C-015

CNN のハードウェア実装における全結合層のリソース削減手法に関する一検討
An Approach to Reduce Fully Connected Layer Resources in CNN Hardware

川合勇気†，古川 巧†，黒木修隆†，廣瀬哲也‡，沼 昌宏†
Yuhki Kawai†, Takumi Furukawa†, Nobutaka Kuroki†, Tetsuya Hirose‡, and Masahiro Numa†

1．まえがき

近年，ニューラルネットワークを用いることにより，高精度な画像認識や音声認識が実現されている。なかでも，画像認識分野で有効な畳み込みニューラルネットワーク（CNN：Convolutional Neural Networks）に注目が集まっている。これまで CNN の実装に一般的に用いられてきた GPGPU（General Purpose Graphics Processing Unit）は，エッジ・コンピューティングに適用するには消費電力が大きいという問題がある [1]。低消費電力化実現のため，GPGPU の処理内容を，書き換え可能な FPGA（Field Programmable Gate Array）上のハードウェアで実装する研究が行われている。一方で，FPGA の回路規模には制限があり，ハードウェア化を考慮した演算手法や回路構成が必要となる。

そこで本稿では，高精度かつ低リソースの CNN ハードウェア実現を目的として，重みの 2 のべき乗近似を用いた CNN モデル [2] を対象として，分類問題に直結する全結合層における一部のユニットにのみ 2 のべき乗近似を行うことで，認識精度低下を抑制する手法について検討する。また，全結合層アーキテクチャにおいて，近似を適用するユニット数に対する FPGA リソース数の変化に関する評価結果を報告する。

2. 重みの 2 のべき乗近似を用いた CNN

2.1 近似による回路規模の小型化

CNN は主に 2 種類の層，すなわち入力画像から特徴を抽出して新たな特徴マップを生成する畳み込み層と，ニューロンを模したユニット層間の結合によって信号伝搬を行う全結合層によって構成される。いずれの層においても，膨大な数の積和演算を必要とする。積和演算において，入力との乗算係数は，重みと呼ばれる。ハードウェア化に際して，この積和演算を高速かつ小規模の回路で実現することが求められる。重みを 2 のべき乗で近似し，乗算をシフト演算に置き換えることで，規模の大きな乗算回路の代わりに，マルチプレクサとレジスタのみで構成されるビットシフト回路で演算を実現でき，小型化が期待できる。

2.2 全結合層における演算及び 2 のべき乗近似

全結合層の N 個の層をL_n $(n = 1, 2, \cdots, N)$ と定義する。L_{n-1} 層のユニット数を I, L_nのユニット数を Jとし，L_n層がL_{n-1}層から受け取る入力ユニット値を u_i $(i \in I)$，出力ユニット値を S_j $(j \in J)$，重みを $w_{i,j}$，バイアスを b_j とおくと，全結合層の演算は

$$S_j = \sum_{i \in L_{n-1}} (w_{i,j} \cdot u_i) + b_j \tag{1}$$

と表される。この出力ユニット値が活性化され，次の層への入力となる。ここで，2 のべき乗近似における，べき指数を表す関数を $C(x)$ と定義すると，2 のべき乗近似を行う場合，式 (1) は

$$S_j = \sum_{i \in L_{n-1}} \{u_i \ll C(w_{i,j})\} + b_j \tag{2}$$

と書き換えられる。

† 神戸大学　Kobe University
‡ 大阪大学　Osaka University

3. 提案手法

3.1 各ユニットの優先度と演算簡略化対象の決定

提案手法においては，比較的出力への影響力が小さいと考えられるユニットに対して近似を行う。その影響力の大きさを決めるための指標として，He らによるユニットのプルーニングに関する研究 [3] において提案された *inorm* を用いる。*inorm* は，そのユニットの値を算出するために用いられる重み係数の絶対値の合計によって決定される指標であり，この値が大きいほどユニットの値が大きいことが見込まれる。式 (1) と同様の文字を用いると，L_n層のユニット u_j に関する *inorm* は，

$$inorm(u_j) = \sum_{i \in L_{n-1}} \left| w_{i,j} \right| \tag{3}$$

で表される。この *inorm* の降順にユニットをソートした上で，下位 N_{shift} 個のユニットに対して近似を行うことで，本来の演算結果からの誤差を小さく抑え，本来のモデルの精度維持を図る。ここで，L_nのユニット数 J に対する，2 のべき乗近似適用ユニット数 N_{shift} の割合を，2 のべき乗近似適用比率 $R_{\text{shift}} = N_{\text{shift}} / J$ と定義する。R_{shift} の決定法については，4.1 節において検討を行う。

