УДК 004.032.26

І.Г. Цмоць, д.т.н., О.В. Скорохода, В.Б. Красовський, Національний університет «Львівська політехніка», м. Львів

ОПЕРАЦІЙНИЙ БАЗИС НЕЙРОКОМП'ЮТЕРНИХ СИСТЕМ

Анотація. Проаналізовано особливості апаратної реалізації штучних нейронних мереж, сформовано та проаналізовано операційний базис нейрокомп'ютерних систем.

Аннотация. Проанализированы особенности аппаратной реализации искусственных нейронных сетей, сформирован и проанализирован операционный базис нейрокомпьютерных систем.

Abstract. Features of a hardware implementation of artificial neural networks have been analyzed; operational basis of neurocomputing systems has been formed and analyzed.

Ключові слова: нейрокомп'ютинг, штучні нейронні мережі, нейрокомп'ютерні системи, операційний базис.

Ключевые слова: нейрокомпьютинг, искусственные нейронные сети, нейрокомпьютерные системы, операционный базис.

Keywords: neurocomputing, artificial neural networks, neurocomputing systems, operational basis.

Вступ.

На сучасному етапі розвитку штучних нейромережевих (ШНМ) 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 розробки [1]. Особливістю таких задач є:

- регулярність і рекурсивність алгоритмів;
- постійність та висока інтенсивність надходження даних;
- великий об'єм обчислень, у яких переважають обчислювальні, а не логічні операції;
- можливість розпаралелювання обчислень як в часі, так і в просторі.

При обробці інтенсивних потоків даних за складними алгоритмами режим реального часу забезпечується розпаралелюванням і конвеєризацією процесів обчислень та використанням нових технологічних досягнень в області розробки надвеликих інтегральних схем (НВІС). Тому актуальною задачею є розробка високоефективних нейрокомп'ютерних систем реального часу, орієнтованих на НВІС-реалізацію.

Основна частина.

Аналіз галузей застосування нейротехнологій реального часу, архітектур, типових задач і нейромережевих алгоритмів показав, що вони

мають такі особливості:

- високу інтенсивність та постійність вхідних потоків даних;
- постійне ускладнення алгоритмів обробки та підвищення вимог до точності результатів;
- можливість розпаралелення обробки як у часі, так і у просторі;
- здатність до узагальнення та абстрагування;
- навчання, самонавчання та самоорганізації.

Проведений аналіз показав, що забезпечення нейромережевої обробки інтенсивних потоків даних у реальному часі можливо апаратними засобами. Апаратні нейромережі реального часу ґрунтуються на операційному базисі, який наведений на рис. 1.

Операційний базис нейрокомп'ютерних систем складається із нейроопераций попередньої обробки, процесорних нейроопераций та обчислення елементарних функцій.

На етапі попередньої обробки даних початкові дані, які потрібно подати на входи мережі, необхідно перетворити до вигляду, який дасть найкращі результати. Навчальний вектор містить по одному значенню на кожний вхід мережі і, у залежності від типу навчання (з вчителем або без), по одному значенню для кожного виходу мережі. Навчання мережі на «сирому» наборі, як правило, не дає якісних результатів. Існує ряд способів покращити «сприйняття» мережі [2]:

- нормалізація виконується тоді, коли на різні входи мережі подаються дані різної розмірності. Наприклад, на перший вхід мережі подаються величини зі значеннями від нуля до одиниці, а на другий – від ста до тисячі. При відсутності нормування значення на другому вході завжди будуть мати набагато більший вплив на вихід мережі, ніж значення на першому вході. При нормалізації розмірності всіх вхідних та вихідних даних зводяться до одного діапазону;
- квантування виконується над неперервними величинами, для яких виділяється скінченний набір дискретних значень. Наприклад, квантування використовуються для задання частот звукових сигналів при розпізнаванні мови;
- фільтрація виконується для «зашумлених» даних і полягає у відкиданні значень, які, швидше за все, є некоректними.

Нормалізацію вхідних даних рекомендовано робити завжди. Нормалізація – це процедура попередньої обробки вхідних даних (навчальних, тестових і робочих вибірок), при якій значення ознак, які формують вхідний вектор, приводиться до деякого заданого діапазону. Після нормалізації всі значення вхідних ознак будуть приведені до деякого вузького діапазону (зазвичай, [0, 1] або [-1, 1].

