Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа, обработки и интерпретации больших данных.**

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Анализ методов распознавания дипфейков

Студент ИУ6-12М — 13.12.2023 — С.В. Астахов
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель — 13.12.2023 — А.А. Сотников
(Подпись, дата) (И.О.Фамилия)

2023 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Анализ методов распознавания дипфейков

Студент группы ИУ6-12М
Астахов Сергей Викторович
(Фамилия, имя, отчество)

Магистерская программа 09.04.01/07 Интеллектуальные системы анализа, обработки и интерпретации больших данных.

Направленность НИР (исследовательская, практическая, производственная, др.)
исследовательская

Источник тематики (кафедра, предприятие, НИР) кафедра

График выполнения НИР: 25% к 4 нед., 50% к 7 нед., 75% к 11 нед., 100% к 14 нед.

Техническое задание выполнить анализ существующих систем и методов распознавания дипфейков, рассмотреть их достоинства, недостатки и границы области применения для дальнейшего формирования архитектурных требований к собственной системе определения дипфейков

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 25-30 листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
нет

Дата выдачи задания « 1 » сентября 2023 г.

Руководитель НИР 01.09.2023 А.А. Сотников
(Подпись, дата) (И.О.Фамилия)

Студент 01.09.2023 С.В. Астахов
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах.

РЕФЕРАТ

Отчет — 29 с., 29 рис., 19 ист.

ДИПФЕЙК, НЕЙРОННАЯ СЕТЬ, КОМПЬЮТЕРНОЕ ЗРЕНИЕ, МАШИННОЕ ОБУЧЕНИЕ, СТЕГОАНАЛИЗ, ФОРЕНСИКА.

Объектом исследования являются системы и методы определения дипфейков.

Цель работы — анализ структуры и методов, используемых в рассматриваемых системах для дальнейшего формирования архитектурных требований к собственной системе определения дипфейков.

В процессе работы был проведен анализ существующих систем и методов распознавания дипфейков, рассмотрены их достоинства, недостатки и границы области применения.

Актуальность работы обусловлена тем, что с увеличением сложности технологии дипфейков становится все труднее отличить настоящий контент от поддельного. Это вызывает опасения относительно потенциального использования дипфейков для злонамеренных целей, таких как распространение ложной информации или манипулирование общественным мнением. В результате возникает необходимость в эффективных методах их обнаружения и идентификации.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Визуальный артефакт - это графический дефект в цифровом изображении, который может возникнуть в результате ошибок при съемке, обработке или хранении изображений;

Дипфейк — технология синтеза изображений на основе искусственного интеллекта и нейросетей;

Нейронная сеть — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей;

Область интереса — фрагмент набора данных, выделяемый для дальнейшего детального анализа;

Стегоанализ — извлечение скрытой информации из содержащего её сообщения и дальнейшая её дешифровка.

ВВЕДЕНИЕ

В наше время социальные сети и мессенджеры стали неотъемлемой частью жизни многих людей. Однако, вместе с возможностью общения и обмена информацией, появилась и проблема фейковых новостей и фотографий.

Дипфейки - это фотографии или видео, созданные с помощью искусственного интеллекта, которые могут быть использованы для создания фальшивых новостей или дезинформации.

В связи с этим, возникает необходимость в разработке методов определения дипфейков, которые позволят бороться с распространением фальшивой информации. В данном реферате рассмотрены основные методы определения дипфейков.

Данный реферат является предпроектным исследованием, предворяющим проектирование собственной системы определения дипфейков. Очевидно, что прежде, чем приступить к разработке системы распознавания дипфейков, необходимо произвести анализ существующих методов обнаружения дипфейков. Приведенная ниже классификация методов составлена на основе статьи инженерно-исследовательского центра цифровой криминалистики [1].

1. Методы на основе сверточных и капсульных нейросетей

Так, как нейросети являются довольно универсальной и относительно простой в конфигурации математической моделью ИИ, не требующей предобработки данных, первые алгоритмы определения дипфейков использовали именно их. Такие модели обрабатывают предоставленное видео в покадровом режиме, оценивают вероятность появления дипфейка в каждом кадре и затем обобщают полученные результаты для всего видео. В этой группе можно выделить две подгруппы: модели на основе трансферного обучения и специально созданные нейросети.

Трансферное обучение — технология, позволяющая уменьшить набор данных, необходимый для тренировки глубокой нейросети, за счет использования предварительно подготовленной сети, обученной на другом наборе данных, но выполняющей задачу, аналогичную требуемой [2].

Пример применения такого подхода: необходимо обучить нейросеть классифицировать изображения еды. Пусть, при этом, существует уже обученная нейросеть для классификации изображений животных. В сети для классификации животных более глубокие слои будут отвечать за определение общих паттернов, необходимых для дальнейшего “понимания” изображений (например, определение форм и границ предметов). Необходимо будет лишь заменить и обучить самые верхние слои сети, которые буду отвечать за интерпретацию промежуточных результатов (например, определять по геометрической форме тип животного или блюда). Иллюстрация работы технологии трансферного обучения приведена на рисунке 1.

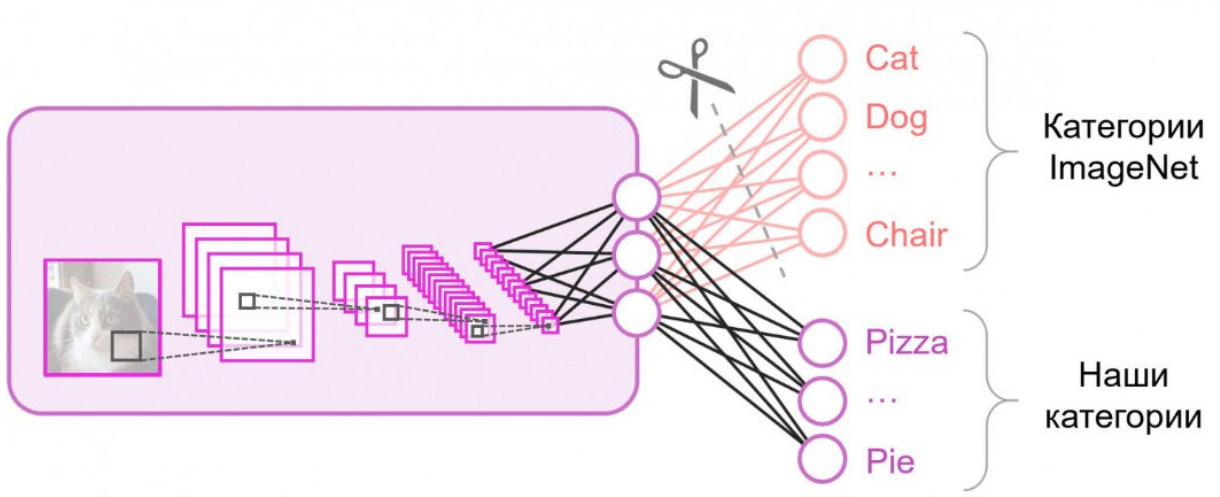

Рисунок 1 — использование технологии трансферного обучения

Примером применения трансферного обучения для решения задачи определения дипфейков является нейросеть, полученная учеными из Мэрилендского университета (рисунок 2) [3].

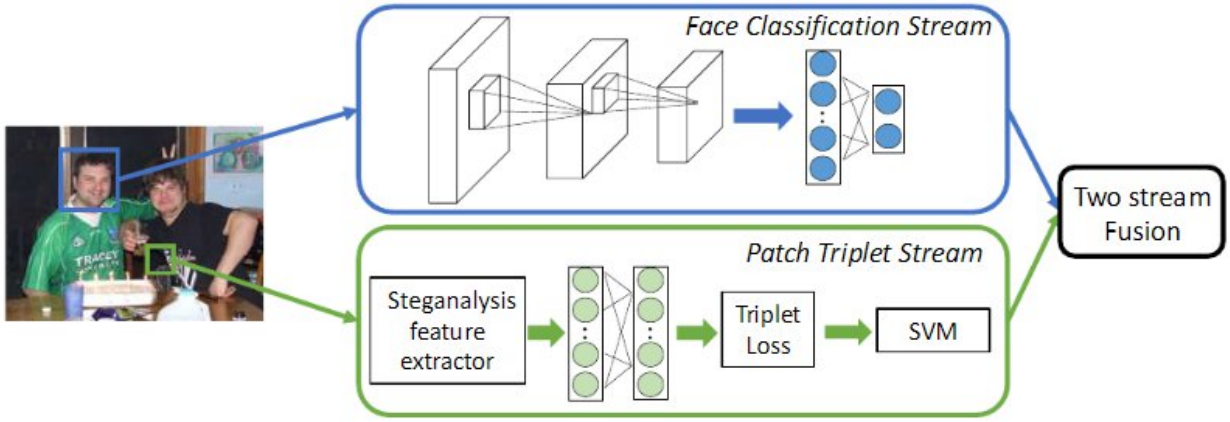

Рисунок 2 — применение трансферного обучения для задачи определения дипфейков

В представленной модели есть два параллельных потока обработки изображения, результаты которых объединяются лишь на самой последней фазе. Верхний поток основывается на методе трансферного обучения на основе нейросети, натренированной для классификации человеческих лиц. Этот поток работает с такой информацией, как, например, форма и геометрия лица. Нижний же поток ищет более скрытые закономерности и артефакты, такие, как локальные изменения уровня зашумленности изображения.

