

Платформа для поиска похожих кроссовок



Состав проекта

Куратор: Георгий Панчук

Распределение ролей участников:

Литвинов Вячеслав: парсинг, eda, модели

Моисеев Даниил: парсинг, исследование моделей, телеграм-бот



Цели и задачи проекта

- Сбор данных (изображения кроссовок) с различных источников
- Обработка, чистка и объединение данных
- Разведочный анализ данных
- Получение эмбеддингов для картинок с помощью различных алгоритмов
- Исследование качества моделей классификации и поиска похожих изображений
- Обертка этих моделей в сервис и телеграмм-бот



Структура проекта

Описание	Путь в проекте
Модуль для парсинга, чистки, объединения данных и создания датасета	sneakers_ml/data
Модуль для получения эмбеддингов картинок с помощью различных методов (SIFT, HOG, ResNet152)	sneakers_ml/features
Модуль для тренировки и применения моделей	sneakers_ml/models
Модуль для telegram-бота	sneakers_ml/bot
Модуль для рантайм-хостинга моделей (API-сервис + Streamlit приложение)	sneakers_ml/app

Датасеты и парсинг

- Спарсили 5 интернет-магазинов кроссовок
- Парсили скрапингом и используя публичные API



POLO RALPH LAUREN
Masters Sport "White / Blue"
€149,95



MERRELL
Moab Speed GTX "Black"
€169,95



MERRELL
Agility Peak 5 "Black"
€159,95



MERRELL
Moab Speed GTX "White"
€169,95



	brand	title	pricecurrency	price	images_path
1	Nike	Blazer Mid Next Nature "White"	EUR	49.99	data/raw/images/sneakerbaas/category-kids/nike/b...
2	New Balance	550 PS "Lavender"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...
3	New Balance	550 "astro dust"	EUR	79.99	data/raw/images/sneakerbaas/category-kids/new-ba...
4	Reebok Classics	CLUB C CARDI "Blue"	EUR	34.99	data/raw/images/sneakerbaas/category-kids/reebok...
5	Nike	Air Max TW "University Red"	EUR	59.95	data/raw/images/sneakerbaas/category-kids/nike/a...
6	New Balance	2002 "TIMBERWOLF"	EUR	94.95	data/raw/images/sneakerbaas/category-kids/new-ba...
7	Nike	Nike Blazer '77 "Jade Ice"	EUR	69.95	data/raw/images/sneakerbaas/category-kids/nike/n...
8	Reebok Classics	Cardi B Club C "Red"	EUR	34.99	data/raw/images/sneakerbaas/category-kids/reebok...
9	Karhu	Albatross 82 "Navy"	EUR	29.99	data/raw/images/sneakerbaas/category-kids/karhu/...
10	Nike	Huarache Run GS "Jade Ice"	EUR	89.95	data/raw/images/sneakerbaas/category-kids/nike/h...
11	New Balance	550 PS "Orange Sea"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...
12	New Balance	550 PS "Marina"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...
13	New Balance	550 "Lavender"	EUR	89.95	data/raw/images/sneakerbaas/category-kids/new-ba...
14	New Balance	2002 "ECLIPSE"	EUR	94.95	data/raw/images/sneakerbaas/category-kids/new-ba...
15	New Balance	2002 "BLACK"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...
16	Jordan	Air Jordan 6 Retro GS "Dutch Blue"	EUR	99.99	data/raw/images/sneakerbaas/category-kids/jordan...
17	New Balance	2002 "SLATE GREY"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...
18	Reebok Classics	CLUB C CARDI "Blue"	EUR	49.99	data/raw/images/sneakerbaas/category-kids/reebok...
19	Vans	Old Skool "Powder Pink"	EUR	39.99	data/raw/images/sneakerbaas/category-kids/vans/o...
20	New Balance	550 "WHITE"	EUR	79.99	data/raw/images/sneakerbaas/category-kids/new-ba...
21	Nike	Air Max Motif "WHITE"	EUR	59.99	data/raw/images/sneakerbaas/category-kids/nike/a...
22	Reebok Classics	CLUB C CARDI "Purple"	EUR	24.99	data/raw/images/sneakerbaas/category-kids/reebok...
23	New Balance	2002 "ECLIPSE"	EUR	99.95	data/raw/images/sneakerbaas/category-kids/new-ba...

