РАЗРАБОТКА МОДЕЛИ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБНАРУЖЕНИЯ ПРИЗНАКОВ СИНТЕЗА ИЗОБРАЖЕНИЯ ЛИЦА ЧЕЛОВЕКА

Bobokhonov Akhmadkhon Kholmirzokhon ugli
Sun'iy intellekt va axborot tizimlari kafedrasi, Sh. Rashidov nomidagi Samarqand davlat universiteti b_akhmadkhon@mail.ru,

Аннотация. В рамках работы рассмотрены существующие подходы к обнаружению синтеза изображения лица человека. Был разработан алгоритм и реализующий его программный комплекс, в котором используется предложенная в работе модификация «HiResCAM++» метода объяснимого искусственного интеллекта GradCAM. С применением моделей машинного обучения ResNeXt и EfficientNet, осуществляется обнаружение признаков синтеза на изображениях, их визуализация, а также расчет вероятности синтеза изображения или видеозаписи. В результате были определены преимущества использования объяснимого искусственного интеллекта для решения задачи выявления синтеза изображений, а также подтверждены преимущества разработанной модификации «HiResCAM++» метода GradCAM для решения поставленной задачи.

Ключевые слова. GradCAM, HiResCAM, HiResCAM++, ResNeXt, EfficientNet, deepfake

Введение

В последние годы генерация синтетических изображений и видеозаписей людей или других живых существ кратно возросла. Эта концепция также известна как «deepfake», где «deep» отражает использование нейронных сетей Deep Learning, а «fake» - отражает подделку и подмену по отношению к исходному вводу.

Такие приложения, как Snapchat и Reddit [1], используют подходы DL для изучения конкретных функций изображения и переноса их на другое изображение или видео. Эти достижения deepfake подчеркнули потенциальные последствия цифровых манипуляций с лицом. Несмотря на то, что это интересно и удобно, существует угроза того, что злоумышленники будут использовать deepfake-атаки, чтобы поставить под угрозу безопасность других.

Хотя многие модели генерации deepfake производят неразличимые репродукции, эти подделки все еще могут быть обнаружены либо специализированными криминалистическими методами, либо методами глубокого обучения [2]. Почти каждый генератор deepfake оставляет на

изображении следы своих операций свертки.

Несмотря на высокую точность некоторых моделей, представленных сегодня для решения этой задачи, технологии подделки изображений так же не стоят на месте, что не дает гарантии, что модель будет так же точно распознавать deepfake изображения, полученные новым методом.

Выбор моделей машинного обучения

Выбор моделей машинного обучения основывался на соревновании по датасету DFDC, проводимому на площадке Kaggle в 2019-2020 годах. В рамках соревнования исследователи предлагали алгоритмы определения синтеза изображения лица человека, которые единообразно оценивались по метрике потерь [3]. В результате анализа наиболее успешных решений были определены типы моделей, наиболее подходящие для решения данной задачи:

1. EfficientNet B7 (1-е и 3-е место в рейтинге) [3];
2. ResNeXt 50, 101 (2-е и 5-е место в рейтинге) [4].

СХЕМА АЛГОРИТМА

Алгоритм обнаружения и интерпретации признаков синтеза лица человека состоит из следующих этапов:

1. Подготовка данных:
   a. Выделение кадров из видеозаписи
   b. Определение маски лиц на кадрах
   c. Обрезка и ресемплинг изображения
2. Применение моделей машинного обучения
3. Применение метода HiResCAM к моделям машинного обучения
4. Ассамблирование предсказаний моделей для каждого кадра
5. Преобразования для получения итогового результата:
   a. Вычисление итоговой вероятности синтеза видеозаписи
   b. Выбор кадров для отображения тепловых карт признаков синтеза

Схема основных этапов работы алгоритма представлена на рис. 1.

