Міністерство освіти і науки, молоді та спорту України

Харківський національний університет радіоелектроніки

ЛЕБЬОДКІНА АЛЛА ЮРІЇВНА

УДК 681.5; 004.272

МЕТОДИ ТА МОДЕЛІ ПРИСКОРЕНОЇ НЕЙРОМЕРЕЖЕВОЇ ОБРОБКИ ДАНИХ У РОЗПОДІЛЕНОМУ ОБЧИСЛЮВАЛЬНОМУ СЕРЕДОВИЩІ

05.13.23 – системи та засоби штучного інтелекту

Автореферат дисертації

на здобуття наукового ступеня кандидата технічних наук

Харків – 2012

Дисертацією є рукопис.

Робота виконана у Харківському національному університеті радіоелектроніки, Міністерство освіти і науки, молоді та спорту України.

Науковий керівник - кандидат технічних наук, старший науковий співробітник **Аксак Наталія Георгіївна,** Харківський національний університет радіоелектроніки, доцент кафедри електронних обчислювальних машин.

Офіційні опоненти: доктор технічних наук, професор **Гороховатський Володимир Олексійович,** Харківський інститут банківської справи Університету банківської справи Національного банку України, завідувач кафедри інформаційних технологій, м. Харків;

доктор технічних наук, професор **Михальов Олександр Ілліч,** Національна металургійна академія України, завідувач кафедри інформаційних технологій і систем, м. Дніпропетровськ.

Захист відбудеться «14» листопада 2012 р. о 13^{00} годині на засіданні спеціалізованої вченої ради Д 64.052.01 у Харківському національному університеті радіоелектроніки за адресою: 61166, м. Харків, просп. Леніна, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: 61166, м. Харків, просп. Леніна, 14.

Автореферат розісланий «12» жовтня 2012 р.

Учений секретар
спеціалізованої вченої ради Є.І. Литвинова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Необхідність вирішення в сучасних дослідженнях складних наукових і виробничих завдань за найкоротші терміни приводить до нарощування потужності обчислювальних ресурсів і появи нових інформаційних технологій, зокрема таких, що застосовують елементи штучного та обчислювального інтелекту. Насамперед, слід зазначити доцільність застосування штучних нейронних мереж, на розвиток яких значно впливають як українські вчені, серед яких Руденко О.Г., Бодянський Є.В., Куссуль Н.М., так і зарубіжні – Хайкін С., Яхно В.Г., Галушкін О.І., Горбань А.Н., Круглов В.В., Кірсанов Е.Ю., Тєрєхов С.О..

Можна виділити галузі науки, які не можуть обійтися без великих обчислювальних ресурсів (метеорологія, ядерна фізика і геофізика, обчислювальна хімія, генна інженерія, екологія та інші). Авторами Петренком А.І., Святним В.А., Мельником А.О., Воєводіним В.В., Хорошевським В.Г., Бухановським О.В., Гергелем В.П. вирішені завдання моделювання та аналізу паралельних обчислень. Дослідження показують успішне впровадження паралельних обчислень з використанням спеціалізованих суперкомп'ютерів і розподілених середовищ для вирішення різних наукових завдань у даних галузях науки.

Сьогодні багато досліджень у наукових і прикладних областях створюють передумови для виникнення нейромережевих систем прискореної обробки даних, оскільки для багатьох типів штучних нейронних мереж із зростанням вхідних даних обчислювальні витрати і обсяг пам'яті збільшуються експоненціально, що значно уповільнює процес навчання мережі, істотно знижуючи тим самим ефективність вирішення поставленої задачі.

Не зважаючи на значну потребу у вирішенні описаних проблем, слід зазначити досить слабку вивченість прискореної нейромережевої обробки даних. Як показали дослідження в цьому напрямі, реалізація процесів навчання нейронних мереж із застосуванням сучасних паралельних і розподілених технологій дозволяє, якщо не запобігти, то значною мірою зменшити негативний вплив збільшення розмірності, обсягу вхідного простору і властивостей нейронних мереж.

У зв'язку з цим розроблення методів і моделей обробки даних нейронною мережею за допомогою високопродуктивних обчислень у розподіленому середовищі є актуальним як з теоретичної, так і з практичної точки зору.

Зв'язок роботи з науковими програмами, темами, планами. Дисертаційну роботу виконано відповідно до тематичного плану Харківського національного університету радіоелектроніки в рамках держбюджетних НДР, згідно з наказами Міністерства освіти і науки, молоді та спорту України: 1) НДР «Синтез методів обробки інформації в умовах невизначеності на основі самонавчання і м'яких обчислень»; розділ «Гібридні моделі, що самонавчаються, в задачах обробки нечіткої інформації» (№ ДР 0107U003028); 2) НДР «Еволюційні гібридні системи обчислювального інтелекту зі змінною структурою для інтелектуального аналізу даних», розділ «Еволюційні гібридні методи і моделі інтелектуальної обробки інформації зі змінною структурою в умовах невизначеності» (№ ДР 0110U000458). Автор був одним із виконавців робіт з даних тем.

Мета і задачі дослідження. Метою дисертаційної роботи є розробка методів і моделей прискореного навчання та функціонування багатошарової нейронної мережі прямого поширення (БШМ), які за рахунок диспетчерування та масштабування обчислювальної системи, а також адаптації та підвищення продуктивності під час реалізації БШМ забезпечують суттєве скорочення часу вирішення задач великої розмірності.

Для досягнення поставленої мети необхідно вирішити такі задачі:

– розробити метод рівномірного розподілу нейромережевої обробки даних (метод РРНОД);

– адаптувати нейронну мережу в розподіленому середовищі;

– розробити метод масштабування обчислювальної системи (метод МОС);

– розробити модель оцінювання прискорення навчання та функціонування (модель ОПНФ) багатошарової нейронної мережі прямого поширення;

– провести експериментальне дослідження запропонованих методів і моделей, здійснити їх реалізацію на основі технологій паралельного та розподіленого програмування для скорішого розв’язання практичних задач великої розмірності.

Об'єктом дослідження є процес навчання та функціонування багатошарової нейронної мережі прямого поширення.

Предметом дослідження є методи та моделі прискореної нейромережевої обробки даних у розподіленому обчислювальному середовищі.

Методи дослідження. Для рівномірного розподілу нейромережевої обробки даних використані теорія штучних нейронних систем, лінійна алгебра, методи алгоритмізації задач, обчислювальні методи; для моделі оцінювання застосовані параметричне моделювання, системний аналіз; для адаптації та прискорення нейрообробки даних – теорія графів, теорія матриць, принципи організації комп’ютерних мереж; для масштабування обчислювальної системи – теоретичні основи побудови високопродуктивних систем, теорія паралельних та розподілених обчислень; для підтвердження ефективності отриманих результатів і розроблення рекомендацій щодо їх застосування – імітаційне моделювання з використанням мов високого рівня програмування.

Наукова новизна результатів дисертаційної роботи:

1. Вперше запропоновано метод рівномірного розподілу нейромережевої обробки даних, який базується на динамічному перерозподілі наборів нейронів БШМ між обчислювачами залежно від обсягу оброблюваних даних, що дозволяє суттєво скоротити час навчання та функціонування БШМ, а також зменшити на порядок обчислювальну складність порівняно з існуючими послідовними методами.

2. Вперше запропоновано модель оцінювання прискорення навчання та функціонування БШМ, яка характеризується можливістю вибору ефективних значень параметрів, що сукупно враховують обсяг і розподіл вхідної інформації залежно від числа передач даних у віртуальних топологіях («зірка», «сітка», «повнозв’язний граф») та апаратні характеристики середовища й обчислювачів, що дозволяє підвищити продуктивність нейрообчислень та значно прискорити обробку великого обсягу даних у розподіленому обчислювальному середовищі.

3. Вперше запропоновано метод масштабування обчислювальної системи, який характеризується точним визначенням часу виконання розподілених між

обчислювачами наборів даних нейропроцедур, що дозволяє ефективно диспетчерувати навантаження ресурсів та оцінити продуктивність подальшого підвищення потужності гетерогенного або гомогенного обчислювального середовища для прискорення вирішення задач.

4. Набула подальшого розвитку модель прискореної нейрообробки даних шляхом урахування обсягу вхідної інформації, топології передачі даних («зірка», «повнозв'язний граф», «сітка») для скорочення кількості передач між обчислювачами, що дозволяє адаптувати структуру БШМ у розподіленому обчислювальному середовищі для скорішого вирішення задач великої розмірності.

Практичне значення результатів дисертаційної роботи. Результати дисертаційної роботи були використані під час побудови інтелектуальних систем, що обробляють суттєвий обсяг вхідної інформації для вирішення великих обчислювальних задач у різних галузях.

