C-003

CNN 推論処理における畳込み層の同値性および連続性に基づく演算量削減手法
Complexity Reduction Based on Equivalence and Continuity of Convolutional Layers in CNN Inference

大森 優也† 小林 大祐† 吉田 周平† 八田 彩希† 鵜澤 寛之† 中村 健† 佐野 公一†
Yuya Omori Daisuke Kobayashi Shuhei Yoshida Saki Hatta Hiroyuki Uzawa Ken Nakamura Kimikazu Sano

1. はじめに

近年，監視カメラやドローン等のエッジデバイス上で実行される AI システムが注目されている．これらのエッジ AI システムは，畳込みニューラルネットワーク（CNN）を用いた AI 推論処理をデバイス上の AI 推論用ハードウェアで行うことで，入力映像に対するリアルタイムのオブジェクト認識や物体検出等をオフラインで実現する．

計算資源の限られるエッジデバイス上で AI 推論処理のスループットと認識精度を両立させるためには，演算精度の劣化無く CNN 推論の演算量を削減し高速化することが必要である．特に CNN 推論演算の大部分を占める畳込み演算の演算量削減が不可欠となる．AI 推論ハードウェアにおける畳込み演算量削減として，ブロック単位の処理を前提として畳込み層のゼロスパース性に注目した演算スキップ手法がある[1][2]．これらの AI 推論ハードウェアの畳込み演算では，積和演算回路を並列に用意することで，(M×N 画素)×1ch を単位ブロックとして毎サイクル 1 ブロックを同時に処理して結果を累積加算していく．畳込み層の入力特徴マップの値がゼロである場合には畳込み演算後の結果もフィルタによらずゼロとなり計算不要であるため，入力特徴マップのある 1 ブロックの内部が全てゼロである場合には演算をスキップして 1 サイクル分の高速化が可能となる．一方で，スループットを高めるためにブロックサイズをある程度大きく設定する場合，入力ブロック内部の M×N 個相当のデータが全てゼロとなることは少なく，スキップによる演算量削減効果が低いという課題がある．また，演算精度を高めるために特徴マップデータを大きなビット深度で保持する場合も，ゼロスパースとなるブロックは少なく十分な高速化が期待できないことが多い．

本研究では，畳込み層の同値性と連続性に注目した演算スキップ手法を提案する．入力ブロックがゼロスパースでない場合でも，ブロック内部が全て同値となる場合や同一ブロックが連続している場合は多い．入力ブロックが同値または連続となる場合に畳込み演算をスキップして内部メモリデータ参照のみで畳み込み結果を出力可能なハードウェア構成により，従来のゼロスパースなブロックをスキップする手法と比べて演算スキップ率を大きく向上させ，AI 推論ハードウェアの処理高速化を実現可能とする．

2. 提案手法

入力ブロック内部が全て同値となるものを同値ブロック，直前の入力ブロックと内部構造が全く一致するものを連続ブロックとする．一例を図 1 に示す．例では同時に出力する単位ブロックのサイズを (4×2 画素)×1ch としている．畳込み出力結果の全体サイズを 8×4 画素とすると 4 ブロックで構成されるため，スキップ無しの場合は 1ch ごとに 4 サイクル必要である．フィルタは対象出力位置の周囲±1 画素を畳み込む 3×3 サイズとする．出力ブロック単位が 4×2 画素である場合，入力ブロック単位はフィルタの±1 画素を考慮して 6×4 画素となる．同様に入力特徴マップサイズは 10×6 画素である．図 1 では，1 クロック目の左上入力 6×4 画素は内部が全て 2 であるため同値ブロックである．同値ブロックは出力内部も全て同値(-2)となる．同値ブロックの 1 つがゼロスパースといえる．2,3 クロック目の右上入力と左下入力は同値や連続ではない．4 クロック目の右下入力は，内部構造が直前の左下入力と全て一致しているため連続ブロックである．連続ブロックの出力は直前ブロックの出力と内部構造が一致する．同値ブロック/連続ブ

