МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение
высшего образования «Самарский национальный исследовательский
университет имени академика С.П. Королева»
(Самарский университет)

Институт информатикии кибернетики
Кафедра геоинформатики и информационной безопасности

ОТЧЕТ ПО ПРАКТИКЕ

Вид практики: производственная практика
(учебная, производственная)

Тип практики: научно-исследовательская работа

Сроки прохождения практики: с 01.09.2024 г. по 09.01.2024 г.
по направлению подготовки 10.05.03 Информационная безопасность
автоматизированных систем
(уровень академического специалитета)
направленность (профиль) «Обеспечение информационной безопасности
распределенных информационных систем»

Студент группы № 6511-100503D Казанцев П.А.

Руководитель практики
от университет профессор Мясников В.В.

Дата сдачи 09.01.2024 г.
Дата защиты 09.01.2024 г.

Оценка _____

Самара 2024

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева»
(Самарский университет)

Институт информатики и кибернетики
Кафедра геоинформатики и информационной безопасности

Индивидуальное задание на практику

Студенту группы № 6511-100503D Казанцеву П.А.

Направление на практику оформлено приказом по университету от 30.08.2023 № 347-ПР на кафедру геоинформатики и информационной безопасности Самарского университета

(наименование профильной организации или структурного подразделения университета)

Планируемые результаты освоения образовательной программы (компетенции)	**Планируемые результаты практики**	**Содержание задания**
ОК-1	Знать: основные направления, проблемы, теории и методы философии, содержание современных философских дискуссий по проблемам общественного развития Уметь: использовать положения и категории философии для оценивания и системного анализа различных социальных тенденций, фактов и явлений и моделирования процессов в научной деятельности; Владеть: - навыками анализа текстов, имеющих философское содержание	Обосновать актуальность проблемы. Изучить основные виды и классификации дипфейков.
ОК-2	Знать: базовые экономические понятия, законы функционирования экономики и поведения экономических агентов, показатели макроэкономического уровня развития страны, экономические показатели,	Рассмотреть архитектуры моделей для генерации и детектирования дипфейк изображений. Изучить математические модели для задач машинного обучения.

	используемые для оценки производственно- хозяйственной деятельности промышленных предприятий Уметь: использовать понятийный аппарат экономической науки для описания экономических и финансовых процессов Владеть: навыками использования экономических знаний в сфере личных финансов и профессиональной деятельности.	
ОК-3	Знать: закономерности и этапы исторического процесса, основные исторические факты, даты, события и имена исторических деятелей России; основные события и процессы отечественной истории в контексте мировой истории Уметь: критически воспринимать, анализировать и оценивать историческую информацию, факторы и механизмы исторических изменений Владеть: навыками анализа причинно-следственных связей в развитии российского государства и общества; места человека в историческом процессе и политической организации общества; навыками уважительного и бережного отношения к историческому наследию и культурным традициям.	Изучить этапы обучения моделей для генерации дипфейков. Рассмотреть признаки изображений лиц, используемые для генерации и детектировании дипфейков. Разновидности и методы обучения для решения задач классификации.
ОК-5	Знать: основные закономерности взаимодействия человека и общества, специфику профессиональной деятельности; основы социологии, структуру общества и социальных институтов; основные этические понятия, историю этических учений, современное положение в сфере этического знания; основные понятия культурологии, типологию культур. Уметь: определять место и роль профессии в социальной сфере, взаимосвязь с другими профессиями; создавать и поддерживать высокую мотивацию к выполнению профессиональной деятельности; ориентироваться в этической проблематике; выявлять основные черты и особенности культурно-исторических ценностей. Владеть: методами выявления мотивов социального поведения; технологиями анализа и прогноза социокультурных	Создать тестовую базу данных. Реализовать алгоритмы для быстрой генерации дипфейк изображений на языке программирования Python.

	процессов для решения практических профессиональных проблем.	
ОК-9	Знать: основные средства и методы физического воспитания Уметь: выбирать и применять методы и средства физической культуры для совершенствования основных физических качеств Владеть: навыками использования методов и средств физической культуры для обеспечения полноценной социальной и профессиональной деятельности.	Разбить полученное задание на подзадачи и последовательно их выполнять. Продемонстрировать промежуточные этапы исследований.
ОПК-6	Знать: правила поиска и содержание основных нормативно-правовых документов регулирующих работу в области обеспечения информационной безопасности информационных систем Уметь: разрабатывать локальные и объектовые нормативно-правовые документы для обеспечения нормативно-правового сопровождения работ по обеспечению информационной безопасности на предприятии. Владеть: навыками систематизации и выбора необходимой нормативно-правовой информации согласно поставленным задачам в области обеспечения информационной безопасности автоматизированных систем.	Проанализировать полученные результаты работы детектирующих моделей. Сделать вывод об использовании старых детектируемых моделей для решения задач классификации.

Дата выдачи задания: 01.09.2023 г.

Срок предоставления на кафедру отчета о практике: 09.01.2024 г.

Руководитель практики от университета, д.ф.-м.н., профессор _____ Мясников В.В.

(подпись)

Задание принял к исполнению студент группы № 6511-100503D _____ Казанцев П.А.

