УДК 004.9

Гліб Середюк

Володимир Гармаш

МЕТОДИ КВАНТИЗАЦІЇ ТА КОМПРЕСІЇ ДЛЯ ОПТИМІЗАЦІЇ ПАМ'ЯТІ В НЕЙРОННИХ МЕРЕЖАХ ДЛЯ ЗАДАЧ КОМП'ЮТЕРНОГО ЗОРУ

Вінницький національний технічний університет

Анотація

У роботі досліджуються методи оптимізації пам'яті при використанні глибоких нейронних мереж для задач комп'ютерного зору. Розглянуто підходи до квантизації, обрізання (pruning) та компресії моделей, що дозволяють суттєво зменшити вимоги до пам'яті без значної втрати точності. Проаналізовано ефективність цих методів при застосуванні до задач розпізнавання та класифікації зображень. Встановлено, що комбінований підхід, який включає обрізання, квантизацію та кодування Гаффмана, здатен зменшити розмір моделі до 35-49 разів при зниженні точності менше ніж на 1%. Представлено порівняльний аналіз алгоритмів квантизації після навчання (PTQ) та квантизації з урахуванням навчання (QAT) для найпоширеніших архітектур нейронних мереж.

Ключові слова: глибокі нейронні мережі, квантизація, обрізання моделей, компресія, комп'ютерний зір, оптимізація пам'яті.

Abstract

This paper investigates memory optimization methods for deep neural networks in computer vision tasks. The approaches to quantization, pruning, and model compression are considered, which significantly reduce memory requirements without substantial accuracy loss. The effectiveness of these methods is analyzed when applied to image recognition and classification tasks. It has been established that a combined approach that includes pruning, quantization, and Huffman coding can reduce model size by 35-49 times with accuracy degradation of less than 1%. A comparative analysis of Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) algorithms for the most common neural network architectures is presented.

Keywords: deep neural networks, quantization, model pruning, compression, computer vision, memory optimization.

Вступ

Сучасні задачі комп'ютерного зору, такі як розпізнавання об'єктів, сегментація зображень та виявлення облич, вимагають використання складних глибоких нейронних мереж. Ці моделі демонструють високу точність, але потребують значних обчислювальних ресурсів і пам'яті для зберігання мільйонів параметрів. Архітектури типу VGG-16, ResNet-50 чи YOLO можуть займати сотні мегабайтів пам'яті, що суттєво ускладнює їх застосування на пристроях з обмеженими ресурсами, таких як мобільні телефони, вбудовані системи або IoT-пристрої.

Оптимізація використання пам'яті є критичною для розширення можливостей застосування нейронних мереж у реальних сценаріях. Актуальність даної проблеми зростає з тенденцією до розгортання моделей штучного інтелекту на кінцевих пристроях замість хмарних серверів, що забезпечує приватність даних та зменшує залежність від мережевого підключення.

Метою даного дослідження є аналіз методів оптимізації пам'яті при використанні нейронних мереж для задач комп'ютерного зору та оцінка їх ефективності з точки зору компромісу між зменшенням розміру моделі та збереженням точності.

Результати дослідження

Результати експериментальних досліджень показують, що найбільш ефективними підходами до оптимізації пам'яті є обрізання мереж, квантизація параметрів та їх кодування. При обрізанні мережі видаляються зв'язки з найменшими за абсолютним значенням вагами, що перетворює щільну матрицю ваг на розріджену. Наприклад, для архітектури VGG-16 обрізання дозволяє зменшити кількість параметрів з 138 млн до 11,3 млн.

Квантизація параметрів передбачає зниження точності представлення ваг з 32-бітних чисел з плаваючою комою до 8-бітних цілих чисел, що зменшує розмір моделі в 4 рази. Розрізняють квантизацію після навчання (PTQ) та квантизацію з урахуванням навчання (QAT), де остання демонструє кращі результати збереження точності.

Кодування Гаффмана, яке присвоює коротші коди частіше вживаним значенням ваг, забезпечує додаткове зменшення розміру моделі на 20-30%. Комбінований підхід "Deep Compression", що послідовно застосовує всі три методи, дозволяє досягти вражаючих результатів. Для архітектури AlexNet це зменшує вимоги до пам'яті з 240 МБ до 6,9 МБ (стиснення в 35 разів) при зниженні точності на 0,58%.

Таблиця 1 - Порівняння методів оптимізації пам'яті для архітектури ResNet-50

Метод	**Розмір моделі, МБ**	**Зменшення розміру, рази**	**Зниження точності, %**
Оригінальна модель	97,8	1,0	0,0
8-біт PTQ	24,5	4,0	0,5
8-біт QAT	24,5	4,0	0,3
Обрізання (90%)	9,8	10,0	0,7
Обрізання + 8-біт QAT	2,4	40,0	0,9

Висновки

Проведене дослідження показує, що методи квантизації та компресії дозволяють суттєво зменшити вимоги до пам'яті при використанні нейронних мереж для задач комп'ютерного зору. Експериментальні результати підтверджують, що комбінований підхід, який включає обрізання, квантизацію та кодування Гаффмана, забезпечує стиснення моделей в 35-49 разів при зниженні точності менше ніж на 1%. Найкращим компромісом між розміром моделі та збереженням точності є поєднання обрізання мережі з квантизацією з урахуванням навчання, що дозволяє зменшити вимоги до пам'яті у 40 разів для архітектури ResNet-50. Отримані результати відкривають можливості для розгортання складних нейромережевих архітектур на пристроях з обмеженими ресурсами, таких як мобільні телефони, камери спостереження та IoT-пристрої.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

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

Середюк Гліб Анатолійович — аспірант, Вінницький національний технічний університет, м. Вінниця, e-mail: hlibsered@gmail.com

Гармаш Володимир Володимирович — кандидат технічних наук, доцент, Вінницький національний технічний університет, м. Вінниця

Seredyuk Hlib — PhD Student, Vinnytsia National Technical University, Vinnytsia, e-mail: hlibsered@gmail.com

Garmash Volodymyr — Candidate of Technical Sciences, Associate Professor, Vinnytsia National Technical University, Vinnytsia