Секция «19.8 Дискретная математика и математическая кибернетика»

# Дифференцируемый поиск схемы квантизации для нейросетевых моделей

**Научный руководитель – Половников Владимир Сергеевич**

***Давыдова Дарья Николаевна***
*Аспирант*
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра математической теории
интеллектуальных систем, Москва, Россия
*E-mail: d.dawydowa2017@yandex.ru*

Поиск архитектуры нейронной сети - это метод, позволяющий автоматически подобрать наилучшую архитектуру нейронной сети для данной задачи.

Одной из разновидностей этого метода является DARTS [3] - дифференцируемый поиск архитектуры нейронной сети. DARTS позволяет перейти от дискретного пространства поиска к непрерывному и использовать метод градиентного спуска для подбора параметров, отвечающих за архитектуру сети.

В исследовании предлагается использовать метод DARTS для подбора оптимальной схемы квантизации генеративной языковой модели. В качестве методов-кандидатов рассматриваются LSQ+ [1], PACT [2], равномерная симметричная квантизация, подбираемые разрядности - 4, 8, 16 и 32 бит.

Для каждого линейного слоя $\ell$ задаётся набор квантизаторов-кандидатов для весов $\{Q_i^w\}_{i=1}^{N_w}$ и активаций $\{Q_j^a\}_{j=1}^{N_a}$ с битностями $b_i^w$ и $b_j^a$ соответственно. Подбор квантизатора и разрядности для весов и активаций осуществляется с помощью параметров $\boldsymbol{\alpha}^w, \boldsymbol{\alpha}^a$, обучаемых градиентным спуском.

Активации и веса квантизуются выбранными квантизаторами с заданными битностями по формулам:
$\tilde{x} = \sum_{j=1}^{N_a} p_j^a \cdot Q_j^a(x)$, $\tilde{W} = \sum_{i=1}^{N_w} p_i^w \cdot Q_i^w(W)$,
где $p_i$ - вероятность выбора i-го квантизатора.

Таким образом, полная смешанная операция на линейных слоях модели задается следующей формулой:

$$Y = \left( \sum_{i=1}^{N_a} p_i^a \, Q_i^a(X) \right) \cdot \left( \sum_{j=1}^{N_w} p_j^w \, Q_j^w(W) \right)^{\mathsf{T}} + b$$

Как и в классическом DARTS, оптимизация параметров модели и параметров квантизаторов проводится на двух уровнях. На внутреннем уровне оптимизируются параметры сети с помощью кросс-энтропии, на внешнем - параметры квантизаторов с помощью функции $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{cost} \cdot \mathcal{L}_{cost} + \lambda_{err} \cdot \mathcal{L}_{\text{err}},$$

где $\mathcal{L}_{CE}$ - кросс-энтропия, $\mathcal{L}_{cost}$ - штраф за вычислительную стоимость модели, $\mathcal{L}_{\text{err}}$ - штраф за ошибку квантизации, а $\lambda_{cost}, \lambda_{err}$ - обучаемые коэффициенты.

Эксперименты показали эффективность градиентных методов для автоматического подбора схемы квантизации: достигнута высокая степень сжатия модели и большая точность, чем при подборе схемы квантизации вручную.

## Источники и литература

1) 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. – С. 696-697.
2) Choi J. et al. Pact: Parameterized clipping activation for quantized neural networks //arXiv preprint arXiv:1805.06085. – 2018.
3) Liu H., Simonyan K., Yang Y. Darts: Differentiable architecture search //arXiv preprint arXiv:1806.09055. – 2018.