Методи та моделі прискореної нейромережевої обробки даних у розподіленому обчислювальному середовищі доведені до рівня програмної реалізації, що дозволило здійснити:

1) розбракування безшовних труб різного призначення за якістю у НПФ ТОВ «Технологія» м. Харкова, Україна (акт впровадження від 18.05.09 р.);

2) прогнозування екологічної обстановки в санітарно-захисній зоні ПАТ «АрселорМіттал Кривий Ріг» для ТОВ "АТОМЕКОСИСТЕМА" м. Харкова, Україна (акт впровадження від 30.05.11 р.).

Проведені експериментальні дослідження для оцінки розроблених методів і моделей підтверджують основні положення, що виносяться на захист.

Результати роботи були впроваджені в навчальний процес Харківського національного університету радіоелектроніки (акт впровадження від 15.03.10 р.).

Особистий внесок здобувача. Усі основні результати, що виносяться на захист, отримано автором самостійно. У роботах, опублікованих у співавторстві, здобувачеві належать: [1] – паралельна нейропроцедура для класифікації даних; [2] – нова модель ОПНФ БШМ з топологіями «зірка» і «повнозв'язний граф»; [3] – новий метод МОС і нова модель ОПНФ БШМ з топологією «сітка»; [4] – модель ОПНФ двошарової нейронної мережі для декомпозиції на рівнях реалізації функцій нейронів і скалярних операцій без урахування топології передачі даних; [5] – модель ОПНФ БШМ для різних архітектур обчислювальних систем з використанням мережного закону Амдала; [6] – новий метод РРНОД та удосконалена модель прискореної нейрообробки даних з віртуальною топологією «зірка»; [7] – новий метод РРНОД та удосконалена модель прискореної нейрообробки даних з віртуальними топологіями «повнозв'язний граф» і «сітка»; [8] – порівняльний аналіз послідовного та прискореного навчання БШМ для різних топологій передачі даних з використанням бібліотеки MPI; [9] – паралельна модель БШМ з топологією «повнозв'язний граф»; [10] – обґрунтований вибір топологій «зірка», «повнозв'язний граф» і «сітка» для побудови паралельної моделі БШМ; [11] – модель оцінювання прискорення та ефективності навчання двошарової нейронної мережі для багатопроцесорних і багатоядерних систем з використанням мережного закону Амдала; [12] – порівняльний аналіз послідовного та прискореного навчання БШМ для різних топологій передачі даних з технологіями MPI та PLINQ;

[13] – модель оцінювання ефективності навчання двошарової нейронної мережі для декомпозиції на рівнях реалізації функцій нейронів і скалярних операцій; [14] – порівняльний аналіз послідовного та прискореного навчання БШМ для різних топологій передачі даних зі спеціальними опціями компілятора; [15] – дослідження розмірів пакетів під час передачі даних з використанням бібліотеки MPI; [16] – порівняльний аналіз послідовного та прискореного навчання БШМ для різних топологій передачі даних з бібліотеками MPI та MS-MPI; [17] – дослідження паралельних моделей нейроалгоритму БШМ з різними топологіями передач даних залежно від доступної архітектури обчислювальної системи; [18] – нейромережева модель прискореного розбракування труб; [19] – модель ОПНФ БШМ з використанням методу МОС; [20] – визначення обчислювальної складності методу РРНОД; [21] – обґрунтована актуальність застосування прискореної БШМ.

Апробація результатів дисертації. Основні положення та результати дисертаційної роботи доповідалися й обговорювалися на: 12, 15 Міжнародних молодіжних форумах «Радиоэлектроника и молодежь в XXI веке» (м. Харків, 2008, 2011); 7, 8 Міжнародних конференціях-семінарах «Высокопроизводительные параллельные вычисления на кластерных системах» (м. Нижній Новгород, 2007, м. Казань, 2008); 7 Міжнародній науково-технічній конференції «Проблемы информатики и моделирования-2007» (м. Харків, 2007); Міжнародній науковій конференції «Моделирование-2008» (м. Київ, 2008); Всеросійських конференціях «Технологии Microsoft в теории и практике программирования» (м. Нижній Новгород, 2008, 2009); 10 Міжнародній науково-технічній конференції «Системний аналіз і інформаційні технології» (м. Київ, 2008); Міжнародній молодіжній науковій конференції «XXXIV Гагаринские чтения» (м. Москва, 2008); Міжнародній науково-практичній конференції «Информатика, математическое моделирование, экономика» (м. Смоленськ, 2011).

Публікації. За темою дисертаційної роботи опубліковано 21 наукова праця, з них: 7 статей у наукових фахових виданнях України з технічних наук, 14 публікацій у збірниках матеріалів і тез доповідей на міжнародних наукових конференціях, семінарах, форумах.

Структура та обсяг дисертаційної роботи. Дисертація складається із вступу, чотирьох розділів, висновків, списку використаних літературних джерел зі 150 найменувань на 18 сторінках, 3 додатків на 14 сторінках. Робота містить 45 рисунків, 10 таблиць. Загальний обсяг роботи складає 175 сторінок, з них 139 основного тексту.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі стисло розглянуто стан досліджень у галузі прискореного навчання та функціонування БШМ, обґрунтовано актуальність теми, сформульовано мету та задачі дослідження, охарактеризовано наукову новизну і практичне значення отриманих результатів, наведено кількість публікацій за темою роботи, виділено особистий внесок здобувача.

У першому розділі досліджені існуючі методи прискорення навчання та функціонування БШМ на різних однорідних обчислювальних системах, у яких завдання мінімізації часових витрат вирішуються під час проектування архітектури мережі шляхом реалізації повного паралельного перебору структур БШМ і типів активаційних функцій нейронів, а під час навчання мережі, застосовуючи паралельну реалізацію складних активаційних функцій, декомпозицію стислої навчальної вибірки, розбиття матриць вагових коефіцієнтів або наборів нейронів кожного шару.

Проведений аналіз показав, що існуючі методи прискорення нейрообробки даних мають недоліки, обумовлені тим, що не дають істотної можливості скоротити час обробки і є трудомісткими, оскільки, по-перше, неефективно використовують обчислювальні ресурси; по-друге, потребують велику кількість передач даних; по-третє, спричиняють труднощі під час обробки великого обсягу вхідних даних.

Виявлені недоліки під час використання нейрокомп'ютерів, що полягають у високій вартості апаратної частини, складності реалізації та масштабування, не сумісності і закритості розробок.

Оскільки обробка значного обсягу даних вимагає тривалого часу навчання та функціонування БШМ, виникає необхідність розробки методів і моделей, які дозволять ефективно диспетчерувати та масштабувати обчислювальну систему під час реалізації БШМ, а також адаптувати та підвищити продуктивність БШМ у розподіленому середовищі, що відповідно значно скорочує час вирішення великих обчислювальних завдань і дає змогу уникнути прокляття розмірності.

На основі проведеного аналізу визначено сукупність перспективних напрямків адаптації та прискорення нейрообробки даних за допомогою високопродуктивних обчислень у розподіленому обчислювальному середовищі і сформульовано задачі дисертаційної роботи.

У другому розділі вперше запропоновано метод РРНОД; набула подальшого розвитку модель прискореної нейрообробки даних.

БШМ задається архітектурою $n_1 - n_2 - \ldots - n_L$, де n_1 – кількість нейронів у вхідному шарі, n_ℓ, $\ell = \overline{2, L-1}$ – кількість нейронів у прихованих шарах, n_L – кількість нейронів у вихідному шарі. Навчальна вибірка $\{(X(1), D(1)), (X(2), D(2)), \ldots, (X(K), D(K))\}$ складається з K прикладів $X(k) = \left[x^1(k), \ldots, x^{n_1}(k)\right]^T$, і відповідних ним значень цільової ознаки $D(k) = \left[d^1(k), d^2(k), \ldots, d^{n_L}(k)\right]^T$, $k = \overline{1, K}$, тестова вибірка містить T прикладів. Для обробки навчальної і тестової вибірок потрібно U епох навчання.

У загальному випадку БШМ навчається методом зворотного поширення помилки відповідно до таких виразів.

Вектор виходів нейронів $Y_\ell(k) = \left[y_\ell^1(k), y_\ell^2(k), \ldots, y_\ell^{n_\ell}(k)\right]^T$ шарів $\ell = \overline{2, L}$ обчислюється як

$$Y_\ell(k) = f\left(W_\ell^T(k) Y_{\ell-1}(k) + W_0^T(k)\right), \qquad (1)$$

причому $Y_1(k) = X(k)$; k – поточний приклад навчальної вибірки; $f(\bullet)$ – активаційна функція; $W_\ell(k) = (w_{ij}(k))_{i=0,\,j=1}^{n_{\ell-1},\,n_\ell}$ – матриця вагових коефіцієнтів шарів $\ell = \overline{2,L}$, в якій елементи нульового рядка $W_0(k) = [w_{01}(k), w_{02}(k), \ldots, w_{0n_\ell}(k)]$ матриці $W_\ell(k)$ є зміщенням відповідного шару.

