

Bit 注入カウンタ型ニューロンを用いた SNN の推論性能評価

2110095 岩瀬レオ 成見研究室

1 はじめに

スパイキングニューラルネットワーク (SNN) は、ニューロン間でスパイクと呼ばれる短い信号を伝達するニューラルネットワークである。スパイクの到来に応じて処理するイベント駆動実装により、低消費電力な推論が期待される。

Chandarana らは、学習済みネットワークを SNN に変換し、専用チップ Loihi で画像分類を行った [1]。論文の報告条件では、NCS2 上の通常のニューラルネットワークと比較して、分類精度はやや低下し、推論時間は増加する一方、推論エネルギーを削減している。

ニューロンに到来するスパイクの処理方式として、逐次処理と並列処理がある。非同期回路の μ Brain は、入力を保持し、1 件ずつ選択して重みを積算する構成で、学習済み SNN の推論を実証している [2]。一方、Li・Zhang は、FPGA 上で 400 入力の重みを加算器ツリーにより並列に合算する構成を用いている [3]。

2 研究目的

入力を 1 件ずつ処理する方式では、未処理の入力を保持する必要がある。入力が集中すると処理待ちが増加し、保持容量を超えた入力を失う場合がある。そこで、Gomes らのカウンタ型ニューロン [4] に着想を得て、入力の保持や処理順の調整を介さず、重みの各 bit に応じてカウンタの各桁を直接更新する方式を提案する。この方式では入力衝突による誤計数が課題となる。

本研究の目的は、提案する bit 注入カウンタ型ニューロンを用いた学習済みネットワークの推論性能を評価することである。比較用に、入力を保持して逐次処理する加算器方式も実装し、入力競合による推論精度への影響、回路資源量および推論時の消費電力を調べる。

3 ニューロン回路の構成

本研究は、漏れのない IF (Integrate-and-Fire) ニューロンを対象とする。膜電位と重みを符号なし整数とし、興奮性入力のみを扱う。入力ごとの重みを膜電位に加算し、閾値以上になると発火して膜電位をリセットする動作を基本とする。比較する 2 つの方式は、どちらもニューロンの動作にグローバルクロックを用いない。

3.1 bit 注入カウンタ方式

図 1 に全体構成を示す。入力ごとの遅延・重みマスク回路から、各桁に向かう信号を XOR で合流し、カウンタを駆動する。カウンタが保持する膜電位 V を閾値 V_θ と直接比較し、閾値以上になるとパルス生成器が出力スパイクを生成する。発火後は一定期間カウンタをクリア状態に保つ。

入力 i の W bit 重みを

$$w_i = \sum_{k=0}^{W-1} w_{i,k} 2^k, \quad w_{i,k} \in \{0, 1\} \quad (1)$$

と表す。第 k 桁を起点にカウンタを 1 回インクリメントすると、衝突やリセットがない条件では計数値が 2^k 増加する。この性質を用いて、重みの各 bit に対応した入力経路を設ける。

図 2(a) のように、各入力に専用の遅延列を設け、入力パルスを上位桁から順に遅延させる。図中の Δ は隣接する桁への注入間隔を表す。各経路は重み bit との AND で有効・無効を切り替える。図 2(b) で同じ桁に向かう信号を XOR で合流し、図 2(c) の T-FF からなるカウンタを駆動する。下位桁の反転出力を桁上がり用に用いる。

重みを W bit、閾値を $1 \leq V_\theta \leq 2^W$ とし、内部の膜電位は $W + 1$ bit で保持する。図 2(c) では $W = 4$ であり、 q_4 を含めた 5 bit を閾値判定に用いる。複数入力や桁上りを仲裁せずに各桁へ作用させるため、信号の競合により誤計数が生じ得る。

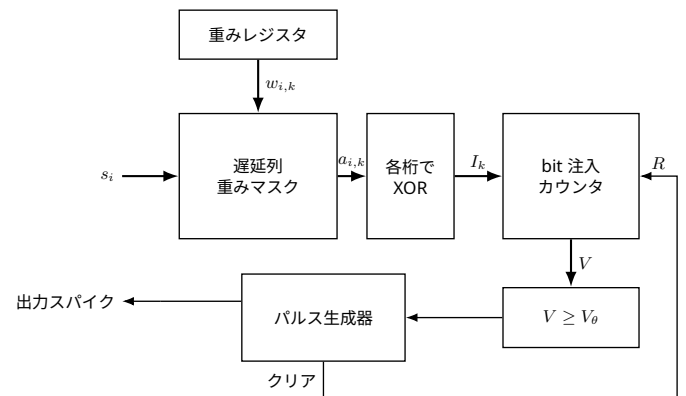
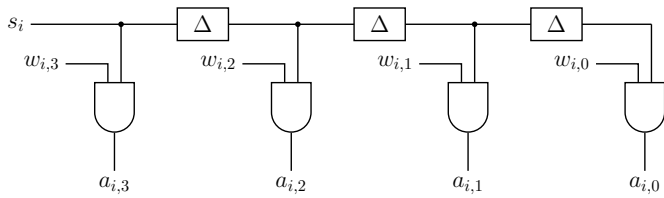
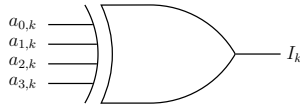
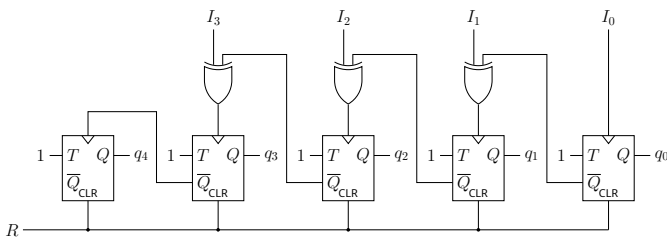


