УДК 621.396

М. В. Васильківський[1]
Н. В. Грабчак[1]
М. І. Антонюк[1]

ОПТИМІЗАЦІЯ МЕРЕЖ 5G ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

[1] Вінницький національний технічний університет

Анотація

У роботі досліджено сучасні методи підвищення продуктивності алгоритмів штучного інтелекту, що виконуються безпосередньо на мобільних та вбудованих пристроях. Розглянуто спеціалізовані підходи до оптимізації, включаючи layer-wise профілювання, ефективне виконання операцій згортки через General Matrix Multiplication (GEMM), operator fusion для зменшення звернень до пам'яті, низькоточні обчислення (FP16, INT8) та оптимізацію використання оперативної пам'яті за допомогою liveness analysis і спільних пулів пам'яті. Досліджено переваги інтеграції сучасних архітектур глибокого навчання, таких як Transformer та генеративні моделі (GAN), для прогнозування параметрів мереж і синтезу навчальних даних. Зроблено висновки щодо підвищення швидкодії інференсу, скорочення обсягів пам'яті та енергоспоживання, а також покращення ефективності використання обчислювальних ресурсів GPU та апаратних прискорювачів AI.

Ключові слова: штучний інтелект, глибоке навчання, мобільний та вбудований пристрій, оптимізація продуктивності, квантизація, управління пам'яттю.

Abstract

This work investigates modern methods for enhancing the performance of artificial intelligence algorithms executed directly on mobile and embedded devices. Specialized optimization approaches are considered, including layer-wise profiling, efficient execution of convolution operations using General Matrix Multiplication (GEMM), operator fusion to reduce memory accesses, low-precision computations (FP16, INT8), and memory usage optimization through liveness analysis and shared memory pools. The advantages of integrating modern deep learning architectures, such as Transformers and generative models (GANs), for network parameter prediction and training data synthesis are explored. Conclusions are drawn regarding improved inference speed, reduced memory footprint and energy consumption, and enhanced efficiency in utilizing GPU and AI accelerator resources.

Keywords: artificial intelligence, deep learning, mobile and embedded devices, performance optimization, quantization, memory management.

Вступ

Сучасний етап розвитку інформаційно-комунікаційних технологій характеризується стрімким зростанням обсягів передавання даних, розширенням спектра цифрових сервісів та збільшенням кількості підключених пристроїв. Це обумовлює підвищені вимоги до продуктивності, надійності та ефективності функціонування телекомунікаційних мереж. У цьому контексті важливу роль відіграють мобільні мережі п'ятого покоління (5G), які забезпечують високі швидкості передавання даних, мінімальні затримки та підтримку масового підключення пристроїв [1].

Технології 5G створюють основу для розвитку таких перспективних напрямів, як Інтернет речей (IoT), автономний транспорт, інтелектуальні міські системи, дистанційна медицина, хмарні обчислення та системи доповненої і віртуальної реальності. Проте зростання складності мережевої інфраструктури та значні обсяги трафіку потребують нових підходів до управління мережами та розподілу ресурсів. Одним із найбільш перспективних інструментів вирішення цих завдань є використання методів штучного інтелекту. Алгоритми машинного навчання та аналізу даних дозволяють автоматизувати процеси управління мережевими ресурсами, прогнозувати навантаження, оптимізувати маршрутизацію трафіку, покращувати якість обслуговування користувачів та підвищувати ефективність використання радіочастотного спектра.

Інтеграція технологій штучного інтелекту в архітектуру мереж 5G забезпечує можливість створення інтелектуальних телекомунікаційних систем, здатних адаптуватися до змінних умов функціонування мережі, автоматично виявляти несправності та оптимізувати параметри роботи у реальному часі. Це сприяє підвищенню продуктивності мережі, зниженню затримок передачі даних та забезпеченню стабільної якості зв'язку [2].