Вектор помилок реакції мережі $\Sigma_\ell(k) = [\delta_\ell^1(k), \delta_\ell^2(k), \ldots, \delta_\ell^{n_\ell}(k)]^T$, $\ell = \overline{L,2}$ визначається відповідно до виразів для вихідного шару

$$\Sigma_L(k) = Y_L(k)(1 - Y_L(k))(D(k) - Y_L(k)), \tag{2}$$

де роль одиниці відіграє одиничний вектор $1 = [1, 1, \ldots, 1]^T$, і для прихованих шарів нейронів $j = \overline{1,n_\ell}$

$$\delta_\ell^j(k) = y_\ell^j(k)(1 - y_\ell^j(k))\sum_{m=1}^{n_{\ell+1}}(\delta_{\ell+1}^m(k) w_{jm}(k)). \tag{3}$$

Налаштування вагових коефіцієнтів шарів $\ell = \overline{2,L}$ нейронів $j = \overline{1,n_\ell}$ здійснюється відповідно до

$$w_{ij}(k+1) = \eta \delta_\ell^j(k) y_\ell^j(k) + \alpha w_{ij}(k), \tag{4}$$

де η – параметр швидкості навчання, α – коефіцієнт інерційності.

Визначення поточної середньоквадратичної помилки навчання E здійснюється на основі

$$E = \frac{1}{2}\sum_{k=1}^{K}(D(k) - Y_L(k))^2.$$

Визначення помилки узагальнення під час тестування нейронних мереж здійснюється відповідно до

$$E_g = \frac{1}{K}\sum_{t=1}^{T}\Delta_t,$$

де Δ_t визначається як

$$\Delta_t = \begin{cases} 1, \text{ якщо } D(t) \neq y(t), \\ 0, \text{ якщо } D(t) = y(t). \end{cases}$$

Загальна кількість скалярних операцій, необхідних для послідовного навчання БШМ, визначається відповідно до виразу

$$\begin{aligned} Q = U \times (K \times (&\sum_{m=2}^{L}(2n_{m-1} + 3)n_m + \sum_{m=L-1}^{2}(2n_{m+1} + 7)n_{m-1}n_m + \\ &+ 8n_L n_{L-1}) + T \times (\sum_{m=2}^{L}(2n_{m-1} + 3)n_m)), \end{aligned} \tag{5}$$

для послідовного функціонування БШМ – $Q_f = \sum_{m=2}^{L}(2n_{m-1} + 3)n_m$.

Аналіз скалярних операцій, необхідних для послідовного навчання та функціонування БШМ, показав, що градієнтний метод навчання БШМ та його модифікації складаються з послідовних взаємозалежних етапів матричної обробки, причому кожен етап є сукупністю незалежних операцій, які можна виконувати паралельно й одночасно.

Метод рівномірного розподілу нейромережевої обробки даних з P обчислювачами мають вигляд у таких обчислювальних середовищах:

- з топологією «зірка»

$$g_h = \begin{cases} (h-1)\left(\left[n_\ell / P-1\right]+1\right)+\varphi,\ \varphi=\overline{1,\left(\left[n_\ell / P-1\right]+1\right)}, & \text{якщо}\quad 1\le h\le b; \\ (P-1-h_d)\left(\left[n_\ell / P-1\right]+1\right)+\left(h-P-1+b-1\right)\left[n_\ell / P-1\right]+\varphi,\ \varphi=\overline{1,\left[n_\ell / P-1\right]}, & \text{якщо}\ b<h\le P-1; \end{cases} \tag{6}$$

де $g_h, h=\overline{1,P-1}$ – зростаюча послідовність, елементами якої є номери нейронів $\ell=\overline{2,L}$ шарів, що обробляються h-м процесором; $b=n_\ell \bmod(P-1)$ – номер обчислювача, після якого для $h_d=P-1-b$ комп'ютерів, починаючи з $(b+1)$-го процесора, кількість оброблюваних нейронів зменшується на один для рівномірного завантаження;

- з топологією «повнозв'язний граф»

$$r_h = \begin{cases} h\left(\left[n_\ell / P\right]+1\right)+\varphi,\ \varphi=\overline{1,\left(\left[n_\ell / P\right]+1\right)}, & \text{якщо}\quad 0\le h<b; \\ (P-h_d)\left(\left[n_\ell / P\right]+1\right)+\left(h-P+b\right)\left[n_\ell / P\right]+\varphi, \quad \varphi=\overline{1,\left[n_\ell / P\right]}, & \text{якщо}\quad b\le h\le P-1; \end{cases} \tag{7}$$

де r_h, $h=\overline{0,P-1}$ – зростаюча послідовність, елементами якої є номери нейронів $\ell=\overline{2,L}$ шарів, що обробляються h-м обчислювачем; $b=n_\ell \bmod P$, $h_d=P-b$;

- з топологією «сітка»

$$a_{h_z} = \begin{cases} h_z\left(\left[n_\ell / P_z\right]+1\right)+\varphi, \varphi=\overline{1,\left(\left[n_\ell / P_z\right]+1\right)}, & \text{якщо}\quad 0\le h_z<b; \\ (P_z-h_d)\left(\left[n_\ell / P_z\right]+1\right)+\left(h_z-P_z+b\right)\left[n_\ell / P_z\right]+\varphi, \quad \varphi=\overline{1,\left[n_\ell / P_z\right]}, & \text{якщо}\quad b\le h_z\le P_z-1; \end{cases} \tag{8}$$

де a_{h_z} $h_z=\overline{0,P_z-1}$ – зростаюча послідовність, елементами якої є номери нейронів $\ell=\overline{2,L}$ шарів, що обробляються h_z-м обчислювачем на z-му рівні Z-мірної сітки $(z=\overline{0,Z-1})$; $b=n_\ell \bmod P_z$, $h_d=P_z-b$.

Для скорочення часу навчання БШМ і можливості обробки великого обсягу вхідних даних декомпозиція навчальної вибірки K здійснюється згідно з виразом

$$b_z = \begin{cases} z\left(\left[\dfrac{K}{Z}\right]+1\right)+\varphi, \varphi=\overline{1,\left(\left[\dfrac{K}{Z}\right]+1\right)}, & \text{якщо}\quad 0\le z<b; \\ (Z-h_d)\left(\left[\dfrac{K}{Z}\right]+1\right)+\left(z-Z+b\right)\left[\dfrac{K}{Z}\right]+\varphi, \varphi=\overline{1,\left[\dfrac{K}{Z}\right]}, & \text{якщо}\quad b\le z\le Z-1, \end{cases} \tag{9}$$

де b_z, $z=\overline{0,Z-1}$ – зростаюча послідовність, елементами якої є номери прикладів

навчальної вибірки, що обробляються на z-му рівні сітки; $b = K \bmod Z$ – номер рівня, після якого для $h_d = Z - b$ рівнів, починаючи з $(b+1)$-го, кількість оброблюваних прикладів зменшується на один для рівномірного завантаження.

Набула подальшого розвитку *модель прискореної нейрообробки даних*, яка адаптує нейрообчислення шляхом організації структури віртуальних зв'язків різних топологій передач даних між обчислювачами.

Модель прискореної нейрообробки даних з топологією передачі даних «зірка» подана у вигляді графа, вершинам якого відповідають незалежні паралельні операції, а спрямованим дугам – передачі даних між обчислювачами, спільно з розкладом:

Такти $1 \div 3L-1$. Головний процесор здійснює розсилку робочим обчислювачам $h = \overline{1, P-1}$ початкових значень елементів підматриць W_{ℓ_h} розмірністю $n_{\ell-1} \times |g_h|$ матриць $W_\ell = [W_{\ell_1}, W_{\ell_2}, \ldots, W_{\ell_{P-1}}]$, $\ell = \overline{2, L}$, визначуваних відповідно до наборів нейронів у послідовності g_h (6). Для поточного k-го прикладу навчальної вибірки, набувши від 0-го процесора значень виходів нейронів попереднього шару і цільових ознак $D(k) = [D_1(k), D_2(k), \ldots, D_{P-1}(k)]^T$, паралельно визначаються виходи g_h-х наборів нейронів $Y_\ell(k) = [Y_{\ell_1}(k), Y_{\ell_2}(k), \ldots, Y_{\ell_{P-1}}(k)]^T$, $\ell = \overline{2, L}$ на основі (1) на робочих обчислювачах $h = \overline{1, P-1}$ і повертаються для формування загального результату $Y_\ell(k)$ на головний комп'ютер.

Такти $3L \div (3L+2)$. Робочі обчислювачі $h = \overline{1, P-1}$ паралельно обчислюють локальні помилки $\Sigma_L(k) = [\Sigma_{L_1}(k), \Sigma_{L_2}(k), \ldots, \Sigma_{L_{P-1}}(k)]^T$ для g_h-х наборів нейронів вихідного шару відповідно до (2) і повертають набуті значення 0-му обчислювачу для формування загального вектора $\Sigma_L(k)$.

