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 (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

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

О Т Ч Е Т

по рубежному контролю № 1

Название: Анализ методов определения дипфейков

Дисциплина: Искусство системного инжиниринга и менеджмента организаций

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

Преподаватель _____ Д.В. Березкин
(Подпись, дата) (И.О. Фамилия)

Москва, 2023

СОДЕРЖАНИЕ

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.

Данные 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.

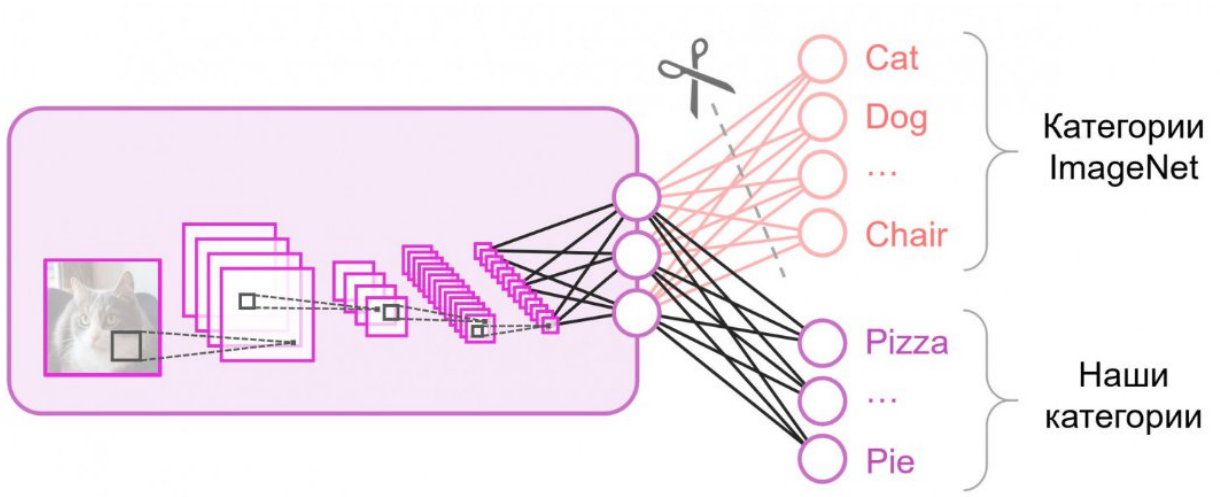

Рисунок 1 — использование технологии трансферного обучения

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

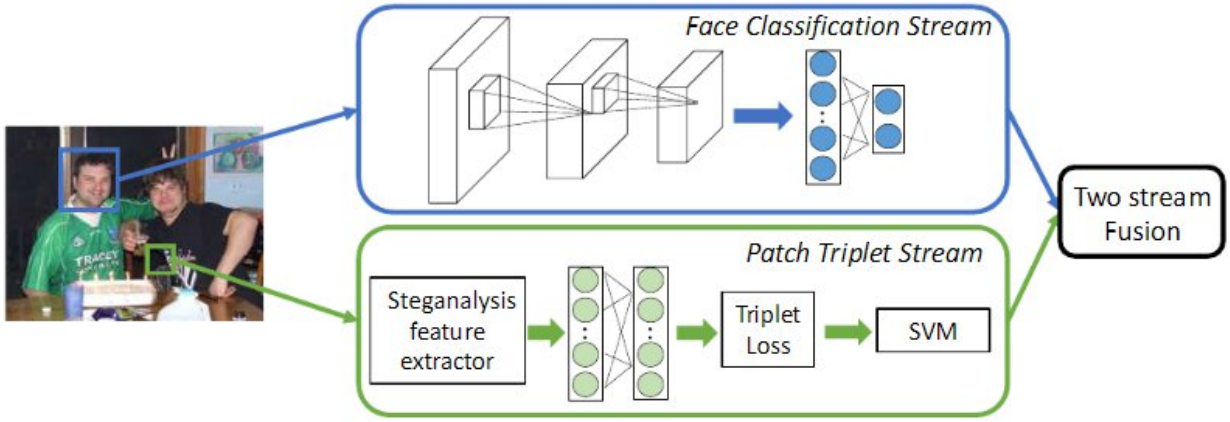

Рисунок 2 — применение трансферного обучения для задачи определения дипфейков

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).