(подпись)

О Т Ч Е Т
о выполнении индивидуального задания
по научно-исследовательской работе

ВВЕДЕНИЕ

Дипфейк (англ. deepfake от deep learning «глубинное обучение» + fake «подделка») – это технология генерации искусственно созданного контента, такого как видео, аудио и изображения с использованием методов глубокого обучения и искусственного интеллекта. У дипфейк технологий есть много положительных сторон. Например, использование в кинематографе или набирающая популярность генерация аудио файлов на разных языках. Но также у этой технологии есть и отрицательные стороны. Злоумышленники могут использовать эту технологию для мошенничества и дезинформации. Политические и социокультурные последствия также вызывают опасения. Важно продолжать исследования и развитие в этой области, чтобы обеспечить баланс между инновациями и обеспечением безопасности и надежности информации.

При прохождении практики по научно-исследовательской работе, руководителем были поставлены следующие задачи:

1) Обзор современного состояния по теме исследования.

2) Поиск базы данных для проведения экспериментов и подготовка существующих программных решений для генерации и детектирования дипфейков.

3) Подготовка с использованием найденных готовых программных решений своей базы данных с дипфейк изображениями.

4) Проведение экспериментального исследования по качеству детектирования дипфейк изображений на общедоступной и своей базе данных.

Задания необходимо было выполнять последовательно в течение всего времени практики, предоставляя руководителю промежуточные отчеты.

ВЫПОЛНЕНИЕ ЗАДАНИЯ

Для выполнения поставленных руководителем задач был проведен анализ тем и обоснована актуальность выбранной темы. Определены цель и задачи исследования.

Целью и задачей исследования стало исследование эффективности детектирования дипфейк изображений с использованием открытых программных решений.

Объектом исследования стали модели для генерации и детектирования дипфейк изображений.

Структура исследования представляет собой следующие этапы:

1) Обзор предметной области.

2) Подготовка базы данных для исследования.

3) Проведение исследования эффективности дипфейк изображений.

4) Анализ результатов исследования.

ОГЛАВЛЕНИЕ

1 Обзор предметной области

В настоящее время тематика дипфейк технологий обрела пик популярности. Дипфейк контент становится все более реалистичным и может использоваться во многих областях. В данной работе будут рассмотрены направления исследования дипфейк технологий, отраженные в научных статьях, находящихся в открытом доступе. Для каждого направления описаны некоторые методы для генерации и детектирования дипфейк объектов.

1.1 Введение

Дипфейк (англ. deepfake от deep learning «глубинное обучение» + fake «подделка») – это технология генерации искусственно созданного контента, такого как видео, аудио и изображения с использованием методов глубокого обучения и искусственного интеллекта. У дипфейк технологий есть много положительных сторон. Например, использование в кинематографе или набирающая популярность генерация аудио файлов на разных языках. Но также у этой технологии есть и отрицательные стороны. Злоумышленники могут использовать эту технологию для мошенничества и дезинформации. Политические и социокультурные последствия также вызывают опасения. Важно продолжать исследования и развитие в этой области, чтобы обеспечить баланс между инновациями и обеспечением безопасности и надежности информации.

1.2 Генерация и обнаружения дипфейков

Подходы и направления в области дипфейк генерации охватывают широкий спектр возможностей для создания и манипуляции мультимедийным контентом с использованием различных технологий глубокого обучения и искусственного интеллекта. Каждое из этих направлений представляет собой уникальный аспект трансформации информации, от замены лиц в изображениях и видео до создания уникальных объектов искусства и синтезирования голосов. Далее будет подробно рассмотрены дипфейки изображений.

Дипфейк изображений наиболее распространенная и обширная область для исследований, которая охватывает различные области:

1) Создание подмены лиц: замена лица человека на исходном изображении лицом другого человека.

2) Изменение внешности и стиля: дипфейк технологии позволяют изменять стиль фотографий или внешность людей. Например, фотографии можно "состарить" или "омолодить", удалить шумы, увеличить разрешение или улучшить цветовую палитру. По отношению к людям изменить цвет волос, добавить макияж, очки, бороду и многое другое.

3) Генерация уникальных фотографий: создание фотографии вымышленных людей, синтез арт-изображений, которые могут сымитировать стиль известных художников или создавать уникальные абстрактные произведения.

1.2.1 Создание и детектирование фальшивых фотографий лиц

Как правило для смены лиц между исходным и целевым изображением используют стратегию использования двух пар кодеров и декодеров. В каждой паре обучение происходит на своем наборе изображений и используется одна и та же сеть кодеров. Общий кодер позволяет находить сходство между наборами изображений лиц. Данная задача относительно проста, потому что лица имеют общие черты, такие как глаза, нос и т.д. Данная стратегия изображена на рисунке 1.

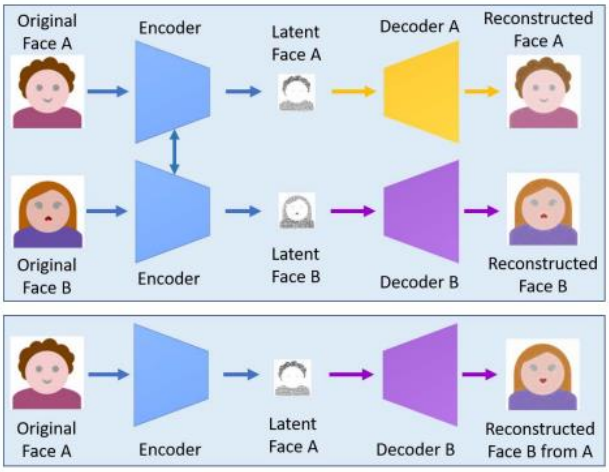

Рисунок 1 – Модель для генерации дипфейк изображения с двумя парами кодеров и декодеров.