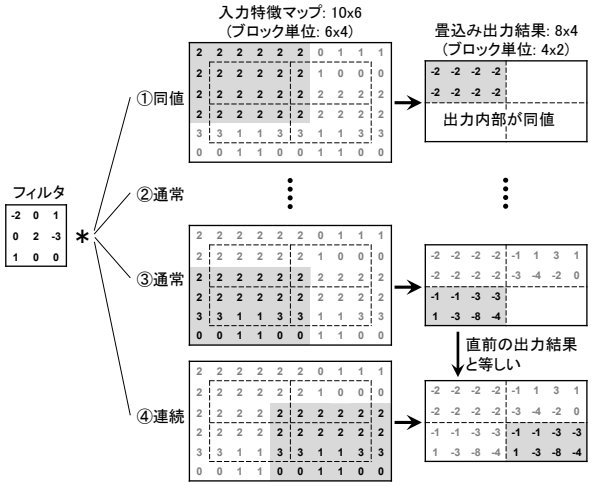

図 1　同値ブロック・連続ブロック

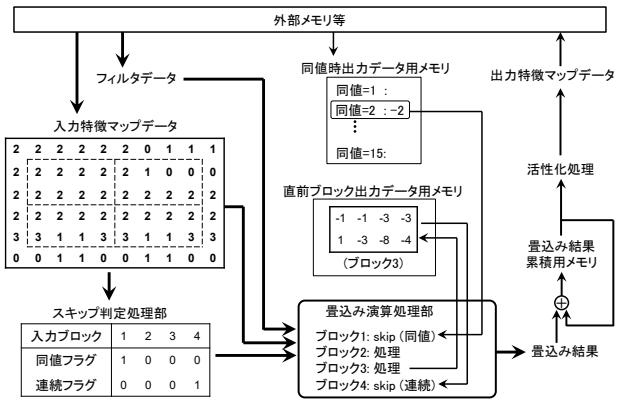

図 2　システム構成図

† 日本電信電話株式会社，NTT デバイスイノベーションセンタ

表 1　ブロックスキップ率

出力ブロック単位	スキップ手法	評価画像 dog	eagle	horses	kite	person	Average
4×2	ゼロブロック	0.6%	1.1%	1.2%	1.9%	1.8%	1.3%
	同値ブロック	3.9%	6.5%	7.3%	10.3%	9.7%	7.5%
	連続ブロック	5.7%	9.2%	9.9%	11.4%	11.1%	9.5%
	同値または連続ブロック	6.1%	9.8%	10.8%	13.9%	13.1%	10.7%
4×4	ゼロブロック	0.6%	1.0%	1.1%	1.1%	1.1%	1.0%
	同値ブロック	3.4%	5.9%	6.5%	6.5%	6.2%	5.7%
	連続ブロック	5.5%	8.9%	9.6%	9.3%	9.1%	8.5%
	同値または連続ブロック	6.0%	9.5%	10.4%	10.2%	9.9%	9.2%
8×4	ゼロブロック	0.4%	0.8%	0.8%	0.9%	0.9%	0.8%
	同値ブロック	2.5%	4.6%	4.9%	5.4%	5.3%	4.5%
	連続ブロック	4.0%	6.7%	7.0%	7.3%	7.3%	6.5%
	同値または連続ブロック	4.3%	7.1%	7.6%	8.1%	8.0%	7.0%

