УДК 004.056.53

ОГЛЯД СУЧАСНИХ МЕТОДІВ АДАПТАЦІЇ НЕЙРОННИХ МЕРЕЖ ДЛЯ МАЛОПОТУЖНИХ ПРИСТРОЇВ

І. В. Струнін[1], Д. О. Прогонов[1]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

Анотація

Забезпеченню надійної автентифікації користувачів сьогодні приділяється особлива увага при розробці, впровадженні та супроводі комплексних систем захисту інформації. До систем автентифікації висуваються високі вимоги щодо низького рівня помилок та можливості роботи в реальному часі. Вирішення даної задачі потребує залучення комплексних методів обробки даних, зокрема штучних нейронних мереж, та потужних обчислювальних комплексів необхідних для їх роботи. Це призводить до високої складності інтеграції таких систем з існуючими системами безпеки. В роботі проведено огляд сучасних методів адаптації шутчних нейронних мереж для роботи на малопотужних обчислювальних пристроях. За результатами дослідження встановлено, що комплексне використання декількох методів адаптації забезпечує високу точність роботи штучних нейронних мереж на малопотужних пристроях при збереженні заданої точності автентифікації особи.

Ключові слова: нейронні мережі, IoT, паралельні обчислення, квантування вагів, автентифікація, ідентифікація, захист інформації

Вступ

Біометричні дані людини все частіше використовуються для вирішення задач аутентифікації в сучасних корпоративних системах контролю доступу (СКД). Ключовими вимогами до таких систем є рівень помилок першого (хибне спрацювання, англ. False Acceptance Rate, FAR) та другого (хибна відмова, англ. False Rejection Rate, FRR) роду, швидкість обробки поступаючих даних (можливість роботи в режимі реального часу або наближеного до нього), складність інтеграції СКД з існуючими системами безпеки та вартість системи контролю доступу.

Одним з поширених типів біометричних систем аутентифікації у складі сучасних СКД є системи аутентифікації за обличчям [1], зокрема засновані на використанні штучних нейронних мереж (ШНМ). Дані системи дозволяють мінімізувати рівень помилок при збереженні високої швидкості роботи [2]. Проте обмеженням практичного застосування даних систем є високі вимоги щодо якості оброблюваних даних, зокрема необхідність детектування та вирівнювання обличчя людини на фотографії, забезпечення заданого рівня яскравості/контрастності зображення. Це обумовлює необхідність використання обчислювально складних методів обробки даних, що унеможливлює використання ШНМ для автентифікації користувачів в режимі реального часу. Тому актуальною та важливою задачею є розробка методів адаптації систем біометричної автентифікації користувачів за обличчям для роботи з даними, що не проходили процедури попередньої обробки.

Вирішення даної задачі потребує внесення змін до структури застосовуваних ШНМ, наприклад збільшення кількості/ширини шарів штучних нейронів, що призводить до зміни об'ємів обчислень необхідних для роботи системи автентифікації. Це підвищує вимоги до швидкодії пристроїв, залучених до обробки біометричних даних користувачів. З іншого боку, застосування високопотужних пристроїв, наприклад кластерів відеопроцесорів, ускладнює інтеграцію із існуючими системами та збільшує кошторис СКД. Тому представляє інтерес пошук методів адаптації сучасних ШНМ для використання на малопотужних обчислювальних пристроях (МОП), що використовуються в СКД [3]. Вирішення даної задачі потребує попереднього огляду існуючих методів адаптації штучних нейронних мереж для запуску на МОП, що є метою даної роботи.

1. Методи адаптації нейронних мереж для запуску на малопотужних пристроях

За результатами огляду літератури в області пристосування ШНМ для роботи на МОП можливо виділити наступні підходи до вирішення даної задачі [4]:

- Розподілення графу обчислень ШНМ між декількома вузлами автоматизованої системи (АС) — в даному випадку проводиться поділ ШНМ на окремі частини, що розподіляються між обчислювальними вузлами (ОВ) для збільшення швидкості обчислень, при збереженні одного каналу вхідних даних. Дане розділення може проводитися як на рівні окремих шарів

нейронної мережі, так і окремих штучних нейронів.