Основные столбцы полученных метаданных



Предобработка и объединение данных

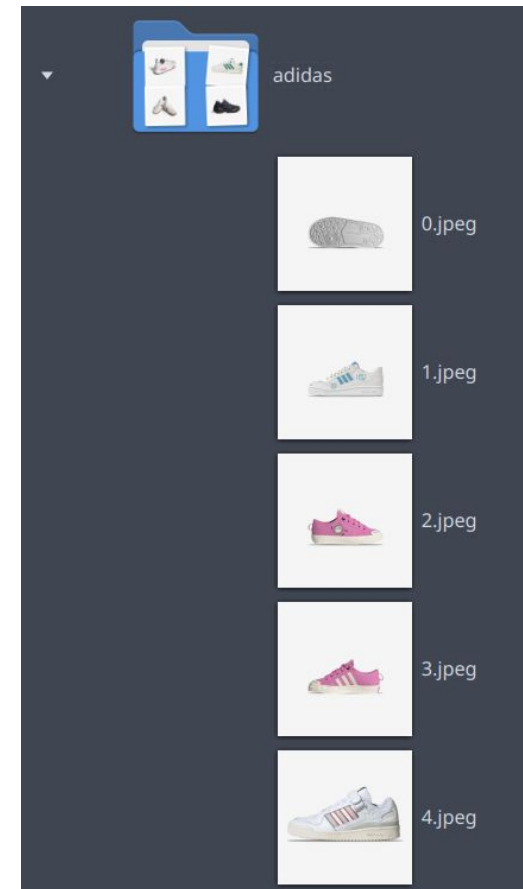
- Для каждого датасета провели базовую предобработку
 - Колонки: удалили ненужные, привели к одному названию
 - Строки: удалили дубликаты, почистили от лишних символов
- Привели название и бренд к одному формату
- После объединения получили два датасета

Объединили по моделям: 2067 моделей кроссовок объединились в 877

Объединили по брендам: 30 брендов в итоге

Предобработка изображений

- Удалили идентичные картинки
- Перевели в палитру RGB
- Привели к одному формату .jpeg

