



І. Г. Цмоць, Б. В. Штогрінець, І. Я. Казимира, А. А. Литвин
Національний університет "Львівська політехніка", м. Львів, Україна

МОДЕЛЬ ТА МЕТОД СИНТЕЗУ НЕЙРОННОГО ЕЛЕМЕНТА ПАРАЛЕЛЬНО-ПОТОКОВОГО ТИПУ

Розроблено узагальнену модель нейронного елемента паралельно-потокowego типу, основними компонентами якої є перетворювач форматів даних, пристрій обчислення скалярного добутку та блок реалізації функції активації, яка забезпечує високу ефективність використання обладнання внаслідок вибору кількості розрядів для формування групових часткових добутків і узгодження частоти введення даних з частотою роботи конвеєра. Проаналізовано алгоритми для обчислення скалярного добутку та обґрунтовано вибір паралельного вертикально-групового методу обчислення скалярного добутку з аналізом k розрядів для формування групових часткових добутків. Для групового підсумовування вибрано каскадний паралельно-паралельний алгоритм, який забезпечує конвеєризацию процесу підсумовування шляхом введення буферних регістрів між ярусами суматорів. Розроблення нейронного елемента паралельно-потокowego типу запропоновано виконувати на підставі таких принципів: використання базису елементарних арифметичних операцій; організація процесу обчислення скалярного добутку як виконання єдиної операції; локалізація та спрощення зв'язків між сходишками конвеєра; забезпечення балансу між введенням-виведенням і обчисленнями; використання конвеєризації та просторового розпаралелювання процесу обчислення; мінімізація зовнішнього інтерфейсу зв'язку. Запропоновано оцінювання апаратної реалізації нейронного елемента здійснювати за критерієм ефективності використання обладнання, який враховує складність, час і витрати обладнання на реалізацію алгоритму та дає оцінку внеску кожного вентиля в загальну продуктивність нейронного елемента. Показано, що синтез нейроелемента паралельно-потокowego типу зводиться до мінімізації апаратних затрат при забезпеченні обчислення скалярного добутку та функції активації в реальному часі. Розроблено метод синтезу нейронного елемента паралельно-потокowego типу, який, за рахунок розбиття його на сходишки та узгодження часу надходження даних з тактом конвеєра, забезпечує синтез цього нейронного елемента з високою ефективністю використання обладнання та обробленням потоків даних у реальному часі.

Ключові слова: скалярний добуток; групове підсумовування; сходишка конвеєра; метод; структура.

Вступ / Introduction

Сучасний етап розвитку нейромережових технологій характеризується розширенням галузей застосування, значна частина з яких вимагає опрацювання в реальному часі інтенсивних потоків даних на комп'ютерних засобах, що задовольняють обмеженням відносно габаритів, маси, енергоспоживання та мають високу ефективність використання обладнання [9, 10, 34]. Нейромережові засоби, які відповідають таким вимогам, широко використовують у промисловості (управління технологічними процесами та складними об'єктами) [2, 28], енергетиці (оптимізація навантаження в електромережах) [11, 22], військовій справі (технічний зір, управління рухом мобільного робота) [19, 29], транспорті (управління рухом і двигуном), медицині (діагностика захворювань) [27, 36] і приладобудуванні (розпізнаван-

ня образів і оптимізація управління) [5, 13]. Створення таких нейромережових засобів вимагає широкого використання сучасної елементної бази (напівзамовних і замовних надвеликих інтегральних схем (НВІС)), розроблення нових методів, алгоритмів і НВІС-структур для реалізації базових компонент (нейронних елементів, пристроїв обчислення скалярного добутку, групового підсумовування) нейромереж реального часу.

Для нейромережового опрацювання інтенсивних потоків даних у реальному часі в основному використовують конвеєрні нерекурсивні нейронні елементи, пристрої обчислення скалярного добутку та групового підсумовування, особливістю яких є відсутність обернених зв'язків й отримання результату при проходженні даних із входу на вихід. Конвеєризація передбачає розподіл пристроїв на сходишки, кожна із яких складається із

Інформація про авторів:

Цмоць Іван Григорович, д-р техн. наук, професор, кафедра автоматизованих систем управління.

Email: ivan.tsmots@gmail.com; <https://orcid.org/0000-0002-4033-8618>

Штогрінець Богдан Володимирович, аспірант, кафедра автоматизованих систем управління.

Email: bohdan.v.shtohrinets@lpnu.ua; <https://orcid.org/0009-0001-4956-3862>

Казимира Ірина Ярославівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління.

Email: iryna.y.kazymyra@lpnu.ua; <https://orcid.org/0000-0003-1597-5647>

Литвин Андрій Андрійович, студент, кафедра інформаційних технологій видавничої справи.

Email: andriylytvyn123@gmail.com; <https://orcid.org/0009-0008-4993-7132>

Цитування за ДСТУ: Цмоць І. Г., Штогрінець Б. В., Казимира І. Я., Литвин А. А. Модель та метод синтезу нейронного елемента паралельно-потокowego типу. Науковий вісник НЛТУ України. 2023, т. 33, № 2. С. 92–100.

Citation APA: Tsmots, I. H., Shtohrinets, B. V., Kazymyra, I. Ya., & Lytvyn, A. A. (2023). Model and method for synthesis of neural element of parallel-streaming type. *Scientific Bulletin of UNFU*, 33(2), 92–100. <https://doi.org/10.36930/40330213>

двох компонент – буферної пам'яті та операційного блоку. Конвеєрні нерекурсивні пристрої обчислення скалярного добутку можна розділити на два класи: перший – матричні, які використовують паралельне формування та підсумовування всіх часткових добутків, другий – паралельно-потоківі, які використовують послідовне формування та додавання макрочасткових добутків із відповідним зсувом. Основним недоліком конвеєрних матричних пристроїв обчислення скалярного добутку є неоднорідність структури та нерегулярність зв'язків, що ускладнює НВІС-реалізацію. Конвеєрні пристрої паралельно-потоківі обчислення скалярного добутку використовують часове розпаралелення процесу обчислення, мають однорідну структуру з регулярними зв'язками та більше орієнтовані на НВІС-реалізацію. У таких пристроях управління процесом обчислення скалярного добутку зводиться до видачі тактових імпульсів, які просувають інформацію із входу на вихід, записуючи проміжні результати в буферну пам'ять. Частота надходження тактових імпульсів визначається часом звертання до буферної пам'яті та часом затримки на операційному блоці. Для забезпечення високої ефективності використання обладнання всі сходинок конвеєра повинні виконувати приблизно однакові за складністю операції, а тривалість введення вхідних даних і вагових коефіцієнтів, такт роботи конвеєра та тривалість виведення результатів повинні бути узгоджені. Кількість сходінок конвеєра залежить від таких параметрів: розрядності вхідних даних; кількості розрядів, які аналізуються для формування макрочасткових добутків; часу надходження вхідних даних і вагових коефіцієнтів.

Для паралельно-потоківі НВІС-реалізації алгоритмів базових компонент нейромереж вони повинні бути добре структурованими, рекурсивними та локально залежними. У рекурсивному алгоритмі всі сходинок конвеєра повинні виконувати приблизно однакові операції. При реалізації рекурсивного алгоритму кожна із сходінок конвеєра буде повторювати виконання фіксованого набору операцій над послідовністю даних, що надходять. Ефективність відображення алгоритму на апаратуру безпосередньо зв'язана зі способом декомпозиції алгоритму на незалежні операції, що виконуються паралельно, або на залежні, що виконуються у конвеєрному режимі. Паралельно-потоківі реалізація нейромереж забезпечує високу ефективність завантаження апаратних засобів і найбільше відповідає умовам роботи у реальному часі.