- Розподілення вхідних даних між декількома ОВ — проводиться паралельна обробка вхідних даних декількома копіями вихідної ШНМ. Для збільшення пропускної спроможності системи обробки, вхідні дані можуть передаватися ШНМ по окремим каналам зв'язку. В більшості випадків даний підхід використовується на етапі налаштування ШНМ для отримання проміжних результатів обчислень.
- Комбінований підхід — поєднує попередні методи, вирішуючи проблеми швидкості обчислень та пропускної спроможності каналів зв'язку. Згідно даного підходу проводиться паралельна обробка вхідних даних з використанням декількох шарів ШНМ, розподілених між окремими ОВ. Прийняття рішення щодо автентифікації користувача проводиться за мажоритарним принципом з врахуванням результатів роботи окремих копій ШНМ.
- Квантування вагів — особливістю даного методу є внесення змін до ШНМ, зокрема бітності представлення вагових коефіцієнтів окремих штучних нейронів. В більшості випадків дані вагові коефіцієнти представлені у вигляді множини чисел із плаваючою комою (32 біти на кожне число), що можуть бути заміщені числами меншої розрядності (8 бітів на кожне число). Це дозволяє зменшити вимоги щодо об'єму пам'яті для роботи ШНМ, що розширює коло пристроїв які можуть застосовуватися для роботи ШНМ.

Дані підходи широко використовуються при налаштуванні систем з використанням ШНМ для роботи на малопотужних обчислювальних пристроях [5]. Розглянемо переваги та обмеження даних підходів більш детально.

1.1. Методи розподілення штучних нейронних мереж

Дані методи засновані на розподілення між декількома ОВ обчислень, що проводяться як окремим шаром ШНМ, так і окремими штучними нейронами в кожному шарі. Таку модель використовують, коли пропускна спроможність C каналу передачі даних значно більша за швидкість обробки пакету даних окремим штучним нейроном V_n:

$$C > V_n. \tag{1}$$

Приклад реалізації даного підходу для роботи на МОП наведений на Рис. 1.

Поділ ШНМ між ОВ (Рис. 1) може бути реалізований шляхом розділення як окремих шарів нейронної мережі, так і окремих груп нейронів між декількома ОВ. Варто зазначити, що обробка вхідних даних буде проводитися одним обчислювальним пристроєм, без поділу ШНМ на окремі складові. Обчислення в проміжних (прихованих) шарах можуть проводитися на окремих ОВ. Результати обробки об'єднуються

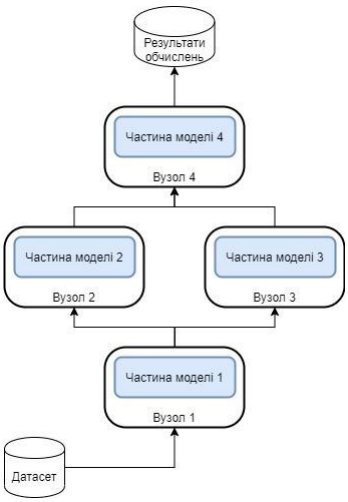

Рис. 1. Схема реалізації підходу розподілення штучної нейронної мережі

на одному обчислювально пристрої для прийняття рішення щодо автентифікації користувача (Рис. 1).

1.2. Методи розподілення даних

Даний підхід заснований на розподілі вхідних даних між декількома копіями ШНМ [4]. Підхід використовується у випадку, коли пропускна спроможність каналу передачі вхідних параметрів C менша за швидкість обробки пакету даних окремим штучним нейроном V_n

$$C < V_n. \tag{2}$$

Схема реалізації даного підходу до розподілення даних зображена на Рис. 2.