Альтернативой трансферному обучению является разработка собственных архитектур нейросетей. Примером такой сети может быть система, разработанная в Национальном институте информационных и коммуникационных технологий Японии [4].

Схема их разработки представлена на рисунке 3. Сначала изображение обрабатывается предобученной сверточной нейросети VGG-19 [5]. На данном этапе, как и в случае трансферного обучения, целью является выделение базовых паттернов в изображении. Затем данные передаются в капсульную нейросеть, которая анализирует изображение на более высоком уровне (рисунок 3).

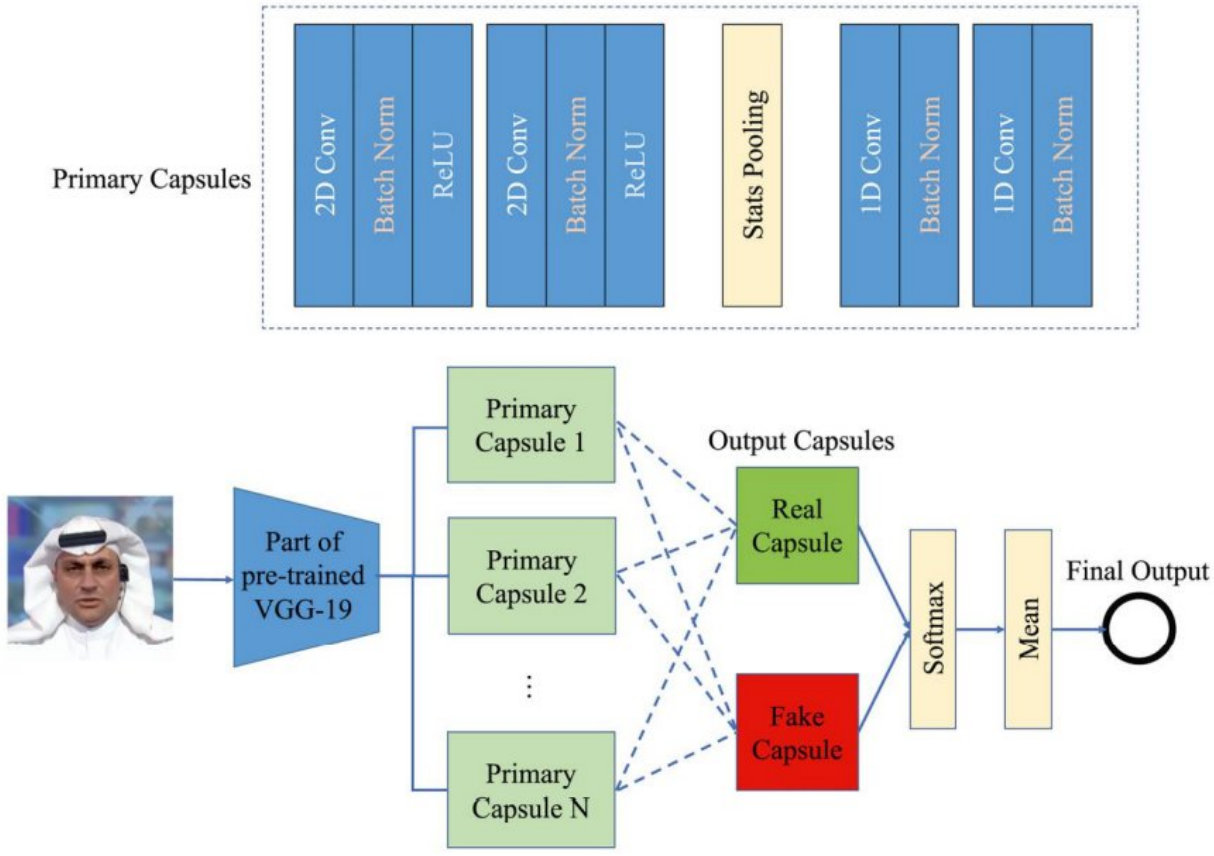

Рисунок 3 — структура системы обнаружения дипфейков с использованием капсульной нейросети

Капсулы используются для представления объектов или их частей, которые затем объединяются для формирования более высокоуровневых представлений. Капсулы - это группы нейронов, которые кодируют свойства

объекта, такие как его положение, ориентация и размер, а также вероятность того, что объект присутствует. Капсулы разработаны таким образом, чтобы быть эквивариантными, то есть они могут распознавать объект независимо от его положения или ориентации. Капсульные сети показали свою эффективность в улучшении точности задач распознавания изображений, особенно в случаях, когда объекты затенены или имеют несколько ориентаций [6].

2. Методы на основе временной согласованности

2.1. Комбинация сверточной и рекуррентной нейронных сетей

Видео представляет из себя последовательность кадров, в которой соседние кадры сильно коррелированы между собой. При генерации дипфейка в режиме покадровой обработки исходного видео корреляция между последовательными кадрами нарушаются, могут происходить мерцания цвета или резкий сдвиг положения лица.

В данном методе как правило используется комбинация сверточных нейросетей, о которых было рассказано выше, с рекуррентным нейросетями (в частности, нейросетями долгой краткосрочной памяти).

Пример такой системы был разработан в Имперском колледже Лондона [7]. Структура системы представлена на рисунке 4.

На первом этапе обработки каждый из кадров обрабатывается с помощью сверточной нейросети (англ. CNN — Convolutional neural network), таким образом отсекается информация о фоне видео, определяется геометрия лица и другие метрики изображения в его зоне (рисунок 5).

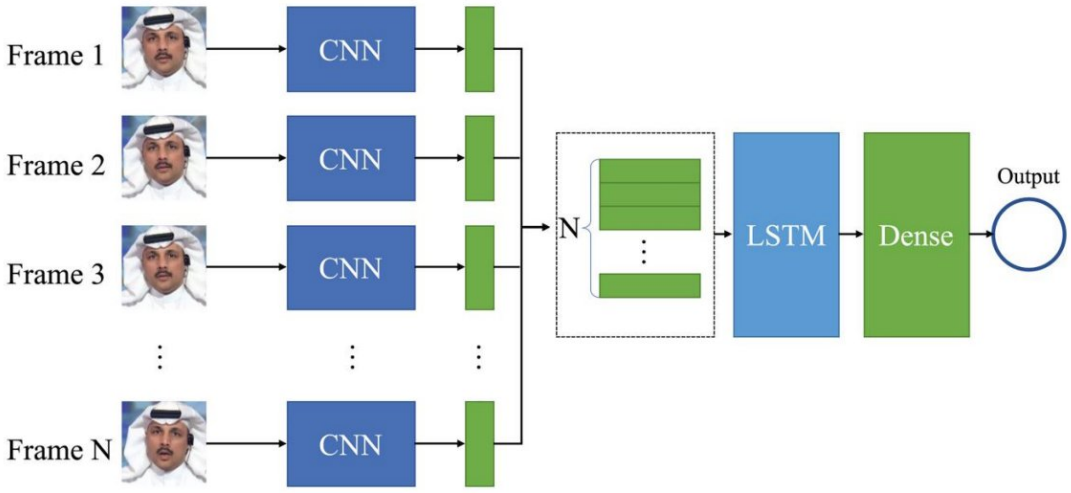

Рисунок 4 — структура системы обнаружения дипфейков с использованием рекуррентной нейросети

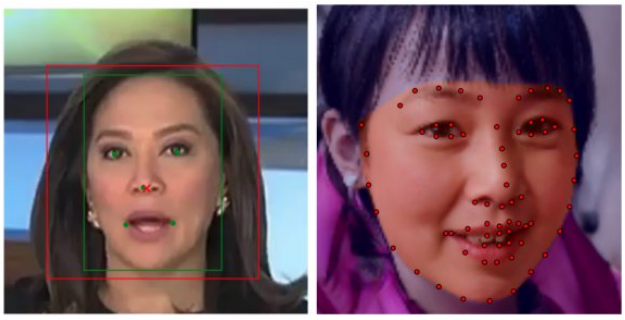

Рисунок 5 — предобработка данных с помощью сверточной неросети

Затем последовательность предобработанных кадров обрабатывается с помощью нейросети долгой краткосрочной памяти (англ. LSTM — Long short-term memory) [8].

Они являются подвидом рекурентных нейросетей (рисунок 6), где нейроны в скрытых слоях имеют ячейки памяти, которые позволяют им запоминать прошлые входные данные и использовать их как часть процесса принятия решений. Это позволяет выявлять закономерности и понимать контекст данных более точно, чем традиционные алгоритмы машинного обучения. Рекуррентные сети могут быть использованы для задач языкового моделирования, машинного перевода, распознавания речи, генерации текста и музыки, анализа временных рядов и других задач, связанных с последовательными данными.

