КОМП'ЮТЕРНА ІНЖЕНЕРІЯ

ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
https://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2021.03.063
Article received 09.04.2021 p.
Article accepted 01.06.2021 p.
UDK 004.272.3

Correspondence author
I. H. Tsmots
ivan.tsmots@gmail.com

І. Г. Цмоць[1], Ю. А. Лукащук[1], І. В. Ігнатєв[2], І. Я. Казимира[1]

[1] Національний університет "Львівська політехніка", м. Львів, Україна
[2] Західноукраїнський національний університет, м. Тернопіль, Україна

КОМПОНЕНТИ АПАРАТНИХ НЕЙРОННИХ МЕРЕЖ УЗГОДЖЕНОГО ПАРАЛЕЛЬНО-ВЕРТИКАЛЬНОГО ОБРОБЛЕННЯ ДАНИХ У РЕАЛЬНОМУ ЧАСІ

Сформовано операційний базис нейронних мереж і вибрано для апаратної реалізації такі операції: пошуку максимального і мінімального значень, обчисленням суми квадратів різниць і скалярного добутку. Визначено вимоги до апаратних компонентів нейронних мереж з узгодженою вертикально-паралельним обробленням даних, основними з яких є забезпечення: високої ефективності використання обладнання, адаптації до вимог конкретних застосувань, узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень у апаратній компоненті, роботу у реальному часі, структурної орієнтації на НВІС-реалізацію, малого часу розробки та невисокої вартості. Показано, що основними шляхами управління інтенсивністю обчислень у апаратних компонентах є вибір кількості та розрядності трактів опрацювання даних, зміна тривалості такту роботи шляхом вибору елементної бази та складності операцій, які реалізуються сходинками конвеєра. Запропоновано для реалізації апаратних компонент нейронних мереж з узгодженою вертикально-паралельним обробленням управління використовувати паралельні вертикально-групові методи опрацювання даних, які забезпечують управління інтенсивністю обчислень, зменшення апаратних затрат і НВІС реалізацію. Розроблено паралельний вертикально-груповий метод обчислення максимальних і мінімальних чисел у масивах, який за рахунок паралельного опрацювання зрізу з групи розрядів всіх чисел забезпечує зменшення часу обчислення. Розроблено паралельний вертикально-груповий метод і структуру компоненти обчислення суми квадратів різниць, яка за рахунок розпаралелення та вибору кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання. Розроблено паралельний вертикально-груповий метод та структуру компоненти обчислення скалярного добутку, яка порівняно з відомими за рахунок вибору розрядності трактів оброблення та кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання. Показано, що використання розроблених компонентів для синтезу нейронних мереж з узгодженою вертикально-паралельним обробленням даних у реальному часі забезпечить зменшення часу і вартості їх реалізації.

Ключові слова: нейронні мережі; апаратні компоненти; узгоджена паралельно-вертикальна оброблення; вертикально-групові методи; реальний час.

Вступ

При використанні нейронних технологій реального часу у промисловості (управління технологічними процесами та складними об'єктами), енергетиці (оптимізація навантаження в електромережах), військовій справі (технічний зір, управління рухом мобільного робота, криптографічний захист даних), транспорті (управління рухом і двигуном), медицині (діагностика захворювань) і приладобудуванні (розпізнавання образів і оптимізація управління) вимагається опрацювання інтенсивних потоків даних засобами, що задовольняють обмеженням відносно габаритів, маси, енергоспоживання та мають високу ефективність використання обладнання. Для забезпечення широкого спектру застосувань необхідно виділити та розробити апаратні компоненти, які просто адаптуються до вимог конкретних застосувань і можуть використовуватися для синтезу широкого спектру апаратних нейронних мереж реального часу. Синтез високоефективних компонентів апаратних нейронних мереж реального часу потребує широкого використання сучасної елементної (напівзамовних і замовних надвеликих інтегральних схем (НІВС), однокристальних нейропроцесорів, систем на кристалі, мікрокомп'ютерів, мікроконтролерів), розроблення нових методів і НВІС-структур. Орієнтація нейронних апаратних компонентів на НВІС-реалізацію з високою ефективністю використання обладнання вимагає зменшення кількості виводів інтерфейсу, міжнейронних зв'язків, узгодження інтенсивності обчислень з інтенсивністю надходження

даних шляхом вибору величини такту роботи, кількості та розрядності трактів опрацювання даних.

Забезпечити ці вимоги можна шляхом використанням паралельних вертикально-групових методів опрацювання та структур, які адаптуються до вимог конкретного застосування.

У зв'язку з цим особливої актуальності набуває проблема розроблення компонентів апаратних нейронних мереж з узгодженою вертикально-паралельним обробленням даних у реальному часі, які адаптуються до інтенсивності надходження даних, забезпечують високу ефективність використання обладнання та орієнтовані на синтез широкого спектру апаратних нейронних мереж.

Об'єкт дослідження – узгодження, пошук виконання обчислень у апаратних компонентах нейронних мереж узгоджено паралельно-вертикальним обробленням даних у реальному часі.

Предмет дослідження – паралельні вертикально-групові методи і структури компонентів пошуку максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць і скалярного добутку з високою ефективністю використання обладнання.

Мета роботи – розроблення паралельних вертикально-групових методів і структуру апаратних компонентів нейронних мереж узгодженої паралельно-вертикального оброблення даних у реальному часі з високою ефективністю використання обладнання.

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

- аналіз останніх досліджень та публікацій;
- виділення операційного базису для реалізації апаратних нейронних мереж;
- розроблення методу узгодження інтенсивності надходження даних з інтенсивністю обчислень у апаратних компонентах нейронних мереж;
- розроблення паралельних вертикально-групових методів і структур компонентів пошуку максимальних і мінімальних чисел у масивах, обчисленні суми квадратів різниць і скалярного добутку.

Наукова новизна дослідження полягає в такому:

- розроблений паралельний вертикально-груповий метод і структура пошуку максимальних і мінімальних чисел у масивах, який за рахунок паралельного опрацювання зрізу з групи розрядів всіх чисел забезпечує зменшення часу обчислення, який в основному залежить від розрядності чисел.
- розроблений паралельний вертикально-груповий метод обчислення суми квадратів різниць, який за рахунок розпаралелення та вибору кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання.
- розроблений паралельний вертикально-груповий метод обчислення скалярного добутку, який порівняно з відомими за рахунок вибору розрядності трактів оброблення та кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання.

Практична значущість результатів дослідження – розроблені паралельні вертикально-групові методи пошуку максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць і скалярного добутку дають змогу реалізовувати компоненти апаратних нейронних мереж реального часу з високою ефективністю використання обладнання.

Матеріали та методи дослідження. У роботі використано: методи розпаралелення – для розроблення нейромережевих алгоритмів, орієнтованих на апаратну реалізацію; теорію проектування комп'ютерних систем реального часу – для розроблення компонентів апаратних нейронних мереж узгодженого паралельно-вертикального оброблення даних у реальному часі.

Аналіз останніх досліджень та публікацій. Аналіз засобів реалізації нейронних мереж показує [1], [3], [4], [6], [13], що для опрацювання інтенсивних потоків даних у реальному часі доцільно використовувати апаратну реалізацію нейронних мереж з використанням конвеєризації та просторового паралелізму.