Набор признаков лица A соединяется с декодером B для создания фейкового изображения лица B из лица A. Такой подход используется в популярных инструментах для генерации дипфейк изображений: DeepFaceLab, DFaker, DeepFake_tf.

Данная модель была улучшена путем обучения на генеративно состязательной сети (GAN). В случае генерации видео путем добавления функции потери восприятия (perceptual loss) улучшается поведение глаз и сглаживаются артефакты в маске сегментации. GAN модель состоит из двух нейронных сетей: генератор и дискриминатор. Цель дискриминатора классифицировать сгенерированные изображения генератора от реальных изображений. Для этого генератор обучается для минимизации вероятности классификации сгенерированных изображений как синтетических, а дискриминатор для улучшения способности правильной классификации.

В статье [1] рассмотрен метод обнаружения дипфейк изображений основанный на сравнении внутренней области лица, которую изменяют и подменяют, а также внешним контекстом, который остается неизменным, такой как шея, уши, волосы и т.д. Для идентификации замены не нужно предварительно знать личности человека, которого заменили. Сравниваются

представления одной или двух личностей, полученных на основе лица и его контекста на исследуемом изображении.

Представления генерируются двумя специально обученными нейросетями (E_f, E_c), основанными на архитектуре Xception. Первая сеть сопоставляет изображения размером 299x299, содержащие пиксели из области лица в векторе псевдовероятностей, связанном с набором данных лица. Вторая сеть отображает оставшиеся пиксели из ограничивающей рамки обнаружения (контекста) в вектор псевдовероятностей тех же классов. Сети обучались на наборе данных VGGFace2, который содержит около 3.3 миллиона изображений, содержащих более 9000 личностей разных возрастов и представленных в различных позах. Для обучения использовались изображения разрешения больше 128x128 пикселей. Проверка возможности распознавания дипфейков осуществлялась на тестовом наборе данных VGGFace2 и LFW (Labeled Faces in the Wild) – набора данных для верификации лиц, содержащего около 13000 изображений, представляющих реальные сцены и условия, что делает набор данных более репрезентативным. При этом для набора данных из LFW не применялось дополнительное обучение или более тонкая настройка модели. Точность проверок представлена в таблице 1.

Таблица 1 – Точность обученных моделей

Метод обнаружения	Обучающий набор VGGFace2	Тестируемый набор VGGFace2	Последний слой LFW	Предпоследний слой LFW
Контекст	99.90	87.06	98.04	96.79
Лицо	99.89	95.10	99.28	99.48
Все изображение	99.98	96.98	99.59	99.76
Лицо в сравнении с контекстом	-	-	87.49	57.98

Для распознавания лиц обычно используются активации предпоследнего слоя LFW, но они не точно совпадают для одного и того же

человека, потому что сети обучались независимо друг от друга. Для улучшения точности добавляются две бинарные сети (E_s и E_r), также основанные на архитектуре Xception. Одна сеть обучена различать реальные и манипулированные изображения методами подмены лиц, другая является опциональной и обучалась для детектирования изображений, связанных с методами реконструкции лица (поза и выражение лица). Детектирование реконструкций лица не является предметом данной статьи, но оно необходимо для тестирования на бенчмарке FaceForensics++.

Сеть E_s обучалась на подмножестве видео из набора данных FaceForensic++ полученных при помощи методов замены лиц FaceSwap и Deepfakes. Сеть E_r также обучалась на основе видео, но с использованием методов реконструкции лиц: Face2Face и NeuralTextures. При этом использовались сжатые версии этих видеороликов, со сжатием C23 (HQ) и C40 (LQ).

После обучения всех 4 сетей, для сетей E_f и E_c замораживаются веса, чтобы гарантировать, что сигналы идентичности останутся доминирующими. Далее обучается окончательная классификационная сеть на тех же фрагментах видеозаписей из FaceForensic++.

Полученный классификатор может обрабатывать изображения, содержащие несколько лиц, но классифицируются лица, имеющие высоту более 64 пикселей. Остальные лица помечаются как фоновые, за исключением ситуации, когда самое большое изображение принимает меньшую высоту. В таком случае обрабатывается только самый большой обнаруженный объект. Для удаления ложных обнаружений используется пороговая обработка для количества пикселей лица в маске сегментации. Порог по умолчанию равен 15%, если после обработки нет обнаружений, то порог уменьшается вдвое и рассматривается участок лица с максимальным количеством детектируемых пикселей. К одному или нескольким областям применяется составная сеть, результатом которой является оценка для каждой области.