Тому актуальним завданням є розроблення високо-ефективних компонент для синтезу паралельно-потоківіх нейромереж реального часу з високою ефективністю використання обладнання.

Об'єкт дослідження – процеси моделювання та обчислення алгоритмів базових компонент нейронних мереж реального часу.

Предмет дослідження – моделі, методи, алгоритми та структури реалізації базових компонент нейронних мереж реального часу для оброблення інтенсивних потоків даних.

Мета роботи – розробити узагальнену модель, алгоритми і структуру нейронного елемента паралельно-потоківі типу.

Для досягнення зазначеної мети визначено такі *основні завдання дослідження*:

- розробити узагальнену модель нейронного елемента паралельно-потоківі типу;
- розробити алгоритм апаратно-орієнтованих методів обчислення скалярного добутку та групового підсумовування;
- розробити базову структуру нейронного елемента паралельно-потоківі типу;
- розробити метод синтезу нейронного елемента з паралельно-потоківі обробленням інтенсивних потоків даних у реальному часі.

Аналіз останніх досліджень та публікацій. Аналіз засобів реалізації нейромереж показує [25, 31, 32, 37], що переважна кількість засобів, які використовують для реалізації нейромереж є програмними. Основним недоліком програмних засобів є невисока швидкодія, яка не забезпечує опрацювання інтенсивних потоків даних. Підвищення швидкодії програмних засобів можна досягнути з використанням високопродуктивних комп'ютерних систем, недоліком яких є те, що вони не відповідають обмеженням щодо габаритів й енергоспоживання. Тому для нейромережевого опрацювання інтенсивних потоків даних у реальному часі доцільно використовувати апаратну НВІС-реалізацію, яка одночасно забезпечує малі габарити та високу швидкодію.

Основними компонентами, на базі яких синтезуються апаратні нейромережі, є нейронні елементи. Існують різні моделі формального нейрона, які відрізняються способами надходження вхідних даних і вагових коефіцієнтів, способами обчислення сигналу постсинаптичного збудження та функції активації [12, 18, 35]. У роботах [1, 3, 26] розглянуто паралельно-вертикальні та вертикально-групові моделі формального нейрона, особливістю яких є їх орієнтація на НВІС-реалізацію та на узгодження інтенсивності надходження даних із інтенсивністю обчислень. Недоліком розглянутих моделей формального нейрона є невисока швидкодія, яка істотно залежить від розрядності операндів.

Характеристики нейроелемента значно залежать від підходів до апаратної реалізації операції обчислення скалярного добутку. Аналіз робіт [6, 8] показує, що існують два підходи до апаратної реалізації операції обчислення скалярних добутків. Перший з них ґрунтується на операціях множення, додавання, а другий – на елементарних арифметичних операціях додавання, інверсії та зсуву. Перший підхід, переважно, застосовують до обчислення скалярного добутку як до сукупності операцій множення і додавання. Такий підхід не забезпечує оптимізації структури пристрою та його часових параметрів. Використання базису елементарних арифметичних операцій та багатооперандного підходу, при якому обчислення скалярного добутку розглядається як виконання єдиної операції, забезпечує оптимізацію структури пристрою за швидкодією та апаратними витратами. Основою алгоритмів обчислення скалярного добутку з використанням багатооперандного підходу та базису елементарних арифметичних операцій є формування часткових добутків з наступним їх додаванням.

Проведений аналіз структур пристроїв [17, 23, 24], які використовують для реалізації алгоритмів обчислення скалярного добутку з використанням багатооперандного підходу та базису елементарних арифметичних операцій показав, що для реалізації використовують два типи структур: рекурсивні та нерекурсивні. Структурною особливістю рекурсивних пристроїв є присутність обернених зв'язків. У таких пристроях обчислення скалярного добутку здійснюється за декілька ітерацій,

кількість яких визначається алгоритмом формування часткових добутоків. Недоліком рекурсивних пристроїв обчислення скалярного добутку є відносно невисока швидкодія.

Більшу швидкодію мають нерекурсивні пристрої, особливістю яких є відсутність обернених зв'язків [7]. Такі пристрої поділяють на два класи: перший – матричні, які використовують паралельне формування та підсумовування всіх часткових добутоків; другий – паралельно-потоківі, які використовують послідовне формування та додавання із відповідним зсувом часткових добутоків. Недоліком матричних пристроїв обчислення скалярного добутку є неоднорідність структури. Паралельно-потоківі пристрої обчислення скалярного добутку мають однорідну структуру з регулярними зв'язками і більше орієнтовані на НВІС-реалізацію.

З аналізу публікацій випливає, що синтез високоєфективних нейроелементів вимагає розроблення нових моделей формального нейрона, методів, алгоритмів, НВІС-структур обчислення скалярного добутку та групового підсумовування

Результати дослідження та їх обговорення / Research results and their discussion

Узагальнена модель та принципи синтезу нейронного елемента паралельно-потоківого типу. При паралельно-потоківому опрацюванні даних у нейромережах вхідні дані $X_j, j = \overline{1, N}$ та вагові коефіцієнти $W_j, j = \overline{1, N}$, де N – кількість входів даних і вагових коефіцієнтів) неперервно надходять одночасно на всі входи паралельним двійковим кодом. Особливістю паралельно-потоківого опрацювання є його орієнтація на опрацювання неперервних потоків даних у реальному часі. Основними компонентами, на базі яких синтезуються нейромережі, є нейронні елементи, які функціонують згідно з формулою

$$y = f \left(\sum_{j=1}^N W_j X_j \right), \quad (1)$$

де: y – вихідний сигнал нейронного елемента; X_j – вхідні дані; W_j – вагові коефіцієнти; N – кількість входів даних і вагових коефіцієнтів; $f(\cdot)$ – функція активації, $j = \overline{1, N}$. З формули (1) випливає, що робота нейронного елемента ґрунтується на таких математичних операціях обчислення скалярного добутку y та функції активації $f(Z)$. Для забезпечення паралельно-потоківого опрацювання даних у нейромережі потрібно, щоб нейронні елементи працювали за конвеєрним принципом. Конвеєризація при реалізації нейронного елемента передбачає розділення його на сходинки шляхом введення буферної пам'яті. При цьому кожна сходинка конвеєра повинна складатися з двох компонент – операційних блоків і буферної пам'яті.