Такти $(3L+3) \div (6L-4)$. Робочі обчислювачі $h = \overline{1, P-1}$ паралельно визначають на основі (3) локальні помилки $\Sigma_\ell(k) = [\Sigma_{\ell_1}(k), \Sigma_{\ell_2}(k), \ldots, \Sigma_{\ell_{P-1}}(k)]^T$ g_h-х наборів нейронів прихованих шарів $\ell = \overline{L-1, 2}$ і повертають готові значення головному процесору для створення векторів $\Sigma_\ell(k)$.

Такти $(6L-3) \div (8L-6)$. Паралельно коректуються елементи підматриць $W_{\ell_h}(k+1)$ розмірністю $n_{\ell-1} \times |g_h|$ матриць $W_\ell(k+1)$ шарів $\ell = \overline{2, L}$ відповідно до (4), після чого робочі обчислювачі $h = \overline{1, P-1}$ повертають результати головному процесору, на якому формуються матриці $W_\ell(k+1)$.

Обмін даними між нейронами в структурі БШМ можна відобразити за допомогою топології передачі даних «повнозв'язний граф», на основі чого побудована модель прискореної нейрообробки даних у вигляді графа спільно з розкладом, що дозволяє зменшити обсяг переданих даних:

Такти $1 \div 2\mathrm{L}$**.** Головний процесор здійснює розсилку обчислювачам $\mathrm{h} = \overline{0, \mathrm{P}-1}$ початкових значень елементів підматриць $\mathrm{W}_{\ell_\mathrm{h}}$ розмірністю $\mathrm{n}_{\ell-1} \times |\mathrm{r_h}|$ матриць $\mathrm{W}_\ell = \left[\mathrm{W}_{\ell_0}, \mathrm{W}_{\ell_1}, \ldots, \mathrm{W}_{\ell_{\mathrm{P}-1}}\right]$ шарів $\ell = \overline{2, \mathrm{L}}$, для поточного k-го прикладу навчальної вибірки значень цільових ознак $\mathrm{D(k)} = \left[\mathrm{D}_0(\mathrm{k}), \mathrm{D}_1(\mathrm{k}), \ldots, \mathrm{D}_{\mathrm{P}-1}(\mathrm{k})\right]^\mathrm{T}$, визначуваних відповідно до наборів нейронів у послідовності $\mathrm{r_h}$ (7), а також значень виходів нейронів $\mathrm{Y}_1(\mathrm{k})$. Паралельно на $\mathrm{h} = \overline{0, \mathrm{P}-1}$ процесорах обчислюються виходи $\mathrm{r_h}$-х наборів нейронів $\mathrm{Y}_\ell(\mathrm{k}) = \left[\mathrm{Y}_{\ell_0}(\mathrm{k}), \mathrm{Y}_{\ell_1}(\mathrm{k}), \ldots, \mathrm{Y}_{\ell_{\mathrm{P}-1}}(\mathrm{k})\right]^\mathrm{T}$, $\ell = \overline{2, \mathrm{L}}$ на основі (1) і визначені часткові значення передаються кожному обчислювачу.

Такти $(2\mathrm{L}+1) \div (2\mathrm{L}+2)$**.** Обчислювачі $\mathrm{h} = \overline{0, \mathrm{P}-1}$ паралельно визначають локальні помилки $\Sigma_\mathrm{L}(\mathrm{k}) = \left[\Sigma_{\mathrm{L}_0}(\mathrm{k}), \Sigma_{\mathrm{L}_1}(\mathrm{k}), \ldots, \Sigma_{\mathrm{L}_{\mathrm{P}-1}}(\mathrm{k})\right]^\mathrm{T}$ для $\mathrm{r_h}$-х наборів нейронів вихідного шару відповідно до (2) і передають набуті часткові значення кожному обчислювачу.

Такти $(2\mathrm{L}+3) \div (4\mathrm{L}-2)$**.** Усі обчислювачі паралельно визначають на основі (3) локальні помилки $\Sigma_\ell(\mathrm{k}) = \left[\Sigma_{\ell_0}(\mathrm{k}), \Sigma_{\ell_1}(\mathrm{k}), \ldots, \Sigma_{\ell_{\mathrm{P}-1}}(\mathrm{k})\right]^\mathrm{T}$ $\mathrm{r_h}$-х наборів нейронів прихованих шарів $\ell = \overline{\mathrm{L}-1, 2}$ і розсилають готові часткові значення кожному обчислювачу $\mathrm{h} = \overline{0, \mathrm{P}-1}$ середовища.

Такти $(4\mathrm{L}-1) \div (6\mathrm{L}-4)$**.** Паралельно на $\mathrm{h} = \overline{0, \mathrm{P}-1}$ процесорах налаштовуються елементи підматриць $\mathrm{W}_{\ell_\mathrm{h}}(\mathrm{k}+1)$ розмірністю $\mathrm{n}_{\ell-1} \times |\mathrm{r_h}|$ матриць $\mathrm{W}_\ell(\mathrm{k}+1)$ шарів $\ell = \overline{2, \mathrm{L}}$ відповідно до (4), після чого дробові результати пересилаються кожному обчислювачу середовища.

Враховуючи особливості методу зворотного поширення помилки, побудована модель прискореної нейрообробки даних з топологією передачі даних «сітка» у вигляді графа спільно з розкладом, наведеним нижче, яка більш ефективно реалізує в розподіленому середовищі одночасне навчання на різних навчальних прикладах на окремих рівнях сітки і взаємодію сусідніх обчислювачів між рівнями, що скорочує кількість передач даних для прискорення часу обробки значного обсягу даних.

Такти $1 \div (2\mathrm{L}+3)$**.** Головний процесор $\mathrm{h}_0 = 0$ здійснює розсилку матриць вагових коефіцієнтів W_ℓ шарів $\ell = \overline{2, \mathrm{L}}$, для поточного прикладу з послідовності $\mathrm{b_z}$ (9), що обробляється на z-му рівні сітки, вектора значень цільових ознак $\mathrm{D}(\mathrm{b_z})$, виходів першого шару $\mathrm{Y}_1(\mathrm{b_z})$ обчислювачам $\mathrm{h_z} = 0$, $\mathrm{z} = \overline{1, \mathrm{Z}-1}$ по «лінійці». Обчислювачі, що керують, $\mathrm{h_z} = 0$, $\mathrm{z} = \overline{0, \mathrm{Z}-1}$ розподіляють початкові значення елементів підматриць $\mathrm{W}_{\ell_{\mathrm{h_z}}}$ розмірністю $\mathrm{n}_{\ell-1} \times |\mathrm{a}_{\mathrm{h_z}}|$ матриць $\mathrm{W}_\ell = \left[\mathrm{W}_{\ell_0}, \mathrm{W}_{\ell_1}, \ldots, \mathrm{W}_{\ell_{\mathrm{P_z}-1}}\right]$ шарів $\ell = \overline{2, \mathrm{L}}$, значення цільових ознак $\mathrm{D}(\mathrm{b_z}) = \left[\mathrm{D}_0(\mathrm{b_z}), \mathrm{D}_1(\mathrm{b_z}), \ldots, \mathrm{D}_{\mathrm{P_z}-1}(\mathrm{b_z})\right]^\mathrm{T}$, визначувані наборами нейронів у послідовності $\mathrm{a_{hz}}$ (8) між обчислювачами $\mathrm{h_z} = \overline{1, \mathrm{P_z}-1}$, $\mathrm{z} = \overline{0, \mathrm{Z}-1}$ по «лінійці».

Отримавши від процесорів одного рівня значення виходів нейронів попереднього шару, паралельно на робочих обчислювачах $h_z = \overline{0, P_z - 1}, \quad z = \overline{0, Z-1}$ визначаються виходи a_{h_z}-х наборів нейронів $Y_\ell(b_z) = \left[Y_{\ell_0}(b_z), Y_{\ell_1}(b_z), \ldots, Y_{\ell_{P_{z-1}}}(b_z)\right]^T$, $\ell = \overline{2, L}$ на основі (1) і обчислені часткові значення передаються кожному обчислювачу поточного рівня.

Такти $(2L+4) \div (2L+6)$**.** Паралельно визначаються відповідно до (2) обчислювачами $h_z = \overline{0, P_z - 1}, \quad z = \overline{0, Z-1}$ локальні помилки $\Sigma_L(b_z) = \left[\Sigma_{L_0}(b_z), \Sigma_{L_1}(b_z), \ldots, \Sigma_{L_{P_{z-1}}}(b_z)\right]^T$ для a_{h_z}-х наборів нейронів вихідного шару і передаються набуті часткові значення кожному обчислювачу одного рівня і на різних рівнях сітки.

Такти $(2L+7) \div 5L$**.** Обчислювачі одного рівня паралельно визначають на основі (3) локальні помилки $\Sigma_\ell(b_z) = \left[\Sigma_{\ell_0}(b_z), \Sigma_{\ell_1}(b_z), \ldots, \Sigma_{\ell_{P_{z-1}}}(b_z)\right]^T$ a_{h_z}-х наборів нейронів прихованих шарів $\ell = \overline{L-1, 2}$ і розсилають готові часткові значення кожному обчислювачу одного рівня і на різних рівнях сітки.

