УДК 004.89 **doi: 10.26906/SUNZ.2024.3.046**

Ю. І. Олімпієва

Державний університет інформаційно-комунікаційних технологій, Київ, Україна

ЗАБЕЗПЕЧЕННЯ ФУНКЦІОНАЛЬНОЇ СТІЙКОСТІ ВИРОБНИЧИХ ПРОЦЕСІВ ПРОМИСЛОВИХ ПІДПРИЄМСТВ НА ОСНОВІ НЕЙРОАДАПТИВНОЇ СИСТЕМИ

Анотація. У статті розглядається проблема забезпечення функціональної стійкості виробничих процесів промислових підприємств на основі нейроадаптивних систем. В умовах індустрії, де виробничі процеси стають дедалі складнішими та автоматизованими, важливість надійних і стійких систем управління зростає. Проводиться аналіз сучасних підходів до використання нейронних мереж для адаптивного управління виробничими процесами, що дозволяє підвищити їхню ефективність та надійність. Аналізуються різні методи підвищення функціональної стійкості, включаючи адаптивні алгоритми навчання, регуляризацію та техніки виявлення і корекції помилок. Особлива увага приділяється інтеграції нейроадаптивних систем з існуючими технологіями та виробничими лініями, а також їх здатності до швидкого відновлення після збоїв і адаптації до змінних умов експлуатації. На основі проведеного аналізу запропоновано новий алгоритм діагностування на основні нейроадаптивної системи, що сприятиме підвищенню стійкості та надійності виробничих процесів. При проектуванні апаратної реалізації нейромереж враховуються такі характеристики нейрочіпів, як масштабованість, вартість розробки, сумісність з минулими та майбутніми версіями. Тому серед існуючих нейрочіпів було обрано Google TPU v4, оскільки у нього високий коефіцієнт продуктивності, а фреймворки машинного навчання, такі як TensorFlow та PyTorch, надають набір операцій для управління мережевим взаємодією вузлів системи. Важливо зазначити, що нейронні мережі можуть використовуватися в комплексі з іншими методами та системами діагностування, такими як експертні системи, фізичні моделі, статистичні методи.

Ключові слова: функціональна стійкість, діагностування, нейронні мережі, нейроадаптивні системи, алгоритм, нейрочіпи, виробничий процес, промислові підприємства.

Вступ

Більшість інформаційних систем, які використовуються сьогодні в різноманітних сферах життя, є складними. Вони утворюють багаторівневі конструкції, які не можна описати звичайною сумою взаємодії її елементів. Такі складні технічні системи створюються для розв'язання спеціальних задач. Стрімкий розвиток таких систем приводять до ускладнення їх функціонування та виникнення непередбачуваних збоїв. Тому підтримка безперервної роботи складних технічних систем, виявлення та локалізація різних несправностей набуває все більшого значення.

Для забезпечення автономного функціонування системи можна досягти за допомогою властивості функціональної стійкості [1]. Дана властивість забезпечує надійну роботи системи, незважаючи на різноманітні потоки відмов і несправності протягом часу її роботи.

Постановка завдання. Поняття функціональної стійкості вперше було введено науковцем Машковим О.А. в 1990-х роках, які на той час розв'язував конкретні прикладні задачі. Інші науковці, такі як Барабаш О.В., Собчук В.В. та ін., продовжують вести дослідження в даному напрямі [2-3]. У своїх роботах вони описують методи забезпечення функціональної стійкості для вдосконалення технічних характеристик складних технічних систем, що працюють в екстремальних умовах. Також приділяється увага зменшенню кількості порушень при роботі систем завдяки вчасному діагностуванню таких збоїв та перерозподілі задач між обчислювальними вузлами [4-6]. Однак розвиток елементної бази обчислювальних систем, складності сучасних і, особливо, перспективних автономних динамічних систем, дозволяє розширити область застосування методів забезпечення функціональної стійкості. У сучасних умовах дуже активно розвивається область нейронауки, де вже представлено багато різних методів для вирішення різних інформаційно-технологічних проблем, серед яких одним з найбільш ефективних є штучні нейронні мережі. Тому для забезпечення високого рівня функціональної стійкості потрібно створювати нові методи або вдосконалювати існуючі. Нейронні мережі можуть бути ефективним інструментом, який дозволить створити глибоку ієрархію прийняття рішень з врахуванням рекомендацій, отриманих з моніторингу роботи виробничих процесів у системі. Буде розглянуто особливості основних положень теорії штучного інтелекту, а саме нейронних мереж, з порівняльними характеристиками апаратної реалізації нейромереж.

Аналіз останніх досліджень і публікацій. Забезпеченню функціональної стійкості присвячено низка робіт. В роботі [7] автори описують початковий етап забезпечення функціональної стійкості, а саме контроль стану інформаційної системи, і якщо є відмова, то включається самодіагностування системи. У [8] науковці розробили методологію побудови ефективної системи самодіагностики інформаційних систем на прикладі підприємств металургійної та енергетичної промисловості, за результатами сформульовано критерії достатності діагностичної інформації за відсутності обмежень щодо виконання елементарних перевірок. У статті [9] розглядається методологія забезпечення функціональної стійкості об'єктів критичної інфраструктури через представлення функціонування системи у вигляді формалізованого процесу. У [10] розкрито механізми самоорганізації інформаційних неоднорідних мереж, проаналізовано нові показники та критерії визначення функціонально стійких систем, запропоновано мережі відповідно до концепції SON. Розглянуто

особливості застосування методу обернених задач динаміки для керування відновленням, отримано вираз для керуючої сили та проведено моделювання для етапу визначення керуючої сили [11]. Для об'єктів критичної інфраструктури запропоновано метод побудови закону управління безпекою критичних об'єктів інфраструктури в умовах зовнішніх неконтрольованих впливів, а саме метод побудови надійної системи управління, що дозволяє компенсувати неконтрольовані зовнішні впливи [12]. У розглянутих роботах для забезпечення функціональної стійкості не використовувалися нейронні мережі