Обчислення скалярного добутку в нейронному елементі паралельно-потоківого типу виконується на базі однотипних операцій, які зводяться до формування макрочасткових добутоків P_{Mh} і їх додавання до раніше накопичених сум згідно з формулою

$$Z_h = 2^{-h} Z_{h-1} + P_{Mh}, \quad (2)$$

де $Z_0 = 0$. Операція формування макрочасткового добутку P_{Mh} та його додавання за формулою (2) реалізуються h -тою сходинкою конвеєра $h = \overline{1, m}$ де $m = \lceil n/k \rceil$, $\lceil \cdot \rceil$ – знак заокруглення до більшого цілого числа, n – розряд-

ність операндів; k – кількість розрядів, які аналізуються для формування макрочасткових добутоків. Кількість сходинок конвеєра визначається розрядністю операндів n і кількістю розрядів k , які аналізуються для формування макрочасткових добутоків P_{Mh} . Формування h -го макрочасткового добутку P_{Mh} , виконується підсумовуванням групових часткових добутоків P_{jh} , які отримують при аналізі k розрядів відповідно до формули

$$P_{Mh} = \sum_{j=1}^N \sum_{l=1}^k 2^{-(l-1)} W_j X_{jhl} = \sum_{j=1}^N P_{jh}. \quad (3)$$

Для паралельно-потоківого обчислення скалярного добутку доцільно використовувати алгоритми з аналізу молодших розрядів для формування часткових добутоків. Від кількості розрядів k , що аналізуються, та алгоритмів формування групових часткових добутоків P_{jh} залежить структура сходинки конвеєра.

Узагальнена аналітична модель нейронного елемента паралельно-потоківого типу запишеться так:

$$y = f_{a(Z)}(f_{3nZ}(f_{PMn}(f_{(P_{m1}, \dots, P_{mN})}(f_{3nm}(X_1, \dots, X_N, W_1, \dots, W_N))^{(\dots)}))^{(\dots)}))^{(\dots)}). \quad (4)$$

де: y – вихід нейронного елемента; $f_{a(Z)}$ – функція активації; f_{3nZ} – буферне запам'ятовування результату скалярного добутку; f_{PMn} – обчислення m -го макрочасткового добутку; $f_{(P_{m1}, \dots, P_{mN})}$ – формування групових часткових добутоків у m -ий конвеєрній сходинці; $f_{3nm}(X_1, \dots, X_N, W_1, \dots, W_N)$ – буферне запам'ятовування операндів у m -ий конвеєрній сходинці; f_{Z1Rk} – обчислення 1-го часткового результату $Z_1 = P_{M(n-1)}$, $f_{P_{M1}}$ – обчислення 1-го макрочасткового добутку; $f_{(P_{11}, \dots, P_{1N})}$ – формування групових часткових добутоків у 1-ий конвеєрній сходинці; $f_{3m}(X_1, \dots, X_N, W_1, \dots, W_N)$ – буферне запам'ятовування операндів у 1-ий конвеєрній сходинці; $f_{(qW \rightarrow NW, qX \rightarrow NX)}$ – послідовно-паралельне формування даних для паралельно-потоківого обчислювача скалярного добутку, тривалість якого залежить від частоти надходження даних F_d , кількості каналів q_d та їх розрядності n_d .

Структуру узагальненої моделі паралельно-потоківого нейронного елемента, яка реалізує вираз (3), подано на рис. 1. Основними компонентами цієї моделі є: перетворювач форматів даних, який шляхом узгодження частоти введення даних з частотою роботи конвеєра забезпечує мінімізацію кількості виводів інтерфейсу; m однотипних сходинок конвеєра, в кожній з яких обчислюється частковий результат відповідно до формули (2); блок реалізації функції активації. Особливістю цієї моделі є суміщення у часі одночасного опрацювання m вхідних масивів даних. Основним шляхом підвищення швидкодії роботи паралельно-потоківого нейроелемента є зменшення конвеєрного такту роботи.

Апаратно-орієнтовані алгоритми для реалізації скалярного добутку та групового підсумовування. Обчислення скалярного добутку зводиться до виконання N множень і $N-1$ додавань. Операція множення є складною і для своєї реалізації вимагає значних апаратних затрат. Існують два підходи до апаратної реалізації обчислення скалярного добутку. Перший з них ґрунтується на операціях множення та додавання, другий – на операціях додавання та зсуву. Перший підхід, переважно, використовують при синтезі пристроїв обчислення скалярних добутоків на базі окремих мікросхем

(помножувачів, суматорів), а другий – при реалізація на програмованих логічних інтегральних схем (ПЛІС). Тому для ефективної апаратної реалізації операції скалярного добутку на ПЛІС типу FPGA необхідно розробити апаратно-орієнтовані методи та алгоритми. Такі методи та алгоритми повинні використовувати базис елементарних арифметичних операцій, бути добре структурованими та виконувати обчислення, як виконання єдиної операції. Основою апаратно-орієнтованих

методів обчислення скалярних добутків є операції формування та підсумовування часткових добутків P_{ji} . Алгоритми формування часткових добутків P_{ji} можна розділити на дві групи: з прямим формуванням часткових добутків і з формуванням часткових добутків з використанням попередніх обчислень. Для апаратної реалізації на базі ПЛІС доцільно використовувати пряме формування часткових добутків, оскільки вони є регулярними та добре структурованими.

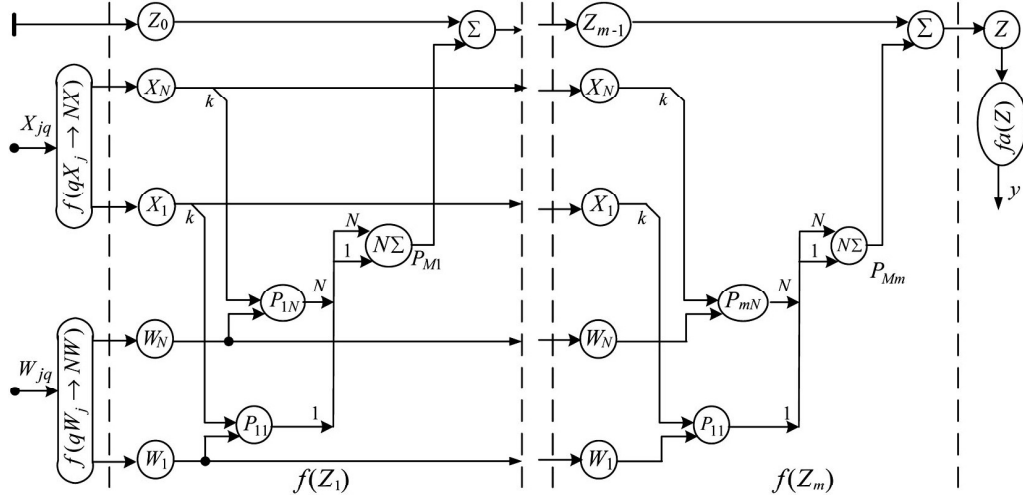


Рис. 1. Узагальнена модель нейронного елемента паралельно-потокowego типу / Generalized model of neural element of parallel-streaming type

Алгоритм обчислення скалярного добутку з аналізом одного розряду. Алгоритм обчислення скалярного добутку з аналізом одного розряду записуємо так:

$$Y = \sum_{j=1}^N W_j X_j = \sum_{i=1}^n 2^{-i} \sum_{j=1}^N W_j X_{ji} = \sum_{i=1}^n 2^{-i} \sum_{j=1}^N P_{ji} = \sum_{i=1}^n 2^{-i} P_{Mi}, \quad (5)$$

де: P_{ji} – i -ий частковий добуток, який визначається так:

$$P_{ji} = W_j X_{ji} = \begin{cases} 0, & \text{якщо } X_{ji} = 0; \\ W_j, & \text{якщо } X_{ji} = 1, \end{cases} \quad i = \overline{1, n}, j = \overline{1, N}, \quad (6)$$

P_{Mi} – i -ий макрочастковий добуток, який обчислюють:

$$P_{Mi} = \sum_{j=1}^N P_{ji}. \quad (7)$$

З формули (5) випливає, що обчислення скалярного добутку виконується за n тактів, в кожному i -му такті виконуються такі операції:

- формування N часткових добутків P_{ji} для i -го розрядного зрізу X_{ji} ;
- формування i -го макрочасткового добутку P_{Mi} відповідно до формули (7);
- підсумовування макрочасткових добутків P_{Mi} згідно з виразом $Y_i = 2^{-1} Y_{i-1} + P_{Mi}$, де $Y_0 = 0$.

