C-001

FPGA 実装に向けた超解像 CNN のリソース削減手法適用による低消費電力化
Low Power Super-Resolution CNN for FPGA Implementation Using Resource Reduction Method

辰己　守† Mamoru Tatsumi　森　一紀† Kazuki Mori　黒木 修隆† Nobutaka Kuroki　沼　昌宏† Masahiro Numa

1. はじめに

近年，物体認識や画像・映像の高画質化を，畳み込みニューラルネットワークにより実現する技術が注目を集めている。膨大な計算量を必要とする畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）[1] の処理を高速に行うためのアクセラレータとして，従来は GPGPU（General Purpose Graphic Processing Unit）が一般的に利用されていたが，汎用プロセッサのため無駄な処理が多く，消費電力が大きいという問題があった。そこで，GPU の約 10 分の 1 程度の消費電力で動作可能な FPGA 上にニューラルネットワーク専用の回路を実装することで，低消費電力化を実現する研究が注目されている。FPGA 上のリソースで実装可能な超解 CNN として，SRCNN-FP（Super Resolution CNN with Fewer Parameters）が提案されている [3]。しかし，多くの乗算を必要とするため，消費電力が高いという問題がある。そこで本稿では，精度を保ちながら低消費電力動作可能なネットワーク構造を提案する。まず，グループ化畳み込みを導入し，パラメータ数の抑制を図る。次に，固定小数点化を導入し，乗算回数の抑制を図る。

2. ネットワーク軽量化手法

2.1 グループ化畳み込み

図 1 にグループ化畳み込みの概略を示す。畳み込み演算を行う際にパラメータ数が多くなる要因として，チャネル方向の畳み込みが挙げられる。そこで，入力をチャネル方向に分割することでパラメータ数を削減する手法として，グループ化畳み込みが提案されている。しかし，各グループ内で畳み込みを実行するため，グループ間の関係性が低下し，精度が低下する可能性がある。

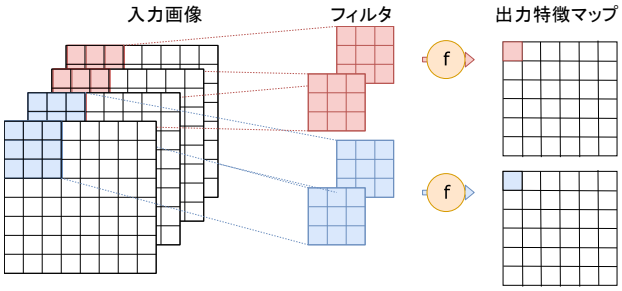

図 1 グループ化畳み込み

2.2 固定小数点化

固定小数点化は，量子化手法の 1 つである。ニューラルネットワークにおける量子化とは，通常 32 bit もしくは 16 bit 精度の浮動小数点数で表現されている重み等のパラメータを，より少ないビット幅で表現することであり，計算の高速化や省メモリ化等の利点がある。

† 神戸大学，Kobe University

固定小数点化では，パラメータの整数部分（IL：Integer Length）と小数部分（FL：Fluctuating Length）をそれぞれ任意のビット幅に量子化できるため，精度とのトレードオフがとれる。本稿で利用している固定小数点化のビット幅は，SRCNN-FP において精度低下を 0.1 dB 程度に抑えるビット幅であり，重みについては IL = 3 bit , FL = 7 bit とし，特徴マップについては IL = 4 bit , FL = 9 bit に量子化した。

2.3 提案ネットワーク

図 2 に提案するネットワーク構造を示す。パラメータ数を削減しつつ，精度の低下を維持するために，グループ化畳み込みを Residual Block 内に限り適用した。また，固定小数点化を各畳み込み層の重みと特徴マップに適用した。このようにして，低消費電力かつ精度低下を抑制したネットワークを実現する。

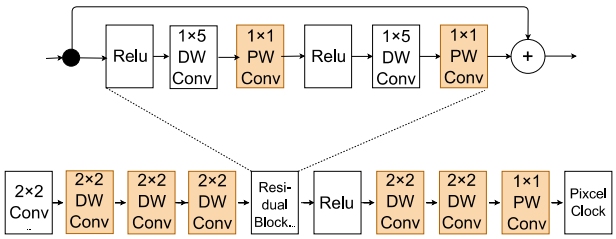

図 2 提案ネットワーク構造

3. 評価実験と考察

3.1 実験内容

本稿では，SRCNN-FP をもとにグループ化畳み込みと固定小数点化を適用したネットワークを提案とした。SRCNN-FP を従来手法として，精度，利用リソース数，消費電力に関して提案手法との比較を行った。また，提案手法に導入したグループ化畳み込みと固定小数点化それぞれの有用性を検証するために，次の 2 段階に分けて実験を行った。

