Адаптивная схема инициализации для повышения качества и стабильности методов квантизации

Жариков Илья Николаевич[1]
Кузькина Юлия Сергеевна[2]

1: *Аспирант, Физтех-школа прикладной математики и информатики МФТИ, Москва, Россия*
2: *Исследователь, Лаборатория машинного интеллекта МФТИ, Москва, Россия*
E-mail: ilya.zharikov@phystech.edu, juliakuzkina1@gmail.com

Глубокие нейронные сети показывают высокое качество в различных областях, таких как, классификация изображений, понимание текста, распознавание речи. Однако использование глубоких нейронных сетей требует выполнения множества арифметических операций и большое количество постоянной памяти для хранения параметров сети. Таким образом, использовать нейронные сети на устройствах с небольшой вычислительной мощностью не всегда представляется возможным. В настоящее время существует множество методов обеспечения эффективной работы глубоких сетей на устройствах с ограниченными ресурсами. К таким методам относятся поиск эффективной архитектуры сети (NAS), дистилляция знаний (англ. knowledge distillation), прунинг (англ. pruning) и квантизация (англ. quantization).

Данная работа посвящена методам квантизации, то есть способам перевода весов и активаций нейронной сети в представление с меньшей точностью. Это позволяет ускорить модель и уменьшить её физический размер. Основным параметром этого метода является шаг квантизации. С помощью данного параметра контролируется диапазон значений, внутри которого будет выполняться квантизация. Последние работы по квантизации [1–3] показывают результаты, сравнимые с оригинальными моделями, которые проводят вычисления в высокой точности. Для достижения этих результатов процесс квантизации применяется совместно с обучением модели. Во время обучения модели есть возможность сделать шаг квантизации обучаемым, определив производную по данному параметру [1]. Для уменьшения ошибки квантизации значений весов модели из распределения с отличным от нуля средним значением вводится дополнительный параметр, который может контролировать это среднее значение [2].

Аналогичный обучаемый метод [3] предлагает использовать неравномерную структуру уровней квантизации, которая адаптируется к колоколообразному распределению весов модели и более точно учитывает веса, находящихся вокруг среднего значения этих весов. Также для определения шага квантизации могут использоваться численные методы совместно с оценкой параметров распределения весов и активаций модели [4].

В данной работе решается проблема инициализации параметров квантизации (шаг квантизации). Параметры квантизации чувствительны к инициализации и обновлениям, так как контролируют диапазон значений внутри которого будет выполняться квантизация, поэтому не могут быть инициализированы случайным образом. Общепринятым подходом к инициализации данных параметров является вычисление различных статистик весов и активаций модели, что не всегда является оптимальным в смысле минимизации ошибки квантизации способом. Во многих работах предлагаются свои подходы к инициализации [1–4], но они не являются универсальными или стабильными при обучении. В данной работе предлагается универсальный способ инициализации параметров квантизации, с помощью которого можно получать стабильно высокие результаты квантизации при использовании обучаемых методов, ускорить их сходимость и повысить стабильность совместного с квантизацией процесса обучения.

Литература

1. Steven K. et al. Learned step size quantization // International Conference on Learning Representations (ICLR), 2020.
2. Bhalgat Y. et al. LSQ+: Improving low-bit quantization through learnable offsets and better initialization // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, P. 696–697.
3. Li Y., Dong X., Wang W. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks // arXiv preprint arXiv:1909.13144, 2019.
4. Banner R., Nahshan Y., Soudry D. Post training 4-bit quantization of convolutional networks for rapid-deployment // Advances in Neural Information Processing Systems. 2019. Vol. 32.