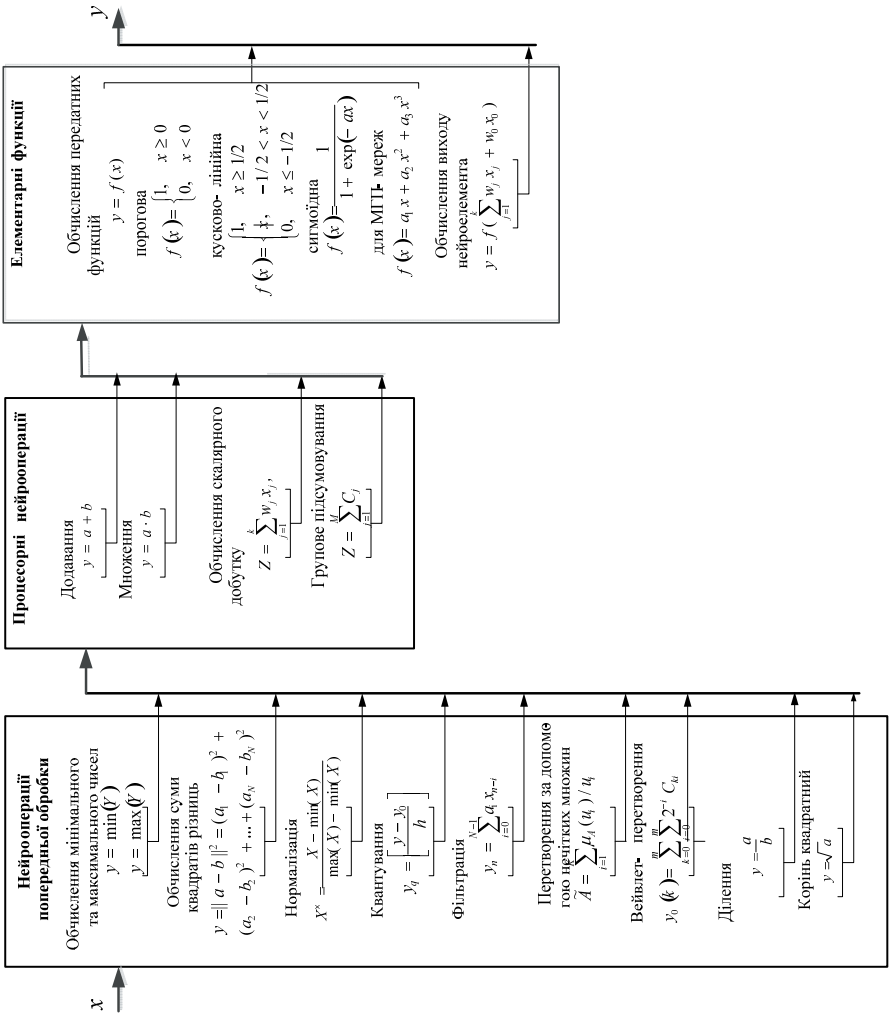

Рис. 1. Операційний базис апаратних нейромереж реального часу

Існує велика кількість способів нормалізації вхідних значень. Найпростішою, але у більшості випадків ефективною, є лінійна нормалізація. Якщо початкові дані потрібно привести до діапазону [0, 1], то вона виконується так:

$$X^{\times} = \frac{X - \min(X)}{\max(X) - \min(X)}. \qquad (1)$$

Для приведення початкових даних до діапазону [-1, 1] лінійна нормалізація здійснюється таким чином:

$$X^{\times} = \frac{X}{\max(|X|)} \tag{2}$$

Якщо вхідні дані X щільно заповнюють певний інтервал, то використання лінійної нормалізації оптимальне, оскільки вона не потребує здійснення складних обчислень.

Але лінійна нормалізація ефективна не завжди, наприклад, у випадках, коли існують рідкісні відхилення, які значно більші за типові значення. У таких випадках лінійна нормалізація призведе до того, що більшість значень початкових даних будуть близькими до нуля.

Цього недоліку не має нормалізація за допомогою стандартного відхилення:

$$X^{\times} = \frac{X - \overline{X}}{\sigma},\ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \sigma = \frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 . \tag{3}$$

При такій нормалізації досягається більш рівномірний розподіл вхідних даних. Але нормалізація за допомогою стандартного відхилення має суттєвий недолік – отримані нормалізовані значення не обов'язково будуть належати до одиничного інтервалу. Для вхідних даних це не важливо, але вихідні дані можуть використовуватися як еталонні значення для виходів нейронів. А це недопустимо у випадку, коли функція активації нейрона є сигмоїдною, оскільки вона приймає значення тільки в одиничному діапазоні.