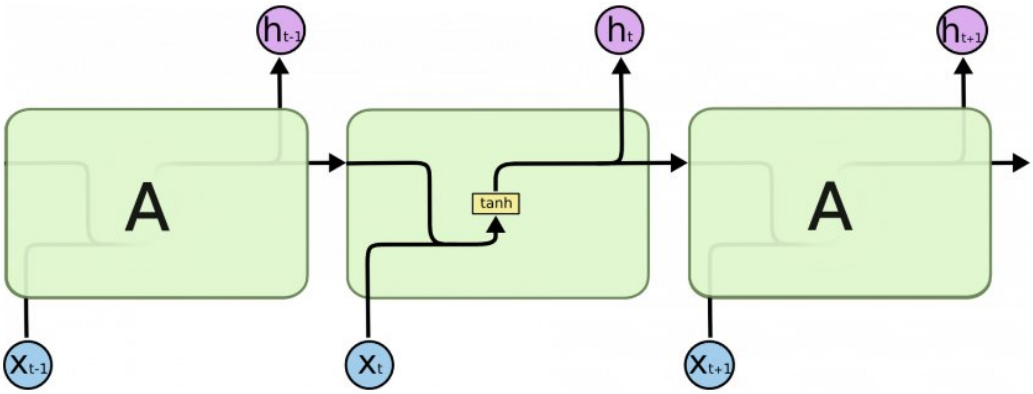

Рисунок 6 — модули рекуррентной нейронной сети

Основной проблемой стандартных рекуррентных нейросетей является сложность настройки для просчета истории на большую глубину, так как параметры со временем затираются. Сети с долгой краткосрочной памятью были созданы специально для решения проблемы долговременных зависимостей (рисунок 7).

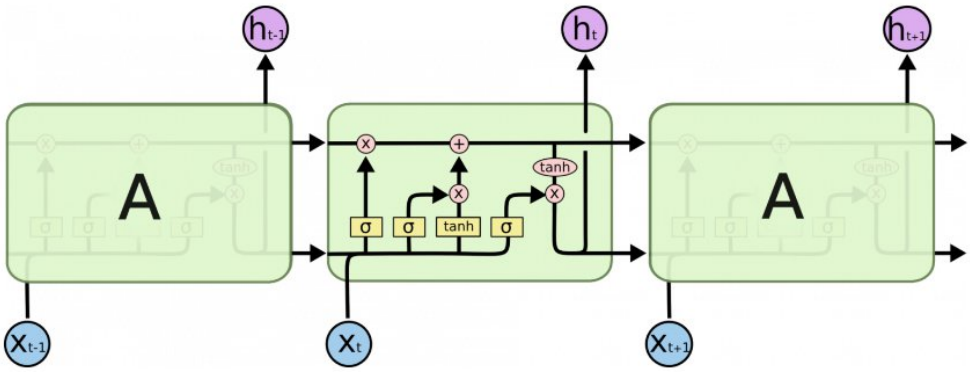

Рисунок 7 — модули LSTM сети

В отличии от обычной рекуррентной сети, LSTM-сеть в составе каждого модуля четыре слоя, которые призваны сделать управление состоянием ячейки более гибким:

- слой фильтра забывания — определяет, какая информация должна быть удалена из состояния ячейки;
- слой входного фильтра (является комбинацией сигмоидального слоя и гиперболического тангенса) — определяет, какая информация должна быть сохранена;
- слой выходного фильтра — определяет, какие данные будут поданы на управляющий вход следующего.

2.2. Улучшенные системы анализа временной согласованности

Несмотря на то, что использование классической LSTM дает хорошие результаты только на видео высокого качества, в то время как видео низкого разрешения с использованием плохого освещения или ракурсов будет обрабатываться с низкой точностью.

Эту проблему частично удалось решить ученым из лаборатории VIPER [9]. Архитектура системы в сущности остается неизменной, но каждому кадру анализируемого видео в соответствие выставляется “вес” — коэффициент, обозначающий качество кадра, вычисляемый на основе количества шумов, равномерности освещения и т.п. В конце обработки видео считается средневзвешенная вероятность наличия дипфейка (рисунок 8).

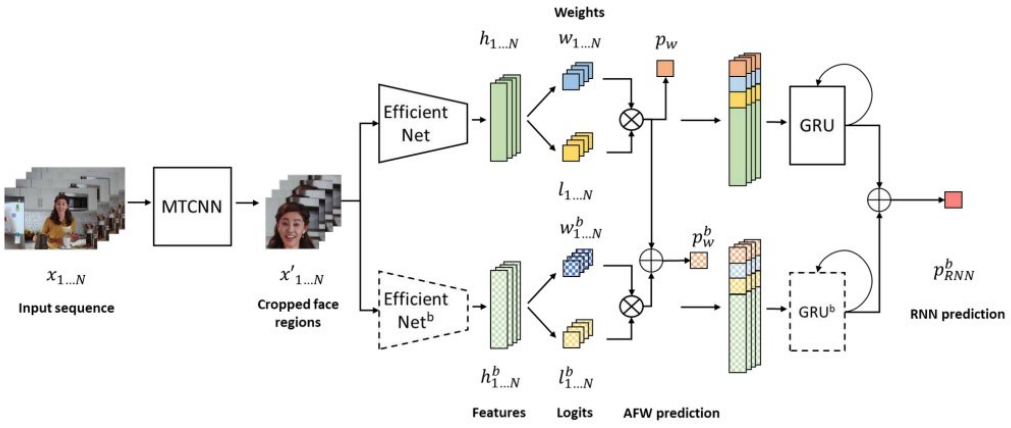

Рисунок 9 — структура системы обнаружения дипфейков с использованием рекуррентной нейросети и системы весов

Кроме того, в данной реализации, LSTM-сеть была заменена на сеть управляемых рекуррентных блоков (англ. GRU — Gated Recurrent Units). По своей идее эти два типа сетей довольно близки, но в GRU фильтры «забывания» и входа объединяют в один фильтр «обновления» . Кроме того, состояние ячейки объединяется со скрытым состоянием, есть и другие небольшие изменения. Построенная в результате модель проще, чем стандартная LSTM, и популярность ее неуклонно возрастает (рисунок 10).

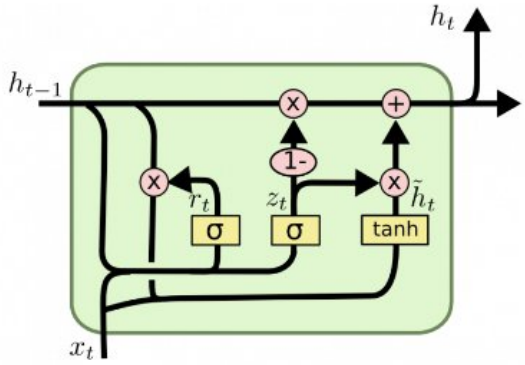

Рисунок 10 — управляемый рекуррентный блок

Упрощенный метод, избегающий использования рекуррентных нейронных сетей, был предложен учеными из Уханя [10]. В представленной системе на вход сверточной сети подается не только кадр исходного изображения, но и оптический поток, т.е., по сути, разность состояний пикселей в соседних кадрах (рисунок 11).

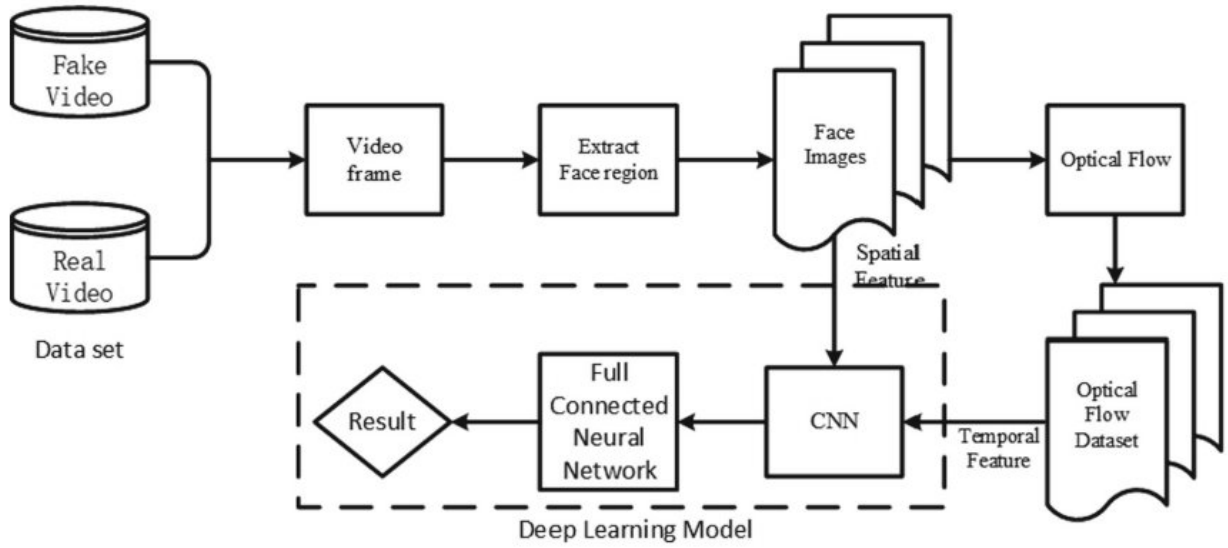

Рисунок 9 — структура системы обнаружения дипфейков с использованием оптического потока

Еще один вариант системы, использующей два потока вычислений, был представлен в Институте информационных наук Южной Калифорнии [11]. Структурная схема системы представлена на рисунке 10.

