Техническая архитектура Сервиса инференса MarQus

Сервис инференса нейросетевых моделей MarQus для автоматизации процессов распознавания и сортировки различных типов ТБО.

2023

Содержание

Аннотация

Документ описывает архитектуру программного обеспечения как комплекс взаимоувязанных решений по основополагающим принципам выбора технологий для создания сервиса инференса моделей «Marqus», интеграции сервиса с модулями заказчика, а также требований к необходимым для разработки и функционирования этих интеграций техническим средствам и иным видам обеспечения.

1. Основные положения

Сервис служит для автоматизации процессов по распознаванию и сортировке различных типов ТБО при помощи технологий машинного зрения и нейронных сетей. Основное назначение сервиса заключается в обработке входных данных видеопотока, поступающих с различных источников технологической линии сортировки ТБО (видеокамер) и передаче нужной информации в управляющую физическими манипуляторами систему данной линии с целью распределения и сортировки ТБО.

2. Структура и логика работы {#2}

Сервис представляет собой набор из нескольких модулей, каждый из которых выполняет свою функцию:

- Очередь запросов на распознавание - принимает фреймы изображений, передаваемых с камер по протоколу TCP с помощью ZMQ.
- Модуль приема и предобработки – забирает из очереди запросов на распознавание фреймы видеопотока и производит первичную обработку, пропуск лишних, не значащих кадров, нормализацию изображений.
- Нейромодуль детекции и сегментации – выполняет задачу определения позиции предмета на конвейерной ленте, определяет контуры предметов, производит обрезку изображения с целью дальнейшей классификации по типу выявленной категории ТБО.
- Нейромодуль классификации по спектру – выполняет задачу распознавания и присваивания типов в разрезе выявленных категорий ТБО с помощью анализа спектра.
- Нейромодуль классификации по форме - выполняет задачу распознавания и присваивания типов в разрезе выявленных категорий ТБО с помощью анализа формы предметов.
- Модуль трекинга – собирает информацию от нейромодулей и передает в управляющую систему технологической линии координаты выделенных из общего

потока данных объектов ТБО и их метаданные для дальнейших физических манипуляций над ними.

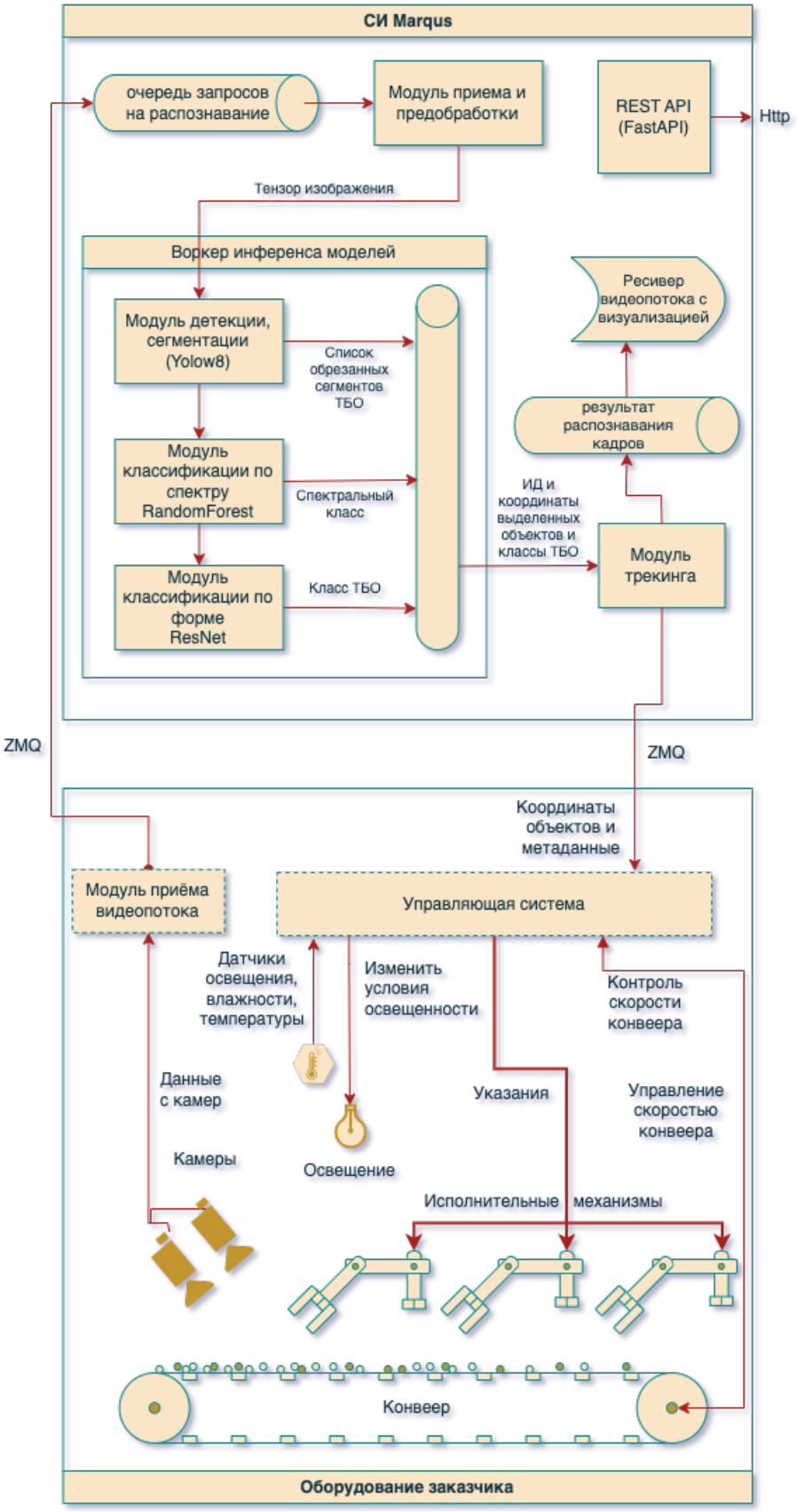
- Ресивер видеопотока с визуализацией (вспомогательный модуль) - выполняет задачу предоставления видеопотока с наложенной информацией детекции, сегментации, классификации и треккинга для визуального контроля работоспособности сервиса.
- Модуль REST API - служит для предостовления API для мониторига, контроля и тестирования сервиса.