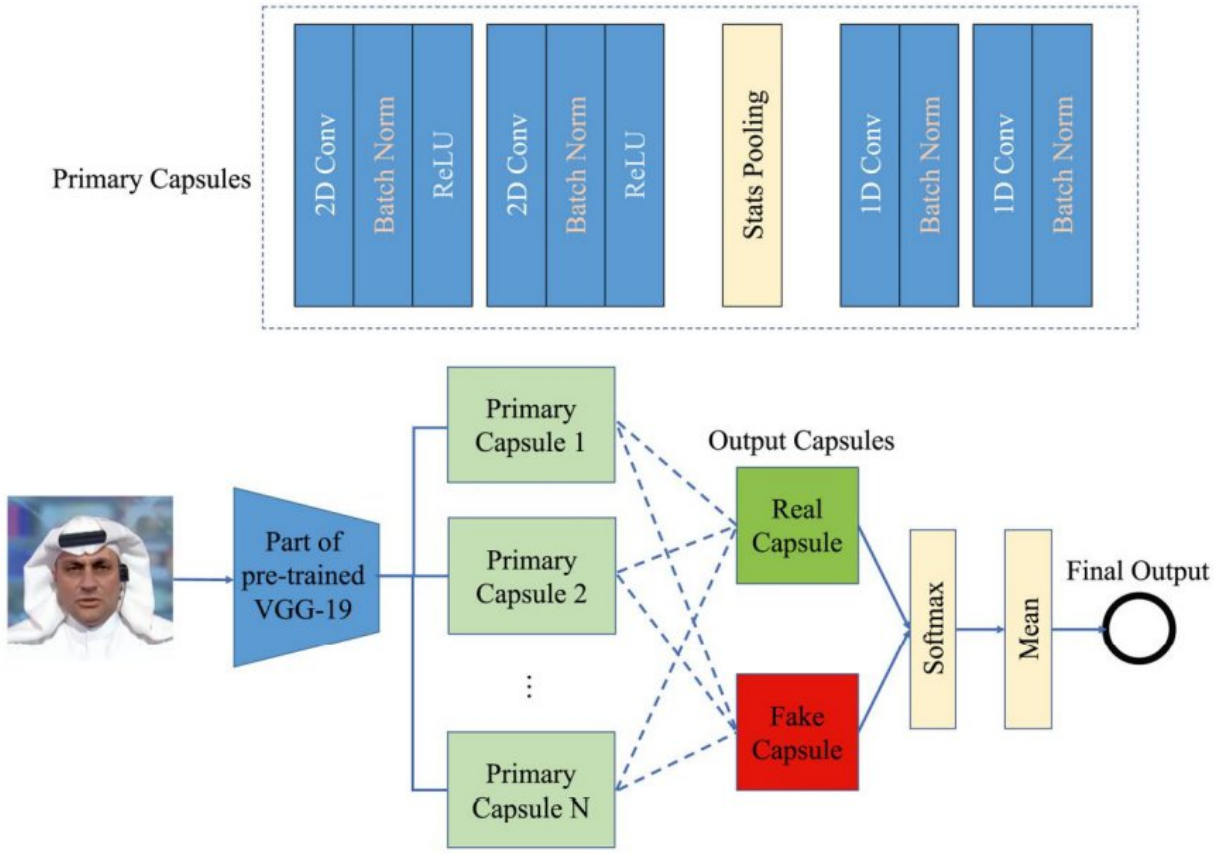

Рисунок 3 — структура системы обнаружения дипфейков с использованием капсульной нейросети

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).