У статті [13] вже розглянуто особливості основних положень теорії штучного інтелекту, а саме нейронних мереж для забезпечення функціональної стійкості виробничих процесів промислових підприємств, досліджено можливості застосування нейронних мереж для діагностики стану систем та практичного застосування інструментарію нейронних мереж для виявлення та локалізації дефектів в роботі систем, що є запорукою забезпечення функціональної стійкості виробничих процесів підприємства. Також удосконалено методику забезпечення властивості функціональної стійкості інформаційної системи підприємства. У статті [14] запропоновано новий підхід до визначення технічного стану мікропроцесорних систем в базисі програмно-реконфігурованої логіки, які є основою термінальних компонентів технологічної системи. З метою автоматизації системи контролю та діагностування мікропроцесорних систем запропоновано реалізувати принцип самодіагностування, в основу якого покладено ідеї штучного інтелекту. У [15] досліджуються процеси глобальної трансформації інфраструктури інформаційних технологій на фоні масового впровадження кіберфізичних систем та проривних здобутків в галузях штучного інтелекту і робототехніки у виробництво та технологічні процеси. Описано способи застосування нейромереж в моделюваннях процесів механічної обробки металів різанням. Дано універсальну методику побудови нейромережевих моделей процесу механічної обробки на базі штучної нейронної мережі зустрічного поширення. У [16] досліджується стійкість класу нейронних мереж із затримкою, що змінюється в часі. Для GSC мережевих мереж, коли кількість змінних рішення залишається незмінною, інформація про затримку часу та її похідні додатково викопується за допомогою підходу розподілу затримки. На основі цих методів наведено покращені критерії стабільності.

Детальний аналіз останніх досліджень і результатів досліджень вказує на те, що задача забезпечення функціональності стійкості за допомогою нейронних мереж є новою та перспективною в сучасному світі.

Метою дослідження є забезпечення функціональної стійкості виробничих процесів промислових підприємств на основі нейроадаптивної системи.

Виклад основного матеріалу

Автоматизація процесу контролю параметрів виробничих процесів на сучасних підприємствах є важливим фактором для організації випуску якісної продукції (рис. 1).

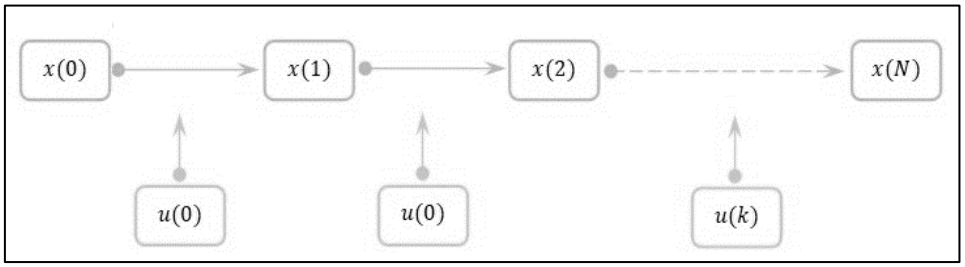

Рис. 1. Топологія лінійного технологічного виробничого процесу

Випуск продукції складається з декількох етапів, на кожному з яких висуваються вимоги до параметрів та характеристик сировини та готової продукції [15]. Нехай $x(i)$ – це набори параметрів на кожному i-му етапі, i=1, 2, …, N. На кожному етапі для гарантування досягнення параметрів $x(i)$ впливають зовнішні параметри $u(i)$, наприклад, ефект від роботи, енергетичний ефект, хімічні та ін. Вважаємо, що якість продукції на кожному етапі залежить від дотримання технології та забезпечення контролю за необхідними параметрами на кожному попередньому кроці.

Ще введемо додаткові позначення:

- $A(i)$ — матриця залежності показників якості продукції на (i+1)-му етапі від показників на i-му етапі (матриця виробничого процесу);
- $C(i)$ — матрицю, яка визначає структуру впливу на виробничий процес $u(i)$.

Математична модель технологічного процесу, що забезпечується інформаційними системами виробничого підприємства, представляється так:

$$x(i+1) = A(i)x(i) + C(i)u(i),\ i = 1,2,...,N,$$
$$x(i) \in R^n,\ A(i) \in R^{n\times n},\ C(i) \in R^{n\times m},\ u(i) \in R^m, \quad (1)$$

Відтак забезпечення функціональної стійкості виробничого процесу залежить від здатності забезпечувати керованість процесом в кожному виробничому центрі та контролювати виникаючі несправності. Тому для забезпечення цієї властивості важливо застосовувати діагностику для виявлення несправних елементів у системі та вчасно їх локалізувати. Отже, функціональна стійкість виробничих процесів надає такі переваги:

- безперебійність роботи: стійке функціонування обладнання та технологічних систем гарантує

безперебійний випуск продукції, мінімізуючи простої та втрати;

- якість продукції: стійкість параметрів технологічних процесів забезпечує стабільну якість виробленої продукції, що відповідає стандартам та очікуванням споживачів;
- ефективність виробництва: безвідмовна робота обладнання та мінімізація браку сприяють підвищенню ефективності виробництва, знижуючи витрати та збільшуючи рентабельність;
- безпека праці: стійкість до відмов та аварій забезпечує безпечне робоче середовище для персоналу, мінімізуючи ризики травм та виробничих аварій.

Наразі для діагностування складних технічних систем пропонується застосовувати штучні нейронні мережі, які дозволять оптимізувати виробничі процеси, підвищувати їх надійність та стійкість до відмов. Зазначимо, що нейронні мережі є інструментом, який дозволяє створювати глибоку ієрархію прийняття рішень з врахуванням місця, виду та рівня дефекту, який виник в системі управління і, як наслідок, може використовуватись в діагностуванні. Нейронні мережі можуть аналізувати дані з сенсорів та інших джерел, щоб передбачити можливі несправності обладнання. Це дозволяє проводити обслуговування на основі реальних потреб, а не за фіксованим графіком, що підвищує надійність та зменшує простої.

Виділимо деякі можливості нейронних мереж в контексті забезпечення функціональної стійкості виробничих процесів:

1. Прогнозування відмов обладнання та обслуговування: нейронні мережі можуть аналізувати історичні дані про виробничі процеси та прогнозувати майбутні відмови обладнання або неполадки. Наприклад, рекурентні нейронні мережі (RNN) можуть враховувати часові залежності між даними і передбачати моменти, коли обладнання може вийти з ладу. Це дозволяє планувати профілактичне обслуговування та уникнути непередбачених зупинок у виробництві.

