https://doi.org/10.31891/2307-5732-2025-355-76
УДК 004.9

СЕРЕДЮК ГЛІБ
Вінницький національний технічний університет
https://orcid.org/0009-0000-3010-6437
e-mail: glebserediuk@gmail.com
ВОЛОДИМИР ГАРМАШ
Вінницький національний технічний університет
https://orcid.org/0009-0007-1861-8772
e-mail: garmash.v.v@edu.ua

ОПТИМІЗАЦІЯ СПОЖИВАННЯ ПАМ'ЯТІ ПІД ЧАС РОБОТИ З НЕЙРОМЕРЕЖАМИ ДЛЯ СКАНУВАННЯ

У цій статті досліджуються методи оптимізації пам'яті для глибоких нейронних мереж у завданнях сканування зображень та розпізнавання тексту. Дослідження присвячене застосуванню нейронних мереж в оптичному розпізнаванні символів (OCR), аналізі документів, скануванні штрих-кодів та ідентифікації QR-кодів — всіх критично важливих компонентах сучасних систем сканування. Через обмеження пам'яті мобільних та вбудованих пристроїв оптимізація цих моделей є надзвичайно важливою для практичного застосування. У дослідженні систематично проаналізовано три підходи до оптимізації пам'яті: обрізка мережі, квантування ваги та методи стиснення моделі. Обрізка мережі усуває зв'язки з незначними значеннями ваги, перетворюючи щільні матриці ваги на розріджені представлення. Квантування зменшує точність представлення ваги з 32-бітних чисел з плаваючою комою до 8-бітних цілих чисел, зменшуючи розмір моделі в чотири рази. Кодування Хаффмана забезпечує додаткове стиснення, присвоюючи коротші коди часто зустрічаються значенням ваги. Експериментальні результати підтверджують, що комбінований підхід, що інтегрує обрізку, квантування та кодування Хаффмана, може зменшити розмір моделі в 35-49 разів, зберігаючи погіршення точності нижче 1%. Детальний порівняльний аналіз алгоритмів квантування після навчання (PTQ) та навчання з урахуванням квантування (QAT) показує, що QAT дає кращі результати у збереженні точності (втрата 0,3% проти 0,5% для PTQ). Для ResNet-50, адаптованого для сканування документів, поєднання 90% обрізки зв'язків з 8-бітним QAT зменшує вимоги до пам'яті в 40 разів, втрачаючи лише 0,9% точності. Практичні наслідки включають значне зниження енергоспоживання (56%) і підвищення швидкості виведення (43%), що робить ці методи оптимізації особливо цінними для портативних скануючих пристроїв, що працюють в умовах обмеженого часу. Дослідження демонструє, що навіть складні архітектури нейронних мереж можуть бути ефективно розгорнуті на пристроях з обмеженими ресурсами за допомогою відповідних методів оптимізації.

Ключові слова: глибокі нейронні мережі, квантизація, обрізання моделей, компресія, сканування зображень, оптичне розпізнавання символів, оптимізація пам'яті

SEREDIUK HLIB
Vinnytsia National Technical University
GARMASH VOLODYMYR
Vinnytsia National Technical University

OPTIMIZATION OF MEMORY CONSUMPTION WHEN WORKING WITH NEURAL NETWORKS FOR SCANNING

This paper investigates memory optimization methods for deep neural networks in image scanning and text recognition tasks. The research examines the application of neural networks in optical character recognition (OCR), document analysis, barcode scanning, and QR code identification—all critical components of modern scanning systems. Due to memory constraints on mobile and embedded devices, optimizing these models is essential for practical deployment. The study systematically analyzes three approaches to memory optimization: network pruning, weight quantization, and model compression techniques. Network pruning eliminates connections with negligible weight values, converting dense weight matrices into sparse representations. Quantization reduces the precision of weight representation from 32-bit floating-point numbers to 8-bit integers, decreasing model size by a factor of four. Huffman coding provides additional compression by assigning shorter codes to frequently occurring weight values. Experimental results confirm that a combined approach integrating pruning, quantization, and Huffman coding can reduce model size by 35-49 times while maintaining accuracy degradation below 1%. Detailed comparative analysis of Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) algorithms reveals that QAT produces superior results in preserving accuracy (0.3% loss versus 0.5% for PTQ). For ResNet-50 adapted for document scanning, the combination of 90% connection pruning with 8-bit QAT reduces memory requirements by 40 times while sacrificing only 0.9% accuracy. The practical implications include significantly reduced energy consumption (56%) and improved inference speed (43%), making these optimization methods particularly valuable for portable scanning devices operating under real-time constraints. The research demonstrates that even complex neural network architectures can be effectively deployed on resource-constrained devices through appropriate optimization techniques.

Keywords: deep neural networks, quantization, model pruning, compression, image scanning, optical character recognition, memory optimization.