Такти $(5L+1) \div (6L-2)$**.** Кожним обчислювачем одного рівня визначаються усереднені локальні помилки $\Psi_\ell(b_z) = \left[\Psi_{\ell_0}(b_z), \Psi_{\ell_1}(b_z), \ldots, \Psi_{\ell_{P_{z-1}}}(b_z)\right]^T$ для a_{h_z}-х наборів нейронів шарів $\ell = \overline{2, L}$.

Такти $(6L-1) \div (8L-3)$. Паралельно кожним процесором одного рівня налаштовуються елементи підматриць $W_{\ell_{h_z}}(b_z + 1)$ розмірністю $n_{\ell-1} \times \left|a_{h_z}\right|$ матриць $W_\ell(b_z + 1)$ шарів $\ell = \overline{2, L}$ відповідно до (4), після чого дробові результати пересилаються кожному обчислювачу одного рівня.

Оцінювання обчислювальної складності методу РРНОД підтвердило, що порівняно з послідовним аналогом запропонований метод має меншу на один порядок обчислювальну складність і відповідно менші часові витрати (табл. 1).

Таблиця 1

Порівняльний аналіз послідовної та прискореної нейромережевої обробки даних

Функціонування БШМ	Навчання БШМ
Послідовний метод	
$O\left(\sum_{m=2}^{L} n_{m-1} n_m\right)$	$O(U \times (M \times \sum_{m=L-1}^{2} n_{m-1} n_m n_{m+1} + T \times \sum_{m=2}^{L} n_{m-1} n_m))$
Метод РРНОД з різними віртуальними топологіями мережі	
$O\left(\sum_{\substack{m_1=1 \\ r^{-1}(m_1) \neq 1 \\ r^{-1}(m_1) \neq L}}^{J} n_{r^{-1}(m_1)-1}\right)$	$O(U \times (M \times (\sum_{\substack{m_1=1 \\ r^{-1}(m_1) \neq 1 \\ r^{-1}(m_1) \neq L}}^{J} n_{r^{-1}(m_1)-1} n_{r^{-1}(m_1)+1} + n_L n_{L-1}) + T \times \sum_{\substack{m_1=1 \\ r^{-1}(m_1) \neq 1}}^{J} n_{r^{-1}(m_1)-1}))$

У табл. 1 введені такі позначення: J – номер елемента у перестановці, значення якого є граничним для умов $P < O\left(n_{r^{-1}(j+1)}\right)$ і $P \geq O\left(n_{r^{-1}(j)}\right)$, $j \in \{1,\ldots,L\}$, M – кількість прикладів, що обчислюється за допомогою виразу (9), або обсяг навчальної вибірки K в останніх випадках.

У третьому розділі вперше запропоновано метод МОС і модель ОПНФ БШМ у розподіленому обчислювальному середовищі.

Запропоновано *метод масштабування обчислювальної системи*, який дозволяє оцінити продуктивність подальшого підвищення потужності обчислювальної системи шляхом визначення прискорення розподіленої процедури з урахуванням загальної кількості скалярних операцій і втрат часу при неефективному розподілі паралельних операцій на кожен процесор у гетерогенному і гомогенному обчислювальному середовищі для прискорення рішення задач великої розмірності.

Метод масштабування обчислювальної системи має вигляд

$$S(P) = \frac{Q \times t^0}{\max\limits_{h=0,P-1}\left(Q^0_{\text{пар}} \times t^0, Q^1_{\text{пар}} \times t^1, \ldots, Q^h_{\text{пар}} \times t^h\right) + \max\limits_{h=0,P-1}\left(Q_{\text{пос}} t^h\right) + Q_c \times t_c}, \qquad (10)$$

де $Q^h_{\text{пар}}$, $h = \overline{0,P-1}$ – кількість скалярних паралельних операцій підзадач, що виконуються на h-му обчислювачі; Q – загальна кількість скалярних операцій у послідовній процедурі; $Q_{\text{пос}}$ – кількість скалярних послідовних операцій у розподіленій процедурі; Q_c – загальна кількість викликів методів передач даних; t^h, $h = \overline{0,P-1}$ – час виконання однієї обчислювальної операції h-го обчислювача, від якого залежить продуктивність розподілених нейрообчислень для гетерогенного середовища, сек; t_c – час однієї передачі даних, сек.

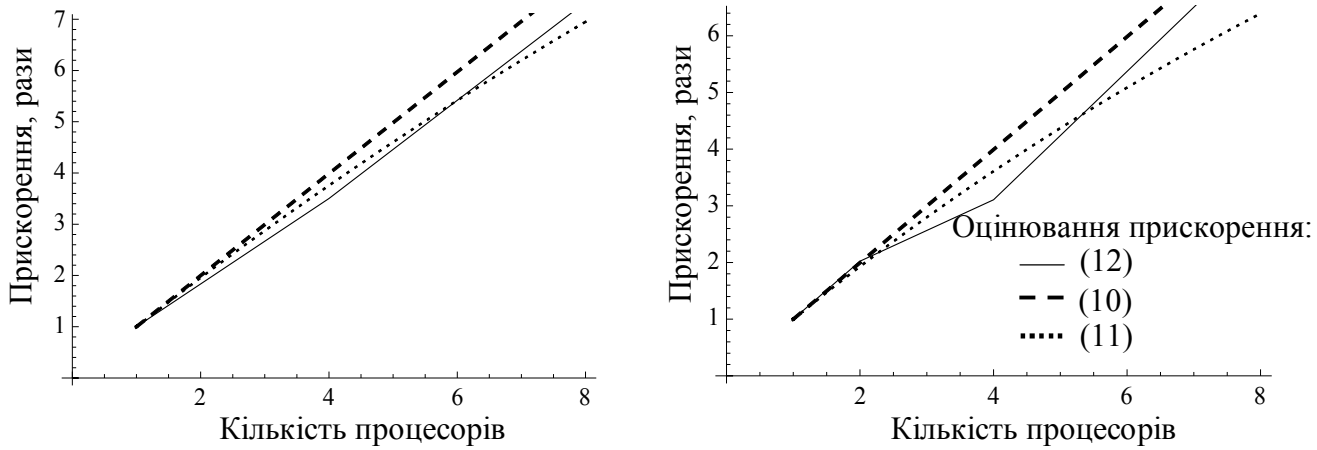
На тестових завданнях матрично-векторного множення (матриця розмірністю 5000×5000) і матричного множення (матриці розмірністю 3000×3000) стрічкової схеми розподілу даних по рядках проведений порівняльний аналіз методу МОС (10), традиційного закону Амдала, що розраховує загальну кількість скалярних операцій розподіленої процедури Q_p на підставі значення $Q_{\text{пос}}$ на основі

$$S(P) = \frac{1}{a + \frac{1-a}{P} + c}, \qquad (11)$$

де $a = \frac{Q_{\text{пос}}}{Q}$ – питома вага послідовних операцій у розподіленій процедурі; $c = \frac{Q_g t_c}{Qt}$ – коефіцієнт мережної деградації обчислень; Q_g – загальна кількість передач даних, біт; t – час виконання однієї скалярної операції, сек, і методу оцінювання прискорення на основі співвідношення часу виконання послідовної T і розподіленої T_p процедур

$$S(P) = \frac{T}{T_p}. \qquad (12)$$

Порівняльний аналіз, результати якого подані на рис. 1, показав точнішу відповідність експериментальним даним запропонованого методу (10) під час визначення прискорення розподіленої обчислювальної процедури, оскільки значення теоретичного прискорення є оцінкою зверху, а закон Амдала показує навіть на малому обчислювальному середовищі пониження графіку прискорення порівняно з експериментальними даними.



а б

Рис. 1. Залежності прискорення вирішення задачі від кількості обчислювачів: а – матрично-векторне множення; б – матричне множення

Для визначення прискорення S побудована *модель оцінювання прискорення* розподіленого навчання L-шарової нейронної мережі

$$S\left(U,K,T,L,P,h,n_1,\ldots,n_L,t^h,t_c\right)=\frac{Q\left(U,K,T,L,n_1,\ldots,n_L,t^h\right)}{Q_p\left(U,K,T,L,n_1,\ldots,n_L,h,t^h\right)+Q_c\left(U,K,T,L,P,t_c\right)}$$

та розподіленого функціонування БШМ

$$S\left(L,P,h,n_1,\ldots,n_L,t^h,t_c\right)=\frac{Q\left(L,n_1,\ldots,n_L,t^h\right)}{Q_p\left(L,n_1,\ldots,n_L,h,t^h\right)+Q_c\left(L,P,t_c\right)}.$$

Обчислення загальної кількості паралельних скалярних операцій Q_p при паралелізмі процедури на рівні реалізації функцій нейронів здійснюється за допомогою виділення тактів обробки незалежних операцій, якими представлені функції обчислення виходу нейрона, помилки реакції мережі і налаштування вагових коефіцієнтів, а скалярні операції для обчислення перерахованих функцій виконуються послідовно на окремих обчислювачах. Тоді для реалізації розподіленого навчання БШМ потрібна загальна кількість скалярних операцій відповідно до виразу

$$\begin{aligned}Q_p = Q_u + Q_{пос} = U\times(M\times(\sum_{m=2}^{L}\max_{h=0,P-1}(V_h\left(2n_{m-1}+3\right)t^h)+\\ +8\max_{h=0,P-1}(V_h n_{L-1}t^h)+\sum_{m=L-1}^{2}\max_{h=0,P-1}(V_h\left(2n_{m+1}+7\right)n_{m-1}t^h))+\\ +T\times(\sum_{m=2}^{L}\max_{h=0,P-1}(V_h\left(2n_{m-1}+3\right)t^h)))+\max_{h=0,P-1}\left(Q_{пос}t^h\right),\end{aligned} \qquad (13)$$

для розподіленого функціонування БШМ –

$$Q_p = Q_u + Q_{пос} = \sum_{m=2}^{L} \max_{h=\overline{0,P-1}} (V_h (2n_{m-1} + 3) t^h) + \max_{h=\overline{0,P-1}} (Q_{пос} t^h),$$

де V_h – набір нейронів, визначуваний за допомогою методу (6)-(8); $P = P_z$ під час використання топології передачі даних «сітка».

