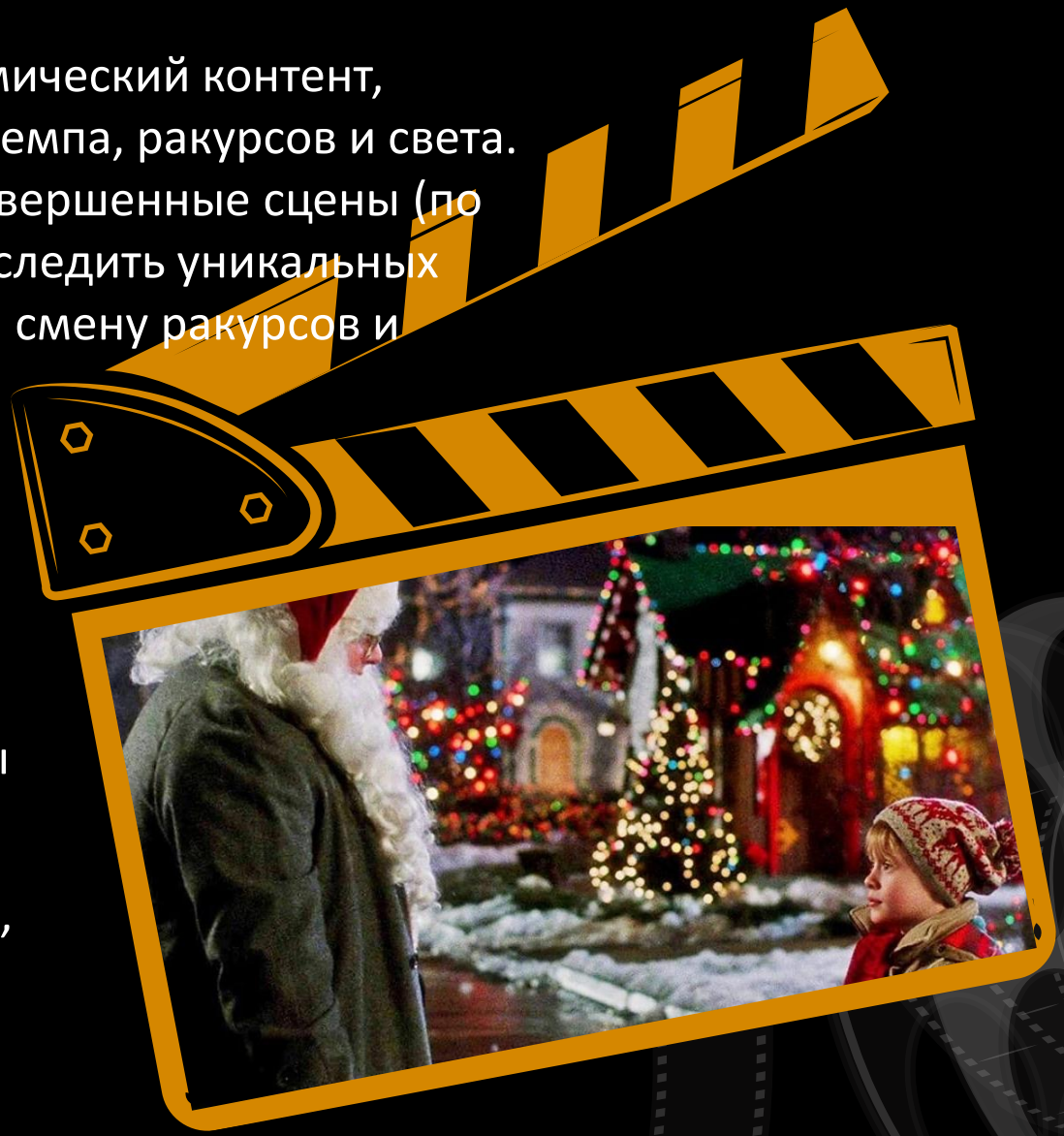
SKILLFACTORY × National Research Tomsk State University

Сквозная идентификация объектов и сегментация видеоконтента KION на логические сцены

Команда: Христофорова Полина Андреевна

Проблема

Онлайн-видео в KION — это мультикамерный, динамический контент, который включает диалоги, сцены действия, смену темпа, ракурсов и света. Как автоматически разделить видео на логически завершенные сцены (по смыслу, а не по смене кадров)? Как определить и отследить уникальных персонажей или объекты сквозь сцены, несмотря на смену ракурсов и изменения внешности?