Решение представлено в виде набора микросервисов, реализованных и запускаемых в платформе docker.

2. Общая логика работы сервиса

Поступающий на вход видеопоток данных с камер трансформируется в набор тензоров и подаётся на вход нескольких нейросетевых моделей, выполняющих каждая свою функцию по распознованию. Передача происходит через очереди, что взаимоисключает блокировку обработки видеопотока в целом, в случае если для обработки полного потока не хватает вычислительных ресурсов. Логика устроена так, что кадры пропускаются, что позволяет выдерживать необходимую скорость обработки в целом.

Сервис производит логирование информации, что позволяет производить анализ ситуации для последующей оптимизации кода сервиса и построения рекомендаций по увеличению вычислительных ресурсов, либо уменьшению потока ТБО на ленте конвейера.


СИ Marqus
очередь запросов на распознавание
Модуль приема и предобработки
REST API (FastAPI)
Http
Тензор изображения
Воркер инференса моделей
Модуль детекции, сегментации (Yolow8)
Список обрезанных сегментов ТБО
Модуль классификации по спектру RandomForest
Спектральный класс
Модуль классификации по форме ResNet
Класс ТБО
Ресивер видеопотока с визуализацией
результат распознавания кадров
ИД и координаты выделенных объектов и классы ТБО
Модуль трекинга
ZMQ
ZMQ
Координаты объектов и метаданные
Модуль приёма видеопотока
Управляющая система
Датчики освещения, влажности, температуры
Изменить условия освещенности
Контроль скорости конвеера
Данные с камер
Указания
Управление скоростью конвеера
Камеры
Освещение
Исполнительные механизмы
Конвеер
Оборудование заказчика

3. Нейромодули и нейросетевые модели

Нейромодули детекции и сегментации и классификации по форме для инференса моделей используют библиотеку torch, которая показывает высокую производительность при инференсе моделей на GPU.

Нейромодуль классификации по спектру для построения классических моделей машинного обучения использует библиотеку sklearn.