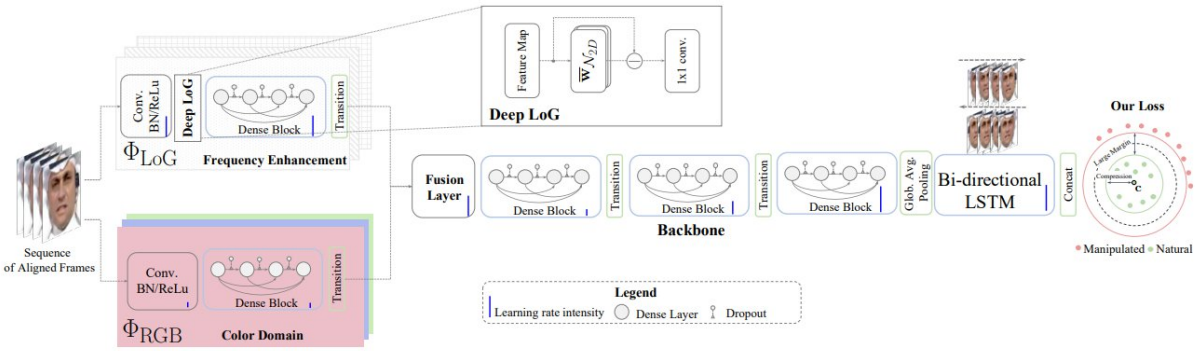

Рисунок 10 — структура системы обнаружения дипфейков с использованием оператора LoG

Данное техническое решение сначала раскладывает изображение на цвета, а параллельно ищет границы предметов с помощью оператора LoG (Лапласиан от фильтра Гаусса, оператор Марра-Хилдрета — см. рисунок 11).

Рисунок 11 — определение контуров лица с помощью оператора LoG

Затем, данные о цветном изображении и о границах единожды проходят через блок свертки отдельно, объединяются, проходят еще несколько блоков свертки. После этого производится анализ изменения изображения во времени с помощью LTSM-сети. Использование такой архитектуры должно позволить на основе информации о контурах лица выделить зоны, подлежащие наиболее тщательно анализу на цветных слоях изображения.

Кроме того, исследование подтверждает, что точность анализа видео значительно повышается с ростом числа анализируемых кадров: даже если анализ отдельных кадров или применение LSTM-сети на малом фрагменте видео дают неточный результат, то агрегация вероятности наличия признаков дипфейка по всем кадрам видео дает более показательную оценку (рисунок 12).

Рисунок 12 — видимость признаков дипфейка в разных кадрах видеоряда

На представленном рисунке моменты, когда система “уверена”, что кадр – изображение реального человека — зеленые, что на видео фейк — красные, вероятность равна примерно 0.5 — серые. Очевидно, что значительную часть времени система затрудняется определить тип отдельного кадра, однако анализ видеоряда в целом дает куда большую точность.

Несмотря на все произведенные улучшения, алгоритмы, основанные на временной согласованности все еще нуждаются в улучшении, так как они чувствительны к дрожанию камеры и изменению сцены.

3. Методы на основе визуальных артефактов

В большинстве случаев генерации дипфейков изображение сгенерированного лица может быть смешано с фрагментами фона или одного из исходных лиц (рисунок 13). При использовании качественных моделей эти дефекты могут быть неразличимы человеческим глазом, однако они все равно могут быть проанализированы алгоритмически.

Рисунок 13 — визуальные артефакты при генерации дипфейка

Первым типом визуальных артефактов являются артефакты, искажающие форму лица. Так как при генерации дипфейка исходное изображение проходит афиннное преобразование (рисунок 14) для получения нужного положения лица, возникают искажения цвета, формы и качества в определенных участках изображения.

Исследователи из Университета штата Нью-Йорк в Олбани используют для поиска таких артефактов понятие RoI — region of interest — регион интереса [12]. RoI — зоны лица, наиболее подверженные искажению при генерации дипфейков. Исследователи с помощью технологий компьютерного зрения выделяют необходимые зоны, а затем анализируют их с помощь сверточных нейросетей.

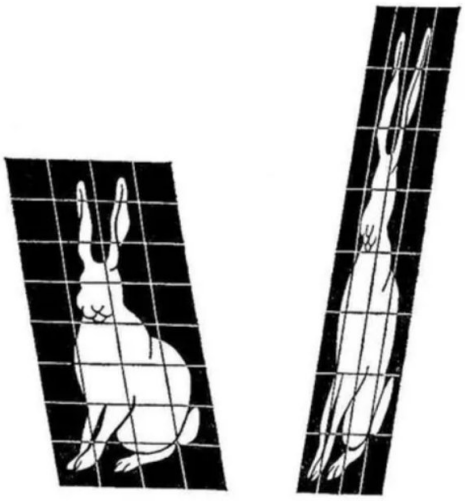

Рисунок 14 — аффиное преобразование изображения

Так как этот метод не требует анализа во времени, он куда более устойчив к изменениям сцены в видео и может быть применен для анализа статических изображений.

Исследователи из Microsoft research предлагают несколько другой метод определения дипфейков на основе визуальных артефактов, который они называют “лицевым рентгеном” [13]. Метод является универсальным для различных реализаций генераторов дипфейков, более того, модель может быть обучена вообще без использования образцов дипфейков, только на изображениях, подвергшихся более простым преобразованиям. Такая

универсальность обусловлена тем, что по сути модель должна лишь определять границу склейки двух изображений (или ее отсутствие) и конкретный механизм склейки не имеет большого влияния на точность классификации. Границ склейки определяются на основе различий в уровне шумов в разных зонах изображения (рисунки 15-16).

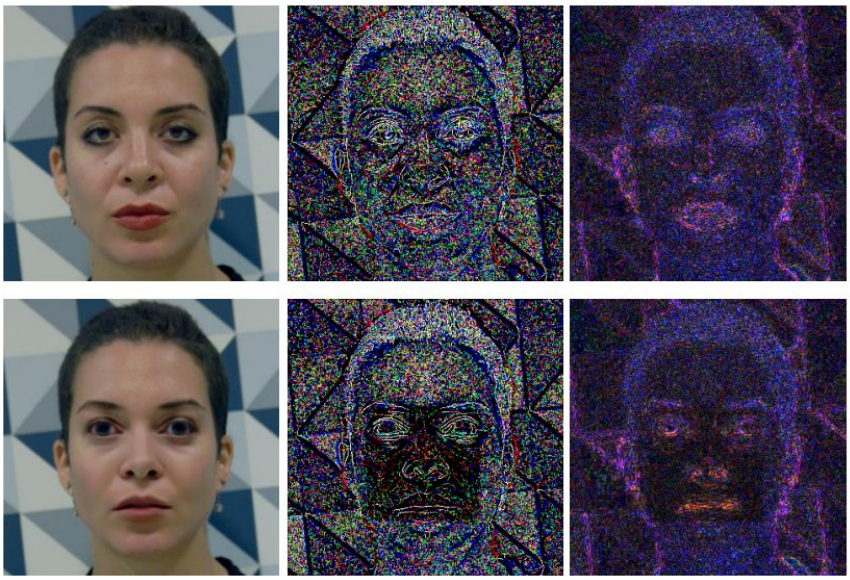

Рисунок 15 — уровень шумов на обычном и модифицированном фото

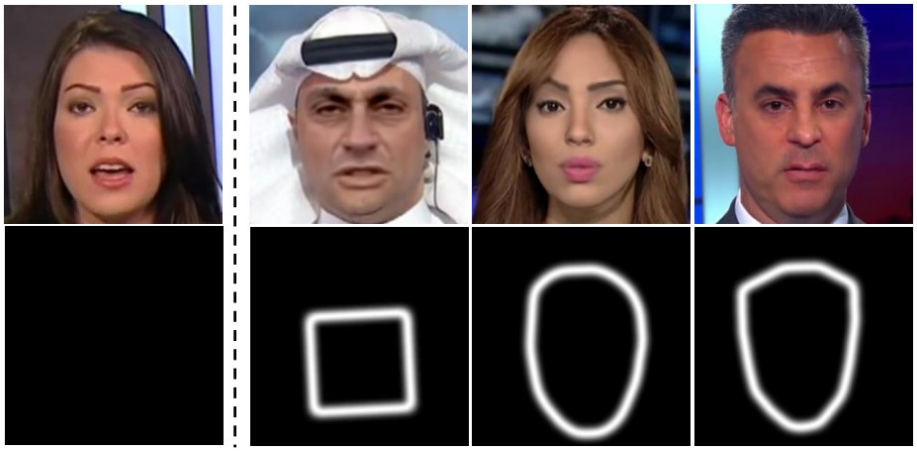

Рисунок 16 — границы наложения изображений

Авторы акцентируют внимание на том, что их метод не анализирует специфические изменения, происходящие непосредственно в зоне лица, при проекции текстуры одного изображения на геометрию другого. Вместо этого они фокусируются на более поздней стадии — склейки измененной части изображения с фоном (рисунок 17).

