0.1. *Чудаков Д.С.* Итеративная адаптация к шуму квантования в нейронных сетях

В настоящий момент нейронные сети повсеместно применяются в современных продуктовых решениях, начиная от простых классификаторов изображений заканчивая комплексными системами включающими в себя классификацию, детекцию и сегментацию изображений.

Нейронные сети крайне требовательны к вычислительным ресурсам не только на этапе обучения, но и на этапе исполнения (Inference). В современных решениях необходимо использовать новые архитектуры, новые подходы к обучению и применять последующие оптимизации [1]. Это позволяет использовать в реальных продуктах точные и лёгкие сети, для которых нет необходимости в дорогостоящем оборудовании и даже запускать их на мобильных процессорах. Один из подходов к оптимизации скорости и размера сети после обучения является квантование (quantization) [2].

Существенная часть методов квантования нейронных сетей после обучения нацелены на уменьшение шума от квантования благодаря подбору оптимальных порогов. Данные методы страдают от большой просадки точности итоговой сети, особенно это заметно при использовании малого числа бит (6 и 4 бита). В первую очередь это связано с тем что нет возможности достаточно уменьшить шум после дискретизации в случаях с низким числом бит.

Был предложен метод итеративной адаптации к шуму от квантования [3], который последовательно уменьшает шум от дискретизации, а затем обучает последующие слои сети работать с возникшим шумом минимизируя среднеквадратичное отклонение между квантованной и исходной сетью. Данный метод позволил значительно улучшить точности при квантовании нейронных сетей в малое число бит, по сравнению с методами квантования после обучения, при этом метод затрачивает меньше времени чем полное обучение квантованной сети с нуля (Метод обучения с учётом квантования).

Результаты работы алгоритма были проверены на нескольких самых популярных архитектурах нейронных сетей, был измерен шум от квантования в результате работы нескольких алгоритмов квантования и сделаны выводы о важности шума от квантования.

Список литературы

[1] Tan, M., Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks
In International Conference on Machine Learning. 2019. (pp. 6105-6114). PMLR.

[2] Jacob, B., Kligys, S., Chen, B., «et al.» Quantization and training of neural networks for efficient integer-arithmetic-only inference // In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. (pp. 2704-2713).

[3] Chudakov, D., Alyamkin, S., Goncharenko, A., Denisov, A. Iterative Adaptation to Quantization Noise
In International Work-Conference on Artificial Neural Networks. 2021. (pp. 303-310). Springer, Cham.