

УДК 004.85

Діденко А.Є.¹, Субботін С.О.²

¹ асп. НУ «Запорізька політехніка»

² д-р техн. наук, проф. НУ «Запорізька політехніка»

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ПРИСКОРЕННЯ НАВЧАННЯ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Прискорення навчання нейронних мереж має важливе значення у сфері глибинного навчання. По-перше, це дозволяє дослідникам проводити більше експериментів за певну одиницю часу. По-друге, скорочення часу навчання зменшує фінансові витрати проєкту, такі як оренда необхідного обладнання. Це особливо важливо для дослідників і організацій з обмеженим бюджетом. Крім того, прискорення навчання дозволяє впроваджувати нейронні мережі у сферах, де швидкодія має вирішальне значення.

Зміна параметрів середовища навчання може з однієї сторони прискорити навчання нейронної мережі, а з іншої – погіршити її точність, що в свою чергу спонукає до знаходження компромісу між цими двома величинами. В даній роботі розглядаються методи прискорення навчання нейронних мереж і те, як вони впливають на час навчання та точність моделі.

Для тестування обрано модель загорткової нейронної мережі, що виконує задачу класифікації зображень. Нейронна мережа складається з двох загорткових шарів, одного агрегувального шару та трьох повнозв'язних шарів. У якості функції активації обрано ReLU. В якості фреймворку для навчання нейронної мережі обрано PyTorch, оскільки він має велику кількість параметрів для тонкого налаштування процесу навчання. У якості навчальної і тестової вибірок було обрано датасет CIFAR-10 [1], що містить 60 тис. кольорових зображень у 10 класах.

Популярним методом прискорення навчання є зміна розміру вхідного батчу. Зазвичай, використання більшого розміру батчу дозволяє прискорити навчання, оскільки модель проходить через меншу кількість ітерацій. З іншої сторони, збільшення розміру батчу часто призводить до погіршення точності моделі, оскільки через зменшення кількості кроків циклу оптимізаційного алгоритму знижується здатність моделі до узагальнення.

Іншим методом прискорення навчання є використання змішаної точності. На сьогоднішній день, більшість моделей використовують тип даних float32, що займає 32 біти пам'яті. Але існують такі типи даних, як float16 і bfloat16, які займають 16 бітів пам'яті. Сучасні процесори прискорених обчислень швидше виконують операції з 16-бітними типами даних. Однак, використання таких типів даних також може призвести до зниження точності результатів.

Злиття шарів [2] також може бути використане для прискорення навчання моделі. Це техніка стиснення нейронної мережі, яка дозволяє поєднати певні шари в один, наприклад згортку і функцію активації.

У табл. 1 наведено результати навчання у залежності від розміру вхідного батчу. У табл. 2 наведено порівняння швидкості і точності моделі різних методів прискорення (розмір батчу – 128). Кожна модель навчалась 10 епох, метод оптимізації і функція втрат однакові для кожної моделі (AdamW і крос-ентропія відповідно).

Таблиця 1 – Порівняння швидкості навчання і точності залежно від розміру батчу

Розмір батчу	Час навчання, с	Точність, %
32	105,2	63,8
64	94,7	62,7
128	88	62,6
256	86,1	60,4

Таблиця 2 – Порівняння методів прискорення навчання

Метод	Час навчання, с	Точність, %
Оригінальна модель	88	62,6
Float16	87,1	61,3
Bfloat16	86,1	61,7
Злиття шарів	86,92	61,8

Таким чином, ефективно прискорити навчання згорткової нейронної мережі можна декількома способами, але необхідно брати до уваги компроміс між швидкістю навчання і точністю результатів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. The CIFAR-10 dataset [Electronic recourse]. – Access mode: <https://www.cs.toronto.edu/~kriz/cifar.html>;
2. Compressing Deep Neural Networks via Layer Fusion [Electronic recourse] / J.O. Neill, G.V. Steeg, A. Galstyan. – 2020. – Access mode: <https://arxiv.org/abs/2007.14917>.