УДК 004.27:004.8

В. А. Шатайло

РОЛЬ СПЕЦІАЛІЗОВАНИХ NPU У ПРИСКОРЕННІ ОБЧИСЛЕНЬ ШТУЧНОГО ІНТЕЛЕКТУ

Вінницький Національний Технічний Університет

Анотація

У роботі проаналізовано нейронні процесори (NPU) як спеціалізоване рішення для прискорення алгоритмів штучного інтелекту. Розглянуто ключові архітектурні відмінності NPU від традиційних CPU та GPU, зокрема принципи потоку даних та квантування, що забезпечують високу швидкодію. Визначено роль цих процесорів у розвитку технологій Edge AI, які дозволяють виконувати енергоефективні та безпечні обчислення безпосередньо на пристроях, зменшуючи залежність від хмарних сервісів.

Ключові слова: нейронний процесор, NPU, штучний інтелект, енергоефективність, глибоке навчання.

Abstract

This paper analyzes Neural Processing Units (NPUs) as a specialized solution for accelerating artificial intelligence algorithms. Key architectural differences between NPUs and traditional CPUs/GPUs are examined, specifically dataflow principles and quantization, which ensure high performance. The study defines the role of these processors in the development of Edge AI technologies, enabling energy-efficient and secure on-device computing while reducing reliance on cloud services.

Keywords: neural processor, NPU, artificial intelligence, energy efficiency, deep learning.

Вступ

Інтеграція алгоритмів штучного інтелекту в повсякденні технології створила безпрецедентне навантаження на апаратне забезпечення. Традиційна архітектура комп'ютерів, яка десятиліттями розвивалася для виконання послідовних логічних операцій, виявилася неоптимальною для специфічних потреб машинного навчання. Нейронні мережі вимагають одночасної обробки величезних масивів даних, що для класичних процесорів стає «вузьким місцем». Це зумовило необхідність переходу від універсальних рішень до неоднорідних обчислювальних систем, де різні типи процесорів спеціалізуються на конкретних класах завдань заради досягнення максимальної продуктивності та енергоефективності [1].

Огляд та аналіз

Щоб зрозуміти необхідність появи нового класу пристроїв, варто розглянути обмеження існуючих компонентів. Центральний процесор (CPU) — це універсальний інструмент, оптимізований для послідовної обробки складних інструкцій [2]. Його архітектура орієнтована на мінімізацію затримок (latency) при виконанні різних задач — від запуску операційної системи до керування логікою програм. CPU має потужні блоки передбачення розгалужень і великі кеші, що дозволяє йому ефективно працювати в умовах невизначеності. Проте, коли мова йде про монотонні математичні операції над мільйонами точок даних, універсальність CPU стає його недоліком: він витрачає надто багато ресурсів на керування процесом, а не на саме обчислення. Графічний процесор (GPU) частково вирішив цю проблему. Його архітектура побудована за принципом SIMD (Single Instruction, Multiple Data — «одна інструкція, багато даних»). Маючи тисячі обчислювальних ядер, GPU здатний паралельно виконувати ідентичні операції, що ідеально підходить як для рендерингу графіки, так і для тренування нейронних мереж. Однак GPU все ще залишається відносно універсальним пристроєм, спроєктованим для роботи з числами високої точності (наприклад, FP32), що є надлишковим для багатьох завдань експлуатації вже навчених моделей.

Нейронний процесор (NPU) становить наступний еволюційний етап у розвитку мікроелектроніки, являючи собою яскравий приклад доменно-специфічної архітектури. Якщо універсальні процесори змушені підтримувати сумісність із застарілим кодом та виконувати різнорідні інструкції, то NPU спроєктовано виключно для фізичного прискорення тензорних операцій [3]. В основі роботи будь-якої глибокої нейронної мережі лежать дві фундаментальні математичні дії — перемноження матриць та

згортка, і NPU перетворює ці абстрактні моделі на апаратну реальність, виконуючи їх з максимальною ефективністю. Ключова архітектурна відмінність, що дозволяє досягати такої продуктивності, полягає у відмові від класичної моделі фон Неймана на користь архітектури потоку даних. У традиційному центральному процесорі обробка інформації часто страждає від так званого «вузького місця пам'яті», коли чіп витрачає більше часу та енергії на переміщення даних з оперативної пам'яті до регістрів і назад, ніж на саме обчислення. NPU вирішує цю проблему радикально, адже його серцем є масиви тисяч обчислювальних модулів, відомих як блоки множення-накопичення (MAC) [4]. Вони організовані таким чином, що результати обчислення одного блоку передаються безпосередньо сусідам, не звертаючись до основної пам'яті, що нагадує конвеєр на заводі, де дані проходять через чіп хвилею, обробляючись на кожному кроці за один тактовий цикл.

