Разработка модуля поиска человека по словесному описанию

Руководитель
Мамедов Тимур Закирович

Закончил бакалавриат ВМК МГУ в 2021 году, в 2023 году там же получил степень магистра. В настоящее время активно работает над подготовкой кандидатской диссертации в качестве соискателя в НИУ ВШЭ. В 2022 году стал лауреатом стипендии МГУ имени М.В. Ломоносова для молодых сотрудников, аспирантов и студентов, добившихся значительных результатов в педагогической и научно-исследовательской деятельности, которая присуждается 200 людям со всего университета. С 2020 года работает в ООО “Технологии видеоанализа” в качестве руководителя двух исследовательских направлений в области видеоаналитики – Силуэтов и Ситуационной аналитики. С сентября 2024 года участвует в проведении курса “Компьютерное зрение”, читаемом в МГУ, МФТИ и ШАД. Является автором 7 научных публикаций, индексируемых в Scopus и РИНЦ, среди которых можно выделить статью в журнале Q1, а также работу, представленную на конференции WACV 2025 (уровень A). С 2023 года является рецензентом в высокоуровневом журнале IEEE Transactions on Image Processing. Область научных интересов включает задачи реидентификации людей, однокамерного и многокамерного трекинга, словесного описания людей и событий, детекции.

Зачем делаем?
Так как нашей конечной целью является разработка мультимодальной LLM для интеллектуальной видеоаналитики, которая будет способна анализировать данные с реальных камер видеонаблюдения, в сценах которых чаще всего встречаются люди, то на начальном этапе возникает необходимость в разработке нейросетевого модуля, способного искать в видео человека, соответствующего словесному описанию, заданному пользователем.

Что делаем?
Разрабатываем нейросетевой модуль поиска человека по словесному описанию, основанный на мультимодальной LLM. Суть модуля заключается в том, что ему на вход подается видеопоследовательность и текстовый запрос, содержащий описание человека. Разрабатываемый модуль должен найти всех людей на видео и в качестве выхода выдать временные интервалы, когда искомый человек был на сцене, а также его bbox’ы с заданной частотой.

Как достигнем цель?
Данная задача является усложненным вариантов Text-based Person Re-identification. В упомянутой задаче имеется база данных из изображений людей и в ней нужно найти те картинки, которые соответствуют текстовому запросу пользователя. Решаемая в рамках данного проекта проблема является более универсальной: она предполагает обработку “сырого” видео, то есть нет ни выходов детектора, ни базы данных. В связи с чем работу над этим проектом предлагается построить следующим образом:

- Изучить текущие подходы Text-based Person Re-identification
- Разработать модуль, способный искать людей по словесному запросу в рамках базы данных (т.е. модуль для решения задачи Text-based Person Re-identification)
- Изучить существующие мультимодальные LLM, выбрать из них базовую
- Собрать данные с реальных камер видеонаблюдения и разметить их для задач детектирования, словесного описания людей, а также выполнить разметку на временные интервалы появления того или иного человека на сцене
- Дообучить выбранную LLM для задачи детектирования людей

- Интегрировать модуль, способный искать людей по словесному запросу в рамках базы данных, в дообученную LLM для решения целевой проблемы

Как измерить достижение результата?
Для задачи Text-based Person Re-identification общепринятой метрикой является Rank-1, с помощью которой имеется возможность оценивать качество модуля, способного искать людей по словесному запросу в рамках базы данных, а также сравниваться с текущими state-of-the-art подходами. Оценка качества детекции производится с помощью метрики mAP. Для оценки решения конечной задачи будет разработана отдельная метрика.