2. Виявлення аномалій у виробничому процесі: нейронні мережі засновані на зворотному розповсюдженні помилок (backpropagation), можуть навчитися виявляти аномалії у виробничих процесах. Вони можуть виявляти відхилення від нормальної роботи обладнання або процесів, що може свідчити про потенційні проблеми, такі як пошкодження або несправності.

3. Оптимізація параметрів технології: нейронні мережі можуть бути використані для оптимізації параметрів виробничих процесів з метою підвищення продуктивності та зниження витрат. Наприклад, з використанням глибоких нейронних мереж (DNN) можна розробити моделі, які адаптуються до різноманітних умов виробництва та швидко знаходять оптимальні режими роботи.

4. Автоматизоване управління виробничим процесом: НМ можуть використовуватися для автоматизованого управління виробничим процесом, реагуючи на зміни в режимі реального часу та підтримуючи його стійкість. Один з ефективних прикладів нейронної мережі для автоматизованого управління виробничим процесом – це нейронні контролери або адаптивні контролери, які використовуються для автоматичного керування об'єктами зі складною динамікою або змінюваними параметрами. Одним із найпоширеніших прикладів таких мереж є нейронні мережі зі зворотнім поширенням помилок (Backpropagation Neural Networks – BPNN).

5. Прогнозування попиту та управління запасами сировини та готової продукції: нейронні мережі можуть бути використані для прогнозування попиту на продукцію та управління запасами сировини та готової продукції. Це допомагає уникнути надлишкових запасів або нестачі, що може призвести до перебоїв у виробництві. Наприклад, може бути ефективним використання глибокої нейронної мережі, такої як Long Short-Term Memory (LSTM) або Gated Recurrent Unit (GRU). Вони часто використовуються для аналізу часових рядів та здатні ефективно моделювати довгострокові залежності у даних.

Кожен конкретний випадок вимагає вибору підходящої архітектури нейронної мережі з урахуванням конкретних характеристик виробничого процесу. Наприклад, для аналізу часових рядів можуть бути ефективними рекурентні нейронні мережі, а для класифікації даних – згорткові нейронні мережі. Вибір конкретної архітектури також може залежати від обсягу даних, складності моделі та доступних обчислювальних ресурсів. Отже, для забезпечення функціональної стійкості виробничих процесів можна широко застосовувати різні класи нейронних мереж для забезпечення діагностики стану обладнання на кожному виробничому центрі. Навчаючи нейронні мережі з урахуванням умови функціональної стійкості технологічного процесу, буде забезпечена ефективна робота як виробничого обладнання, так і поточний контроль дотримання якості продукції згідно визначеної системи толерансів.

Для досягнення функціональної стійкості системи важливо проводити її діагностування, що дозволить враховувати всі особливості системи та визначити, чи працюють модулі системи належним чином, оперативно виявляти несправності та вживати відповідні заходи для забезпечення надійності та продуктивності системи. Система діагностування повинна складатися з об’єкту управління, області діагностичних ознак та алгоритму прийняття рішень. Від об’єкту до області надходять сигнали, які аналізуються алгоритмом та видається висновок про правильність функціонування досліджуваного об’єкту. В залежності від місця та способу використання нейронної мережі можна отримати різні структури систем діагностування (рис. 2). На рис. 2, а представлена структура, де в якості області ознак використовуються змінні параметри об’єкту (наприклад, вхідні сигнали). Нейронна мережа тут використовується в якості пристрою прийняття рішень. На рис. 2, б представлена структура, де нейронна мережа використовується як модель динамічної системи, а в якості алгоритму прийняття рішень – будь-який алгоритм класифікації, наприклад, алгоритм виявлення розлагодження. Структура, яка представлена

на рис. 2, в, використовує стандартний підхід з використанням моделі для генерації нев'язок, в той час як нейронний класифікатор розв'язує задачу виявлення та локалізації дефектів. У структур, які зображені на рис. 2, б та 2, в, відбувається суміщення аналітичних методів з нейромережевим підходом до розв'язання пов'язаних між собою задач генерації ознак та прийняття рішень. На рис. 2, г представлена структура для випадку, коли система діагностування виявляє дефекти, але не буде їх локалізувати.

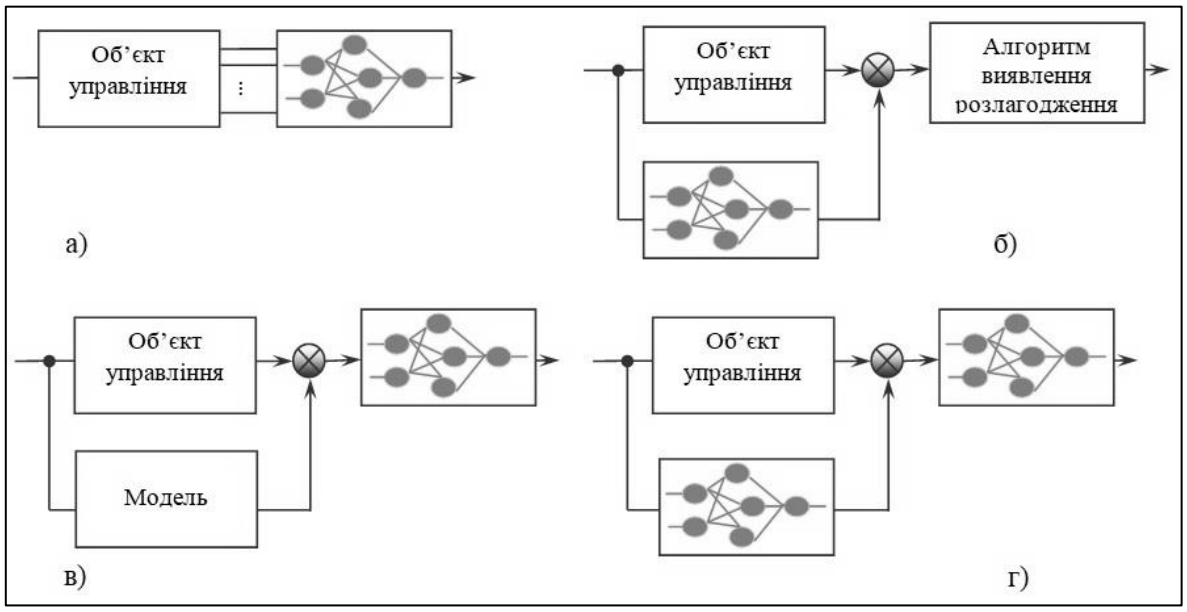

Рис. 2. Структури діагностування на основі нейронних мереж

У випадку використання нейронної мережі в якості алгоритму класифікації незалежно від структури і якщо нейронна мережа має два виходи, то розв'язується задача виявлення. Коли ж входів більше двох, розв'язується задача виявлення та локалізації дефектів і, можливо, визначення величини дефекту.