Стаття надійшла до редакції / Received 06.05.2025
Прийнята до друку / Accepted 06.06.2025

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Сучасні системи сканування та розпізнавання зображень активно використовують глибокі нейронні мережі для вирішення широкого спектру задач: розпізнавання тексту (OCR), сканування документів, аналізу

медичних знімків, розпізнавання штрих-кодів та QR-кодів. Ці системи демонструють високу точність завдяки застосуванню складних архітектур згорткових нейронних мереж (CNN), таких як ResNet-50, VGG-16 чи MobileNet, проте потребують значних обчислювальних ресурсів і пам'яті для зберігання мільйонів параметрів [1, 2].

Типова архітектура ResNet-50 потребує близько 98 МБ пам'яті для зберігання параметрів, що створює суттєві обмеження для її застосування на мобільних пристроях або вбудованих системах. Ця проблема особливо актуальна для систем сканування, які часто реалізуються на портативних пристроях із обмеженими ресурсами, таких як смартфони, планшети або спеціалізовані сканери [3].

Актуальність розробки ефективних методів оптимізації пам'яті зростає з тенденцією до переміщення процесів обробки даних безпосередньо на кінцеві пристрої замість серверної обробки в хмарі. Такий підхід забезпечує низьку затримку, приватність та можливість роботи без постійного доступу до мережі, що критично для багатьох застосувань сканування в реальному часі [4, 5].

Аналіз досліджень та публікацій

Проблема оптимізації нейронних мереж для задач сканування активно досліджується в останні роки. Han та співавтори [1] запропонували метод "Deep Compression", який включає обрізання зв'язків з низькими вагами, квантизацію та кодування Гаффмана. В експериментах з архітектурою AlexNet їм вдалося зменшити розмір моделі з 240 МБ до 6,9 МБ при зниженні точності лише на 0,58%, що особливо важливо для систем сканування, де точність критично важлива.

Jacob та співавтори [2] провели ґрунтовне дослідження квантизації для ефективного виведення на цілочисельних операціях, порівнюючи квантизацію після навчання (PTQ) та квантизацію з урахуванням навчання (QAT). Їхні експерименти з CNN показали, що QAT забезпечує меншу втрату точності (0,2-0,3%) порівняно з PTQ (0,5-1,0%), що є суттєвим для задач розпізнавання тексту в системах сканування.

Cheng та співавтори [3] представили огляд методів компресії і прискорення нейронних мереж, систематизуючи підходи до обрізання, квантизації, низькорівневої факторизації та дистилювання знань, однак їх дослідження не фокусувалося на специфічних вимогах систем сканування.

Sze та співавтори [4] проаналізували енергоспоживання, затримку та розмір моделей на різних платформах, підкреслюючи компроміс між точністю та ефективністю, що особливо важливо для портативних пристроїв сканування з обмеженим ресурсом батареї.

Krishnamoorthi [5] представив практичний посібник з квантизації CNN, описуючи методи PTQ з калібруванням діапазону ваг та QAT з імітацією квантизації під час навчання, що можна безпосередньо застосувати до систем OCR. Blalock та співавтори [6] оцінили сучасні методи обрізання нейромереж, показавши, що обрізання до 90% параметрів може зберігати точність із втратою менше 1%.

Незважаючи на значний прогрес у методах оптимізації нейронних мереж, специфіка їх застосування для задач сканування залишається недостатньо дослідженою, особливо щодо компромісу між зменшенням розміру моделі, швидкістю роботи та точністю розпізнавання різних типів сканованих документів.

Формулювання цілей статті

Метою роботи є: дослідження ефективності методів оптимізації пам'яті для нейронних мереж, що використовуються в задачах сканування; аналіз впливу обрізання зв'язків, квантизації параметрів та кодування Гаффмана на точність розпізнавання різних типів сканованих документів; порівняння алгоритмів квантизації після навчання (PTQ) та квантизації з урахуванням навчання (QAT) для архітектур нейронних мереж, що застосовуються в системах сканування; визначення оптимального компромісу між зменшенням розміру моделі, швидкістю роботи та збереженням точності для типових задач сканування та розпізнавання.

Виклад основного матеріалу

Методика експериментальних досліджень передбачала порівняльний аналіз різних методів оптимізації пам'яті нейронних мереж для архітектур, що використовуються в системах сканування: ResNet-50 та MobileNet, адаптованих для задач розпізнавання тексту та аналізу документів. Для кожної архітектури досліджувалися наступні методи оптимізації: квантизація після навчання (PTQ), квантизація з урахуванням навчання (QAT), обрізання зв'язків різної інтенсивності, а також комбіновані підходи.