Обчислення скалярного добутку за модифікованим алгоритмом Бута. Модифікований алгоритм Бута передбачає розбиття множників X_j на групи з трьох розрядів, так що сусідні групи мають один спільний розряд. Розбиття множників X_j на групи здійснюється з молодших розрядів, причому молодший розряд наймолодшої групи завжди доповнюється нулем. Обчислення скалярного добутку за модифікованим алгоритмом Бута записується так:

$$\begin{aligned} Y &= \sum_{j=1}^N W_j X_j = \sum_{i=g}^p 2^{-2g} \sum_{j=1}^N W_j (X_{j(g+1)} X_{jg} X_{j(g-1)}) = \\ &= \sum_{i=g}^p 2^{-2g} \sum_{j=1}^N P_{jg} = \sum_{i=1}^p 2^{-i} P_{Mg}, \end{aligned} \quad (8)$$

де $p = \lceil (n+1)/2 \rceil$.

Для кожної g -ої групи розрядів множників $X_{j(g+1)} X_{jg} X_{j(g-1)}$ формування часткового добутку здійснюється множенням W_j на відповідний коефіцієнт K_{jg} . Значення K_{jg} визначається як сума ваг ненульових цифр групи розрядів множника $X_{j(g+1)} X_{jg} X_{j(g-1)}$, де $X_{j(g+1)}$ має вагу мінус два, а X_{jg} та $X_{j(g-1)}$ – одиницю. Формування коефіцієнта K_{jg} здійснюється за таким виразом:

$$\begin{aligned} K_{jg} &= \sum_{j=1}^N W_j X_j = \sum_{j=g}^p 2^{-2g} \sum_{j=1}^N W_j K_{jg} = \\ &= \sum_{j=g}^p 2^{-2g} \sum_{j=1}^N P_{jg} = \sum_{j=g}^p 2^{-i} \sum_{j=1}^N M_{jg}, \quad g = \overline{1, G}, j = \overline{1, N}. \end{aligned} \quad (9)$$

Під час формування часткових добутків P_{jg} операцію множення на два доцільно реалізовувати зсувом на один розряд вліво, а зміну знаку – інверсією всіх розрядів, множеного з наступним додаванням одиниці до молодшого розряду. Замінивши у формулі (8) розряди $X_{j(g+1)} X_{jg} X_{j(g-1)}$ на коефіцієнт K_{jg} , отримаємо:

$$K_{jg} = \begin{cases} 2, & \text{якщо } X_{j(g+1)} = 0, X_{jg} = X_{j(g-1)} = 1; \\ 1, & \text{якщо } X_{j(g+1)} = 0, X_{jg} \neq X_{j(g-1)}; \\ 0, & \text{якщо } X_{j(g+1)} = X_{jg} = X_{j(g-1)}; \\ -1, & \text{якщо } X_{j(g+1)} = 1, X_{jg} \neq X_{j(g-1)}; \\ -2, & \text{якщо } X_{j(g+1)} = 1, X_{jg} = X_{j(g-1)} = 0. \end{cases} \quad (10)$$

З формули (10) бачимо, що обчислення скалярного добутку за модифікованим алгоритмом Бута зводиться до формування та підсумовування часткових добутків P_{jg} , причому кожний P_{jg} частковий добуток зсунутий відносно $P_{j(g-1)}$ на два розряди вправо. Обчислення скалярного добутку за модифікованим алгоритмом Бута забезпечує зменшення у 2 рази кількості тактів, необхідних для отримання результату скалярного добутку.

З формули (10) випливає, що для обчислення скалярного добутку необхідно виконати p тактів, у кожному з яких виконуються такі операції:

- формування для кожного g -ї групи розрядних зрізів $X_{j(g+1)}X_{jg}X_{j(g-1)}$ часткових добутків P_{jg} ;
- обчислення g -го макрочасткового добутку P_{Mg} підсумовуванням g -тих часткових добутків P_{jg} ;
- підсумовування макрочасткових добутків P_{Mg} відповідно до виразу $Y = 2^{-2}Y_{g-1} + P_{gi}$, де $Y_0 = 0$.

Паралельний вертикально-груповий метод обчислення скалярного добутку. Зменшити кількість тактів роботи, а відповідно тривалість обчислення можна шляхом використання паралельного вертикально-групового опрацювання даних групами із k розрядів. Для реалізації такого опрацювання вхідні дані X_j та вагові коефіцієнти W_j записуються так:

$$W_j = \sum_{h=1}^m 2^{-(h-1)k} \prod_{l=1}^g W_{jhl}, X_j = \sum_{h=1}^m 2^{-(h-1)k} \prod_{l=1}^g X_{jhl}, j = \overline{1, N}, \quad (11)$$

де $m = \lceil n/k \rceil$.

Обчислення скалярного добутку за паралельним вертикально-груповим методом виконується за формулою

$$Y = \sum_{j=1}^N W_j X_j = \sum_{h=1}^m 2^{-(h-1)k} \sum_{j=1}^N \sum_{l=1}^k 2^{-(k-l)} W_j X_{jhl} = \sum_{i=1}^n 2^{-(h-1)k} \sum_{j=1}^N P_{jh} = \sum_{h=1}^m 2^{-(h-1)k} P_{Mh}, \quad (12)$$

де: P_{jh} – jh -ий груповий частковий добуток, P_{Mh} – h -ий груповий макрочастковий добуток, який формують шляхом додаванням N групових часткових добутків P_{jh} ,

тобто $P_{Mh} = \sum_{j=1}^N P_{jh}$.

З формули (12) випливає, що обчислення скалярного добутку виконується за m тактів, у кожному h -му такті виконуються такі операції:

- формування h часткових добутків відповідно до виразу $P_{jhr} = W_j X_{jhr}$, де $r = \overline{1, k}$;
- обчислення групового часткового добутку P_{jh} згідно з формулою

$$P_{jh} = \sum_{r=1}^k 2^{-(r-1)} W_j X_{jhr}, h = \overline{1, m}; j = \overline{1, N};$$

- обчислення h -го макрочасткового добутку P_{Mh} підсумовуванням h -х часткових добутків P_{jh} згідно з формулою

$$P_{Mh} = \sum_{j=1}^N P_{jh};$$

- підсумовування макрочасткових добутків P_{Mh} відповідно до виразу $Y = 2^{-h}Y_{h-1} + P_{Mh}$, де $Y_0 = 0$.

Алгоритми групового підсумовування. При виконанні операції групового підсумовування $C = \sum_{j=1}^N Z_j$ данки Z_j є двійковими n -розрядними додатними числами, меншими за одиницю, які записуються так:

$$Z_j = \sum_{i=1}^n 2^{-i} Z_{ji}, j = \overline{1, N}. \quad (13)$$

Підставивши значення (13) у формулу для групового підсумовування, отримаємо:

$$C = \sum_{j=1}^N \sum_{i=1}^n 2^{-i} Z_{ji}. \quad (14)$$

Формула (14) відображає горизонтальну модель обчислення оператора групового підсумовування. Най-

швидшим варіантом реалізації горизонтальної моделі групового підсумовування є паралельно-паралельний метод обчислення, граф алгоритму якого подано на рис. 2.

