



І. Г. Цмоць, Ю. В. Опотяк, Б. В. Штогрінець, Т. Б. Мамчур, В. М. Голубець

Національний університет "Львівська політехніка", Львів, Україна

МОДЕЛЬ, СТРУКТУРА ТА МЕТОД СИНТЕЗУ НЕЙРОННОГО ЕЛЕМЕНТА МАТРИЧНОГО ТИПУ

Зростаюче застосування нейромереж для опрацювання даних потребує пошуку нових і вдосконалення наявних засобів для їх розроблення. Застосування таких засобів на мобільних платформах вимагає врахування додаткових обмежень щодо габаритів, маси та енергоспоживання. З'ясовано, що перспективним способом вирішення цієї проблеми є застосування надвеликих інтегральних схем (НВІС), що, однак, потребує розроблення нових моделей, структур і методів синтезу нейронних елементів (НЕ) як основи для побудови нейронних та нейроподібних мереж. У роботі запропоновано й описано узагальнену модель, структуру та метод синтезу НЕ матричного типу. Описано алгоритми обчислення скалярного добутку, групового підсумовування часткових добутків, які використано для реалізації базової структури НЕ. Виконано оцінювання параметрів НЕ матричного типу з формуванням часткових добутків за модифікованим алгоритмом Бута. Описано розроблену узагальнену модель НЕ матричного типу з формуванням часткових добутків за модифікованим алгоритмом Бута, компонентами якої є перетворювач форматів даних, пристрій обчислення скалярного добутку та блок реалізації функції активації. Розроблена модель НЕ за рахунок узгодження тривалості надходження вхідних даних і вагових коефіцієнтів з тактом роботи конвеєра забезпечує опрацювання даних у реальному часі. Для зменшення у два рази кількості обчислень часткових добутків запропоновано використати матричний метод обчислення скалярного добутку з формуванням часткових добутків за допомогою модифікованого алгоритму Бута. Описано розроблені алгоритми та структури для групового підсумовування часткових добутків, які за рахунок розпаралелення процесу підсумовування часткових добутків на підставі дерев Уоллеса та каскадного паралельно-паралельного алгоритму забезпечують зменшення часу підсумовування. Описано розроблену базову структуру НЕ матричного типу з формуванням часткових добутків з використанням модифікованого алгоритму Бута, яка за рахунок паралельно-паралельного формування та підсумовування часткових добутків на підставі дерев Уоллеса і каскадного алгоритму з використанням конвеєрних регістрів для узгодження інтенсивності надходження даних з інтенсивністю обчислення, забезпечує оброблення потоків даних у реальному часі та реалізацію на НВІС. Оцінено витрати обладнання для реалізації НЕ матричного типу та тривалості конвеєрного такту для визначених конфігурацій обчислювальних засобів, а саме, зрядності операндів та кількості пар добутків.

Ключові слова: нейронна мережа; конвеєрний метод обчислень; алгоритм Бута; конвеєрний такт; витрати обладнання; опрацювання даних у реальному часі.

Вступ / Introduction

Нейромережеві технології дедалі ширше застосовують у різних галузях науки і техніки та потребують опрацювання інтенсивних потоків даних у реальному часі. Окрім цього, застосування нейромережевих засобів опрацювання даних вимагає врахування додаткових обмежень стосовно габаритів, маси та енергоспоживання [4, 10, 27]. Апаратна реалізація зазначених нейромережевих засобів дає змогу отримати високі показники продуктивності, проте вимагає розроблення нових моделей та базових структур нейронних елементів (НЕ), вдоско-

налення методів та алгоритмів їх реалізації, організації обчислювального процесу тощо. Реалізація конвеєризації обчислень є одним з методів забезпечення обчислень у реальному часі, однак вимагає розроблення відповідних обчислювальних структур, зокрема і нейронного елемента (НЕ) для побудови високопродуктивних нейронних і нейроподібних мереж.

Розроблення вказаних НЕ можливе з використанням сучасної елементної бази на базі НВІС (надвелика інтегральна схема VLSI (англ. *Very Large-Scale Integration*)), однак, водночас, вимагає розроблення нових методів, удосконалення відповідних алгоритмів та НВІС-

Інформація про авторів:

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

Email: ivan.h.tsmots@lpnu.ua; <https://orcid.org/0000-0002-4033-8618>

Опотяк Юрій Володимирович, канд. техн. наук, доцент, кафедра автоматизованих систем управління.

Email: yurii.v.opotyak@lpnu.ua; <https://orcid.org/0000-0001-9889-4177>

Штогрінець Богдан Володимирович, магістр, аспірант, кафедра автоматизованих систем управління.

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

Мамчур Тарас Борисович, магістр, аспірант, кафедра автоматизованих систем управління.

Email: taras.b.mamchur@lpnu.ua; <https://orcid.org/0009-0006-0593-7937>

Голубець Володимир Мирославович, магістр, аспірант, кафедра автоматизованих систем управління.

Email: volodymyr.m.holubets@lpnu.ua; <https://orcid.org/0009-0003-6856-147X>

Цитування за ДСТУ: Цмоць І. Г., Опотяк Ю. В., Штогрінець Б. В., Мамчур Т. Б., Голубець В. М. Модель, структура та метод синтезу нейронного елемента матричного типу. Науковий вісник НЛТУ України. 2024, т. 34, № 4. С. 68–77.

Citation APA: Tsmots, I. G., Opotyak, Yu. V., Shtohrinets, B. V., Mamchur, T. B., & Holubets, V. M. (2024). Model, structure and synthesis method of matrix-type neural element. *Scientific Bulletin of UNFU*, 34(4), 68–77. <https://doi.org/10.36930/40340409>

структур для реалізації вказаних НЕ з метою оброблення даних у режимі реального часу.

Для забезпечення опрацювання у реальному часі потоків даних за допомогою нейромережових засобів потрібно вважати перспективним застосування конвеєрних нерекурсивних НЕ, що містять вузли обчислення скалярного добутку та вузли групового підсумовування результатів. Особливістю таких НЕ є відсутність обернених зв'язків і отримання результату неітераційних обчислень при виконанні розрахунків. Для реалізації конвеєрних обчислень необхідний розподіл елементів обчислювального пристрою на окремі сходинки, які містять буферну пам'ять та операційний блок. Буферна пам'ять на кожній сходинці забезпечує узгодження інтенсивності надходження даних з часом їх оброблення в операційному блоці.

Для забезпечення високої продуктивності НЕ і мережі на їх основі загалом, усі сходинки конвеєра повинні виконувати однакові за складністю операції при забезпеченні узгодження тривалості введення вхідних даних та вагових коефіцієнтів, такту роботи конвеєра та тривалості виведення результатів обчислень. Обрана розрядність вхідних даних, кількість розрядів операційного блоку, що задіяні у формуванні макрочасткових добутоків, тривалість надходження вхідних даних і вагових коефіцієнтів визначають необхідну кількість сходинок конвеєра реалізованої на НЕ мережі.

Для ефективної реалізації паралельно-потокowego оброблення даних на НВІС алгоритми реалізації НЕ мають бути добре структурованими. Ефективність відображення алгоритму на апаратуру безпосередньо зв'язана зі способом декомпозиції алгоритму роботи НЕ на незалежні операції для паралельного їх виконання, або на залежні при конвеєрному режимі виконання. Паралельно-потокова реалізація НЕ і відповідних мереж здатна забезпечити їх апаратне виконання на НВІС й обчислення у реальному часі.

Тому актуальним завданням є розроблення узагальненої моделі та базової структури НЕ матричного типу для оброблення даних у режимі реального часу.

Об'єкт дослідження – обчислення скалярного добутку, групового підсумовування часткових добутоків у режимі реального часу.

Предмет дослідження – моделі, методи, алгоритми та структури для реалізації НЕ матричного типу, що дасть змогу узгодити тривалість надходження даних з конвеєрним тактом обчислення для забезпечення обчислень у режимі реального часу.

