Применение тензорного разложения для сжатия свёрточных нейронных сетей

Гарипов Тимур Исмагилевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: timgaripov@gmail.com

Алгоритмы машинного обучения, основанные на свёрточных нейронных сетях, в настоящее время демонстрируют передовые результаты в задачах компьютерного зрения, обработки естественных языков и в других областях. Но для применения таких алгоритмов требуются значительные вычислительные ресурсы. Современные архитектуры нейронных сетей включают миллионы настраиваемых параметров, для хранения которых требуются сотни мегабайт. Это обстоятельство не позволяет оптимально использовать энергоэффективные запоминающие устройства при работе со свёрточными нейронными сетями, а следовательно, ограничивает их применение на мобильных устройствах.

В работе [1] был предложен метод сжатия матрицы весов полносвязных слоёв нейронной сети, использующий разложение тензорного поезда [2] (Tensor-Train, TT). Было показано, что описанный метод позволяет значительно уменьшить число параметров в полносвязных слоях без значительной потери в качестве работы алгоритма. При применении указанного метода снижается доля параметров полносвязных слоёв и основная часть потребляемой памяти расходуется на хранение параметров свёрточных слоёв. Помимо того, в последнее время появляется все больше успешных архитектур свёрточных сетей, в которых число параметров полносвязных слоёв незначительно по сравнению с числом параметров свёрточных слоёв.

В данной работе предложен метод сжатия свёрточных слоёв нейронных сетей, основанный на специальном способе представления параметров свёрточного слоя в виде многомерного тензора и уменьшения числа параметров, необходимого для хранения полученного тензора с помощью TT-разложения. Ниже перечисленны основные результаты работы.

- Экспериментально показано, что наивное применение TT-разложения к 4-х мерному массиву параметров свёрточного слоя приводит к значительному снижению точности работы нейронной сети.

- Предложен специальный способ представления параметров свёрточного слоя в виде многомерного тензора, направленный на возможность эффективного применения TT-разложения.

- Проведены эксперименты, показавшие возможность успешного применения предложенного подхода. В задаче классификации изображений на наборе данных CIFAR-10 удалось достичь четырёхкратного уменьшения числа параметров нейронной сети, все параметры которой относятся к свёрточным слоям, при уменьшении качества классификации на 2%.

- Показано, что предложенный подход допускает успешное сочетание с предложенным ранее методом сжатия полносвязных слоёв. Совместным сжатием свёрточных и полносвязных слоёв удалось снизить число параметров сети, в которой присутствуют как полносвязные, так и свёрточные слои, в 82 раза при уменьшении качества классификации на 1%.

Литература

1. Novikov A., Podoprikhin D., Osokin A., and Vetrov D. Tensorizing neural networks. // In Advances in Neural Information Processing Systems 28 (NIPS), Montreal, Canada, 2015, P. 442–450.
2. Oseledets I. Tensor-train decomposition // SIAM Journal of Scientific Computing. 2011. Vol. 33, No. 5. P. 2295–2317.