Нейронні мережі синтезуються на базі апаратних компонент, які реалізують найскладніші операції та орієнтовані на узгодження інтенсивності обчислень з інтенсивністю вхідних даних. У роботах [2], [5], [14], [15] розглянуті методи обчислень та структури апаратних компонентів нейронних мереж. Недоліком розглянутих методів і структур є невисока швидкодія, яка у значні мірі визначають розрядністю операндів.

Характеристики нейронних мереж в значні мірі залежать від варіантів апаратної реалізації операцій пошуку максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць і скалярного добутку. Проведений аналіз структур для реалізації операцій пошуку максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць і скалярного добутку [7], [8], [9], [10], [11], [12], [14] показав, що для апаратної реалізації використовуються два типи структур: рекурсивні та нерекурсивні. Структурною особливістю рекурсивних пристроїв є присутність обернених зв'язків. У таких пристроях обчислення операцій здійснюється за декілька ітерацій, кількість яких в основному залежить від розрядності операндів. Недоліком рекурсивних пристроїв є відносно невисока швидкодія. У нерекурсивних пристроях відсутні обернені зв'язки, вони мають більшу швидкодію та для реалізації вимагають великих апаратних витрат. Недоліком нерекурсивних пристроїв, які реалізують операції пошуку максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць і скалярного добутку є велика кількість виводів інтерфейсу.

Результати дослідження та їх обговорення

1. ***Виділення операційного базису для реалізації апаратних нейронних мереж.*** На підставі аналізу алгоритмів роботи нейронних мереж виділений операційний базис, який наведений на рис. 1.

Операційний базис нейронних мереж складається із трьох груп операцій:

- перша – операції попереднього оброблення даних;
- друга – процесорні операції;
- третя – функції активації.

Операції попереднього оброблення даних направлені на перетворення вхідних даних до вигляду, який забезпечить найкращі результати при їх подальшому нейромережевому опрацюванні. До операцій попереднього оброблення даних відносяться такі операції:

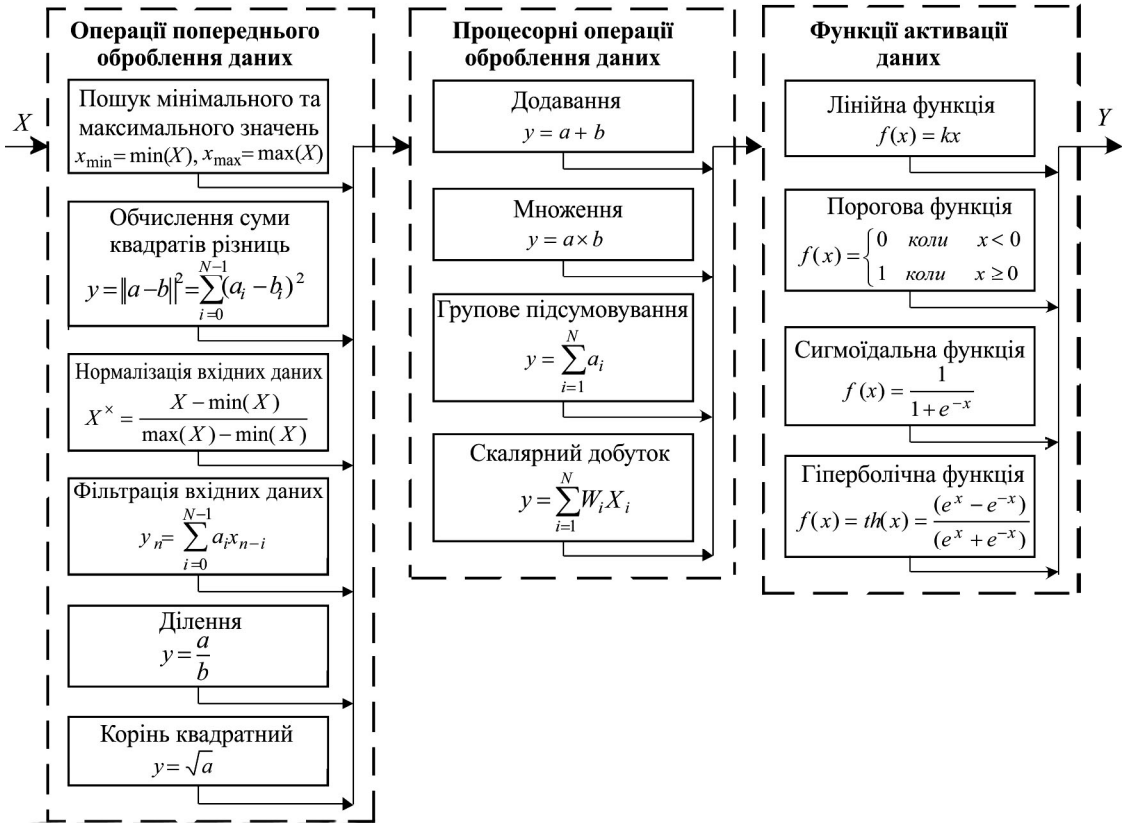

Рис. 1. Операційний базис нейронних мереж

- пошук мінімального та максимального значень;
- нормалізація вхідних даних; обчислення суми квадратів різниць;
- фільтрація вхідних даних; ділення та корінь квадратний.

Процесорні операції над вхідними даними та ваговими коефіцієнтами виконуються безпосередньо у самій нейронній мережі у процесі навчання та функціонування та зводяться до обчислення зваженої суми. При обробці даних у самій нейронній мережі можуть використовуються такі операції: додавання, множення, групове підсумовування та обчислення скалярного добутку.

Значення зваженої суми перетворюється у вихідний сигнал через алгоритмічний процес, відомий під назвою функція активації або передатна функція. У нейронних мережах можуть використовуватися різні функції активації, які вибираються залежно від задач, що розв'язуються та типу нейронної мережі. Найчастіше у нейронних мережах використовуються такі функції активації: лінійна, порогова, сигмоподібна та гіперболічна.

2. *Розроблення методу узгодження інтенсивності надходження даних з інтенсивністю обчислень у апаратних компонентах нейронних мереж.* Із аналізу операційного базису нейронних мереж (рис. 1) видно, що найбільше швидкодія апаратних нейронних мереж залежить від таких компонент: пошуку максимальних і мінімальних значень, обчислення суми квадратів різниць і скалярного добутку. Для реалізації нейроопераций у апаратних паралельним вертикально-груповим опрацюванням даних Апаратні компоненти, які орієнтовані на синтез нейронних мереж повинні забезпечувати такі вимоги:

- високу ефективність використання обладнання;
- адаптацію до вимог конкретних застосувань;
- узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень у апаратній компоненті;
- роботу у реальному часі;
- бути орієнтованими на НВІС-реалізацію;
- малий час розробки та невисоку вартість;
- зменшення кількості виводів інтерфейсу та між нейронних зв'язків.