Мета роботи – розробити узагальнену модель НЕ матричного типу, базової структури та методу його синтезу, для забезпечення реалізації нейромереж реального часу на НВІС.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- розробити модель НЕ матричного типу, що дасть змогу отримати апаратну реалізацію НЕ для оброблення інтенсивних потоків даних у режимі реального часу;
- вибрати для обчислення скалярного добутку матричний метод, що дасть можливість паралельного формування та підсумовування часткових добутоків з використанням дерева Уоллеса та каскадного методу;
- розробити алгоритми та відповідні структури вузлів для групового підсумовування часткових добутоків, що забезпечить зменшення такту конвеєра способом просторового та часового розпаралелення;

- розробити базову структуру НЕ матричного типу, у якому обчислення часткових добутоків виконують з використанням модифікованого алгоритму Бута;
- удосконалити метод синтезу НЕ матричного типу для забезпечення оброблення інтенсивних потоків даних у режимі реального часу, що забезпечить апаратну реалізацію на НВІС з мінімальними витратами обладнання;
- оцінити параметри розробленого НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення обчислення часткових добутоків.

Аналіз останніх досліджень та публікацій. Виконаний аналіз засобів для розроблення нейромереж показує [16, 19, 21, 26], що у переважній більшості вони реалізують програмно. Зокрема, у роботі [16] наведено порівняльне дослідження програмних реалізацій нейронної мережі радіального базису (точна відповідність), радіальної нейронної мережі "k-середніх" і нейронної мережі прямого зв'язку із зворотним розповсюдженням помилок Левенберга-Марквардта, що розроблені для прогнозування вихідної потужності сонячних панелей. У роботі [21] описано тришарову нейронну мережу для розпізнавання мультиспектральних зображень за допомогою Python і бібліотеки TensorFlow, яка також реалізована виключно програмними засобами. Проте основними недоліками програмної реалізації є невисока продуктивність, що не забезпечує опрацювання інтенсивних потоків даних. Підвищення швидкості нейромережових засобів можна досягнути способом їх апаратної реалізації на НВІС.

Основними компонентами для синтезу апаратних реалізацій нейромереж є нейронні елементи. Існує низка моделей формального нейрона, що відрізняються способами представлення вхідних сигналів, надходження даних і опису вагових коефіцієнтів, способами обчислення функції активації тощо.

У роботі [13] описано схему нейрона на основі модифікованої математичної моделі Іжикевича, яка покликана спростити її цифрову апаратну реалізацію. Однак сьогодні ще не розроблено ефективних правил навчання таких мереж. Автори у роботі [20] пропонують розроблення імпульсного нейрона Морріса-Лекара в оптимізованій аналоговій реалізації, що спрямована на розроблення ефективних рішень для побудови аналогових імпульсних нейронних мереж низької складності. У роботі [2] описано аналогову схему без використання вузла помножувача для реалізації нейронної моделі Гіндмарша-Роуза без множників. Проте, аналогова реалізація нейромереж потребує додаткових ресурсів для узгодження з цифровими елементами сучасних пристроїв.

У роботі [9] досліджено моделі біологічних нейронів для розроблення ефективних біоімітаційних процесорів. У цій статті модель нейрона Вільсона реалізована як наближення біологічної моделі Ходжкіна-Хакслі, яка налаштована для ефективної цифрової реалізації на НВІС. Нову реалізацію адаптивно-експоненціальної (AdEx) нейронної моделі на основі методу повної відповідності під назвою PBAM (англ. *Power-2 Based AdEx Model*) описано у роботі [7]. Ця модель може точно відтворювати різні спайкові поведінки, подібні до біологічних нейронів. Запропонована модель фізично реалізована на FPGA як доказ концепції, а експериментальні результати демонструють високу схожість з оригінальною моделлю. Проте зазначені реалізації орієнтовані скоріше на імітацію біологічної поведінки нейронів для дослідження їх функціонування, а не для виконання саме обчислень.

Значна кількість досліджень пов'язана з апаратною реалізацією НЕ на підставі НВІС. У роботах [3, 11, 12, 14, 25] розглянуто варіанти реалізацій НЕ на базі FPGA, що є наразі одним з найперспективніших напрямків такої реалізації. Зокрема, у дослідженні [14] розглянуто використання програмованих вентильних матриць FPGA (англ. *Field-Programmable Gate Array*) для схематехнічного дизайну нейронної моделі на основі біологічно вірогідної квадратичної спайкової моделі нейрона, що може моделювати поведінку нейронних спайків нейронів таламуса та пірамідальних нейронів гіпокампа CA1. У роботі [11] представлено модифіковану модель нейрона Міхаласа-Нібура, придатну для цифрової реалізації на FPGA компактного нейрона Міхаласа-Нібура без множника, який використовує оптимізацію довжини слова, та оператори порозрядного зсуву для множення. У роботі [25] представлено модифіковану модель нейрона Іжикевича, яка замінює складні операції множення та ділення на прості двійкові операції зсуву та яка реалізована на FPGA. Для спрощення реалізації моделі нейрона у роботі [3] представлено кусково-лінійну (NC-PWL) схему підгонки для реалізації нейронної моделі Гіндмарша-Роуза (HR) без множників. За допомогою методу логічного зсуву розроблено цифрову схему без помножувача з низьким споживанням ресурсів для цієї відповідної моделі на базі FPGA. У роботі [12] представлено цифрову реалізацію нейронної моделі Морріса-Лекара, що використовує алгоритм CORDIC (англ. *COordinate Rotation Digital Computer*) для створення реалізації з фіксованою точкою, а її конструкція перевірена на FPGA.

У публікаціях розглядають питання ефективної реалізації структурних елементів НЕ, орієнтованих на НВІС. У роботі [8] запропоновано нейронний пришвидшувач як структурний елемент нейромережі, що навчається за допомогою інфраструктури TensorFlow, та реалізований на основі FPGA. У роботі [23] запропоновано вдосконалення структури НЕ способом використання алгоритму обчислення скалярного добутку операндів у форматі з рухомою комою таблично-алгоритмічним методом та алгоритму формування таблиць макрочасткових добутків, та описано реалізацію на НВІС. У роботі [6] досліджено схеми реалізації НЕ з вертикальним обробленням даних на основі порозрядного конвеєрного оброблення обчислень з погляду успішності їх реалізації у проектах FPGA та виконано порівняльний аналіз отриманих схем з бібліотечними рішеннями за показниками їх складності, пропускну здатності та енергоємності. У роботі [28] досліджено реалізацію в НЕ структури для експоненціального обчислення функції активації та запропоновано модуль експоненціального обчислення для числа з фіксованою комою на основі теорії Fast InvSqrt для реалізації на FPGA.

Результати аналізу показують, що характеристики розглянутих реалізацій НЕ, зокрема, визначаються принципом виконання операції для обчислення скалярного добутку і її реалізації апаратно [18, 24]. Використання для реалізації зазначених операцій базису елементарних арифметичних операцій та багатоперандного підходу, забезпечує реалізацію оптимальної структури отриманих засобів за показниками швидкодії та необхідними апаратними витратами. Основою вказаних алгоритмів обчислення скалярного добутку є виконання

операцій формування часткових добутків та операцій їх додавання [17].

Отже, з аналізу публікацій випливає актуальність розроблення удосконалених моделей нейронних елементів матричного типу та відповідних алгоритмів обчислення скалярного добутку та групового підсумовування часткових добутків, необхідних для такої реалізації.

Результати дослідження та їх обговорення / Research results and their discussion

Модель нейронного елемента матричного типу.

Основними компонентами для синтезу нейронних мереж є нейронні елементи (НЕ), які функціонують згідно з формулою

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

де: Y – обчислені вихідні дані з НЕ; X_j – вхідне j -те значення даних НЕ; W_j – вагове значення j -го коефіцієнта НЕ; N – кількість входів даних НЕ і відповідних вагових коефіцієнтів, $j = \overline{1, N}$; $f()$ – функція активації НЕ. З аналізу формули (1) робимо висновок, що для реалізації НЕ необхідне виконання математичних операцій обчислення скалярного добутку $Z = \sum_{j=1}^N W_j X_j$ та відповідної

функції активації $f(Z)$. У цьому випадку синтез нейронної мережі для роботи у режимі реального часу з можливістю опрацювання інтенсивних потоків даних передбачає розроблення НЕ матричного типу розпаралеленням процесів обчислення у просторі і часі.

Аналітичну модель НЕ матричного типу з використанням для формування часткових добутків модифікованого алгоритму Бута можна записати так:

$$Y = f_a\left(f_z(f_{p_j}(f_{(W_j \rightarrow NW, X_j \rightarrow NX)}))\right), \quad (2)$$

де: Y – вихід результату обчислень НЕ; $f_a(Z)$ – функція активації НЕ; $f_z()$ – функція, що реалізує групове підсумовування часткових добутків P_{jg} ; $f_{p_j}()$ – функція для забезпечення формування часткових добутків P_{jg} , $f_{(W_j \rightarrow NW, X_j \rightarrow NX)}$ – функція послідовно-паралельного перетворення форматів представлення для j -го вагового коефіцієнта W_j і вхідного j -го значення даних X_j НЕ.