Для визначення кількості скалярних непаралельних операцій у розподіленій процедурі $Q_{пос}$ використовується різниця між загальною кількістю скалярних операцій Q у послідовній процедурі і кількістю незалежних скалярних операцій у розподіленій процедурі Q_p, отримані шляхом збільшення кількості обчислювачів, що прагнуть до деякого граничного значення, яке доцільно використовувати в розрахунках.

Загальний час передач даних $Q_{c_i}, i = \overline{1,3}$ під час реалізації розподіленого навчання БШМ з топологією передач даних «зірка» визначається відповідно до виразу

$$Q_{c_1} = U \times (K \times (6PL - 5P - 6L + 5) + T \times (2PL - 2P - 2L + 2)) \times t_c, \qquad (14)$$

з топологією передач даних «повнозв'язний граф» як

$$Q_{c_2} = U \times (K \times (3P^2L - 3P^2 - 2PL + 4P - L - 1) + T \times (P^2L - P^2 - PL + 2P - 1)) \times t_c, \qquad (15)$$

з топологією передач даних «сітка» –

$$\begin{aligned} Q_{c_3} = U \times (M \times (3ZP_z^2L - 3ZP_z^2 - ZP_zL + 3ZP_z - L - 1) + \\ + T \times (P_z^2L - P_z^2 - P_zL + 2P_z - 1)) \times t_c. \end{aligned} \qquad (16)$$

За аналогією з попереднім під час реалізації процедур розподіленого функціонування БШМ потрібен загальний час передач даних

$$Q_{c_{f_1}} = (2PL - 2P - 2L + 2) \times t_c,$$

$$Q_{c_{f_2}} = (P^2L - P^2 - PL + 2P - 1) \times t_c,$$

$$Q_{c_{f_3}} = (P_z^2L - P_z^2 - P_zL + 2P_z - 1) \times t_c.$$

Таким чином, найбільша кількість передач даних здійснюється під час використання топології «повнозв'язний граф», під час реалізації топології передачі даних «сітка» кількість передач у декілька разів менше передач даних порівняно з топологією «повнозв'язний граф», а мінімальна кількість передач досягається під час застосування топології передачі даних «зірка».

На основі моделі ОПНФ отримані найбільш ефективні моделі БШМ для вирішення практичних завдань. На рис. 2 наведені графіки прискорення, отримані за допомогою моделей оцінювання прискорення навчання БШМ на основі методу МОС (10), методу оцінювання прискорення (12) та традиційного закону Амдала (11), вирішення задачі класифікації рукодрукованих символів за допомогою БШМ структури 2500-120-33 з використанням навчальної вибірки обсягом 650000 прикладів на $P = \overline{1,28}$ обчислювачах розподіленого середовища з топологіями передачі даних «зірка», «повнозв'язний граф», «сітка».

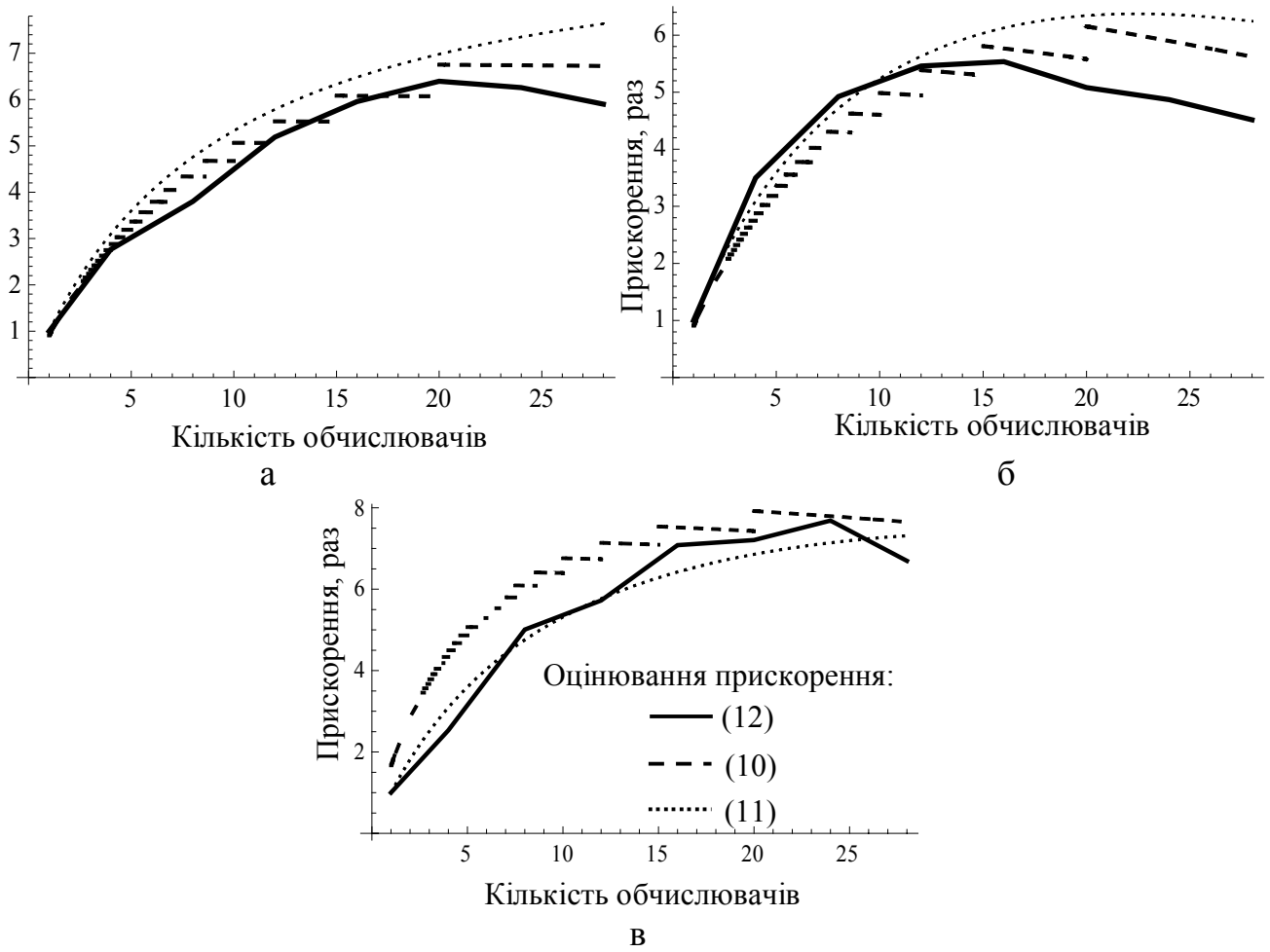


Рис. 2. Залежність прискорення розподіленого навчання БШМ від кількості обчислювачів: а – топологія передачі даних «зірка»; б – топологія передачі даних «повнозв'язний граф»; в – топологія передачі даних «сітка»

Порівняння рис. 2а, 2б та 2в показує, що модель з різними топологіями (5,13,14), (5,13,15), (5,13,16) дозволяє побудувати точну залежність прискорення розподіленого навчання БШМ від кількості обчислювачів, крім того при малій кількості процесорів залежності прискорення розподіленого навчання БШМ співпадають, проте при збільшенні обчислювачів у розподіленому середовищі – розрізняються в рази. Це підтверджує достовірність та доцільність використання запропонованих моделей ОПНФ L-шарової нейронної мережі.

У **четвертому розділі** для проведення експериментальних досліджень і вирішення практичних задач розроблено та реалізовано інструментальне середовище, яке забезпечує засоби завантаження вибірок, моделювання продуктивності БШМ та її адаптації на високопродуктивне середовище за допомогою технологій паралельного і розподіленого програмування.

