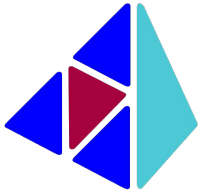


Center for
Cognitive
Modeling

Применение мультимодальных языковых моделей к задаче визуального вопросно-ответного анализа на видеоданных

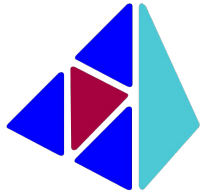
В. Семенов, С. Линок[1], Д. Юдин[2]
Лаборатория интеллектуального транспорта НКБ ВС

[1]Научный сотрудник лаборатории интеллектуального транспорта МФТИ - НКБ ВС
[2]Заведующий лабораторией интеллектуального транспорта МФТИ - НКБ ВС

Актуальность

- Рост интереса к системам VideoQA в задачах интеллектуального анализа видео
- Огромные объемы визуальных данных, требующие эффективной обработки и интерпретации
- Современные MLLM позволяют совмещать текстовое и визуальное представление данных, но плохо применимы к длинным видео
- Необходима система, способная из видео автоматически извлекать знания и отвечать на вопросы, например в планировании

Существующие подходы и их проблемы



MIST Glance and Focus SeViLA KcGA

1) MIST: Multi-Modal Iterative Spatial-Temporal Transformer (CVPR 2023)

   Модель обрабатывает длинные видео через итеративный отбор важных сегментов и регионов кадров с последующим вниманием. **Проблема:** при анализе долгих видео с множеством событий пропускаются важные кадры.

2) Glance and Focus: Memory Prompting for Multi-Event Video QA (NeurIPS 2023)

   Двухэтапная стратегия: сначала модель создает «воспоминания» о событиях (glance), затем фокусируется на релевантных фрагментах для ответа. **Проблема:** на тот момент модели не умели эффективно работать с вопросами, требующими анализа нескольких событий или их последовательности во времени.

3) SeViLA: Self-Chained Image-Language Model for Video QA (NeurIPS 2023)

   Модель использует предобученную image-language модель (BLIP-2) дважды: для выбора ключевых кадров (локализация) и генерации ответа. Обучается без разметки через self-training. **Проблема:** требуется полная разметка кадров.

4) KcGA: Knowledge-Constrained Answer Generation (AAAI 2023)

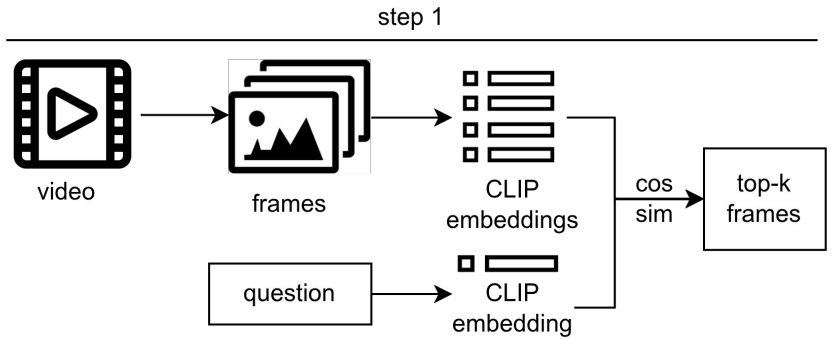
   Объединяет признаки видео и внешние знания (ConceptNet) для генерации открытых, осмысленных ответов на вопросы, даже если нужной информации нет явно в видео. **Проблема:** на тот момент стандартные модели были ограничены видимым контентом и не справляются с вопросами, требующими фоновых знаний или логического вывода.

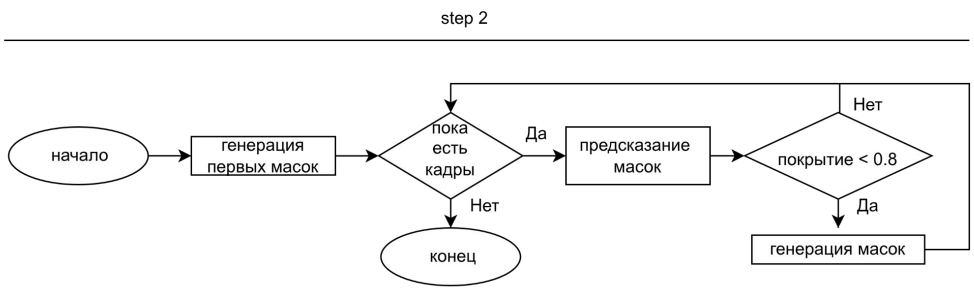
Предложенный подход (пайплайн)