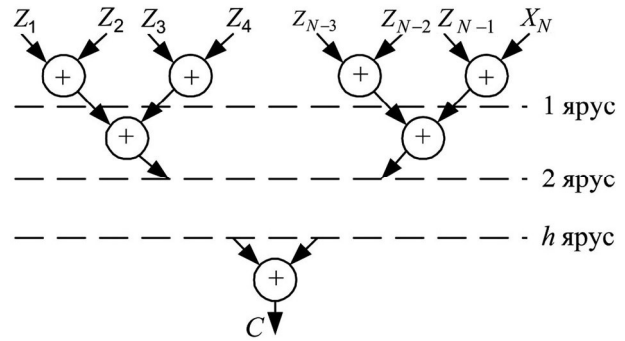


Рис. 2. Граф алгоритму паралельно-паралельного групового підсумовування / Graph of the algorithm for parallel-parallel group summation

Алгоритм паралельно-паралельного групового підсумовування є каскадним. Тривалість обчислення суми за таким алгоритмом залежить від висоти графа (кількості ярусів), яка обчислюється за формулою

$$h = \lceil \log_2 N \rceil. \quad (15)$$

У кожному ярусі операнди розбиваються на пари, для кожної з яких обчислюють суму. Загальна кількість операцій додавання за групового підсумовування N чисел дорівнює $N - 1$.

Базова структура нейронного елемента паралельно-потокowego типу. Для найповнішого використання переваг НВІС технології та забезпечення високої ефективності використання обладнання пропонуємо розробляти нейронний елемент паралельно-потокowego типу за такими принципами:

- використання базису елементарних арифметичних операцій;
- організація процесу обчислення скалярного добутку як виконання єдиної операції;
- однотипності сходнок конвеєра;
- локалізації та спрощення зв'язків між сходінками конвеєра;
- забезпечення балансу між введенням-виведенням і обчисленнями;
- використання конвеєризації та просторового розпаралелювання процесу обчислення;
- мінімізації зовнішнього інтерфейсу зв'язку.

Для апаратної реалізації моделі нейронного елемента паралельно-потокowego типу розроблено базову структуру (рис. 3). Основними компонентами нейронного елемента паралельно-потокowego типу є: перетворювач форматів даних, який формує N вхідних даних X і N вагових коефіцієнтів W для паралельно-потокowego обчислювача скалярного добутку, m сходнок конвеєра паралельно-потокowego обчислювача скалярного добутку та табличний обчислювач передавальної функції.

У нейронному елементі паралельно-потокowego типу вхідні дані X_{jq} та вагові коефіцієнти W_{jq} надходять на входи перетворювача форматів даних, на виходах якого отримуємо N вхідних n -розрядних даних X і N вагових n -розрядних коефіцієнтів W . З виходів перетворювача форматів даних N вхідних n -розрядних даних X і N вагових n -розрядних коефіцієнтів W записуються регістри БФГЧД першої сходнки конвеєра СК1 обчислювача скалярного добутку. Кількість сходнок конвеєра m залежить від таких параметрів: розрядності вхідних даних X , кількості розрядів k , які аналізуються для формування групових часткових добутків P_{jh} та часу форму-

вання вхідних даних і вагових коефіцієнтів на виходах перетворювача форматів даних. Управління процесом оброблення даних у нейронному елементі паралельно-

потокowego типу зводиться до видачі тактових імпульсів, які просувають інформацію із входу на вихід, запи-суючи проміжні результати в буферні регістри.

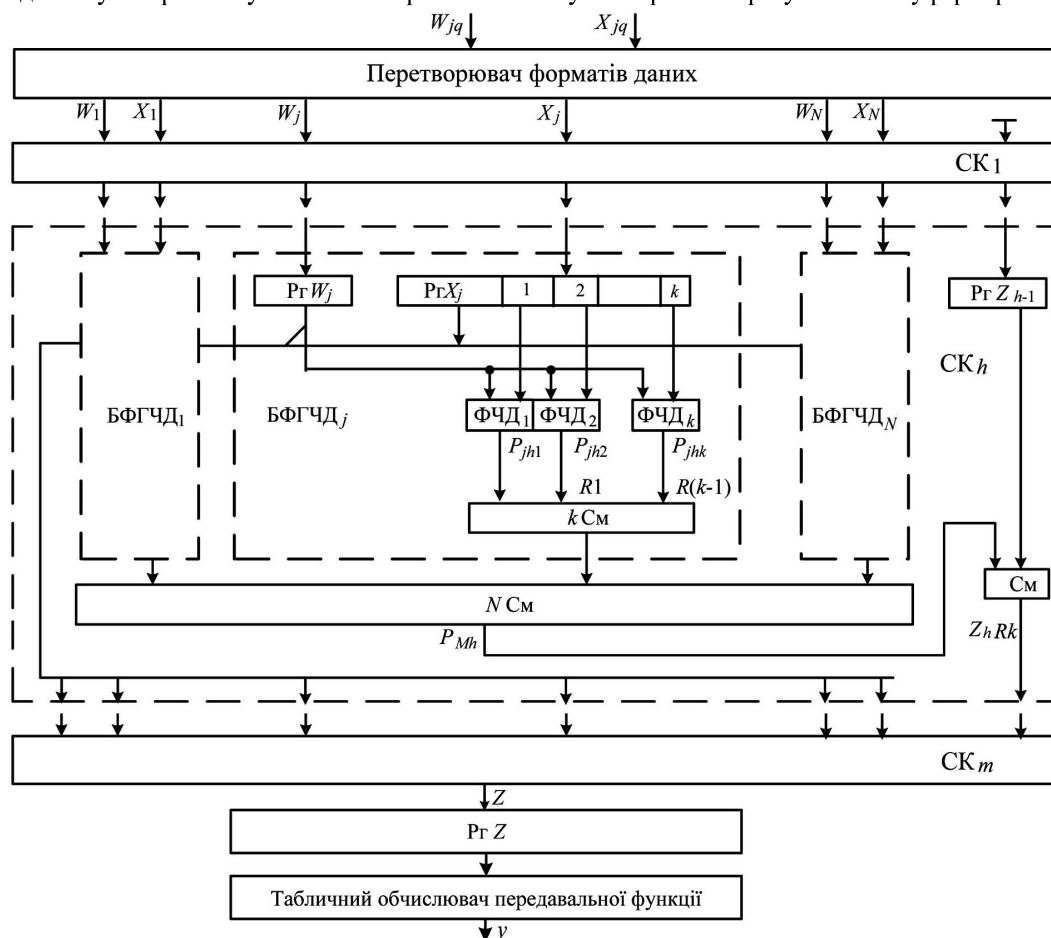


Рис. 3. Базова структура нейронного елемента паралельно-потокowego типу / Basic structure of a neural element of parallel-streaming type: СК – сходи́нка конве́єра / conveyor step; Пр – ре́гістр / register; ФЧД – формувач часткових добу́тків / partial product generator; БФГЧД – блок формування групових часткових добу́тків / group partial product generator; СМ – сумато́р / adder; NСМ – N-входо́вий сумато́р / N-input adder; W_{jq}, X_{jq} – входи нейронного елемента / inputs of the neural element; y – вихід нейронного елемента / output of the neural element

На виході $СК_m$ після m тактів у кожному наступному такті будемо отримувати скалярний добуток Z , який записується у реєстр PrZ . Скалярний добуток Z з виходів реєстра PrZ надходить на адресні входи табличного обчислювача передавальної функції, на виході якого отримуємо вихідний сигнал нейронного елемента y .

