ГИД

Как с помощью AI в тысячах видео найти нужный кадр

Спикер: Соколов Александр

ГИД Дата

- Обработка данных в интересах Газпром Медиахолдинга

40+ телеканалов

10+ радиостанций

Суббота! телеканал

ТНТ

PREMIER

ПЯТНИЦА!

COMEDY RADIO

ЮМОР FM

ГИД Дата

- Обработка данных в интересах Газпром Медиахолдинга
- Входит в экосистему ГИД

ГАЗПРОМ ID

Ваш ключ к вселенной возможностей и персональному пространству

ГПМ Дата

- Обработка данных в интересах Газпром Медиахолдинга
- Входит в экосистему ГИД
- Один из продуктов ГИД Дата «Video Tagging System»

40 000 видео

18 000 часов контента

25 000 000 векторов в базах данных


ПОИСК
ПО ЦВЕТУ

01

Постановка задачи

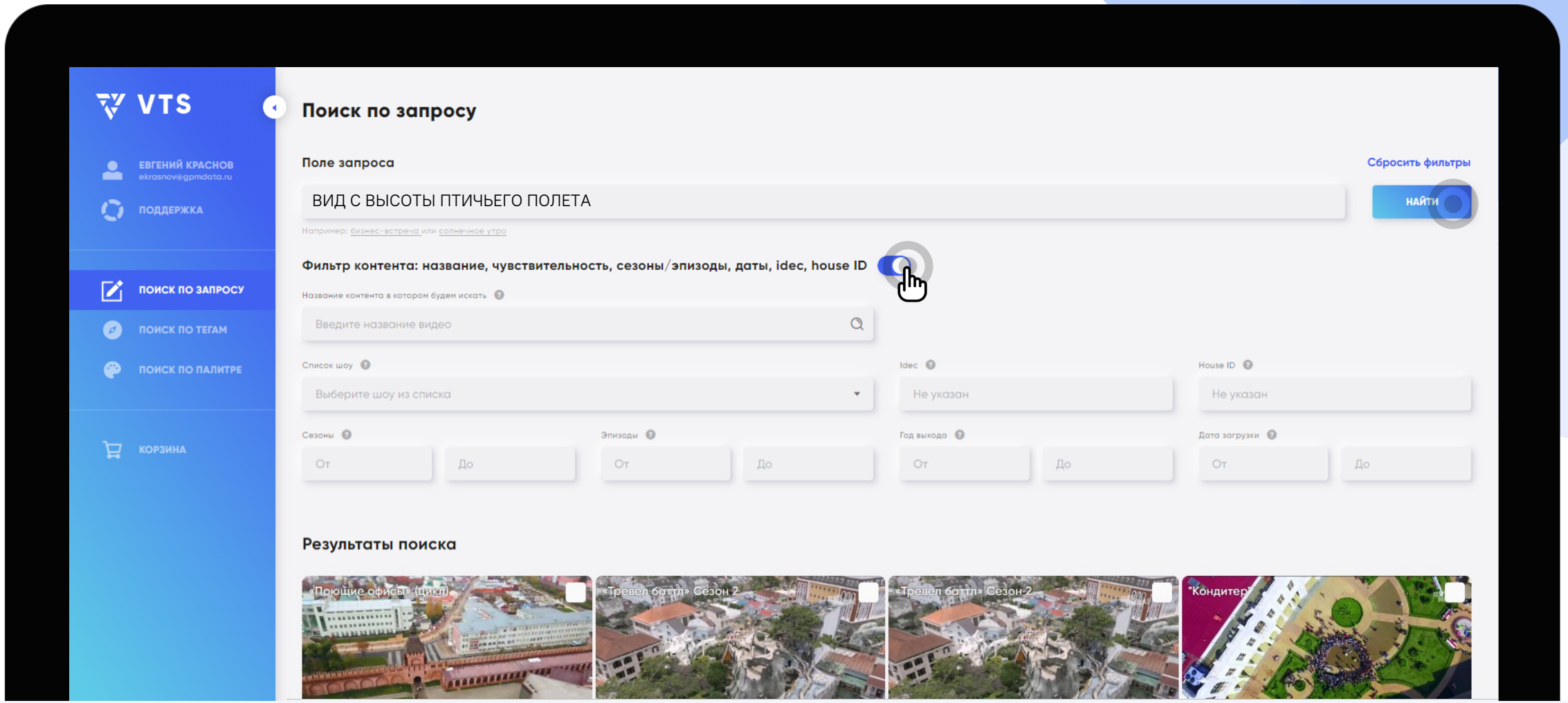
Постановка задачи

Запрос от телеканалов
на поиск по видеоконтенту.

Все функции
с исключительно
прикладным значением.