Цього недоліку позбавлена нелінійна нормалізація, наприклад:

$$X^{\times} = f\left(\frac{X - \overline{X}}{\sigma}\right),\ f(a) = \frac{1}{1 + e^{-a}} \tag{4}$$

Дана функція нормалізує більшість значень рівномірно і гарантує, що нормалізоване значення буде в діапазоні [0, 1]. Недоліком такої нормалізації є складність для апаратної реалізації.

У більшості випадків, якщо вхідні дані є більш-менш рівномірними, для апаратної реалізації найкраще використовувати лінійну нормалізацію (2). Для апаратної реалізації такої нормалізації потрібно розробити методи та структури для обчислення наступних базових операцій:

- визначення максимального числа з групи чисел;
- ділення.

Після нормалізації даних, залежно від типу мережі, можуть використовуватися інші процедури попередньої обробки даних.

Зокрема, після нормалізації даних у RBF- та GRNN-мережах потрібно здійснити обчислення евклідової відстань від кожного вхідного вектора до всіх інших. Для цього обчислення використовується базова операція обчислення суми квадратів різниць:

$$y = \| x_i^e - x_i^b \|^2 = (x_1^e - x_1^b)^2 + (x_2^e - x_2^b)^2 + ... + (x_N^e - x_N^b)^2 . \tag{5}$$

Для інших типів нейронних мереж можуть використовуватися інші види попередньої обробки даних. Наприклад, для нейро-нечітких мереж над

вхідними даними здійснюються перетворення за допомогою нечітких множин та правил нечіткої логіки, які мають добрі апроксимуючі здатності. У вейвлет нейронних мережах для аналізу різних частотних компонентів вхідних даних використовуються вейвлет-перетворення, наприклад, вейвлет Хаара, або вейвлети Добеші [3].

Перетворювач кодів повинен забезпечувати паралельно-послідовне та послідовно-паралельне перетворення. Паралельно-послідовне перетворення здійснюється шляхом послівного приймання N чисел у кожному такті та видачу в кожному такті розрядних зрізів всіх чисел. Таке перетворення використовується для завантаження вертикальних операційних пристроїв, у яких здійснюється вертикальна обробка, яка передбачає у кожному такті послідовну обробку i-х розрядів всіх чисел, тобто, здійснюється послідовна обробка розрядів при паралельній обробці чисел. Послідовно-паралельне перетворення використовується для перетворення виходів результатів обробки вертикальних операційних пристроїв. При послідовно-паралельному перетворенні на вхід результати надходять розрядними зрізами, а на вихід видаються послівно.

На другому етапі операції над вхідними даними виконуються безпосередньо у самій мережі у процесі навчання та функціонування. Аналіз існуючих алгоритмів показав, що всі основні обчислювальні операції в нейронних мережах можна звести до таких базових операцій:

- обчислення скалярного добутку;
- групове підсумовування;
- обчислення передатних функцій;

Серед усієї сукупності операцій, які найчастіше використовуються у нейроалгоритмах, особливої уваги заслуговує операція обчислення сум парних добутків [4-6]. Традиційно обчислення такої операції здійснюється за наступною формулою:

$$Z = \sum_{j=1}^{k} W_j X_j, , \qquad (6)$$

де k – кількість входів нейроелемента, W_j – j-й ваговий коефіцієнт, X_j – значення j-го входу.

Існують два підходи до апаратної реалізації обчислення сум парних добутків [7, 8]. Перший з них ґрунтується на операціях множення і додавання, другий – на операціях додавання, інверсії та зсуву. Перший підхід переважно використовують при синтезі пристроїв обчислення сум парних добутків на базі окремих мікросхем (помножувачів, суматорів), а другий – при НВІС-реалізаціях. Використання для НВІС-реалізацій алгоритмів на базі операцій додавання, інверсії та зсуву дозволяє оптимізувати пристрій за швидкодією, апаратними витратами та збільшити регулярність його структури. Основою таких алгоритмів є формування часткових добутків з наступним їх додаванням.

Основним елементом пристрою обчислення скалярного добутку є багатовходовий суматор. Тобто, у загальному випадку обчислення скалярного добутку в базисі елементарних арифметичних операцій зводиться до макрооперації групового підсумовування макрочасткових добутків:

$$Z = \sum_{j=1}^{M} C_j \ , \tag{7}$$

де M – кількість доданків; C_j – j-й доданок [7].

Сигнал, отриманий після обчислення скалярного добутку, перетворюється у вихідний сигнал через алгоритмічний процес, відомий під назвою передатна функція. У передатній функції для визначення виходу нейрона загальна сума порівнюється з деяким порогом. Якщо сума є більшою за значення порога, елемент обробки генерує сигнал, в іншому випадку сигнал не генерується або генерується гальмуючий сигнал.