i）実験 1：従来手法の Residual Block 内外に場合分けして，グループ化畳み込みを適用した場合の精度評価

ii）実験 2：実験 1 で精度が向上する回路に対し，固定小数点化を適用して精度評価

実験 1, 2 で共通して，超解像は画像の Y 成分（輝度成分）にのみ適用し，評価も Y 成分のみについて行う。元画像を Bicubic 法により 1/2 にダウンサンプリングした画像を入力とし，超解像 CNN により 2 倍拡大した画像を出力する。CbCr 成分（色成分）に対しては，ダウンサンプリングした入力に対して，Nearest Neighborhood 法を利用して 2 倍拡大して出力する。また，パラメータ数を削減するため，バイアスは 0 とする。

表 1 実験 1：グループ化畳み込み適用場所による評価

適用場所	パラメータ数 (比)	PSNR [dB]			
		Set5	Set5	Set14	Urban 100
適用なし	2746 (1.00)	36.19	36.67	32.18	29.39
Residual Block 外	2448 (0.89)	36.19	36.67	32.18	29.39
Residual Block 内	2224 (0.81)	35.95	36.45	32.08	29.23

表 2 実験 2：PSNR 評価

手法	PSNR [dB]			
	General 100	Set5	Set14	Urban 100
従来手法	36.19	36.67	32.18	29.39
提案手法（量子化）	35.95	36.45	32.08	29.23

表 3 実験 2：リソース利用数

リソース	利用数	
	従来回路	提案回路（量子化後）
LUT	28,888	17,495
LUTRAM	528	432
FF	21,614	17,896
BRAM	199	199
DSP	2,734	2,189

表 4 実験 2：消費電力評価

リソース	消費電力[W]	
	従来回路	提案回路（量子化後）
合計（比）	6.274（1.000）	3.925（0.6256）
待機電力（比）	0.310（1.000）	0.287（0.9258）

3.1.1 実験 1：グループ化畳み込みの評価

提案手法では，Residual Block 内にのみグループ畳み込みを適用している。これは，Residual Block の出力を加算する構造により，グループ化畳み込み適用による精度低下の影響が少なくなると考えたためである。比較手法として，Residual Block 外にのみ適用したネットワークを用意し，精度を比較した。表 1 にグループ化畳み込みの適用場所による精度の変化を示す。

3.1.2 実験 2：提案手法の精度評価

実験 2 では，提案手法の有用性を確認するために，従来手法である SRCNN-FP との比較実験を行った。表 2，表 3 および表 4 に PSNR，リソースの利用数および消費電力に関する評価結果を示す。

3.2 結果と考察

表 1 より，提案手法である Residual Block 内にのみ適用する場合が精度とパラメータのトレードオフがとれていることがわかった。そのため，グループ畳み込みの適用場所は Residual Block 内を優先するべきだといえる。

次に表 2 より，精度低下は 0.1~0.2 dB 程度に抑えられていることが確認された。また表 3 より，全体的にリソース数を削減できていることがわかる。

特に，LUT を 40%，FF を 18%，DSP を 20% 削減可能となり，低リソースで FPGA に実装できることが判明した。また表 4 より，消費電力を約 38% 低減する効果を確認した。これより，精度を保ちつつ，消費電力を低減した超解像 CNN を構築できたといえる。

4. まとめ

本稿では，限られた FPGA 上のリソースでリアルタイム処理を実現可能とする SRCNN-FP において，乗算回数の多さから消費電力が高い問題を解決することを目的として，精度を保ちながら低消費電力動作可能なネットワーク構造を提案した。提案ネットワークについて，グループ化畳み込みと固定小数点化を導入することにより，精度の低下を抑制しつつ低消費電力で実現可能な構造とした。提案したネットワークに対して実験評価を行った結果，従来手法と比べて PSNR の低下を約 0.1 dB 程度に抑えつつ，消費電力を約 38% 低減する効果が確認できた。これより，ネットワーク軽量化手法であるグループ畳み込みと固定小数点化が有用であることが確認できた。今後の課題として，提案ネットワークの FPGA 実装が挙げられる。

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[2] Y. Kim, J.-S. Choi, and M. Kim, “A real-time convolutional neural network for super-resolution on FPGA with application to 4K UHD 60fps video services,” IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[3] 森 一紀, “FPGA 実装に向けた CNN のパラメータ数削減”, FIT2020, C-006, 2020 年 9 月.