Цель

Разработать систему для:

- Разбиения видео на шоты (кадры одной камеры)
- Объединения шотов в сюжетные сцены
- Идентификации и трекинга объектов через сцены

Данные

Задача решалась на фильме и трейлере ‘Home alone’

Блок	Модуль	Задачи	Технологии
Сценарная сегментация	Сегментация по аудио	Выделение метаданных в аудио	RMS, Spectral centroid, ZCR, librosa
		Сегментация на сцены по метаданным	
	Сегментация по видео	Определение сцен по видео	PySceneDetect, VideoFileClip
		Объединение визуальных сцен на основе выделенных отрезков	
	Интеллектуальное объединение аудио и видео	Семантическое объединение	CLIP, cv2
		Финальное объединение на визуальных эмбеддингах	
Детекция и трекинг	Основной подход	Детекция лица	YOLO(yolov8n-face), FaceAnalysis(buffalo_l)
		Определение персонажа	
	Финальное объединение лиц	Объединение найденных персонажей на основе мультиэмбеддинга	networkx

Сегментация

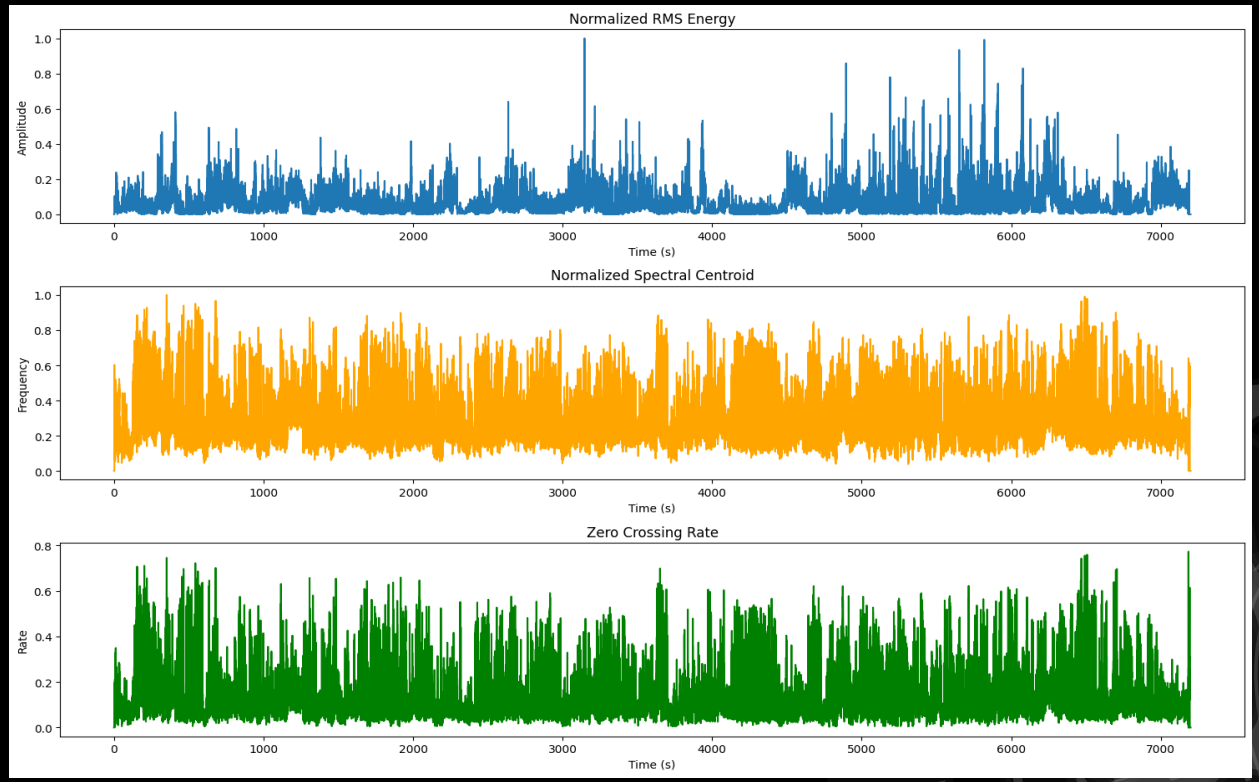
Сегментация по аудио

На первом шаге была выделена основная аудио метаинформация

- RMS - среднеквадратичное значение энергии / громкость
- Spectral centroid - спектральный центроид (яркость/шаг)
- ZCR - Частота пересечения нуля (тип шума/звука)