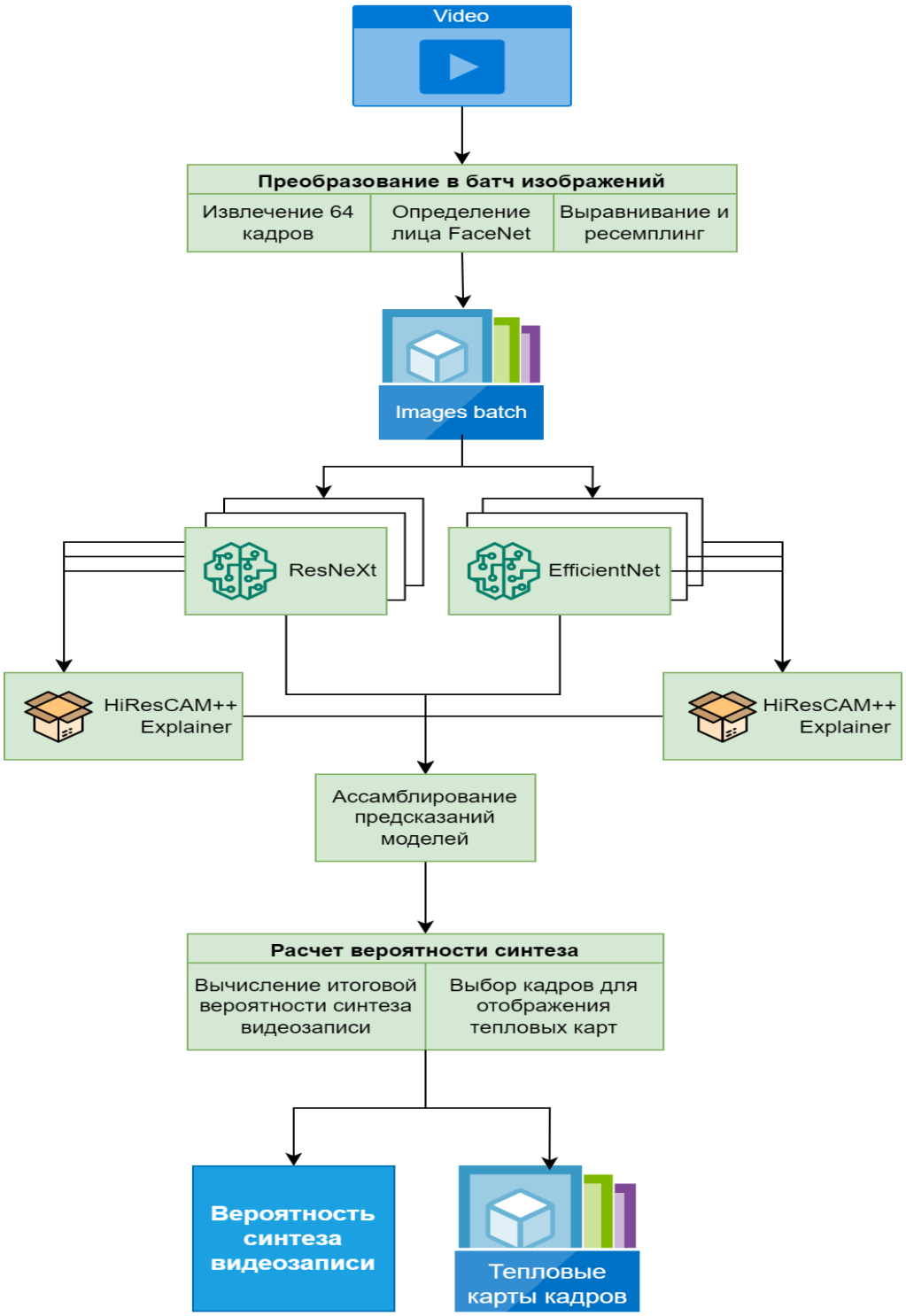

Рисунок 1 – Схема алгоритма

Работа алгоритма на обучающем наборе данных

Для апробации были выбраны ранее описанные инференсы моделей, которые были обучены в рамках соревнования Deepfake Detection Challenge и набрали высокий балл в тестовом прогоне:

1. EfficientNet B7
2. ResNeXt-101

Для оценки корректности работы объяснимого искусственного

интеллекта (ИИ), а также алгоритмов ансамблирования и точности предсказания вероятности дипфейка для всего видео, были использованы видеозаписи из тестового набора данных, для которых известна истинная метка подлинности видеозаписи.