Переважно застосовують нелінійну передатну функцію, оскільки лінійні (прямолінійні) функції обмежені і вихід є прямо пропорційним до входу. Застосування лінійних передатних функцій було проблемою у ранніх моделях мереж, і їх обмеженість та недоцільність була доведена в [9].

Перед обчисленням передатної функції до вхідного сигналу інколи додають однорідно-розподілений випадковий шум, джерело та кількість якого визначається режимом навчання. У літературі цей шум згадується як «температура» штучних нейронів, яка надає математичній моделі елемент реальності.

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 багатооперандні (визначення мінімального та максимального чисел, багатооперандне підсумовування, обчислення скалярного добутку, обчислення суми квадратів різниць). Існуючі апаратні нейроелементи та нейромережі є в основному одно- і двоопераднями, це пов'язано з можливостями елементної бази. Еволюція розвитку апаратних нейроелементів та нейромереж тісно пов'язана з структурною одиницею обробки, тобто з розрядністю і кількістю операндів, які одночасно опрацьовує операційний пристрій. З розвитком інтегральної технології намітилась тенденція зміни структурної одиниці обробки з одно- та двооперандної на багатооперандну, яка виконується паралельно.

Особливістю багатооперандних нейронних операцій є те, що вони виконуються над множиною операндів і результатом опрерації є одне число. Багатооперандні нейрооперації пропонується виконувати на основі багатооперандного підходу, при якому процес обчислення нейрооперації розглядається як виконання єдиної операції, що ґрунтується на елементарних арифметичних операціях.

Висновки.

1. Найхарактернішими особливостями нейротехнологій реального часу та задач є: постійність і висока інтенсивність надходження вхідних даних; великий обсяг обчислень з перевагою обчислювальних операцій над логічними; регулярність і рекурсивність нейромережевих алгоритмів обробки даних; постійне ускладнення алгоритмів обробки та підвищення вимог до точності результатів; можливість розпаралелення обробки як у часі, так і у просторі; здатність до узагальнення та абстрагування; навчання, самонавчання та самоорганізації під впливом зовнішнього середовища.

2. Операційним базисом апаратних нейромереж реального часу є операції: додавання, множення, ділення, добування квадратного кореня, групове підсумовування, обчислення скалярного добутку, передатної функції, мінімальних і максимальних чисел, фільтрація, перетворення за допомогою нечітких множин, вейвлет перетворення, квантування та нормалізація.

3.Для підвищення якості результатів нейрообробки доцільно використати попередню обробку, яка передбачає нормалізацію, квантування та фільтрацію вхідних даних.

1. *Цмоць І.Г.* Інформаційні технології та спеціалізовані засоби обробки сигналів і зображень у реальному часі. – Львів: УАД, 2005.- 227с.

2. Искусственная нейронная сеть [Електронний ресурс]. – Режим доступу: http://ru.wikipedia.org/wiki/искусственная_нейронная_сеть.

3. *Буй Т.Т.* Алгоритмическое и программное обеспечение для классификации цифровых изображений с помощью вейвлет-пребразования Хаара и нейронных сетей / *Т.Т. Буй, Н.Х. Фан, В.Г. Спицын* // Известия ТПУ. – 2011. – №5. – С. 103–106.

4. *Круглов В.В.* Искусственные нейронные сети. Теория и практика. 2-е изд. / *В.В. Круглов, В.В. Борисов.* – М.: Горячая линия-Телеком, 2002. – 382 с.

5. Галушкин А.И. Нейрокомпьютеры. Книга 3. / А.И. Галушкин. – М.: ИПРЖР, 2000. – 528 с.

6. *Руденко О.Г., Бодянський Є.В.* Штучні нейронні мережі / *О.Г. Руденко, Є.В. Бодянський.* – Харків: ТОВ «Компанія СМІТ», 2006. – 404 с.

7. *Цмоць І.Г.* Інформаційні технології та спеціалізовані засоби обробки сигналів і зображень у реальному часі/*І.Г. Цмоць.* – Львів: УАД, 2005. – 227 с.

8. Рабинович З.Л. Типовые операции в вычислительных машинах / *З.Л. Рабинович, В.А. Раманаускас.* – К.: Техніка, 1980, –240 с.

9. *Минский М.* Персептроны / *М. Минский, С. Пейперт.* – М.: Мир, 1971. – 266 с.

Поступила 18.02.2013р.