Для оцінювання розроблених апаратних компонент нейронних мереж використовують критерій ефективності використання обладнання E_{AK}, який враховує складність алгоритму реалізації компоненти, кількість виводів зовнішнього інтерфейсу, однорідність структури компоненти та зв'язує тривалість виконання базової нейрооперації з витратами обладнання і дає оцінку елементам компоненти за продуктивністю. Кількісно величина ефективності використання обладнання для апаратних компонент нейронних мереж визначають так:

$$E_{AK}=\frac{R_{БО}}{t_{БО}(k_1 \mathrm{W_{AK}}+\mathrm{k_2 Q})}, \qquad (1)$$

де: $R_{БО}$ – складність алгоритму реалізації базової нейрооперації; $t_{БО}$ – тривалість виконання базової нейрооперації; W_{AK} – витрати обладнання на реалізацію апаратної компоненти; k_1 – коефіцієнт врахування однорідності структури, k_3 – коефіцієнт врахування кількості виводів зовнішнього інтерфейсу, Q – кількість виводів зовнішнього інтерфейсу.

Висока ефективність використання обладнання при апаратній реалізації апаратних нейронних мереж досягається шляхом узгодженням інтенсивності надходження вхідних даних і вагових коефіцієнтів P_d з інтенсивністю обчислень D_{HM}, яку забезпечує нейронна мережа. Інтенсивність надходження вхідних даних і вагових коефіцієнтів визначають так:

$$\mathrm{P_d}=(gn_g+hn_h)F_d, \qquad (2)$$

де: g – кількість каналів надходження вхідних даних X_j; h – кількість каналів надходження вагових коефіцієнтів W_j; n_g – розрядність каналів надходження вхідних даних X_j; n_h – розрядність каналів надходження вагових коефіцієнтів W_j; F_d – частота надходження вхідних даних і вагових коефіцієнтів. Інтенсивність обчислень нейронної мережі визначають так:

$$D_{HM} = \frac{vn_v}{T_к}, \quad (3)$$

де: v – кількість трактів оброблення; n_v – розрядність трактів оброблення; $T_к$ – такт роботи нейронної мережі.

Із формули (3) видно, що основними шляхами управління інтенсивністю обчислень у апаратних компонентах є:

- вибір кількості трактів опрацювання v;
- вибір розрядності трактів оброблення n_v;
- зміна тривалості такту роботи $T_к$.

Залежно від інтенсивності надходження даних P_d вибирається кількість трактів v оброблення, розрядність трактів n_v, а також тривалість такту роботи $T_к$. Для забезпечення роботи апаратних компонентів у реальному часі вибір v, n_v і $T_к$ повинно забезпечуватися виконання наступної умови:

$$P_d \le D_{AK}. \quad (4)$$

Вибір конкретної кількість трактів v, розрядності трактів n_v і тривалості такту роботи $T_к$ формально зводиться до мінімізації апаратних затрат на реалізацію апаратної компоненти при забезпеченні умови (4).

Для зменшення кількості виводів інтерфейсу та можливості вибору розрядності трактів оброблення n_v пропонується, щоб дані надходили та опрацьовувалися паралельно групами розрядних зрізів (вертикально) з використанням багатоопераційного підходу. Паралельний вертикально-груповий метод опрацювання даних передбачає, що вагові коефіцієнти W_j та вхідні дані X_j надходять паралельно зрізами із k розрядів згідно з наступними формулами:

$$W_j = \sum_{i=1}^{n} 2^{-(i-1)} w_{ji} = \sum_{g=1}^{m} 2^{-(g-1)k} \sum_{l=1}^{k} 2^{-(l-1)} w_{j[(g-1)k+l]}, j = \overline{1,N}, \quad (5)$$

$$X_j = \sum_{i=1}^{n} 2^{-(i-1)} x_{ji} = \sum_{g=1}^{m} 2^{-(g-1)k} \sum_{l=1}^{k} 2^{-(l-1)} x_{j[(g-1)k+l]}, j = \overline{1,N}, \quad (6)$$

де: w_{ji}, x_{ji} – значення i-х розрядів вагових коефіцієнтів та вхідних даних; n – розрядність вагових коефіцієнтів та вхідних даних; m – кількість груп розрядів $m = \lceil n/k \rceil$, на які розбивається вагові коефіцієнти W_j та вхідні дані X_j; k – кількість розрядів у групі.

Основним методом управління тривалістю такту роботи $T_к$ є зміна кількості сходинок конвеєра та вибір швидкодії елементної бази. Збільшення кількості сходинок конвеєра веде до зменшення тривалості такту за рахунок зменшення складності операцій, які виконуються у сходинці. При зменшенні кількості сходинок збільшується складність операцій, які виконуються у сходинці, а відповідно і збільшується тривалість такту. Апаратні витрати на реалізацію компонент зростають із збільшенням кількості сходинок конвеєра.

Інтенсивність обчислень D_{AK} у апаратній компоненті залежить від кількості трактів опрацювання v. Збільшення кількості трактів опрацювання v веде до збільшення інтенсивності обчислень D_{AK} і до збільшення апаратних витрат.

Паралельний вертикально-груповий метод пошуку максимального і мінімального значень. Паралельний вертикально-груповий метод пошуку максимального D_{max} і мінімального D_{min} значень у одновимірному $\{D_k\}_{h=1}^{N}$ масиві передбачає у кожному g-му такті ($g = 1,\ldots,m$, де $m = \lceil n/k \rceil$, k – кількість розрядів у групі, n – розрядність чисел) паралельне надходження N чисел старшими розрядами уперед зрізом із k розрядів [14]. Пошук максимального D_{max} і мінімального D_{min} чисел у одновимірному масиві $\{D_k\}_{h=1}^{N}$ за даним методом ґрунтується на виконанні для кожного r-го розрядного зрізу ($r = 1,\ldots,k$) однотипних базових макроопераций, які виконуються на базі трьох простих операцій.

Для пошуку максимального числа D_{max} у одновимірному масиві $\{D_k\}_{h=1}^{N}$ використовуються такі операції:

1) формування значення r-го розрядного зрізу P_r за формулою:

$$P_r = \bigvee_{h=1}^{N} D_{rh} \wedge y_{rh}, \quad (7)$$

де D_{rh} – значення r-го розряду h-го числа масиву, y_{rh} – значення h-го розряду r-го слова управління, значення 1-го слова управління дорівнює $y_{11} = y_{12} = \ldots = y_{1N} = 1$;

2) визначення r-го розряду максимального числа D_{maxr} за виразом:

$$D_{\max r} = \begin{cases} 0, & коли \quad P_r = 0 \\ 1, & коли \quad P_r = 1 \end{cases}; \quad (8)$$

3) формування h розрядів (r+1)-го слова управління за формулою:

$$y_{(r+1)h} = \begin{cases} 0, & коли \quad P_r = 1, \quad D_{hr} \ne y_{hr} \\ 1, & коли \quad P_r = D_{hr} = y_{hr} = 1 \\ y_{hr}, & коли \quad P_r = 0 \end{cases}. \quad (9)$$

Для пошуку мінімального числа D_{min} у одновимірному масиві $\{D_k\}_{h=1}^{N}$ використовуються такі операції:

1) формування значення r-го розрядного зрізу P_r, яке виконується за формулою:

$$P_r = \bigvee_{h=1}^{N} \overline{D}_{rh} \wedge y_{rh}; \quad (10)$$

2) визначення r-го розряду мінімального числа D_{minr} за виразом:

$$D_{\min r} = \begin{cases} 0, & коли \quad P_r = 1 \\ 1, & коли \quad P_r = 0 \end{cases}; \quad (11)$$

3) формування h розрядів (r+1)-го слова управління, яке здійснюється за виразом:

$$y_{(r+1)h} = \begin{cases} 0, & коли \quad P_r = 1, \quad \overline{D}_{khr} \ne y_{rh}; \\ 1, & коли \quad P_r = \overline{D}_{hr} = y_{hr} = 1; \\ y_{hr}, & коли \quad P_r = 0, \end{cases}. \quad (12)$$

де: $\overline{D}_{rh}$ – інверсне значення r-го розряду h-го числа масиву, y_{rh} – значення h-го розряду r-го слова управління, значення 1-го слова управління становить $y_{11} = y_{12} = \ldots = y_{1N} = 1$.