У зв'язку з цим дослідження методів оптимізації мереж 5G із використанням технологій штучного інтелекту є актуальним науково-практичним завданням, спрямованим на підвищення ефективності функціонування сучасних бездротових телекомунікаційних систем [3].

Метою роботи є дослідження можливостей застосування методів штучного інтелекту для оптимізації функціонування мереж п'ятого покоління та підвищення ефективності використання мережевих ресурсів.

Результати дослідження

Сучасні наукові дослідження у сфері бездротових телекомунікацій свідчать про активний розвиток методів оптимізації мереж п'ятого покоління із застосуванням технологій штучного інтелекту. Зростання складності архітектури 5G-мереж, збільшення обсягів мережевого трафіку та необхідність підтримки великої кількості підключених пристроїв зумовлюють потребу у використанні інтелектуальних алгоритмів управління мережевими ресурсами [1].

У сучасних наукових роботах розглядаються різні підходи до оптимізації функціонування мереж 5G. Значна увага приділяється застосуванню алгоритмів машинного навчання та глибокого навчання для управління мережевими ресурсами, прогнозування навантаження та оптимізації маршрутизації трафіку. Дослідження показують, що традиційні методи керування мережею не завжди здатні ефективно адаптуватися до динамічних умов функціонування телекомунікаційних систем, тоді як алгоритми штучного інтелекту забезпечують більш гнучке та адаптивне управління.

У низці робіт розглядається використання методів глибокого навчання та підкріплювального навчання (Reinforcement Learning) для оптимізації розподілу ресурсів, управління трафіком та реалізації механізмів network slicing у мережах 5G. Такі підходи дозволяють у реальному часі аналізувати стан мережі, прогнозувати навантаження та ефективно розподіляти пропускну здатність між різними сервісами та користувачами. Це сприяє зменшенню затримок, підвищенню пропускної здатності та забезпеченню необхідної якості обслуговування (QoS) [2, 3].

Окремі дослідження присвячені інтеграції штучного інтелекту на різних рівнях архітектури мережі — у радіомережі доступу (RAN), периферійних обчислювальних вузлах (edge computing) та ядрі мережі. Використання AI/ML-алгоритмів дозволяє реалізувати концепцію самоорганізованих мереж (Self-Organizing Networks), які автоматично виконують оптимізацію параметрів мережі, балансування навантаження та виявлення несправностей.

Також значну увагу приділено розробці інтелектуальних моделей прогнозування мережевих ресурсів. Наприклад, сучасні дослідження пропонують використання гібридних нейронних мереж, зокрема моделей CNN-BiLSTM, для прогнозування потреб у мережевих ресурсах і підвищення ефективності їх розподілу. Результати таких досліджень демонструють високу точність прогнозування та можливість застосування цих моделей у реальних телекомунікаційних системах [1, 2].

У наукових оглядових роботах також зазначається, що використання технологій штучного інтелекту є одним із ключових напрямів розвитку бездротових мереж покоління Beyond 5G та майбутніх мереж 6G. AI-алгоритми дозволяють вирішувати складні задачі управління мережею, включаючи управління мобільністю користувачів, оптимізацію радіочастотного спектра, зменшення завад та підвищення енергоефективності мережевої інфраструктури [3].

Таким чином, аналіз сучасних наукових публікацій показує, що застосування технологій штучного інтелекту є одним із найбільш перспективних напрямів підвищення ефективності функціонування телекомунікаційних мереж. Незважаючи на значну кількість досліджень у цій галузі, питання розробки ефективних методів оптимізації мереж 5G з використанням алгоритмів штучного інтелекту залишаються актуальними та потребують подальшого дослідження [3].

Підвищення продуктивності алгоритмів штучного інтелекту, що виконуються безпосередньо на мобільних або вбудованих пристроях, потребує застосування спеціалізованих методів опти-

мізації. Основою таких оптимізацій є глибоке розуміння обчислювально інтенсивних операцій, які використовуються в математичних моделях глибокого навчання. Для аналізу продуктивності нейронних мереж застосовують профілювання окремих шарів моделі (layer-wise profiling), що дозволяє визначити, які саме операції потребують найбільших обчислювальних ресурсів під час виконання моделі [2].