Далее на основе этой информации фильм был разбит на логические аудиосцены по параметрам:

- rms_threshold=0.3
- centroid_threshold=0.4
- min_scene_duration=3.0



Сегментация по видео

Итог конвертации шотов в сцены (1923 -> 293)

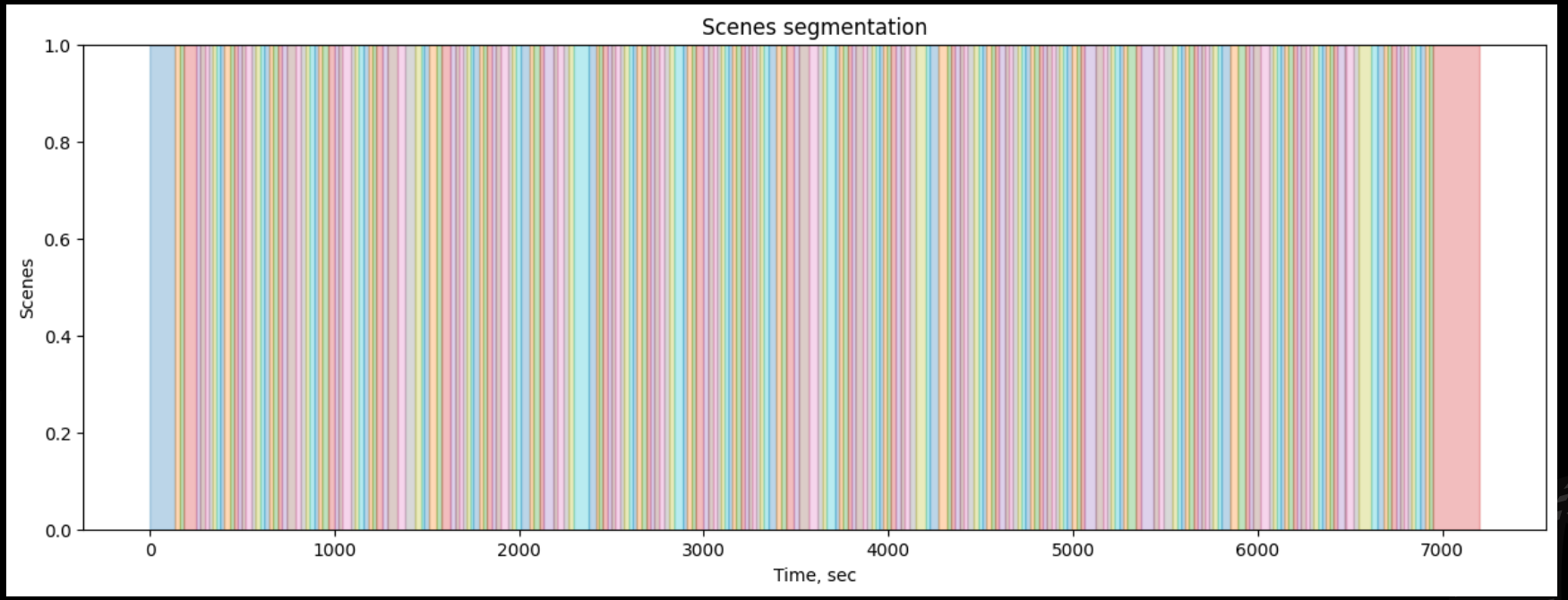
```
100%|██████████| 1923/1923 [01:21<00:00, 23.74it/s]
Detected 293 scenes in visual:
        - scene #1: 0.00 - 51.01 (51.01 sec)
        - scene #2: 51.01 - 137.89 (86.88 sec)
        - scene #3: 137.89 - 161.12 (23.23 sec)
        - scene #4: 161.12 - 185.23 (24.11 sec)
        - scene #5: 185.23 - 206.54 (21.31 sec)
        - scene #6: 206.54 - 230.40 (23.86 sec)
        - scene #7: 230.40 - 252.54 (22.15 sec)
        - scene #8: 252.54 - 273.77 (21.23 sec)
        - scene #9: 273.77 - 296.05 (22.27 sec)
        - scene #10: 296.05 - 316.98 (20.94 sec)
```

Сцены выделяются, основываясь на изменениях, когда камера переключается на новый кадр, изменение местоположения, освещения или движения.