В целом на тренировочном наборе точность ансамбля моделей (при округлении вероятности до 0 и 1) составила 97%, простая функция потерь (без округления вероятностей) – 0.14. Сравнение с референсами – лучшими решениями соревнования – представлено на рис. 2, где EN+ResNeXt – разработанный ансамбль моделей.

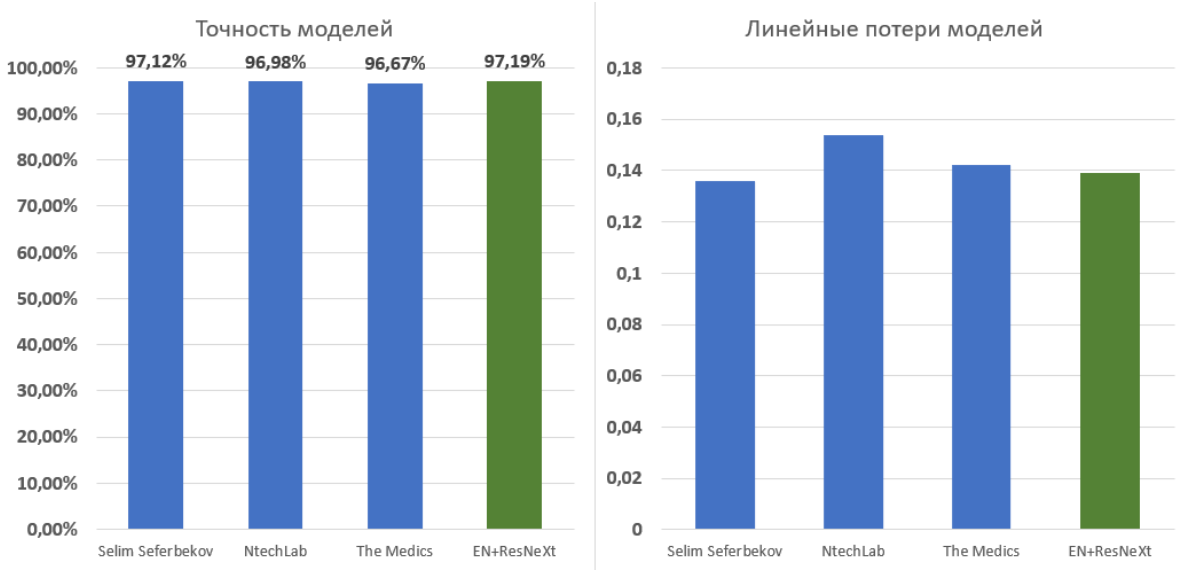

Рисунок 2 – Точность и потери решения в сравнение с другими

Таким образом, разработанное решение удовлетворяет требованию к алгоритму по доказанной точности не ниже, чем у решений, использующих аналогичные модели машинного обучения. Использование метода объяснимого искусственного интеллекта не влияет на работу самой модели.

Интерпретация признаков синтеза методом HiResCAM++

Для демонстрации работы метода HiResCAM++ были выбраны несколько видеозаписей с синтезом лица человека, на которых можно визуально отличить артефакты и искажения, вызванные синтезом. Выбором видеозаписей с синтезом лица человека, в которых наглядно прослеживаются артефакты и искажения, обосновано с целью наглядного сравнения эффективности методов HiResCAM++ и GradCAM [5]. Рассмотрим пример кадра из синтезированного видео, представленный на рис. 3.

