VLM для анализа стримингового видео, применение методов динамической разреженности LLM в этой задаче

Руководитель
Чекалина Виктория Александровна

Закончила Физический факультет МГУ в 2014 году, в 2017 закончила Школу Анализа Данных, 2023 году получила степень PhD в Сколковском Институте наук и технологий по специальности “Информатика и вычислительная техника”. В Церн создала программу для генерации синтетических откликов для заданного типа частиц. Это позволило получать необходимые синтетические данные для исследований, даже когда коллайдер не был запущен. В Сколково занималась исследованиями о применении методов линейной и мультилинейной алгебры для разработки эффективных нейросетевых архитектур. В AIRI продолжила эту тему, добавив исследования в мультимодальных LLM. Автор и лектор курса по Трансформерным архитектурам, семинарист курса Deep Vision and Graphic в ШАД.
Опыт научной работы 12 лет, автор 123 научных работ по Ядерной физике, Искусственному Интеллекту, Обработке естественного языка, в т.ч. 2 работы на конференциях с рейтингом A*, 10 - в журнале квантиля Q1.

Зачем делаем?
Анализ стримингового видео в реальном времени становится все более важной задачей в таких областях, как видеонаблюдение, автоматический мониторинг и интерактивные системы. Современные большие языковые модели (LLM), способны анализировать такие видеопотоки, однако стандартные подходы к обработке таких данных требуют значительных вычислительных ресурсов, что делает их неэффективными для работы в реальном времени.

Что делаем?
Пробуем несколько подходов к уменьшению вычислительных ресурсов, среди них:
1) Dynamic Memory Compression (Динамическое сжатие памяти) [5]
Позволяет сокращать объем занимаемой памяти при обработке длинных последовательностей. Это особенно полезно при работе с длинными видеопотоками, где важно сохранять ключевые события без хранения всей истории.
2) Dynamic Sparsity (Динамическая разреженность) [4]
Позволяет исключать малозначимые части данных и фокусироваться на наиболее важных фрагментах. В случае стримингового видео это означает возможность фильтрации нерелевантных кадров.

Как достигнем цель?
Определим набор моделей, с которыми будем сравниваться, определим набор видео-стриминговых бенчмарков, на которым будем мерить качество, последовательно попробуем методы ускорения инференса и уменьшения вычислительных ресурсов.

Как измерить достижение результата?
Разработанный метод анализа стримингового видео должен обеспечивать баланс между эффективностью работы модели и качеством обработки данных. Результат считается достигнутым, если разработанный метод ускоряет время обработки видеосигнала относительно известных SOTA, при этом на видео бенчмарках [1],[2],[3] качество не падает.

[1] Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge

[2] MSR-VTT: A Large Video Description Dataset for Bridging Video and Language
[3] EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding
[4] Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer
[5] Dynamic Memory Compression: Retrofitting LLMs for Accelerated Inference (P. Nawrot, A. Łancucki, M. Chochowski, D. Tarjan, E. Ponti)