Метод просторово-часового відображення нейроалгоритмів обробки сигналів і зображень в узгоджено-паралельні НВІС-структури реального часу

Д.т.н., проф. І. Цмоць[1], асп. І. Ізонін[2], асп. В. Антонів[1]

[1]Lviv Polytechnic National University, Automated control systems department
[2]Lviv Polytechnic National University, Publishing information technology department
Bandera Street 12, 79013, Lviv, Ukraine
E-mail: omalyk@mail.lviv.ua

Abstract. The purpose of the synthesis of real-time neural algorithms of parallel vertical type is modular and getting regular structure focused on VLSI technology. When designing or selecting for VLSI implementations of neural network algorithms for the signals and images proccesing in real time simultaneously you need to consider many interrelated factors. In this article authors described developed method space-time mapping neural-algorithms which focused on VLSI structures

Keywords: neuro algorithm, VLSI-structures, layer-parallel form, flow graph, real time.

Вступ

Алгоритми нейромережевої обробки сигналів і зображень реального часу, які орієнтовані на НВІС-реалізацію повинні бути добре структурованими, орієнтованими на реалізацію на множині взаємозв'язаних процесорних елементів (ПЕ), та забезпечувати детерміноване переміщення даних. Структура та операції, які виконують ПЕ залежить від вимог, що висуваються до часу обчислення алгоритму. В більшості випадків ПЕ реалізують нейромережний операційний базис, що складається із трьох груп базових операцій: попередньої обробки, процесорних багатоопераднних операцій, елементарних функцій та арифметичних операцій Вихідною інформацією для нейромереж реального часу паралельно-вертикального типу реального часу є: алгоритми навчання та функціонування нейромережі; графове відображення нейромережі; кількість вхідних даних N; інтенсивність надходження вхідних даних і вагових коефіцієнтів; вимоги до інтерфейсу; розрядність вхідних даних, вагових коефіцієнтів, тощо.

Виклад основниго матеріалу

При синтезі нейромереж реального часу необхідно забезпечити її функціонування з мінімальними затратами в реальному часі. Перехід від графового відображення нейромережі до апаратної структури неромережі формально зводиться до мінімізації апаратних затрат:

$$W_{HM} = \sum_{j=1}^{M} W_{EIIj} + \sum_{i=1}^{N} W_{HE_i} + k_1 Y + k_2 P, \qquad (1)$$

де W_{EIIj} – витрати обладнання на реалізацію j-о елемента попередньої обробки; M – кількість елементів попередньої обробки; W_{HEi} – витрати обладнання на реалізацію i-го нейроелемента; N – кількість нейроелементів; Y – кількість виводів інтерфейсу; k_1 – коефіцієнт врахування кількості виводів інтерфейсу $k_1=f(Y)$, P – кількість міжнейронних зв'язків; k_2 – коефіцієнт врахування міжнейронних зв'язків $k_2=f(P)$, при забезпеченні наступної умови:

$$T_{обм} \geq T_p, \qquad (2)$$

де $T_{обм}$ - час обміну, T_p – реалізації алгоритмів навчання та функціонування нейромережі.

Процес синтезу нейромереж реального часу з високою ефективністю використання обладнання можна звести до виконання таких етапів:

1) вибрати нейромережу та представити її у вигляді конкретизованого узгодженого потокового графу;
2) перейти з врахуванням техніко-економічних вимог і обмежень від конкретизованого узгодженого потокового графу до структури апаратної нейромережі;
3) вибрати модель нейроелемента паралельно-вертикального типу, елементів попередньої оброки та здійснити їх синтез;
4) розробити інтерфейс та систему обміну між шарами нейромережі;
5) визначити при потоковій структурі нейромережі порядок реалізації у часі шару нейромережі та синтезувати пристрої управління.