Требования

- [x] Поиск по произвольному запросу

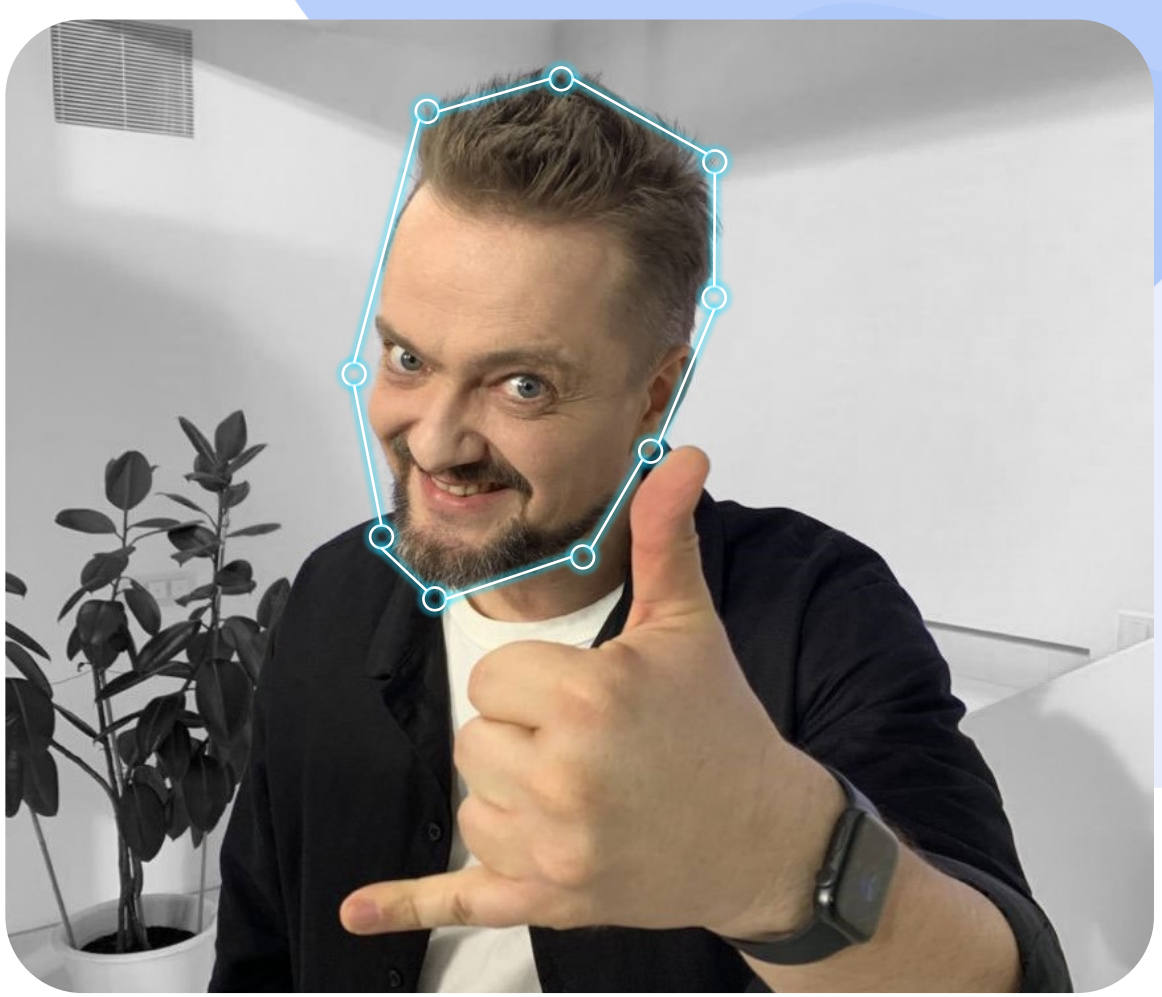


Требования

- [x] Поиск по произвольному запросу

Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
  - База персон
  - Получение видео и кадра найденной персоны
  - Поиск неограниченного круга персон, расширение библиотеки самостоятельно пользователем



Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи

  - Создание субтитров
  - Поиск по тексту/фразе
  - Модерация (мат, ругань, прочие нежелательные выражения – имена, названия компаний и пр.)
  - Подготовка монтажных листов
  - Поиск по целым словарям, строгий/нестрогий поиск

Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
  - Классы COCO
  - Поиск нужных объектов

Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы

Требования

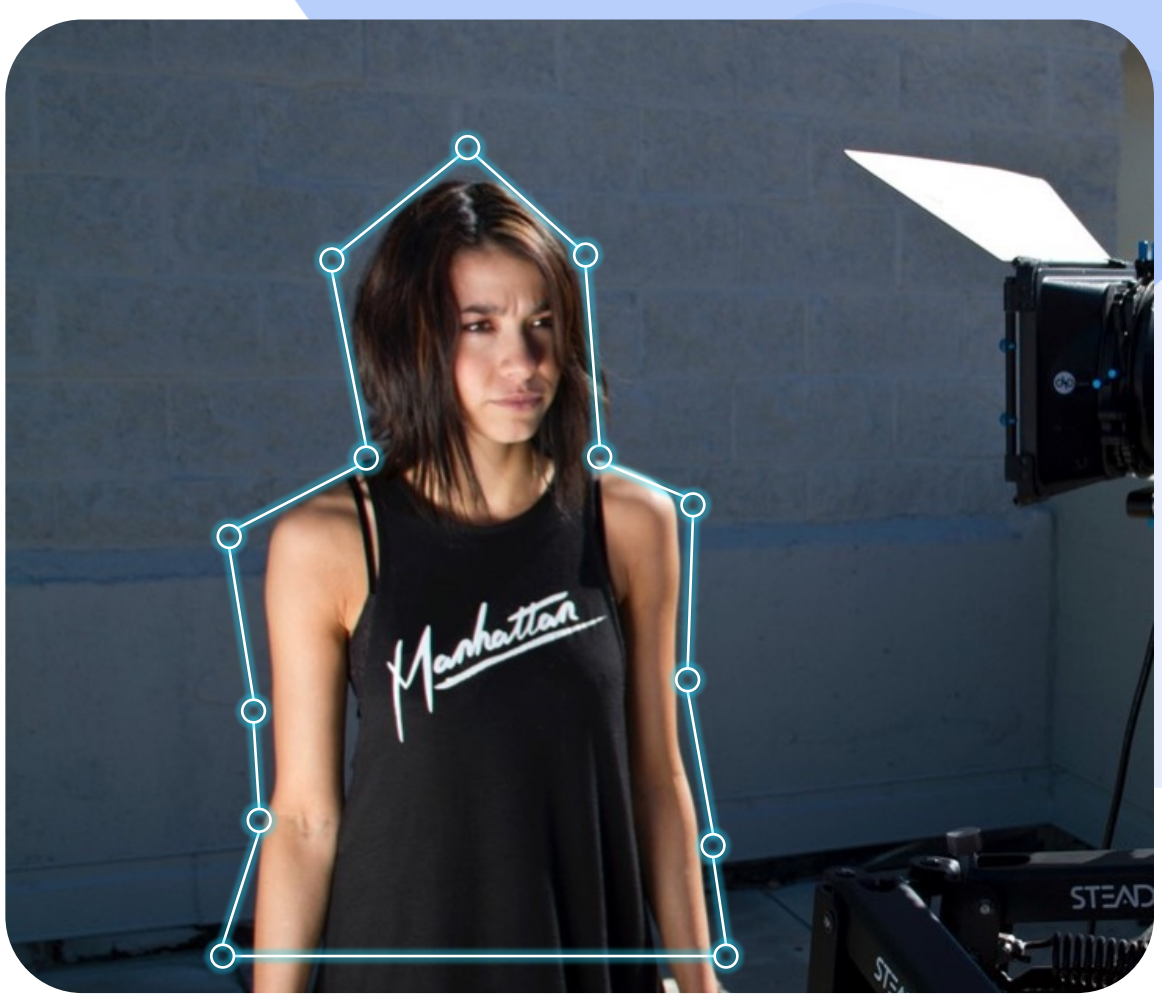
- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы
- [x] Текст в кадре
  - Распознавание надписей
  - Поиск по распознанному тексту

Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы
- [x] Текст в кадре
- [x] Преобладающая палитра
    - Пример:
    - 60% голубой,
    - 30% зеленый