Граф моделі НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення формування часткових добутків наведено на рис. 1.

Основними компонентами НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення формування часткових добутків є: перетворювач форматів представлення даних послідовно-паралельного типу $f_{(W_j \rightarrow NW, X_j \rightarrow NX)}$, матриця $(N \times r)$ блоків обчислення часткових добутків P_{jg} , блок групового підсумовування матриці $(N \times r)$ часткових добутків P_{jg} і блок обчислення функції активації $f_a(Z)$.

Особливістю пропонованої моделі НЕ матричного типу є забезпечення суміщення у часі виконання операцій вводу даних наступного масиву із N вагових коефіцієнтів W_j НЕ і N вхідних даних X_j НЕ з виконанням операцій обчислення для попереднього масиву даних. Для забезпечення реалізації процесу паралельно-потокowego оброблення даних НЕ у режимі реального часу необхідно виконати умову функціонування НЕ за конвеєрним принципом оброблення. Для реалізації цієї вимоги у НЕ передбачається поділ процесу оброблення

даних на окремі сходинки та реалізація конвеєрного методу оброблення способом використання буферної пам'яті для узгодження тактів роботи конвеєру. У цьому випадку забезпечення режиму реального часу оброблення даних НЕ отримуємо унаслідок узгодження часу вводу масивів вагових коефіцієнтів НЕ та вхідних даних у НЕ з тактом роботи конвеєра, який реалізує виконання операцій обчислення даних у НЕ.

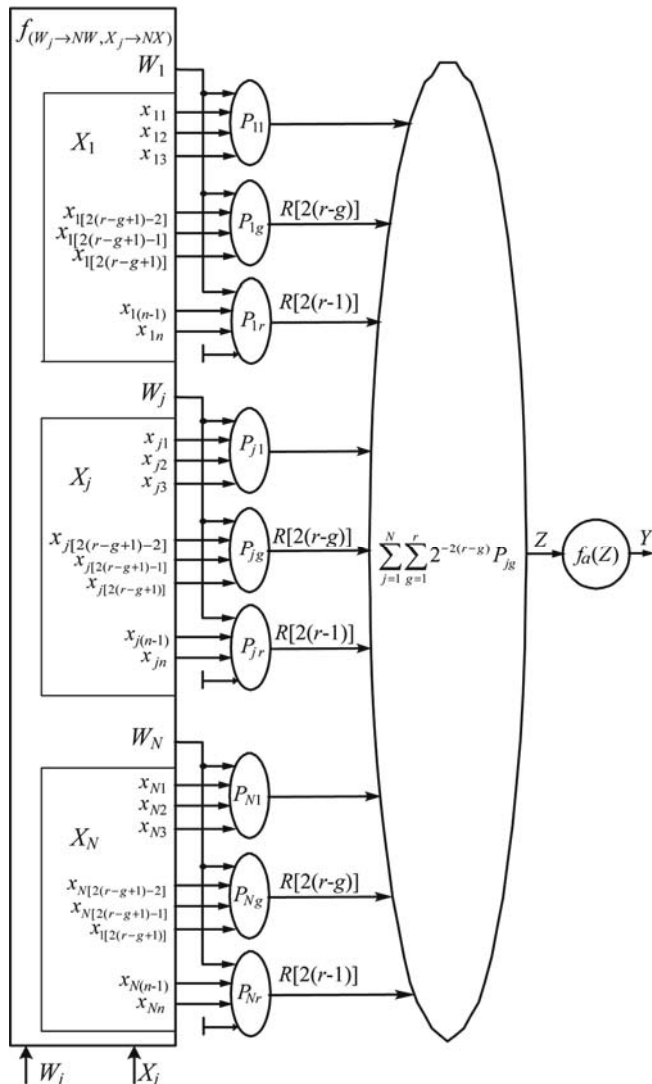


Рис. 1. Граф моделі НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення формування часткових добутоків / Graph of the model of the matrix-type NE using a modified Booth's algorithm to ensure the formation of partial products

Вибір матричного методу обчислення скалярного добутку для НЕ. Для реалізації матричного пристрою розрахунку скалярного добутку для застосування у НЕ необхідно удосконалити процедуру реалізації паралельно-паралельного алгоритму для обчислення і сумування часткових добутоків та розробити відповідну апаратну структуру пристрою. У процесі обчислення окремі часткові добутки можна розраховувати безпосередньо або використовуючи операції їх попереднього розрахунку. У випадку реалізації пристрою за матричним методом доцільно використати алгоритм з формуванням часткових добутоків безпосередньо, оскільки такий алгоритм є структурованим та регулярним. Для виконання операції прямого обчислення часткових добутоків доцільно використати алгоритм множення з виконанням аналізу од-

ного розряду у поєднанні з модифікованим алгоритмом Бута.

У такому випадку операція перемноження двох чисел X_j і W_j , що подані у двійковому доповняльному коді, та виконання аналізу розряду множника зводиться до операцій обчислення часткових добутоків та їх подальшого сумування, що можна подати формулою

$$C_j = W_j X_j = \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} W_j x_{ji} = \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} P_{ji}, \quad (3)$$

де: P_{ji} – значення i -го часткового добутку, використаного для розрахунку добутку j -х чисел; x_{ji} – значення i -го розряду j -го множника; n – розрядність множника.

З використанням наведеного вище алгоритму перемноження двох чисел розроблено алгоритм для реалізації розрахунку скалярного добутку відповідно до такої формули:

$$Z = \sum_{j=1}^N W_j X_j = \sum_{j=1}^N \sum_{i=0}^{n-1} (-1)^{2^i} 2^{-i} P_{ji}. \quad (4)$$

У цьому випадку розрахунок скалярного добутку, враховуючи формулу (3), вимагає обчислення $N \times n$ окремих часткових добутоків. Кількість вказаних операцій розрахунку часткових добутоків можна зменшити у два рази застосувавши для виконання вказаної операції модифіковану версію алгоритму Бута. При цьому операцію перемноження двох чисел з використанням такого модифікованого алгоритму Бута можна реалізувати відповідно до формули

$$C_j = W_j X_j = \sum_{g=1}^r 2^{-2(r-g)} W_j (x_{j2(r-g+1)-2} x_{j2(r-g+1)-1} x_{j2(r-g+1)}) = \sum_{g=1}^r 2^{-2(r-g)} P_{jg}, \quad (5)$$

де: P_{jg} – g -ий частковий добуток, що використовують для розрахунку добутку j -х чисел; $r = \lceil n/2 \rceil$, де $\lceil \cdot \rceil$ – знак заокруглення до більшого цілого числа. У такій реалізації для кожної g -тої групи розрядів множника $x_{j2(r-g+1)-2} x_{j2(r-g+1)-1} x_{j2(r-g+1)}$ розрахунок часткового добутку P_{jg} виконується за допомогою операції перемноження W_j на відповідний йому коефіцієнт K_{jg} . Величина коефіцієнта K_{jg} розраховується шляхом обчислення суми ваг ненульових цифр у групі розрядів множника $x_{j2(r-g+1)-2} x_{j2(r-g+1)-1} x_{j2(r-g+1)}$. При цьому розряд $x_{j2(r-g+1)-2}$ має вагомість два, а розряди $x_{j2(r-g+1)-1}$ та $x_{j2(r-g+1)}$ – одиницю, враховуючи такий вираз:

$$K_{jg} = \begin{cases} 2, & \text{коли } x_{j2(r-g+1)-2} = 0, x_{j2(r-g+1)-1} = x_{j2(r-g+1)} = 1; \\ 1, & \text{коли } x_{j2(r-g+1)-2} = 0, x_{j2(r-g+1)-1} \neq x_{j2(r-g+1)}; \\ 0, & \text{коли } x_{j2(r-g+1)-2} = x_{j2(r-g+1)-1} = x_{j2(r-g+1)}; \\ -1, & \text{коли } x_{j2(r-g+1)-2} = 1, x_{j2(r-g+1)-1} \neq x_{j2(r-g+1)}; \\ -2, & \text{коли } x_{j2(r-g+1)-2} = 1, x_{j2(r-g+1)-1} = x_{j2(r-g+1)} = 0. \end{cases} \quad (6)$$

У такому випадку операцію множення на два виконують шляхом зсуву на один розряд вліво, а зміну знаку реалізують виконанням інвертування всіх розрядів множеного з подальшим додаванням до молодшого розряду значення одиниці.

