Когда слова видят, а изображения читают: исследование подходов обучения больших мультимодальных моделей

Курцев Дмитрий Витальевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: s02230428@gse.cs.msu.ru

Научный руководитель *— Кравцова Ольга Анатольевна*

Ответ на вопрос по изображению (Visual Question Answering, VQA) представляет собой сложную задачу в области мультимодального обучения, которая требует от моделей эффективно интегрировать визуальную и текстовую информацию для получения точных ответов. Для этой задачи необходимо не только глубокое понимание визуального контента, но и способность интерпретировать и с точностью отвечать на текстовые запросы. Недавние достижения в картиночно-языковых моделях (Vision Language Models, VLMs), такие как GPT4-o, Llama [1], LLaVa [2], InternVL2.5 [3], Qwen2.5-VL [4], продемонстрировали большой потенциал в решении проблем VQA. Однако успех этих моделей в значительной степени зависит от их способности обобщать знания.

В данной работе была исследована эффективность обучения VLM, усиленных с помощью Image-to-Image Retrieval-Augmented Generation (RAG) для решения задачи VQA. Конкретно, было сосредоточенно внимание на запоминании информации о картинах, включая их авторов и названия, а также имена достопримечательностей. Используя структуру RAG Image-to-Image, целью является расширение возможностей модели по извлечению и интеграции соответствующей визуальной и текстовой информации из базы знаний, что способствует повышению её эффективности при выполнении данной задачи.

Была проведена серия экспериментов, в которых различные компоненты VLMs, такие как визуальный кодировщик, языковая модель и адаптер, были обучены индивидуально и в сочетании. Этот подход позволяет определить наиболее эффективную стратегию запоминания информации. Производительность каждой конфигурации оценивается на открытом эталонном наборе данных, что позволяет провести всестороннее сравнение их точности, надёжности и возможностей запоминания.

Эта работа вносит двойной вклад: во-первых, она способствует

пониманию того, как отдельные компоненты VLM влияют на запоминание знаний, а во-вторых, она предоставляет практические знания по применению Image-to-Image RAG для специализированных задач VQA. Данные выводы служат основой для будущих исследований, направляя оптимизацию мультимодальных моделей в сценариях, которые требуют запоминания и извлечения знаний.

Литература

1. Grattafiori A. The Llama 3 Herd of Models // Computer Vision and Pattern Recognition, Nashville, USA, 2025
2. Haotian L., Chunyuan L., Qingyang W. Visual instruction tuning // In Neural Information Processing Systems, New Orleans, USA, 2023
3. Zhe C. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling // In Computer Vision and Pattern Recognition, Nashville, USA, 2025
4. Yang B. Qwen2.5 Technical Report // In Computation and Language, 2025