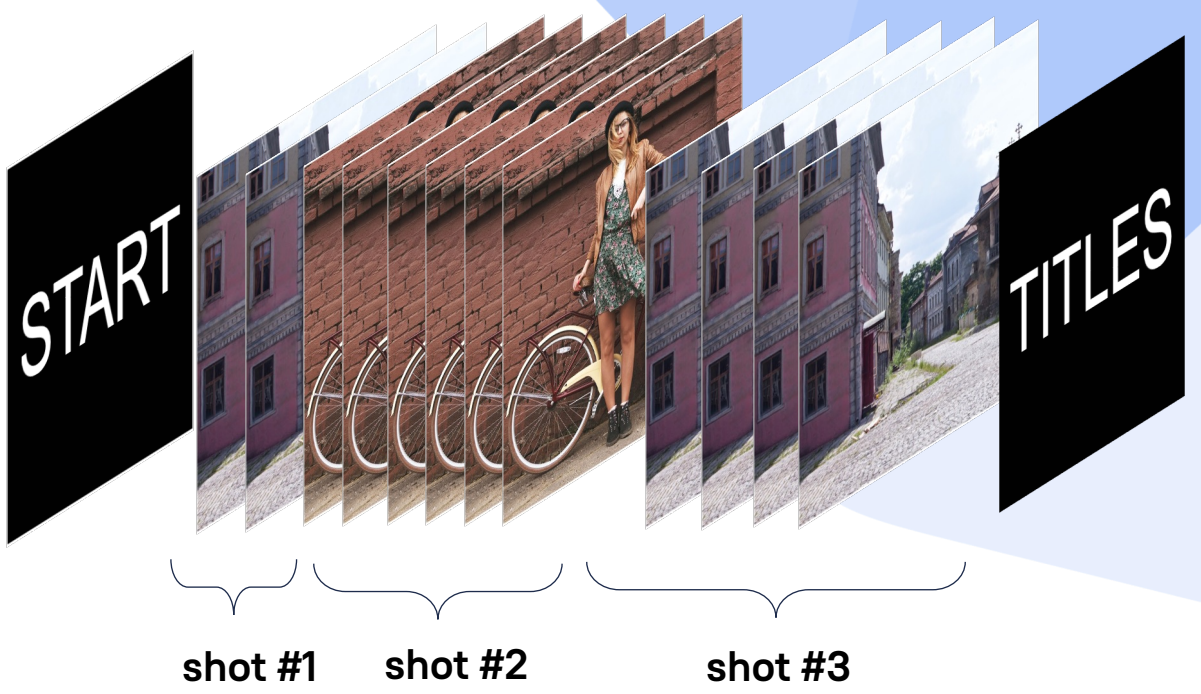
Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы
- [x] Текст в кадре
- [x] Преобладающая палитра
- [x] Кинематографический план



Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы
- [x] Текст в кадре
- [x] Преобладающая палитра
- [x] Кинематографический план
- [x] Границы сцен (shots)



Требования

- [x] Поиск по произвольному запросу
- [x] Распознавание персон
- [x] Распознавание речи
- [x] Распознавание объектов
- [x] Действия людей/позы
- [x] Текст в кадре
- [x] Преобладающая палитра
- [x] Кинематографический план
- [x] Границы сцен (shots)
- [x] Количество людей в кадре

02

Фичи

Фичи

- Эмбеддинги лиц
- Speech2Text
- Классы действий
- Распознанные надписи
- Эмбеддинг палитры
- Кинематографический план
- Границы сцен (shots)
- Эмбеддинг описания кадра
- Классы объектов и их количество в кадре

03

Модели

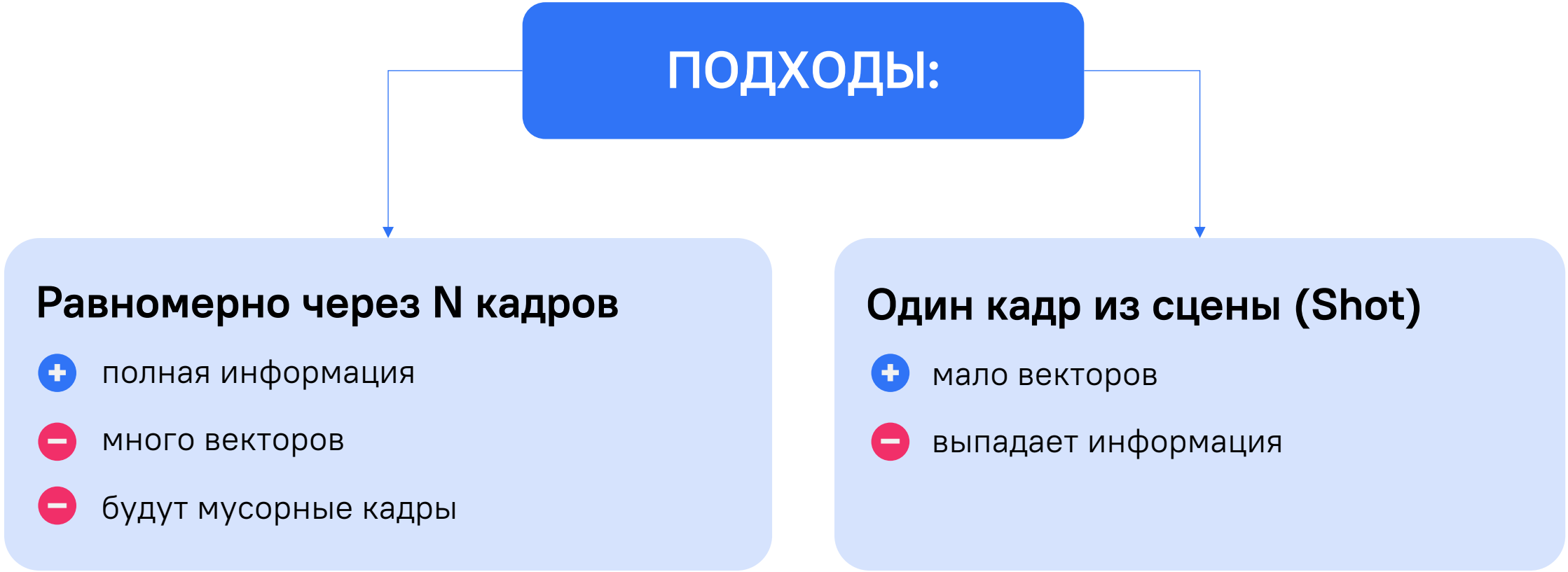
Поиск
по описанию

Модели: поиск по описанию

- Поиск нужного кадра по его описанию
- Модель: ruCLIP (СБЕР)
- Нативный русский язык
- Датасет 240 м пар текст-картинка
- Эмбеддинг 512 X float