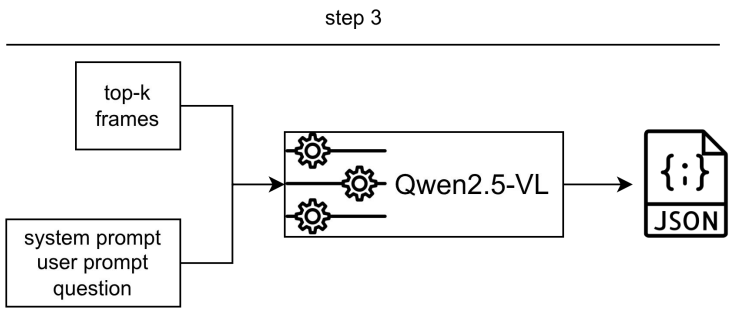
1) CLIP: отбор топ-5 релевантных кадров по вопросу

2) SAM2 Tracker: сегментация и отслеживание объектов с сохранением ID

3) Qwen2.5-VL (Vision-Language Model): генерация {reasoning, answer}







Экспериментальные результаты

Выбор базовой MLLM

1) Проведено сравнение нескольких MLLM (таблица 1), таких как Ferret [1], Shikra [2], SPHINX-Tiny [3] и Qwen2.5-VL [4] на задаче предсказания отношения между объектами. Лучшая модель Queen 2.5-VL.

2) Для данной модели дополнительно исследовалось влияние формата подачи данных об объектах на точность предсказания отношения между ними, вопрос содержал список из 50 возможных ответов (таблица 2).

Таблица 1. Сравнение MLLM

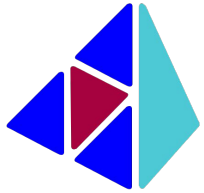
	Exact Match	BLEU	METEOR	BERTScore
Ferret	0.0565	0.0011	0.0535	0.7793
Shikra	0.0983	0.0000	0.0530	0.7825
Sphinx-Tiny	0.0491	0.0042	0.0783	0.8351
Qwen2.5-VL	0.1130	0.0018	0.1033	0.8041

Таблица 2. сравнение формата подачи данных об объектах

	top1 acc	top2 acc	top3 acc
номера на объектах	0.0513	0.1026	0.1282
названия объектов	0.1795	0.3077	0.359
номера + названия	0.2564	0.4359	0.4872

Экспериментальные результаты

Запуск на датасете LongViTU



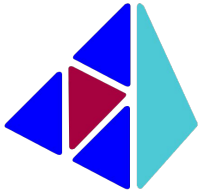
- LongViTU - VideoQA датасет на базе видео Ego4D, снятых от первого лица.
- В тестовой выборке 100 видео, длтельностью от 10 до 80 минут.
- Всего 6158 вопросов различного типа: Action (1233), Location(1006), Attribute(679), Object (657), Transition(505), Motivation(503), Function(352), Causality(342), Planning(248), Affordance(224), Interaction(195), Risk(168), Other(46).

Пример:

```
{'qa_uid': '119686',
  'question_type': 'Action',
  'question': 'What did you do to the maroon scarf on the
ironing board?',
  'answer': 'I adjusted the maroon scarf.',
  'gt answer': 'You adjusted the maroon scarf.'}
{'qa_uid': '119684',
  'question_type': 'Location',
  'question': 'Where did I adjust the maroon sweater?',
  'answer': 'I pulled the maroon sweater down.',
  'gt answer': 'In front of a wooden cabinet.'},
```

	Exact Match	BLEU	ROUGE-L	METEOR	BERTScore
0	0.0	0.071	0.203	0.238	0.89

Сравнение с существующими подходами



Для сравнения были взяты датасеты AGQA, STAR, NExT-QA и измерена точность ответа. Каждый элемент датасета представляет собой вопрос и 4 варианта ответа.

	AGQA v2 test set	STAR val set	NExT-QA val set
MIST	54.39	51.13	57.18
Glance and Focus	55.08	53.94	58.83
SeViLA	-	44.6	63.6
KcGA	-	-	28.21
(Ours)	53.12	51.85	54.40

Выводы и будущее направление

1) Модульный пайплайн решает проблему улучшает интерпретируемость
2) Сегментация + консистентные ID важны для восстановления сцены
3) Перспективы:
    a) Применение для алгоритмов планирования автономных роботов
    b) Автоматическое извлечение графа событий
    c) Интеграция с языковым агентом для цепочки рассуждений
    d) использование аудио/стерео и других видов информации

Благодарю за внимание

Контакты:

Вадим Семенов semenov.vr@phystech.edu

Благодарности:

Дмитрий Юдин (Заведующий лабораторией интеллектуального транспорта МФТИ - НКБ ВС),

Сергей Линок (Научный сотрудник лаборатории интеллектуального транспорта МФТИ - НКБ ВС).