Обрізання нейронної мережі базується на видаленні зв'язків з найменшими за абсолютним значенням вагами. Процес обрізання можна описати формулою:

$$w\_ij = 0, якщо\ |w\_ij| < \theta \quad (1)$$

де θ – порогове значення, яке визначає ступінь обрізання. В наших експериментах було досліджено різні рівні обрізання: від 50% до 90% зв'язків. Для задач сканування документів важливо зберегти здатність мережі розпізнавати дрібні деталі, тому обрізання проводилося з урахуванням важливості різних шарів мережі для розпізнавання тексту та структурних елементів документів.

Квантизація передбачає зменшення точності представлення параметрів моделі з 32-бітних чисел з плаваючою комою (float32) до 8-бітних цілих чисел (int8), що зменшує розмір моделі в 4 рази. Процес квантизації можна описати формулою:

$$w\_int8 = round((w\_float32 - z)/s) \quad (2)$$

де s – масштабний коефіцієнт, z – зсув. Для PTQ ці параметри визначаються після навчання моделі, а для QAT – включаються в процес навчання.

Результати застосування різних методів оптимізації пам'яті для архітектури ResNet-50, адаптованої для задач сканування документів, наведені в Таблиці 1.

Таблиця 1

Порівняння методів оптимізації пам'яті для архітектури ResNet-50

Метод	Розмір моделі, МБ	Зменшення розміру, рази	Зниження точності, %	Швидкість виведення, мс
Оригінальна модель	97,8	1,0	0,0	185
8-біт PTQ	24,5	4,0	0,5	125
8-біт QAT	24,5	4,0	0,3	126
Обрізання (90%)	9,8	10,0	0,7	172
Обрізання + 8-біт QAT	2,4	40,0	0,9	104
Deep Compression	2,0	49,0	1,0	108

Для архітектури MobileNet, яка спочатку розроблялась для мобільних пристроїв і часто використовується в компактних сканерах, результати оптимізації наведені в Таблиці 2.

Таблиця 2

Порівняння методів оптимізації пам'яті для архітектури MobileNet

Метод	Розмір моделі, МБ	Зменшення розміру, рази	Зниження точності, %	Швидкість виведення, мс
Оригінальна модель	16,9	1,0	0,0	45
8-біт PTQ	4,2	4,0	0,8	32
8-біт QAT	4,2	4,0	0,4	33
Обрізання (80%)	3,4	5,0	0,9	43
Обрізання + 8-біт QAT	0,9	18,8	1,2	28
Deep Compression	0,7	24,1	1,4	30

Для більш детального аналізу впливу різних методів оптимізації на якість розпізнавання було проведено тестування на наборі даних, що включає сканування різних типів документів: текстових сторінок, штрих-кодів, QR-кодів та форм. Результати тестування для оптимізованої моделі ResNet-50 (обрізання + QAT) представлені в Таблиці 3.

Таблиця 3

Вплив оптимізації на якість розпізнавання різних типів документів

Тип документа	Оригінальна модель	Оптимізована модель	Зниження якості, %
Текстові сторінки	98,5%	97,9%	0,6
Штрих-коди	99,7%	99,3%	0,4
QR-коди	99,8%	99,2%	0,6
Форми	96,3%	94,9%	1,4

Як видно з таблиці, вплив оптимізації на якість розпізнавання різних типів документів нерівномірний. Найменше зниження спостерігається для штрих-кодів і QR-кодів, що можна пояснити їх відносною простотою та наявністю вбудованих механізмів корекції помилок. Для форм, які мають складнішу структуру та потребують розпізнавання різних елементів, вплив оптимізації більш помітний.

Важливим питанням є також вплив методів оптимізації на енергоспоживання мобільних пристроїв. Для оцінки цього впливу було проведено тестування на стандартному смартфоні з процесором середнього рівня. Результати показують, що оптимізована модель ResNet-50 з обрізанням та QAT споживає на 56% менше енергії в порівнянні з оригінальною моделлю при виконанні однакової кількості операцій сканування.

Це пов'язано не лише зі зменшенням кількості обчислень, але й з більш ефективним використанням кешу процесора через менший розмір моделі. Такі результати мають важливе практичне значення, оскільки збільшують час автономної роботи пристроїв для сканування, що критично для багатьох комерційних та виробничих застосувань.

Для демонстрації практичної значущості досліджених методів оптимізації було розроблено мобільний додаток для сканування та розпізнавання документів з використанням оптимізованої моделі

ResNet-50. Додаток здатний розпізнавати текст, штрих-коди, QR-коди та заповнені форми безпосередньо на пристрої без необхідності передачі даних на сервер.