Модели: поиск по описанию

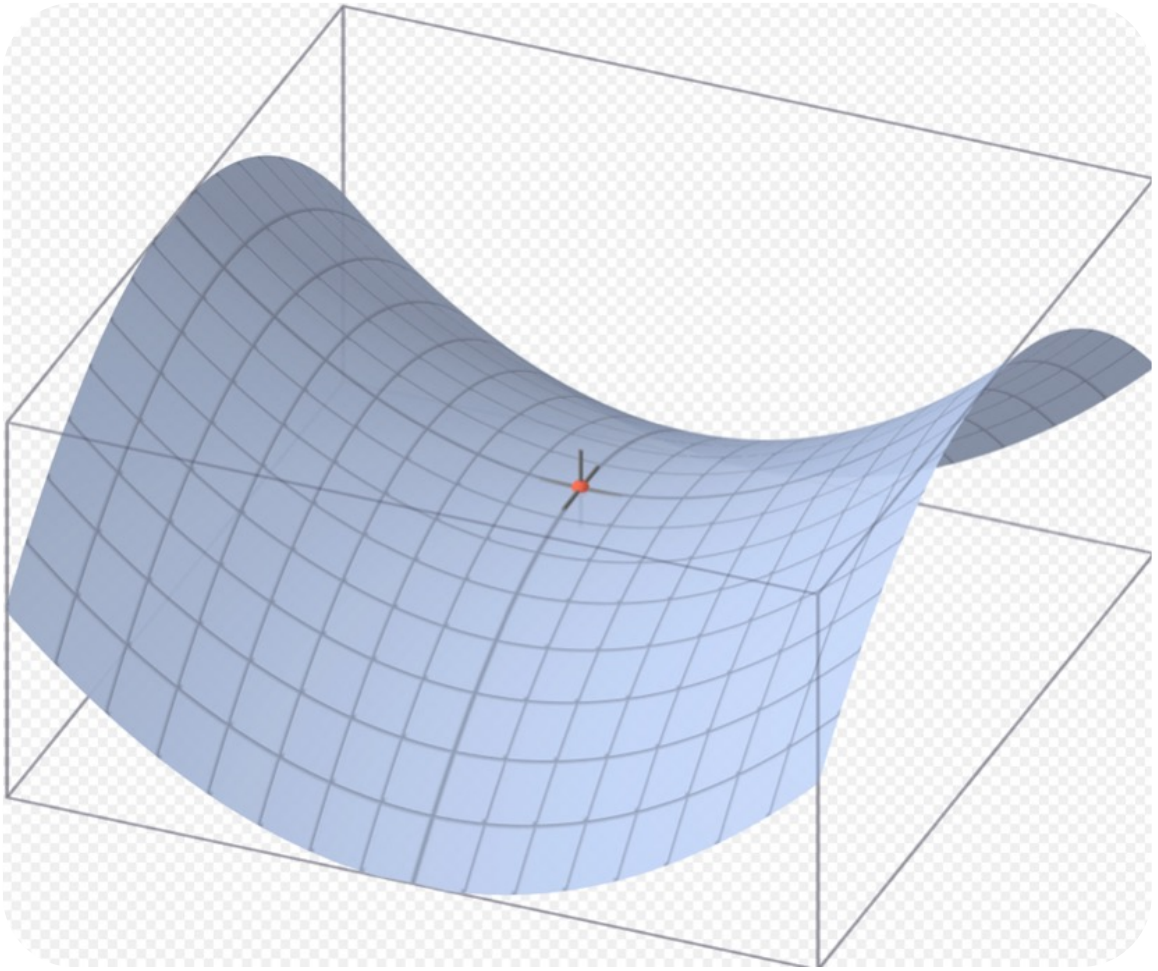
Проблема выбора кадров из видео для эмбеддингов



Модели: поиск по описанию

Задача:

- Минимальное количество векторов в базе данных
- Максимальная полнота информации из видео



Модели: поиск по описанию

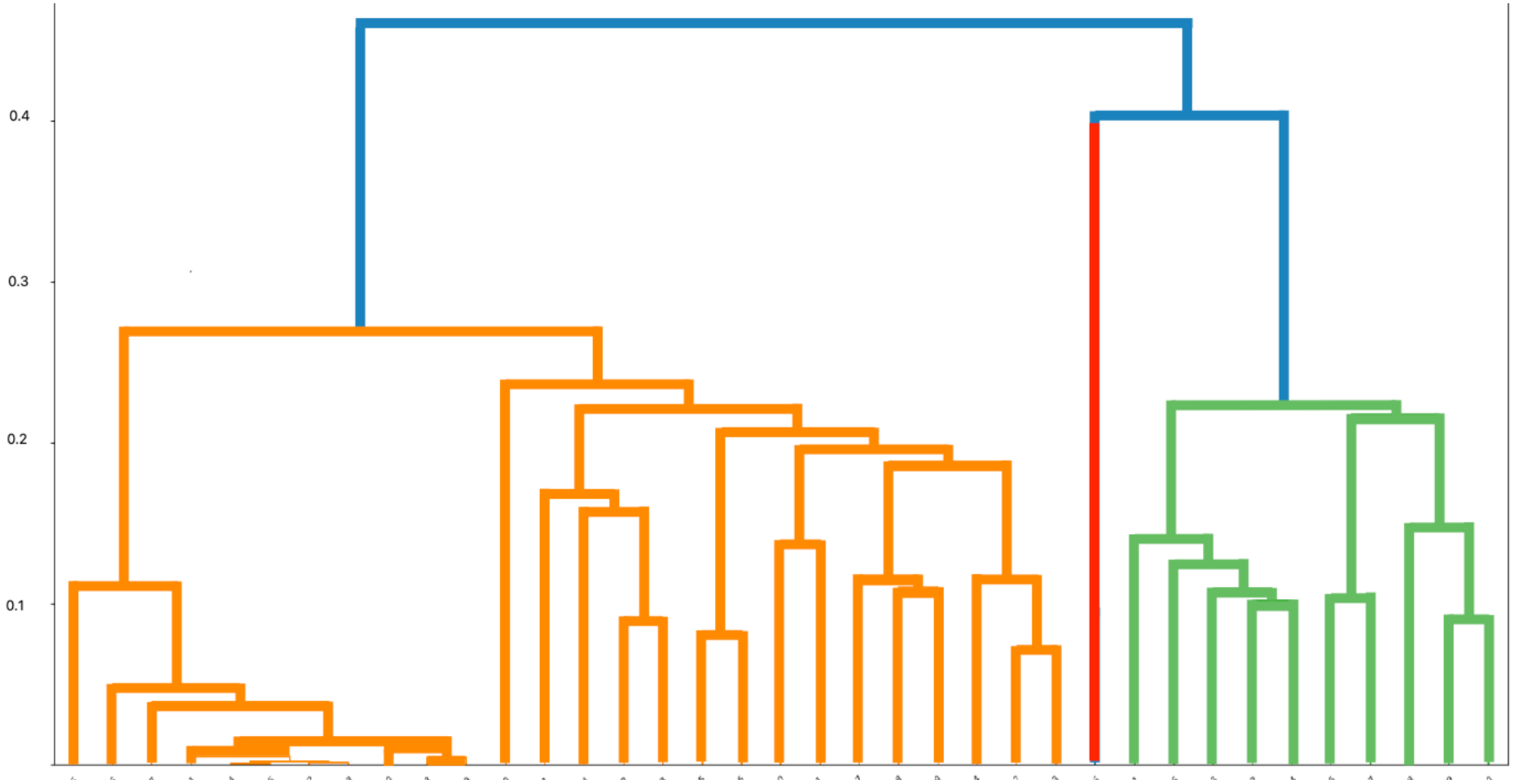
Задача:

- Минимальное количество векторов в базе данных
- Максимальная полнота информации из видео

Решение – кластеризация

- Метод Hierarchical agglomerative clustering (HAC)
- Вычислительная сложность SLINK for single-linkage O(n2)
- Кластеризация «скользящим окном»

Модели: поиск по описанию:
дерево кластеризации кадров по эмбеддингам



Модели: поиск по описанию

Группировка кадров в кластер:

для аналога shot separation. margin 0.26

для границ «смысловых сцен» margin 0.15

Модели: поиск по описанию.
Кадры с границ кластеров

Модели: поиск по описанию.
Кадры с границ кластеров

Модели: поиск по описанию.
Кадры с границ кластеров

Модели: поиск по описанию.
Кадры с границ кластеров

Модели: поиск по описанию.
Кадры с границ кластеров

Модели: поиск по описанию.
Детекция монтажных склеек

Модели: поиск по описанию.
Детекция монтажных склеек

Модели: поиск по описанию

Выгоды кластеризации

- Плюсы обоих подходов (малое количество векторов и полная информация из видео)
- Получение границ «смысловых сцен»
- Удаление кадров монтажных склеек