Метод синтезу нейронного елемента з паралельно-потокowym обробленням інтенсивних потоків даних у реальному часі. Задача синтезу нейронного елемента паралельно-потокowego типу зводиться до формування таких множин: вимог $R = \{R_j, j = \overline{1, k}\}$, характеристик $H = \{H_j, j = \overline{1, m}\}$, і обмежень $B = \{B_j, j = \overline{1, l}\}$, та знаходження такого вектора характеристик $H^* = \{H_j^*, j = \overline{1, m}\}$, $H_i^* = f_i(R, H, B)$, який забезпечить максимальне значення ефективності використання обладнання $E = \max f(R, H^*, B)$.

Множина вимог R складається з: R_1 – кількість каналів надходження даних m_d ; R_2 – розрядність каналів надходження даних n_d ; R_3 – частота надходження даних F_d ; R_4 – швидкодія елементної бази, яка визначається часом затримки вентиля t_β ; R_5 – кількість вхідних даних X_j і вагових коефіцієнтів W_j ; R_6 – розрядність операндів n .

До множини характеристик H входять: H_1 – загальна кількість зв'язків Z ; H_2 – просторова зв'язкова віддаль

Δj ; H_3 – конвеєрний такт t_k ; H_4 – витрати обладнання W ; H_5 – кількість видів функціональних вузлів S ; H_6 – кількість каналів введення; H_7 – розрядність каналів введення; H_8 – кількість виводів інтерфейсу зв'язку Y .

Обмеження B , які необхідно враховувати при синтезі нейронного елемента паралельно-потокowego типу, є такими: B_1 – точність обчислення, яка визначається розрядністю результату n_D ; B_2 – тривалість конвеєрного такту роботи повинна бути.

Для вибору варіанта нейронного елемента паралельно-потокowego типу використовується критерій ефективності використання обладнання E , який враховує кількість виводів інтерфейсу, однорідність структури, кількість і локальність зв'язків, зв'язує продуктивність з витратами обладнання та дає оцінку елементам (вентилям) компонента за продуктивністю [3]. Кількісний показник для оцінювання ефективності використання обладнання визначається за формулою

$$R = \frac{R_{HE}}{t_{CK} W_{HE}}, \quad (16)$$

де: R_{HE} – складність алгоритму обчислення, представлена кількістю елементарних арифметичних операцій; t_{CK} – тривалість спрацювання сходи́нки конве́єра; W_{HE} – витрати обладнання (у вентилях) на реалізацію нейронного елемента.

При виборі та розробленні методів і алгоритмів реалізації скалярного добутку та функції активації враховують вимоги R і характеристик H , але визначальним є забезпечення обмежень B .

Задача синтезу нейронного елемента паралельно-потокowego типу зводиться до опрацювання даних у реальному часі за мінімальних апаратних затрат на його реалізацію. Вихідними даними для синтезу нейронного елемента паралельно-потокowego типу є інтенсивність надходження вхідних даних, яка визначається так:

$$P_d = (g_x n_x + g_w n_w) F_d, \quad (17)$$

де: g_x – кількість каналів надходження вхідних даних X ; g_w – кількість каналів надходження вагових коефіцієнтів W ; n_x – розрядність каналів надходження вхідних даних X ; n_w – розрядність каналів надходження вагових коефіцієнтів W ; F_d – частота надходження вхідних даних X і вагових коефіцієнтів W . Кількість каналів g_x та g_w може бути від 1 до N , а розрядність n_x та n_w також може бути від 1 до n .

Для забезпечення роботи нейронного елемента паралельно-потокowego типу реального часу вимагається виконання такої умови:

$$t_d \geq T_k, \quad (18)$$

де: t_d – час надходження вхідних даних і вагових коефіцієнтів; T_k – конвеєрний такт роботи пристрою. Час надходження вхідних даних і вагових коефіцієнтів визначається так:

$$t_d = \frac{2Nn}{(g_x n_x + g_w n_w) F_d}, \quad (19)$$

де: N – кількість вхідних даних X і вагових коефіцієнтів W ; n – розрядність вхідних даних X і вагових коефіцієнтів W .

Конвеєрний такт T_k роботи нейронного елемента паралельно-потокowego типу визначається часом затримки даних у сходінці конвеєра. Цю тривалість визначають за формулою

$$T_k = t_{p_2} + t_{OB} = t_{p_2} + t_{фцд} + t_{БСм} + t_{См}, \quad (20)$$

де: t_{p_2} – тривалість затримки даних регістром; t_{OB} – тривалість спрацювання операційного блоку; $t_{фцд}$ – тривалість формування часткових добутків; $t_{БСм}$ – тривалість підсумовування часткових добутків; $t_{См}$ – тривалість додавання двох чисел.

Тривалість конвеєрного такту роботи t_k , в основному, залежить від:

- швидкодії елементної бази;
- алгоритму та засобів формування часткових добутків;
- кількості часткових добутків, алгоритму та засобів їх підсумовування;
- алгоритму та засобів реалізації підсумовування макрочасткових добутків.

Кількість сходінок конвеєра визначається складністю операцій, які реалізуються даними сходінками. Зменшення складності операцій, які реалізуються сходінкою конвеєра, призводить до збільшення їх кількості та зменшення тривалості конвеєрного такту. При зменшенні кількості сходінок збільшується складність операцій, які виконуються у сходінці, а відповідно, збільшується тривалість конвеєрного такту. Апаратні витрати на реалізацію нейронного елемента паралельно-потокowego типу зростають із збільшенням кількості сходінок конвеєра.

Синтез нейронного елемента паралельно-потокowego типу з високою ефективністю використання обладнан-

ня досягається узгодженням часу надходження вхідних даних і вагових коефіцієнтів t_d з конвеєрним тактом T_k роботи пристрою. Таке узгодження може досягатися шляхом збільшення або зменшення T_k . Основними шляхами зменшення T_k є:

- використання алгоритмів, які зменшують кількість часткових добутків;
- конвеєризація багатовходового суматора шляхом розбиття його на сходінки;
- паралельне включення двох і більше пристроїв обчислення скалярного добутку, кількість яких визначається, в основному, часом надходження вхідних даних і вагових коефіцієнтів.

Основними шляхами збільшення T_k є:

- використання алгоритмів, які зменшують кількість сходінок конвеєра (алгоритм Бута, алгоритм з формуванням групових часткових добутків);
- реалізація однією сходінкою конвеєра двох і більше ітерацій алгоритму.

Синтез нейронного елемента паралельно-потокowego типу зі заданим часом t_d вимагає виконання таких етапів:

- 1) оцінювання тривалості такту T_k конвеєра, затрат обладнання та ефективності використання обладнання для структур з різною кількістю розрядів k , які використовують для формування групових часткових добутків P_{jh} ;
- 2) вибір алгоритму та кількості розрядів k для формування групових часткових добутків P_{jh} , які забезпечать синтез нейронного елемента паралельно-потокowego типу з високою ефективністю використання обладнання;
- 3) узгодження конвеєрного такту T_k роботи нейронного елемента паралельно-потокowego типу з часом t_d надходження вхідних даних і вагових коефіцієнтів;
- 4) оцінювання затрат обладнання та ефективності використання обладнання для нейронного елемента паралельно-потокowego типу з узгодженим конвеєрним тактом T_k роботи.