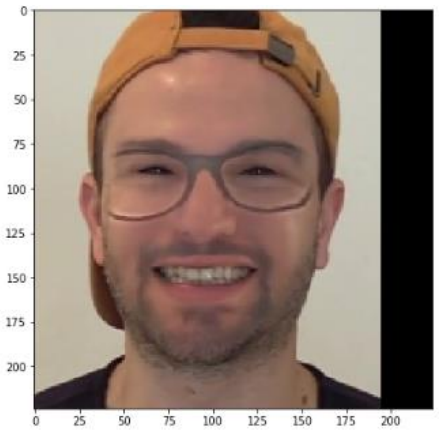

Рисунок 3 – Пример синтезированного кадра

В кадре явно наблюдается искажение лица, которое проявляется в следующих признаках:

1. Дужки очков отличаются справа и слева, имеют неестественную форму;
2. Несоразмерно маленькие и черные глаза;
3. Неестественное расположение очков относительно носа.

Рассмотрим объяснения следующих методов:

4. «GradCAM N» – объяснение методом GradCAM на последнем сверточном слое;
5. «GradCAM N-1» – объяснение методом GradCAM на предпоследнем сверточном слое;
6. «HiResCAM++» - объяснение разработанным методом HiResCAM++.

На рис. 4 изображены интерпретации (тепловые карты) данных методов.

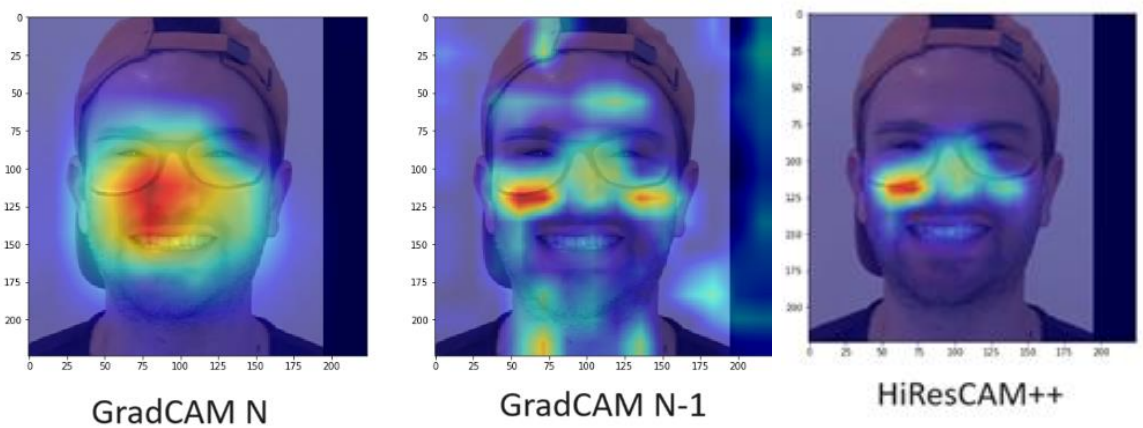

Рисунок 4 – Интерпретация признаков синтеза

На тепловых картах, где цветовая гамма тяготеет к красному, выделены области, предположительно соответствующие зонам синтеза изображения,

отображающие влияние признаков на более высокие значения выхода модели.

В данном контексте, рассматривается применение методов объяснимого искусственного интеллекта:

Метод GradCAM-N показал адекватную локализацию области синтеза, преимущественно сосредоточенной в области выше центра лица (глаза и очки). Однако, разрешение интерпретации оказалось слишком низким, что препятствует детальной идентификации особых черт (части лица, предметы), свидетельствующих о синтезе.

Метод GradCAM-N-1 выявил необычную форму дужек очков, но также выделил части шеи, подбородка и кепки, несмотря на отсутствие заметных искажений в этих областях изображения [6]. Это следствие игнорирования карт признаков последнего слоя свертки, хотя они несут значимую информацию о закономерностях на изображении в целом. Таким образом, этот метод предоставляет более детализированное объяснение, но может также выделять неприменимые области.

