Методи паралельно-вертикального опрацювання даних у нейромережах

Д.т.н., проф. І. Цмоць, к.т.н., асист. О. Скорохода

Lviv Polytechnic National University, 12 Bandera St., Lviv-13, 79013, Ukraine

Abstract. Methods of parallel-vertical data processing in neural networks have been developed. These methods are based on calculation of group summation operator and provide the VLSI-implementation of artificial neural networks with the high efficiency of equipment use.

Keywords: parallel-vertical data processing, neural elements, VLSI implementation, group summation.

Широке впровадження штучних нейронних мереж (ШНМ) у різних областях науки, техніки і виробництва вимагають від них високих технічних характеристик. Однією з найбільш широко розповсюджених вимог, що ставиться до засобів реалізації ШНМ є забезпечення високої швидкодії [1, 2]. Подібна проблема виникає, як правило, при використанні ШНМ для розв'язання задач в реальному часі, який накладає певні обмеження на процес опрацювання інформації.

Застосування ШНМ у галузях, де апаратура є бортовою, тобто такою, що возиться, носиться, літає і плаває, накладає жорсткі обмеження на їхні масогабаритні характеристики. Одночасно до засобів реалізації ШНМ висуваються жорсткі вимоги до споживаної потужності, яка впливає на габарити джерел живлення та засобів відводу тепла. Необхідність задоволення вимог забезпечення масогабаритних характеристик, енергоспоживання, вартості змушують при розробці ШНМ під заданий клас задач дуже строго підходити до вибору параметрів, що визначають апаратні затрати на їхнє створення [3]. Це проявляється в бажанні зменшити довжину розрядної сітки, використовувати фіксовану кому для представлення операндів, скоротити перелік команд, що використовуються, і число ліній адресної шини, що визначають доступну користувачу ємність пам'яті.

Зменшення масогабаритних характеристик, енергоспоживання, підвищення надійності ШНМ та забезпечення режиму реального часу можна досягнути за допомогою їхньої НВІС-реалізації. При НВІС-реалізації ШНМ повинні забезпечити високу ефективність використання обладнання, яка враховує кількість виводів інтерфейсу, однорідність структури, кількість і локальність зв'язків, зв'язує продуктивність з витратами обладнання та дає оцінку елементам пристрою за продуктивністю.

Задача синтезу ШНМ реального часу, орієнтованих на НВІС-реалізацію, з високою ефективністю використання обладнання зводиться до мінімізації апаратних затрат, кількості виводів інтерфейсу, збільшення однорідності структури та регулярності зв'язків.

Аналіз операційного базису нейромереж показав, що нейромережеві операції за кількістю операндів, що одночасно опрацьовуються, можна розділити на одно- (корінь квадратний, передатні функції), дво- (додавання, ділення, множення) і багатооперандні (визначення мінімального та максимального чисел, багатоопераднe підсумовування, обчислення скалярного добутку, обчислення суми квадратів різниць). Відомі апаратні нейроелементи та нейромережі є, зазвичай, одно- і двоопераднимиі, 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 склалася тенденція зміни структурної одиниці опрацювання з одно- та двоопераднної на багатоопераднy, яка виконується паралельно.

Особливістю багатоопераднних нейроопераций є те, що вони виконуються над множиною операндів і результатом опрерації є одне число. Багатоопераднні нейрооперації пропонується виконувати на основі багатоопераднного підходу, при якому процес обчислення нейрооперації розглядається як виконання єдиної операції, що ґрунтується на елементарних арифметичних операціях.

Паралельна НВІС-реалізація нейроелементів і нейромереж на основі багатоопераднного підходу вимагає великих затрат обладнання і значної кількості виводів інтерфейсу, які залежать як від кількості операндів, так і від їхньої розрядності. Вартість і швидкодія паралельних НВІС-реалізацій нейроелементів і нейромереж залежить як від рівня технології, розміру кристалу, так і від кількості виводів інтерфейсу. Для забезпечення високої швидкодії НВІС-реалізації та зменшення кількості виводів інтерфейсу пропонується опрацювання даних у нейроелементах і нейромережах здійснювати паралельно розрядними зрізами (вертикально) на основі багатоопераднного підходу. Таке опрацювання даних у нейроелементах і нейромережах будемо називати паралельно-вертикальним.

Оскільки, нейромережа є сукупністю N нейронних елементів, які певним чином зв'язані між собою, то особливості паралельно-вертикального опрацювання даних доцільно розглянути на прикладі p-го нейроелемента (p=1,…, N). При паралельно-вертикальному опрацюванні даних у нейроелементі вхідні дані X_j та вагові коефіцієнти W_j (j=1,…, N, де N – кількість входів даних і вагових коефіцієнтів) представляться у порозрядному вигляді згідно з формулою:

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

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

Нейроелемент у загальному випадку здійснює перетворення відповідно до формули:

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

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

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

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

де P_{ji} – ji-ий частковий добуток (результат); P_{Mi} – i-ий макрочастковий добуток (результат), який формується додаванням N часткових добутків.

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:

$$Z = \sum_{j=1}^{M} C_j, \quad (4)$$

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:

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

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

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

Формула (6) відображає горизонтальну модель обчислення оператора групового підсумовування.

Замінивши у формулі (6) порядок підсумовування переходимо до вертикальної моделі обчислення оператора групового підсумовування, яка записується так:

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

де M_i – кількість доданків у i-у розрядному зрізі.

Ця модель групового підсумовування зводить процес підсумовування до перетворення багаторядного коду в однорядний.

Методи реалізації паралельно-вертикального опрацювання даних у нейроелементі залежать від:

1. Способу надходження даних:
- паралельним порозрядним надходженням вхідних даних X_{ji} і вагових коефіцієнтів W_{ji};
- почерговим паралельним порозрядним надходження вхідних даних X_{ji} і вагових коефіцієнтів W_{ji};
- суміщенням 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 основі попередніх обчислень.

3. Формування макрочасткових результатів P_{Mi}:
- послідовне;
- паралельне;
- послідовно-паралельне.

4. Формування результату обчислення:
- послідовне;
- паралельне;
- послідовно-паралельне.

Підвищення швидкодії паралельно-вертикального опрацювання даних у нейроелементі можна досягнути такими шляхами:
- 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}.

Для підвищення швидкодії паралельно-вертикального опрацювання даних у нейроелементах і нейромережах потрібно використовувати поряд з просторовим розпаралелюванням часове (конвеєр). За допомогою зменшення складності операцій, які реалізуються сходинкою конвеєра, підвищується тактова частота роботи конвеєра, за допомогою чого досягається підвищення швидкодії.

Використання паралельно-вертикального опрацювання даних у нероелементах та нейромережах забезпечує зменшення кількості виводів нейроелементів, розрядності міжнейронних зв'язків та зменшує витрати обладнання.

[1]. Misra J. Artificial neural networks in hardware: A survey of two decades of progress / J. Misra, I. Saha // Neurocomputing. – 2010. – Vol. 74, Issue 1. – pp. 239–255.

[2]. Haykin S. Neural networks and learning machines. Third Edition. / S. Haykin. – New York: Prentice Hall, 2009. – 936 p.

[3]. Цмоць І.Г. Особливості проектування спеціалізованих комп'ютерних систем для обробки інтенсивних потоків даних / І.Г. Цмоць // Збірник наукових праць. Моделювання та інформаційні технології. – 1999. – Випуск 4. – С. 113–118.