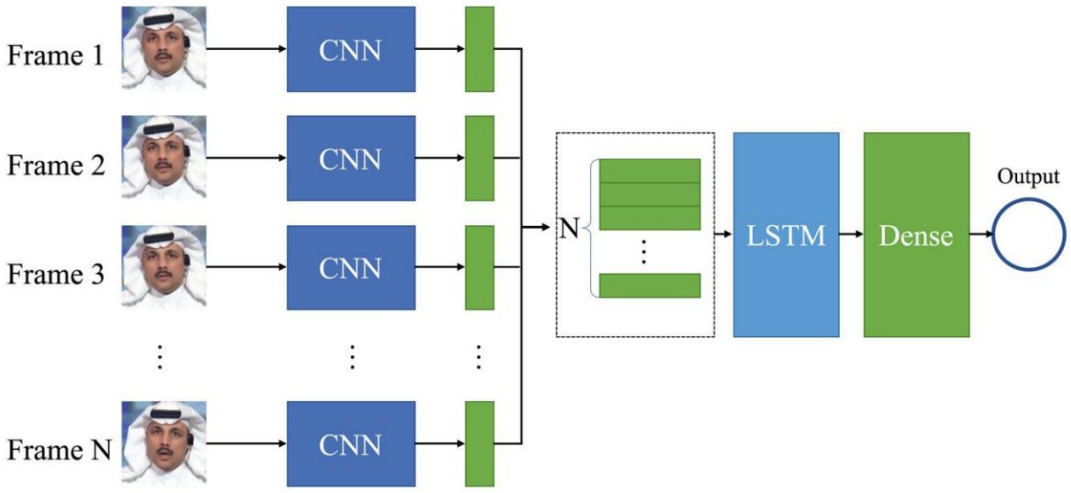

Рисунок 4 — структура системы обнаружения дипфейков с использованием рекуррентной нейросети

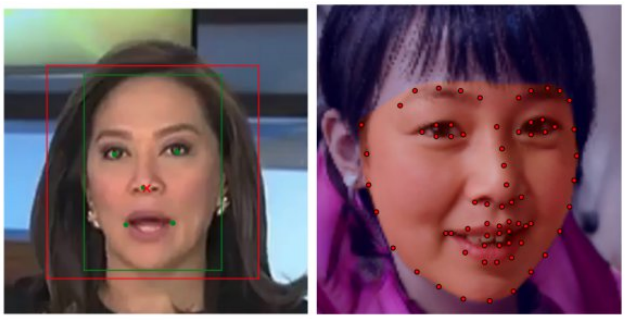

Рисунок 5 — предобработка данных с помощью сверточной неросети

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.

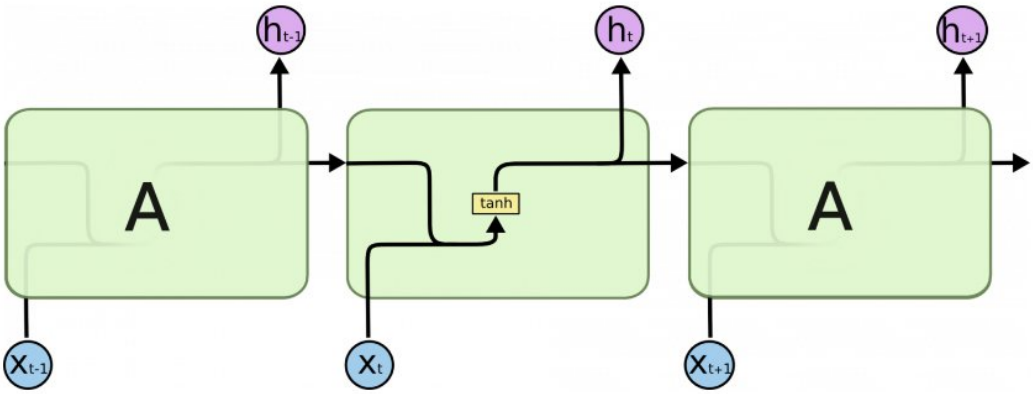

Рисунок 6 — модули рекуррентной нейронной сети

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).

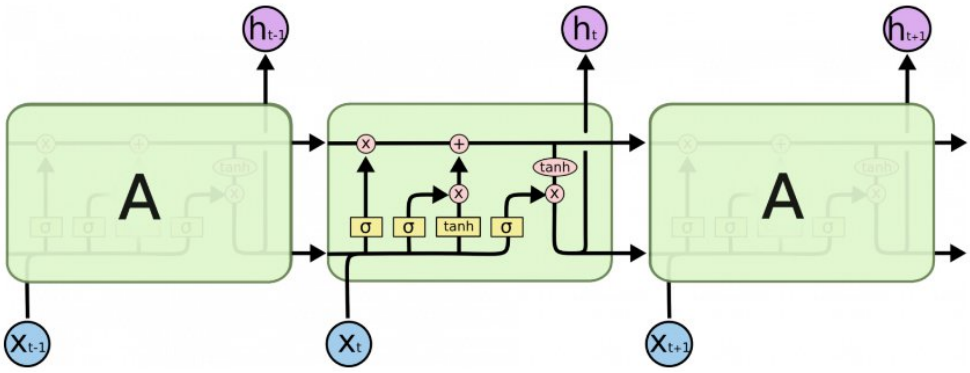

Рисунок 7 — модули LSTM сети

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).

