2 値化ニューラルネットワークにおける ポップカウントのシフトレジスタ実装

Shift Register Implementation of Pop-Count in Binarized Neural Network

中道 凌　　石浦 菜岐佐
Ryo Nakamichi　　Nagisa Ishiura

関西学院大学 理工学部　School of Science and Technology, Kwansei Gakuin University

1　はじめに

2 値化ニューラルネットワーク (BNN) は, ニューロンの入出力と重みを 2 値に制限したニューラルネットワークである [1]. BNN はハードウェア化した場合の回路規模を大幅に削減するが, それでもなお回路規模は大きい. 本稿では, BNN のニューロンにおける逐次的なポップカウント処理をシフトレジスタで実装することによって, 回路規模を削減する手法を提案する.

2　2 値化ニューラルネットワーク (BNN)

ニューラルネットワークのニューロン i に入力しているニューロンの集合を I_i とし, i のバイアスを b_i, ニューロン j から i のシナプス結合に関する重みを $w_{i,j}$ とする. i の活性値 a_i は $a_i = f(b_i + \sum_{j \in I_i} w_{i,j} \cdot a_j)$ と表現できる. BNN では, a_i と $w_{i,j}$ は $\{-1, +1\}$ に制限され, i の出力 $f(x)$ は, $0 \leq x$ ならば $f(x) = +1$, そうでなければ $f(x) = -1$ となる.

-1 と $+1$ をそれぞれ 0 と 1 で符号化すると, 上式における乗算は排他的論理和否定 (EXNOR) 演算に置き換えられる. ニューロンの出力は, 全入力について EXNOR 演算結果が 1 であるものを数え (ポップカウント), それがしきい値 t_i を超えるか否かにより決定できる. この計算を入力に対して逐次的に行う場合, ニューロンの回路は 図 1 に示すように, EXNOR ゲート, 加算器, レジスタ, および比較回路で構成できる.

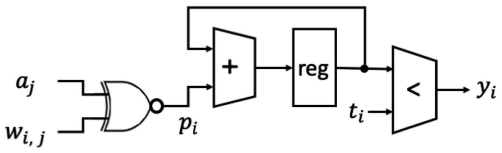


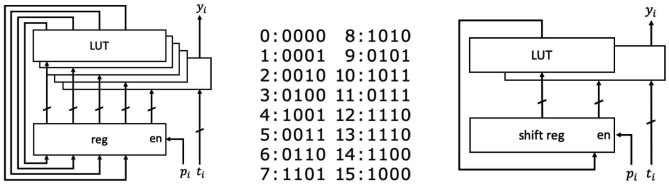
図 1　BNN のニューロンの回路

3　シフトレジスタを用いたポップカウントの実装

本稿では, ポップカウントを加算器ではなくシフトレジスタで実装することにより, FPGA 実装時の LUT 数を削減する手法を提案する.

計数の表現に通常の 2 進符号を用いると, 図 2 (a) のように, 加算器の実装にレジスタのビット数の分だけ LUT が必要になる. これに対し, (b) のような符号化を用いれば, 計数処理はシフトレジスタにより実装でき, (c) のように LUT 数を削減できる.

ただし, レジスタのビット数が増えると, 比較回路が複雑になって LUT 数が増えることがある. これは, レジスタを 4 ～ 5 ビット程度に分割する構成をとることにより回避できる.



(a)　(b)　(c)

図 2　ポップカウントの FPGA 実装

表 1　2 値化 LeNet5 [2] の LUT 数

層	入力 ×出力	従来法	本手法
1	25× 6	17.25	15.43
3	150× 16	40.94	37.60
5	400× 120	32.10	26.60
6	120× 84	16.85	13.88
7	84× 10	16.32	15.06
全体	–	38,151	36,654

Synthesizer: Xilinx Vivado (2016.4)
Target: Xilinx Artix-7 (xc7a100tcsg324-3)

4　実装結果

提案手法に基づき, パラメータ埋め込み型の 2 値化 LeNet5 [2] を実装した. 8 ～ 9 ビットのカウンタを要するニューロンは, 4 ～ 5 ビットのカウンタを組み合わせて実装した. 論理合成は Xilinx Artix-7 をターゲットに Vivado (2016.4) で行った. BNN のパラメータ (重みとしきい値) には乱数を使用した.

合成結果 (LUT 数) を 表 1 に示す. 1 ～ 7 層の LUT 数は 100 個のニューロンの平均である. この BNN の実装では, 全パラメータをの記憶に多数の LUT を使用するため, 従来法に対する回路規模削減効果は約 4% であるが, パラメータを RAM で記憶する方式ではより大きな効果が得られると考える.

5　むすび

本稿では, ポップカウントのシフトレジスタ実装による BNN の回路規模削減手法を提案した. カウンタ回路の更なる最適化が今後の課題である.

謝辞 本研究は一部キオクシア (株)「奨励研究」の助成による.

参考文献

[1] M. Courbariaux, et al.: “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1,” *Computer Research Repository*, arXiv:1602.02830 (Mar. 2016).

[2] R. Sugimoto and N. Ishiura: “Parameter embedding for FPGA implementation of binarized neural networks,” in *Proc. SASIMI 2019*, pp. 41–45 (Oct. 2019).