При виборі варіанту апаратної нейромережі реального часу використовується критерій ефективності використання обладнання E [1], який обчислюється за наступною формулою:

$$E = \frac{R}{t_o(\sum_{j=1}^{M} W_{EIIj} + \sum_{i=1}^{N} W_{HE_i} + k_1 Y + k_2 P)}, \qquad (3)$$

де R – складність алгоритмів навчання та функціонування нейромереж; t_o – час роботи алгоритмів щодо навчання та роботи нейромережі. Він оцінює елементи за критерієм продуктивність/апаратні затрати і враховує кількість міжнейронних звязків, кількість виводів інтерфейсу

Для апаратної реалізації нейромержі з високою ефективністю використання обладнання необхідно граф нейромережі подати у просторово-часовому відображенні у вигляді конкретизованого узгодженого потокового графа на рівні одно-, дво- і багатоопераднних нейрооперацій [3, 4]. Потоковий граф нейромережі паралельно-вертикального типу, де $\Phi_{(s-p)}$ – функціональний оператор послідовно-паралельного перетворення; Φ_{Pji} – функціональний оператор формування часткових результатів; Φ_{PMi} – функціональний оператор формування макрочасткового результату;

Φ_Z – функціональний оператор підсумовування макрочасткових результатів; Φ_a – функціональний оператор функції активації; $\Phi_{(p\text{-}s)}$ – функціонал паралельно-послідовного перетворення; x_j – j-й вхідний сигнал, w_{ij} – ij-й ваговий коефіцієнт.Для забезпечення просторово-часового відображення нейромережі враховуючи всі форми паралелізму використовується ярусно-паралельна форма (ЯПФ) [4]. При такій формі подання нейромережі здійснюється розподіл всіх її функціональних операторів Φ_i за ярусами наступним чином: в j-му ярусі розміщені ті функціональні оператори, які залежать від функціональних операторів *(j-1)*-о ярусу але не залежать від операторів інших ярусів. Всі функціональні оператори одного ярусу виконуються незалежно один від одного.

Кількість ярусів у ЯПФ є її висотою h, а l – ширина, яка визначається максимальною шириною ярусів. Використовуючи ЯПФ для відображення орієнтованого графа нейромережі вводяться деякі позначення: j - номер яруса, який розглядаємо як часовий індекс, k - номер функціональних операторів в відповідному ярусі, який розглядаємо як просторовий індекс. На основі цих позначень відбувається розміщення функціональних операторів. Таке відображення орієнтованого графа алгоритму будемо називати потоковою паралельною формою або потоковим графом [3]. Складність функціональних операторів Φ_i, ширина l і висота h є взаємно залежними параметрами потокового графу, а отже зміна одного зумовлює зміни інших.

Відображення потокових графів нейромережі може здійснюватись з різною степеню деталізації, в залежності від засобів реалізації. При НВІС-реалізації як правило використовується відображення нейромережі на рівні одно-, дво- і багатооперандних арифметичних операцій.

Для синтезу апаратної нейромережі паралельно-вертикального типу з високою ефективністю використання обладнання використовується метод адекватного апаратного відображення структури графів алгоритмів і процесів функціонування, у якому кожному функціональному оператору ставиться у відповідність операційний блок, а дугам між функціональними операторами - каналами передачі даних [2,3]. Синтезовані таким чином апаратні нейромережі є алгоритмічними. Забезпечення високої ефективності використання обладнання досягаються шляхом зміни параметрів потокового графу (степені деталізації функціональних операторів, висоти h і ширини l графу) та орієнтацією його на використання нейроелементів паралельно-вертикального типу.

Синтез подібних нейромереж вимагає розробки конкретизованих узгоджених потокових графів, які повинні забезпечити виконання умови:

$$P_d \le D_к, \qquad (4)$$

Процес розробки конкретизованих узгоджених потокових графів для синтезу апаратних нейромереж паралельно-вертикального типу можна розбити на наступні чотири етапи:

1) декомпозиція алгоритму навчання та функціонування нейромережі;
2) проектування комунікацій (обмінів даними) між нейроелементами сусідніх шарів нейромережі;
3) укрупнення функціональних операторів у кожному шарі нейромережі;
4) планування обчислень при реалізації укрупнених функціональних операторів.