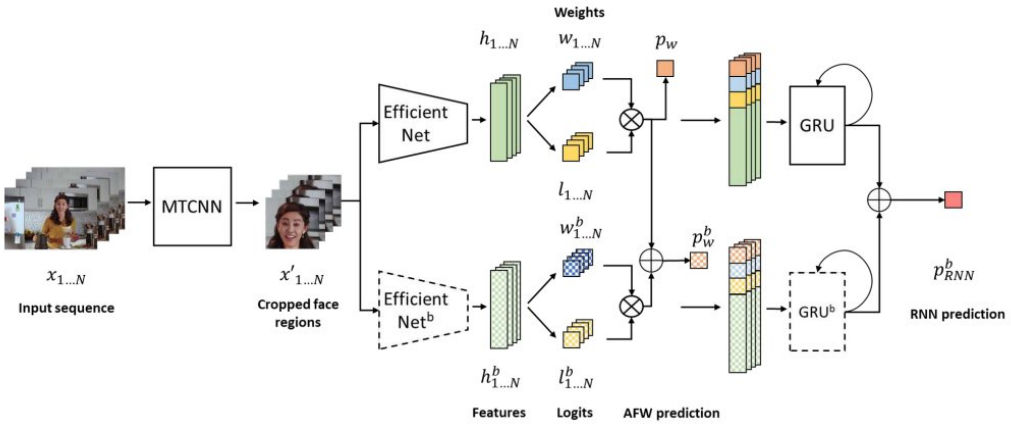

Рисунок 9 — структура системы обнаружения дипфейков с использованием рекуррентной нейросети и системы весов

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).

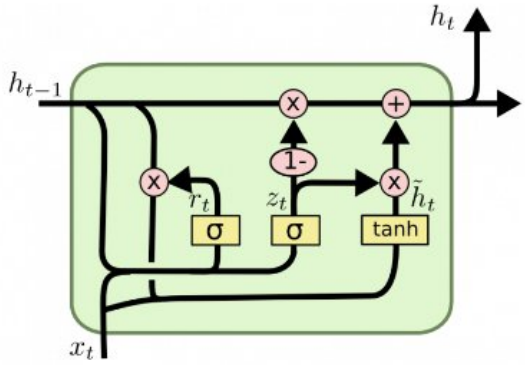

Рисунок 10 — управляемый рекуррентный блок

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).

