#deepfake, explainable AI, synthetic data, face recognition

Разработка модели объяснимого ИИ для выявления синтетических изображений

Облизанов Александр, гр. 8304

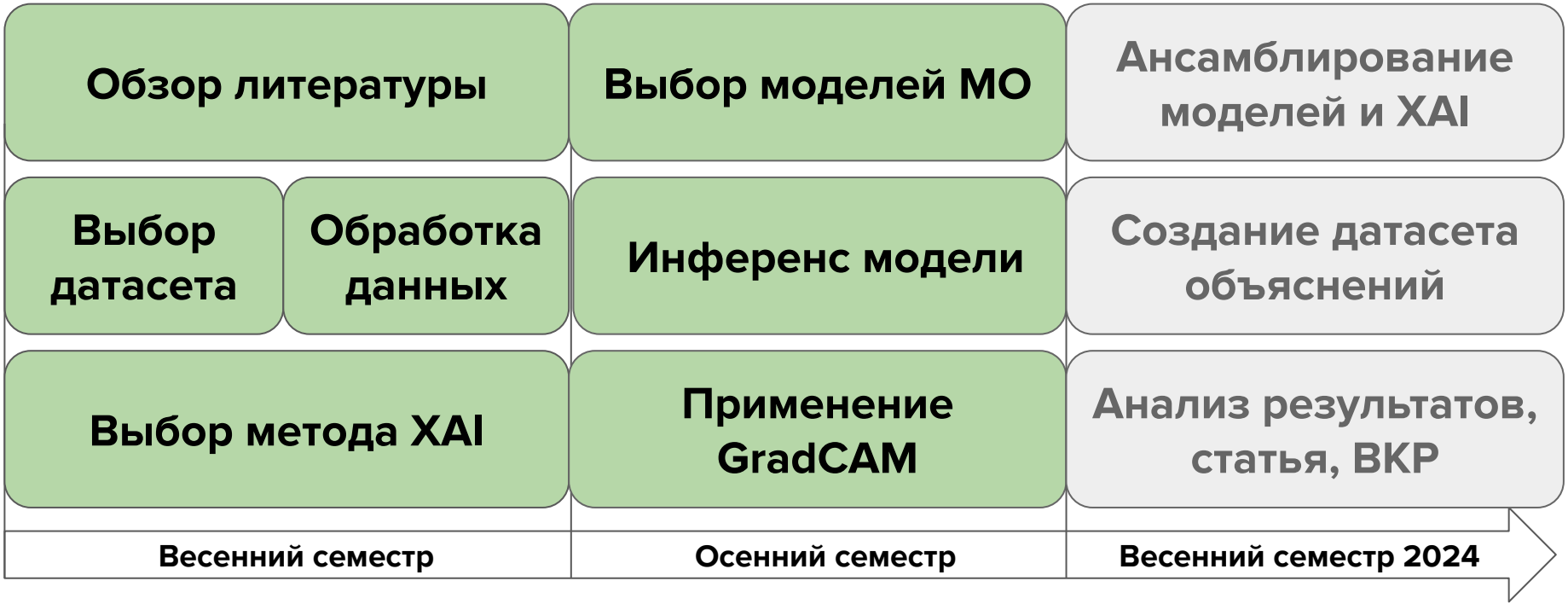
Постановка задачи

Объектом исследования являются подходы к классификации синтетических изображений.

Предметом исследования являются модели объяснимого искусственного интеллекта.

Целью работы является применение моделей объяснимого искусственного интеллекта для классификации синтетических изображений и реальных снимков.