Для проведения экспериментов были задействованы наборы данных из FaceForensics++. Случайным образом были выбраны 1000 пар видео и использованы для создания дополнительных 1000 обработанных видеороликов, представляющих четыре метода изменения лица. Два метода выполняют полную замену лиц: метод замены лиц на основе 3D с использованием традиционного графического конвейера и смешивания, а также метода на основе GAN, который использует изображения пар субъектов для вычисления сопоставления между ними. Два дополнительных метода выполняют реконструкцию лица на основе алгоритма Face2Face, основанном на изменении мимики лица путем изменения коэффициентов экспрессии и алгоритма NeuralTextures, который изучает нейронную текстуру лица из видео и использует ее для реалистичной визуализации реконструированной 3D-модели лица. Результаты теста на бенчмарках FaceForensics++ представлены в таблице 2.

Таблица 2 – Точность обнаружения на бенчмарке FaceForensics++

Метод / Модель	DeepFakes	DeepFakes	FaceSwap	NeuralTextures	Pristine	Общая оценка
Steg. Features	73.6	73.7	68.9	63.3	34.0	51.8
Cozzolino et al.	85.4	67.8	73.7	78.0	34.4	55.2
Rahmouni et al.	85.4	64.2	56.3	60.0	50.0	58.1
Bayar and Stamm	84.5	73.7	82.5	70.6	46.2	61.6
MesoNet	87.2	56.2	61.1	40.6	72.6	66.0
Xception	96.3	86.3	90.3	80.7	52.4	71.0
Исследуемая	94.5	80.3	84.5	74.0	67.6	75.0

Исследуемая модель показала наивысший общий бал, который является значимой метрикой для данного бенчмарка.

1.2.2 Создание и детектирование фейковых атрибутов лица

Эффективным средством для манипуляции лиц на основе GAN моделей является использование скрытого, или латентного пространства. Пространство называют скрытым, так как оно содержит скрытые, сложные признаки данных. Скрытое представление объекта – это вектор, содержащий основную информацию об объекте. Успех такого подхода зависит от врожденного распутывания (disentanglement) скрытых пространственных осей генератора. Но, как правило, при манипуляции лицами предполагают воздействие на локальные области, а обычные генераторы не обладают необходимым пространственным распутыванием. По этой причине, в статье [9] предлагается метод, основанный использовании карт внимания. При этом генерация атрибутов лица происходит на основе текстового описания и обращения к только соответствующим областям.

Данный метод использует латентное пространство StyleGAN2 для редактирования изображений. StyleGAN2 – модель для генерации изображений высокого разрешения. Генератор в StyleGAN2 работает по сверточной модели: в нейронной сети генератора находятся несколько слоёв, каждый из которых больше предыдущего (до требуемого разрешения изображения) и последовательно за предыдущим уточняет изображение, моделируя его и добавляя в него новые детали.

При наличии латентного кода и целевого редактирования обучается сеть, которая предсказывает смещение в латентном пространстве. В процессе обучения изучается карта внимания, чтобы объединить функции, полученные с помощью исходного скрытого кода, с функциями смещенного скрытого кода. Далее обучается модуль внимания, который объединяет особенности всех слоев в единую карту внимания. Карта внимания применяется к объектам целевого слоя во время создания отредактированного изображения. Смешение замаскированных объектов с исходными приводит к манипулированию только намеченными областями. Для управления редактированием используется

нейросеть CLIP, разработанная OpenAI, для выявления связи между текстом и изображением, генерации описаний и сравнении фото. Результат работы алгоритма представлен на рисунке 2.

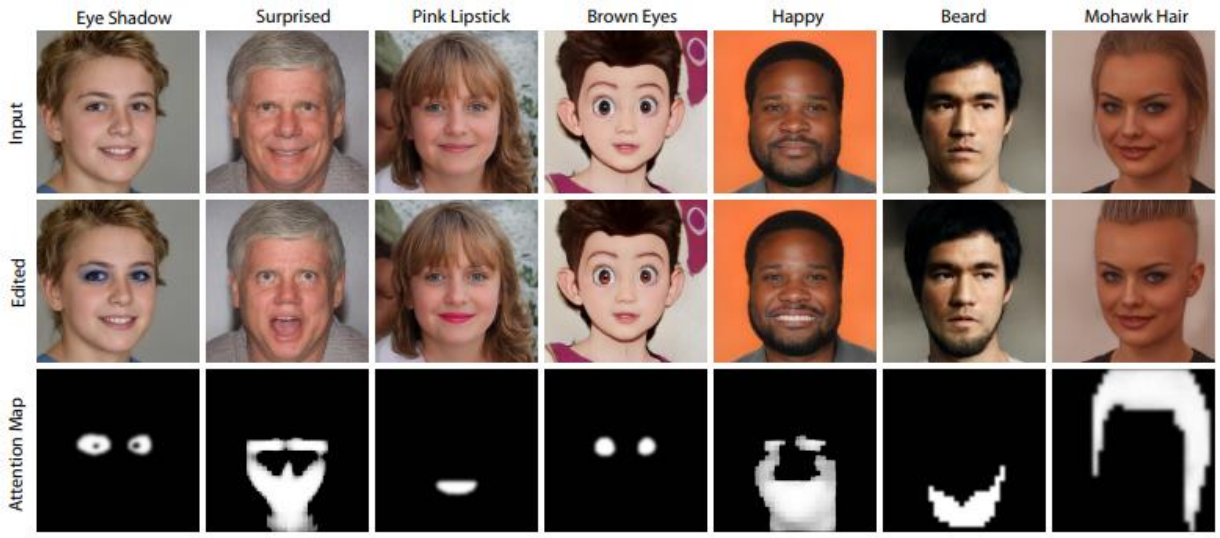

Рисунок 2 – Результат работы алгоритма для входных текстовых атрибутов и карта внимания для каждого изображения