Унаслідок реалізації таких дій алгоритм розрахунку скалярного добутку з використанням для виконання обчислення часткових добутоків модифікованої версії алгоритму Бута (5) можна записати такою формулою:

$$Z = \sum_{j=1}^N W_j X_j = \sum_{j=1}^N \sum_{g=1}^r 2^{-2(r-g)} P_{jg}. \quad (7)$$

Для реалізації у цьому випадку паралельно-паралельного алгоритму обчислення скалярного добутку відповідно до формул (4), (7) необхідно виконати операції паралельного обчислення з подальшим сумуванням часткових добутків P_{jg} або P_{jg} , тобто реалізувати інтегральну макрооперацію групового сумування.

Отже, як показано, при синтезі НЕ матричного типу для виконання операцій розрахунку окремих часткових добутків доцільно застосувати їх обчислення з використанням модифікованого алгоритму Бута. Цей алгоритм забезпечує зменшення у два рази кількості необхідних розрахунків часткових добутків порівняно з алгоритмом, що здійснює аналіз одного розряду множника. При цьому, необхідна для розрахунку скалярного добутку, кількість окремих обчислень часткових добутків, що розраховують з використанням модифікованого алгоритму Бута, становить $N \times r$.

Розроблення алгоритмів і структур групового розрахунку часткових добутків для НЕ. Швидкодія НЕ матричного типу, що використовує для розрахунку часткових добутків модифікований алгоритм Бута, визначається тривалістю виконання інтегральної макрооперації групового сумування:

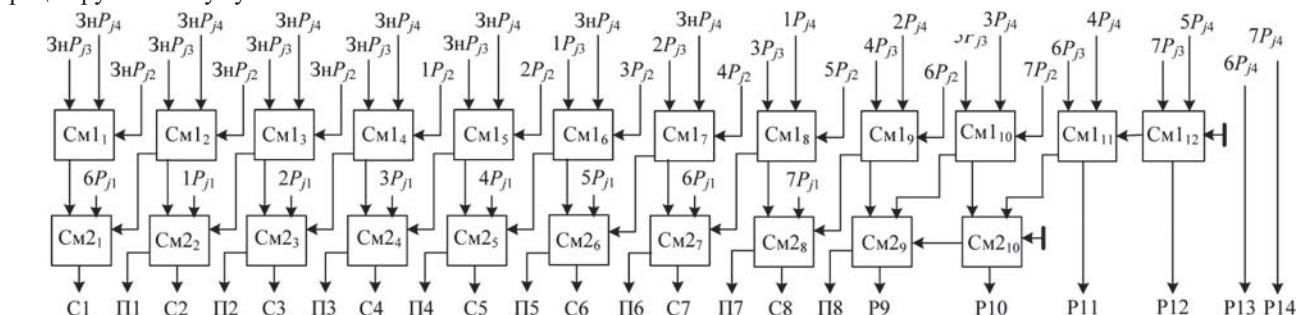


Рис. 2. Схема додавання часткових добутків / Scheme of addition of partial products

Реалізація підсумовування часткових добутків P_{jg} за схемою дерева Уоллеса зі збереженням переносів здійснюється на базі двох рядів однорозрядних суматорів. Необхідну кількість однорозрядних суматорів для реалізації перетворювача $n/2$ рядного коду у 2-рядний можна визначити за формулою

$$Q_{1CM} = \frac{n^2 - 4n}{2} + \sum_{k=1}^{n/2-2} 2k. \quad (9)$$

У такому випадку тривалість перетворення $n/2$ -рядного коду у дворядний код визначається за формулою

$$t_{n/2 \rightarrow 2} = (n/2 - 2)t_{1CM}, \quad (10)$$

де t_{1CM} – тривалість операції додавання однорозрядним суматором.

Використання N перетворювачів $n/2$ рядного коду у дворядний код забезпечує одночасне виконання перетворення N груп з $n/2$ часткових добутків у N дворядних кодів розрядністю n . На виході j -го перетворювача отримуємо дворядний код, який складається із n розрядів суми та n розрядів переносів.

Для додавання N дворядних кодів розрядністю n використовують каскадне паралельно-паралельне підсумовування. Розроблену схему для реалізації інтегральної макрооперації виконання групового сумування часткових добутків за формулою (8) наведено на рис. 3. У вказаній реалізації інтегральної макрооперації розрахунок групової суми часткових добутків власне процес його обчислення розглядають як виконання єдиної операції. Для реалізації макрооперації обчислення сум

$$Z = \sum_{j=1}^N \sum_{g=1}^r 2^{-2(r-g)} P_{jg}, \quad (8)$$

де: P_{jg} – значення g -го часткового добутку j -их чисел; $N \times r$ – кількість необхідних операцій обчислення часткових добутків.

Розглянемо принципову схему додавання часткових добутків. Для виконання операції сумування часткових добутків P_{jg} , які розраховують у процесі перемноження чисел $W_j \times X_j$, необхідно використати N перетворювачів з $n/2$ рядного коду у дворядний код. Операцію перетворення $n/2$ часткових добутків ($n/2$ рядного коду) у дворядний код (сумування та переноси) можна реалізувати на підставі структури дерева Уоллеса та алгоритму зі збереженням переносів. У пропонованій реалізації вагові коефіцієнти W_j НЕ та вхідні дані X_j НЕ представлені двійковим доповняльним кодом, тому для виконання операції сумування потрібно вказані часткові добутки $P_{j1}, \dots, P_{j(r-1)}$ доповнити знаковими розрядами. Розроблену схему перетворення 4-рядного коду у 2-рядний для чисел розрядністю 8 з використанням дерева Уоллеса та алгоритму зі збереженням переносів наведено на рис. 2.

парних добутків переважно використовують інший підхід, який дає змогу оптимізувати пристрій за швидкодією, апаратними витратами та збільшити однорідність і регулярність структури.

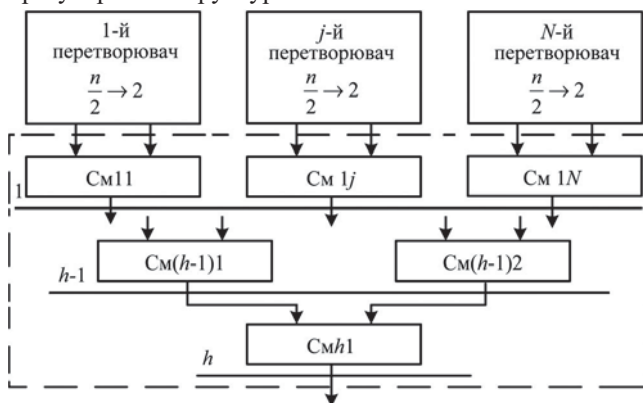


Рис. 3. Схема реалізації інтегральної макрооперації групового підсумовування часткових добутків / Scheme of implementation of integral macrooperation of group summation of partial products

Кількість ярусів, необхідних для реалізації операції каскадного сумування $2N$ чисел, обчислюється відповідно до формули

$$h = \lceil \log_2 2N \rceil, \quad (11)$$

де $\lceil \rceil$ – операція заокруглення до більшого цілого числа. У кожному ярусі каскадного сумування операнди поділяють на пари, для кожної з яких обчислюють су-

му. Загальна кількість n розрядних суматорів, необхідних для підсумовування $2N$ чисел, визначають як

$$Q_{nCM} = N - 1. \quad (12)$$

Розроблення базової структури НЕ матричного типу з формуванням часткових добутоків за модифікованим алгоритмом Бута. За організацією обчислювального процесу НЕ матричного типу, що використовує для обчислення часткових добутоків модифікований алгоритм Бута для операндів з рухомою комою, можна віднести до класу нерекурсивних пристроїв. У нерекурсивних НЕ матричного типу з використанням операндів з рухомою комою відсутні обернені зв'язки. Такі пристрої реалізують у вигляді матричних і паралельно-потоківих структур з використанням методів просторового та часового розпаралелення алгоритмів для реалізації операції розрахунку скалярного добутку. Результат обчислення у нерекурсивному НЕ матричного типу отримується за один прохід даних із входу до виходу через всі елементи НЕ.