Особливість розглянутого паралельного вертикально-групового методу пошуку максимального (мінімального) значення числа є те, що у кожному *g*-у такті роботи визначаються *k* розрядів максимального D_{max} (мінімального D_{min}) значення числа.

Особливістю паралельного вертикально-групового пошуку максимальних і мінімальних значень чисел у масиві є:

- використання однієї базової макрооперації;
- можливість використання розпаралелення та конвеєризації обчислень;
- можливість одночасного опрацювання *N* розрядних зрізів;
- тривалість обчислення в основному визначають як кількістю розрядів у групі *k*, так і розрядністю чисел *n*, а не їх кількістю *N*.

Структура апаратної компоненти пошуку максимальних і мінімальних значень. Апаратна компонента пошуку максимальних і мінімальних значень реалізується на базі однотипних процесорних елементів (ПЕ). Кожний ПЕ апаратно реалізує *k* базових макроопрерацій обчислення максимального і мінімального чисел.

Структура апаратної компоненти пошуку максимальних і мінімальних чисел у одновимірному масиві $\{D_k\}_{h=1}^{N}$ за паралельним вертикально-груповим методом наведена на рис. 2, де: *ТІ* – тактові імпульси, *ПУ* – початкова установка, *Тг* – тригер, *Рг* – регістр, $D_{h1} - D_{hk}$ – *h*-й вхід групи із *k* розрядів, $D_{h1min} - D_{hkmin}$ та $D_{h1max} - D_{hkmax} D_{imin}$ – вихід груп із *k* розрядів відповідно максимального та мінімального значент чисел.

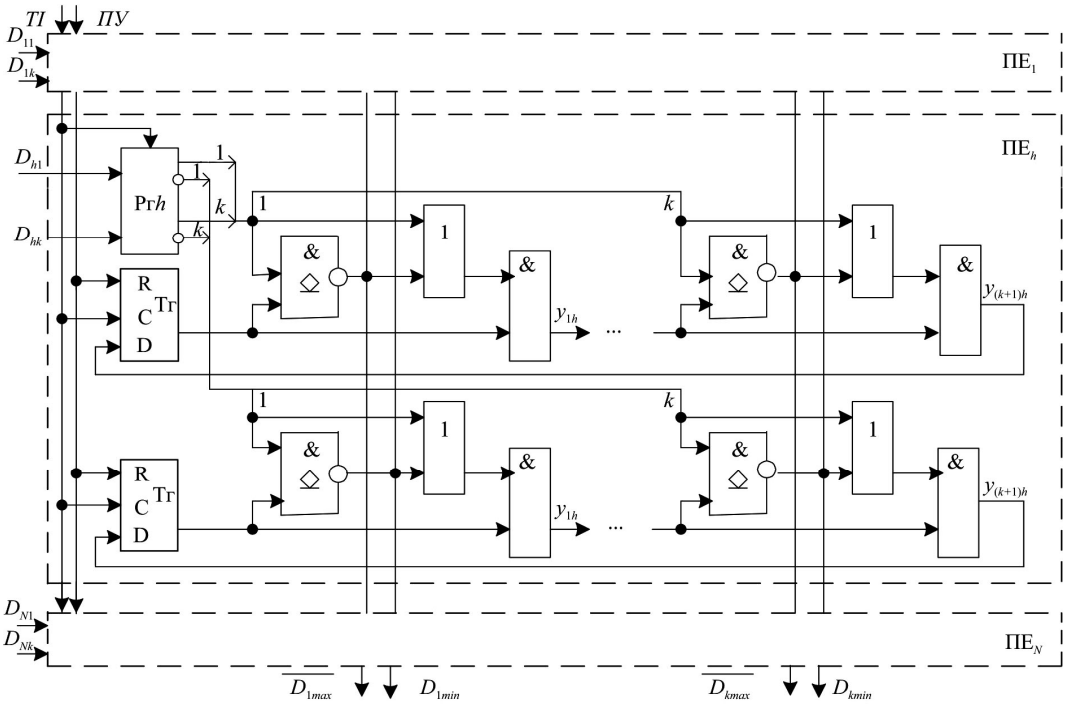

Рис. 2. Структура апаратної компоненти пошуку максимальних і мінімальних значень

Кількість ПЕ, що підключаються до спільної шини результатів, при одночасному пошуку максимального та мінімального значень чисел для одновимірного масиву $\{D_k\}_{k=1}^{N}$ визначають його розміром. Використання спільних шин результатів забезпечує розпаралелення процесу опрацювання розрядного зрізу, тривалість опрацювання якого визначає такт роботи пристрою. Пошук максимальних і мінімальних чисел за паралельним вертикально-груповим методом у такому пристрої здійснюють за час, який визначають за таким виразом:

$$t_{\Pi} = m\left(t_{Tz} + 3kt_I\right), \qquad (13)$$

де t_{Tz} та t_I час спрацювання відповідно тригера та логічних елементів типу *АБО*, *І*, *І-НЕ*, *k* – кількість розрядів у групі.

Паралельний вертикально-груповий метод обчислення суми квадратів різниць. Паралельний вертикально-груповий метод обчислення суми квадратів різниць вимагає, щоб кожний операнд був представлений у вигляді груп із *k* розрядів. При такому представлені операнди записуються так:

$$X_j = \sum_{i=1}^{n} 2^{-(i-1)} x_{ji} = \sum_{g=1}^{m} 2^{-(g-1)k} \sum_{l=1}^{k} 2^{-(l-1)} x_{j[(g-1)k+l]}, j = \overline{1,N}, \quad (14)$$

де: x_{ji} – значення *i*-го розряду *j*-го операнда; *n* – розрядність операнду, $m = \lceil n/k \rceil$ – кількість груп, на які розбивається операнд.

