Звуковое описание видео для людей с нарушениями зрения

Литвинов Иван Борисович
Факультет космических исследований, Москва
E-mail: litvan007@gmail.com
Соавторы: Совелько Д., Семенов А., Михальцов Д.

Развитие онлайн-видеосервисов выявило проблему доступности видеоконтента в Интернете для людей с ограниченными возможностями, особенно для тех, у кого есть нарушения зрения. Очень важно, чтобы видео на экране устройства было понятным, доступным и удобным для людей, страдающих этим заболеванием. При современном развитии вычислительных ресурсов, технологий машинного обучения и больших данных эта проблема может быть решена с помощью перечисленных инструментов. Целью этого исследования является создание инструмента, который поможет людям с нарушениями зрения понять, что происходит в конкретной сцене фильма.

Описанная проблема может быть решена с помощью методов image2text, которые включают в себя большие языковые модели (LLM). Однако существуют и другие подходы к описанию происходящего на сцене, и было решено использовать архитектуру слияния vision-language, ориентированную на объектно-семантическую подготовку для задач Vision-Language (OSCAR) [1], с архитектурой обнаружения объектов, придающей значение визуальным представлениям в моделях Vision-Language (VinVL) [2]. Этот подход основан на использовании предварительно обученных моделей для распознавания объектов и их семантических связей на изображениях, а также генерации текстовых описаний сцен на основе их визуального содержания. Для создания голоса за кадром использовались модели синтеза речи Silero [3]. Для разработки использовались Python (pyscenedetect, opencv, torch, flask и другие), React JS, Docker и nginx.

Чтобы протестировать эффективность предложенного подхода, было создано веб-приложение, позволяющее пользователю загружать любое видео с YouTube и смотреть его голосом за кадром. Мы также получили экспертную оценку от человека с полным отсутствием зрения. В частности, он оценил высокую детализацию описания сцены, а также отметил, что для большего погружения необходимо лучше проработать вопрос соединения звуков происходящего с закадровым голосом, что будет учтено в дальнейшей работе. Однако в ходе дальнейших исследований планируется провести эксперименты с участием большей группы респондентов, имеющих проблемы со зрением, которые будут оценивать наш инструмент по различным критериям, таким как понятность, информативность, естественность и привлекательность описаний.

В этой работе мы представили первую версию нашего инструмента [4], который позволяет людям с нарушениями зрения получать текстовые описания сцен видео на основе их визуального содержания. Наш инструмент имеет микросервисную структуру, которая позволяет встраивать различные алгоритмы описания сцен, в том числе с использованием современного LLM. Наша концепция также может быть применена не только к биомедицинским задачам, но и для обучения людей с отсутсвием зрения на любых онлайн курсах.

Источники и литература

1) Li, Xiujun et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020, 2020.
2) Zhang, Pengchuan et al. VinVL: Making Visual Representations Matter in Vision-Language Models. CVPR 2021. 2021.

3) Silero Team. (n.d.). Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks. GitHub. Retrieved April 6, 2023, from https://github.com/snakers4/silero-models.

4) Ivan Litvinov et al. Audio description of video for visually impaired people [Computer software]. GitHub. https://github.com/litvan007/Audio-description-of-video-for-visually-impaired-people.