Отже, при використанні нейромережевого підходу можливо проводити обчислення паралельно, а це в свою чергу дає можливість реалізувати системи управління вищого порядку за прийнятних показників збіжності і, отже, досягти вищої якості управління. Нейромережевий підхід до реалізації багатовимірних просторових систем управління дозволяє вирішувати проблеми, що стояли перед розробниками щодо необхідності виконання векторно-матричних операцій високої розмірності в реальному часі.

Проектування нейронної системи представляє собою складний і трудомісткий процес, у якому вибір алгоритму діагностування системи є дуже важливим кроком. Розглянемо алгоритм діагностування системи за допомогою нейроадаптивної системи.

Алгоритм діагностування

Крок 1. Дослідження та аналіз процесу з точки зору нейромережевих технологій.

Крок 2. Визначення типу нейронів в мережі.

Крок 3. Визначення кількості шарів нейронної мережі і число нейронів в шарах в залежності від заданої кількості областей розбиття області ознак при умові мінімізації числа нейронів в мережі.

Крок 4. Синтез нейромережі зі змінною структурою:

1) так – вибір алгоритму побудови структури багатошарової нейромережі з метою досягнення найкращої відповідності складності розв'язуваної задачі;

2) ні – синтезувати структуру нейромережі з визначенням способу створення зв'язків між шарами: прямими, перехресними, оберненими, комбінованими.

Крок 5. Вибір алгоритму адаптації багатошарової нейромережі.

Крок 6. Діагностика синтезованої нейромережі.

Крок 7. Оцінка функціональної стійкості системи.

Крок 8. Вибір варіанту перерозподілу задач між працюючими модулями.

Крок 9. Оцінка результатів діагностування.

Промисловий попит на комп'ютери для машинного навчання зріс у мільйон разів за останні шість років і щороку продовжує збільшитись удесятеро. Для апаратної реалізації нейромережі потрібно обрати нейрочіп із вбудованими нелінійними перетвореннями. Їх виробляють в багатьох країнах світу й вони створюються для конкретних систем. Розглянемо класифікацію чипів [17]:

- за способом подання інформації – цифрові, аналогові та гібридні;
- за типом реалізації нейроалгоритмів – з апаратною та програмно-апаратною реалізацією;
- за характером нелінійних перетворень – апаратно реалізовані та перепрограмуючі;
- за побудовою мереж – з жорсткою та змінною нейромережевою структурою;
- систолічні;
- нейросигнальні.

Орієнтація у виконанні нейромережевих операцій зумовлює з одного боку підвищення швидкостей обміну між пам'яттю та паралельними арифметичними пристроями, а з іншого боку – зменшення часу вагового підсумовування (множення та накопичення) за рахунок застосування фіксованого набору команд типу регістр-регістр.

Відмітимо, що при проектуванні апаратної реалізації нейромереж враховуються такі характеристики, як масштабованість, вартість розробки, сумісність з минулими та майбутніми версіями. І якщо для розробника критичні терміни розробки і вартість, то слід звернути увагу на сигнальні чіпи або чіпи для каскадних архітектур. Отже, загальноприйнятих рекомендацій по вибору конкретної бази в наш час не існує. Все залежить від вимог розробленої системи.

Для оцінки продуктивності нейрочипів будемо використовувати такі показники.

- Теоретична максимальна продуктивність нейрочипа в операціях з рухомою комою з одинарною точністю (FP32) за секунду. Високий показник TFLOPS вказує на те, що нейрочіп може обробляти великі обсяги даних з високою швидкістю, що робить його придатним для складних завдань машинного навчання. Однак важливо зазначити, що TFLOPS – це лише теоретичний показник, і реальна продуктивність може відрізнятися залежно від конкретного завдання та програмного забезпечення.
- Максимальна тактова частота – це максимальна швидкість, на якій може працювати нейрочіп.
- Розмір кристала – це загальна площа кремнієвої пластини, на якій розміщені транзистори нейрочипа. Більший розмір кристала може дозволити розмістити більше транзисторів, що може призвести до кращої продуктивності.
- Кількість транзисторів – це загальна кількість транзисторів на нейрочипі. Більша кількість транзисторів зазвичай вказує на те, що нейрочіп є більш складним і може виконувати більше операцій.
- Чіпів на CPU-хост- це кількість нейрочипів, які можна підключити до одного CPU. Більша кількість нейрочипів може призвести до кращої масштабованості та загальної продуктивності системи.
- Теплова розсіююча потужність – це максимальна кількість тепла (теоретичний максимум), яку може генерувати нейрочіп. Низький показник вказує на те, що нейрочіп є більш енергоефективним.
- Найбільша масштабована конфігурація – це максимальна кількість нейрочипів, які можна використовувати разом у одній системі. Більша масштабована конфігурація може призвести до кращої продуктивності для дуже складних завдань.

В [17] виділяють дві базові лінії розвитку обчислювальних високопродуктивних систем з масовим паралелізмом (ВСМП): з модифікованими послідовними алгоритмами та надпаралельних нейромережевих алгоритмів вирішення різних завдань. До цієї інформації слід додати, що сучасні тенденції розвитку ВСМП характеризуються наявністю більшої кількості базових ліній розвитку, які неможна чітко окреслити, бо вони часто перетинаються та поєднуються одна з одною.