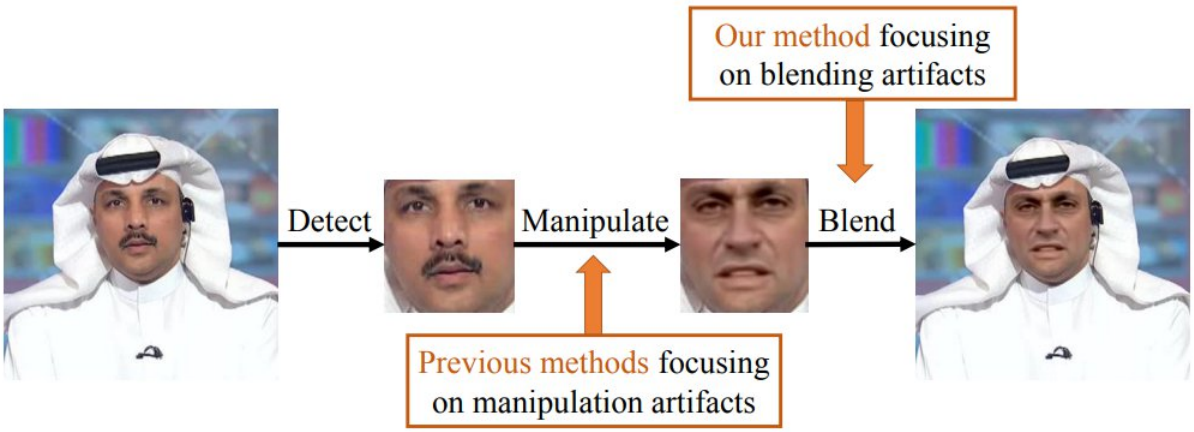

Рисунок 17 — стадии генерации дипфейка

Еще один метод был предложен в университет штата Нью-Йорк в Олбани. Этот метод основан на сравнении направления центральной части лица и краевой зоны лица [14]. Идея состоит в том, что при наложении модифицированной центральной части лица на исходное изображение положение плоскостей центральной и краевой части лица будут отличаться (рисунок 18).

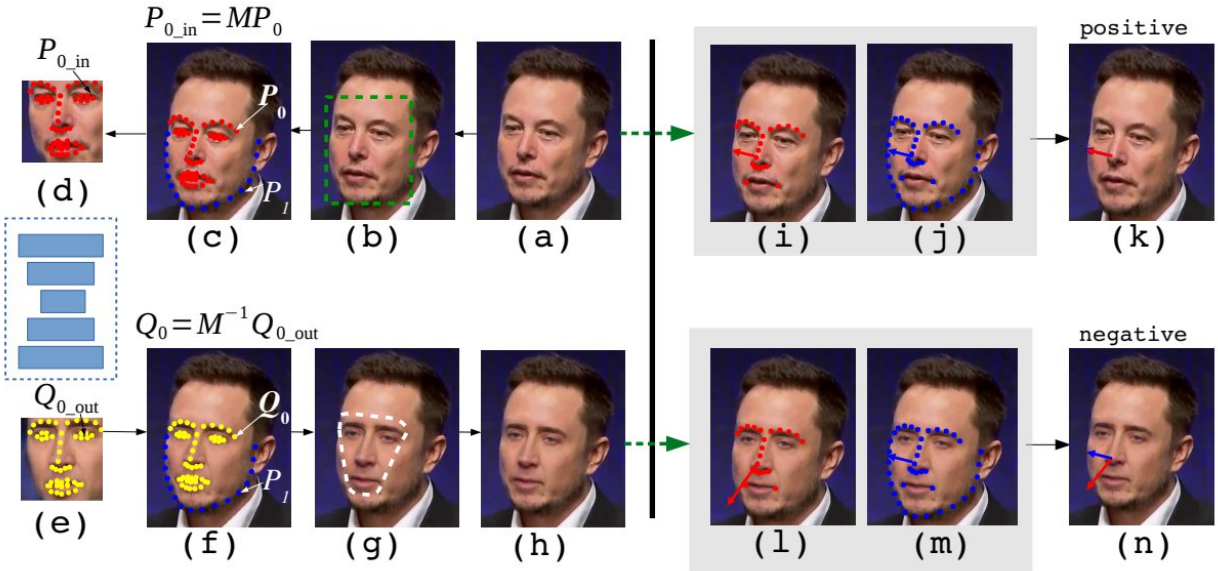

Рисунок 18 — определение положения плоскости лица

На рисунке 19 красным цветом показаны точки, определяющие положение плоскости центральной части лица, синим — плоскости краевой части лица.

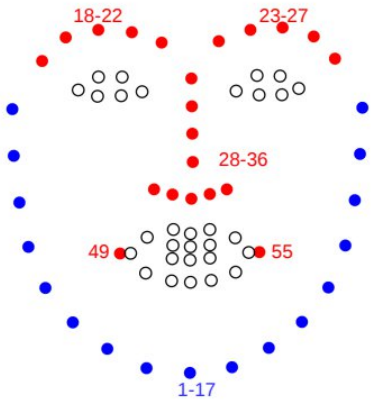

Рисунок 19 — анализируемые точки

На рисунке 20 показаны результаты тестирования модели: по оси абсцисс отложены косинусные расстояния между векторами направления двух плоскостей, по оси ординат — количество кадров с данным расстоянием, результаты измерений для реальных изображений показаны синим, для дипфейков — красным.

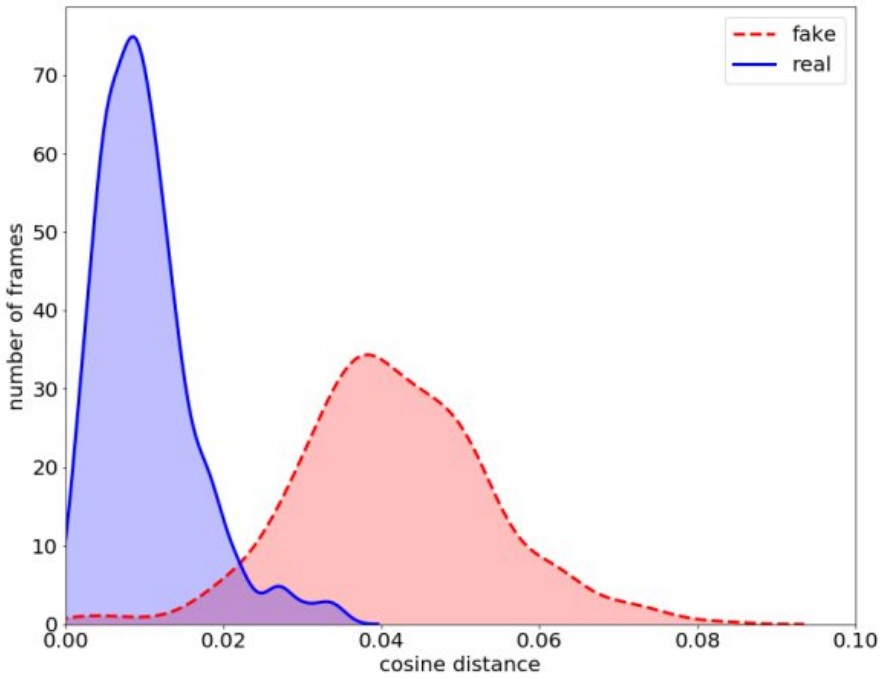

Рисунок 20 — результаты тестирования модели

4. Методы на основе “отпечатков” камеры

Под “отпечатком” камеры понимается слабый шум, возникающий в изображении из-за аппаратных особенностей использованной фото- и видеотехники и играющий большую роль в форензике.

Разновидностями этого типа методов являются:

- анализ неоднородности изображения (photo response nonuniformity — PRNU);
- анализ паттернов шума в отдельном кадре;
- анализ паттернов шума в видеоряде.

Модель, разработанная исследователями из Нидерландов использует PRNU. Идея данного метода состоит в том, чтобы взять несколько последовательных кадров видео и определить корреляцию паттернов шума в них [15]. Так как в реальных видео рисунки шума будут зависеть, среди прочего, от аппаратных особенностей видеокамеры, которые остаются постоянными от кадра к кадру, корреляция между ними будет выше, чем в случае дипфейков, где рисунок шума будет произвольным в каждом кадре. Нормированные значения корреляции для реальных (синий) и модифицированных (красный) видео показаны на рисунке 21.