Использование карт внимания позволяет правильно локализовать глаза, губы, бороду и волосы, избегая ненужных изменений в нерелевантных областях. Модуль внимания используется для создания карты вероятностей, определяющей регионы для изменения. Без использования карты внимания при генерации только одного атрибута происходит переплетение нескольких атрибутов лица. Также необходимо учитывать влияние манипуляций в разных слоях при смешивании объектов. Например, для манипулирования цветом подходит 18 слой, а для структурных манипуляций предпочтителен 8 слой.

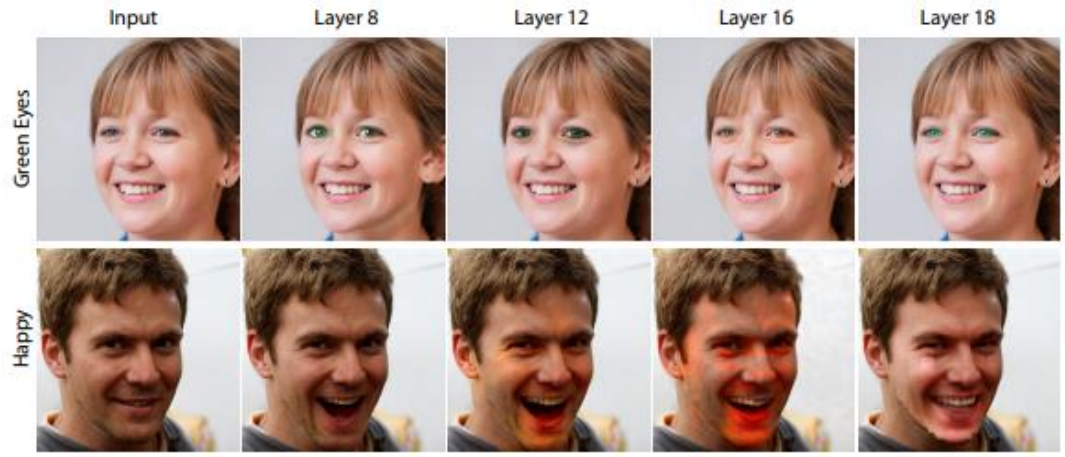

Рисунок 3 – Визуальное сравнение путем смешивания различных векторных слоев

В публикации [10] был произведен анализ существующих методов детектирования изменения лиц. Опираясь на другие работы по данной тематике, автор выявил ряд проблем. Изменение атрибутов лица может привести к увеличению частоты ошибок автоматизированных систем проверки лиц на 50%. При этом коммерческие системы распознавания также теряют свою эффективность на 40% при манипулировании лицами с помощью макияжа и на 60-80% при подделке лиц. Например, электронные паспорта (ePass), содержащие измененные фотографии, могут быть использованы двумя лицами. Кроме того, дети могут изменять образцы своего лица, чтобы выглядеть моложе, а пожилые люди могут редактировать свое лицо, чтобы выглядеть моложе для обмана систем проверки возраста.

Для создания наборов данных использовались бесплатные инструменты, позволяющие накладывать разные фильтры. Для тестирования использовались уникальные атрибуты, которые не доступны в известных наборах данных, таких как DeepfakeTIMIT и FaceForensics++, которые в большей степени содержат примеры для полной подмены лиц.

Для проведения эксперимента использовалось 6 методов обнаружения. Результаты исследования для каждого типа манипуляций приведены на рисунке 3. Базовый метод уступает в точности детектирования

при одной манипуляции, потому что остальные сети имеют больше слоев и обучались на более крупных наборах данных.

Manipulation Type	**DeepFake Detection Method**						Manipulation Type Detection Average
	Baseline	**VGG16**	**SqueezeNet**	**DenseNet**	**GoogLeNet**	**ResNet**	
Age: Cool Old	70.73	98.82	96.47	99.41	97.50	98.82	93.63
Age: Young	61.77	97.94	98.09	98.82	96.76	98.82	92.03
Age: Young2	62.65	98.53	98.67	98.97	97.50	99.26	92.59
Gender	86.77	98.52	95.29	99.85	98.38	99.11	96.32
Glasses	92.06	98.83	96.91	98.97	96.76	98.82	97.06
Hair: Color Black	99.41	99.05	99.85	100.00	96.99	98.38	98.94
HairStyle: Long Hair	99.85	100.00	100.00	100.00	99.70	100.00	**99.92**
Morphing	73.56	84.29	98.23	98.21	73.71	95.10	87.18
Smile	80.29	98.97	97.94	98.87	98.97	99.26	95.71
Tattoo	88.52	99.95	99.85	100.00	97.11	98.24	97.27
Detection Method Average	81.56	97.49	98.13	**99.31**	95.34	98.58	-

Рисунок 4 – Точность методов детектирования изменения лиц.