Этапы выполнения



Датасет и обработка данных

DFDC Dataset

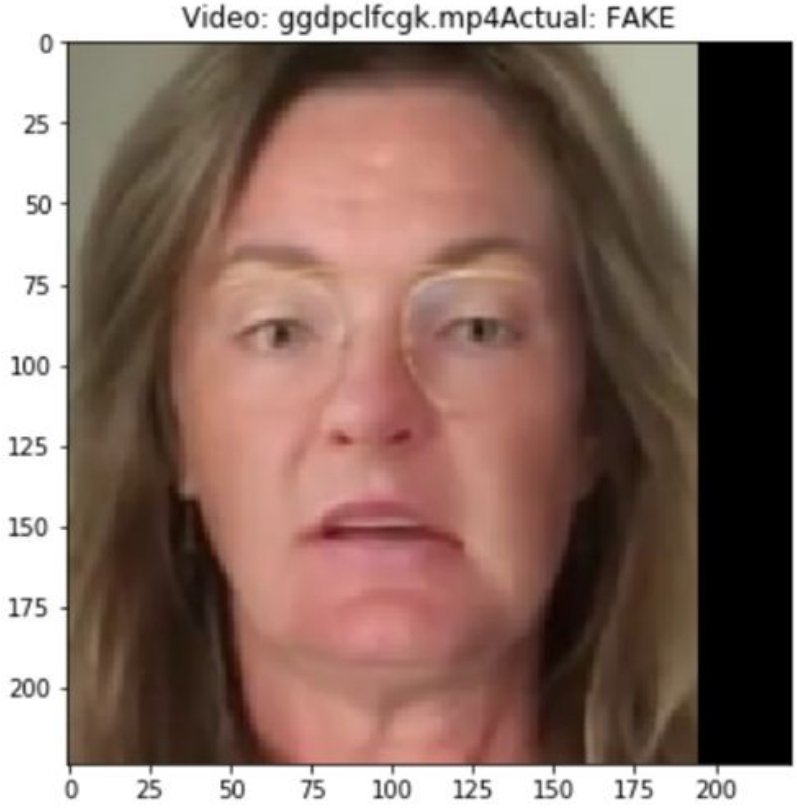
Input: видеозаписи (истинные и фейк) с людьми в любой части кадра

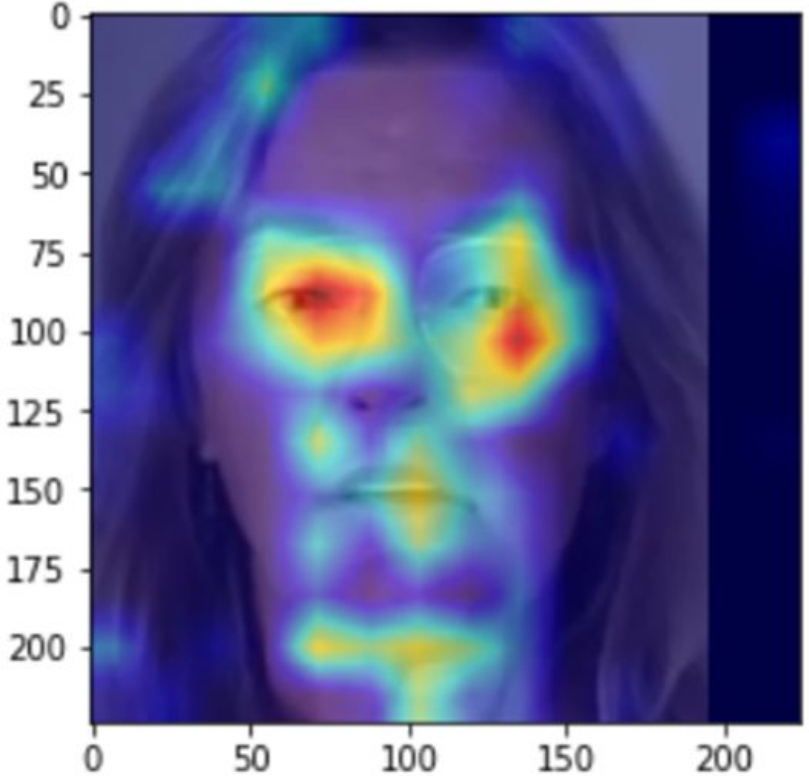
Process: многопоточное чтение кадров из видео, применение модели MTCNN и доп. обработка