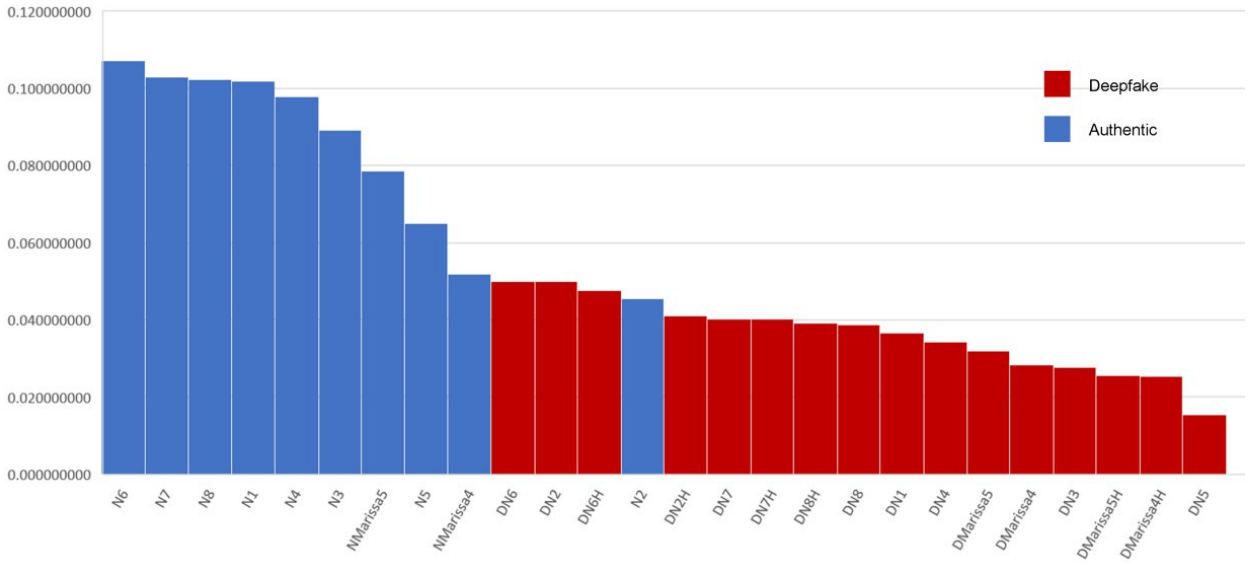

Рисунок 21 — Нормированные значения корреляции для реальных и модифицированных видео

Логическим продолжением этого метода было использовать для определения шума не сравнение последовательных кадров, а нейронные сети [16]. На первом этапе для извлечения шума используется сверточная нейронная сеть, которая обучается на наборе зашумленных фото, а в качестве

обучающего набора выходных данных представлены “отпечатки” исходного шума (зашумленные фото сгенерированы из исходных посредством наложение заранее известных “отпечатков” шума). Данный процесс представлен на рисунках 22-23.

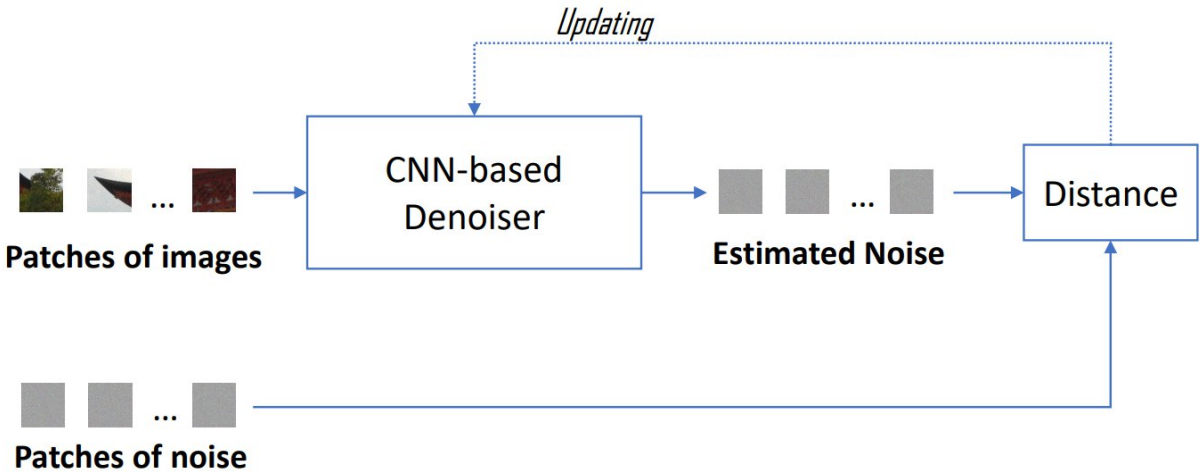

Рисунок 22 — обучение нейронной сети извлечению паттернов шума

Рисунок 23 — изображение и его паттерн шума

Затем полученный паттерн шума обрабатывается сиамской нейросетью для определения класса фото: “дипфейк” или “реальное изображение” (рисунок 24).

Сиамские нейросети - это класс нейронных сетей, которые используются для сравнения двух объектов. Сиамские нейросети состоят из двух идентичных подсетей, которые имеют одинаковую архитектуру и веса. Каждая из этих подсетей принимает на вход один из объектов, которые нужно

сравнить. Затем, используя общие веса, сети вычисляют векторные представления для каждого из объектов.

Далее, эти векторы сравниваются с помощью некоторой метрики расстояния, например, евклидова или косинусная метрика. Если расстояние между векторами мало, то это означает, что объекты похожи друг на друга. Если же расстояние большое, то объекты различны.

Сиамские нейросети широко используются в задачах распознавания образов, идентификации лиц, поиска дубликатов и т.д. Они позволяют получать более точные результаты, чем традиционные методы сравнения объектов [17].

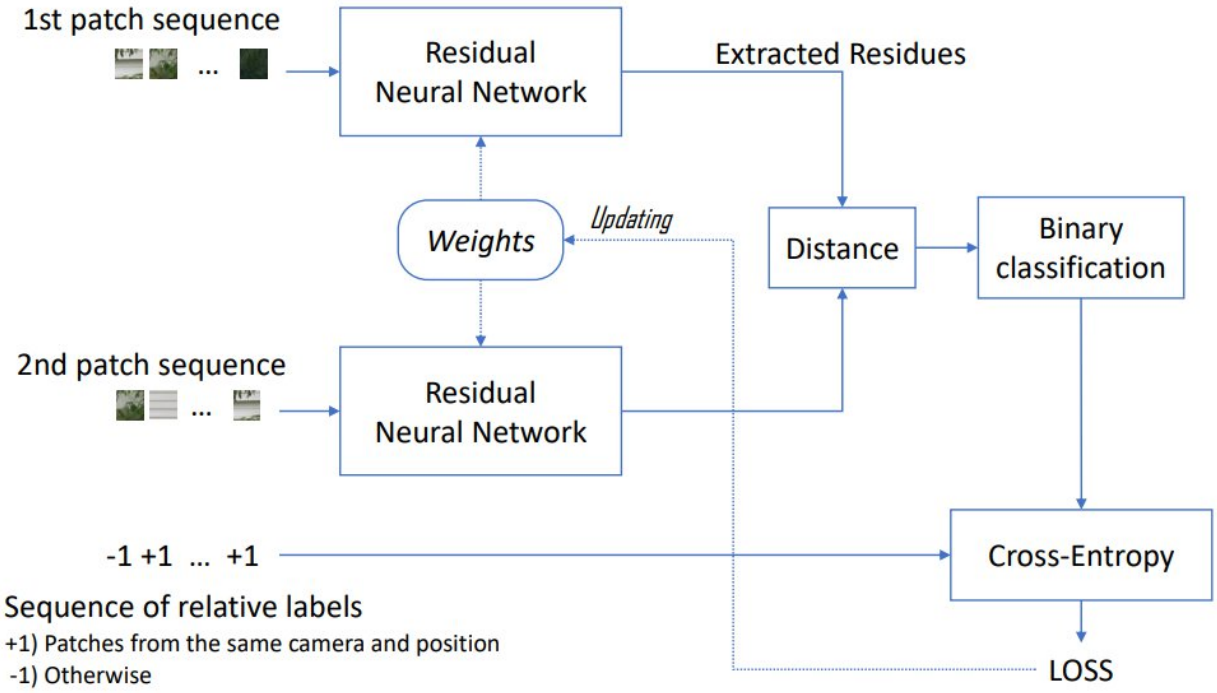

Рисунок 24 — обучение сиамской нейросети

Однако, сиамские нейросети могут лишь определять принадлежность фрагментов проверяемого изображения к одному или различным исходным. Для применения их к задаче определения дипфейков необходимо сравнивать паттерны шума именно в зоне лица с остальной частью изображения. Для этого к рассмотренной системе был добавлен модуль, определяющий положение лица на изображении (рисунок 25).