Для забезпечення вимоги оптимального використання обладнання структура НЕ матричного типу повинна бути орієнтована на реалізацію з використанням прог-

рамованих логічних інтегральних схем (ПЛІС, англ. *Programmable Logic Device, PLD*). Для повного використання переваг ПЛІС НЕ матричного типу потрібно реалізувати з використанням таких принципів:

- реалізація НЕ у базисі елементарних арифметичних операцій;
- реалізація операції обчислення скалярного добутку у вигляді інтегральної макрооперації групового сумування часткових добутоків;
- реалізація інтегральної макрооперації групового сумування як єдиної операції;
- узгодження у НЕ інтенсивності надходження вхідних даних з інтенсивністю виконання обчислень та інтенсивністю видачі результатів;
- використання масового паралелізму оброблення даних у НЕ.

На базі граф моделі НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення обчислення часткових добутоків розроблено базову структуру НЕ (рис. 4). На цьому рисунку W_j , X_j – входи відповідно вагових коефіцієнтів НЕ і вхідних даних НЕ, Рг – регістр, БРг – буферний регістр, ФЧД – формувач часткового добутку, БФЧД – блок формування часткових добутоків, Y – вихід НЕ.

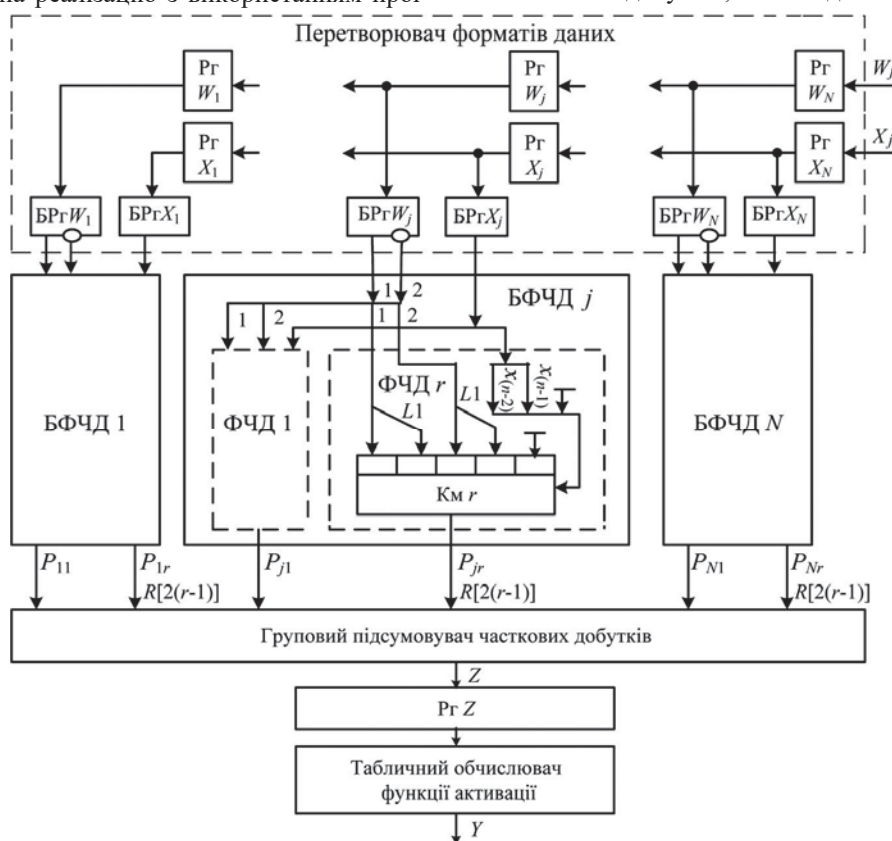


Рис. 4. Базова структура НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення формування часткових добутоків / The basic structure of the matrix type NE using a modified Booth algorithm to ensure the formation of partial products

Основними компонентами НЕ матричного типу з використанням модифікованого алгоритму Бута для забезпечення формування часткових добутоків є: перетворювач форматів даних, N БФЧД, груповий підсумовувач часткових добутоків і табличний обчислювач функції активації.

Перетворювач форматів даних здійснює перетворення N вхідних даних X_j і N вагових коефіцієнтів W_j , що надходять послідовно двома каналами, у два паралельні масиви із N вагових коефіцієнтів $W_1, \dots, W_N$ та N вхідних даних $X_1, \dots, X_N$. Кожен канал надходження даних складається із N послідовно з'єднаних регістрів Рг W_1 ,

$\dots$, Рг W_N (Рг $X_1, \dots$, Рг X_N), з яких інформація переписується у буферні регістри БРг $W_1, \dots$, БРг W_N (БРг $X_1, \dots$, БРг X_N). З виходів буферних регістрів БРг $W_1, \dots$, БРг W_N та БРг $X_1, \dots$, БРг X_N дані надходять на входи відповідно БФЧД $_1, \dots$, БФЧД $_N$.

Для обчислення добутку чисел $W_j X_j$ у БФЧД $_j$ здійснюють формування r часткових добутоків $P_{j1}, \dots, P_{jr}$. Кожен g -ий, де $g = 1, \dots, r$, частковий добуток P_{jg} надходить зсувом на $2(g-1)$ розрядів вправо на входи групового підсумовувача часткових добутоків. Реалізація групового підсумовувача часткових добутоків (формула (7)) здійснюють за схемою (див. рис. 2).

На виході групового підсумовувача часткових добутоків отримуємо скалярний добуток, який записується у регістр PrZ. Скалярний добуток Z з виходів регістра PrZ надходить на адресні входи табличного обчислювача функції активації, на виході якого отримуємо вихідний сигнал нейронного елемента Y.

Удосконалення методу синтезу НЕ матричного типу для оброблення інтенсивних потоків даних у реальному часі. Задача синтезу НЕ матричного типу полягає у реалізації структури, що забезпечує обчислення у режимі реального часу з урахуванням мінімальних апаратних затрат на його реалізацію. Початковими даними для синтезу НЕ матричного типу є тривалість надходження N вагових коефіцієнтів W_j і N вхідних даних X_j , який визначається так:

$$t_d = N \frac{1}{f_d}, \quad (13)$$

де f_d – частота надходження вагових коефіцієнтів W_j і вхідних даних X_j .

Для забезпечення роботи НЕ матричного типу у режимі реального часу необхідно використовувати методи часового розпаралелення процедури виконання операцій (конвеєризацію). Конвеєризація передбачає розбиття процесу обчислення результатів у НЕ на сходинок, кожна із яких містить дві необхідні компоненти – буферну пам'ять та операційний блок. Для забезпечення обчислень у НЕ матричного типу у реальному часі необхідне виконання умови $t_d \geq T_k$, де T_k – конвеєрний такт роботи НЕ. Конвеєрний такт T_k роботи НЕ матричного типу визначається як тривалість оброблення даних на сходінці конвеєра.

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

- швидкодії, застосованої для реалізації НЕ елементної бази;
- кількості вагових коефіцієнтів і вхідних даних НЕ та алгоритму формування часткових добутоків;
- алгоритму та засобів групового підсумовування часткових добутоків.

Кількість сходінок конвеєра визначається складністю обчислювальних операцій, які реалізують на цій сходінці. Зменшення складності операцій, які реалізують окремою сходінкою конвеєра, призводить до збільшення їх кількості та зменшення тривалості конвеєрного такту. Зі зменшенням кількості сходінок підвищується складність операцій, які необхідно виконати на окремій сходінці, та одночасно зростає тривалість конвеєрного такту.

Апаратні витрати на реалізацію НЕ матричного типу зростають із збільшенням необхідної для реалізації обчислень кількості сходінок конвеєра.

Синтез НЕ матричного типу з виконанням вимог мінімальних витрат обладнання та забезпечення оброблення інтенсивних потоків даних у режимі реального часу можливий за умови узгодження тривалості надходження вхідних даних і вагових коефіцієнтів t_d у НЕ з конвеєрним тактом T_k роботи НЕ. Цієї умови можна досягнути зменшенням конвеєрного такту T_k оброблення даних у НЕ. Основним шляхом зменшення T_k є конвеєризація операцій групового сумування часткових добутоків у НЕ способом розбиття на сходинок.