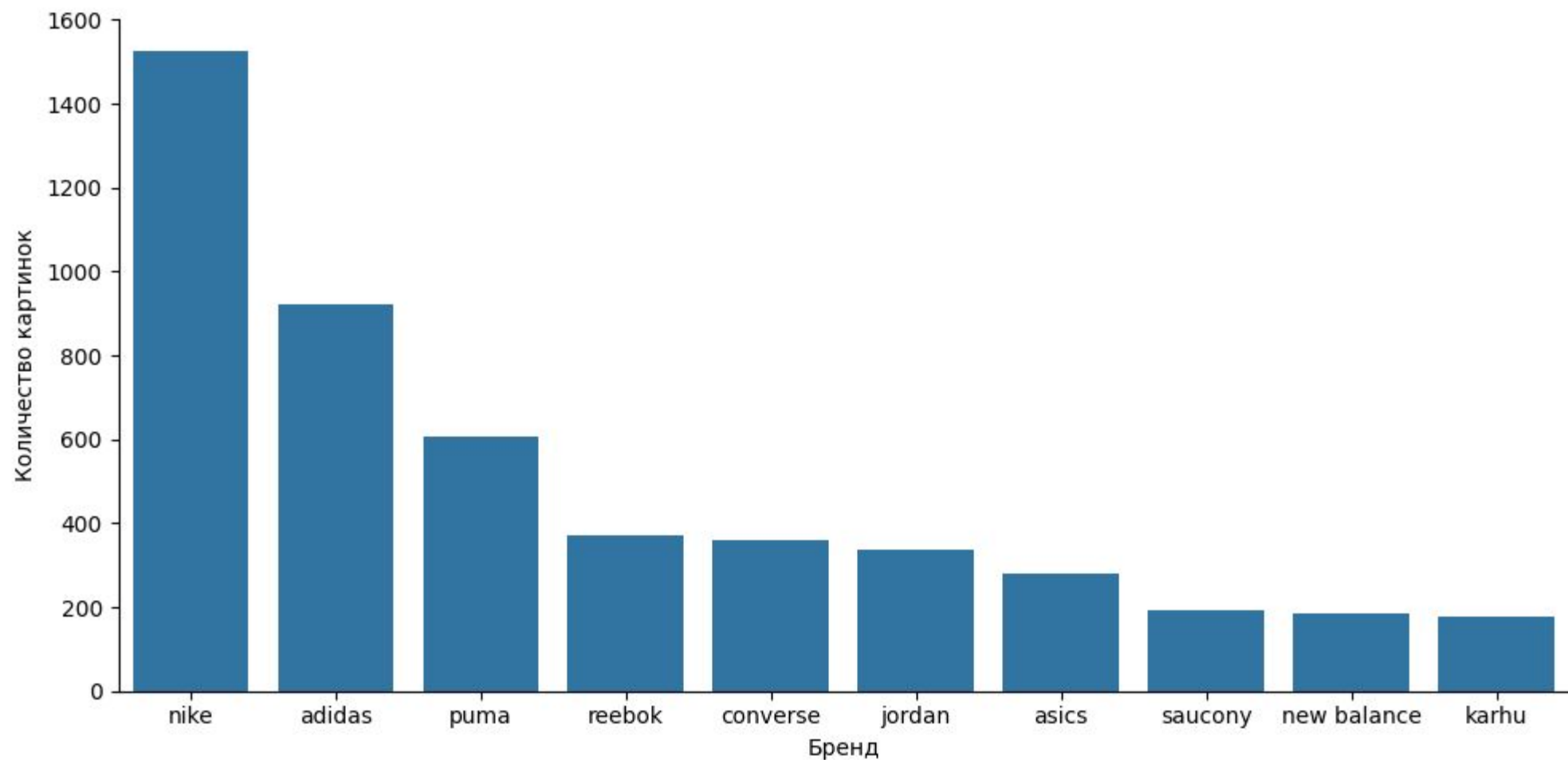




EDA

- Вывели как выглядят метаданные и их основные статистики
- Убедились, что в данных нету пропусков
- Убедились, что все картинки в одинаковом формате
- Посмотрели размеры картинок - они различные, от 400x400 до 2000x2000
- Общее количество картинок - 5852

