

Deepfake Classification

Команда 61

Тимонин Андрей Сергеевич

Рябков Иван Юрьевич

Куратор

Блуменау Марк Ильич

Постановка задачи





Требуется реализовать приложение для классификации фейковых изображений*.

Цели по проекту на первое полугодие:

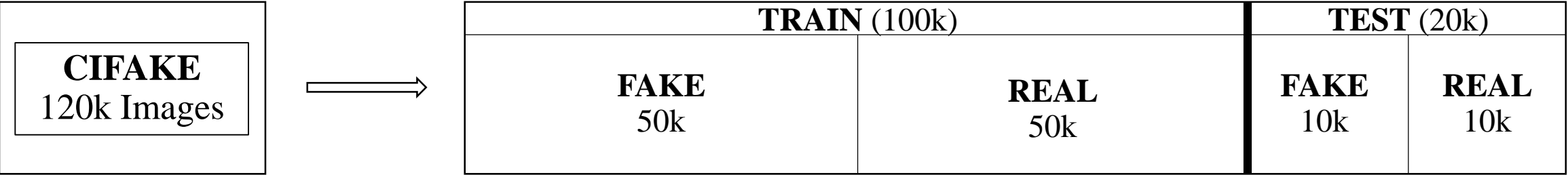
- Определить датасет для решения задачи классификации фейковых изображений.
- Провести EDA на выбранном наборе данных.
- Построить бейзлайн с использованием линейных моделей и постараться его улучшить.
- Реализовать сервис на FastApi и web-приложение на Streamlit для нашей задачи.

***Фейковое изображение** - это изображение, отредактированное или созданное с помощью нейронных сетей, либо других инструментов для обработки изображений.

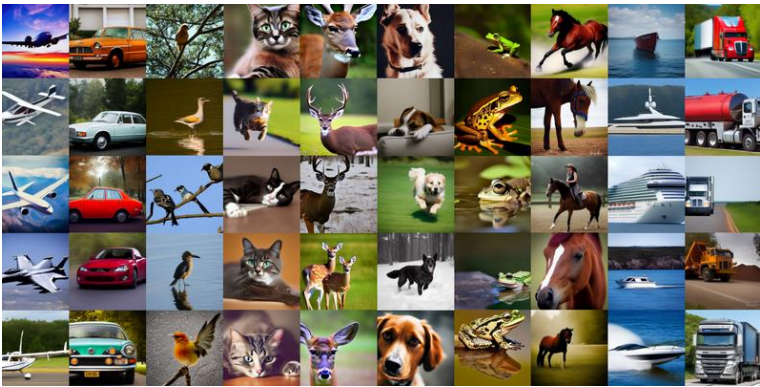
Фейковые изображения делятся на две группы, а именно:

-   → Изображения, относящиеся к классу человеческих лиц (a.k.a. Deepfake)
-   → Изображения, не относящиеся к классу человеческих лиц (a.k.a. Fake Image)

Датасет для обучения. EDA



- CIFAKE составлен из изображений, не относящихся к классу человеческих лиц.
- 60k реальных изображений были взяты из датасета CIFAR-10.
- 60k синтетических изображений были сгенерированы с помощью Stable Diffusion v1.4.



Примеры изображений из датасета CIFAKE

→

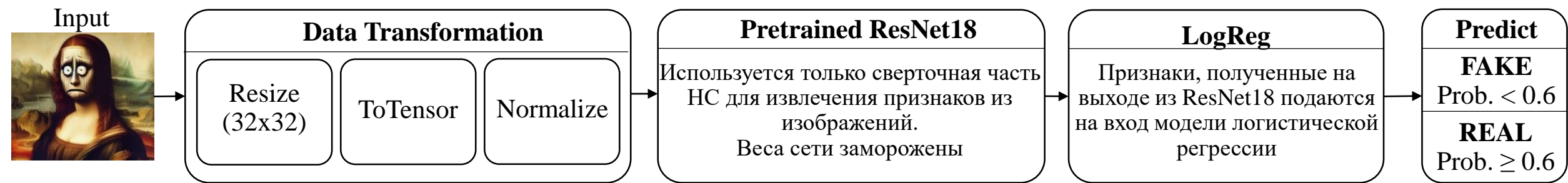
EDA (CIFAKE)	Mean			Std			Average Image Size
	Red	Green	Blue	Red	Green	Blue	
Train	0.472	0.463	0.418	0.238	0.237	0.266	32 x 32
Test	0.473	0.464	0.419	0.238	0.237	0.266	32 x 32
ImageNet	0.485	0.456	0.406	0.299	0.244	0.225	469 x 387