Оцінювання параметрів НЕ матричного типу з формуванням часткових добутоків за модифікованим

алгоритмом Бута. Основними функціональними вузлами, на базі яких синтезується НЕ матричного типу з використанням модифікованого алгоритму Бута для розрахунку часткових добутоків, є: пам'ять, суматори, регістри, комутатори, дешифратори та логічні елементи. Оскільки розроблена структура НЕ матричного типу орієнтована на НВІС реалізацію, за одиницю розрахунку витрат обладнання обрано логічний вентиль (елемент типу інвертор, I, АБО), а для оцінювання часових параметрів – величину затримки логічного вентиля τ . Витрати обладнання на реалізацію функціональних вузлів у вентилях та їхню швидкодію наведено у таблиці, де ПЗП – постійний запам'ятовувальний пристрій.

Таблиця. Витрати обладнання на реалізацію функціональних вузлів та їх швидкодія / Equipment costs for the implementation of functional nodes and their speed

№ з/п	Назва функціонального вузла	Витрати обладнання (вентилів)	Кількість каскадів затримки (τ вентилів)
1	тригер	6	3
2	n -розрядний регістр	$7n$	3
3	однорозрядний суматор	18	7
4	n -розрядний суматор	$20n$	$7 \log_2 n$
5	m -вхідний n -розрядний комутатор	$3mn$	m
6	m -адресний n -розрядний ПЗП	$2_m n$	$(m+3)$

З використанням розроблених аналітичних виразів (див. таблицю) виконано оцінювання апаратних витрат обладнання на реалізацію компонент НЕ (перетворювача форматів даних, БФЧД, групового підсумовувача часткових добутоків, табличного обчислювача передавальної функції), яке виконують за формулою

$$W_{KHEr} = \sum_{j=1}^{k_r} W_{\Phi B r j} q_{rj}, \quad (14)$$

де: W_{KHEr} – витрати обладнання на реалізацію r -тої компоненти НЕ; k_r – кількість типів функціональних вузлів у r -ий компоненті НЕ; $W_{\Phi B r j}$ – витрати обладнання на j -ий тип функціонального вузла r -ої компоненти НЕ; q_{rj} – кількість функціональних вузлів j -го типу r -тої компоненти НЕ.

Оцінювання витрат обладнання на апаратну реалізацію НЕ виконують за формулою

$$W_{HM} = \sum_{r=1}^M \sum_{j=1}^{k_r} W_{\Phi B r j} q_{rj}, \quad (15)$$

де M – кількість компонент НЕ.

Витрати обладнання на реалізацію пропонованого НЕ матричного типу визначають за таким виразом:

$$W_{HE} = W_{ПФД} + N W_{БФЧД} + W_{ГПЧД} + W_{P_2} + W_{ТОПФ}, \quad (16)$$

де: $W_{ПФД}$ – витрати обладнання на перетворювач форматів даних; $W_{БФЧД}$ – витрати обладнання на блок формування часткових добутоків; $W_{ГПЧД}$ – витрати обладнання на груповий підсумовувач; $W_{ТОПФ}$ – витрати обладнання на табличний обчислювач функції активації. У формулу (16) підставляємо кількість витрат обладнання на реалізацію окремих функціональних вузлів, які використовують для синтезу компонент НЕ й отримуємо

$$\begin{aligned} W_{HE} &= W_{ПФД} + N W_{БФЧД} + W_{ГПЧД} + W_{P_2} + W_{ТОПФ} = \\ &= 4N W_{P_2 n} + N r W_{K m 5 \rightarrow n} + W_{C m 1} \left(\frac{n^2 - 4n}{2} + \sum_{k=1}^{n/2-2} 2k \right) + \\ &\quad + (N - 1) W_{C m n} + W_{P_2 n} + W_{ПЗП n \rightarrow n}, \end{aligned} \quad (17)$$

де: $W_{Pz,n}$ – витрати обладнання на реалізацію n -розрядного регістра; $r = \lceil n/2 \rceil$, $W_{Km5 \rightarrow n}$ – витрати обладнання на реалізацію 5-входового n -розрядного комутатора; W_{Cmn} – витрати обладнання на реалізацію n -розрядного суматора; $W_{ПЗПn \rightarrow n}$ – витрати обладнання на реалізацію n -адресного n -розрядного ПЗП; $W_{Pz,n}$ – витрати обладнання на реалізацію n -розрядного регістра.

Підставляючи у формулу (17) значення витрат обладнання на реалізацію окремих функціональних вузлів у вентилях (див. таблицю) отримаємо вираз

$$W_{HE} = (28Nn + 10Nrn + 18(\frac{n^2 - 4n}{2} + \sum_{k=1}^{n/2-2} 2k) + 20n(N-1) + 7n + 2^n n) \times \text{вентилів}. \quad (18)$$

Конвеєрний такт T_k роботи НЕ матричного типу визначається тривалістю формування та групового підсумовування часткових добутків і оцінюють за такою формулою:

$$T_k = t_{Pz} + t_{ФЧД} + t_{ГПЧД} = t_{Pz} + t_{Km5} + t_{n/2 \rightarrow 2} + \lceil \log_2 2N \rceil t_{Cmn}, \quad (19)$$

де: t_{Pz} – тривалість запису в регістр; t_{Km5} – тривалість затримки даних на 5-ти входовому комутаторі; $t_{n/2 \rightarrow 2}$ – тривалість перетворення $n/2$ рядного коду у 2-рядний; t_{Cmn} – тривалість додавання двох n -розрядних чисел.

Підставляючи у формулу (19) значення кількості каскадів затримки при проходженні даних через функціональні вузли у τ вентилів отримаємо вираз

$$T_k = 3 + 5 + 7(n/2 - 2) + 7 \lceil \log_2 2N \rceil \log_2 n = \{8 + 7(n/2 - 2) + \lceil \log_2 2N \rceil \log_2 n\} \tau. \quad (20)$$

З аналітичних виразів (18) і (20) видно, що витрати обладнання і конвеєрний такт НЕ залежать від розрядності n та кількості N операндів.

З використанням розробленого аналітичного виразу (18) побудовано графіки витрат обладнання на реалізацію розробленого НЕ матричного типу (рис. 5,а).

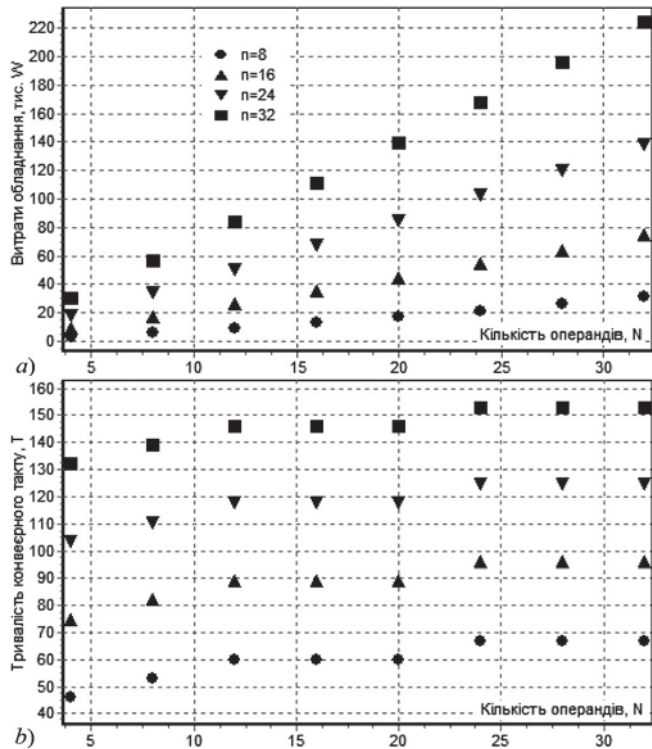


рис. 5. Графіки витрат обладнання (а) та тривалості конвеєрного такту T_k (б) на реалізацію НЕ матричного типу з формуванням часткових добутків за модифікованим алгоритмом Бута / Graphs of equipment costs (а) of the duration of the T_k

conveyor cycle (b) for the implementation of matrix type NE with the formation of partial products according to the modified Booth's algorithm

На підставі аналітичного виразу (20) побудовано графіки оцінювання тривалості конвеєрного такту T_k розробленого НЕ матричного типу (рис. 5,б).

Згідно з даними рис. 5, витрати обладнання та тривалість конвеєрного такту залежать як від розрядності операндів n , так і від їх кількості N . На підставі отриманих значень витрат обладнання і тривалості конвеєрного такту обирають ПЛІС для реалізації НЕ із заданими параметрами.