Классификация брендов

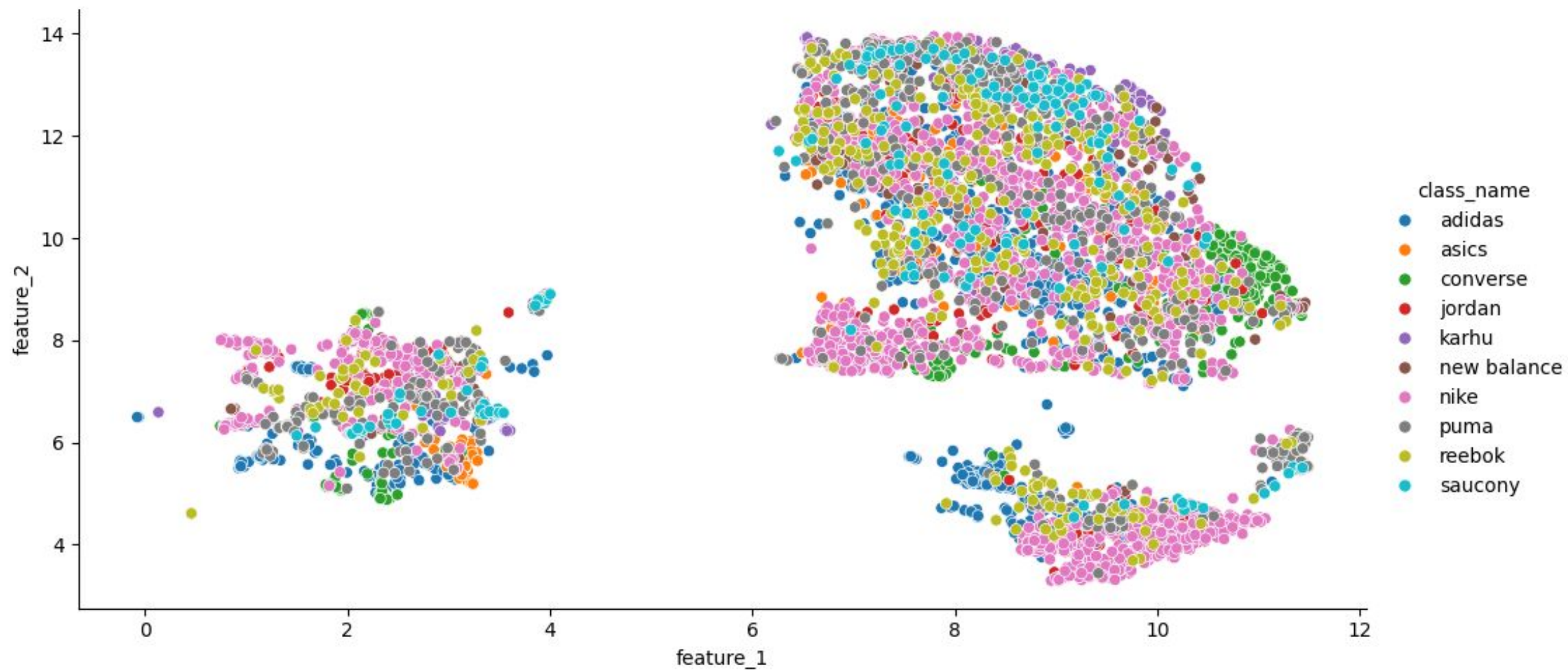


Зависимость числа картинок от бренда - наблюдается дисбаланс классов

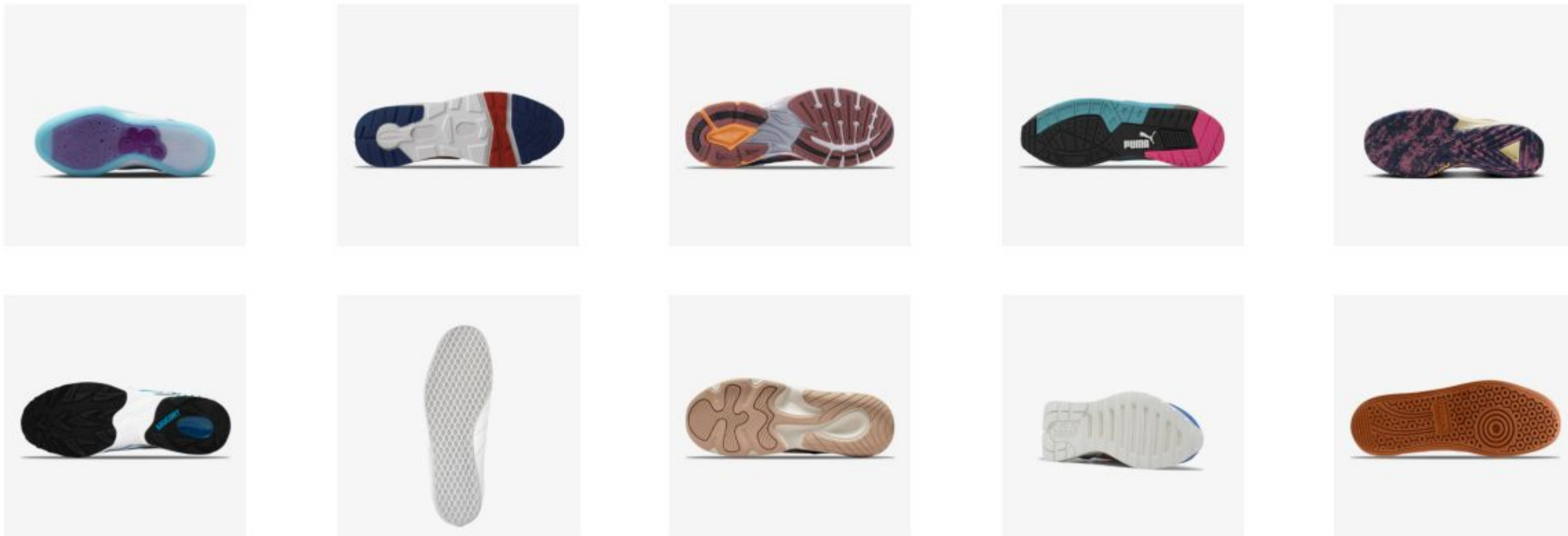


Features

- Поделили датасет на train/val/test в пропорциях 60/20/20
- Отобрали только те бренды, у которых больше 100 картинок
- Получилось 13 брендов для классификации
- Для всех картинок взяли эмбединги с помощью ResNet152, HOG, SIFT
- Для визуализации эмбедингов использовали SVD, T-SNE, PCA, UMAP



Визуализация ResNet features с помощью UMAP. Картинки кластеризуются по ракурсам изображений



Визуализация ResNet features с помощью UMAP - левый кластер



Метрики и классические ML-модели

- По результатам EDA решили попробовать удалить картинки из кластера “**подошвы**”, так как по подошве вряд ли можно определить бренд
- Считали метрику F1-weighted и F1-macro
- Baseline модель предсказывает самый частый класс
- В ML-части использовали модели SVM, SGD, CatBoost, LogReg, DecisionTree, RandomForest над эмбедингами ResNet152, HOG, SIFT
- Перебирали параметры по сетке с кросс-валидацией
- Лучшее качество среди ML-моделей показывает hog-svm
- По DL части зафайнтюнили целиком ResNet152 - она работает лучше всех



Результаты ML-моделей

Датасет со всеми картинками

Feature-model_name	f1-macro	f1-weighted
baseline	0.03	0.13
hog-log_reg	0.72	0.76
hog-random_forest	0.54	0.63
hog-decision_tree	0.34	0.45
hog-svm	0.78	0.81