Обговорення результатів дослідження. У роботах [4, 20, 33] показано, що для нейромережевого опрацювання інтенсивних потоків даних у реальному часі використовуються як програмна реалізація з використанням технології CUDA і графічних процесорів, так і апаратна реалізація компонентів нейронних мереж на FPGA. Недоліком таких засобів [16, 21] є складність узгодження інтенсивності надходження даних з інтенсивністю опрацювання даних та значні апаратні витрати на їх реалізацію. Перевагою розробленої узагальненої моделі та базової структури нейронного елемента паралельно-потокowego типу є орієнтація на узгоджене паралельно-потокове оброблення інтенсивних потоків даних реального часу, яка забезпечує підвищення ефективності використання обладнання.

Подальшими напрямками досліджень є розроблення засобів автоматизації процесу синтезу нейронного елемента паралельно-потокowego зі заданими технічними параметрами.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження:

- розроблено узагальнену модель нейронного елемента паралельно-потокowego типу, основними компонентами якої є перетворювач форматів даних, пристрій обчислення скалярного добутку та блок реалізації функції активації, яка за рахунок узгодження часу надходження вхід-

них даних і вагових коефіцієнтів з тривалістю конвеєрного такту, забезпечує вибір параметрів нейронного елемента для опрацювання потоків даних у реальному часі з високою ефективністю використання обладнання;

- розроблено метод синтезу нейронного елемента паралельно-потокowego типу для оброблення потоків даних у реальному часі, який за рахунок вибору кількості розрядів для формування групових часткових добутків і тривалості такту роботи конвеєра забезпечує синтез нейронного елемента з мінімальними апаратними затратами.

Практична значущість результатів дослідження – використання розробленої моделі та методу синтезу нейронного елемента паралельно-потокowego типу забезпечує синтез нейронного елемента для оброблення потоків даних у реальному часі з мінімальними апаратними затратами.

Висновки / Conclusions

Унаслідок проведеного дослідження розроблено орієнтовану на опрацювання інтенсивних потоків даних у реальному часі узагальнену модель нейронного елемента паралельно-потокowego типу, алгоритми обчислення скалярного добутку, метод синтезу та базову структуру нейронного елемента. За результатами дослідження можна зробити такі основні висновки:

1. Розроблено узагальнену модель нейронного елемента паралельно-потокowego типу, основними компонентами якої є перетворювач форматів даних, пристрій обчислення скалярного добутку та блок реалізації функції активації, яка, за рахунок вибору кількості розрядів для формування групових часткових добутків і узгодження частоти введення даних з частотою роботи конвеєра, забезпечує високу ефективність використання обладнання.
2. Проаналізовано алгоритми для обчислення скалярного добутку та вибрано паралельний вертикально-груповий метод обчислення скалярного добутку з аналізом k розрядів для формування групових часткових добутків.
3. Вибрано для групового підсумовування каскадний паралельно-паралельний алгоритм, який забезпечує конвеєризацию процесу підсумовування введенням буферних регістрів між ярусами суматорів.
4. Запропоновано здійснювати розроблення нейронного елемента паралельно-потокowego типу на підставі таких принципів: використання базису елементарних арифметичних операцій; організація процесу обчислення скалярного добутку як виконання єдиної операції; локалізація та спрощення зв'язків між сходами конвеєра; забезпечення балансу між введенням-виведенням і обчисленнями; використання конвеєризації та просторового розпаралелювання процесу обчислення; мінімізація зовнішнього інтерфейсу зв'язку.
5. Запропоновано виконувати оцінювання апаратної реалізації нейронного елемента за критерієм ефективності використання обладнання, який враховує складність, час і витрати обладнання на реалізацію алгоритму та дає оцінку внеску кожного вентиля в загальну продуктивність нейронного елемента.
6. Показано, що синтез нейроелемента паралельно-потокowego типу зводиться до мінімізації апаратних затрат при забезпеченні обчислення скалярного добутку та функції активації в реальному часі.
7. Розроблено метод синтезу нейронного елемента паралельно-потокowego типу, який, за рахунок розбиття його на сходи конвеєра з узгодження часу надходження даних з тактом конвеєра, забезпечує синтез нейронного елемента з високою ефективністю використання обладнання та обробленням потоків даних у реальному часі.

References

1. Abdoli, S., Cardinal, P., & Koerich, A. L. (2019). End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems With Applications*, 136, 252–263 p.
2. Allam, Z. (2019). Achieving neuroplasticity in artificial neural networks through smart cities. *Smart Cities* 2, 118–134.
3. Anumula, J., Neil, D., Delbruck, T., & Liu, S.-C. (2018). Feature Representations for Neuromorphic Audio Spike Streams. *Frontiers in Neuroscience*, 12(23), 1–12. <https://doi.org/10.3389/fnins.2018.00023>
4. Araujo, G., Griebler, D., Rockenbach, D.A., Danelutto, M., Fernandes, L.G. (2023). NAS Parallel Benchmarks with CUDA and beyond. Software – Practice and Experience this link is disabled, 53(1), pp. 53–80.
5. Ashraf, I., Zikriya, Y. B., Hur, S., et al. (2021). Localizing pedestrians in indoor environments using magnetic field data with term frequency paradigm and deep neural networks. *Int International Journal of Machine Learning and Cybernetics* 12, 3203–3219. <https://doi.org/10.1007/s13042-021-01279-8>
6. Bohrer, J. S., Grisci, B. I., & Dorn, M. (2020). Neuroevolution of Neural Network Architectures Using CoDeepNEAT and Keras. *Preprint arXiv.org*, 29. URL: <https://arxiv.org/abs/2002.04634>
7. Bouvier, M., Valentian, A., Mesquida, T., Rummens, F., Reyboz, M., Vianello, E., & Beigne, E. (2019). Spiking Neural Networks Hardware Implementations and Challenges: A Survey. *ACM Journal on Emerging Technologies in Computing Systems*, 15(2), 22 p.
8. Cecchi, S., Spinsante, S., Terenzi, A., & Orcioni, A. A. (2020). Smart Sensor-Based Measurement System for Advanced Bee Hive Monitoring. *Sensors*, 20(9), <https://doi.org/10.3390/s20092726>
9. Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2017). Eyeriss, an energy-efficient reconfigurable accelerator for deep convolutional neural networks. *Journal of Solid-State Circuits*, 52(1), 127–138.
10. Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2019). Eyeriss v2: a flexible accelerator for emerging deep neural networks on mobile devices. *J Emerg Sel Top Circuits Syst*, 9(2), 292–308.
11. Duka, A. V. (2014). Neural network based inverse kinematics solution for trajectory tracking of ai robotic arm. *Procedia Technol* 12, 20–27.
12. Frank, Seide & Amit, Agarwal. (2016). Cntk: Microsofts open-source deep-learning toolkit. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 16, New York, NY, USA. ACM, 2135–2135.
13. Gadekallu, T. R., Khare, N., Bhattacharya, S., et al. (2020). Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*. <https://doi.org/10.1007/s12652-020-01963-7>
14. Gadekallu, T. R., Rajput, D. S., Reddy, M. P. K., et al. (2020). A novel PCA – whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *J Real-Time Image Proc*. <https://doi.org/10.1007/s11554-020-00987-8>
15. Geche, F., Mulesa, O., Buchok, V. (2017). Verification of realizability of boolean functions by a neural element with a threshold activation function. *Eastern-Eur J Enterp Technol*, 1(4), 30–40.
16. Koehn, T., & Athanas, P. (2019). Data staging for efficient high throughput stream processing. Parallel Computing this link is disabled, 90, 102566.
17. Krestinskaya, O., James, A. P., & Chua, L. O. (2020). Neuromemristive Circuits for Edge Computing: A Review. *Transactions on Neural Networks and Learning Systems*, 31, 4–23.
18. Kryvokhata, A. G., Kudin, O. V., & Choporov, S. V. (2020). Neural network mathematical models in problems of sound signal processing: monograph. Kherson: "Helvetyka" Publishing House, 120 p.
19. Lotric, U., Bulic, P. (2012). Applicability of approximate multipliers in hardware neural networks. *Neurocomputing* 96, 57–65.
20. Mencagli, G., Griebler, D., Danelutto, M. (2022). Towards Parallel Data Stream Processing on System-on-Chip CPU+GPU Devices. Proceedings-30th Euromicro International Conference on Pa-