Метод HiResCAM++ выделил неестественные дужки очков и область носа, расположенные в пределах области, выделенной методом GradCAM-N, с разрешением, соответствующим методу GradCAM-N-1 [7]. Данный метод исключает излишние признаки, отфильтрованные на последнем слое свертки модели. Таким образом, метод HiResCAM++ обеспечивает наиболее понятную интерпретацию работы модели и помогает выявить признаки синтеза лица человека.

Следует отметить, что ни один из методов не выявил глаза, подвергшиеся синтезу, что, возможно, связано с особенностями обучающих данных и работы модели. В дополнительных исследованиях других интерпретаций выявлено, что при наличии очков в синтезированном видео алгоритм склонен выделять именно этот атрибут.

Сравнение HiResCAM++ с методом HiResCAM

Для подтверждения корректности работы алгоритма HiResCAM++ проведем сравнение интерпретации данного метода с методом HiResCAM, примененным на последний слой свертки модели машинного обучения. Для сравнения будет использована модель ResNeXt-101, входящая в состав ансамбля алгоритма.

Авторами метода HiResCAM доказана его интуитивность, то есть выведена математическая зависимость между работой модели машинного обучения на полносвязном слое нейронной сети после свертки и весами важности признаков на тепловой карте метода [8]. Метод гарантирует, что все подсвеченные признаки в действительности увеличивают значение на выходе

модели машинного обучения.

В таком случае если области, подсвеченные методом HiResCAM++, будут входить в подмножество ненулевых областей метода HiResCAM, то практически можно гарантировать, что все они также увеличивают значение на выходе модели машинного обучения. Если области, выделенные методом HiResCAM++, подтверждаются как важные с помощью анализа методом HiResCAM, это подтверждает согласованность двух методов и увеличивает уверенность в том, что выделенные признаки действительно являются ключевыми для принятия решений моделью [9].

Рассмотрим на примере работу двух методов, интерпретация которых и исходное изображение из синтезированного видео представлена на рис. 5.

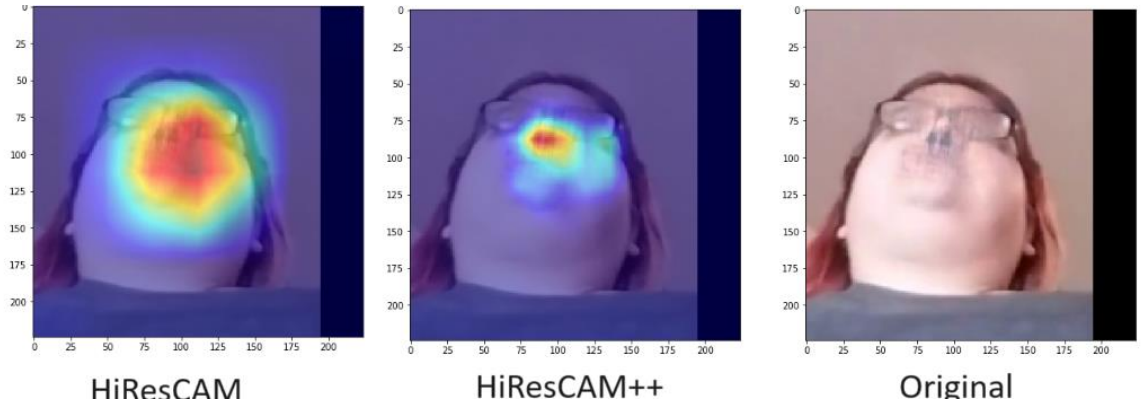

Рисунок 5 – Сравнение работы метода HiResCAM и HiResCAM++

Из данного примера видно, что область ненулевых весов тепловой карты метода HiResCAM++ действительно входит в подмножество области ненулевых весов тепловой карты HiResCAM. При этом, также как и на прошлом сравнении, на изображении явно видны артефакта, вызванные синтезом лица. Эти артефекты наиболее ярко проявляются в области носа человека. Именно метод HiResCAM++, за счет более высокого разрешения, приводит более детальную интерпретацию.