В первом приближении создается большое количество шотов, которые объединяются в сцены на основе визуальной схожести (hist_threshold=0.35), резких изменений и продолжительности

Интеллектуальное объединение

Финальный шаг заключается в формировании единых сцен на основе выделенных из аудио и видео, где audio scenes определяют крупные смысловые блоки, а video scenes отвечают за вспомогательное уточнение



Грубое объединение дополнительно сглаживается дополнительным объединением сцен на основе визуальных эмбеддингов CLIP модели с threshold=0.8. В итоге получаем сегментацию из 254 сцен (на графике наглядно)

Кадры из итоговых рандомных сцен

Детекция

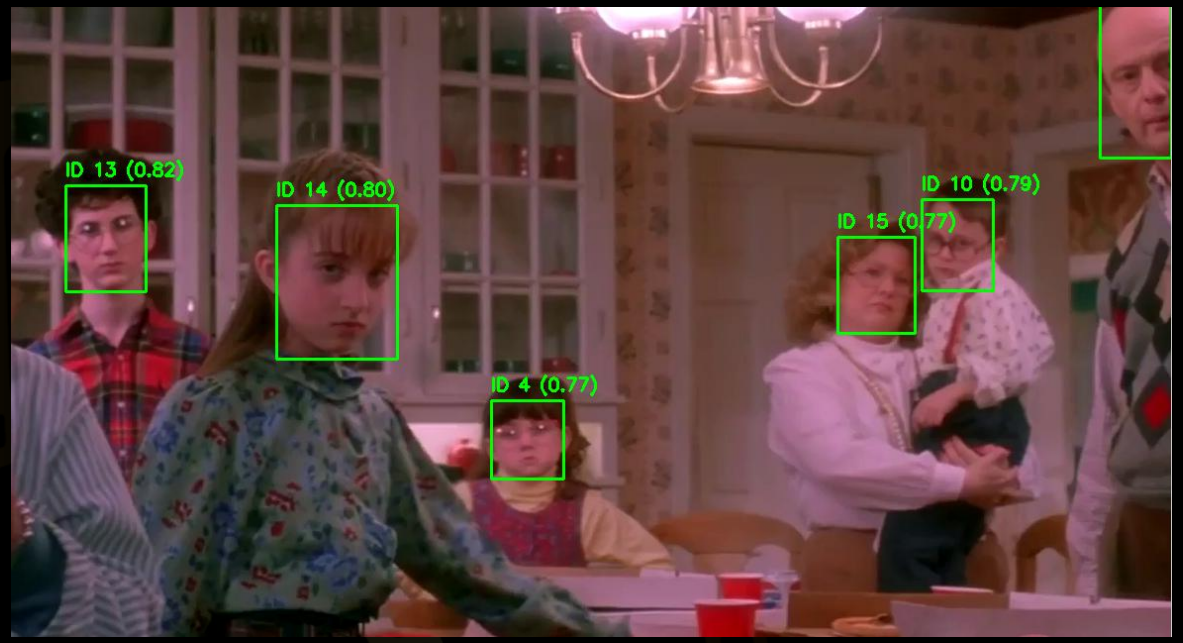
Основной подход

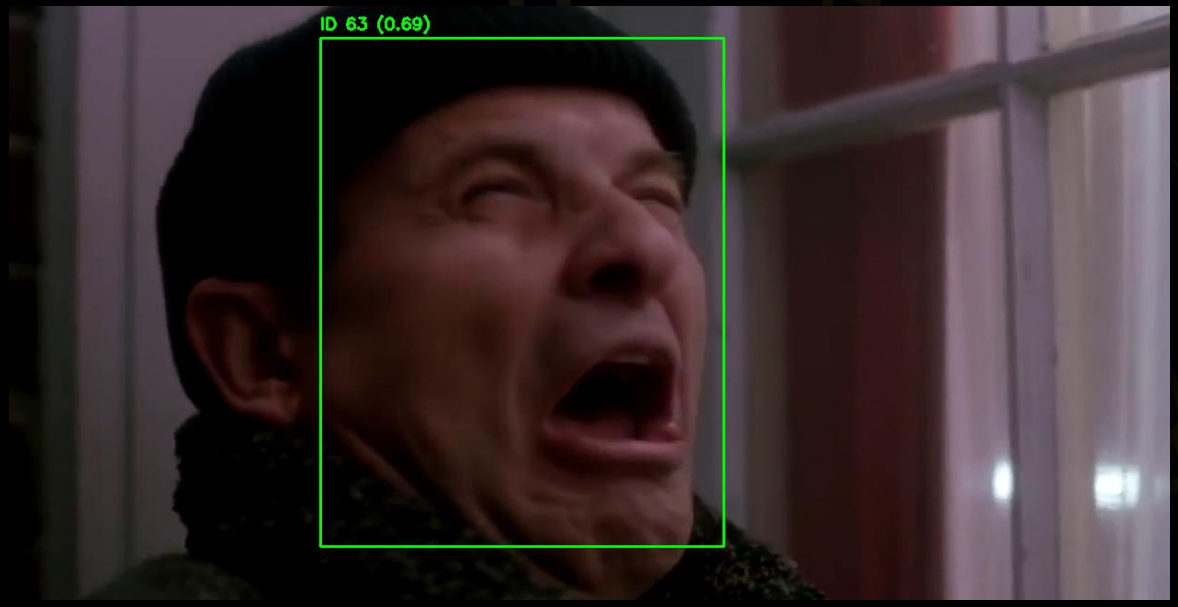