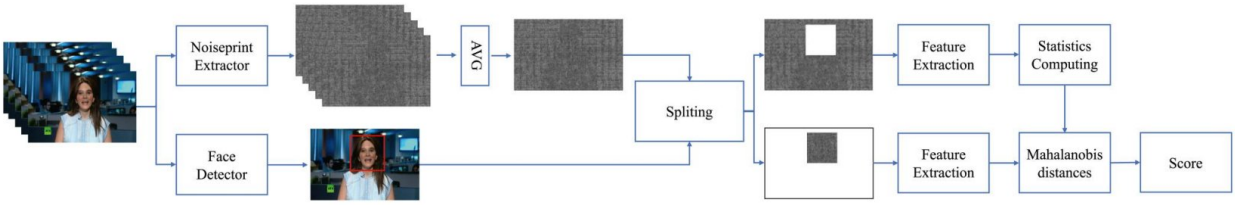

Рисунок 25 — применение сиамской нейросети для определения дипфейков

Стоит отметить, что методы на основе “отпечатков” камеры неэффективны, если модифицированное изображение было дополнительно зашумлено или напротив предпринимались усилия по удалению из него шума.

5. Методы на основе биологических показателей

Несмотря на то, что генеративно-состязательные нейросети способны генерировать дипфейки с высоким уровнем реалистичности, подделать биологические сигналы, такие как моргание и сердцебьение все еще тяжело.

Одним из способов распознавания дипфейков является анализ частоты моргания, так как нейросети, генерирующие дипфейки, в основном обучаются на изображениях с открытыми глазами и плохо справляются с воспроизведением моргания.

Для анализа видео применяется сверточная нейросеть с долгосрочной памятью (LRCN — Long-term Recurrent Convolutional Network), которая является комбинацией нескольких модулей LSTM [18]. Пример применения LRCN показан на рисунке 26.

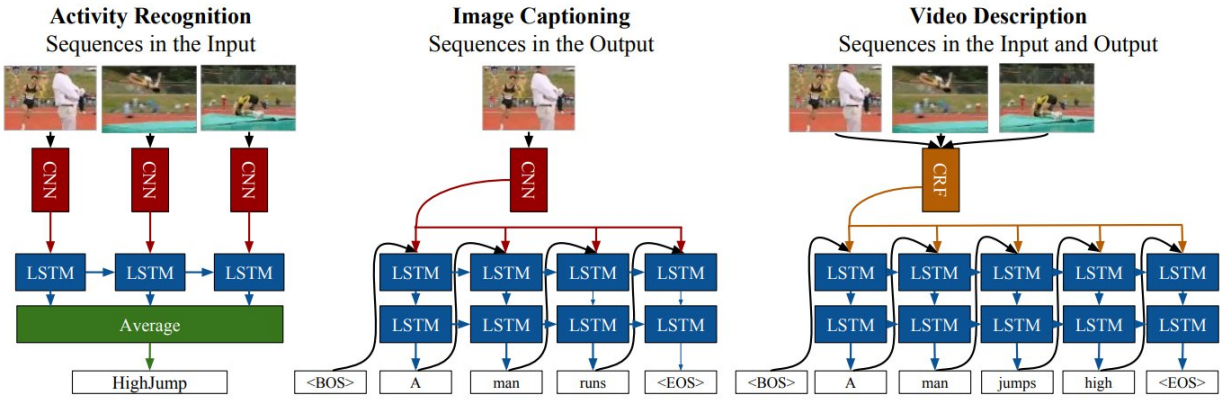

Рисунок 26 — использование LRCN

Для распознавания дипфейков разработанная исследователями из Беркли система сначала определяет положение глаз, затем определяет закрыты они или открыты, после этого определяется и анализируется паттерн моргания во времени (рисунок 27).

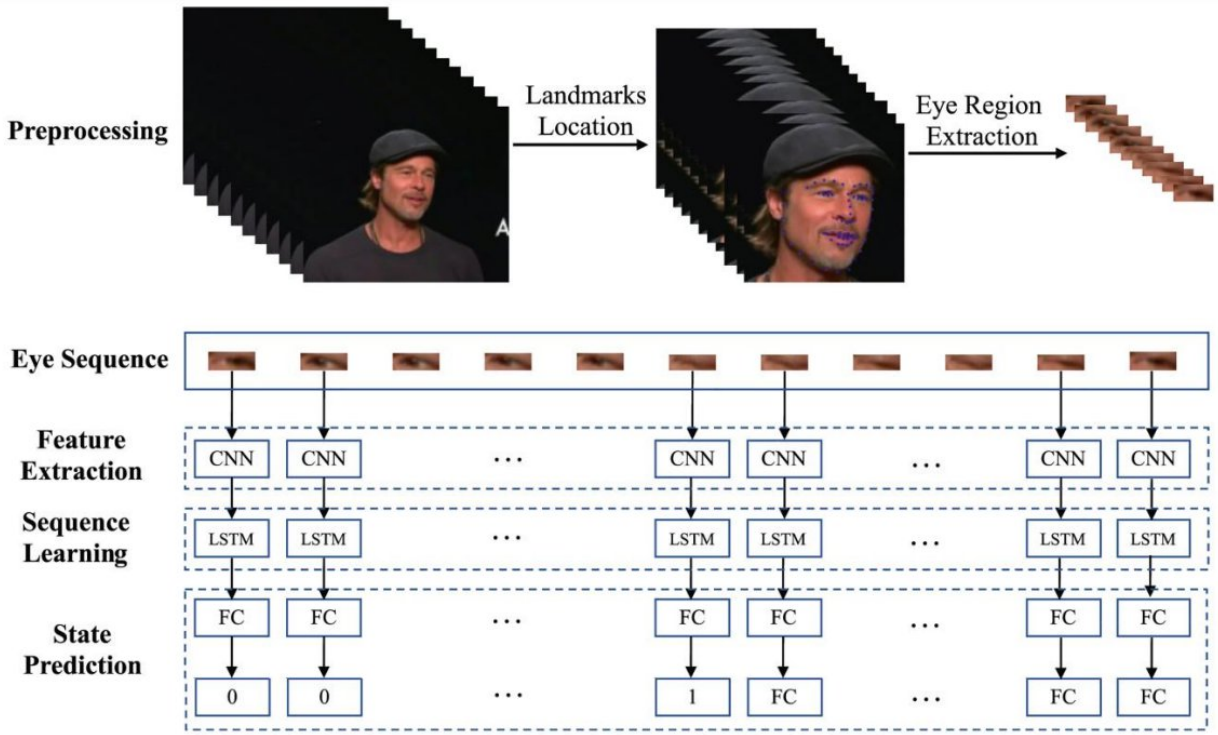

Рисунок 27 — рапознавание дипфейков с помощью LRCN

Основной проблемой этого метода является тот факт, что его точность значительно уменьшается, если для обучения нейросети, сгенерировавшей дипфейк использовалось достаточно изображений с закрытыми глазами, что происходит все чаще, так как изначально метод казался перспективным и создатели генераторов дипфейков приложили значительные усилия для борьбы с ним.

Еще одним методом обнаружения дипфейков является анализ частоты сердцебиения человека на видео с помощью фотоплетизмографии. Фотоплетизмография — это технология, позволяющая определить прилив крови в капилляры оптическим методом, что делает ее применимой для анализа частоты пульса по видео. Для фотоплезмографии по видео применяются методы iPPG и rPPG (рисунок 28) [19].

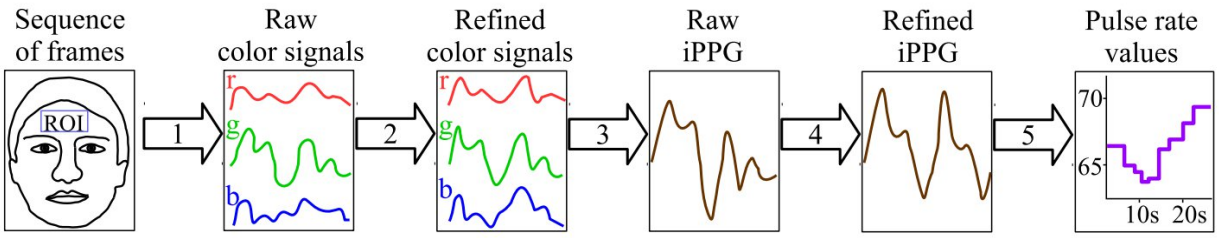

Рисунок 28 — фотоплезмография по методу iPPG

В ходе процесса вычисления частоты пульса выделяются “зоны интереса” (ROI) на изображении — обычно, это все лицо или зона скул и щек. Затем извлекаются цветовые составляющие изображения. После этого временная диаграмма цветовых составляющих очищается, например, с помощью скользящего среднего. Затем непосредственно измеряется частота пульса с помощью одного из следующих методов:

- G-метод — оценивает частоту пульса по изменению зеленого компонента;
- GRD-метод — оценивает частоту пульса по изменению разницы между интенсивностью зеленого и красного компонента (подразумевается, что красный сигнал содержит только артефакты);
- aGRD-метод — оценивает частоту пульса по изменению разницы между интенсивностью зеленого и красного компонента с учетом отличия их предобработанных значений от изначальных;
- Методы CHROM и POS — использует метрику стандартного отклонения.

