

Отчет по проекту «Автоматическое выделение таймкодов начала и конца содержания эпизода сериала»

Тип проекта - Статья

Автор: Гранкина Елизавета Григорьевна

1. Описание

Проект выполняется на основе научной статьи «Automatic Detection of Intro and Credits in Video using CLIP and Multihead Attention» ([ссылка на статью](#)).

Основная цель проекта — разработка методов и моделей, которые автоматически определяют временные метки:

- начала основного контента эпизода сериала (после вступительной заставки),
- конца основного контента (перед финальными титрами).

2. Что уже сделано

2.1. Сбор и разметка датасета

- Вручную собрано более 100 серий (около 100 часов видео) из различных сериалов.
- Для каждой серии размечены таймкоды:
 - начала основного контента,
 - конца основного контента.
- Применены методы борьбы с дисбалансом классов:
 - взвешенная функция потерь,
 - Balanced Sampler в DataLoader,
 - аугментация для класса «титры».

2.2. Предобработка данных

- Используется FFmpeg для декодирования видео (1 кадр/сек) и извлечения аудио.
- Для аудио строится мел-спектrogramма, для видео — эмбеддинги через CLIP.

2.3. Разработка видео-модели

- **Реализована модель на основе CLIP и Multi-Head Attention Transformer.**
Именно такая модель используется в статье для извлечения тайм кодов.
- Извлечение признаков: кадры видео обрабатываются с частотой 1 FPS, каждый кадр кодируется в 512-мерный вектор.
- Архитектура включает **16 голов внимания и 60 линейных классификаторов для покадровой классификации**.
- **Модель обучена и протестирована** на собранном датасете.

2.4. Разработка аудио-модели

- **Реализована дополнительная модель обработки аудиодорожки:**
 - преобразование аудио в мел-спектрограммы,
 - использование CNN-энкодера для извлечения признаков,
 - дополнительно рассчитана косинусная схожесть с шаблоном заставки.
- Оценки аудио и видео объединяются с весами (70% видео + 30% аудио) для формирования финального прогноза.

3. Что еще не сделано

3.1. Отсутствует слияние признаков в единую бимодальную модель

- На текущий момент видео- и аудио-модели работают отдельно, их прогнозы объединяются только на уровне постобработки.
- Планируется реализация бимодальной архитектуры с использованием Random Forest или другого классификатора, учитывающего оба типа признаков с весами классов.
- Это ключевой этап, который должен улучшить точность и снизить ошибки.

3.2. Не проведено полноценное сравнение моделей

- Предварительные оценки показывают, что видео-модель точнее аудио-модели, но систематическое сравнение на тестовой выборке еще не завершено.
- Не рассчитаны окончательные метрики (MAE, Precision, Recall, F1) для комбинированного подхода.

4. Планы до завершения проекта

1. Завершение мультимодальной модели
 - Объединение видео- и аудио-признаков в единую архитектуру.
 - Обучение и валидация комбинированной модели.
2. Проведение сравнительного анализа
 - Сравнение точности видео-, аудио- и мультимодальной моделей.
 - Фиксация итоговых метрик (MAE, F1, точность).
3. Подготовка финальной документации и кода
 - Оформление репозитория на GitHub.
 - Подготовка презентации для защиты.

5. Заключение

На текущий момент выполнены ключевые этапы проекта:

- Собран и размечен датасет
- Реализованы и обучены видео- и аудио-модели
- Настроен пайплайн предобработки

Осталось завершить интеграцию моделей, провести финальное тестирование и оптимизацию. **Проект находится на стадии 60–70% завершенности.**