Окрім архітектурної оптимізації, NPU ефективно використовують техніку квантування. Традиційні обчислення в науці чи графіці зазвичай вимагають високої 32-бітної або 64-бітної точності чисел з плаваючою комою для уникнення похибок, проте природа нейронних мереж є ймовірнісною, тому для розпізнавання об’єктів алгоритму не потрібна математична точність до десятого знаку після коми. NPU апаратно оптимізовані для роботи зі спрощеними форматами даних, наприклад, 8-бітними цілими числами, і таке зменшення розрядності дозволяє не лише економити пропускну здатність пам'яті, але й розміщувати на кристалі значно більше обчислювальних блоків, кратно збільшуючи продуктивність без зростання енергоспоживання.

У практичній площині ця технологія стала фундаментом концепції Edge AI, тобто штучного інтелекту на периферії [5]. Завдяки інтеграції NPU в системи на кристалі, сучасні смартфони, розумні камери та IoT-пристрої отримали автономність від хмарних серверів, що вирішує одразу кілька критичних проблем. По-перше, це гарантує приватність, оскільки біометричні дані або приватні розмови обробляються локально і не залишають пристрій. По-друге, зникає проблема затримки, що критично важливо, наприклад, для систем автономного водіння, де автомобіль не може чекати відповіді від сервера і має приймати рішення миттєво. Нарешті, локальна обробка забезпечує високу енергоефективність, адже вона витрачає значно менше енергії, ніж постійна передача великих обсягів інформації через мобільні мережі.

Висновки

Підсумовуючи, можна стверджувати, що виокремлення нейронних процесорів у самостійний обчислювальний клас стало не просто інженерним рішенням, а необхідною відповіддю на запит сучасності. Цей крок дозволив подолати фізичні обмеження універсальних архітектур CPU та GPU, забезпечивши критично важливий баланс між високою продуктивністю алгоритмів штучного інтелекту та енергоефективністю. У найближчому майбутньому роль NPU лише посилюватиметься: вони можуть стати обов’язковим стандартом для всіх типів електроніки, від найпростіших побутових датчиків до високопродуктивних робочих станцій. Очікується, що наступні покоління цих чіпів стануть настільки потужними, що дозволять запускати складні генеративні моделі та великі мовні системи локально, без жодного звернення до хмарних серверів. Такий вектор розвитку остаточно закріпить перехід до гібридних обчислень, де межа між можливостями суперкомп’ютера та кишенькового гаджета ставатиме дедалі менш помітною, роблячи технології ще більш автономними та приватними.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. CPU vs GPU vs NPU: What's the difference? [Електронний ресурс]. – Режим доступу: https://www.corsair.com/us/en/explorer/diy-builder/power-supply-units/cpu-vs-gpu-vs-npu-whats-the-difference/

2. A guide to CPU, GPU, NPU, and Windows [Електронний ресурс]. – Режим доступу: https://www.microsoft.com/en-us/windows/learning-center/cpu-gpu-npu-windows

3. ШІ-прискорювач [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/ШІ-прискорювач

4. Architecture for ML [Електронний ресурс]. – Режим доступу: https://developers.google.com/coral/guides/hardware/arch_ml

5. What is a neural processing unit (NPU)? [Електронний ресурс]. – Режим доступу: https://www.ibm.com/think/topics/neural-processing-unit

Шатайло В'ячеслав Андрійович — студент групи 2КІ-25м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький Національний Технічний Університет, Вінниця, e-mail: viacheslavshatailo@gmail.com

Shatailo Viacheslav Andriyovych — student of group 2KI-25m, faculty of information technologies and computer engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: viacheslavshatailo@gmail.com