Умовно ці лінії розвитку можна подати таким чином:

1) підхід використання модифікованих послідовних алгоритмів як і раніше, використовується для адаптації класичних алгоритмів до багатопроцесорних систем, проте з'являються нові методи оптимізації та розробки алгоритмів, які враховують специфіку сучасних архітектур, таких як GPU, TPU та нейроморфні процесори;

2) підхід використання нейромережевих алгоритмів продовжує розвиватися, з'являються нові архітектури та алгоритми навчання, що значно розширюють можливості нейромереж. Нейромережі використовуються для нових задач, таких як генерація тексту, переклад мов та управління складними системами;

3) підхід використання нейроморфних обчислень прагне створити комп'ютерні системи, які імітують роботу людського мозку. Нейроморфні системи мають потенціал значно перевершити традиційні комп'ютери за енергоефективністю та продуктивністю при вирішенні певних задач;

4) підхід використання квантових обчислень. Квантові комп'ютери мають потенціал вирішувати деякі задачі, які неможливо вирішити на традиційних комп'ютерах, наприклад, факторизацію великих чисел;

5) підхід використання гетерогенних обчислювальних систем, які поєднують в собі різні типи процесорів, такі як CPU, GPU, TPU та нейроморфні процесори.

Наведемо декілько найбільш популярних нейрочіпів у країнах Європи та США:

NVIDIA A100 (США, 2020) використовується у широкому спектрі застосунків, включаючи штучний інтелект, машинне навчання, обробку природної мови та наукові обчислення;

Intel Habana Labs Gaudi (США, 2020) спеціально розроблений для Intel AI Framework, він пропонує високу продуктивність та енергоефективність, що робить його популярним вибором для застосунків штучного інтелекту, які потребують низької затримки;

Graphcore Colossus GC30 (Великобританія, 2020) пропонує високу продуктивність та масштабованість, що робить його популярним вибором для великих проектів машинного навчання; Google TPU v4 (США, 2021) та наступні версії v5e, v5p спеціально розроблені для TensorFlow, платформи машинного навчання Google;

AMD MI25X (США, 2021) є конкурентом NVIDIA A100 і пропонує схожу продуктивність за нижчою ціною;

Inoveo NeuroScale 2 (Франція, 2021) спеціально розроблений для штучного інтелекту в галузі охорони здоров'я, популярний вибір для медичної діагностики та обробки зображень;

NVIDIA H200 (США, 2023) – це перший графічний процесор із пам'яттю HBM3E, яка відрізняється від звичайної HBM3 (high bandwidth memory, HBM) вищою швидкістю, нова пам'ять дозволить прискорювачу швидше працювати з величезними обсягами даних для генеративного ШІ та високопродуктивних обчислювальних навантажень;

Cerebras Wafer-Scale Engine 3 (WSE-3) (США, Cerebras Technologies, 2024) – це третє покоління суперкомп'ютерної платформи Cerebras, є одним із найпотужніших нейрочіпів у світі, пропонує безпрецедентну продуктивність, що робить його популярним вибором для дослідницьких проектів штучного інтелекту, які потребують обчислення великих обсягів даних;

Google TPU Trillium (США, очікується в кінці 2024) – це шосте покоління фірмового тензорного процесора, новий чіп майже вп'ятеро продуктивніший за попередника TPU v5e, при цьому Trillium на 67% енергоефективніший, ніж TPU v5e [18-21].

Найбільш перспективними є нейрочіпи розробників NVIDIA та Google.

В табл. 1 наведено їх характеристики [21-24].

Таблиця 1 – **Порівняльна характеристика деяких нейрочипів**

Характеристики	Nvidia A100	Graphcore MK2 IPU	Google TPU v4	Google TPU v3
Вихід на ринок	2020	2021	2020	2018
Пікова продуктивність, TFLOPS	312 (bf16), 624 (i8)	250 (bf16)	275 (bf16 або int8)	123 (bf16)
Тактова частота, МГц	1095/1410	1850	1050	940
Тех.процес, нм Розмір кристала, мм^2	7 826	7 832	7 <600	16 <700
Кількість транзисторів, мільярдів	54	59	22	10
Мікросхем на хост ЦП	4	4	4	8
Номінальна теплова потужність, Вт	400	300	інформація не доступна	
Енергоспоживання нейрочіпа, коли він не використовується, мін/сер/макс, Вт	інформація не доступна		90 121/170/192	123 175/220/262
Міжчіпове з'єднання: кількість з'єднань та швидкість передачі даних, ГБ/с	12 25	3 64	6 50	4 70
Конфігурація найбільшого масштабу, к-ть чипів	4216	256	4096	1024
Архітектура процесора	SIMT	MIMD	SIMD 2D	
Процесори / Чип	108	1472	2	2
Потоки / Ядра	32	6	1	1
Розріджені ядра/ Чип			4	2
Оперативна пам'ять, МБ	38	865	123 (CMEM)+ 30 (VMEM) + 10 (spMEM)	30 (VMEM) + 5 (spMEM)
Розмір файлу регістрів	26	1,3	0,24	0,24
Ємність HBM2, ГБ пропускна здатність, ГБ/с	74 2039	0	30 1200	30 900

В графі «Архітектура процесора» (табл. 1) розшифруємо абревіатури:

Single Instruction Multiple Threads (SIMT) – це архітектура процесора, яка дозволяє йому виконувати одну інструкцію над декількома потоками даних одночасно. Це може значно прискорити обробку даних, адже процесор може виконувати багато операцій одночасно.

Принцип роботи SIMT.

Крок 1. Процесор отримує одну інструкцію.

Крок 2. Процесор розбиває інструкцію на декілька менших частин.

Крок 3. Процесор виконує кожну частину інструкції над декількома потоками даних одночасно.

Multiple Instruction Multiple Data (MIMD) – це архітектура процесора, яка дозволяє йому виконувати декілька інструкцій над декількома потоками даних одночасно, що робить MIMD-процесори дуже універсальними, адже вони можуть виконувати широкий спектр задач.

Вони більш гнучкі, адже можуть виконувати декілька різних інструкцій одночасно для неоднорідного набору даних.

Принцип роботи MIMD.

Крок 1. Процесор отримує декілька інструкцій.

Крок 2. Процесор розбиває кожну інструкцію на декілька менших частин.

Крок 3. Процесор виконує кожну частину інструкції над декількома потоками даних одночасно.

Його можна використовувати в ЦП, ГП, нейрочіпах та інших комп'ютерних системах, може бути менш ефективним для задач машинного навчання,

адже йому може бути складно оптимізувати виконання декількох різних інструкцій.