Модели: поиск по описанию

Тонкости реализации

- [x] Адаптивное окно кластеризации
    - размер буфера 64 фрейма
    - защита от «разделения» последней сцены по границе буфера
    - фреймы из последней смысловой сцены не попадают в дальнейшую обработку – остаются в буфере

Модели: поиск по описанию

Тонкости реализации

- [x] Адаптивное окно кластеризации
- [x] Предварительный отбор кадров на кластеризацию
  - Из окна длиной 6 кадров отбирается 2, 4 и 6 кадры
  - Из них выбирается один кадр с наибольшей резкостью
  - Отобранный кадр помещается в адаптивный буфер

Модели: поиск по описанию

Кластеризация, к сожалению, не заменяет Shot Separation

- Большое пороговое значение может приводить к склеиванию сцен
- Малое – частому дроблению сцен

Модели: поиск по описанию.
Проблема: склеивание сцен

Косинусное расстояние < 0.1

04

Модели

Готовим Whisper
правильно

Модели: Whisper

Speech-to-text – применение

- Субтитрование видео и создание монтажных листов
- Поиск по ключевым фразам
- Поиск по смыслу
- Модерация по речи

Модели: Whisper

- Whisper-v2-large
- Версия FasterWhisper
- Whisper stable-ts

Модели: Whisper

Модель от OpenAI

- ✓ Высокое качество транскрибации
- ✓ Мультиязычность
- ✓ Пунктуация + орфография
- ✓ Таймстемпы
- ✗ Галлюцинации

Модели: Whisper. Галлюцинации

- Фиксированные фразы
  - Появляются во время пауз в речи
  - Вызваны обучением на открытых данных

Спасибо за просмотр!

Продолжение следует

Ставь лайк, коммент

Субтитры подготовлены...

Модели: Whisper. Галлюцинации

- Фиксированные фразы
- Зацикливание на одном звуке
  - Появляются на растянутых гласных звуках
  - Десятки повторений
  - Последовательность 1-5 символов

Да-а-а-а-а-а-а-а-а

Ээээээээээээээээ

Модели: Whisper. Галлюцинации

- Фиксированные фразы
- Зацикливание на одном звуке
- Повторение фраз
  - Бесконечное повторение одной фразы
  - Остановка распознавания речи

```
00:00:55.080 --> 00:00:55.800
Как Решка ляжет.
00:00:55.940 --> 00:00:56.400
Орел.
00:00:57.300 --> 00:00:57.680
Решка.
00:01:23.720 --> 00:01:27.560
Решка.
00:01:28.040 --> 00:01:29.959
Решка.
00:01:43.420 --> 00:01:44.099
Решка.
```

Модели: Whisper.
Метрики качества

- ☑ Собран собственный датасет с эталонными субтитрами
- ☑ Взвешенный WER

  - Без обработки. Сравнение слов и знаков препинания.
  - Удаление знаков препинания. Сравнение только слов.
  - Удаление знаков препинания и приведение к нижнему регистру. Сравнение только слов.
  - Посимвольное сравнение приведенного к нижнему регистру текста без знаков препинания и пробелов.

Модели: Whisper.
Борьба с галлюцинациями

- [x] Whisper из коробки WER 0.55

  - Практически всегда начинает галлюцинировать
  - Иногда не выходит из галлюцинирования до конца аудио

Модели: Whisper.
Борьба с галлюцинациями

- [x] Whisper из коробки WER 0.55
- [x] VAD Voice Activity Detection WER 0.21
  - PyAnnote-audio
  - Нарезка аудио на фрагменты
  - Обработка 100% фрагментов
  - Средний WER уменьшился с 0.55 до 0.21

Модели: Whisper.
Борьба с галлюцинациями

- [x] Whisper из коробки WER 0.55
- [x] VAD Voice Activity Detection WER 0.21
- [x] Фильтрация «стоп-слов» WER 0.20
    - Словарь «стоп-слов» («подписывайтесь на канал», «ставьте лайки»...)
    - Regexp
    - Средний WER уменьшился до 0.20

Модели: Whisper.
Борьба с галлюцинациями

- [x] Whisper из коробки WER 0.55
- [x] VAD Voice Activity Detection WER 0.21
- [x] Фильтрация «стоп-слов» WER 0.20
- [x] Детекция и фильтрация зацикливаний WER 0.192
    - Зацикливание на звуке
    - Зацикливание на фразе
    - Улучшение WER до 0.192

Модели: Whisper.
Борьба с галлюцинациями

- [x] Whisper из коробки WER 0.55
- [x] VAD Voice Activity Detection WER 0.21
- [x] Фильтрация «стоп-слов» WER 0.20
- [x] Детекция и фильтрация зацикливаний WER 0.192
- [x] Переобработка мест зацикливаний WER 0.18
    - Определение момента начала зацикливания
    - Обработка заново с момента начала зацикливаний

Модели: Whisper.
Борьба с галлюцинациями

- Whisper из коробки WER 0.55
- VAD Voice Activity Detection WER 0.21
- Фильтрация «стоп-слов» WER 0.20
- Детекция и фильтрация зацикливаний WER 0.192
- Переобработка мест зацикливаний WER 0.18

Модели: Whisper.
Детали реализации

- [x] Предобработка аудио
    - Динамическое выравнивание уровня громкости

Модели: Whisper.
Детали реализации

- [x] Предобработка аудио
- [x] ~~Применение фильтров для подавления шума~~
  - Фильтры шумоподавления в среднем ухудшают качество
  - Фильтры для выделения частот голоса не докидывают качество

Модели: Whisper.
Детали реализации

- [x] Предобработка аудио
- [x] ~~Применение фильтров для подавления шума~~
- [x] Обработка 100% аудио
    - VAD не всегда распознает речь
    - Фильтрация фрагментов без речи с «особым пристрастием»

Модели: Whisper.
Детали реализации

- [x] Предобработка аудио
- [x] ~~Применение фильтров для подавления шума~~
- [x] Обработка 100% аудио
- [x] Манипуляции с промптом для борьбы с галлюцинациями
  - Нулевой прирост качества от использования промпта
  - Промпт провоцирует галлюцинации
  - Использование в промпте набора символов для легкой детекции галлюцинаций. Пример «-,-,»