Обговорення результатів дослідження. Порівняння розробленої моделі НЕ матричного типу з використанням модифікованого алгоритму Бута для формування часткових добутків із моделлю НЕ, поданою у роботі [22], дає змогу зробити висновок, що в розробленій моделі НЕ процес обчислення скалярного добутку зведений до операції групового підсумовування часткових добутків з використанням дерев Уоллеса з використанням каскадного паралельно-паралельного алгоритму та процесу конвеєризації, що забезпечує апаратну реалізацію НЕ реального часу на ПЛІС.

У роботі [5] описано реалізацію на ПЛІС штучного нейрона з чотирма входами та сигмоподібною функцією активації з використанням 16-розрядних чисел із фіксованою комою. Представлена у цій роботі реалізація НЕ матричного типу дає змогу застосовувати різні функції активації застосовуючи їх табличну реалізацію.

У роботі [15] представлено бітово-последовну апаратно-оптимізовану реалізацію НЕ для імпульсної нейронної мережі. Однак запропонований бітовий последовний дизайн добре працює для даних з невеликою розрядністю. Для більшої розрядності даних така конструкція вимагає багато додаткових тактів оброблення даних. Для запропонованого у роботі НЕ розрядність даних може становити 8-32 біт.

Автори у роботі [1] досліджують підхід до розроблення комплексно значної нейронної мережі CVNN (англ. *Complex-Valued Neural Network*) на FPGA (англ. *Field-Programmable Gate Array* – програмована користувачем вентильна матриця, ПКВМ). Однак наведена реалізація призначена для класифікації 2D-зображень, перетворених у полярну форму, а дослідження було обмежено набором даних рукописних цифр MNIST (англ. *Modified National Institute of Standards and Technology* – об'ємна база даних зразків рукописного написання цифр). Пропонуваний у роботі НЕ матричного типу орієнтований на універсальне оброблення даних і не прив'язаний до конкретної реалізації нейромережі.

Подальші дослідження з розроблення НЕ матричного типу пов'язані зі збільшенням діапазону зміни величин вхідних даних і вагових коефіцієнтів шляхом їх представлення у форматі з рухомою комою. Робота з вхідними даними і ваговими коефіцієнтами у форматі з рухомою комою передбачає використання послідовно-паралельного перетворення та вирівнювання порядків вхідних даних і вагових коефіцієнтів.

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

Наукова новизна отриманих результатів дослідження – розроблено аналітичну та графову моделі НЕ

матричного типу з використанням модифікованого алгоритму Бута для обчислення часткових добутоків, апаратна реалізація НЕ на підставі якої забезпечує оброблення інтенсивних потоків даних у режимі реального часу з використанням НВІС. Удосконалено метод синтезу НЕ матричного типу, який за рахунок узгодження тривалості надходження вхідних даних і вагових коефіцієнтів у НЕ з тривалістю конвеєрного такту забезпечує зменшення апаратних затрат та оброблення інтенсивних потоків даних у режимі реального часу.

Практична значущість результатів дослідження – використання розроблених алгоритмів і структур для розрахунку часткових добутоків на підставі дерев Уоллеса та каскадного паралельно-паралельного алгоритму забезпечує зменшення тривалості оброблення. Використання розроблених моделей та удосконаленого методу синтезу НЕ матричного типу на підставі використання модифікованого алгоритму Бута для обчислення часткових добутоків забезпечує реалізацію НЕ, що дає змогу здійснити оброблення інтенсивних потоків даних у режимі реального часу та застосувати у НВІС.

Висновки / Conclusions

Широке застосування нейромереж для опрацювання даних вимагає розроблення нових і вдосконалення наявних засобів для їх розроблення. При цьому, за умови застосування таких засобів на мобільних платформах, потрібно враховувати додаткові обмеження стосовно габаритів, маси та енергоспоживання. Одним із перспективних шляхів вирішення цієї проблеми є застосування НВІС, що, однак, вимагає розроблення нових моделей, структур та методів синтезу НЕ як основи для побудови нейронних та нейроподібних мереж. У роботі проаналізовано підхід до розроблення відповідного НЕ. За результатами дослідження можна зробити такі основні висновки.

1. Розроблено аналітичну та графову моделі НЕ матричного типу. Для синтезу вказаного НЕ обрано матричний метод розрахунку скалярного добутку, розроблено алгоритми і структури для виконання групового сумування часткових добутоків у процесі обчислень. Розроблено базову структуру для апаратної реалізації НЕ матричного типу, що використовує для обчислення часткових добутоків модифікований алгоритм Бута. У роботі вдосконалено метод синтезу НЕ матричного типу для забезпечення оброблення інтенсивних потоків даних у режимі реального часу та виконано оцінювання необхідних для реалізації НЕ витрат обладнання та тривалості такту конвеєра у НЕ.
2. Запропонована у роботі аналітична та графова моделі НЕ матричного типу дають можливість отримати апаратну реалізацію НЕ, що забезпечує оброблення потоків даних у режимі реального часу, мінімальні витрати обладнання та реалізацію на ПЛІС.
3. Обрано матричний метод для виконання розрахунку скалярного добутку та обчислення часткових добутоків з використанням модифікованого алгоритму Бута, що дало змогу зменшити у два рази необхідну кількість розрахунків часткових добутоків.
4. Запропоновано нові алгоритми та структури для реалізації групового сумування часткових добутоків, які внаслідок розпаралелення процесу підсумовування на підставі дерев Уоллеса та використання каскадного паралельно-паралельного алгоритму обчислень забезпечують зменшення відповідного часу підсумовування.
5. Розроблено базову структуру НЕ матричного типу з використанням для обчислення часткових добутоків моди-

фікованого алгоритму Бута, яка за рахунок узгодження тривалості надходження вхідних даних і даних про вагові коефіцієнти з тривалістю конвеєрного такту забезпечує оброблення у режимі реального часу інтенсивних потоків даних. Розроблена базова структура НЕ дає можливість ефективної реалізації її на ПЛІС.

6. Розроблено аналітичні вирази для оцінювання витрат обладнання на реалізацію НЕ та швидкодії функціональних вузлів, які використовують при синтезі НЕ матричного типу. Виконано оцінювання витрат обладнання на реалізацію НЕ, розраховано тривалість конвеєрного такту НЕ залежно від розрядності операндів та кількості парних добутоків.

References

1. Ahmad, Maruf, Lei Zhang, & Muhammad, E.H. Chowdhury. (2024). FPGA Implementation of Complex-Valued Neural Network for Polar-Represented Image Classification. *Sensors*, 24(3), 897 p. <https://doi.org/10.3390/s24030897>
2. Cai, J., Bao, H., Chen, M., Xu, Q., & Bao, B. (2022). Analog/Digital Multiplierless Implementations for Nullcline-Characteristics-Based Piecewise Linear Hindmarsh-Rose Neuron Model. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(7), 2916–2927. <https://doi.org/10.1109/TCSI.2022.3164068>
3. Cai, J., Bao, H., Chen, M., Xu, Q., & Bao, B. (2022). Analog/Digital Multiplierless Implementations for Nullcline-Characteristics-Based Piecewise Linear Hindmarsh-Rose Neuron Model. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(7), 2916–2927. <https://doi.org/10.1109/TCSI.2022.3164068>
4. Chen, K.-H., Su, C.-W., & Wang, J.-H. (2022). Energy-efficient and Accurate Object Detection Design on an FPGA Platform. *IET International Conference on Engineering Technologies and Applications (IET-ICETA)*, Changhua, Taiwan, 1–2. <https://doi.org/10.1109/IET-ICETA56553.2022.9971590>
5. Doroshenko, A., Shymkovych, V., Mamedov, T., & Yatsenko, O. (2022). Automated design of an artificial neuron for field-programmable gate arrays based on an algebra-algorithmic approach. *International Scientific Technical Journal "Problems of Control and Informatics"*, 67(5), 61–72. <https://doi.org/10.34229/2786-6505-2022-5-6>
6. Drozd, J., Drozd, O., Nikul, V., & Sulima, J. (2018). FPGA implementation of vertical addition with a bitwise pipeline of calculations. *9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, Kyiv, Ukraine, 239–242. <https://doi.org/10.1109/DESSERT.2018.8409136>
7. Haghir, S., & Ahmadi, A. (2020). A Novel Digital Realization of AdEx Neuron Model. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(8), 1444–1448. <https://doi.org/10.1109/TCSII.2019.2938180>
8. Hong, S., Lee, I., & Park, Y. (2018). Optimizing a FPGA-based neural accelerator for small IoT devices. *International Conference on Electronics, Information, and Communication (ICEIC)*, Honolulu, HI, USA, 1–2. <https://doi.org/10.23919/ELINFO-COM.2018.8330546>
9. Imani, M. A., Ahmadi, A., RadMalekshahi M., & Haghir, S. (2018). Digital Multiplierless Realization of Coupled Wilson Neuron Model. *Transactions on Biomedical Circuits and Systems*, 12(6), 1431–1439. <https://doi.org/10.1109/TBCAS.2018.2869319>
10. Jung, J., & Lee, K. J. (2020). An Energy-Efficient Deep Neural Network Accelerator Design. *54th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 272–276. <https://doi.org/10.1109/IEEECONF51394.2020.9443508>
11. Kongpoon, M., & Leelavattananon, K. (2019). Multiplier-less and compact FPGA implementation of Mihalas-Niebur neuron. *IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Bangkok, Thailand, 321–324. <https://doi.org/10.1109/APCCAS47518.2019.8953116>
12. Leigh, A. J., Heidarpur, M., & Mirhassani, M. (2022). A High-Accuracy Digital Implementation of the Morris-Lecar Neuron With Variable Physiological Parameters. *Transactions on Circuits*