Для проведення експериментальних досліджень використовувалася комп'ютерна мережа Fast Ethernet з фізичною топологією «зірка» зі швидкістю передачі 100 Мбіт, що складається з групи $P = \overline{1,7}$ чотириядерних обчислювачів Intel Core2 Quad Q8200 2,33 ГГц.

Порівняльний аналіз застосування різних реалізацій інтерфейсу передач даних на $P = 20$ обчислювачах показав найбільше прискорення розподіленого навчання

БШМ структури 2500-120-33 з топологією передачі даних «зірка» у 7 разів під час використання пакета Compute Cluster Pack, оскільки технологія MPI 2.0 дозволяє прискорити вирішення даної задачі тільки в 3 рази, а спільне використання бібліотеки MPI 2.0 і технології PLINQ – в 4 рази.

Розпаралелювання за допомогою бібліотеки MS MPI із застосуванням опції Full Optimization компілятора Intel® C++ Compiler 10.1 під керуванням операційної системи Microsoft Windows Compute Cluster Server 2003 показало прискорення розподіленого навчання БШМ структури 2500-120-33 з використанням навчальної вибірки обсягом 650000 прикладів, 4225-245-33 з вибіркою обсягом 1050000 прикладів, 10000-455-33 з вибіркою із 2460000 прикладів відповідно у 8, 7, 9 разів для топології передач даних «зірка», у 11, 10, 12 разів для топології передач даних «повнозв’язний граф», у 18, 16, 18 разів для топології передач даних «сітка». Розроблені рекомендації щодо найбільш доцільної адаптації БШМ у доступному розподіленому обчислювальному середовищі.

Розроблені в роботі методи та моделі були використані для розв'язання практичних нейромережевих задач: розбракування труб і прогнозування екологічної обстановки. Достовірність результуючих даних перевірена на тестових вибірках, отримане відхилення не перевищує допустимого значення (0,38%).

ВИСНОВКИ

У дисертаційній роботі наведені результати, які, відповідно до мети дослідження, в сукупності є вирішенням актуальної науково-практичної задачі – розробки методів і моделей прискореного навчання та функціонування багатошарової нейронної мережі прямого поширення у розподіленому обчислювальному середовищі, що має велике значення для суттєвого скорочення часу вирішення задач великої розмірності в доступному розподіленому середовищі з різними топологіями передачі даних. Отримані наукові результати:

1. Новий метод рівномірного розподілу нейромережевої обробки даних, який поетапно та рівномірно завантажує обчислювачі роботою шляхом автоматичного розподілу оброблюваних даних у задачах великої розмірності під час реалізації БШМ з топологіями передачі даних «зірка», «повнозв’язний граф» і «сітка» в розподіленому середовищі, що дозволяє значно скоротити час навчання та функціонування БШМ.

2. Нова модель оцінювання прискорення навчання та функціонування БШМ за рахунок вибору ефективних значень як часових параметрів виконання нейропроцедури, так і апаратних характеристик мережі та обчислювачів дозволяє підвищити продуктивність нейрообчислень та значно прискорити обробку великого обсягу даних у розподіленому середовищі.

3. Новий метод масштабування обчислювальної системи за рахунок визначення прискорення розподіленої нейропроцедури на підставі загальної кількості скалярних операцій та врахування часу їх виконання надає можливість ефективно використовувати доцільну кількість обчислювачів. Це дозволяє прискорити вирішення обчислювальних задач.

4. Набула подальшого розвитку модель прискореної нейрообробки даних, яка представлена у вигляді графа спільно з розкладом розподіленого потактового завантаження обчислювачів у високопродуктивному середовищі з різними віртуальними топологіями передач даних. Це дозволяє скоротити кількість передач, які суттєво впливають на загальний час вирішення задач великої розмірності.

5. Проведено порівняльний аналіз обчислювальної складності методів послідовного, а також розподіленого навчання та функціонування БШМ з топологіями мережі передачі даних «сітка», «повнозв'язний граф», «зірка». Обчислювальна складність методу РРНОД з різними топологіями зменшена на порядок порівняно з існуючим послідовним методом.

6. Проведено експериментальні дослідження запропонованих методів і моделей, які показали значне скорочення часу їх виконання та високу ефективність. На основі процедур розподіленого та паралельного навчання БШМ розроблено і реалізовано інструментальне середовище із застосуванням високопродуктивних технологій MPI, MS MPI і PLINQ. Результати роботи були впроваджені та показали свою ефективність для розв'язання задачі розбракування труб, а також для прогнозування екологічної обстановки.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Аксак Н.Г. Распознавание изображений антропоморфного объекта / Н.Г. Аксак, А.Ю. Тыхун, О.Ю. Барковская, А.С. Солдатов // Біоніка інтелекту: наук.-техн. журнал. – Харків: ХНУРЕ, 2009. – №1(70). – С.102-105.

2. Аксак Н.Г. Система параллельно-распределенного экспертного оценивания / Н.Г. Аксак, А.Ю. Лебёдкина, М.В. Кушнарёв // Вісник ХНТУ. – Херсон: ХНТУ, 2011. – №2 (41). – С. 403-407.

3. Аксак Н.Г. Методы и модели производительности обучения многослойных нейронных сетей в распределенных компьютерных средах / Н.Г. Аксак, А.Ю. Лебёдкина //Науково-теоретичний журнал «Штучний інтелект». – Донецьк: ІПШІ МОН і НАН України «Наука і освіта», 2011. – Випуск 4. – С.481-488.

4. Аксак Н.Г. Анализ моделей производительности реализации нейроалгоритма / Н.Г. Аксак, А.Ю. Тыхун // Системи обробки інформації: зб. наук. пр. – Харків: ХУПС імені Івана Кожедуба, 2008. – Вип.1 (68). – С. 2-5.

5. Аксак Н.Г. Модели производительности нейроалгоритма, адаптированного на высокопроизводительные системы / Н.Г. Аксак, А.Ю. Тыхун // Питання прикладної математики і математичного моделювання: зб. наук. пр. – Д.: Вид-во Дніпропетр. нац. ун-ту, 2009. – С. 29-38.

6. Аксак Н.Г. Процедура параллельного обучения многослойной нейронной сети. Топология передачи данных «звезда» / Н.Г. Аксак, А.Ю. Лебёдкина, О.В. Хоменко // Науковий вісник Чернівецького національного університету імені Юрія Федьковича. Серія: Комп'ютерні системи та компоненти. – Чернівці: ЧНУ, 2010. – Том 1, випуск 2. – С. 95-103.

7. Аксак Н.Г. Метод равномерного распределения параллельных операций ускоренного обучения многослойной нейронной сети с различными топологиями передач данных / Н.Г. Аксак, А.Ю. Лебёдкина // Збірник наукових праць «Системи

управління, навігації та зв'язку. – Київ: ДЦ «Центральний науково-дослідний інститут навігації і управління», 2011. – Випуск 2(18). – С. 66-73.

8. Аксак Н.Г. Сравнительный анализ программной реализации параллельных алгоритмов / Н.Г. Аксак, А.Ю. Тыхун, А.С. Солдатов // Сборник трудов конференции «Моделирование-2008». – Киев: Институт проблем моделирования в энергетике им. Пухова НАН Украины, 2008. – Том 2. – С. 377-382.

9. Аксак Н.Г. Вычислительная модель нейроалгоритма многослойного персептрона / Н.Г. Аксак, А.Ю. Тыхун // Материалы Седьмой Международной конференции-семинара «Высокопроизводительные параллельные вычисления на кластерных системах». – Нижний Новгород: Изд-во Нижегородского госуниверситета, 2007. – С. 11-18.

10. Аксак Н.Г. Анализ эффективности применения вычислительной модели многослойного персептрона / Н.Г. Аксак, А.Ю. Тыхун // Материалы Седьмой Международной научно-технической конференции «Проблемы информатики и моделирования – 2007». – Харьков: НТУ "ХПИ", 2007. – С. 30.

11. Аксак Н.Г. Модели производительности синтеза нейроалгоритма на высокопроизводительные архитектуры / Н.Г. Аксак, А.Ю. Тыхун, И.В. Новосельцев // Материалы Восьмой Международной конференции «Высокопроизводительные параллельные вычисления на кластерных системах (HPC-2008)». – Казань: Изд. КГТУ, 2008. – С. 199-202.

12. Аксак Н.Г. Сравнительный анализ программных реализаций параллельного алгоритма с применением технологии MPI и стандарта OpenMP / Н.Г. Аксак, А.Ю. Тыхун, А.С. Солдатов // Материалы Всероссийской конференции «Технологии Microsoft в теории и практике программирования». – Нижний Новгород: Изд-во Нижегородского госуниверситета, 2008. – С. 340-343.

13. Тыхун А.Ю. Адаптация нейроалгоритма для MPP-систем / А.Ю. Тыхун // Материалы XII Международного молодежного форума «Радиоэлектроника и молодежь в XXI веке». – Харьков: ХНУРЭ, 2008. – Ч.2. – С. 117.