05

Модели

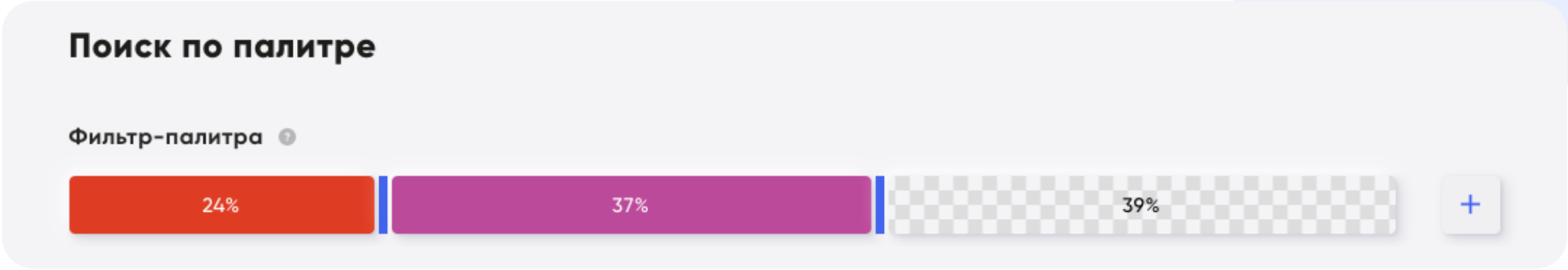
Поиск
по палитре

Модели: поиск по палитре

- Модель без применения нейронок
  - все расчеты в numpy
  - ускорение – numba

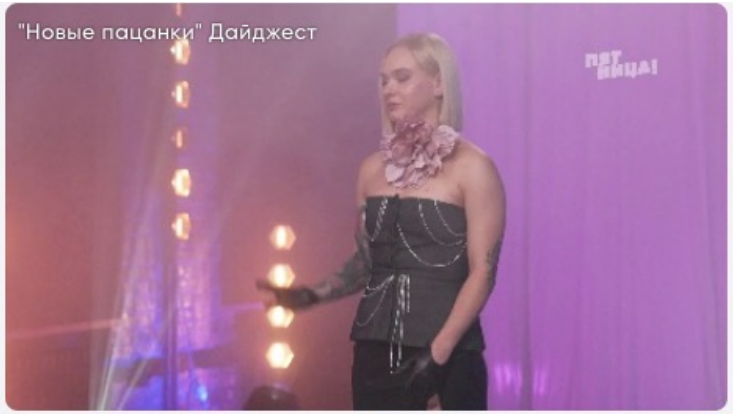
Модели: поиск по палитре

- [x] Модель без применения нейронок
- [x] Расчет эмбеддинга
  - Сокращение палитры до 64 цветов (4 градации RGB)
  - Подсчет статистики по каждому из цветов
  - Нормализация



Модели: поиск по палитре

- [x] Модель без применения нейронок
- [x] Расчет эмбеддинга
- [x] Поиск по косинусному сходству векторов



06

Модели

Прочие
модели

Модели: прочие

- Определение плана – Resnet50
- Детекция лиц RetinaFace
- Эмбеддинги лиц arcface
- Детекция сцен TransnetV2
- Детекция текста EasyOCR
- Детекция объектов RT-DETR
- Детекция действий mmaction2