図 1 bit 注入カウンタ方式の全体構成

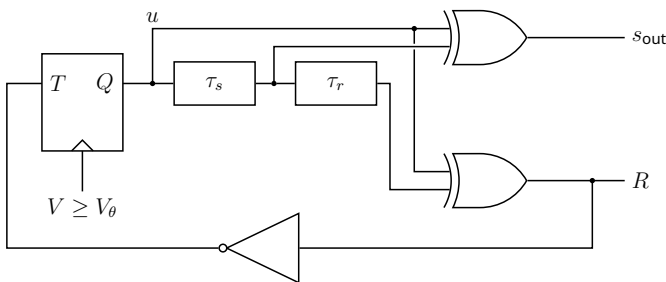
(a) 入力 i の遅延経路と重みマスク(b) 第 k 桁への入力の合流

(c) 各桁への入力と桁上がり

s_i : 入力パルス, $w_{i,k}$: 重みの各 bit
 $a_{i,k}$: 遅延・マスク後の信号
 I_k : 同じ桁の信号を XOR で合流した入力
 q_k : 膜電位の第 k 桁, R : 非同期クリア
 Δ : 隣接する桁への注入間隔

図 2 入力経路とカウンタ (4 入力, 重み 4 bit)

図 3 はパルス生成器の構成を示しており、閾値到達時に FF の保持値を反転し、遅延前後の信号の XOR から出力スパイクとクリア信号を生成する。出力スパイク終了後に長さ τ_r の不応期を設け、その終了後にカウンタのクリアを解除して膜電位更新を再開する。発火後は不応期が終了するまで FF の反転を抑える。



u : FF の保持値, s_{out} : 出力スパイク
 R : クリア信号, τ_s : 出力スパイク幅
 τ_r : 不応期の長さ

図 3 パルス生成器の構成

表 1 2 方式の入力処理の違い

項目	bit 注入カウンタ	加算器
入力保持	なし	各入力 1 件
状態更新	各桁への bit 注入	1 件ずつ加算
発火判定	計数値を直接比較	加算結果を比較
入力集中時	信号の競合	保持容量の超過

3.2 加算器方式

各入力に 1 件分の保持枠を設け、固定優先順位で 1 件を選ぶ。重み全体を膜電位に加算し、遅延列で結果の収束を待つ。閾値判定と状態更新を行う。発火時に保持した入力を破棄する。不応期中の入力と、保持枠が埋まった入力への再入力 は受理しない。

3.3 入力処理方式の比較

表 1 に両方式の違いを示す。カウンタ方式では計数値の変化に応じて、加算器方式では 1 件の加算結果で発火を判定する。

4 進捗

両方式を SystemVerilog で記述し、DE10-Nano Rev. E に実装した。4 入力、重み 4 bit、閾値 16 とし、重みは停止中に書き込み、動作中は固定した。

動作の確認として、重みを 3 種類に設定し、各設定で 8 ケースの入力を与えた。入力同士の衝突が起らないように入力間隔を十分に空けた条件で、両方式とも全 24 ケースの最終膜電位と発火数が期待値に一致した。

5 今後の評価

単体回路では入力間隔を変え、膜電位と発火数の誤差を調べる。各方式の発火・不応期・入力受理の仕様を反映した参照モデルを定め、方式間のモデル差と回路の誤計数を区別する。また、遅延段数を調整した上で資源量を比較する。

その後には、非負整数重みなどの回路の制約に対応した学習済みネットワークを構成し、入力競合が推論精度に与える影響と推論時の消費電力を評価する。

参考文献

- [1] P. Chandarana et al., "Energy-Efficient Deployment of Machine Learning Workloads on Neuromorphic Hardware," IGSC, 2022. arXiv:2210.05006v2.
- [2] J. Stuijt et al., "μBrain: An Event-Driven and Fully Synthesizable Architecture for Spiking Neural Networks," Front. Neurosci., vol. 15, 664208, 2021.
- [3] Q. Li and L. Zhang, "System-level FPGA validation of a trainable and robust multiplier-free spiking neural network," iScience, vol. 29, 115985, 2026.
- [4] E. O. Gomes and D. Rontani, "Scalable neuromorphic computing from autonomous spiking dynamics in a clockless reconfigurable chip," arXiv:2605.16114v1, 2026.