После получения фотоплезмограммы необходима постобработка для устранения шума. Но фотоплезмограмма по сути содержит лишь информацию о количестве крови в капиллярах, которую надо обработать для вычисления частоты пульса. Для этого используются следующие методы:

- поиск пиков на графике iPPG — самый интуитивно понятный метод, по сути просто считывающий количество приливов и отливов крови,

однако он является весьма неточным из-за зашумленности сигнала iPPG и используется редко;

- использование максимальной спектральной плотности мощности — для расчета используется дискретное преобразование фурье или авторегрессионная модель, которая дает большую точность на небольшом наборе данных;
- использование непрерывного вейвлет-преобразования.

Далее, полученные значения частоты пульса используются для анализа с помощью методов машинного обучения. Разница между характеристиками реального видео и дипфейка в данном случае видна и невооруженным глазом (рисунок 29).

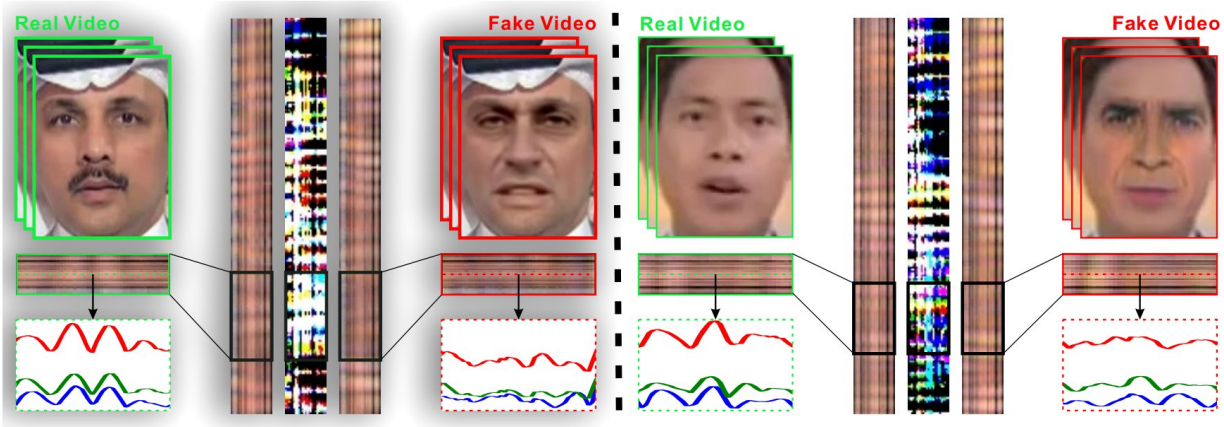

Рисунок 29 — изменение частоты пульса в реальных видео и дипфейках

Хотя подходы к обнаружению дипфейков, основанные на биологических сигналах, показали хорошую производительность на различных наборах данных, естественным недостатком такого рода методов является то, что, на информацию, отражаемую биологическим сигналом, серьезно влияет качество видео, поэтому существуют естественные недостатки и ограниченный диапазон применения подходов, основанных на биологических сигналах.

ЗАКЛЮЧЕНИЕ

Проблема фейковых новостей, видео и фотографий становится все более актуальной в нашем информационном обществе. Дипфейки могут нанести серьезный ущерб как отдельным людям, так и обществу в целом.

Однако, существуют различные методы определения дипфейков, которые позволяют бороться с распространением фальшивой информации.

В данном реферате были рассмотрены основные методы определения дипфейков. Каждый из этих методов имеет свои преимущества и недостатки, и их эффективность зависит от конкретной ситуации. В целом, разработка и совершенствование методов определения дипфейков является важной задачей для обеспечения достоверности информации в нашем информационном обществе.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. A Survey on Deepfake Video Detection / P. Yu, Z. Xia, J. Fei, Y. Lu // IET Biometrics. – 2020. – № 10. – С. 1-18.
2. Habr: Transfer Learning : сайт. – URL: https://habr.com/ru/companies/binarydistrict/articles/428255/ (дата обращения: 05.10.2023).
3. Two-Stream Neural Networks for Tampered Face Detection / P. Zhou, X. Han, L. Davis [и др.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops : электронный журнал. – URL: https://www.researchgate.net/publication/319284586_Two-Stream_Neural_Networks_for_Tampered_Face_Detection (дата обращения: 05.10.2023).
4. Nguyen, H. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos / H. Nguyen, J. Yamagishi, I. Echizen // ICASSP : электронный журнал. – URL: https://www.researchgate.net/publication/332790927_Capsule-forensics_Using_Capsule_Networks_to_Detect_Forged_Images_and_Videos (дата обращения: 07.10.2023).
5. Simonyan, K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // arXiv : электронный журнал. – URL: https://www.researchgate.net/publication/265385906_Very_Deep_Convolutional_Networks_for_Large-Scale_Image_Recognition (дата обращения: 07.10.2023).
6. NeuroHive: CapsNet : сайт. – URL: https://neurohive.io/ru/osnovy-data-science/kapsulnaja-nejronnaja-set-capsnet/ (дата обращения: 07.10.2023)
7. Tarasiou, M. Extracting deep local features to detect manipulated images of human faces / M. Tarasiou, S. Zafeiriou // ResearchGate : электронный журнал. – URL: https://www.researchgate.net/publication/339326111_Extracting_deep_local_features_to_detect_manipulated_images_of_human_faces (дата обращения: 11.10.2023).

8. LSTM – сети долгой краткосрочной памяти // Habr : сайт. – URL: https://habr.com/ru/companies/wunderfund/articles/331310/ (дата обращения: 11.10.2023)
9. Deepfakes Detection with Automatic Face Weighting / D. Montserrat, H. Hao, S. Yarlagadda [и др.] // CVPRW : электронный журнал. – URL: https://www.researchgate.net/publication/343275679_Deepfakes_Detection_with_Automatic_Face_Weighting (дата обращения: 12.10.2023).
10. Capturing the Persistence of Facial Expression Features for Deepfake Video Detection / Y. Zhao, W. Ge, W. Li, R. Wang // Information and Communications Security : электронный журнал. – URL: https://www.researchgate.net/publication/339315933_Capturing_the_Persistence_of_Facial_Expression_Features_for_Deepfake_Video_Detection (дата обращения: 15.10.2023).
11. Two-Branch Recurrent Network for Isolating Deepfakes in Videos / I. Masi, A. Killekar, R. Mascarenhas [и др.] // Computer Vision – ECCV 2020 : электронный журнал. – URL: https://www.researchgate.net/publication/346765088_Two-Branch_Recurrent_Network_for_Isolating_Deepfakes_in_Videos (дата обращения: 17.10.2023).
12. Li, Y. Exposing DeepFake Videos By Detecting Face Warping Artifacts / Y. Li, S. Lyu // ResearchGate : электронный журнал. – URL: https://www.researchgate.net/publication/328736629_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts (дата обращения: 17.10.2023).
13. Face X-ray for More General Face Forgery Detection / L. Li., J. Bao, T. Zhang [и др.] // OpenAccess : электронный журнал. – URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.pdf (дата обращения: 25.10.2023).

14. Yang, X. Eposing deepfakes using unconsistent head poses / X. Yang, Y. Li, S. Lyu // Arxiv.org : электронный журнал. – URL: https://arxiv.org/pdf/1811.00661.pdf (дата обращения: 01.11.2023).
15. Koopman, M. Detection of Deepfake Video Manipulation / M. Koopman, A. Macarulla, Z. Geradts // ResearchGate : электронный журнал. – URL: https://www.researchgate.net/publication/329814168_Detection_of_Deepfake_Video_Manipulation (дата обращения: 05.11.2023).
16. Cozzolino, D. A CNN-based camera model fingerprint / D. Cozzolino, L. Verdoliva // Arxiv.org : электронный журнал. – URL: https://arxiv.org/pdf/1808.08396.pdf (дата обращения: 07.11.2023).
17. Habr: распознавание лиц с помощью сиамских сетей : сайт. – URL: https://habr.com/ru/companies/jetinfosystems/articles/465279/ (дата обращения: 09.11.2023)
18. Long-term Recurrent Convolutional Networks for Visual Recognition and Description / J. Donahue, L. Hendricks, M. Rohrbach [и др.] // Arxiv.org : электронный журнал. – URL: https://arxiv.org/pdf/1411.4389.pdf (дата обращения: 11.11.2023).
19. Unakafov, A. Pulse rate estimation using imaging photoplethysmography: generic framework and comparison of methods on a publicly available dataset / A. Unakafov // Arxiv.org : электронный журнал. – URL: https://arxiv.org/pdf/1710.08369.pdf (дата обращения: 14.11.2023).