Выводы

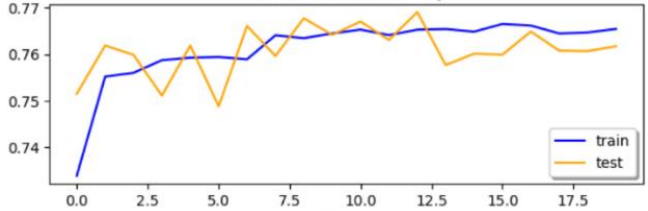
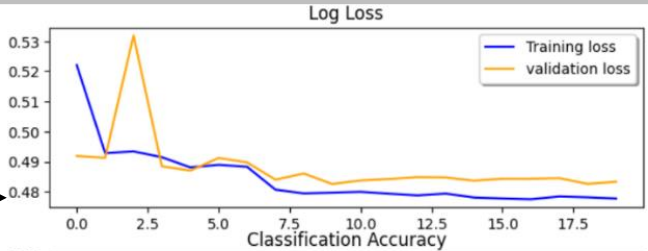
EDA показал отсутствие выбросов в разрезе цветовых профилей в CIFAKE.

EDA показал, что цветовой профиль датасета CIFAKE близок к цветовому профилю датасета ImageNet.

Архитектура бейзлайна

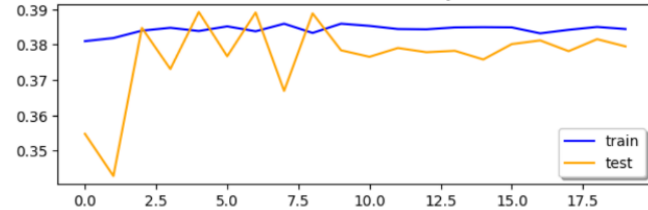
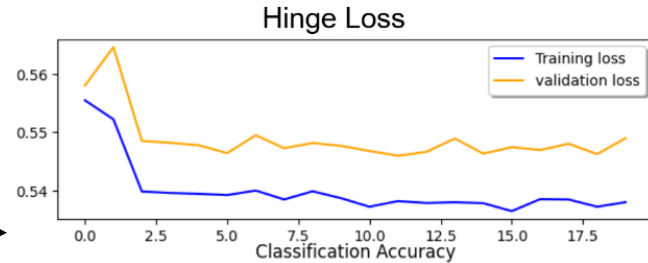


LogReg	Accuracy, %	Log Loss
Train (15 epoch)	76.54	0.478
Test (15 epoch)	76.17	0.483
Train (+finetune, 15 epoch)	80.46	0.454
Test (+finetune, 15 epoch)	80.24	0.457



↪ Метрика качества - Accuracy. Баланс классов 50/50

SVM	Accuracy, %	Hinge Loss
Train (15 epoch)	38.44	0.538
Test (15 epoch)	37.95	0.549
Train (+finetune, 10 epoch)	41.51	0.508
Test (+finetune, 10 epoch)	40.62	0.519



Реализация сервиса

Backend (FastApi)

POST /fit – обучение НС на основе изображений, подаваемых клиентом на вход модели

POST /set – установка модели (среди тех, что были обучены через /fit) в качестве модели для инференса

POST /predict – предсказание метки класса (REAL/FAKE) для изображения моделью для инференса

GET /models – вывод списка обученных моделей с подробной информацией о них, а именно: значения гиперпараметров, ассигасу, вид предобученной модели, функция активации на последнем слое

GET /eda – расчет профиля данных, подаваемых клиентом на вход модели

+

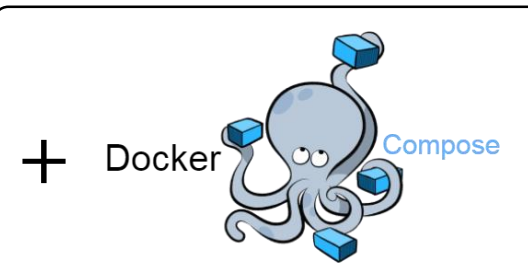
Frontend (Streamlit)

Пример работы web-приложения



https://disk.yandex.ru/i/GXvscBSP_ndJnA

Схема работы web-приложения



Распределение работы в команде

Тимонин Андрей Сергеевич	Анализ подходов к решению задачи классификации фейковых изображений.
	Изучение релевантных датасетов. EDA.
	Построение бейзлайна и его последующее улучшение.
	Реализация backend и frontend части web-приложения. Докеризация приложения.
Рябков Иван Юрьевич	Анализ подходов к решению задачи классификации фейковых изображений.
	Изучение релевантных датасетов. EDA.

Планы на второе полугодие

1. Проверить гипотезу о том, что модель, обучаясь на фейковых изображениях, сгенерированных одной моделью, не способна классифицировать фейковые изображения, созданные другой моделью.
2. Реализовать метод классификации фейковых изображений на основе «отпечатков (fingerprints)» генеративных моделей.
3. Реализовать приложение в Telegram-боте.

Спасибо за внимание!