На першому етапі розробки конкретизованого узгодженого потокового графу виконується декомпозиція алгоритму Φ навчання та функціонування нейромережі. Декомпозиція передбачає розбивається алгоритму навчання та функціонування нейромережі Φ на функціональні оператори Φ_i, між якими встановлюються зв'язки, у відповідності із алгоритмом. Від результату цього кроку великою мірою залежить легкість узгодження алгоритму, тобто виконання умови (4).

Спосіб і час реалізації функціонального оператора Φ_i є визначальним при оцінці конвеєрного такту роботи $T_к$ нейромережі. Після виконання першого етапу розробки отримуємо граф нейромережі, де складність функціональних операторів Φ_i визхначають засобами реалізації.

На другому етапі розробляються засоби комунікацій для нейромережі паралельно-вертикального типу, яка працює в конвеєрному режимі. Для цього виконується перехід від графу нейромережі до потокового графу. На основі структури зв'язків у між функціональними операторами Φ_{jk} сусідніх ярусів потокового графу можна визначити кількість каналів надходження даних.

Після другого етапу розробки конкретизованого узгодженого потокового графу отримуємо потоковий граф, який забезпечує визначення обчислювальної здатності $D_к$ нейромережі.

Вихідними даними для визначення обчислювальної здатності нейромережі паралельно-вертикального типу $D_к$ є:

- кількість трактів обробки (нейроелементів) g і каналів надходження даних в цих трактах обробки s;
- складність функціональних операторів Φ_i;
- швидкодія елементної бази, на якій буде реалізовуватися нейромережа.

Оцінити узгодженість інтенсивності надходження даних P_d проводять на основі коефіцієнта узгодженості, який визначається так:

$$L = \left\lceil \frac{P_d}{D_k} \right\rceil, \qquad (5)$$

де $\lceil \ \rceil$ – знак округлення до більшого цілого, $D_к$ - обчислювальна здатність.

Коефіцієнт узгодженості L може бути L=1, L>1 та L<1. Коли L=1, то розроблений граф нейромережі є узгодженим і забезпечує перехід до структури описаної нейромережі.

Коли L<1, то розроблений граф нейромережі не є узгодженим і для його узгодження необхідно збільшувати обчислювальну здатність $D_к$. Підвищення обчислювальної здатності $D_к$ може бути досягнуте шляхом збільшенням кількості трактів обробки (нейроелементів) g і каналів надходження даних в трактах обробки s або зменшенням складності функціональних операторів Φ_i. Інколи, підвищення обчислювальної здатності досягається паралельним використанням декількох нейромереж.

У випадку коли $L<1$, то розроблений граф нейромережі не є узгодженим і для його узгодження необхідно зменшити обчислювальну здатність $D_к$. Зменшення обчислювальної здатності $D_к$ можна досягнути шляхом об'єднання функціональних операторів $Ф_{ji}$, як у межах ярусу так і між ярусами.

На третьому етапі виконується укрупнення операцій за рахунок об'єднання функціональних операторів $Ф_{jk}$ і каналів передачі даних. Результатом такої операції буде граф нейромережі, який будемо називати конкретизованим потоковим графом нейромережі. Для кожного j-го ярусу коефіцієнт об'єднання V_j визначається кількістю об'єднаних операцій і каналів передачі даних. Коефіцієнт об'єднання у кожному ярусі V_j повинен забезпечувати максимальне завантаження обладнання шляхом узгодження обчислювальної здатності $D_{кj}$ у кожному ярусі графа. Величина коефіцієнта об'єднання у кожному ярусі V_j визначається так:

$$V_j \le \left\lfloor \frac{D_{kj}}{P_{dj}} \right\rfloor, \qquad (6)$$

де $\lfloor\ \rfloor$ – знак округлення до меншого цілого.