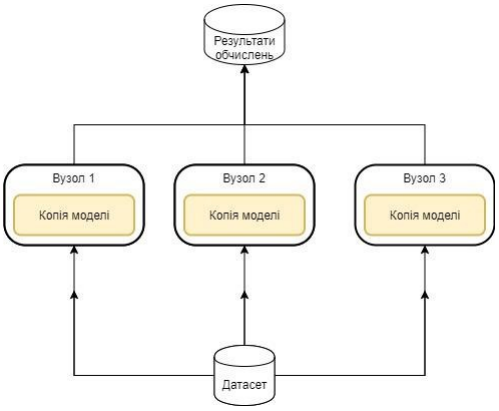

Рис. 2. Схема реалізації підходу розподілення даних

Даний підхід заснований на можливості розділення значних об'ємів вхідних даних між декількома копіями ШНМ з використанням каналів передачі, що мають низьку пропускну здатність. Наприклад, вхідні дані можуть бути розподілені однаково між ОВ (Рис. 2), проте швидкість обробки даних кожним ОВ буде пропорційно залежати як від ширини каналу, так і обчислювальних можливостей ОВ.

Альтернативним підходом до реалізації даного підходу є ієрархічне представлення вхідних даних (біо-

метричних ознак людини), з використанням множини більш простих об'єктів (Рис. 3).

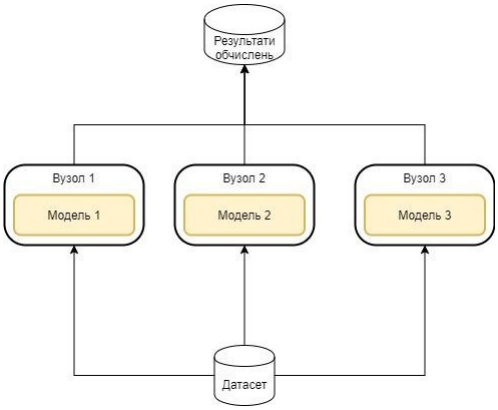

Рис. 3. Схема реалізації ієрархічного розподілення даних

Наприклад, виділення характерних рис обличчя, зокрема зображень очей, вух, носу тощо. В такому випадку, в обчислювальних вузлах системи автентифікації будуть знаходитись ШНМ, налаштовані з використанням конкретного типу даних (зображення очей, вух тощо). При цьому прийняття рішення системою буде проводитися за мажоритарним принципом — переважаючий серед ШНМ прогноз (рішення щодо автентичності користувача) буде використовуватися в якості результату роботи системи. В даному підході використовується припущення, що кожна ознака має однаковий вплив на рішення системи. В деяких випадках, вплив окремих ознак може бути визначальним при прийнятті рішення (наприклад, камера бачить лише частину обличчя), тому для таких випадків може використовуватись принцип вагових коефіцієнтів — результати прогнозів окремих моделей будуть зважені за величиної їх впливу на загальний результат. Вагові коефіцієнти, в свою чергу, можуть визначатись за результатами попереднього налаштування системи.

1.3. Комбінована модель

Комбінована розподілена модель обчислень поєднує переваги розглянутих раніше підходів: розподілення обчислень дозволяє зменшити необхідні обчислювальні потужності, розподілення даних, в свою чергу, дозволяє зменшити пропускну спроможність між джерелом вхідних даних (датасетом) та обчислювальними вузлами. Кількість вузлів, що займаються розподіленням даних, визначається за результатами аналізу пропускної спроможності каналів зв'язку між вузлами системи обробки даних, а кількість обчислювальних вузлів вибирають згідно вимог, щодо швидкодії системи обробки даних [5]

Як зображено на схемі на Рис. 4, вхідні параметри поступають по декільком каналам із використанням методу розподілу даних, на розподілену модель, яка в свою чергу в якості результату видає прогноз.