У більшості сучасних архітектур глибоких нейронних мереж операції згортки реалізуються за допомогою узагальненого множення матриць (General Matrix Multiplication, GEMM). Ефективність виконання таких операцій значною мірою залежить від оптимальної організації доступу до пам'яті процесора, зокрема правильного завантаження матриць у кеш-пам'ять і використання векторизованих інструкцій для прискорення обчислень.

Додаткового підвищення продуктивності можна досягти шляхом паралельного виконання обчислень із використанням багатопотокових технологій та багатоядерних обчислювальних систем, які входять до складу сучасних систем-на-кристалі (SoC). Значне прискорення обчислень забезпечують графічні процесори (GPU), які оптимізовані для паралельної обробки великих масивів даних.

Використання оптимізованих обчислювальних ядер, реалізованих, наприклад, за допомогою технології OpenCL, може забезпечити підвищення продуктивності приблизно у десять разів. Ще більшого прискорення можна досягти завдяки використанню спеціалізованих апаратних прискорювачів штучного інтелекту, які реалізують відповідні обчислювальні операції на рівні апаратного забезпечення.

Одним із ефективних методів оптимізації виконання нейронних мереж є об'єднання декількох операцій в одну обчислювальну процедуру, що отримало назву operator fusion або kernel fusion. Цей підхід широко використовується у сучасних програмних платформах для виконання моделей глибокого навчання, таких як TensorFlow, Apache TVM та Mobile Neural Network.

Сутність методу полягає у поєднанні кількох послідовних операцій нейронної мережі в одну узагальнену операцію [3].

Ще одним важливим напрямом оптимізації є використання обчислень зі зниженою точністю представлення чисел. Ефективність математичних обчислень значною мірою залежить від кількості бітів, які використовуються для представлення числових значень.

Під час навчання моделей глибокого навчання традиційно використовується формат чисел із плаваючою комою з 32-бітною точністю (FP32). Проте дослідження показують, що для багатьох задач достатньою є 16-бітна точність (FP16), що дозволяє суттєво зменшити обчислювальні витрати. Для етапу виконання моделі (inference) часто застосовуються ще компактніші формати представлення даних, зокрема 8-бітні цілі числа (INT8) або навіть 4-бітні значення.

Зменшення розрядності чисел має кілька важливих переваг: прискорення виконання арифметичних операцій; зменшення обсягу пам'яті, необхідної для зберігання моделі; ефективніше використання кеш-пам'яті; зниження енергоспоживання.

Перехід від високої точності представлення даних до низької реалізується за допомогою процедури квантизації (quantization). Основними підходами до реалізації квантизації є: квантизація після завершення навчання (post-training quantization); навчання моделі з урахуванням квантизації (quantization-aware training).

Для ефективного виконання алгоритмів глибокого навчання на мобільних та вбудованих пристроях важливо також оптимізувати використання оперативної пам'яті. Це особливо актуально для задач обробки зображень високої роздільної здатності, наприклад у системах покращення якості зображень, підвищення роздільної здатності (super-resolution) або обробки зображень в умовах низької освітленості.

У таких задачах обробка кожного пікселя потребує значних обчислювальних ресурсів та обсягів пам'яті. Тому для ефективного виконання моделей застосовуються спеціальні методи управління пам'яттю. Одним із таких підходів є аналіз життєвого циклу буферів даних (liveness analysis) з подальшим використанням спільного пулу пам'яті [3].

