ДОПОВІДІ
НАЦІОНАЛЬНОЇ
АКАДЕМІЇ НАУК
УКРАЇНИ

10 • 2014

ІНФОРМАТИКА ТА КІБЕРНЕТИКА

УДК 004.032.26

Член-кореспондент НАН України **В. В. Грицик, І. Г. Цмоць, О. В. Скорохода**

Методи паралельно-вертикального опрацювання даних у нейромережах

Визначено операційний базис нейромереж, обгрунтовано доцільність розроблення апаратних нейромереж паралельно-вертикального типу, розроблено орієнтований на НВІС-реалізацію паралельно-вертикальний метод опрацювання даних у нейроелементах (нейромережах), який забезпечує зменшення кількості виводів інтерфейсу, розрядності міжнейронних зв'язків і затрат обладнання та запропоновано принципи НВІС-реалізації нейромереж.

На сучасному етапі розвитку нейромережевих технологій відбувається розширення галузей застосування, в значній частині з яких потрібно розв'язувати задачі у реальному часі на апаратних засобах, що відповідають обмеженням щодо енергоспоживання, габаритів, часу та вартості розроблення. Створення високоефективних нейромережевих засобів реального часу потребує широкого використання сучасної елементної бази (надвеликих інтегральних схем (НВІС)), розроблення нових методів і алгоритмів опрацювання даних у реальному часі, орієнтованих на апаратну реалізацію.

Аналіз операційного базису нейромереж показує, що нейромережеві операції за кількістю операндів, що одночасно опрацьовуються, можна розділити на одно- (корінь квадратний, передатні функції), дво- (додавання, ділення, множення) і багатоопераднді (визначення мінімального та максимального чисел, багатоопераднде підсумовування, обчислення скалярного добутку, обчислення суми квадратів різниць) [1, 2]. Відомі апаратні нейроелементи та нейромережі в основному є одно- і двоопераднми, це пов'язано з можливостями елементної бази. Еволюція розвитку архітектури нейроелементів та нейромереж тісно пов'язана з структурною одиницею опрацювання, тобто з розрядністю і кількістю операндів, які одночасно опрацьовує операційний пристрій. З розвитком інтегральної технології з'явилася тенденція зміни структурної одиниці опрацювання з одно- та двоопераднодної на багатоопераднду, яка виконується паралельно.

Особливістю багатоопераднодних нейроопераций є те, що вони виконуються над множиною операндів і результатом операції є одне число. Багатоопераднодні нейрооперації пропонується

виконувати на основі багатооперандного підходу, при якому процес обчислення нейрооперації розглядається як виконання єдиної операції, що грунтується на елементарних арифметичних операціях.

Паралельна НВІС-реалізація нейроелементів і нейромереж на основі багатооперандного підходу потребує великих затрат обладнання і значної кількість виводів інтерфейсу, які залежать як від кількості операндів, так і від їхньої розрядності. Вартість і швидкодія паралельних НВІС-реалізацій нейроелементів і нейромереж істотно залежить як від рівня технології, так і від кількості виводів інтерфейсу, які визначають розмір кристала. Для забезпечення високої швидкодії, зменшення кількості виводів інтерфейсу та розрядності міжнейронних зв'язків пропонується опрацювання даних у нейромережах здійснювати паралельно розрядними зрізами на основі багатооперандного підходу, тобто паралельно-вертикальними методами [3]. На основі таких методів опрацювання даних розробляються апаратні нейроелементи та нейромережі, які мають архітектуру паралельно-вертикального типу. Тому метою дослідження є розроблення паралельно-вертикальних методів опрацювання даних у реальному часі, орієнтованих на НВІС-реалізацію.

Методи паралельно-вертикального опрацювання даних у нейромережах. Основними компонентами, на базі яких синтезуються апаратні нейромережі, є нейроелемети. При паралельно-вертикальному опрацюванні даних у нейроелементі вхідні дані X_j та вагові коефіцієнти W_j ($j = 1, \ldots, N$, де N — кількість входів даних і вагових коефіцієнтів) подаються у порозрядному вигляді згідно з формулою

$$W_j = \sum_{i=1}^{n} 2^{-i} W_{ji}, \qquad X_j = \sum_{i=1}^{n} 2^{-i} X_{ji}, \tag{1}$$

де W_{ji}, X_{ji} — значення i-х розрядів множників W_j і X_j; n — розрядність множників.

У загальному випадку нейроелемент здійснює перетворення у відповідності з формулою

$$y_p = f\left(\sum_{j=1}^{N} W_j X_j\right), \tag{2}$$

де y_p — вихідний сигнал p-го нейроелемента; f — функція активації.

Перетворення у p-му нейроелементі з використанням паралельно-вертикального опрацювання даних записується так:

$$y_p = f\left(\sum_{j=1}^{N} W_j X_j\right) = f\left(\sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{N} W_j X_{ji}\right) = f\left(\sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{N} P_{ji}\right) = f\left(\sum_{i=1}^{n} 2^{-i} P_{Mi}\right), \tag{3}$$

де P_{ji} — ji-й частковий результат; P_{Mi} — i-й макрочастковий результат, який формується додаванням N часткових результатів [4].

З формули (3) випливає, що паралельно-вертикальне опрацювання даних у нейроелементах зводиться до виконання таких етапів:

формування для кожного розрядного зрізу часткових результатів P_{ji};

підсумовування часткових результатів та отримання макрочасткового результату P_{Mi};

підсумовування макрочасткових результатів;

обчислення функції активації f.

Аналіз формули (3) показує, що основою паралельно-вертикального опрацювання даних у нейроелементі є операція групового підсумовування

$$Z = \sum_{j=1}^{M} C_j, \tag{4}$$

де M — кількість часткових результатів; C_j — j-й частковий результат.