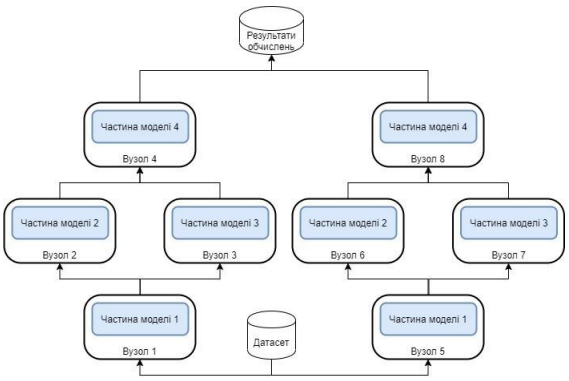

Рис. 4. Схема реалізації комбінованого підходу

1.4. Квантування вагів

Варто зазначити, що розглянуті методи адаптації ШНМ спрямовані на внесення змін до порядку проведення обчислень, без зміни структури самих ШНМ. Це потребує зберігання структури вихідної мережі, що висуває додаткові вимоги щодо розмірів пам'яті, необхідної для роботи ШНМ. Для подолання даного обмеження був запропонований метод квантування вагових коефіцієнтів ШНМ [6].

Даний підхід дозволяє знизити вимоги щодо точності представлення вагових коефіцієнтів штучних нейронів ШНМ при контрольованому зменшенні загальної точності роботи системи обробки даних. Наприклад, переведення чисел із множини 32-бітних чисел із плаваючою комою в 8-бітні цілі числа. Метод квантування вагів широко використовується в реальних системах обробки даних з наступних причин:

1) Використання арифметики з меншою кількістю бітів для представлення чисел дозволяє пришвидшити обчислення на сучасних процесорах [6].
2) Арифметика із 8-бітовими даними зменшує об'єм пам'яті, що виділяється майже в 4 рази порівняно із 32-бітовими числами із плаваючою комою згідно стандарту IEEE 754-2008 [3, 6]. Також зменшується розмір даних, що пересилається по мережі для оновлення вагових коефіцієнтів нейронної мережі, відповідно такі нейронні мережі можуть швидше оновлюватись
3) Менша кількість бітів дозволяє зменшити зменшує частоту доступу до даних та підвищити швидкість роботи за рахунок використання стандартних кешів та регістрів процесору.
4) Малопотужні обчислювальні пристрої не завжди підтримують арифметику із плаваючою комою за рахунок особливостей реалізації архітекури використовуваного процесору. В той же час, підтримка цілих чисел, широко реалізована на даних пристроях [6].

Варто зазначити, що ШНМ, які використовуються в системах біометричної аутентифікації за обличчям, налаштовані виокремлювати та аналізувати ключові ознаки об'єктів в умовах наявності сторонніх шумів. Це дозволяє забезпечити стійкість таких мереж до невеликих змін вагових коефіцієнтів, що виникають

при використанні даного методу. В роботі [6] показано мінімальний вплив квантування вагів окремих нейронів на отримувану точність роботи ШНМ. Це, в поєднанні зі значним зменшенням обсягу пам'яті, споживання енергії та збільшенням обчислювальної швидкості, робить квантування ефективним підходом для розгортання нейронних мереж на МОП.

2. Оцінка швидкодії сучасних алгоритмів адаптації нейронних мереж

В роботі досліджено вплив розглянутих методів адаптації ШНМ на швидкість обробки даних з використанням МОП. В таблиці 1 наведено порівняльну характеристику ШНМ на основі VGG-16, що використовуються в задачах автентифікації користувачів за обличчям. Порівняння швидкості обробки проводилося відносно випадку використання 1 ядра процесору. Дослідження проводились для процесору Intel Core i7-3770 із базовою тактовою частотою 3.4 ГГц та графічним прискорювачем Nvidia GeForce GTX Titan із базовою тактовою частотою 876 МГц.