07

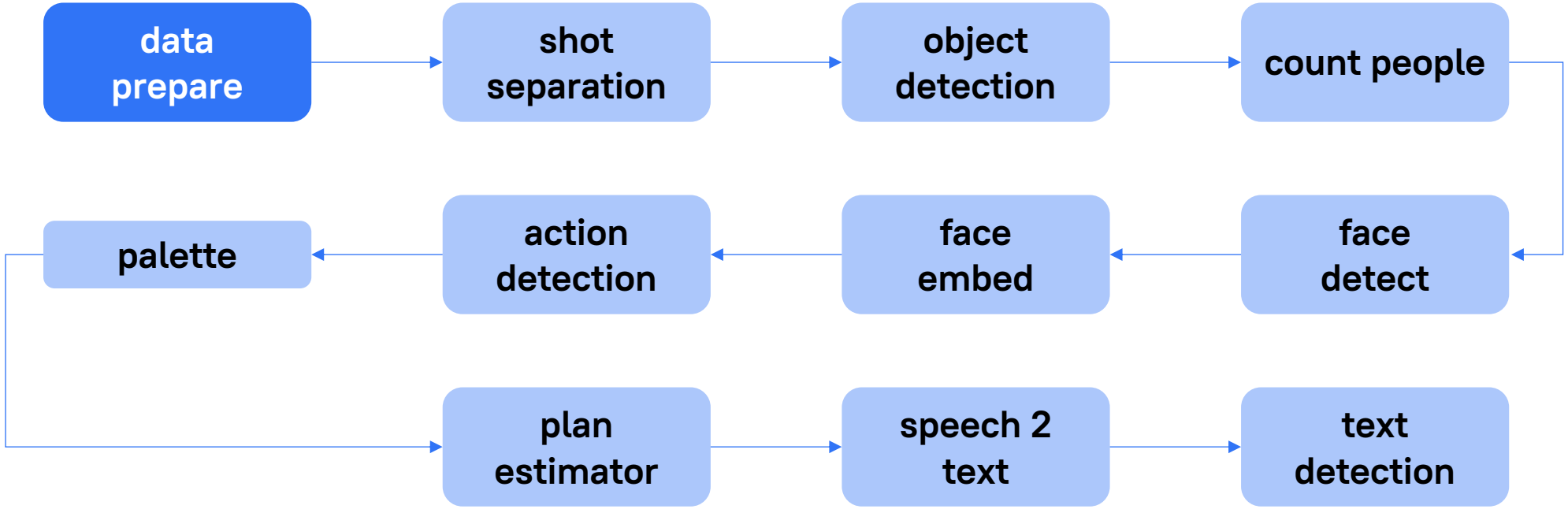
Пайплайн обработки

Ускоряем работу

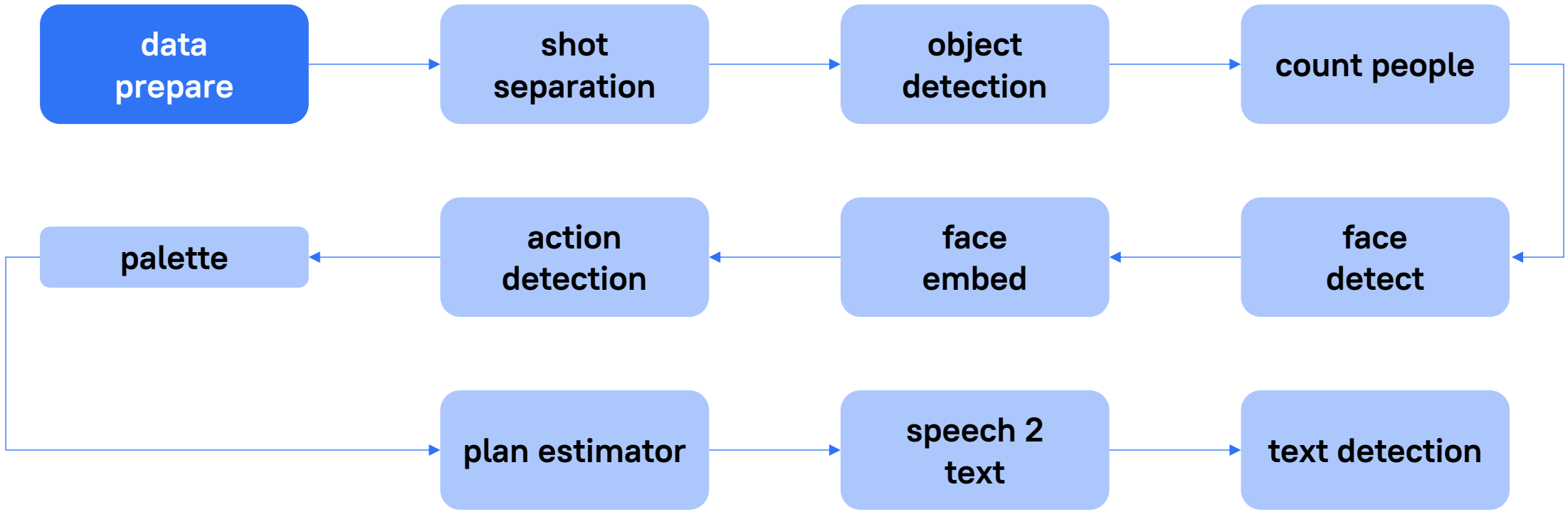
Пайплайн обработки

Первая версия

- [x] 10 стейджей
- [x] 1 час видео = 18 часов



Пайплайн обработки



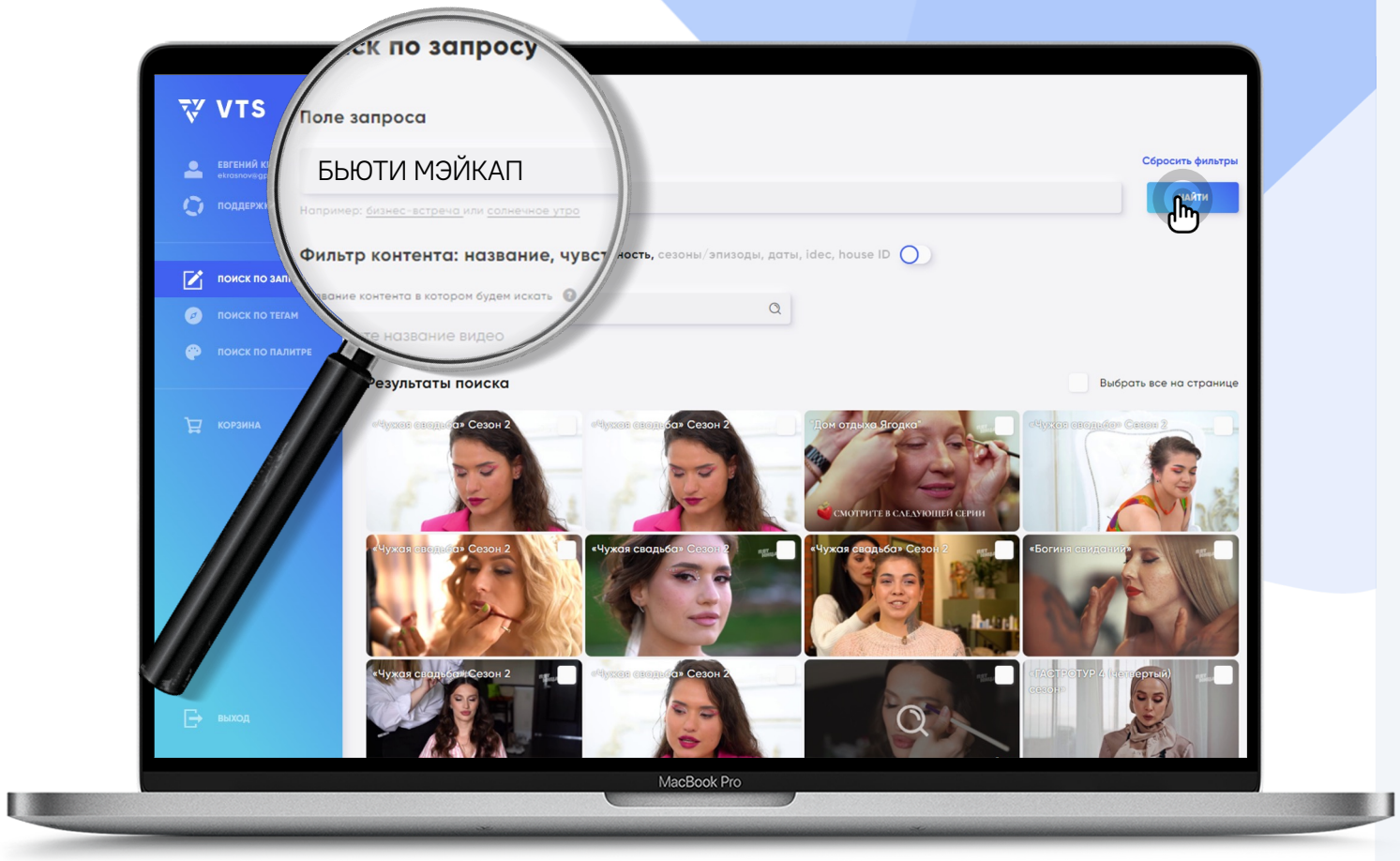
Пайплайн обработки: ускоряем обработку

- [x] Нормализация видео
  - 25 кадров в секунду
  - Кодек H.264
  - Извлечение аудиодорожки

ffmpeg

Пайплайн обработки: ускоряем обработку

- [x] Нормализация видео
- [x] Картинки – preview
  - Быстрое отображение результатов
  - Промежуточная фича
  - Эффект 15 минут → 40 сек