hog-sgd	0.72	0.75
hog-catboost	0.67	0.71
sift-log_reg	0.29	0.36
sift-random_forest	0.18	0.29
sift-decision_tree	0.17	0.24
sift-svm	0.37	0.44
sift-sgd	0.26	0.34
sift-catboost	0.29	0.39
resnet152-log_reg	0.71	0.73
resnet152-random_forest	0.41	0.51
resnet152-decision_tree	0.27	0.36
resnet152-svm	0.76	0.76
resnet152-sgd	0.72	0.74
resnet152-catboost	0.64	0.68
[DL] resnet152-finetune	0.84	0.85

Датасет с удаленными подошвами

Feature-model_name	f1-macro	f1-weighted
baseline	0.03	0.13
hog-log_reg	0.73	0.77
hog-random_forest	0.51	0.62
hog-decision_tree	0.34	0.44
hog-svm	0.79	0.82
hog-sgd	0.71	0.75
hog-catboost	0.66	0.71
sift-log_reg	0.28	0.37
sift-random_forest	0.15	0.3
sift-decision_tree	0.16	0.25
sift-svm	0.36	0.43
sift-sgd	0.25	0.36
sift-catboost	0.32	0.43
resnet152-log_reg	0.74	0.75
resnet152-random_forest	0.4	0.51
resnet152-decision_tree	0.28	0.36
resnet152-svm	0.74	0.75
resnet152-sgd	0.67	0.7
resnet152-catboost	0.64	0.69
[DL] resnet152-finetune	0.83	0.84



Структура проекта

- Данные, эмбединги и модели храним с помощью DVC на S3
- Добавили различные линтеры и форматтеры
- В проекте настроили poetry и pre-commit
- Модели храним в onnx формате, pickle не используем
- Настроили конфигурацию с помощью Hydra
- Реализовали телеграмм-бот, CI/CD Docker-образа в Registry
- Документация проекта в .md файлах в /notes
- Основной код в .py файлах, ноутбуки используем только как черновик