Піднесення числа до квадрату є основною операцією обчислення суми квадратів різниць. Для виконання такої операції, зазвичай, використовують такий вертикальний алгоритм:

$$X^2 = (0.01) \wedge x_1 + 2^{-1}(0.x_1 01) \wedge x_2 + 2^{-2}(0.x_1 x_2 01) \wedge x_3 + \ldots + 2^{-(n-1)}(0.x_1 x_2 \ldots x_{n-1} 01) \wedge x_n = \sum_{i=1}^{n} 2^{-(i-1)} Q_i, \qquad (15)$$

де Q_i – частковий результат піднесення числа до квадрату, який визначають так:

$$Q_i = (0.x_1 x_2 \ldots x_{i-1} 01) \wedge x_i . \qquad (16)$$

Розвитком розглянутого алгоритму є формування для групи із *k* розрядів Q_{Ig} групового часткового результату піднесення числа до квадрату:

$$Q_{Mg} = Q_{g1} + 2^{-1} Q_{g2} + \ldots + 2^{-(k-1)} Q_{gk} = \sum_{r=1}^{k} 2^{-(r-1)} Q_{gr}, \qquad (17)$$

де Q_{gr} – частковий результат піднесення числа до квадрату.

Алгоритм піднесення числа до квадрату з використанням формування групових часткових результатів Q_{Ig} записується так:

$$X^2 = \sum_{g=1}^{m} 2^{-(g-1)k} Q_{Ig} . \quad (18)$$

Обчислення N сум квадратів різниць будемо здійснювати на підставі багатооперандного підходу, який полягає у одночасному опрацюванню всіх операндів і формуванні для них групових часткових результатів суми квадратів різниць. Обчислення N сум квадратів різниць будемо здійснювати за паралельним вертикально-груповим методом, який записується так:

$$Y = \sum_{j=1}^{N} (X_j^e - X_j^b)^2 = \sum_{j=1}^{N} \Delta X_j^2 = \sum_{j=1}^{N} 2^{-(g-1)k} Q_{j,Ig} = \\ = \sum_{j=1}^{N} \sum_{g=1}^{m} 2^{(g-1)k} Q_{j,Ig} \sum_{g=1}^{m} 2^{(g-1)k} \sum_{j=1}^{N} Q_{j,Ig} = \sum_{g=1}^{m} 2^{(g-1)k} Q_{Mg}, \quad (19)$$

де Q_{Mg} – g-й макрочастковий результат суми квадратів різниць.

Основними етапами паралельного вертикально-групового методу обчислення суми квадратів різниць є:

- одночасне послідовно-групове надходження операндів X_j^e, X_j^b і обчислення N модулів різниці ΔX_j;
- формування для кожного j-го модуля ΔX_j у g-у такті k часткових результатів піднесення до квадрату $Q_{n-(kg-1)},\ldots,Q_{n-k(g-1)}$;
- підсумовування $N \times k$ часткових результатів піднесення до квадрату;
- формування макрочасткового результату суми квадратів різниць Q_{Mg} шляхом підсумовуванням $N(k)$ часткових результатів піднесення до квадрату;
- отримання результату суми квадратів різниць шляхом підсумовування макрочасткових результатів піднесення до квадрату Q_{Mg} із зсувом вправо на k розрядів.

Структура апаратної компоненти паралельного вертикально-групового обчислення суми квадратів різниць. Залежно від способу формування та підсумовування макрочасткових результатів суми квадратів різниць Q_{Mg} можливі такі варіанти реалізації компоненти обчислення суми квадратів різниць:

- з послідовним формуванням і підсумовуванням макрочасткових результатів піднесення до квадрату Q_{Mg};
- з паралельним формуванням і послідовним підсумовуванням макрочасткових результатів піднесення до квадрату Q_{Mg};
- з паралельним формуванням і підсумовуванням макрочасткових результатів піднесення до квадрату Q_{Mg}.

Структура апаратної компоненти обчисленням суми квадратів різниць з паралельним формуванням і послідовним підсумовуванням макрочасткових результатів піднесення до квадрату Q_{Mg} наведена на рис. 3, де Рг – регістр, КБСм – конвеєрний багатовходовий суматор, См – суматор, ПЕ – процесорний елемент.

Основними компонентами даної структури є N процесорних елементів $ПЕ_j$ та $N \times k$ -вхідний конвеєрний суматор КБСм. Підсумовування $N \times k$ часткових результатів піднесення до квадрату будемо виконувати за каскадним алгоритмом. Кількість сходинок, яка необхідна для реалізації такого підсумовування обчислюється за такою формулою:

$$s = \lceil \log_2(N \times k) \rceil . \quad (20)$$

У кожній сходинці операнди розбивають на пари, для кожної пари обчислюється сума. Для НВІС-реалізацій можуть використовуватися модифіковані каскадні алгоритми підсумовування без розповсюдження переносу.

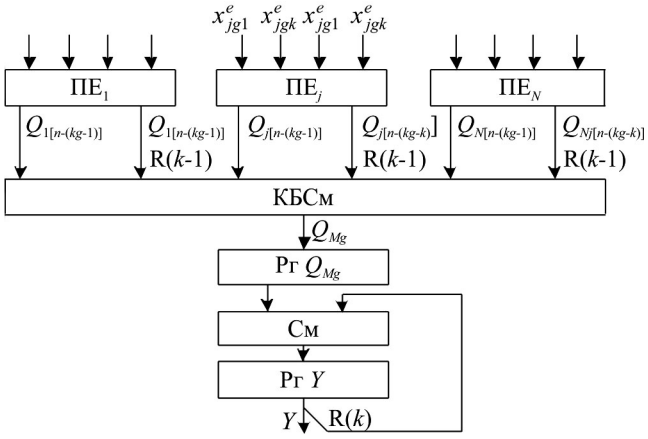

Рис. 3. Структура апаратної компоненти обчисленням суми квадратів різниць

Для обчислення модуля різниці ΔX_j та формування k часткових результатів піднесення до квадрату $Q_{j[n-(kg-r)]}$ розроблений $ПЕ_j$ структура, якого наведена на рис. 4, де *Від* – віднімач, *Тг* – тригер, *Рг* – регістр, $|\Delta X_j|$ – модуль різниці.

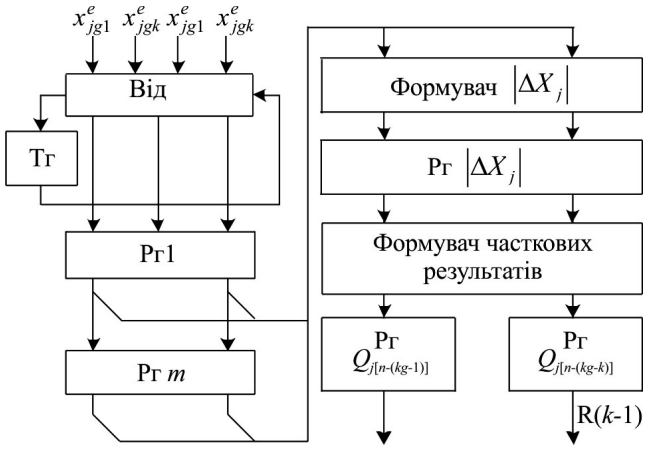

Рис. 4. Структура процесорного елемента (П*е*)

Розроблена структура $ПЕ_j$ орієнтована на суміщення у часі процесів обчислення модуля різниці $|\Delta X_j|$ для одного вхідного масиву та формування k часткових результатів піднесення до квадрату $Q_{j[n-(kg-r)]}$ для іншого вхідного масиву.