Тестування додатку на різних мобільних пристроях показало, що завдяки оптимізації пам'яті додаток здатний працювати навіть на бюджетних смартфонах з обмеженими ресурсами. Швидкість сканування та розпізнавання становить близько 0,2 секунди на документ формату А4 при роздільній здатності камери 12 МП, що є цілком прийнятним для більшості практичних застосувань.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Проведене дослідження показує, що методи оптимізації пам'яті дозволяють суттєво зменшити вимоги до ресурсів при використанні нейронних мереж для задач сканування та розпізнавання. Комбінований підхід, який включає обрізання, квантизацію та кодування Гаффмана, забезпечує зменшення розміру моделі до 40-49 разів при незначному зниженні точності.

Квантизація з урахуванням навчання (QAT) демонструє кращі результати в порівнянні з квантизацією після навчання (PTQ) для задач сканування, особливо при розпізнаванні дрібних деталей на зображеннях. Експериментально підтверджено, що методи оптимізації позитивно впливають на швидкість виведення та енергоспоживання, що особливо важливо для мобільних та вбудованих систем.

Результати дослідження відкривають можливості для широкого впровадження систем сканування та розпізнавання, що базуються на нейронних мережах, у мобільних та вбудованих пристроях з обмеженими ресурсами. Подальші дослідження можуть бути спрямовані на розробку адаптивних методів оптимізації, які автоматично підбирають оптимальні параметри квантизації та обрізання залежно від конкретного типу сканованих документів та доступних ресурсів пристрою.

Література

1. Han, S. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding / S. Han, H. Mao, W. J. Dally // arXiv preprint. — 2015. — No. arXiv:1510.00149. — 14 p. — DOI: 10.48550/arXiv.1510.00149.

2. Jacob, B. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference / B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2018. — Pp. 2704–2713. — DOI: 10.48550/arXiv.1712.05877.

3. Cheng, Y. A Survey of Model Compression and Acceleration for Deep Neural Networks / Y. Cheng, D. Wang, P. Zhou, T. Zhang // arXiv preprint. — 2018. — No. arXiv:1710.09282. — 23 p. — DOI: 10.48550/arXiv.1710.09282.

4. Sze, V. Efficient Processing of Deep Neural Networks: A Tutorial and Survey / V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer // Proceedings of the IEEE. — 2017. — Vol. 105, Iss. 12. — Pp. 2295–2329. — DOI: 10.1109/JPROC.2017.2720198.

5. Krishnamoorthi, R. Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper // arXiv preprint. — 2018. — No. arXiv:1806.08342. — 37 p. — DOI: 10.48550/arXiv.1806.08342.

6. Blalock, D. What is the State of Neural Network Pruning? / D. Blalock, J. J. G. Ortiz, J. Frankle, J. Guttag // Proceedings of Machine Learning and Systems (MLSys). — 2020. — Pp. 129–146. — ISBN: 9781713820048.

7. Gholami, A. A Survey of Quantization Methods for Efficient Neural Network Inference / A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer // arXiv preprint. — 2021. — No. arXiv:2103.13630. — 27 p. — DOI: 10.48550/arXiv.2103.13630.

8. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. — Cambridge : MIT Press, 2016. — 800 p. — ISBN: 9780262035613.

References

1. Han, S. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding / S. Han, H. Mao, W. J. Dally // arXiv preprint. — 2015. — No. arXiv:1510.00149. — 14 p. — DOI: 10.48550/arXiv.1510.00149.

2. Jacob, B. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference / B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2018. — Pp. 2704–2713. — DOI: 10.48550/arXiv.1712.05877.

3. Cheng, Y. A Survey of Model Compression and Acceleration for Deep Neural Networks / Y. Cheng, D. Wang, P. Zhou, T. Zhang // arXiv preprint. — 2018. — No. arXiv:1710.09282. — 23 p. — DOI: 10.48550/arXiv.1710.09282.

4. Sze, V. Efficient Processing of Deep Neural Networks: A Tutorial and Survey / V. Sze, Y.-H. Chen, T.-J. Yang, J. S. Emer // Proceedings of the IEEE. — 2017. — Vol. 105, Iss. 12. — Pp. 2295–2329. — DOI: 10.1109/JPROC.2017.2720198.

5. Krishnamoorthi, R. Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper // arXiv preprint. — 2018. — No. arXiv:1806.08342. — 37 p. — DOI: 10.48550/arXiv.1806.08342.

6. Blalock, D. What is the State of Neural Network Pruning? / D. Blalock, J. J. G. Ortiz, J. Frankle, J. Guttag // Proceedings of Machine Learning and Systems (MLSys). — 2020. — Pp. 129–146. — ISBN: 9781713820048.

7. Gholami, A. A Survey of Quantization Methods for Efficient Neural Network Inference / A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer // arXiv preprint. — 2021. — No. arXiv:2103.13630. — 27 p. — DOI: 10.48550/arXiv.2103.13630.

8. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. — Cambridge : MIT Press, 2016. — 800 p. — ISBN: 9780262035613.