Детекция лиц была реализована на основе модели yolov8n-face с последующим определением персонажа с применением FaceAnalysis модель buffalo_l.

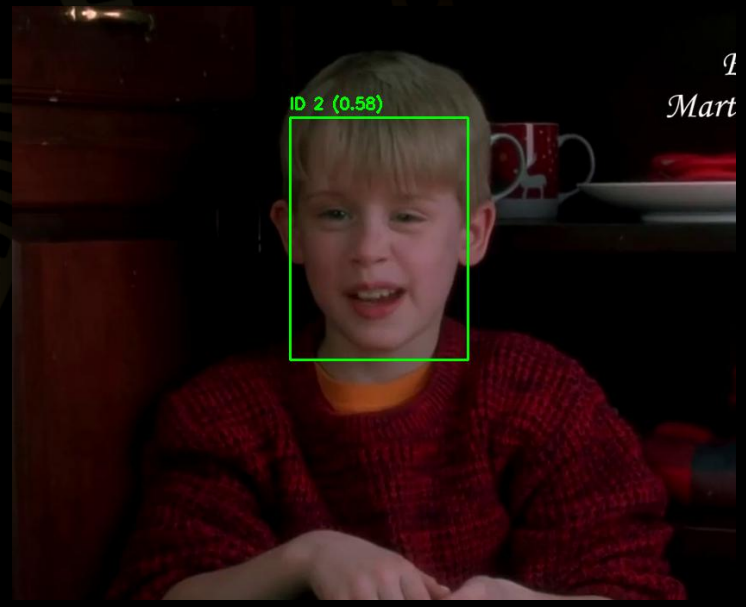
Основные шаги:

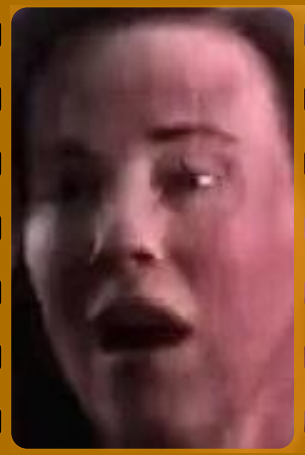
1. Детекция лиц на кадре с помощью YOLO
2. Более точно определяем границу лица
3. Производим нормализацию яркости/контраста и повышаем резкость
4. Вычисляем эмбеддинг с помощью buffalo_l. При этом использовалась технология сглаживание эмбеддинга с использованием экспоненциального скользящего среднего (EMA)
5. Вычисляем косинусное расстояние со всеми накопленными лицами из базы и определяем id персонажа
6. Сохраняем лицо и его эмбеддинг в папку по соотвутствующему id-шнику