Операнди X_j^e і X_j^b на поступають вхід $ПЕ_j$ послідовно групами з k розрядів починаючи з молодших розрядів. В кожному $ПЕ_j$ за допомогою віднімача *Від* за m тактів обчислюється різниця ΔX_j, яка записується в регістри $Рг_1,\ldots, Рг_m$. Обчислена різниця ΔX_i надходить на входи формувача $|\Delta X_j|$, на виході якого отримуємо її модуль $|\Delta X_j|$. У наступних тактах роботи на виходах формувачів часткових результатів формуємо k частко-

вих результатів піднесення до квадрату $Q_{j[n-(kg-r)]}$. Формування часткових результатів піднесення до квадрату $Q_{j[n-(kg-r)]}$ здійснюється починаючи з старших розрядів модуля $|\Delta X_j|$. Сформовані на виході ПЕ$_j$ k часткових результатів піднесення до квадрату $Q_{j[n-(kg-r)]}$ надходять із зсувом вправо на (r–1) розряди входи конвеєрного $N \times k$ -вхідного суматора КБСм. Отримана на виході КБСм сума є макрочастковим результатом піднесення до квадрату Q_{Mg}, який записується у регістр РгQ_M. На суматорі См у кожному такті виконується додавання даних з виходу регістр РгQ_{Mg} до раніше накопиченої суми з регістра РгY зсунутої на k розрядів вправо згідно з наступною формулою:

$$Y_g = 2^{-k} Y_{g-1} + P_{Mg}, \quad (21)$$

де $Y_0 = 0$. Тривалість обчислення суми квадратів різниць визначають за наступною формулою:

$$t_{\text{СКР}} = m(t_{\text{Рг}} + t_{\text{См}}), \quad (22)$$

де: m – кількість груп, $t_{\text{Рг}}$ – тривалість звертання до регістра, $t_{\text{См}}$ – тривалість додавання.

Паралельний вертикально-груповий метод обчислення скалярного добутку. Паралельний вертикально-груповий метод обчислення скалярного добутку реалізується на базі елементарних арифметичних операцій, орієнтований на НВІС-реалізацію та забезпечує зменшення кількості тактів роботи, а відповідно часу обчислення. Обчислення скалярного добутку за даним методом зводиться до формування та підсумовування часткових добутків згідно з наступною формулою:

$$Z = \sum_{j=1}^{N} W_j X_j = \sum_{j=1}^{N} \sum_{g=1}^{m} 2^{-(g-1)k} (W_j X_{j[(g-1)k+1} + \\ + 2^{r-1} W_j X_{j[(g-1)k+r} + \ldots + 2^{-(k-1)} W_j X_{j[(g-1)k+k}), \quad (23)$$

де $r = 1, \ldots, k$.

Зробивши необхідні зміни у формулі (23) обчислення скалярного добутку можна записати так:

$$Z = \sum_{g=1}^{m} 2^{-(g-1)k} \sum_{j=1}^{N} (W_j X_{j[(g-1)k+1} + 2^{r-1} W_j X_{j[(g-1)k+r} + \ldots \\ + 2^{-(k-1)} W_j X_{j[(g-1)k+k}) = \sum_{g=1}^{m} 2^{-(g-1)k} P_{gM}, \quad (24)$$

де P_{gM} – g-й макрочастковий результат обчислення скалярного добутку.

З формули (24) випливає, що для обчислення скалярного добутку виконується за m тактів, у кожному g-у такті виконуються такі операції:

- формування для кожної j-ї пари операндів k часткових добутків згідно з формулою $\text{P}_{j[(g-1)k+r]} = W_j X_{j[(g-1)k+r]}$;
- обчислення g-го макрочасткового результату обчислення скалярного добутку P_{gM} шляхом підсумовуванням $N \times k$ часткових добутків згідно з формулою

$$P_{gM} = \sum_{j=1}^{N} \sum_{l=1}^{j} 2^{-(l-1)} W_j X_{j[(g-1)k+l}, g = \overline{1,G};$$

- додавання g-го макрочасткового результату обчислення скалярного добутку P_{gM} до результату підсумовування, який зсунутий вправо на k розрядів, згідно з виразом $Z_g = 2^{-k} Z_{g-1} + \text{P}_{gM}$, де $Z_0 = 0$.

Структура апаратної компоненти обчислення скалярного добутку. Структура апаратної компоненти, яка реалізує паралельний вертикально-груповий метод обчислення скалярного добутку залежить від наступного:

- використання окремих або мультиплексованих шин для введення вхідних даних X_j та вагових коефіцієнтів W_j;
- паралельний вертикально-груповий метод обчислення скалярного добутку послідовного або паралельного формування g-го макрочасткового результату обчислення скалярного добутку P_{gM};
- розділення або суміщення процесів надходження операндів одного масиву та обчислення скалярного добутку для операндів другого масиву.
- Для НВІС-реалізації паралельно вертикально-групового методу обчислення скалярного добутку вибираємо структуру, яка забезпечує:
- використання для введення вхідних даних X_j та вагових коефіцієнтів W_j – $2N$ каналів розрядністю k;
- введення вхідних даних X_j та вагових коефіцієнтів W_j групами з k розрядів починаючи з молодших розрядів;
- використання N трактів для опрацювання даних;
- паралельне формування у кожному g-у такті $N \times k$ часткових добутків;
- обчислення g-го макрочасткового результату скалярного добутку P_{gM} шляхом паралельно-конвеєрного підсумовування $N \times k$ часткових добутків;
- послідовне підсумовування макрочасткових результатів скалярного добутку P_{gM};
- суміщення у часі процесів надходження вагових коефіцієнтів W_j та обчислення скалярного добутку.

Структура апаратної компоненти, яка реалізує паралельний вертикально-груповий метод обчислення скалярного добутку за наведена на рис. 5, де ПЕ – процесорний елемент, Рг – регістр, КБСМ – конвеєрний багатовходовий суматор, См – суматор.

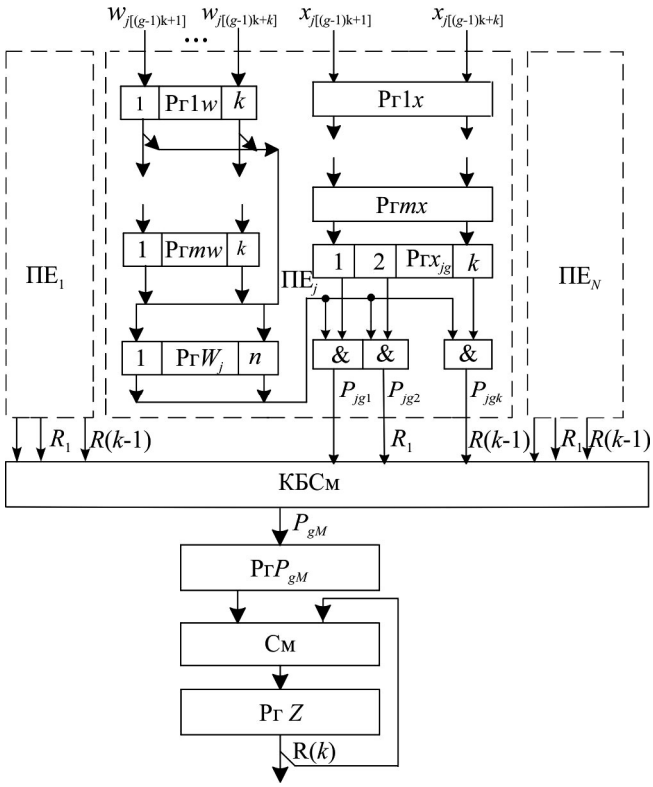

Рис. 5. Структура апаратної компоненти обчислення скалярного добутку

Основним елементом даної структури є $ПЕ_j$, на виході якого формуються k часткових добутків, де кожний r-й частковий добуток $P_{j[(g-1)k+r]}$ зсунутий вправо на (r–1) розряд. Паралельне вертикально-групове обчислення скалярного добутку у даній компоненті розбивається на два етапи, кожний з яких виконується за m тактів.