Single Instruction 2D Data (SIMD 2D) – це архітектура процесора, яка є розширенням класичної SIMT архітектури. Вона дозволяє процесору виконувати одну інструкцію над двовимірними матрицями даних одночасно.

Принцип роботи SIMD 2D.

Крок. Процесор отримує одну інструкцію.

Крок 2. Процесор розбиває інструкцію на декілька менших частин.

Крок 3. Процесор виконує кожну частину інструкції над двовимірними матрицями даних одночасно.

SIMD 2D поєднує в собі переваги SIMT та MIMD, пропонуючи високу продуктивність для задач з двовимірними даними.

В графі «Оперативна пам'ять» (табл. 1) є три типи ОП.

CMEM – це кеш-пам'ять першого рівня (L1), яка використовується для зберігання інструкцій та даних, які використовуються найчастіше.

VMEM – це кеш-пам'ять другого рівня (L2), яка використовується для зберігання даних, які використовуються трохи рідше, ніж дані в CMEM.

spMEM – це спеціальна пам'ять, яка використовується для зберігання даних, які потрібні для роботи певних блоків нейрочіпа.

Отже, отримано наступні результати порівняння характеристик нейрочипів:

- технологічний вузол (розмір транзистора, що використовується в мікросхемі) у перших трьох вдвічі менший за TPU v3, що дозволяє розмістити більше транзисторів на одній мікросхемі, а це може призводити до кращої продуктивності та енергоефективності;
- TPU v4 має вдвічі більше матричних множників, ніж TPU v3. Матричні множники використовуються для виконання основних операцій машинного навчання, таких як множення матриць та згортки, тому вони є важливими для виконання завдань машинного навчання, а отже їх збільшення може суттєво покращити продуктивність.
- Graphcore MK2 IPU має найвищу тактову частоту: на 68%/31% вищу за Nvidia A100, на 76% вищу, ніж TPU v4 та на 97% вищу, ніж TPU v3. Тактова частота стосується кількості циклів, які процесор може виконати за одну секунду. Більша тактова частота може призвести до кращої продуктивності.
- Graphcore MK2 IPU має 12 міжз'єднань (порівнянно з 6 та 4 у Google TPU) на чіп дозволяють створювати мережеві топології з меншим діаметром мережі.
- частина покращення продуктивності на ват (ефективності) походить від самого 7-нм техпроцесу (близько 40%). Решта покращення походить від змін у дизайні TPU v4, таких як балансування конвеєра та реалізація тактового гетування (для динамічного відключення блоків, які не використовуються в даний момент, що допомагає зменшити споживання енергії). Продуктивність на ват – це міра енергоефективності процесора. Вона вимірюється в GFLOPS/W (гігафлопсах на ват) і показує, скільки GFLOPS процесор може виконувати на один ват споживаної потужності.
- HBM (High Bandwidth Memory) - це тип пам'яті, який може дуже швидко передавати дані. TPU v4 має в 1,4 та 1,3 рази більшу пропускну здатність пам'яті HBM порівняно з TPU v3 та MK2 IPU відповідно, в свою чергу Nvidia A100 в 2 рази більшу порівняно з TPU v4, що означає, що він може швидше отримувати доступ до даних з пам'яті.
- TPU v4 має 128 МБ вбудованої пам'яті scratchpad (CMEM), якої немає в TPU v3. Scratchpad-пам'ять – це тип дуже швидкої пам'яті, до якої процесор може отримувати доступ дуже швидко. Це може бути корисно для зберігання даних, які часто використовуються під час обчислень.

На рис. 3 представлені результати експеременту застосування TPU v4 та TPU v3 однакової конфігурації (кількості TPU чипів) для різних застосунків (наскільки суперкомп'ютери TPU v4 перевершують TPU v3).

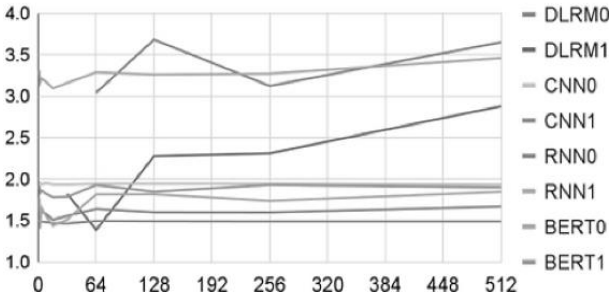

Рис. 3. Прискорення TPU v4 порівнянно з v3 за однакової кількості чипів

Виходячи з порівнянь у таблиці, не дивно, що для більшості застосунків при однаковій конфігурації, TPU v4 працює в 1.5-2.0 рази швидше, ніж TPU v3. Застосунок DLRM0 працює в 3.0-3.5 рази швидше, а DLRM1 - в 2.8 рази швидше на конфігурації з 512 чипами. Це пояснюється тим, що TPU v4 має вдвічі більше обчислювальних блоків (SC) та їхня тактова частота є вищою. Але найбільш неочікуваний результат отримано для застосунку RNN1 - він працює в 3.3 рази швидше на TPU v4. Це зумовлено тим, що малі ваги та малий розмір пакетів даних RNN1 отримують значну перевагу від пропускної здатності внутрішньої пам'яті CMEM порівняно з зовнішньою пам'яттю HBM [21].

TPUv2/v3 мають менші кристали у більш старому напівпровідниковому процесі та нижчі ціни на хмару, незважаючи на те, що вони менш зрілі на багатьох рівнях стеку апаратного/програмного забезпечення, ніж CPU та GPU. Ці добрі результати, незважаючи на технологічні недоліки, свідчать, що підхід TPU є економічно ефективним і може забезпечити високу архітектурну ефективність у майбутньому [22].

В цілому, Google TPU v4 – це достатньо потужний нейрочіп, який розроблений спеціально для задач машинного навчання. Він має значно більшу продуктивність, ніж A100, але дорожчий. TPU v4 добре підходить для вирішення складних задач машинного

навчання, таких як навчання великих мовних моделей та розробка нових алгоритмів штучного інтелекту. TPU v4 підтримує TensorFlow та PyTorch, що робить його зручним для використання з популярними фреймворками машинного навчання. TPU v4 доступний через Google Cloud TPU, що робить його доступним для розробників у всьому світі. Зазначається, що платформа Cloud TPU v4 в 1,2-1,7 рази продуктивніша і витрачає в 1,3-1,9 рази менше енергії, ніж платформи на базі NVIDIA A100 у системах аналогічного розміру. Хоча, поки компанія не порівнювала TPU v4 з більш новими прискорювачами NVIDIA H100 через їхню обмежену доступність і 4-нм архітектури (порівняно з 7-нм у TPU v4) [21-22].