- rallel, Distributed and Network-Based Processing, PDP 2022, pp. 34–38.
21. Mencagli, G., Torquati, M., Cardaci, A., Rinaldi, L., & Danelutto, M. (2021). WindFlow: High-Speed Continuous Stream Processing with Parallel Building Blocks. *IEEE Transactions on Parallel and Distributed Systems* this link is disabled, 32(11), pp. 2748–2763, 9408386.
 22. Nurvitadhi, E., Venkatesh, G., Sim, J., Marr, D., et al. (2017). Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks. In: *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, Monterey, CA, USA, ACM: NY, USA, 5–14.
 23. Passalis, N., & Tefas, A. (2020). Continuous drone control using deep reinforcement learning for frontal view person shooting. *Neural Computing and Applications*, 32, 4227–4238.
 24. Qinghui, Hong, Ya, Li, & Xiaoping, Wang. (2020). Memristive continuous Hopfield neural network circuit for image restoration. *Neural Computing and Applications*, 32, 8175–8185.
 25. Sugianto, I., Axenie, C., & Conradt, J. (2019). FPGA-based hardware accelerator for an embedded factor graph with configurable optimization. *Journal of Circuits, Systems and Computers*, 28(02).
 26. Tkachenko, R. O., Tkachenko, P. R., & Izonin, I. V. (2017). *Neural network tools of artificial intelligence: a study guide*. Lviv: Lviv Polytechnic Publishing House, 207 p.
 27. Tkachenko, R., Izonin, I. (2019). Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. In: Hu ZB, Petoukhov S (eds) *Advances in Computer Science for Engineering and Education, ICCSEE2018, Advances in Intelligent Systems and Computing*. Springer, Cham, 754, 578–587.
 28. Torbati, N., Ayatollahi, A., & Kermani, A. (2014). An efficient neural network based method for medical image segmentation. *Comput Biol Med* 44, 76–87.
 29. Trimberger, S. M. (2015). Three ages of FPGAs: a retrospective on the first thirty years of FPGA technology. *Proc* 103, 318–331.
 30. Tsmots, I. G., Tkachenko, R. O., Teslyuk, V. M., Riznyk, O. Ya., & Kazymira, I. Ya. (2022). *Smart systems: technologies, architectures, data processing, protection and coding: monograph*. Lviv: Publishing House "Lviv Polytechnic", 220 p.
 31. Tsmots, I., Teslyuk, V., Kryvinska, N., Skorokhoda, O., Kazymyra, I. (2023). Development of a generalized model for parallel-streaming neural element and structures for scalar product calculation devices. *Journal of Supercomputing*, 79, 4820–4846. <https://doi.org/10.1007/s11227-022-04838-0>
 32. Tsmots, I., Teslyuk, V., Teslyuk, T., Ihnatyev, I. (2018). Basic components of neuronetworks with parallel vertical group data real-time processing. In: *Advances in Intelligent Systems and Computing II, Advances in Intelligent Systems and Computing* 689. Springer International Publishing AG 2018, 558–576.
 33. Vogel, A., Griebler, D., Danelutto, M., & Fernandes, L. G. (2022). Self-adaptation on parallel stream processing: A systematic review. *Concurrency and Computation: Practice and Experience* this link is disabled, 34(6), 6759.
 34. Wu, R., Guo, X., Du, J., & Li, J. (2021). Accelerating neural network inference on FPGA-based platforms – A survey. *Electronics* 10. <https://doi.org/10.3390/electronics10091025>
 35. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, & Trevor Darrell. (2014). Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv: 1408.5093.
 36. Yarovy, A., Ilchenko, R., Arseniuk, I., Shmet, Y., Kotyra, A., & Smailova, S. (2018). An intelligent system of neural networking recognition of multicolor spot images of laser beam profile. In: *Proceedings of SPIE 10808, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2018, 108081B*. <https://doi.org/10.1117/12.2501691>
 37. Zeng, W., Guo, Z., Shen, Y., et al. (2021). Data-driven management for fuzzy sewage treatment processes using hybrid neural computing. *Neural Comput Appl*. <https://doi.org/10.1007/s00521-020-05655-3>

I. H. Tsmots, B. V. Shtohrinets, I. Ya. Kazymyra, A. A. Lytvyn

Lviv Polytechnic National University, Lviv, Ukraine

MODEL AND METHOD FOR SYNTHESIS OF NEURAL ELEMENT OF PARALLEL-STREAMING TYPE

A generalized model of a parallel-streaming type neural element has been developed. The main components of this model are as follows: a data format converter, a device for calculating scalar product, and a block for implementing the activation function, which ensures high efficiency use of equipment, due to the selection of the number of digits for the formation of group partial products and the coordination of the frequency of data input with the frequency of the conveyor. The algorithms for calculating the scalar product were analyzed, and a parallel vertical group method for calculating the scalar product with the analysis of k digits for the formation of group partial products was selected as well. A cascaded parallel-parallel algorithm was chosen for group summation, which ensures the pipeline of the summation process by introducing buffer registers between the levels of adders. It is proposed to develop the neural element of the parallel-streaming type on the basis of the following principles: the use of the basis of elementary arithmetic operations; the organization of the process of calculating the scalar product as a single operation; the localization and simplification of connections between conveyor steps; the balance ensuring between input-output and calculations; the use of pipelining and spatial parallelization of the calculation process; the minimization of the external communication interface. It is suggested to evaluate the hardware implementation of the neural element according to the criterion of the efficiency of the equipment use, which takes into account the complexity, time and costs of the equipment for the implementation of the algorithm and gives an estimate of the contribution of each gate to the overall performance of the neural element. It is shown that the synthesis of the neural element of the parallel-flow type is reduced to the minimization of hardware costs, provided the calculation of the scalar product and the activation function in real time. A method for the synthesis of a parallel-flow type neural element has been developed, which provides the synthesis of a neural element with high efficiency of equipment use and processing of data flows in real time by dividing it into steps and matching the time of data arrival with the conveyor clock.

Keywords: scalar product; group summation; conveyor step; method; structure.