Output: координаты краев прямоугольника, в котором есть лицо на кадре

GradCAM + ResNeXt Deepfake detection model

Интерпретация важности признаков на изображении

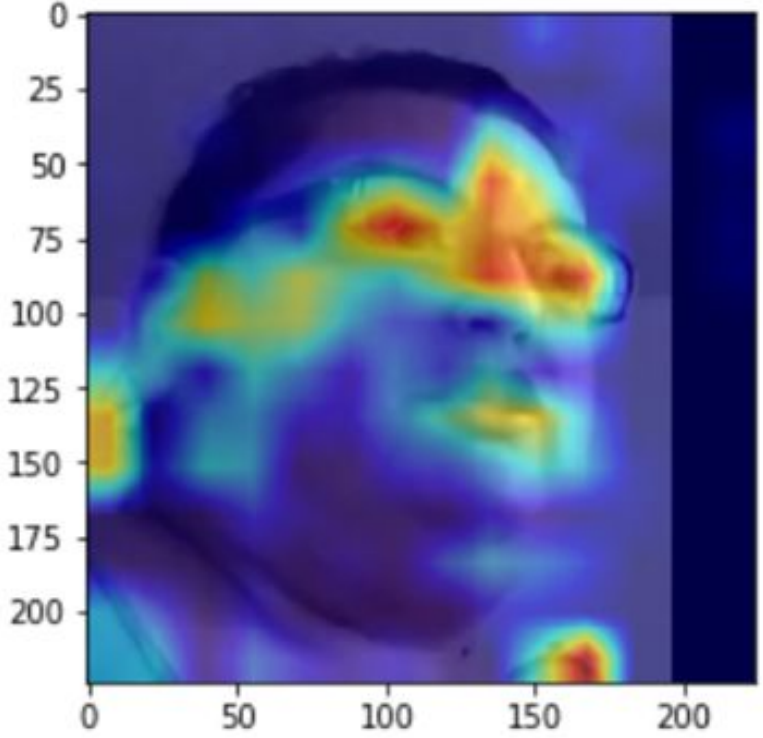




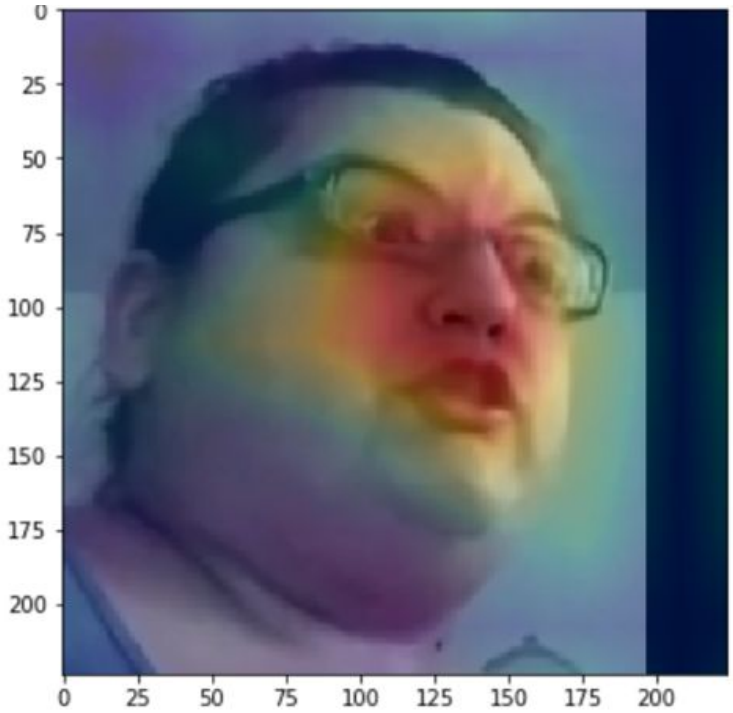
Разрешение интерпретаций GradCAM

Интерпретация производится по функциям активации слоев модели

Предпоследний ResNeXt блок



Последний ResNeXt блок



Архитектура

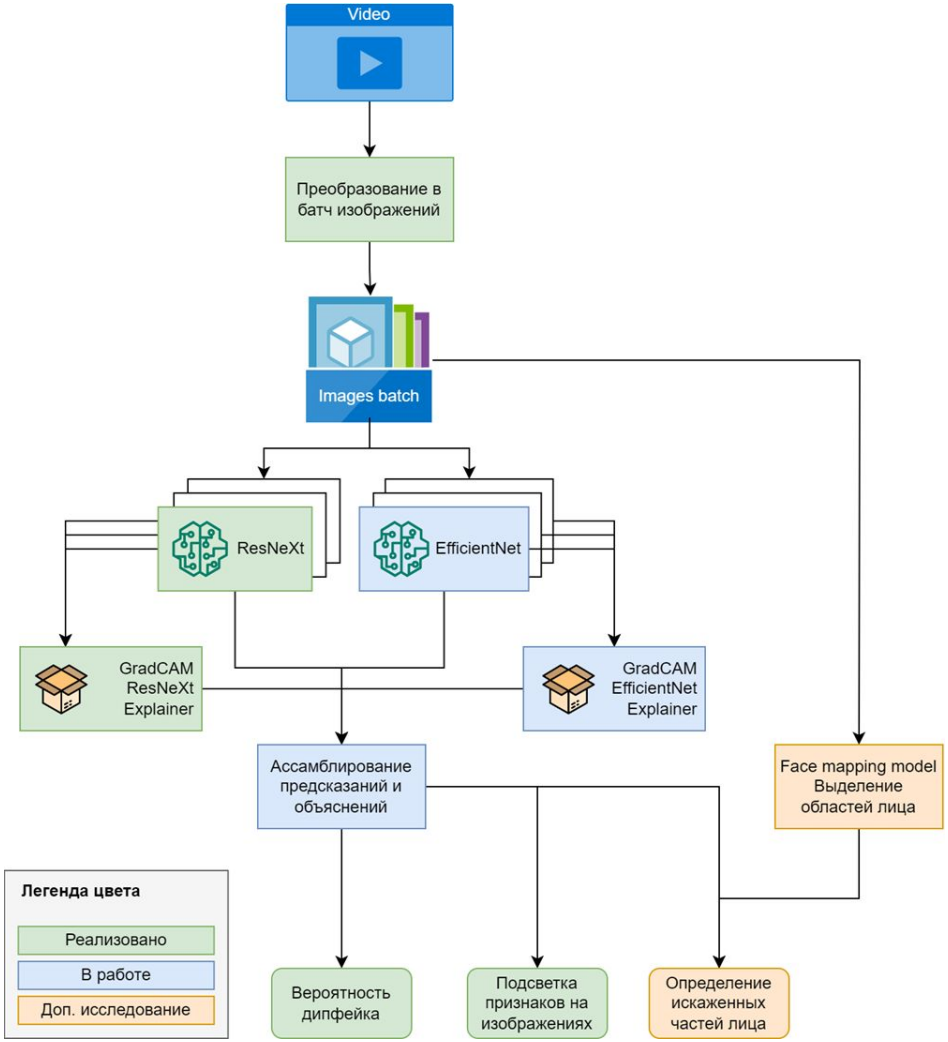
Выделение кадров лиц из видеозаписей

Ансамблирование моделей и интерпретаций

Адаптация GradCAM для моделей ResNeXt и EfficientNet

Доработка метода GradCAM для увеличения размерности интерпретации

Локализации признаков (Face mapping + Deepfake detection)



Результат и дальнейшие исследования

- ПО позволит применять ансамбли моделей ResNeXt и EfficientNet (наиболее эффективные модели по распознаванию deepfake) вместе с методом объяснимого ИИ GradCAM и получать интерпретацию (объяснение) предсказания модели
- Полученные интерпретации формируют датасет для последующего анализа:
  - Уязвимости алгоритмов распознавания и создания дипфейков
  - Сравнение человеческого и машинного “восприятия” изображений
  - Повышение доверия пользователей к системам ИИ с помощью технологий объяснимого ИИ