Нехай доданки C_j є двійковими n-розрядними додатними числами, меншими за одиницю, які записуються так:

$$C_j = \sum_{i=1}^{n} 2^{-i} C_{ji}. \tag{5}$$

Підставивши значення (5) у формулу (4), отримаємо

$$Z = \sum_{j=1}^{M} \sum_{i=1}^{n} 2^{-i} C_{ji}. \tag{6}$$

Формула (6) відображає горизонтальну модель обчислення оператора групового підсумовування. Замінивши у формулі (6) порядок підсумовування, переходимо до вертикальної моделі обчислення оператора групового підсумовування

$$Z = \sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{M_i} C_{ji}, \tag{7}$$

де M_i — кількість доданків у i-му розрядному зрізі.

У цій моделі групового підсумовування процес підсумовування зводиться до перетворення багаторядного коду в однорядний.

Методи реалізації паралельно-вертикального опрацювання даних у нейроелементі залежать від таких чинників.

1. Спосіб надходження даних:

паралельно-порозрядне надходження вхідних даних X_{ji} і вагових коефіцієнтів W_{ji};

почергове паралельно-порозрядне надходження вхідних даних X_{ji} і вагових коефіцієнтів W_{ji};

суміщення процесу паралельного порозрядного надходження вхідних даних X_{ji} і вертикально-табличного формування і підсумовування макрочасткових результатів P_{Mi}.

2. Формування для кожного розрядного зрізу часткових результатів P_{ji}:

з прямим формуванням;

на базі попередніх обчислень.

3. Формування макрочасткових результатів P_{Mi}:

послідовне;

паралельне;

послідовно-паралельне.

4. Формування результату обчислення $\sum_{i=1}^{n} 2^{-i} P_{Mi}$:

послідовне;

паралельне;

послідовно-паралельне.

Підвищення швидкодії паралельно-вертикального опрацювання даних у нейроелементі можна досягнути такими шляхами [5]:

зменшенням часу формування часткових результатів P_{ji};

зменшенням кількості часткових результатів P_{ji};

зменшенням часу формування макрочасткових результатів P_{Mi};

зменшенням часу підсумовування макрочасткових результатів P_{Mi}.

Паралельно-вертикальний метод опрацювання даних у нейроелементах (нейромережах) завдяки використанню багатоопераднього підходу, порозрядного надходження даних та суміщення процесів надходження даних з виконанням обчислень забезпечує зменшення кількості виводів інтерфейсу, розрядності міжнейронних зв'язків і затрат обладнання та підвищує швидкодію обчислень.

При НВІС-реалізації нейромереж паралельно-вертикального типу доцільно використовувати такі принципи: модульності, який передбачає розроблення компонентів нейромереж у вигляді функціонально завершених пристроїв (модулів); однорідності та регулярності архітектури нейромереж; локалізації та спрощення зв'язків між елементами нейросистеми; конвеєризації та просторового паралелізму опрацювання даних; адаптації апаратних засобів до структури алгоритмів опрацювання та інтенсивності надходження даних.

Таким чином, використання паралельно-вертикальних методів опрацювання даних при НВІС-реалізації нейроелементів і нейромереж забезпечує підвищення швидкодії, зменшення кількості виводів інтерфейсу, розрядності міжнейронних зв'язків та затрат обладнання.

1. *Хайкин С.* Нейронные сети: полный курс. – Москва: Вильямс, 2006. – 1104 с.
2. *Цмоць І. Г., Скорохода О. В., Красовський В. Б.* Операційний базис нейрокомп'ютерних систем // Зб. наук. праць Ін-ту проблем моделювання в енергетиці. – 2013. – Вип. 66. – С. 149–155.
3. *Цмоць І. Г., Ткаченко Р. О., Скорохода О. В.* Вертикально-паралельні методи та структури для реалізації базових компонентів нейромережевих технологій реального часу // Техн. вісті. – 2010. – **1(31)**, **2(32)**. – С. 166–169.
4. *Цмоць І. Г., Скорохода О. В., Балич Б. І.* Модель та НВІС-структури формального нейрона паралельно-вертикального типу з використанням мультиплексування шин // Зб. наук. праць "Моделювання та інформаційні технології" Ін-ту проблем моделювання в енергетиці. – 2013. – Вип. 67. – С. 160–166.
5. *Грицик В. В., Цмоць І. Г., Теслюк В. М.* Методологія системного проектування нейрокомп'ютерних засобів мобільних робототехнічних систем // Доп. НАН України. – 2013. – № 1. – С. 30–36.

НУ "Львівська політехніка" *Надійшло до редакції 25.02.2014*

Член-корреспондент НАН Украины **В. В. Грицык, И. Г. Цмоць, О. В. Скорохода**

Методы параллельно-вертикальной обработки данных в нейросетях

Определен операционный базис нейросетей, обоснована целесообразность разработки аппаратных нейросетей параллельно-вертикального типа, разработан ориентированный на СБИС-реализацию параллельно-вертикальный метод обработки данных в нейроэлементах (нейросетях), который обеспечивает уменьшение количества выводов интерфейса, разрядности межнейронных связей и затрат оборудования, и предложены принципы СБИС-реализации нейросетей.

Corresponding Member of the NAS of Ukraine **V. V. Grytsyk, I. G. Tsmots, O. V. Skorokhoda**

Methods of parallel vertical data processing in neural networks

An operational basis of neural networks has been identified. The feasibility of development of parallel-vertical hardware neural networks has been substantiated. A parallel-vertical data processing method in neural elements (neural networks) that is oriented to the VLSI implementation and provides a reduction of the number of interface's pins, the bitness of interneuron connection, and equipment costs has been developed. The principles of the VLSI implementation of neural networks have been proposed.