



Направление Data Scientist, компания ЕВРАЗ

Добро пожаловать на виртуальную стажировку компании ЕВРАЗ!

Предлагаем тебе прокачать свои навыки в Data Science и применить методы машинного обучения в области Computer Vision¹ как в теории, так и на практике.

В рамках проекта ты должен составить алгоритм превращения видео в дата-сет, а также научить модель правильно распознавать рукописные цифры.

Будь готов, что на выполнение задач потребуется 2–3 часа твоего времени. Однако все зависит от уровня владения Python, а также наличия необходимого программного обеспечения (ПО) на компьютере.

Во время работы над проектом ты научишься:

1. Строить алгоритм по превращению видео в дата-сет.
2. Обучать сверточную нейронную сеть (Convolutional Neural Network²) классифицировать рукописные цифры из MNIST дата-сета.

Рекомендуемый тайминг:

1. 20–30 минут на первое задание.
2. 60–80 минут на второе задание.

Желаем удачи!

¹ Computer Vision (компьютерное зрение) — область искусственного интеллекта, связанная с анализом изображений и видео.

² Convolutional Neural Network (CNN, сверточная нейронная сеть) — специальная архитектура искусственных нейронных сетей, нацеленная на эффективное распознавание образов, входит в состав технологий глубокого обучения.

Задание 1. Преврати видео в обучающий дата-сет

Коллектив IT-департамента холдинга ЕВРАЗ рад приветствовать тебя в роли специалиста по работе с данными. И в качестве первого задания ты должен помочь коллегам с пониманием того, как сформировать собственный обучающий дата-сет для проектов видеоаналитики³. Это приоритетное направление цифровой трансформации ЕВРАЗа.

Утром ты обнаружил на почте электронное письмо от руководителя с постановкой задания.

Привет!

Сейчас наш департамент активно помогает службе охраны труда и промышленной безопасности с внедрением системы видеоаналитики для контроля использования средств индивидуальной защиты (СИЗ) в опасных зонах. При этом для обучения системы на основе искусственного интеллекта требуются дата-сеты, полученные с камер видеонаблюдения. Сейчас точность работы цифровой сети достигает 95%, но для повышения точности нам нужно дообучить ее на еще нескольких дата-сетах.

Твоей задачей будет написание алгоритма для определения того, находится человек в опасной зоне или нет, и если да, то надел ли он СИЗ.

В ответном письме ждем одностраничный документ, где ты должен описать основные шаги по составлению подобного алгоритма. Обычно мы включаем в него формирование дата-сета, обучение модели, оценку точности.

Спасибо!

Полезные материалы

- Видео, демонстрирующее работу системы видеоаналитики: [Работа видеоаналитики определение наличия СИЗ – YouTube](#).

Формат конечного результата

Документ формата .docx с описанием алгоритма.

Форма загрузки результата

Пожалуйста, загрузи решение в формате zip-архива и включи в него свой документ.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.

³ Видеоаналитика — технология, использующая методы компьютерного зрения для автоматизированного получения различных данных на основании анализа последовательности изображений, поступающих с видеокамер в режиме реального времени или из архивных записей.

Задание 2. Обучи модель распознавать рукописные цифры

Твое погружение в работу Data Scientist идет полным ходом, и теперь настала пора переходить от теории к практике. Во втором электронном письме от руководителя ты найдешь детали нового задания.

Привет!

Ты отлично справляешься с ролью специалиста по работе с данными. Пришло время дать тебе практическую задачу.

Для лучшего понимания того, как работает система распознавания лиц или объектов в проектах видеоаналитики, предлагаю начать с анализа [MNIST](#) дата-сета. Этот набор данных содержит изображения цифр от нуля до девяти, написанных от руки разными людьми (60 тыс. примеров в обучающей и 10 тыс. – в тестовой выборках).

Для работы с MNIST дата-сетом удобно использовать библиотеку глубокого обучения [Keras](#). В этом случае тебе не нужно даже скачивать файлы MNIST, так как они уже встроены в библиотеку, и их нужно просто импортировать оттуда.

Наш Junior Data Scientist Александр уже начал писать программу для распознавания рукописных цифр с использованием Convolutional Neural Network на Keras, но сейчас он занят на другом важном проекте. Поэтому мы просим тебя подключиться к этому заданию и дополнить код, написанный Александром.

Саша оставил необходимые комментарии в коде о том, что нужно добавить, чтобы программа заработала:

- несколько слоев, включая flatten;
- функцию обучения модели;
- функцию проверки на отложенных данных;
- печать точности.

Также он просил, чтобы ты дополнил код несколькими строками для отображения графика, содержащего первые десять изображений рукописных цифр дата-сета (для этого используй Matplotlib и цикл for).

Надеемся на твою помощь и ждем обновленный код.

Спасибо!

Полезные материалы

- Файл с кодом от Александра.
- Статья о том, как установить Python на компьютер: [Скачать Python 3 – Установка Python 3 на Windows, Ubuntu, macOS \(python-scripts.com\)](#).
- Как установить Keras: [Установка Keras на Linux и Windows, обучение первой модели ~ PythonRu](#).

Форма загрузки результата



Пожалуйста, загрузи решение в формате zip-архива и включи в него свой документ.

Пример решения

У тебя будет возможность ознакомиться с примером решения задания после отправки своей версии.