCM2837B0
CPU	ARM Cortex-A53 (ARM v7)64-bit 1.4GHz
メモリ	1GB LPDDR2 SDRAM

Raspberry Pi3B+

SoC	Broadcom BCM2711
CPU	Quad core Cortex-A72 (ARM v8)64-bit 1.5GHz
メモリ	4GB LPDDR4 SDRAM

Raspberry Pi4B

図１　エッジデバイス概要

表１　実装AI概要

	実装AI①	実装AI②
概要	CIFAR-10[※1]を用いた 10種類の画像分類[※2]	ねじ頭部の傷検出
入力画像	32×32(カラー)	140×140(カラー)
学習画像数	5000[枚]	756[枚]

※１　機械学習のチュートリアル等で使用される公開されたデータセット。

※２　10種の内訳は「飛行機」,「自動車」,「鳥」,「猫」,「鹿」,「犬」,「カエル」,「馬」,「船」,「トラック」。

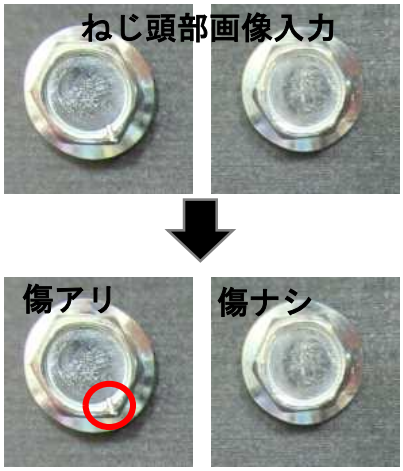

図２　実装AI②ねじ頭部の傷検出

２．３．２　評価方法

評価項目としては，[推論用のAI読み込み]，[テスト用画像１枚の入力]，[結果の出力]の一連の流れを完了するまでの時間及び推論の精度で評価した。ここでの精度の値は，実装AI①では，テスト用画像1000枚における精度，実装AI②では，テスト用画像108枚における精度である。

また，実装AI①，②について，それぞれ圧縮による軽量化前後と実装デバイス（Raspberry Pi3B+, Raspberry Pi4B）による評価結果の比較を行った。

２．３．３　評価結果

評価結果を表２に示す。表２の結果から，圧縮を用いたAI軽量化によって，いずれの実装AI及びデバイスにおいても精度を維持しつつ処理時間を短縮することができた。実装デバイスによる精度，処理時間の比較では，精度については全てのデバイスで同様の圧縮を行ったため，デバイスによる精度の違いは発生していない。処理時間については，処理能力の高いRaspberry Pi4Bがいずれの結果においてもより高速に処理できている。

表２　実装AIモデル評価結果

		軽量化(圧縮)前		軽量化(圧縮)後	
		精度[%]	処理時間[s]	精度[%]	処理時間[s]
Raspberry Pi3B+	実装AI①	90.4	13.2	90.4	8.4
	実装AI②	100	10.6	100	3.4
Raspberry Pi4B	実装AI①	90.4	7.3	90.4	4.1
	実装AI②	100	4.6	100	1.4

３．　おわりに

AI圧縮による軽量化で，処理能力の低いエッジデバイスでも精度を維持しつつ，より短時間でのAI処理が可能となった。今後は，今回詳細な検討を行っていない圧縮手法による軽量化の効果や，今回対象とした実装AI以外での実装性能の比較にも取り組みたい。