Отже, нейронні мережі володіють широким спектром можливостей для застосування в задачах діагностування стану виробничих процесів та складних технічних систем. Їхня здатність до навчання, швидкість обробки інформації, гнучкість, автоматизація, можливість виявлення неявних несправностей та прогнозування несправностей роблять їх цінним інструментом для підвищення надійності, безпеки та ефективності експлуатації систем. Важливо зазначити, що нейронні мережі можуть використовуватися в комплексі з іншими методами та системами діагностування, такими як експертні системи, фізичні моделі, статистичні методи. Їх застосування для діагностування системи промислового підприємства може призвести до значної економічної вигоди, таких як зниження витрат на ремонт, простої та брак, а також до підвищення якості продукції та конкурентоспроможності підприємства.

Висновки

Функціональна стійкість є критичним фактором для успішного впровадження технологій у промислових підприємствах. У контексті швидкого розвитку індустрії, підприємства дедалі частіше використовують передові технології, такі як нейронні мережі та інші методи штучного інтелекту, для оптимізації виробничих процесів, контролю якості та прогнозування технічного обслуговування. Однак, висока залежність від таких технологій робить питання їхньої надійності і стійкості до збоїв надзвичайно актуальними.

Функціональна стійкість у промислових умовах означає здатність систем підтримувати свою продуктивність і точність при змінних робочих умовах, впливі зовнішніх збурень і можливих апаратних відмовах. Це включає стійкість до непередбачуваних змін вхідних даних, адаптивність до нових умов експлуатації, а також здатність до швидкого відновлення після збоїв. Недоліки в стійкості систем можуть призвести до зупинки виробничих ліній, значних фінансових втрат та зниження якості продукції.

У сучасних умовах дуже активно розвивається область нейронауки, де вже представлено багато різних методів для вирішення різних інформаційно-технологічних проблем, серед яких одним з найбільш ефективних є штучні нейронні мережі. Нейронні мережі можуть бути ефективним інструментом, який дозволить створити глибоку ієрархію прийняття рішень з врахуванням рекомендацій, отриманих з моніторингу роботи виробничих процесів у системі. У статті розглянуто особливості основних положень теорії штучного інтелекту, а саме нейронних мереж, з порівняльними характеристиками апаратної реалізації нейромереж. Визначено структури діагностування виробничих процесів, описано алгоритм діагностування системи за допомогою нейроадаптивної системи. Для апаратної реалізації нейромережі обрано нейрочіп із вбудованими нелінійними перетвореннями, а саме Google TPU v4, оскільки у нього високий коефіцієнт продуктивності, системне програмне забезпечення TPU v4 відповідає за розподіл обчислень, оптимізацію передачі даних та моніторинг системи та фреймворки машинного навчання, такі як TensorFlow та PyTorch, надають набір операцій для управління мережевим взаємодією вузлів системи.

Подальшим перспективним шляхом розвитку роботи є застосування експертних систем, які дозволять вдосконалити процес діагностування виробничих процесів, надаючи їм більшої гнучкості та універсальності.

Список літератури

1. Собчук В.В. Основи забезпечення функціональної стійкості інформаційних систем підприємств в умовах впливу дестабілізуючих факторів: монографія / В.В. Собчук, О.В. Барабаш, А.П. Мусієнко. К: Міленіум. – 2022. – 272 с. https://www.researchgate.net/publication/363474851_Basis_for_functional_stability_of_information_systems_businesses_under_the_influence_of_destabilizing_factors
2. Sobchuk A.V. Assessment methods of functional stability of wireless sensor networks / A.V. Sobchuk, O.V. Barabash, A.P. Musienko // Телекомунікаційні та інформаційні технології. – 2019. – № 3 (64). – С. 46–54. https://doi.org/10.31673/2412-4338.2019.034654
3. Собчук В.В. Методи контролю і діагностування інформаційної системи підприємства за принципом адаптивного накопичення діагностичної інформації / В.В. Собчук, О.В. Барабаш, А.П. Мусієнко, О.А. Капустян // Вісник Київського національного університету ім. Т. Шевченка. – 2020. – Вип. 4. – С. 69–78. https://doi.org/10.17721/1812-5409.2020/4.10
4. Barabash O. Ensuring the functional stability of the information system of the power plant on the basis of monitoring the parameters of the working condition of computer devices / O. Barabash, O. Svynchuk, I. Salanda, V. Mashkov, M. Myroniuk // Advanced Information Systems. – 2024. – Vol. 8, no. 2. – P. 107–117. https://doi.org/10.20998/2522-9052.2024.2.12
5. Барабаш О. В. Програмне забезпечення контролю справного стану інформаційних систем в енергетичній галузі для забезпечення функціональної стійкості / О. В. Барабаш, О. В. Свинчук, Бандурка О. І. // Сучасний захист інформації. – 2024. – № 2 (58). – С. 41–49. https://doi.org/10.31673/2409-7292.2024.020005
6. Svynchuk O. Development of the information system for monitoring time changes in forest plantations based on the analysis of space images / O. Svynchuk, O. Bandurka, O. Barabash, O. Ilin, A. Lapin // Eastern-European Journal of Enterprise Technologies. – 2022. – Vol. 5, no. 2 (119). – P. 31–41. https://doi.org/10.15587/1729-4061. 2022. 265039

