УДК 004.8

Сжатие сверточных нейронных сетей с использованием матричного разложения весов и прунинга фильтров.

Литвинцева А. В. (Университет ИТМО)
Научный руководитель – к. т. н., Никитин Н. О.
(Университет ИТМО)

Введение. Современные глубокие нейронные сети требуют высокого потребления памяти и больших вычислительных нагрузок. Развертывание крупных сверточных нейронных сетей на мобильных телефонах или микрокомпьютерах часто является недопустимым с точки зрения времени и занимаемого пространства. Алгоритмы оптимизации структуры нейронных сетей могут уменьшить объем занимаемой памяти и время вывода в условиях ограниченных ресурсов. Целью оптимизации структуры сети является отбрасывание избыточных весов чрезмерно параметризованной сети и создание сжатой модели, производительность которой сравнима с производительностью исходной сети. Обрезку сети также можно использовать для уменьшения нагрузки, связанной с ручным проектированием небольшой сети, путем автоматического определения эффективных архитектур из более крупных сетей.

Методы. В данной работе исследуются два подхода к оптимизации структуры сверточных нейронных сетей: мягкий прунинг фильтров и сингулярное разложение матрицы весов с последующей подрезкой спектра. Реализованные алгоритмы применяются к архитектуре ResNet18, используемой для классификации изображений, и тестируются на двух наборах данных. Оба алгоритма применяются во время обучения нейронной сети, что позволяет получить подготовленную к сжатию нейронную сеть. Результаты сравниваются по нескольким параметрам: степень сжатия модели по отношению к потерям в целевой метрике, применимость метода к предобученным моделям, гибкость настройки алгоритма.

Выводы. В ходе выполнения работы был разработан программный модуль оптимизации структуры сверточных нейронных сетей во время их обучения. Предложенные алгоритмы позволяют эффективно сжимать модели при незначительном снижении точности классификации. На одном из рассмотренных наборов данных алгоритм позволил сжать модель на 80% при потере в точности 1%. Предложенные методы сжатия моделей применимы для развертывания систем компьютерного зрения на микрокомпьютерах, что позволяет использовать их в системах промышленного мониторинга, например, для обнаружения дефектов на производстве.

Список использованных источников:

1. Neill J. O. An overview of neural network compression //arXiv preprint arXiv:2006.03669. – 2020.
2. Yang H. et al. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. – 2020. – С. 678-679.
3. He Y. et al. Soft filter pruning for accelerating deep convolutional neural networks //arXiv preprint arXiv:1808.06866. – 2018.

Литвинцева А. В. (автор) Подпись

Никитин Н. О. (научный руководитель) Подпись