3.1. Нейромодуль классификации по спектру

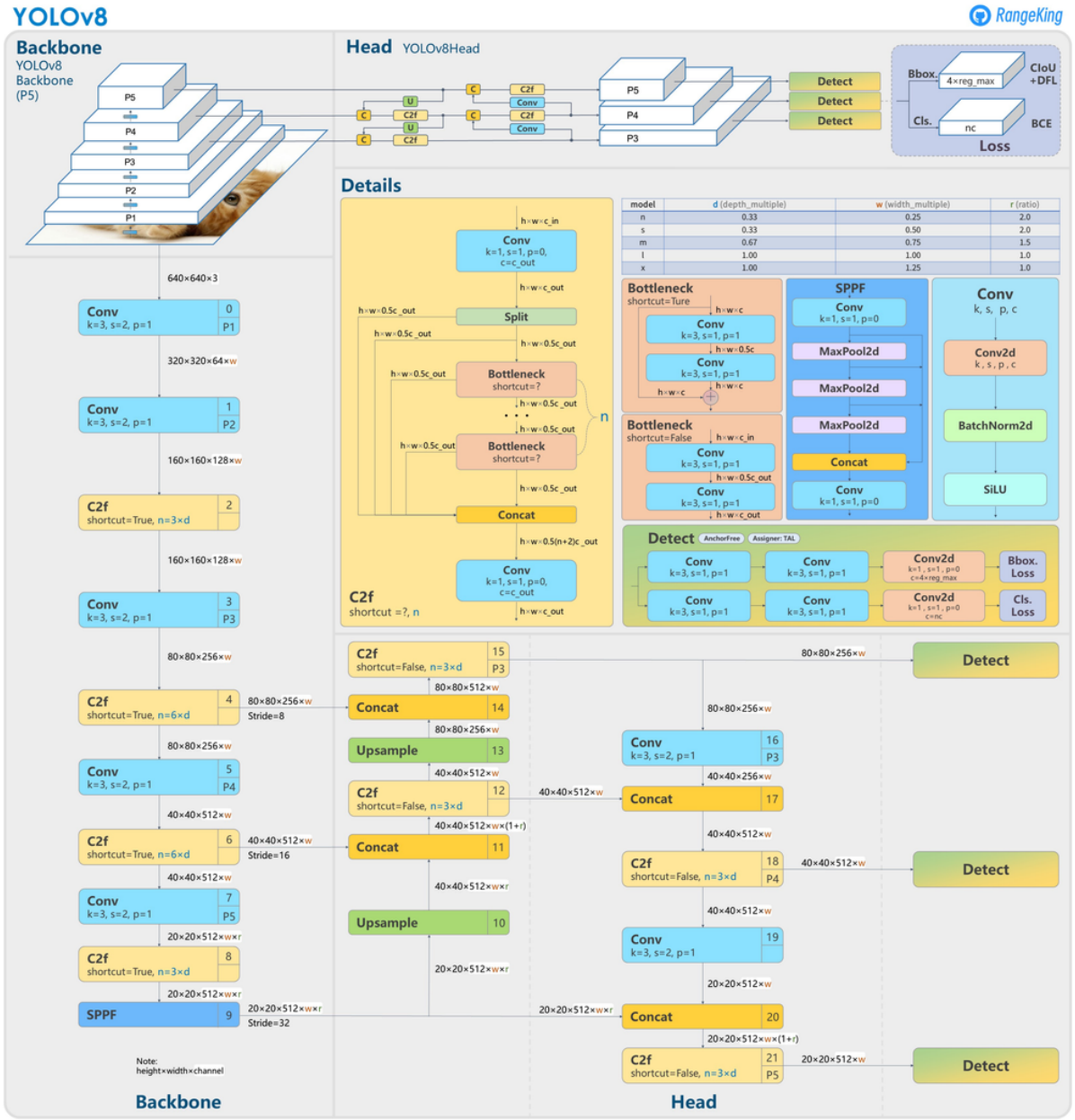
Для классификации на основе информации о спектре выбрана классическая модель машинного обучения - метод случайного леса (Random forest).

Более подробно работу алгоритма можно изучить по ссылке (6)

3.2. Нейромодуль детекции и сегментации

В качестве модели для детекции (сегментации) была выбрана YOLOv8, а непосредственно дообучение происходило на YOLOv8x-seg.

Архитекрура YOLOv8 представлена на изображении.

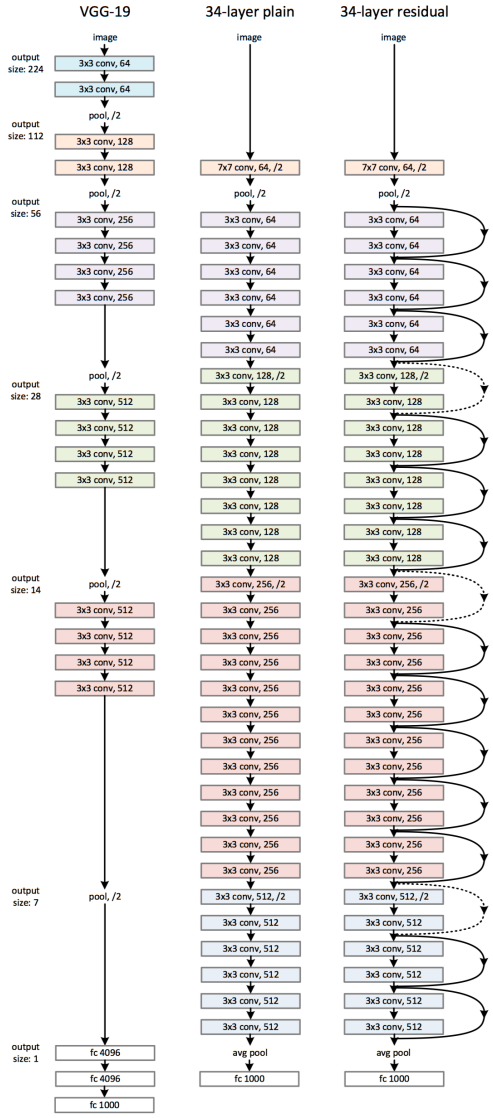


Более подробно устройсво YOLOv8 можно изучить по ссылке (5).