В качестве базового метода обнаружения использовалась трехслойная модель CNN с тремя двумерными сверточными слоями и тремя полносвязными слоями. Двумерные слои были построены с использованием 64, 128 и 256 сверточными фильтрами размером 5×5 с заполнением нулями и шагом равным 1. Каждый сверточный слой использует функцию активации ReLU и операцию подвыборки Max-pooling. Выходные данные третьего сверточного слоя подаются на полносвязные слои. Первый, второй и третий полносвязные слои состояли из 512, 1024 и 2 нейронов с функцией активации ReLU.

Дальнейшие эксперименты проводились с одновременным использованием нескольких видов манипуляции. При таких перекрестных манипуляциях эффективность детектирования значительно ухудшается. Ни один из тестируемых методов не превысил точность обнаружения более 61%.

2 ПОДГОТОВКА БАЗЫ ДАННЫХ ДЛЯ ИССЛЕДОВАНИЯ

В качестве бенчмарка для детектирующих моделей был выбран бенчмарк FaceForensincs, содержащий 1000 изображений.

Для генерации собственной базы данных было выбрано три нейросетевых модели.

GHOST (Generative High-fidelity One Shot Transfer) – модель, использующая в основе архитектуру FaceShifter, для генерации дипфейк изображений и видео. В качестве модификации для улучшения качества передачи добавлена функция потери для глаз, которая позволяет сохранить направление глаз, как у исходного объекта на изображении. Также добавлен алгоритм сглаживания маски лица, новая техника стабилизации для уменьшения колебаний лиц на соседних кадрах и этап увеличения исходного разрешения до сверх разрешения (super-resolution).

ROOP – модель для генерации дипфейк изображений, основанная на использовании моделей GFPGAN и модель Inswapper от проекта InsightFace. Генерация итогового изображения происходит в два этапа. Сначала используется модель Inswapper для смены лиц. Затем используется модель GFPGAN для улучшения качества и детализации лиц на изображении, полученном после смены лиц.

Encoder4editing (e4e) - кодировщик для манипулирования изображениями StyleGAN. Кодер e4e специально разработан для дополнения существующих методов манипулирования изображениями, выполняемых в скрытом пространстве StyleGAN. Это семантически богатое латентное пространство, которое может быть использовано для выполнения различных манипуляций с изображениями. Чтобы применить такие манипуляции к реальным изображениям, необходимо сначала инвертировать данное изображение в латентное пространство, т. е. получить латентный код, так называемый код стиля. Подача полученного кода стиля в качестве входных данных в обученную модель StyleGAN возвращает исходное изображение.

Чтобы определить правильную реконструкцию, используются два свойства: искажение и качество восприятия. Путем балансировки данных свойств достигается лучшее качество изображения.

В качестве основной базы с изображениями лиц для моделей была выбрана база данных Celebrity-Face-Recognition-Dataset. Это набор данных из 800 тысяч изображений, включающий 1100 известных знаменитостей и класс «неизвестные» для классификации неизвестных лиц. Все изображения были извлечены из браузера Google и не содержат дубликатов. Каждый класс (папка) «знаменитости» состоит примерно из 700-800 изображений, а папка без классификации содержит 100 тысяч изображений. Чтобы подготовить более качественные генерации, выбранные папки с изображениями были отсортированы до разрешения не ниже 400x400 пикселей.

Для проведения тестов было сгенерировано по 100 изображений моделей ROOP и GHOST. Для каждой генерации необходимо два изображения. Первое изображение необходимо для выделения исходного лица, а другое для встраивания этого лица в новый контекст. Для точной оценки детектирующих моделей подобраны уникальные изображения для каждого типа входных данных. Итоговое разрешение изображения равно разрешению исходного изображения.

Рисунок 5 – Примеры дипфейк изображений смены лиц

Для генерации атрибутов лица было сгенерированы изображения с такими атрибутами как: закрытие глаз, улыбка, удаление бороды, раскраска

губ, создание седины (старение) и использование нескольких атрибутов сразу случайным образом. Разрешение результирующего изображения для каждого атрибута равно 1024x1024.

Для каждой модели написаны скрипты для массовой генерации изображений с целью дальнейшего использования в будущих работах.

Рисунок 6 – Пример сгенерированных атрибутов лица

3 ПРОВЕДЕНИЕ ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ ДИПФЕЙК ИЗОБРАЖЕНИЙ

Исследование нацелено на анализ детекторов дипфейк изображений и атрибутов лиц. На выбор модели влияли такие факторы как доступность модели и ограничения на запуск. Большинство статей, описывающих методы детектирования, не содержат ссылки на открытые источники данных исходного кода или полученные модели с необходимыми весами.

Для исследования были выбраны три модели для детектирования дипфейк изображений. XceptionNet и Meso-4 – модели, которые показали один из наилучших результатов в таблице 2. Также используется детектирующая модель, основанная на архитектуре EfficientNetB4.

На вход сети EfficientNetB4 подается квадратное цветное изображение. Вместо использования полного кадра в качестве входного сигнала для повышения точности классификации подается только изображение лица. На выходе сети получается вектор признаков. Далее выбираются карты признаков, извлеченные EfficientNetB4 до определенного слоя, выбранного таким образом, чтобы эти признаки предоставляли достаточную информацию о входном кадре, не будучи слишком подробными. Результирующие карты признаков обрабатываются с помощью одного конволюционного слоя с единичным размером ядра и сигмоидальной функцией активации для получения единой карты внимания. В результирующем этапе умножают карту внимания для каждой из карт признаков. Архитектура модели изображена на рисунке 7, с добавленным механизмом внимания, помеченным красным цветом. Модель обучалась на кадрах из набора данных DFDC. Тренировочная выборка содержала 5000 видео. Точность для тестовых данных из исходного репозитория равна 0.951.