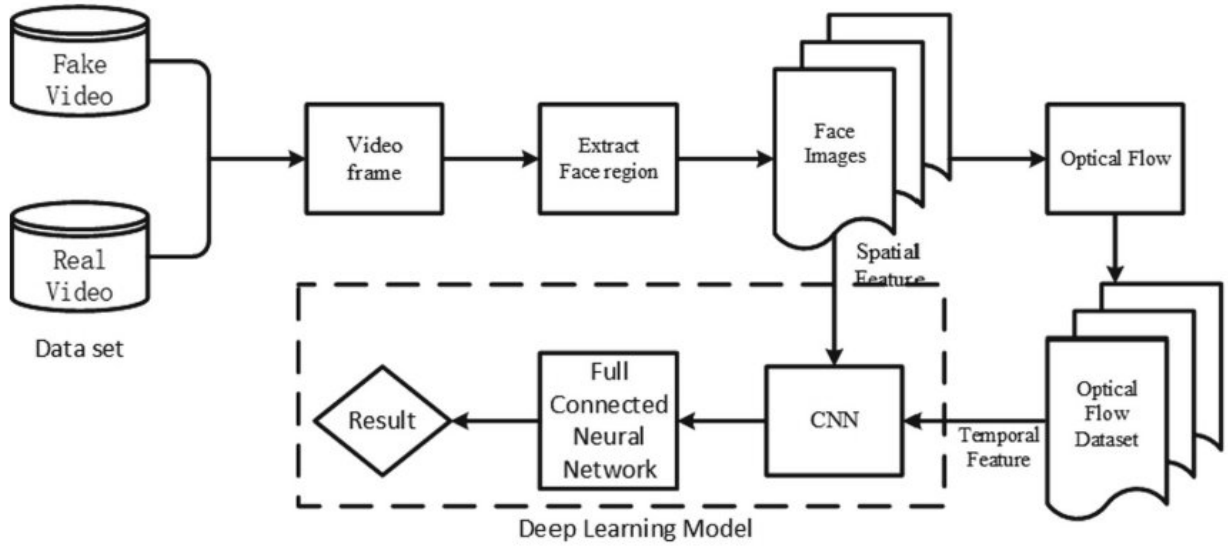

Рисунок 9 — структура системы обнаружения дипфейков с использованием оптического потока

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

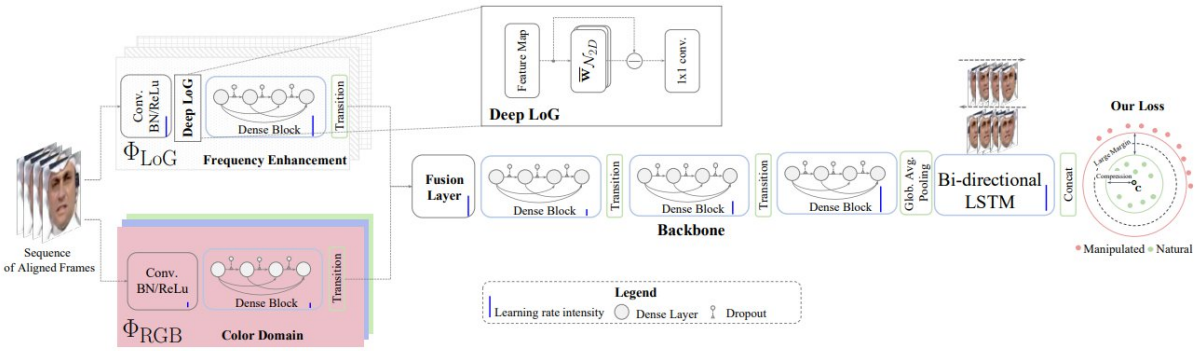

Рисунок 10 — структура системы обнаружения дипфейков с использованием оператора LoG

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.

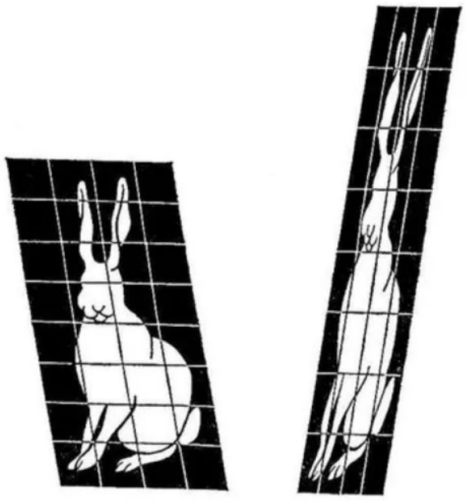

Рисунок 14 — аффиное преобразование изображения

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.

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, такие как анализ пикселей, анализ движения, анализ голоса и другие.

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. https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/bme2.12031
2. https://habr.com/ru/companies/binarydistrict/articles/428255/
3. https://www.researchgate.net/publication/319284586_Two-Stream_Neural_Networks_for_Tampered_Face_Detection
4. https://arxiv.org/pdf/1810.11215.pdf
5. https://arxiv.org/pdf/1409.1556v1.pdf
6. https://neurohive.io/ru/osnovy-data-science/kapsulnaja-nejronnaja-set-capsnet/
7. https://arxiv.org/pdf/1911.13269.pdf
8. https://habr.com/ru/companies/wunderfund/articles/331310/
9. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w39/Montserrat_Deepfakes_Detection_With_Automatic_Face_Weighting_CVPRW_2020_paper.pdf
10. https://wangrun.github.io/paper/ICICS_19.pdf
11. https://arxiv.org/pdf/2008.03412.pdf
12. https://arxiv.org/pdf/1811.00656.pdf