АНАЛИЗ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В НЕЙРОННЫХ СЕТЯХ В КОНТЕКСТЕ ЗАДАЧ ОПТИМИЗАЦИИ НЕЙРОСЕТЕВЫХ ВЫЧИСЛЕНИЙ

А.П. Попов

ИПМ им. М.В. Келдыша РАН, г. Москва

alekspopovr68@mail.ru

Разработка методов решения задач машинного зрения – на сегодняшний день одна из ключевых задач многих исследовательских лабораторий. Алгоритмы компьютерного зрения используются в автономных робототехнических системах, в системах видеонаблюдения и в других прикладных систем. На сегодняшний день применение современных нейросетевых решений сопряжено с использованием больших вычислительных ресурсов. Улучшение нейронных сетей зачастую происходит за счет их масштабирования: увеличение сложности вычислительных операций, увеличение числа вычисляемых параметров и так далее. В контексте такого развития нейронных сетей следует учитывать необходимость их использования на конечных устройствах, которые ограничены в своей производительности и развиваются медленнее, чем нейросетевые модели.

В рамках данной работы проведен обзор методов оптимизации нейронных сетей и анализ работы методов оптимизации с точки зрения работы со знаниями и представлениями знаний внутри обученной нейросетевой архитектуры. Задача оптимизации вычислений внутри нейронной сети ставится следующим образом, требуется получить $A^* = F(A)$, где F – функция оптимизации нейронной сети, а A оптимизируемая нейронная сеть и A^* нейронная сеть, полученная в результате оптимизации, при условии, что

$$\big(E(A,D) - E(A^*,D)\big) \geq Eps_1,\ \big(S(A^*,D) - S(A,D)\big) \geq Eps_2,$$
$$\big(M(A,D) - M(A^*,D)\big) \geq Eps_3,$$

где E, S, M – функционалы оценки ошибки, скорости, потребления памяти нейронной сети на входных данных D соответственно, а Eps_2, Eps_3 заранее заданные ограничения оптимизации, а Eps_1 качественная оценка оптимизации. В работе рассмотрены методы квантизации трансформерных и сверточных нейронных сетей [1, 2], а также больших языковых моделей, в которых точное представление знаний является основной задачей [3]. Также проведен анализ методов прунинга [4, 5] нейронных сетей и его влияния на представление знаний и избыточность нейросетевой модели. Кроме того, особо внимание уделено методам дистилляции знаний [6, 7], как методам оптимизации нейронных сетей, которые осуществляют прямое или непрямое сжатие знаний внутри нейронной сети. В докладе приводятся примеры обучения небольшой

сети ученика при помощи сети учителя за счет передачи внутренних сигналов учителя. Такой пример позволяет понять, как внутренние сигналы передают информацию о целевой решаемой задаче.

На основе анализа работ, посвященных задаче оптимизации нейронных сетей и внутреннего представления знаний, делается вывод о перспективности применения рассмотренных методов оптимизации нейронных сетей. Такой вывод позволяет заключить, что применение больших нейросетевых моделей для решения задач машинного зрения на этапе исследовательской работы вполне обоснованно, так как после получения удовлетворительных результатов экспериментов, можно провести оптимизацию и итоговую интеграцию нейросетевого решения.

Список литературы:

1. Li Y. [et al.]. Q-vit: Accurate and fully quantized low-bit vision transformer //Advances in neural information processing systems. 2022. Vol. 35. Pp. 34451–34463. https://doi.org/10.48550/arXiv.2210.06707
2. Xu K. [et al.]. Eq-net: Elastic quantization neural networks // Proceedings of the IEEE/CVF international conference on computer vision. 2023. Pp. 1505–1514.
3. Wang C. [et al.]. Q-VLM: Post-training Quantization for Large Vision-Language Models // arXiv:2410.08119. 2024.
4. Bai S. [et al.]. Unified data-free compression: Pruning and quantization without fine-tuning // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. Pp. 5876–5885.
5. Frantar E., Alistarh D. Optimal brain compression: A framework for accurate post-training quantization and pruning // Advances in Neural Information Processing Systems. 2022. Vol. 35. Pp. 4475–4488.
6. Wang L., Yoon K.J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks // IEEE transactions on pattern analysis and machine intelligence. 2021. Vol. 44. Issue 6. Pp. 3048–3068. https://doi.org/10.1109/tpami.2021.3055564
7. Touvron H. [et al.]. Training data-efficient image transformers & distillation through attention // International conference on machine learning. PMLR. 2021. Vol. 139. Pp. 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html