Табл. 1. Порівняння швидкодії обробки вхідних даних при розподілі нейронної мережі між декількома обчислювальними пристроями

Модель процесору	Кількість залучених ядер	Приріст швидкості
Intel i7	1	1
Intel i7	4	1,85
GeForce GTX Titan	3072	23,27

Застосування графічного процесору GeForce GTX Titan, що складається із порівняно малопотужних ядер, дозволяє збільшити швидкодію роботи ШНМ в 23.27 рази (Табл. 1) відносно випадку використання одного ядра процесору Intel i7.

Для порівняння був проведений аналіз впливу методу квантування вагів [6] на кількість пам'яті, необхідної для роботи ШНМ (Табл. 2).

Табл. 2. Порівняння швидкості обчислення сучасних нейронних мереж до та після квантування вагових коефіцієнтів

Модель мережі	Помил., %	Розмір пам'яті	Стисн., разів
LeNet-300-100	1,64	1070 KB	40
LeNet-300-100 кв	1,58	27 KB	
LeNet-5	0,80	1720 KB	39
LeNet-5 кв	0,74	44 KB	
AlexNet	42,78	249 MB	35
AlexNet кв	42,78	6,9 MB	
VGG-16	31,50	552 MB	49
VGG-16 кв	31,17	11,3 MB	

За результатами дослідження встановлено, що використання методу квантування вагів дозволяє зменшити використання пам'яті на $30 - 40$ разів у порівнянні з вихідними ШНМ без суттєвих змін точності їх роботи.

Висновки

В роботі проведено аналітичний огляд сучасних методів адаптації штучних нейронних мереж для використання на малопотужних обчислювальних пристроях. Зокрема розглянуто методи реалізації паралельних обчислень, а саме розподілення складових штучних нейронних мереж між окремими обчислюваними вузлами, метод розподілення даних, а також комбінований метод, що поєднує розподілення вхідних даних та складових штучних нейронних мереж. Дані методи дозволяють підвищити швидкість обробки даних адаптованими штучними нейронними мережами при збереженні структури даних мереж.

Для зменшення вимог щодо об'єму пам'яті, необхідного для збереження штучних нейронних мереж розглянуто використання методу квантування вагів. Даний метод заснований на зменшенні бітності представлення вагових коефіцієнтів при контрольованій зміні точності роботи штучних нейронних мереж. За результатами експериментальних досліджень швидкодії та об'ємів використання пам'яті адаптованими ШНМ показано, що комплексне застосування декількох методів адаптації забезпечує високу точність роботи ШНМ при використанні малопотужних обчислювальних пристроїв.

Перелік використаних джерел

1. Castellotti Ricardo, Canali Luca. Distributed Deep Learning for Physics with TensorFlow and Kubernetes. — Access mode: `https://db-blog.web.cern.ch/blog/luca-canali/2020-03-distributed-deep-learning-physics-tensorflow-and-kubernetes` (online; accessed: 04.30.2021).
2. Струнін І.В., Кісіоглова А.І., Прогонов Д.О. Огляд сучасних методів автентифікації користувача за геометрією обличчя // Теоретичні і прикладні проблеми фізики, математики та інформатики: матеріали XVIII Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених. — 2020. — 05.
3. Verhelst M., Moons B. Embedded Deep Neural Network Processing: Algorithmic and Processor Techniques Bring Deep Learning to IoT and Edge Devices // Solid-State Circuits Magazine, vol. 9. — 2020. — 05. — P. 55–65.
4. Offer Steffen. dKeras: Make Keras up to 30x faster with a few lines of code. — Access mode: `https://medium.com/@offercstephen/dkeras-make-keras-faster-with-a-few-lines-of-code-a1792b12dfa0` (online; accessed: 04.30.2021).
5. Improving the speed of neural networks on CPUs // Deep Learning and Unsupervised Feature Learning Workshop. — 2011.
6. Han Song. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // Cornell university conference. — 2015.