На основе всей тренировочной и тестовой выборки был произведен расчет процент пикселей тепловой карты HiResCAM++, которые не равны 0 и находятся вне области пикселей тепловой карты HiResCAM. Также было рассчитано их среднее значение – имеют ли такие пиксели высокий вес. Результаты приведены в таблице 2.

Таблица 1 – Расчет процента отклонений HiResCAM++ от метода HiResCAM

Количество изображений для интерпретации	**82340**
Процент ненулевых пикселей HiResCAM++ вне области ненулевых пикселей HiResCAM	**3%**
Среднее значение пикселей HiResCAM++ вне области ненулевых пикселей HiResCAM (нормализиванное)	**0.09**
Среднее значение пикселей HiResCAM++ внутри области ненулевых пикселей HiResCAM (нормализованное)	**0.71**
Процент совпадений областей	**63%**

По результатам сравнения выводится, что ограниченное количество выделенных пикселей (не более 3%) на тепловой карте метода HiResCAM++ не обязательно оказывает положительное воздействие на прогнозирование модели. При этом их нормализованное значение не превышает 0.1, то есть в интерпретации они являются областями низкого влияния на прогнозы. Процент совпадения областей в 63% свидетельствует о том, что различия в детализации методов приводят к отличиям в их интерпретации, однако это в большей степени не отражается на интуитивности метода HiResCAM++.

В итоге можно сделать практический вывод о том, что метод HiResCAM++ с незначительными отклонениями сохраняет основной принцип интуитивности, демонстрированный методом HiResCAM, при этом обеспечивая более детальную интерпретацию влияния признаков на прогнозы модели.

Вывод

В ходе выполнения данной работы были выполнены поставленные задачи и достигнуты следующие результаты:

Определены критерии к подходам оценивания методов XAI: количество метрик, имеющих математическое или методическое описание, количество оцениваемых технических характеристик, учет разнородности входных данных, виды исследований для вычисления метрик, объем необходимых исследований для вычисления метрик;

Проведено исследование существующих метрик оценки методов XAI, которое выявило, что многие аналоги не учитывают разнородность входных данных, либо же не предлагают математическое или методическое описание метрик;

Разработанное программное обеспечение позволяет проводить эксперименты с двумя методами XAI (SHAP, LIME) и двумя моделями

машинного обучения (модель линейной регрессии и модель деревьев решений), что не позволяет в полной мере оценить работу методов в реальных условиях.

References

1. Mirsky Y., Lee W. The creation and detection of deepfakes: A survey //ACM Computing Surveys (CSUR). – 2021. – Т. 54. – №. 1. – С. 1-41.
2. Guarnera L. et al. Preliminary forensics analysis of deepfake images //2020 AEIT international annual conference (AEIT). – IEEE, 2020. – С. 1-6.
3. Jain N. et al. Imperfect ImaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses //Artificial Intelligence. – 2022. – Т. 304. – С. 103652.
4. Tolosana R. et al. Deepfakes and beyond: A survey of face manipulation and fake detection //Information Fusion. – 2020. – Т. 64. – С. 131-148.
5. Dufour N., Gully A. Contributing data to deepfake detection research //Google AI Blog. – 2019. – Т. 1. – №. 3.
6. Korshunov P., Marcel S., Fakes D. A new threat to face recognition? Assessment and detection. – 2018.
7. Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., & Dorofeeva, A. (2023). Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. Applied System Innovation, 6(1), 26.
8. William I. et al. Face recognition using facenet (survey, performance test, and comparison) //2019 fourth international conference on informatics and computing (ICIC). – IEEE, 2019. – С. 1-6.
9. Draelos R. L., Carin L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks //arXiv preprint arXiv:2011.08891. – 2020.