3.3. Нейромодуль классификации по форме

Для классификации основных типов ТБО и классификации по форме объектов выбрана и обучена сверточная нейронная сеть ResNet50.

Архитекрура ResNet50 представлена на изображении.



Более подробно устройсво ResNet50 можно изучить по ссылке (7).

4. Вспомогательные модули

4.1. Модуль трекинга

Модуль треккинга использует алгоритм BoT-SORT. Алгоритм обеспечивает обнаружение и отслеживание всех объектов в сцене с сохранением уникального идентификатора для каждого объекта.

Более подробно работу алгоритма можно изучить по ссылке (8)

4.2. Модуль REST API

Модуль построен на базе библиотеки FastAPI. Помимо реализации REST методов предоставляет OpenAPI спецификацию, позволяющую проводить генерирование кода клиента для API для более быстрой интеграции с сервисом.

4.3. Ресивер видеопотока с визуализацией

Модуль осуществляет трансляцию видеопотока с визуализацией результатов распознавания для осуществления визуального контроля для анализа качества работы сервиса. Сервис транслирует результат распознавания в виде image/jpeg фреймов.

5. ZMQ

В сервисе используется библиотека imageZMQ для приема потока изображений OpenCV с помощью обмена сообщениями PyZMQ.

Более подробно про захват и передачу изображений можно изучить по ссылке (2).

6. Docker

Для упаковки решения используется Docker — программное обеспечение для автоматизации развёртывания и управления приложениями в средах с поддержкой контейнеризации.

7. Ссылки

1. Протокол передачи сообщений ZeroMQ
2. Использование imageZMQ в проектах распределенного компьютерного зрения
3. YOLOv8 Полное руководство
4. Метод случайного леса (random forest)
5. «Остаточные» CNN для классификации изображений
6. Алгоритм BoT-SORT

Перечень терминов и сокращений

Термин/сокращение	
Сервис	Сервис инференса нейросетевых моделей для автоматизации процессов распознавания и сортировки различных типов ТБО.
API	Программный интерфейс приложения, интерфейс прикладного программирования (англ. application programming interface) – описание способов (набор классов, процедур, функций, структур или констант), которыми одна компьютерная программа может взаимодействовать с другой программой. Используется программистами при написании всевозможных приложений.
REST	(англ. Representational State Transfer — «передача репрезентативного состояния» или «передача „самоописываемого“ состояния») — архитектурный стиль взаимодействия компонентов распределённого приложения в сети.
Docker	ПО с открытым исходным кодом для автоматизации развёртывания и управления Программы с использованием технологии контейнеризации. В состав «Docker» включен пакетный менеджер «Docker Compose», обеспечивающий запуск многоконтейнерных приложений.
JSON	(англ. JavaScript Object Notation) – текстовый формат обмена данными, основанный на JavaScript.
OpenSource	Открытое программное обеспечение (англ. open-source software) – программное обеспечение с открытым исходным кодом. Исходный код таких программ доступен для просмотра, изучения и изменения, что позволяет убедиться в отсутствии уязвимостей и неприемлемых для пользователя функций, принять участие в доработке самой открытой программы, использовать код для создания новых программ и исправления в них ошибок – через заимствование исходного кода, если это позволяет совместимость лицензий, или через изучение использованных алгоритмов, структур данных, технологий, методик и интерфейсов (поскольку исходный код может существенно дополнять документацию, а при отсутствии таковой – сам служит документацией).
TCP	Сетевая модель передачи данных, представленных в цифровом виде. Модель описывает способ передачи данных от источника информации к получателю. В модели предполагается прохождение информации через четыре уровня, каждый из которых описывается правилом (протоколом передачи). Наборы правил, решающих задачу по передаче данных, составляют стек протоколов передачи данных, на которых базируется Интернет. Название TCP/IP происходит из двух основных протоколов семейства – Transmission Control Protocol (TCP) и Internet Protocol (IP), которые были первыми разработаны и описаны в данном стандарте.
ZeroMQ , ZMQ	Высокопроизводительная асинхронная библиотека обмена сообщениями, ориентированная на использование в распределённых и параллельных вычислениях. Библиотека реализует очередь сообщений, которая может функционировать без выделенного брокера сообщений.
ТБО	Твёрдые бытовые отходы.