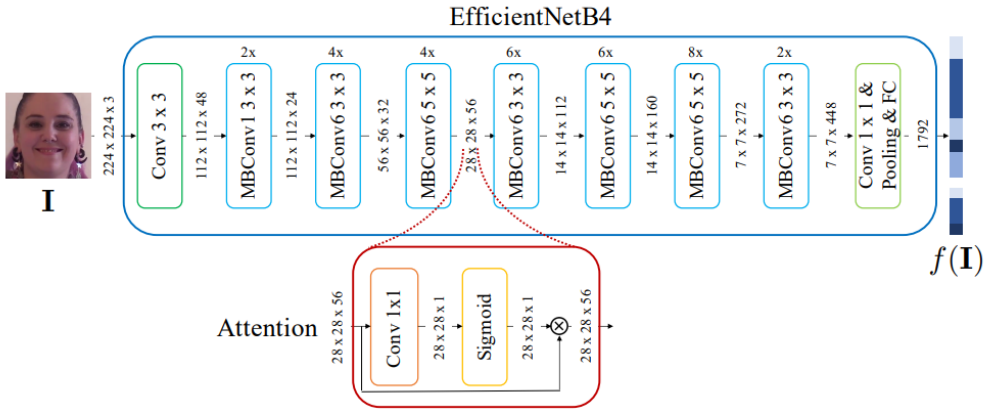

Рисунок 7 – Архитектура EfficientNetB4 с механизмом внимания

Модель XceptionNet обучалась на наборе данных CelebA. Выборка содержит 1600 изображений, по 800 изображений каждого класса. Итоговая точность на тестовых данных достигает – 0.9821. Разрешение обучающих и тестовых изображений 128x128. Слои данной модели изображены на рисунке 8. Архитектура Xception представляет собой линейный стек разделяемых по глубине сверточных слоев с остаточными связями. Это делает архитектуру очень простой для модификаций. Работа модели основана на гипотезе: отображение кросс-канальных корреляций и пространственных корреляций в картах признаков сверточных нейронных сетей может быть полностью развязано. Поскольку эта гипотеза является более сильной версией гипотезы, лежащей в основе архитектуры Inception, она была названа Xception.

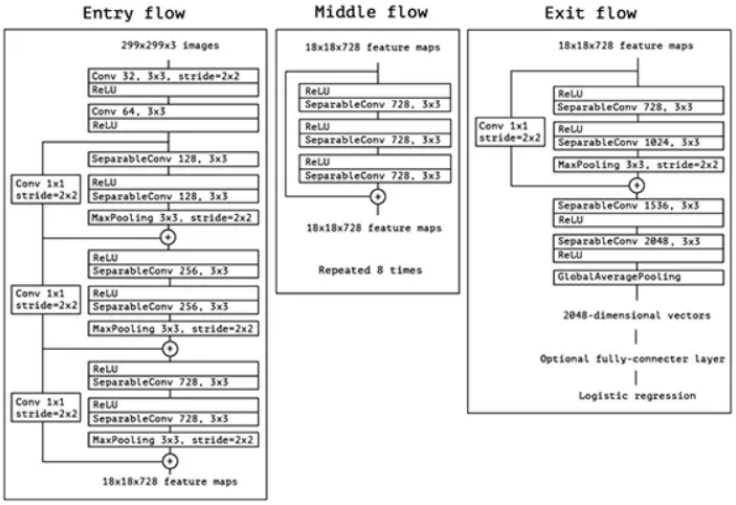

Рисунок 8 – Архитектура модели XceptionNet

Meso-4 – модель, основывающаяся на мезоскопических свойствах изображения. Обучалась на изображениях разного разрешения от 92x92 до 400x400, полученных из видео, находящихся в свободном доступе. Достигла точности равной 0.96. Архитектура данной модели изображена на рисунке 9.

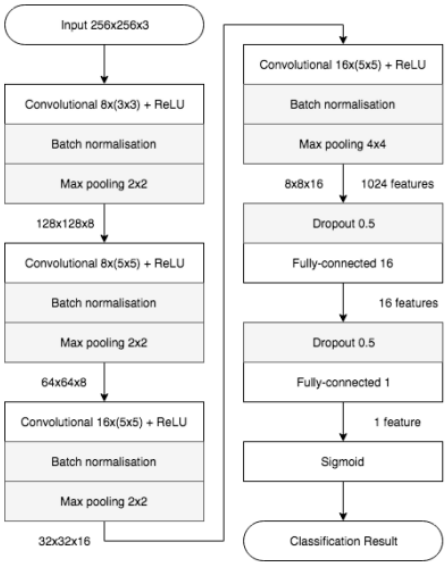

Рисунок 9 – Архитектура Meso-4

Исследование проводилось в три этапа. Каждая модель отдельно тестировалась на бенчмарке FaceForensics, самостоятельно сгенерированных дипфейках полной смены лица и самостоятельно сгенерированных изображениях с измененными атрибутами лица. База данных, содержащая атрибуты лица, включает в себя все манипуляции, перечисленные в разделе 2.