ID 2 (0.78)
ID 4 (0.78)


ID 13 (0.82)
ID 14 (0.80)
ID 4 (0.77)
ID 15 (0.77)
ID 10 (0.79)


ID 63 (0.69)


ID 2 (0.58)

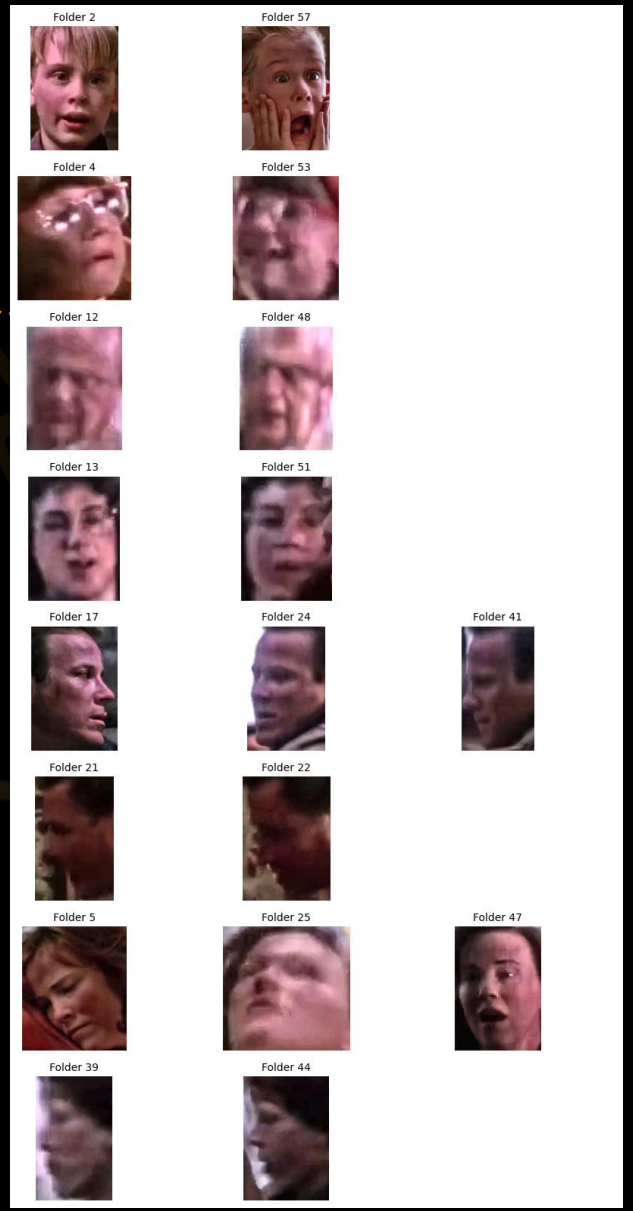


Финальное объединение лиц

На конечном шаге алгоритма, для каждого из найденных 74 персонажей были определены косинусные расстояния каждого к каждому - близкие пары с последующей группировкой полученных пар по принципу связных компонент для объединения найденных совпадений под единый id персонажа

Топ-3 найденных лиц по частоте

ID лиц к объединению



Преимущества решения

- Алгоритм реализован под новую GPU архитектуру
- Быстрая отработка даже на 8 GB
- Применение современных моделей и визуальных эмбеддингов на EMA
- Гибкая архитектура с возможностью масштабирования и дообучения
- Адаптация под сложные мимические и динамические сцены фильма

Зоны развития

Сегментация

При смысловой сегментации можно учесть:

- смену локации с детекцией конкретных предметов
- долгое отсутствие диалога
- начало музыки / экшена
- Automatic Speech Recognition (ASR)

Если же механизм будет использоваться для деления фильма на кадры относительно присутствия тех или иных актеров, то имеет смысл внедрить детекцию лиц в алгоритм сегментации

Детекция

Детекция лиц на трейлире может быть использована в решении сразу нескольких задач:

1. На данных можно сформировать датасет
2. Данные можно разметить и использовать в качестве датасета для дообучения моделей алгоритма
3. На нем можно проверить качество детекции, сравнив стоковые снимки актеров с задетекченными лицами

Спасибо за внимание

И с наступающим новым годом