Пайплайн обработки: ускоряем обработку

- [x] Нормализация видео
- [x] Картинки – preview
- [x] Использование TensorRT

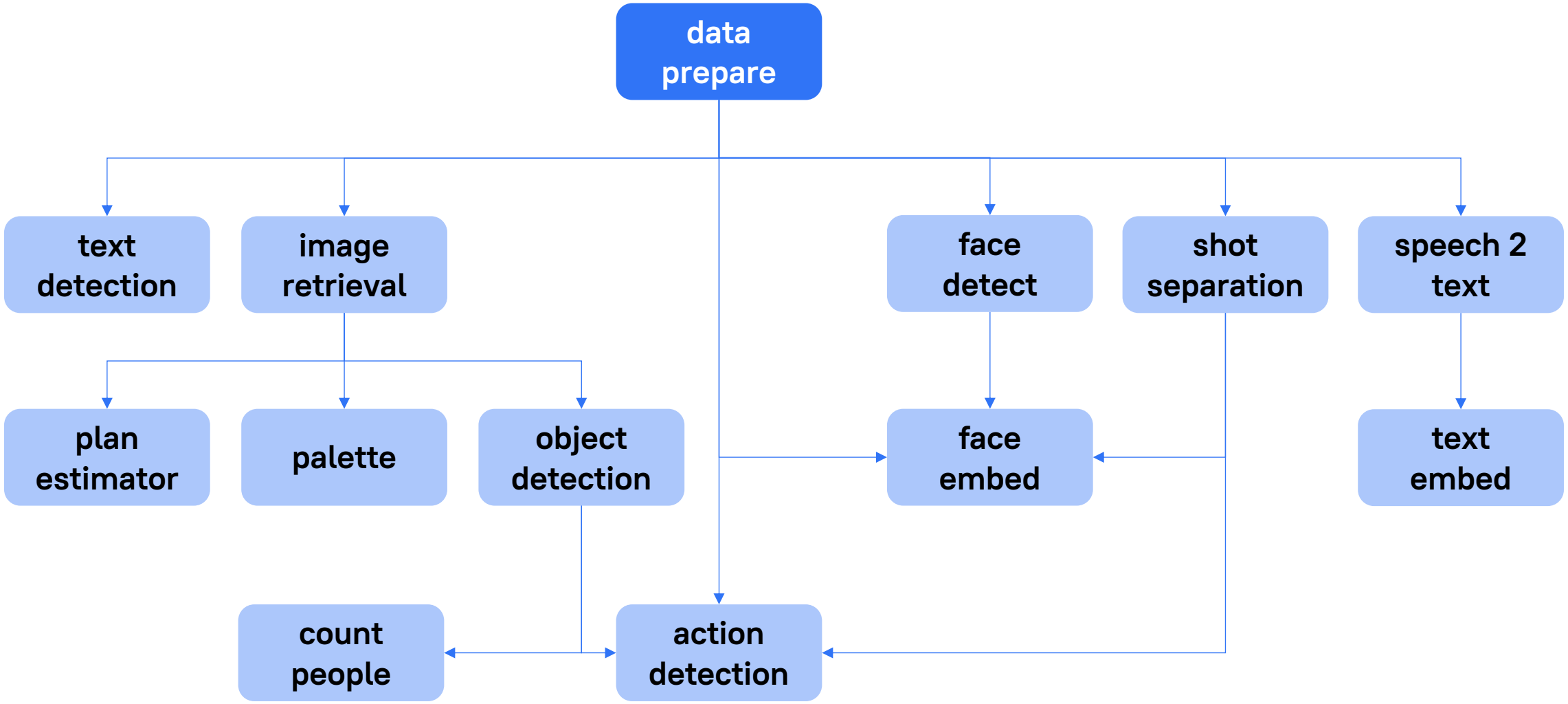
Пайплайн обработки: ускоряем обработку

- [x] Нормализация видео
- [x] Картинки – preview
- [x] Использование TensorRT
- [x] Detection → embedding
  - Face embedding
  - Text embedding

Пайплайн обработки: ускоряем обработку

- [x] Нормализация видео
- [x] Картинки – preview
- [x] Использование TensorRT
- [x] Detection → embedding
- [x] Параллельная обработка

Пайплайн обработки



Пайплайн обработки

13 стейджей

с 18 часов
до 45 минут –
общее время

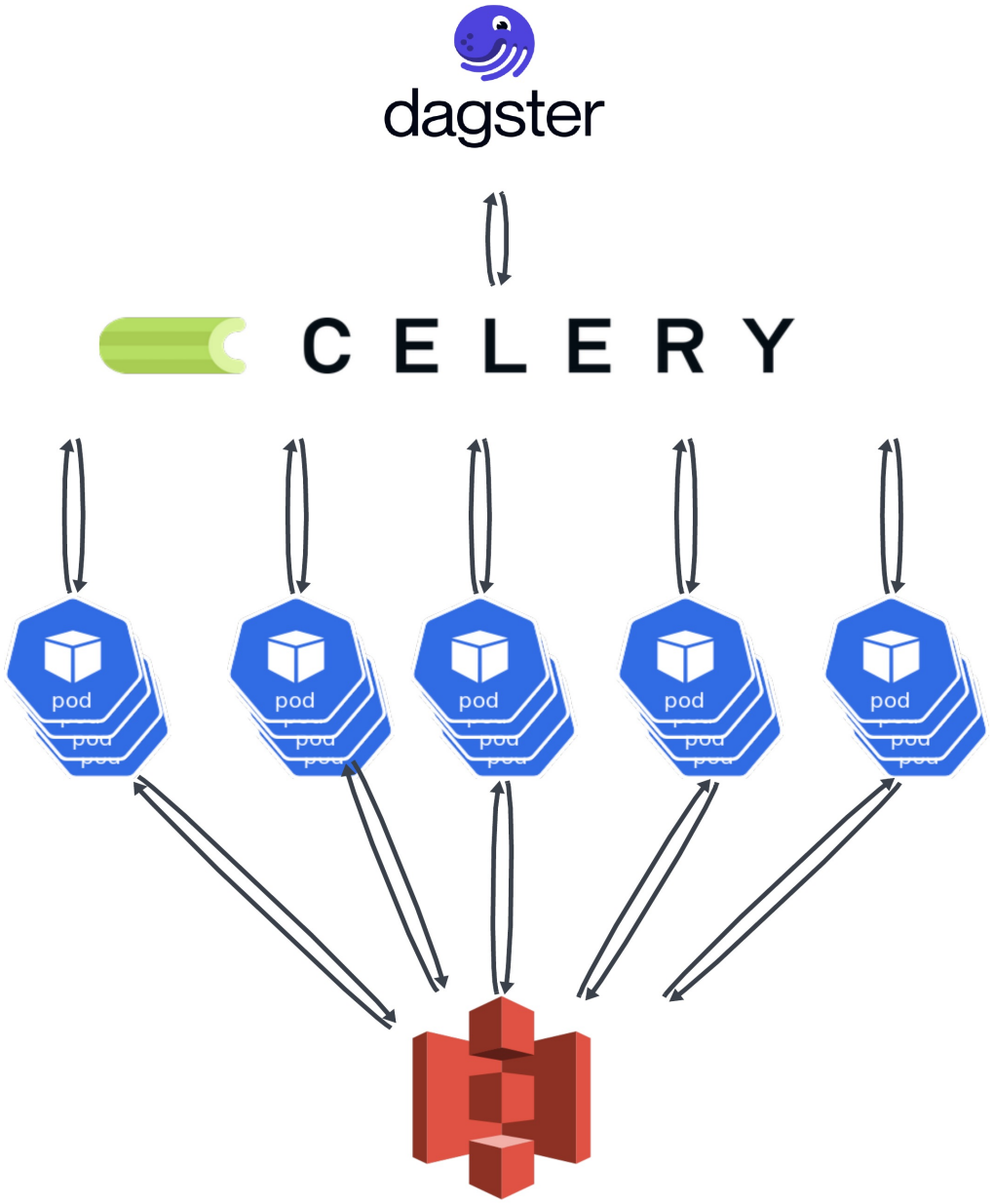
с 10 минут
до 45 сек –
время стейджей

08

MLOps

MLOps

- Приложения развернуты в K8S
- Балансировка количеством POD
- Версионирование образов и алгоритмов
- Оркестратор Dagster+Celery
- 1 стейдж == 1 микросервис
- Хранение артефактов S3



09

Кейсы у заказчиков

Кейс: создание промо-контента

Промоконтент: проморолики, тизеры, трейлеры, шортс и видеопрезентации

x20

сокращение времени подбора материалов

x3

увеличение количества вариантов ролика

Клиент

ПЯТНИЦА!

Кейс: Расстановка рекламных пауз

Для онлайна (рекламные мил-роллы) и для эфира (рекламные спецпроекты)

x3

увеличение рекламных кампаний

Было:

до 3-х недель отсмотра видео

Клиент

Кейс: субтитрование контента

x5

сокращение времени на создание субтитров

- **Закон:** 5% эфира с субтитрами
- **Было:** 60 минут – 8 часов работы
- **Стало:** 60 минут – 1,5 часа работы

Клиент

PREMIER

ГИД

Спасибо за внимание

Соколов Александр,
tg:@AESokolov
«ГИД Дата» Инженер машинного обучения