

Параллельный алгоритм последовательного обучения нейронных сетей*

И.В. Шарун, А.В. Зыкина

Омский государственный технический университет

В настоящее время подходы к ускорению обучения глубоких нейронных сетей строятся на основе параллелизма данных, последовательного обучения [1], либо на основе вертикального (конвейерного) параллелизма [2]. Данные подходы не лишены недостатков: либо происходит простой оборудования, либо возникает сложная система коммуникации между объектами оптимизации.

В рамках данного исследования рассматривается подход к ускорению обучения глубоких нейронных сетей за счет применения параллельного алгоритма последовательного обучения. Суть данного метода заключается в декомпозиции общей задачи обучения сети на множество отдельных оптимизационных задач для каждого слоя. Каждый слой рассматривается как самостоятельный объект оптимизации, что позволяет организовать его обучение независимо от других слоев. После завершения шага оптимизации каждого слоя происходит агрегация полученных параметров, формируется общий набор весов, который используется в качестве начальной точки для следующей итерации.

Такая организация процесса обучения обеспечивает возможность использования параллельных и распределенных вычислений для одновременной работы над отдельными слоями или группами слоев, что значительно сокращает время обучения. Подход позволяет сократить время обучения, особенно для сетей с большим количеством слоев и параметров.

Для проверки эффективности подхода были проведены эксперименты с различными архитектурами нейронных сетей прямого распространения, включая сверточные нейронные сети. В ходе экспериментов использовались известные наборы данных, в том числе MNIST, CIFAR-10 и ImageNet, произведена оценка применимости метода для широкого спектра задач, связанных с классификацией изображений.

Результаты экспериментов подтвердили гипотезу о том, что параллельный алгоритм последовательного обучения способен не только значительно сократить время обучения, но и сохранить уровень точности модели, сравнимый с результатами, полученными при использовании стандартных методов. При этом наблюдается связь между увеличением размера архитектуры сети и масштабируемостью предложенного подхода: чем больше слоев и параметров содержит сеть, тем более выраженным становится эффект ускорения.

Дальнейшие исследования направлены на анализ применимости метода для других видов нейронных сетей, в т.ч. рекуррентных, улучшение методов агрегации параметров для достижения еще большего ускорения без потери качества модели.

Литература

1. Karimi A., Kalhor A., Sadeghi Tabrizi M. Forward layer-wise learning of convolutional neural networks through separation index maximizing // Scientific Reports. 2024. Vol. 14. Article 8576. DOI: 10.1038/s41598-024-59176-3.
2. Núñez G., Romero-Sandí H., Rojas E., Meneses E. A Study of Pipeline Parallelism in Deep Neural Networks // Revista Colombiana de Computación. 2024. Vol. 25, no. 1. P. 48–59. DOI: 10.29375/25392115.5056.

*Работа выполнена в рамках Государственного задания № FSGF-2024-0006 «Разработка новых моделей и методов принятия решений в многоуровневых системах управления».