На першому етапі в кожному g-у такті на вхід $ПЕ_j$ надходять k розрядів вхідних даних X_j та k розрядів вагових коефіцієнтів W_j. Надходження груп вхідних даних X_j та вагових коефіцієнтів W_j починається їз молодших розрядів. Перший етап завершується на (m–1) записом вагового коефіцієнта W_j у регістр $Рг W_j$ та k молодших розрядів у регістр $Рг x_{jg}$.

На другому етапі у кожному g у такті роботи в $ПЕ_j$ для групи розрядів вхідних даних $X_{j[(g-1)k+1]} \dots X_{j[(g-1)k+r]} \dots X_{j[(g-1)k+k]}$ формується k часткових добутків згідно з виразом $P_{j[(g-1)k+r]} = W_j X_{j[(g-1)k+r]}$. Сформовані у $ПЕ_j$ часткові добутки надходять на вхід конвеєрного багатовходового суматора КБСм, причому r-ий (r=1,…, k) частковий добуток $P_{j[(g-1)k+r]}$ зсунутий відносно (r–1)-го часткового добутку $P_{j[(g-1)k+r-1]}$ на один розряд вправо. Шляхом додавання $N \times k$ часткових добутків на виході КБСм отримуємо макрочастковий результат скалярного добутку P_{gM}, який записується у регістр $Рг P_{gM}$. На суматорі См виконується додавання макрочасткового результату скалярного добутку P_{gM} до раніше обчисленої суми зсунутої вправо на k розрядів, згідно з виразом $Z_g = 2^{-k} Z_{g-1} + P_{gM}$, де $Z_0 = 0$.

Розроблена компонента для обчислення скалярного добутку працює за конвеєрним принципом і орієнтована на опрацювання неперервних потоків даних. Конвеєрний такт роботи такого пристрою визначають так:

$$t_{СД} = t_{Рг} + t_{См}, \qquad (25)$$

де: $t_{Рг}$ – час спрацювання регістра, $t_{См}$ – тривалість додавання двох чисел. Обчислення скалярного добутку здійснюється за m конвеєрних тактів.

Обговорення отриманих результатів дослідження. Розроблені паралельні вертикально-групові методи пошуку максимального і мінімального значень, обчисленням суми квадратів різниць і скалярного добутку орієнтовані на опрацювання потоків даних у реальному часі. Особливістю даних методів є можливість здійснювати пошук і обчислення з різною кількістю розрядів у групі. Для розроблених паралельних вертикально-групових методів максимальне значення кількості розрядів у групі може бути $k=n/2$. У випадку коли $k=n$ паралельні вертикально-групові методи перетворюються у паралельно-паралельні, які забезпечать максимальну швидкодію.

При апаратній реалізації компонент пошуку максимального і мінімального значень, обчисленням суми квадратів різниць і скалярного добутку основними шляхами управління інтенсивністю обчислень у таких компонентах є: вибір кількості та розрядності каналів надходження даних, кількості трактів опрацювання даних, зміна тривалості такту роботи шляхом вибору елементної бази та складності операцій, які реалізуються сходинками конвеєра. Особливістю апаратної реалізації для груп з малою кількістю розрядів є малі апаратні витрати та невелика кількість виводів інтерфейсу. Збільшення кількості розрядів у групі веде до збільшення апаратних затрат і кількості виводів. При $k=n$ апаратні компоненти будуть мати матричну структуру, яка характеризується великими апаратними затратами та великою кількістю виводів.

Забезпечити обчисленням суми квадратів різниць і скалярного добутку у реальному часі при інтенсивному надходженні даних можна досягнути шляхом розпаралелення оброблення даних до бітового рівня.

Висновки

1. Сформовано операційний базис нейронних мереж і вибрано для апаратної реалізації такі операції: пошуку максимального і мінімального значень, обчисленням суми квадратів різниць і скалярного добутку.

2. Визначено вимоги до апаратних компонентів нейронних мереж з узгодженою вертикально-паралельним обробленням даних, основними з яких є забезпечення: високої ефективності використання обладнання, адаптації до вимог конкретних застосувань, узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень у апаратній компоненті, роботу у реальному часі, структурної орієнтації на НВІС-реалізацію, малого часу розробки та невисокої вартості.

3. Показано, що основними шляхами управління інтенсивністю обчислень у апаратних компонентах є вибір кількості та розрядності трактів опрацювання даних, зміна тривалості такту роботи шляхом вибору елементної бази та складності операцій, які реалізуються сходинками конвеєра.

4. Запропоновано для реалізації апаратних компонент нейронних мереж з узгодженою паралельно-вертикальним обробленням даних використовувати паралельні вертикально-групові методи опрацювання даних, які забезпечують управління інтенсивністю обчислень, зменшення апаратних затрат і НВІС реалізацію.

5. Розроблено паралельний вертикально-груповий метод і структуру обчислення максимальних і мінімальних чисел у масивах, яка за рахунок паралельного опрацювання зрізу з групи розрядів всіх чисел забезпечує зменшення часу обчислення, який в основному залежить від розрядності чисел.

6. Розроблено паралельний вертикально-груповий метод обчислення суми квадратів різниць, який за рахунок розпаралелення та вибору кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання.

7. Розроблено паралельний вертикально-груповий метод обчислення скалярного добутку, який порівняно з відомими за рахунок вибору розрядності трактів оброблення та кількості сходинок конвеєра забезпечує узгодження інтенсивності надходження вхідних даних з інтенсивністю обчислень, режим реального часу та високу ефективність використання обладнання.

References

[1] Haikin, S. (2016). *Neural networks: full course*, (2nd ed. add. and revised). (Trans. from English). Moscow: Williams, 1104 p.