Висновки

Сучасні досягнення у сфері застосування штучного інтелекту в бездротових комунікаціях створюють передумови для формування комунікаційних систем нового покоління, які поєдну-

ють високий рівень надійності, адаптивності та інтелектуальності. Використання сучасних алгоритмів машинного навчання дозволяє ефективно вирішувати низку складних задач у бездротових мережах, зокрема оптимізацію використання радіоресурсів, підвищення якості обслуговування користувачів та адаптивне управління мережею.

Особливий інтерес для подальших досліджень становлять моделі глибокого навчання, побудовані на основі архітектури Transformer neural network architecture, які характеризуються здатністю враховувати контекстну інформацію та ефективно обробляти складні послідовності даних. Завдяки механізмам уваги (attention mechanisms) такі моделі можуть забезпечувати більш гнучке та точне прогнозування параметрів мережі, що є важливим для інтелектуального управління бездротовими системами зв'язку.

Крім того, перспективним напрямом є дослідження можливостей застосування генеративних моделей, зокрема моделей на основі Generative Adversarial Network, для вдосконалення різних етапів проєктування та оптимізації комунікаційних систем. Генеративні нейронні мережі можуть використовуватися для синтезу навчальних даних, моделювання умов поширення сигналів або створення нових алгоритмів обробки сигналів у бездротових мережах.

З огляду на стрімке зростання складності сучасних моделей глибокого навчання, що характеризуються значною кількістю параметрів і високими вимогами до обчислювальних ресурсів, актуальним напрямом досліджень є оптимізація їх виконання на периферійних пристроях. Особливої уваги потребує розроблення методів ефективного використання обчислювальних можливостей графічних процесорів та спеціалізованих апаратних прискорювачів штучного інтелекту.

Таким чином, подальший розвиток досліджень у галузі штучного інтелекту для бездротових комунікацій має бути спрямований на створення нових алгоритмів глибокого навчання, оптимізацію їх реалізації на пристроях з обмеженими ресурсами, а також інтеграцію інтелектуальних методів управління в архітектуру телекомунікаційних мереж наступних поколінь. Це сприятиме підвищенню ефективності функціонування мереж, покращенню якості сервісів та розширенню можливостей використання інтелектуальних телекомунікаційних технологій.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Singh S., Wu Y., Rao G.N.S., Joshi K., Barnaghi P., Kanagarathinam M.R. AI in Wireless for Beyond 5G Networks. CRC Press, 1st Edition, 2024. xxi + 210 p. (Hardcover), ISBN: 9781032301211/9781032301228.

2. Васильківський М. В. Керування телекомунікаційними мережами з використанням технологій AI/ML [Текст] / М. В. Васильківський, О. Болдирева, Г. Варгатюк, М. Будаш // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2023. – Вип. 1. – С. 89–100.

3. Васильківський М. В. Коригування параметрів мобільних систем MIMO із використанням штучного інтелекту [Текст] / М. Васильківський, О. Болдирева, Г. Варгатюк, Н. Грабчак // Комп`ютерно-інтегровані технології: освіта, наука, виробництво. – 2023. – № 51. – С. 139-147.

Васильківський Микола Володимирович — кандидат технічних наук, доцент, доцент кафедри інфокомунікаційних систем і технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: mvasylkivskyi@gmail.com

Грабчак Назарій Віталійович — аспірант групи 172-23а, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця, e-mail: nazariihrabchak@gmail.com

Антонюк Марія Ігорівна — студентка групи ПЗТ-22б, факультет інформаційних електронних систем, Вінницький національний технічний університет, Вінниця, e-mail: alsomahsa@gmail.com

Vasylkivskyi Mykola V. — candidate of technical sciences, associate professor, associate professor of the Department of Information Communication Systems and Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: mvasylkivskyi@gmail.com

Hrabchak Nazarii V. — graduate student of group 172-23a, Faculty of Information Electronic Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: nazariihrabchak@gmail.com

Antonyuk Mariia I. - student of group TSS-22b, Faculty of Information Electronic Systems, Vinnytsia National Technical University, Vinnytsia, e-mail: alsomahsa@gmail.com