7. Mashkov V.A. Self-Checking of modular systems under random performance of elementary checks. Engineering Simulation / V.A. Mashkov, O.V. Barabash // Amsterdam: OPA. – 1995. – Vol. 12. P. 433–445.
8. Sobchuk V. Adaptive accumulation and diagnostic information systems of enterprises in energy and industry sectors / V. Sobchuk, O. Barabash, A. Musienko, O. Svynchuk // E3S Web of Conferences. – 2021. – Vol. 250. – P. 82–87. https://doi.org/10.1051/e3sconf/202125008002
9. Barabash O. System analysis and method of ensuring functional sustainability of the information system of a critical infrastructure object / O. Barabash, V. Sobchuk, A. Musienko O. Laptiev, V. Bohomia, S. Kopytko // In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Artificial Intelligence. Studies in Computational Intelligence. – 2023. – Vol 1107. – Springer, Cham. – P. 117–192. https://doi.org/10.1007/978-3-031-37450-0_11
10. Maksymuk O.V. A system of indicators and criteria for evaluation of the level of functional stability of information heterogenic networks / O.V. Maksymuk, V.V. Sobchuk, I.P. Salanda, Yu.V. Sachuk // Mathematical modeling and computing. – 2020. – Vol. 7, no. 2. – P. 285–292. http://doi.org/10.23939/mmc2020.02.285
11. Mashkov O. Features of determining controlling effects in functionally-stablesystems with the recovery of a control / O. Mashkov, V. Chumakevych, O. Sokulsky, L. Chyrun // Mathematical Modeling And Computing. – 2019. – Vol. 6, no. 1. – P. 85–91. http://doi.org/10.23939/mmc2019.01.085
12. Laptiev O. The method of construction of the law of safety management of critical infrastructure objects under the conditions of external uncontrolled influences / O. Laptiev, O. Barabash, I. Tsyganivska, D. Obidin, A. Sobchuk // CEUR Workshop Proceedings. – 2023. – Vol. 3624. – P. 291–300. https://ceur-ws.org/Vol-3624/Paper_24.pdf
13. Собчук А.В. Застосування нейромереж для забезпечення функціональної стійкості виробничих процесів / А.В. Собчук, Ю.І. Олімпієва // Телекомунікаційні та інформаційні технології. – 2020. – № 2 (67). – С. 13–28. http://doi.org/10.31673/2412-4338.2020.021328
14. Тюлюпа С.В. Самодіагностування як спосіб підвищення кіберстійкості термінальних компонентів технологічної системи / С.В. Тюлюпа, Ю.Я. Самохвалов, П.В. Хусаінов, С.С. Штатенко // Кібербезпека: освіта, наука і техніка. – 2023. – № 2 (22). С. 134–147. https://doi.org/10.28925/2663-4023.2023.22.134147
15. Собчук В.В. Функціональна стійкість технологічних процесів на основі нелінійної динаміки із застосуванням нейромереж / В.В. Собчук, І.В. Замрій, Ю.І. Олімпієва, С.О. Лаптєв // Сучасні інформаційні системи. – 2021. – Т.5 (2). – С.49–57. https://doi.org/10.20998/2522-9052.2021.2.08
16. Guo-Qiang Kong Stability analysis of delayed neural networks based on improved quadratic function condition / Guo-Qiang Kong, Liang-Dong Guo // Neurocomputing. – 2023. – Vol. 524. – P. 158–166. https://doi.org/10.1016/j.neucom.2022.12.012.
17. Кожем'яко В.П. Сучасний стан, елементна база та порівняльний аналіз характеристик нейрообчислювачів / В.П. Кожем'яко, А.В. Кожем'яко, О.С. Васильківа // Оптико-електронні інформаційно-енергетичні технології. – 2017. – Т. 32 (2). – С. 29–38.
18. The future of AI is Wafer-Scale [Електронний ресурс]. – Режим доступу: https://www.cerebras.net/product-chip/
19. NVIDIA H200 Tensor Core GPU [Електронний ресурс]. – Режим доступу: https://www.nvidia.com/en-us/data-center/h200/
20. NVIDIA DGX H100 [Електронний ресурс]. – Режим доступу: https://nvdam.widen.net/s/kpwzdwrwbv/ai-for-enterprise-dgx-h100-datasheet-nvidia-a4-2146027-r3-web
21. Introducing TPU v4: Googles Cutting Edge Supercomputer for Large Language Models [Електр. ресурс]. – Режим доступу: https://www.kdnuggets.com/2023/04/introducing-tpu-v4-googles-cutting-edge-supercomputer-large-language-models.html
22. Jouppi N.P. A domain-specific supercomputer for training deep neural networks / N.P. Jouppi, D.H. Yoon, G. Kurian, S. Li, N. Patil, J. Laudon, C. Young, D. Patterson // Communications of the ACM. – 2020. – Vol. 63, no. 7. – P. 67–78. https://doi.org/10.1145/3360307
23. Graphcore/ IPU-POD64 Reference Design Datasheet [Електронний ресурс]. – Режим доступу: https://docs.graphcore.ai/projects/ipu-pod64-datasheet/en/latest/overview.html
24. NVIDIA A100 Tensor Core GPU Architecture [Електронний ресурс]. – Режим доступу: https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

Received (Надійшла) 06.05.2024
Accepted for publication (Прийнята до друку) 10.07.2024

Ensuring functional sustainability of production processes of industrial enterprises based on the neuroadaptive system

Yu. Olimpiyeva

Abstract. The article considers the problem of ensuring the functional stability of production processes of industrial enterprises based on neuroadaptive systems. In an industry where production processes are becoming increasingly complex and automated, the importance of reliable and sustainable control systems is growing. An analysis of modern approaches to the use of neural networks for adaptive management of production processes is carried out, which allows to increase their efficiency and reliability. Various methods of increasing functional stability are analyzed, including adaptive learning algorithms, regularization, and error detection and correction techniques. Special attention is paid to the integration of neuroadaptive systems with existing technologies and production lines, as well as their ability to quickly recover from failures and adapt to changing operating conditions. Based on the analysis, a new algorithm for diagnosing the main neuroadaptive systems is proposed, which will contribute to increasing the stability and reliability of production processes. When designing the hardware implementation of neural networks, such characteristics as scalability, development cost, compatibility with past and future versions are taken into account. Therefore, among existing neurochips, Google TPU v4 was chosen because it has a high performance factor and the system part of the software implements set of operations for managing the network interaction of system nodes. It is important to note that neural networks can be used in combination with other diagnostic methods and systems, such as expert systems, physical models, statistical methods.

Keywords: functional stability, diagnostics, neural networks, neuroadaptive systems, algorithm, neurochips, production process, industrial enterprises.