Тестирование проводилось как на общей выборке с применением манипуляций, так и на отдельных наборах данных под каждый вид манипуляции. Вследствие одинаковых показателей под каждый тип манипуляций результаты будут представлены только для общей выборки. Результаты исследований приведены в таблице 3 и 4. Общая точность модели вычислялась с использованием наборов дипфейк данных и реальных изображений с помощью следующей формулы:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positives (TP): Количество изображений, которые были правильно классифицированы как дипфейки.

True Negatives (TN): Количество изображений, которые были правильно классифицированы как реальные изображения.

False Positives (FP): Количество изображений, которые были ошибочно классифицированы как дипфейки.

False Negatives (FN): Количество изображений, которые были ошибочно классифицированы как реальные изображения.

Помимо общей точности модели, приведена точность только для фейковых изображений для более лучшего анализа.

Таблица 3 – Общая точность обнаружения найденных моделей

	FaceForensics	БД полной смены лиц	БД атрибутов лица
EfficientNetB4	0.6974	0.5128	0.4175
Meso-4	0.54	0.51	0.48
Xception	0.68	0.506	0.44

Таблица 4 –Точность обнаружения дипфейк изображений

	FaceForensics	БД полной смены лиц	БД атрибутов лица
EfficientNetB4	0.444	0.1026	0.025
Meso-4	0.10	0.02	0
Xception	0.43	0.0385	0

4 АНАЛИЗ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Исследование выявило, что существующие детекторы дипфейк изображений и атрибутов лиц часто сталкиваются с трудностями при работе с реальными данными. Они могут проявлять чувствительность к качеству изображений, что ограничивает их применимость в реальных сценариях.

Анализ результатов позволяет выделить ключевые проблемы, требования к улучшению и определить направления будущих исследований.

Многие детекторы остаются чрезмерно чувствительными к качеству изображений, что может приводить к неправомерным срабатываниям на низкокачественных или сверхкачественных реальных данных. Также недостаток эффективных моделей с открытым исходным кодом создает проблему доступности для исследователей и разработчиков.

Обучение детекторов на более разнообразных реальных данных станет ключевым фактором для повышения их устойчивости и обобщающей способности. Необходимо продолжать разработку более точных и устойчивых алгоритмов обнаружения дипфейк изображений, способных эффективно работать с реальными данными различного качества.

Подавляющее большинство методов обнаружения дипфейков теряет эффективность при новых способах генерации. Модели обнаружения, как правило, создаются и анализируются на одинаковых бенчмарках. Данный подход способствует высокой точности классификации только на наборах данных, используемых для обучения. Также отсутствие образцов сверхвысокого разрешения накладывает ограничения на некоторые методы детектирования. Дальнейшие направления исследования могут быть связаны с задачей повышения производительности существующих методов детектирования или разработкой собственных алгоритмов детектирования. Одним из подходов для реализации этой задачи может стать задача создания постоянно обновляемого эталонного набора данных дипфейков для проверки новых методов обнаружения. Также при создании собственных методов

детектирования или доработки существующих необходимо использовать собственно сгенерированные синтетические данные.

ЗАКЛЮЧЕНИЕ

В течение практики по научно-исследовательской работе успешно выполнены поставленные задачи: были проанализированы темы в области искусственных нейронных сетей и машинного обучения. Определена и обоснована актуальность выбранной темы. Определена цель и задачи исследований. Определен объект и предмет исследования. Подготовлена кодовая база для массовой генерации дипфейк изображений трех генеративных моделей. Проведено исследование по определению точности классификации с использованием открытых программных решений. Результаты исследования показывают очень низкую эффективность детектирования дипфейк изображений на реальных данных.

За время прохождения практики освоены необходимые компетенции, в частности изучены: теоретические основы нейронных сетей для генерации и детектирования дипфейк изображений; математические модели для задач классификации дипфейк изображений. Дальнейшие исследования будут направлены на разработку собственного метода для распознавания дипфейк изображений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1) Yuval Nirkin; Lior Wolf; Yosi Keller; Tal Hassner. DeepFake Detection Based on Discrepancies Between Faces and Their Context.

2) Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, Bimal Viswanath. Deepfake Text Detection: Limitations and Opportunities.

3) Thanh Thi Nguyena, Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng. Deep Learning for Deepfakes Creation and Detection: A Survey.

4) Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

5) Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 3(1):80–87, 2019.

6) Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 2018.

7) Xin Yang, Yuezun Li and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8261–8265. IEEE, 2019.

8) Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In IEEE Winter Applications of Computer Vision Workshops (WACVW), pages 63–72. IEEE, 2019.

9) Xianxu Hou, Linlin Shen, Or Patashnik, Daniel Cohen-Or, Hui Huang. FEAT: Face Editing with Attention.

10) Zahid Akhtar, Murshida Rahman Mouree, Dipankar Dasgupta. Utility of Deep Learning Features for Facial Attributes Manipulation Detection.

КОД ПРОГРАММЫ

Исходный код программы доступен для скачивания в репозитории GitHub: Dryg1214/DeepFakeNeiroWork (github.com)