ロックがスキップ可能となれば 1,4 クロック目の処理が不要となり，通常 4 サイクルを 2 サイクルまで短縮できる．

同値ブロック/連続ブロックの演算スキップを実現可能なシステム構成について図 2 に示す．外部メモリ等からリードした入力特徴マップデータに対し，事前に各入力ブロックが同値ブロック/連続ブロックかの判定処理を行い，判定結果をフラグ化する．畳込み演算回路には，フィルタデータ，入力ブロックデータと合わせてスキップ判定フラグを入力し，フラグが有効な場合は対象入力ブロックに対する演算処理をスキップする．同値ブロックの場合，演算結果の出力ブロック内部は全て同値となり，さらに演算結果は 2^n 種類 (n:ビット精度，図 2 は 4bit の例) のみに限定される．このため，出力ブロック内のある一点の値のみを 2^n 種類保持するだけで，同値ブロックの演算結果を表現可能である．同値ブロックの演算結果は事前計算可能であるため，現層での同値ブロック結果を予めリードして保持し，同値フラグ有効の場合には対応する値の同値ブロック結果を利用することで，畳込み演算をスキップできる．連続ブロックの場合，演算結果は直前ブロックと一致するため，直前の出力ブロック結果を常にレジスタ等で更新しながら格納しておき，連続フラグ有効時は連続して結果出力するだけでよく，畳込み演算なしで高速化ができる．以上の構成により，本手法はハードウェアリソースの大きな増加無しで同値ブロック/連続ブロックの演算スキップが可能であり，演算精度の劣化無く AI 推論ハードウェアの高速化を実現する．

3. 評価

スキップ可能なブロック数の測定により，本手法の演算高速化性能の評価を行った．ネットワークモデルとして YOLOv3 608×608[3]を使用し，評価画像は図 3 に示す 5 種類とした．8bit 演算精度の AI 推論ハードウェアを想定し，各層の特徴マップおよびフィルタは 8bit データとして演算を行った． 並列出力可能なブロックサイズを 4×2, 4×4, 8×4 の 3 種類とし，それぞれの場合についてブロックスキップ率を算出した．YOLOv3 のフィルタサイズは 3×3 または 1×1 であるため，入力ブロック単位はそれぞれ (6×4 または 4×2)，(6×6 または 4×4)，(10×6 または 8×4)となる．本手法は入力ブロックが同値ブロック/連続ブロックとなる時に演算をスキップし処理高速化が可能なため，同値ブロック・


dog　eagle　horses　kite　person

図 3　評価画像

連続ブロックの比率を算出することで高速化性能を評価した．従来技術との比較として，入力ブロック内部が全てゼロとなるブロックの比率についても算出した．

実験結果を表 1 に示す．出力ブロックサイズ 4×2, 4×4, 8×4 のそれぞれで，従来のゼロスパースなブロックスキップ手法の平均スキップ率が 1.3%, 1.0%, 0.8%であるのに対し，本手法の同値または連続なブロックをスキップする手法の平均スキップ率は 10.7%, 9.2%, 7.0%であり，本手法により AI 推論処理を大きく高速化可能であることが分かる．出力ブロックサイズすなわち演算スループットと平均スキップ率はトレードオフの関係にあるが，スループット最大の 8×4 サイズの場合であっても本手法は平均 7.0%の演算スキップが可能であり，従来手法の平均 0.8%と比較して +6.2%の高速化を達成する．同値ブロック単体での平均スキップ率はブロックサイズ 4×2, 4×4, 8×4 でそれぞれ 7.5%, 5.7%, 4.5%，連続ブロック単体での平均スキップ率はそれぞれ 9.5%, 8.5%, 6.5%であり，本手法では連続ブロックによるスキップ性能がより支配的となっている．

4. おわりに

本研究では，AI 推論用ハードウェアにおける CNN 推論処理の演算量削減および高速化を目的として，畳込み層の同値性と連続性を利用した畳込み演算のスキップ手法を提案した．既存のゼロスパースなブロックスキップ手法と比較して本手法が演算スキップ率を大きく向上可能であることを，YOLOv3 を用いた評価実験により確認した．

参考文献

[1] Angshuman Parashar, et al., “SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks”, arXiv preprint arXiv:1708.04485 (2017).
[2] Zhe Yuan, et al., “Sticker: A 0.41-62.1 TOPS/W 8Bit Neural Network Processor with Multi-Sparsity Compatible Convolution Arrays and Online Tuning Acceleration for Fully Connected Layers”, 2018 IEEE Symposium on VLSI Circuits (2018).
[3] J. Redmon, et al., “YOLOv3: An Incremental Improvement,” arXiv preprint arXiv:1804.02767 (2018).