14. Аксак Н.Г. Анализ способов ускорения нейроалгоритма на основе технологий параллельного программирования / Н.Г. Аксак, А.Ю. Тыхун, И.В. Новосельцев // Матеріали X Міжнародної науково-технічної конференції «Системний аналіз та інформаційні технології». – Київ: НТУУ „КПІ", 2008. – С. 284.

15. Аксак Н.Г. Применение библиотеки MPI для решения задач распознавания образов на основе нейронных сетей / Н.Г. Аксак, А.Ю. Тыхун, А.С. Солдатов // Материалы Международной молодежной научной конференции «XXXIV Гагаринские чтения». – Москва: МАТИ, 2008. – С. 238-239.

16. Аксак Н.Г. Сравнительный анализ производительности нейроалгоритма с применением реализаций MPICH и MS-MPI / Н.Г. Аксак, А.Ю. Тыхун, А.С. Солдатов, В.А. Радченко // Материалы конференции «Технологии Microsoft в теории и практике программирования» – Нижний Новгород: Изд-во Нижегородского госуниверситета, 2008. – С. 14-18.

17. Аксак Н.Г. Анализ производительности нейроалгоритма на высокопроизводительных системах / Н.Г. Аксак, К.А. Лавриненко, А.Ю. Лебёдкина // Матеріали першої міжнародної науково-технічної конференції

«Інформаційні технології в навігації і управлінні: стан і перспективи розвитку». – К.: ДП “ЦНДІ НіУ”, 2010. – С. 62.

18. Лебёдкина А.Ю. Компьютерная система контроля качества выпускаемой продукции с помощью нейронных сетей / А.Ю. Лебёдкина, Д.Н. Росинский // Матеріали першої науково-технічної конференції «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління». – Х.: ДП “ХНДІ ТМ”, К.: ДП “ЦНДІ НіУ”, 2010. – С. 84.

19. Лебёдкина А.Ю. Повышение эффективности нейровычислений с помощью параллельной процедуры обучения / А.Ю. Лебёдкина, О.В. Хоменко // Сб. материалов 15-го Юбилейного Международного молодежного форума «Радиоэлектроника и молодежь в XXI веке». – Харьков: ХНУРЭ, 2011. – Т.9. – С. 658-659.

20. Лебёдкина А.Ю. Методы равномерного распределения параллельных операций распределенных процедур и оценивания их эффективности / А.Ю. Лебёдкина // Матеріали другої міжнародної науково-технічної конференції «Інформаційні технології в навігації і управлінні: стан і перспективи розвитку». – К.: ДП “ЦНДІ НіУ”, 2011. – С. 37.

21. Аксак Н.Г. Методы ускоренной реализации многослойной нейронной сети / Н.Г. Аксак, А.Ю. Лебёдкина // Сборник научных статей по итогам Международной научно-практической конференции «Информатика, математическое моделирование, экономика». – Смоленск: Смоленский филиал АНО ВПО ЦС РФ «Российский университет кооперации», 2011. – Том 1. – С. 71-78.

АНОТАЦІЯ

Лебьодкіна А.Ю. Методи та моделі прискореної нейромережевої обробки даних у розподіленому обчислювальному середовищі. – На правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки Міністерства освіти і науки, молоді та спорту України, Харків, 2012.

Дисертаційна робота присвячена розробці методів і моделей прискореного навчання та функціонування багатошарової нейронної мережі прямого поширення, які здатні істотно скоротити час вирішення завдань великої розмірності в розподіленому середовищі.

Метод рівномірного розподілу нейромережевої обробки даних здійснює диспетчерування обчислювальної системи під час реалізації нейронів кожного шару.

Запропонована модель прискореної нейрообробки даних у розподіленому середовищі дозволяє адаптувати нейрообчислення БШМ з різними топологіями на високопродуктивну архітектуру.

Метод масштабування обчислювальної системи шляхом визначення прискорення розподіленої нейропроцедури з урахуванням загальної кількості скалярних операцій і втрат часу на виконання розподілених операцій дозволяє оцінити продуктивність подальшого підвищення потужності гетерогенного або

гомогенного обчислювального середовища.

Запропонована модель оцінювання прискорення навчання та функціонування БШМ з різними топологіями сукупно враховує часові параметри виконання нейропроцедури та апаратні характеристики мережі, що дозволяє підвищити продуктивність паралельних нейрообчислень під час вирішення завдань з великим обсягом вхідних даних.

Ключові слова: багатошарові нейронні мережі, метод зворотного поширення помилки, паралельні та розподілені обчислення, прискорення, продуктивність, масштабування, топологія мережі передачі даних.

АННОТАЦИЯ

Лебёдкина А.Ю. Методы и модели ускоренной нейросетевой обработки данных в распределенной вычислительной среде. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 05.13.23 – системы и средства искусственного интеллекта. – Харьковский национальный университет радиоэлектроники Министерства образование и науки, молодежи и спорта Украины, Харьков, 2012.

Диссертационная работа посвящена разработке методов и моделей ускоренного обучения и функционирования многослойной нейронной сети прямого распространения (МНС), которые за счет диспетчеризации и масштабирования вычислительной системы, а также адаптации и повышения производительности при реализации МНС обеспечивают существенное сокращение времени решения задач большой размерности.

В результате проведенного анализа было выявлено, что существующие методы ускорения обработки данных многослойной нейронной сетью имеют недостатки, связанные с неэффективным использованием вычислительных ресурсов, большим количеством передач данных и трудностями при обработке большого объема входных данных.

Впервые предложенный метод равномерного распределения нейросетевой обработки данных осуществляет диспетчерование вычислительной системы при реализации нейронов каждого слоя с различными топологиями передачи данных, а также позволяет снизить на один порядок вычислительную сложность процедур обучения и функционирования МНС и существенно сократить время обучения и функционирования МНС.

Получила дальнейшее развитие модель ускоренной обработки данных МНС в распределенной среде с виртуальными топологиями «звезда», «решетка», «полносвязный граф», позволяющая адаптировать нейровычисления МНС на высокопроизводительные архитектуры путем организации структуры виртуальных связей топологий передач данных между вычислителями.

Для ускорения решения задач большой размерности впервые предлагается метод масштабирования вычислительной системы, который позволяет оценить продуктивность последующего повышения мощности вычислительной системы путем вычисления ускорения распределенной нейропроцедуры с учетом общего количества скалярных операций и потерь времени при обработке больших объемов

данных, распределенных на каждый процессор в гетерогенной и гомогенной вычислительной среде.

Впервые предложенная модель оценивания ускорения обучения и функционирования МНС совокупно учитывает временные параметры выполнения нейропроцедуры с виртуальными топологиями, аппаратные характеристики среды и вычислителей, что позволяет повысить производительность параллельных нейровычислений посредством выбора эффективных значений и существенно ускорить обработку большого объема данных в распределенной вычислительной среде.

Проведенные вычислительные эксперименты подтвердили, что распределенная нейросетевая обработка данных позволяет значительно ускорить решение больших задач.

Результаты работы использованы при решении нейросетевых задач разбраковки бесшовных труб различного назначения по качеству и прогнозирования экологической обстановки.

Ключевые слова: многослойные нейронные сети, метод обратного распространения ошибки, параллельные и распределенные вычисления, ускорение, производительность, масштабирование, топология сети передачи данных.

ABSTRACT

Lebodkina A.Y. Methods and models for speed up neural data processing in the distributed computing environment. – On the rights of the manuscript.

The dissertation for candidate's degree in engineering science by specialty 05.13.23 – systems and methods of artificial intelligence. – Kharkiv National University of Radio Electronics of the Ministry of Education and Science, Youth and Sports of Ukraine, Kharkiv, 2012.

The dissertation work is devoted to methods and models for speed up training and functioning of the multilayered feedforward neural network (MFNN) development, that allows to significantly reduce time of the large dimension tasks solving in a distributed environment.

The method of equable processed neural data distribution allows assigning neurons of each layer in MFNN dynamically to the processors in the distributed environment.

The proposed model for speed up data processing allows adopting MFNN structure with different virtual topologies to the high-performance computing system.

The method of computing system scaling allows estimating computing systems performance by determining speedup of distributed neural processing based on general amount of scalar operation and time loss of parallel operations execution.

The model of MFNN training and functionality speedup estimation are proposed, that allows to increase the productivity of neural data processing in the large dimensional tasks. This model uses timing parameters of procedures execution and hardware environment characteristics.

Key words: multilayer neural networks, backpropagation method, parallel and distributed computing, speedup, performance, scalability, virtual topologies.

Відповідальний випусковий В.П. Машталір

Підп. до друку 11.10.2012. Формат 60х84 1/16. Спосіб друку – ризографія.
Умов. друк. арк. 1,2. Облік. вид. арк. 1,1. Тираж 100 прим.
Ціна договірна Зам. №

ХНУРЕ. Україна. 61166, Харків, просп. Леніна, 14

Віддруковано в навчально-науковому
видавничо-поліграфічному центрі ХНУРЕ.
61166, Харків, просп. Леніна, 14