Д. А. Сальников

АЛГОРИТМ TURBOQUANT ДЛЯ ОПТИМИЗАЦИИ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ ПРИ ЛОКАЛЬНОМ ИСПОЛЬЗОВАНИИ LLM В ИНТЕЛЛЕКТУАЛЬНЫХ МЕДИЦИНСКИХ СИСТЕМАХ

Рассматриваются перспективы применения нового алгоритма сжатия памяти TurboQuant для оптимизации работы больших языковых моделей в закрытых контурах медицинских учреждений. Проанализированы технические особенности алгоритма, возможности обработки длинного контекста и барьеры внедрения.

Введение

Интеграция искусственного интеллекта в здравоохранение сопровождается строгими требованиями к конфиденциальности данных, что диктует необходимость локального развертывания моделей внутри контура учреждения здравоохранения, что жестко ограничивается аппаратными ресурсами, в первую очередь – объемом видеопамяти. При анализе объемных историй болезни узким местом становится кэш «ключ-значение» (KV-кэш). Представленный компанией Google алгоритм векторного квантования TurboQuant предлагает метод экстремального сжатия памяти без потери качества.

I. Технические характеристики

TurboQuant решает проблему избыточных затрат памяти за счет двухступенчатой архитектуры сжатия. На первом этапе применяется метод PolarQuant: векторы памяти трансформируются из стандартных декартовых координат в полярные, что исключает хранение избыточных констант. На втором этапе к оставшейся ошибке применяется математический метод преобразования Джонсона-Линденштрауса с использованием всего 1 бита мощности сжатия.

Данный подход позволяет квантовать KV-кэш до 3 бит без необходимости предварительного дообучения модели. Эксперименты на моделях Llama-3.1, Gemma и Mistral (бенчмарки LongBench) показали, что алгоритм способен уменьшить объем рабочей памяти как минимум в 6 раз, сохраняя полноту данных и высокую скорость поиска сходства векторных представлений.

II. Сценарии клинического применения и возможные ограничения

Ключевой сценарий применения TurboQuant в медицине – анализ длинного контекста (медицинских карт пациентов за несколько лет, расшифровок консилиумов и научных статей) на локальных серверах со средними вычислительными мощностями. Тесты типа «иголка в стоге сена» подтверждают способность алгоритма находить единичные факты (например, упоминание редкой аллергии или специфического симптома) в огромном массиве неструктурированного текста. Также технология кратно ускоряет работу векторных поисковых движков (индексов), что критически важно для медицинских систем использующих архитектуру RAG (генерация с дополненной выборкой).

Возможной проблемой может стать ограниченность сферой инференса – алгоритм решает проблему нехватки памяти при использовании модели, но первичное обучение или глубокое дообучение попрежнему требует масштабных вычислительных кластеров. Ввиду того, что технология пока не получила официальный релиз для открытого использования, отсутствуют готовые эталонные реализации для бесшовной интеграции с популярными векторными СУБД.

III. Выводы

Алгоритм TurboQuant представляет собой прорывную технологию для интеллектуальных медицинских систем, позволяющую существенно снизить стоимость аппаратного обеспечения для локального запуска LLM. Сохранение высокой точности при экстремальном сжатии KV-кэша делает возможным безопасный анализ объемных медицинских документов непосредственно в закрытом контуре учреждения здравоохранения.

Список литературы

1. TurboQuant – Extreme Compression for AI Efficiency / Electronic resource. – Mode of access: https://turboquant.net/. – Date of access: 15.03.2026.
2. RAGFlow. Open-source RAG engine for document understanding / Electronic resource. – Mode of access: https://github.com/ragflow. – Date of access: 15.03.2026.

Сальников Даниил Андреевич, ассистент кафедры интеллектуальных информационных технологий БГУИР, d.salnikov@bsuir.by.