[2] Moiseychenko, V. S. (2017). Hardware implementation of artificial neural networks. Part 1. *Young scientist, 12*(146), 69–72.
[3] Palagin, A. V., Boyun, V. P., & Yakovlev, Yu. S. (2017). Problems of creating computer systems using a nanoelement base. *Control systems and machines, 5*, 3–15. https://doi.org/10.15407/usim.2017.05.003
[4] Peleshchak, Roman, Lytvyn, Vasyl, Peleshchak, Ivan, & Vysotska, Victoria. (2020). Development of an artificial neural network with oscillatory neurons for spectral pattern recognition. *Bulletin of the National University "Lviv Polytechnic" "Information Systems and Networks", 7*, 16–23. https://doi.org/10.23939/sisn2020.07.016
[5] Petrushenko, A. M. (2020). The principle of firmware control and automation of design of operating devices. II. *Control Systems and Computers, 2*, 3–11. https://doi.org/10.15407/csc.2020.02.003
[6] Rashkevich, Yu. M., Tkachenko, R. O., Dragon, I. G, & Peleshko, D. D. (2014). *Neuro-like methods, algorithms and structures of signal and image processing in real time: monograph.* Lviv: Lviv Polytechnic Publishing House, 256.
[7] Tsmots, I. G, Skorokhoda, O. V., & Medikovsky, M. O. (2019). *Device for calculating the scalar product*. Patent of Ukraine for the invention № 118596, 11.02.2019, Bull. № 3.
[8] Tsmots, I. G, Teslyuk, V. M, Teslyuk, T. V, Medikovsky, M. O., & Tsymbal, Y. V. (2019). *Device for calculating the sums of paired products*. Patent of Ukraine № 120210, 25.10.2019, blvd. № 20/2019.
[9] Tsmots, I. H., Lukashchuk, Yu. A., Khavalko, V. M., & Rabyk, V. H. (2019). Models of neural elements of parallel-parallel type. *Modeling and Information Technologies, 86*, 119–126.
[10] Tsmots, I., Rabyk, V., Skorokhoda, O., & Teslyuk, T. (2019). *Neural element of parallel-stream type with preliminary formation of group partial products*. Electronics and information technologies (ELIT-2019): proceedings of the XIth International scientific and practical conference, 16–18 September, Lviv, Ukraine, 154–158. https://doi.org/10.1109/ELIT.2019.8892334
[11] Tsmots, Ivan, Skorokhoda, Oleksa, Ignatyev, Ihor, & Rabyk, Vasyl. (2017). *Basic Vertical-Parallel Real Time Neural Network Components*. Proceedings of XIIth International Scientific and Technical Conference CSIT 2017, 5–8 September 2017. Lviv, Ukraine, 344–347. https://doi.org/10.1109/STC-CSIT.2017.8098801
[12] Tsmots, Ivan, Teslyuk, Vasyl, Teslyuk, Taras, & Ihnatyev, Ihor. (2018). *Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing*. Advances in Intelligent Systems and Computing II, Advances in Intelligent Systems and Computing, 689. Springer International Publishing AG, 558–576. https://doi.org/10.1007/978-3-319-70581-1_39
[13] Yakovlev, Yu. S. (2016). About an estimation of efficiency of application of FPGA as a part of PIM-systems. *Control systems and machines, 1*, 56–61. https://doi.org/10.15407/usim.2016.01.056
[14] Zhang, C., Li, P., Sun, G., Guan, Y., Xiao, B., & Cong, Zhang, J. (2015). *Optimizing FPGA-based accelerator design for deep convolutional neural networks*. Proceedings of the 2015 ACM/SIGDA International Symposium on FieldProgrammable Gate Arrays. ACM, 161–170. https://doi.org/10.1145/2684746.2689060
[15] Zoev, Y. V., Beresnev, A. P., Markov, N. H., & Malchukov, A. N. (2017). FPGA-based device for handwriting digit recognition in images. *Computer Optics, 41*(6), 938–949. https://doi.org/10.18287/2412-6179-2017-41-6-938-949

I. H. Tsmots[1], Yu. A. Lukashchuk[1], I. V. Ihnatyev[2], I. Ya. Kazymyra[1]

[1] *Lviv Polytechnic National University, Lviv, Ukraine*
[2] *West Ukrainian National University, Ternopil, Ukraine*

COMPONENTS OF HARDWARE NEURAL NETWORKS FOR COORDINATED PARALLEL-VERTICAL DATA PROCESSING IN REAL TIME

It is shown that for the processing of intensive data flows in industry (management of technological processes and complex objects), energy (optimization of load in power grids), military affairs (technical vision, mobile robot traffic control, cryptographic data protection), transport (traffic management and engine), medicine (disease diagnosis) and instrumentation (pattern recognition and control optimization) the real-time hardware neural networks with high efficiency of equipment use should be applied. The operational basis of neural networks is formed and the following operations are chosen for hardware implementation: the search of the maximum and minimum values, calculation of the sum of squares of differences and scalar product. Requirements for hardware components of neural networks with coordinated vertical-parallel data processing are determined, the main ones of which are: high efficiency of equipment use, adaptation to the requirements of specific applications, coordination of input data intensity with the computation intensity in hardware component, real-time operation, structural focus on VLSI implementation, low development time and low cost. It is suggested to evaluate the developed hardware components of neural networks according to the efficiency of the equipment use, taking into account the complexity of the component implementation algorithm, the number of external interface pins, the homogeneity of the component structure and relationship of the time of basic neurooperation with the equipment costs. The main ways to control the intensity of calculations in hardware components are the choice of the number and bit rates of data processing paths, changing the duration of the work cycle by choosing the speed of the element base and the complexity of operations implemented by the conveyor. The parallel vertical-group data processing methods are proposed for the implementation of hardware components of neural networks with coordinated parallel-vertical control processing, they provide control of computational intensity, reduction of hardware costs and VLSI implementation. A parallel vertical-group method and structure of the component of calculation of maximum and minimum numbers in arrays are developed, due to parallel processing of a slice from the group of digits of all numbers it provides reduction of calculation time mainly depending on bit size of numbers. The parallel vertical-group method and structure of the component for calculating the sum of squares of differences have been developed, due to parallelization and selection of the number of conveyor steps it ensures the coordination of input data intensity with the calculation intensity, real-time mode and high equipment efficiency. The parallel vertical-group method and structure of scalar product calculation components have been developed, the choice of bit processing paths and the number of conveyor steps enables the coordination of input data intensity with calculation intensity, real-time mode and high efficiency of the equipment. It is shown that the use of the developed components for the synthesis of neural networks with coordinated vertical-parallel data processing in real time will reduce the time and cost of their implementation.

Keywords: neural networks; hardware components; coordinated parallel-vertical processing; vertical group methods; real time.

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; ResearcherID: R-2780-2017

Лукащук Юрій Андрійович, аспірант, кафедра автоматизованих систем управління. **Email:** urijlukas@gmail.com; https://orcid.org/0000-0002-8933-8635

Ігнатєв Ігор Васильович, викладач, кафедра комп'ютерної інженерії. **Email:** ignatyevki@gmail.com

Казимира Ірина Ярославівна, канд. техн. наук, доцент, кафедра автоматизованих систем управління. **Email:** iryna.y.kazymyra@lpnu.ua; https://orcid.org/0000-0003-1597-5647; ResearcherID: V-5421-2017

Цитування за ДСТУ: Цмоць І. Г., Лукащук Ю. А., Ігнатєв І. В., Казимира І. Я. Компоненти апаратних нейронних мереж узгодженого паралельно-вертикального оброблення даних у реальному часі. Український журнал інформаційних технологій. 2021, т. 3, № 1. С. 63–72.

Citation APA: Tsmots, I. H., Lukashchuk, Yu. A., Ihnatyev, I. V., & Kazymyra, I. Ya. (2021). Components of hardware neural networks for coordinated parallel-vertical data processing in real time. *Ukrainian Journal of Information Technology*, *3*(1), 63–72. https://doi.org/10.23939/ujit2021.03.063