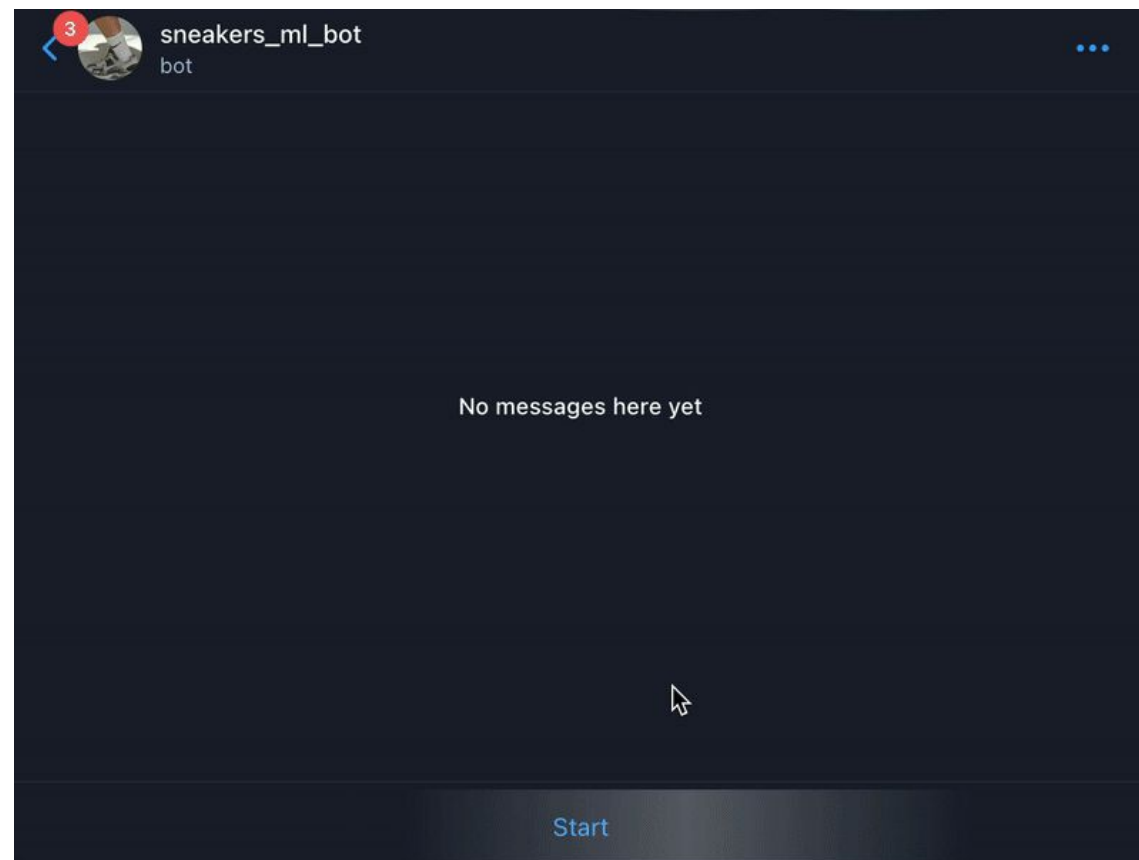


Прогресс и задачи

- На текущий момент обучили лучшую модель классификации Resnet152 в черновике-ноутбуке [notebooks/models/resnet_fine_tune.ipynb](#)
- Планируем раскидать код ноутбука по файлам, возможно использовать PyTorch Lightning
- Хотим попробовать Vision Transformer
- Далее будем решать другие задачи, такие как image2image поиск, similarity learning
- Улучшение telegram-бота, интеграция API с Streamlit
- Сборка датасета из изображений от пользователей через бота
- Логирование, мониторинг и алерты production-окружений (streamlit, tg-bot)



Демонстрация работы бота





Ссылки

- Репозиторий: <https://github.com/miem-refugees/sneakers-ml>
- Телеграм-бот: https://t.me/sneakers_ml_bot