3.2 提案する全結合層のハードウェア構成

図 1 に提案する全結合層回路の全体構成を示す。バイアスおよび重みといったパラメータについては，オフチップの DRAM より読み出す。並列して各ユニットの演算を行う演算ブロックの並列数を K とする。入力ユニット値 1 つに対して，1 ステップで K ユニットの演算を行う。全結合層の演算は，異なる層についてもユニット数以外については共通しているため，回路の再利用により複数層の全結合層演算を 1 つの回路で実現する。具体的には，全ユニットの演算終了後に，負の値を全て 0 とする ReLU 回路を通した後，オンボードの FIFO に演算結果を書き込む。そして，次の層の演算の入力値とするために，FIFO から値を順次読み出す。

図 2 に，ユニット演算ブロックの構成を示す。ユニット演算ブロック内には 1 つの乗算回路またはビットシフト回路に対して，J / K 個のレジスタに出力ユニットの積和演算結果を格納する。1 ステップごとに制御信号とセレクタを用いて演算する対象ユニットを切り替える動作を行う。この結果，1 層の演算に要するステッ

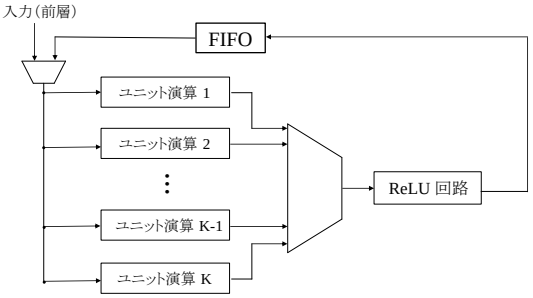

図 1　全結合層回路の全体構成

表 1　1 層の全結合層のみ近似を行った場合の認識精度

R_{shift}	0.25	0.50	0.75	1.00
全結合層 A	71.0%	71.0%	71.0%	71.0%
全結合層 B	70.9%	70.9%	70.9%	71.0%
全結合層 C	70.8%	71.1%	71.1%	70.6%

表 2　全結合層 A にて $R_{shift} = 0.50$ としたときの認識精度

近似対象		全結合層 C				
	R_{shift}	0	0.25	0.50	0.75	1.00
全結合層 B	0	71.0%	71.0%	71.2%	71.0%	70.6%
	0.25	71.0%	71.0%	71.2%	71.0%	70.6%
	0.50	71.0%	71.1%	71.3%	71.0%	70.6%
	0.75	71.0%	71.2%	71.4%	71.1%	70.6%
	1.00	71.0%	70.8%	71.1%	71.2%	70.8%

表 3　全結合層回路の R_{shift} ごとのマッピング結果

R_{shift}	LUT	FF	BRAM
0	37,884（12.5%）	19,252（3.2%）	1（0.1%）
0.25	36,672（12.8%）	19,243（3.2%）	1（0.1%）
0.50	35,052（11.6%）	19,243（3.2%）	1（0.1%）
0.75	34,684（11.4%）	19,276（3.2%）	1（0.1%）
1.00	34,347（11.3%）	19,243（3.2%）	1（0.1%）

プ数は $I \times J / K$ となる。

4. シミュレーション評価と考察

4.1　ソフトウェアシミュレーションによる精度評価

学習済みの畳み込み層 10 層，全結合層 3 層の 13 層からなる CNN を用いて，*inorm* に従って 2 のべき乗近似を各全結合層に適用し，0 から 1 まで 0.25 刻みで R_{shift} を変化させた場合の認識精度をソフトウェアによって比較した。ただし，CNN モデルのプルーニングおよび再学習は行っていない。出力層から遠い全結合層から順に，全結合層 A，全結合層 B，全結合層 C とする。データセットには CIFAR-100 [4] を用い，テストデータ 1,000 例を比較に用いた。ユニット数について全結合層 A と B は 1,024，全結合層 C は 100 とした。近似を適用しなかった場合の認識精度は 70.9% で，すべてに近似を適用した場合は 70.6% であった。途中出力を CNN から抽出し，全結合層の演算をソフトウェア上で行うため，シフト演算を 2 のべき乗近似した重みに対する乗算として扱い，全ての演算を浮動小数点演算で行った。