- and Systems II: Express Briefs, 69(10), 4138–4142. <https://doi.org/10.1109/TCSII.2022.3187623>
13. Leigh, A. J., Mirhassani, M., & Muscedere, R. (2020). An Efficient Spiking Neuron Hardware System Based on the Hardware-Oriented Modified Izhikevich Neuron (HOMIN) Model. *Transactions on Circuits and Systems II: Express Briefs*, 67(12), 3377–3381. <https://doi.org/10.1109/TCSII.2020.2984932>
 14. Lin, X., Lu, H., Pi, X., & Wang, X. (2020). An FPGA-based Implementation Method for Quadratic Spiking Neuron Model. 11th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 0621–0627. <https://doi.org/10.1109/UEMCON51285.2020.9298029>
 15. Losh, M., & Llamocca, D. (2019). A Low-Power Spike-Like Neural Network Design. *Electronics*, 8, 1479 p. <https://doi.org/10.3390/electronics8121479>
 16. Mandal, R. K., & Kale, P. (2018). Development of a Decision-Based Neural Network for a Day-Ahead Prediction of Solar PV Plant Power Output. 4th International Conference on Computational Intelligence & Communication Technology (CICIT), Ghaziabad, India, 1–6. <https://doi.org/10.1109/CICT.2018.8480396>
 17. Rabyk, V., Kryvinska, N., Yatsymirskyy, M., Teslyuk, V. (2022). Design of the Processors for Fast Cosine and Sine Fourier Transforms. *Circuits, Systems, and Signal Processing*, 41(9), 4928–4951. <https://doi.org/10.1007/s00034-022-02012-8>
 18. Shrestha, R. (2017). High-speed and low-power VLSI-architecture for inexact speculative adder. *International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, Hsinchu, Taiwan, 1–4. <https://doi.org/10.1109/VLSI-DAT.2017.7939644>
 19. Song, M., & Pedrycz, W. (2013). Granular Neural Networks: Concepts and Development Schemes. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4), 542–553. <https://doi.org/10.1109/TNNLS.2013.2237787>
 20. Takaloo, H., Ahmadi, A., & Ahmadi, M. (2023). Design and Analysis of the Morris – Lecar Spiking Neuron in Efficient Analog Implementation. *Transactions on Circuits and Systems II: Express Briefs*, 70(1), 6–10. <https://doi.org/10.1109/TCSII.2022.3203929>
 21. Tchyntetskiy, S., Peleshchak, M., Peleshchak, I., & Vysotska, V. (2021). A Neural Network Development for Multispectral Images Recognition. *IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, 278–284. <https://doi.org/10.1109/CSIT52700.2021.9648735>
 22. Tsmots, I. G., Opotyak, Yu. V., & Shtohrinets, B. V. (2023). Method of Synthesis of Devices for Parallel Stream Calculation of Scalar Product in Real Time. *Journal of Lviv Polytechnic National University "Information Systems and Networks"*, 14, 248–266. <https://doi.org/10.23939/sisn2023.14.248>
 23. Tsmots, I., Rabyk, V., Teslyuk, V., & Opotyak, Y. (2023). Floating-Point Number Scalar Product Hardware Implementation for Embedded Systems. 17th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), Jaroslaw, Poland, 6–10. <https://doi.org/10.1109/CADSM58174.2023.10076502>
 24. Tsmots, I., Skorokhoda, O., & Rabyk, V. (2018). Parallel algorithms and matrix structures for scalar product calculation. 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 936–939. <https://doi.org/10.1109/TCSET.2018.8336347>
 25. Yang, S., Liu, P., Xue, J., Sun, R., & Ying, R. (2020). An Efficient FPGA Implementation of Izhikevich Neuron Model. *International SoC Design Conference (ISOC)*, Yeosu, Korea (South), 141–142. <https://doi.org/10.1109/ISOC50952.2020.9333014>
 26. Yang, Y., et al. (2023). A Network Traffic Classification Method Based on Dual-Mode Feature Extraction and Hybrid Neural Networks. *IEEE Transactions on Network and Service Management*, 20(4), 4073–4084. <https://doi.org/10.1109/TNSM.2023.3262246>
 27. Zhang, H., Du, C., & Ko, S.-B. (2023). Energy Efficient FPGA-Based Accelerator for Deep Spiking Neural Networks. *IEEE 3rd International Conference on Computer Systems (ICCS)*, Qingdao, China, 168–172. <https://doi.org/10.1109/ICCS59700.2023.10335532>
 28. Zhang, W., Zhang, C., Niu, L., Din, F. U., Farrukh, & Jiang, H. (2022). An Efficient FPGA Design for Fixed-point Exponential Calculation. *International Conference on Integrated Circuits, Technologies and Applications (ICTA)*, Xi'an, China, 44–45. <https://doi.org/10.1109/ICTA56932.2022.9963050>

I. G. Tsmots, Yu. V. Opotyak, B. V. Shtohrinets, T. B. Mamchur, V. M. Holubets

Lviv Polytechnic National University, Lviv, Ukraine

MODEL, STRUCTURE AND SYNTHESIS METHOD OF MATRIX-TYPE NEURAL ELEMENT

The growing use of neural networks for data processing requires the search for new and improvement of existing means for their creation. The use of such tools on mobile platforms requires considering additional restrictions on dimensions, weight and power consumption. A promising way to solve this problem is the use of ultra-large integrated circuits (UICs), which, however, requires the development of new models, structures and methods of synthesis of neural elements (NE) as a basis for building neural and neuro-like networks. The paper proposes and describes a generalized model, structure and synthesis method of the matrix-type NE. Algorithms for calculating the scalar product, group summation of partial products, which are used to implement the basic structure of NE, are described. The parameters of the matrix-type NE were evaluated with the formation of partial products according to the modified Booth's algorithm. The developed generalized model of the matrix-type NE with the formation of partial products according to the modified Booth's algorithm is described, the components of which are a data format converter, a device for calculating a scalar product, and a block for implementing the activation function. The developed NE model ensures real-time data processing by matching the time of arrival of input data and weighting factors with the conveyor cycle. In order to halve the number of calculations of partial products, it is proposed to use the matrix method of calculating the scalar product with the formation of partial products using a modified Booth's algorithm. Developed algorithms and structures for group summation of partial products are described, which, ensure a reduction in summation time due to the parallelization of the partial product summation process based on Wallace trees and the cascade parallel-parallel algorithm. The developed basic structure of the matrix-type NE with the formation of partial products using the modified Booth's algorithm is described, which provides real-time data streams processing and implementation on the UICs due to the parallel-parallel formation and summation of partial products based on Wallace trees and the cascade algorithm using pipeline registers to match the intensity of data arrival with the intensity of calculation. The hardware costs for the implementation of the matrix-type NE and the duration of the conveyor cycle for the specified configurations of computing devices, namely, the bit rate of the operands and the number of pairs of products, are estimated.

Keywords: neural network; pipeline method of calculations; Booth's algorithm; conveyor cycle duration; equipment costs; real-time data processing.