Для випадку коли коефіцієнт об'єднання V_j є рівний або більший ширини ярусу l_j укрупнення роблять на основі лінійної проекції на вісь передачі даних, при якій всі функціональні оператори j-го ярусу потокового графу подаються у вигляді функціонального макрооператора $Ф_{MO}$, а канали передачі даних подаються у вигляді оператора затримки даних чи їх перестановки $Ф_{ЗП}$.

На четвертому етапі після об'єднання функціональних операторів потокового графа нейромережі здійснюється планування обчислень, визначаються величин затримок і перестановок. Для відтворення послідовності обчислень у кожний ярус конкретизованого графа нейромережі вводяться оператори управління, затримки та перестановки даних.

Лінійна проекція конкретизованого потокового графу нейромережі на вісь, паралельну передачі даних наведена на рис. 1, де $Ф_{MO}$ – функціональний макрооператор, $Ф_{ЗП}$ – операторів затримки та перестановки даних, $Ф_y$ – оператор управління.

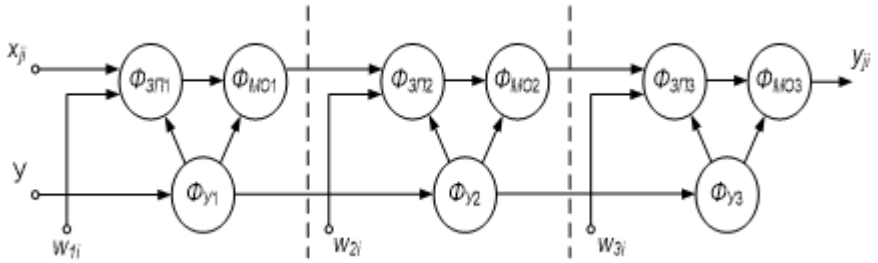

Рис.1 Потоковий граф нейромережі паралельно-вертикального типу.

Наведений на рис. 1 конкретизований потоковий граф нейромережі включає три частини – функціональну, структурну та управляючу. Функціональна - визначає процесорні елементи, які реалізують функціональні макрооператори, структурна – зв'язки між функціональними макрооператорами, величин затримок і перестановки даних, а управляюча – відтворює послідовність обчислень у кожному ярусі відповідно до структури графа.

Висновок

Враховуючи те, що висока продуктивність та ефективність використання обладнання досягається тільки в тих випадках, коли їх архітектура адаптується до інтенсивності надходження потоків даних і адекватно відображає структуру алгоритму розв'язання задачі, авторами запропоновано новий розв'язок актуальної на сьогоднішній день задачі у паралельно-конвеєрних комп'ютерних системах ЦОС. У статті розроблено метод просторово часового відображення нейроалгоритмів в узгоджено-паралельні НВІС-структури реального часу, що дає можливість розробляти ефективні високопродуктивні спеціалізовані нейро- методи, алгоритми та структури обробки сигналів і зображень у реальному часі з високою ефективністю використання обладнання, що орієнтовані на НВІС-технології. Також, даний метод зумовлює зменшення часу і вартості розробки орієнтованих на НВІС-реалізації паралельних операційних пристроїв для виконання базових операцій алгоритмів ЦОС, і забезпечує можливість будувати ефективні спеціалізовані засоби реального часу для паралельного та паралельно-потокового сортування інтенсивних потоків даних.

[1] Tsmots I. Parallel algorithms and VLSI structures for median filtering of images in real time / I. Tsmots, D. Peleshko, I. Izonin // International Journal of Advanced Research in Computer Engineering & Technology, Volume 3 Issue 8, 2014 – 2643-2649 p.

[2] Параллельная обработка информации: в 5-ти т. – Т.4. Высокопроизводительные системы параллельной обработки информации / Л.Б.Авгуль, А.И.Белоус, А.И.Гречишников и др. / Под ред. В.В.Грицыка. – Киев: Наукова думка, 1988.- 272 с.

[3] Вишенчук И.М., Черкаский Н.В. Алгоритмические устройства и супер ЭВМ. – Киев: Техніка, 1991. – 197 с..

[4] http://www.intuit.ru/studies/courses/4447/983/lecture-/14919?page=3