初めに，いずれか 1 つの全結合層のみ全体の R_{shift} 分だけ 2 のべき乗近似を適用した結果を表 1 に示す。全結合層 A，B については，もとの精度とほぼ変わらない結果となる一方，全結合層 C に関しては R_{shift} によって精度に差が生じ，$R_{shift} = 1$ の時に精度が低下している。次に，全結合層 A の R_{shift} を 0.50 に固定した場合の，その他の全結合層の各 R_{shift} に対する変化を表 2 に示す。全結合層 B の R_{shift} を変化させても認識精度はさほど変化しないが，全結合層 C について $R_{shift} = 1$ の場合，1 層のみ近似した時と同様に精度が低下した。また，全結合層 A の R_{shift} を 0.50 から変化させても，精度は表 2 と大差がなかった。これより，各層ごとに認識精度を高める R_{shift} の値が存在すると考えられる一方，全結合層 C については 2 のべき乗近似を多数のユニットに対して行うと，大きな精度低下を招く可能性がある。

ここで，*inorm* と演算精度の関係について考察を加える。2 のべき乗近似の影響を大きく受けうる要因として，重みの大きさと分布のばらつきが考えられる。そこで，各層の全ユニットの *inorm* に関する平均値 μ を求めたところ，全結合層 A, B, C それぞれ μ = 14.6, 7.8, 22.3 となり，各 *inorm* の標準偏差 σ は，全結合層 A, B, C それぞれ σ = 1.48, 6.55, 1.29 であった。そこで，分布の広がりを表す変動係数 σ/μ を求めると，全結合層 A, B, C それぞれ σ/μ = 0.104, 0.839, 0.056 となった。変動係数に関して，全結合層 B が最大，全結合層 C が最小となった。変動係数が小さい全結合層 C はユニットごとの重みに関する変動が小さく，2 のべき乗近似を適用するとその重みの差が消失，もしくは過度に増幅され，認識精度低下に繋がると考えられる。

4.2　FPGA へのマッピングによる評価

3 章で述べた全結合層において，最大ユニット数を 1,024 とし，$K = 128$ とした回路を設計し，各 R_{shift} に対するリソース利用数を評価した結果を表 3 に示す。論理合成には Xilinx 社の Vivado 2016.4 を用いて，同社の VC707 評価ボードに搭載する FPGA（Virtex7：XC7VX485T-2FFG1761C）にマッピングを行った。ただし，DSP スライスは用いていない。表 3 から，R_{shift} を増加させると LUT 利用数が削減できていることが確認できた。

5. まとめ

本稿では，重みの 2 のべき乗近似を用いた CNN モデルに関して，全結合層における近似を重み係数の絶対値 *inorm* に基づいて決定した特定のユニットにのみ適用する手法について検討を行うとともに，そのハードウェア構成を提案した。

提案した手法をソフトウェアシミュレーションにて評価した結果，特定の層では近似によって認識精度が低下するが，R_{shift} を適切に設定することで，全てに近似を適用した場合と比べて認識精度向上が実現できることを確認した。また，ハードウェアの FPGA へのマッピング結果から，R_{shift} を増加させることでリソースの利用数を削減できることが確認できた。

今後の課題として，全結合層回路の DRAM アクセス方式に関する検討が挙げられる。

参考文献

[1] H. Terada and H. Shouno, “B-DCGAN: Evaluation of Binarized DCGAN for FPGA”, https://arxiv.org/abs/1803.10930

[2] 宇都宮誉博, “2 のべき乗近似とプルーニングを用いた CNN 向け FPGA アクセラレータ”, 信学技報, RECONF2017-70, 2017.

[3] T. He, Y. Fan, Y. Qian, T. Tan, and K. Yu, “Reshaping deep neural network for fast decoding by node-pruning,” 2014 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[4] A. Krizhevsky, “The CIFAR-100 dataset”, http://www.cs.toronto.edu/~kriz/cifar.html

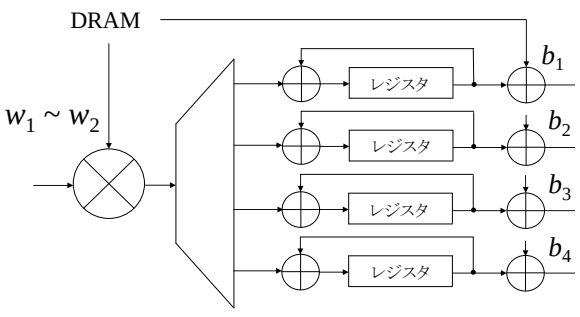

図 2　ユニット演算ブロックの構造