УДК 57.087.1
DOI: 10.25559/SITITO.18.202203.680-690

Оригинальная статья

Обзор технологий выявления модифицированного контента класса DeepFake

А. В. Балашов*, К. Е. Вышегородцев, Д. А. Свичкарь, П. В. Хенкин
ПАО «Сбербанк России», г. Москва, Российская Федерация
Адрес: 117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19
* Balashov.A.Vik@sberbank.ru

Аннотация

В работе приводится обзор публикаций об основных технологиях выявления модифицированного контента класса DeepFake, в том числе, список публично доступных датасетов и результаты тестирования решений, имеющихся в публичном доступе. В работе также приводятся результаты независимого тестирования систем обнаружения DeepFake, полученные в ходе открытого конкурса DFDC – 2019 (DeepFake Detection Contest), а также краткие обзоры проводимых подобных конкурсов в 2020-2021 годах в КНР. Статья также описывает новый подход к обнаружению фото/видеоматериалов, созданных с применением технологий DeepFake. Проведенный авторами анализ решений дал возможность создать новый способ выявления поддельного контента, который был запатентован авторами в Роспатенте. Предложенный авторами механизм выявления DeepFake имеет ряд существенных преимуществ по сравнению с решениями, описанными в обзоре, что позволяет рассчитывать на более эффективное обнаружение попыток обхода биометрических систем.

Ключевые слова: дипфейк, атаки на биометрическое предъявление, лицевая биометрия

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Обзор технологий выявления модифицированного контента класса DeepFake / А. В. Балашов [и др.] // Современные информационные технологии и ИТ-образование. 2022. Т. 18, № 3. С. 680-690. doi: https://doi.org/10.25559/SITITO.18.202203.680-690

Original article

Overview of Technologies for Detecting Modified Content of the DeepFake Class

A. V. Balashov*, K. E. Vyshegorodtsev, D. A. Svichkar, P. V. Khenkin
PJSC "Sberbank of Russia", Moscow, Russian Federation
Address: 19 Vavilov St., Moscow 117312, Russian Federation
* Balashov.A.Vik@sberbank.ru

Abstract
The article provides an overview of publications on the main technologies for detecting modified content of the DeepFake class, including a list of publicly available datasets and the results of testing solutions available in the public domain. The article also presents the results of independent testing of DeepFake detection systems, obtained during the DFDC – 2019 (DeepFakeDetectionContest) and gives brief reviews of similar competitions held in 2020-2021 in China. The article also describes a new approach to detecting photo/video materials created using DeepFake technologies. The authors' analysis of the solutions made it possible to create a new way of detecting fake content, which was patented by the authors in the Federal Service for Intellectual Property for Technology. The DeepFake detection mechanism proposed by the authors has a number of significant advantages over the solutions described in the review, which allows us to count on more efficient detection of attempts to bypass biometric systems.

Keywords: DeepFake, biometric presentation attacks, facial biometrics

Conflict of interests: The authors declare no conflict of interest.

For citation: Balashov A.V., Vyshegorodtsev K.E., Svichkar D.A., Khenkin P.V. Overview of Technologies for Detecting Modified Content of the DeepFake Class. *Modern Information Technologies and IT-Education.* 2022;18(3):680-690. doi: https://doi.org/10.25559/SITITO.18.202203.680-690

Введение

Появление большого числа поддельного (модифицированного) контента вызвало необходимость выявления фактов подделки публикуемых материалов, при этом имеется насущная необходимость использования решений, которые позволили бы выявлять DeepFake по возможности быстро и в автоматическом режиме, без привлечения экспертов для вынесения вердикта о характере предъявленного фото/видео материала. Данная работа посвящена изучению существующих и созданию новых подходов к выявлению фото/видеоматериалов, полученных с помощью технологии DeepFake и преимуществ по сравнению с опубликованными результатами исследований других авторов.

Основная часть

С точки зрения уязвимостей систем лицевой биометрии, которые позволяют реализовать угрозы для автоматизированных систем финансовых учреждений, интерес представляют следующие типы манипуляций с изображением лица:

Типы манипуляции			
Подмена лица (DeepFake – Identity Swap) – замена одного лица на другое в фото/ видеоизображении		Манипуляция атрибутами лица – изменение цвета волос или кожи, изменение пола, изменение возраста	Подмена выражения лица – основано на изменении выражения лица
Датасеты	Механизмы детектирования	Механизмы детектирования	Механизмы детектирования
Таб. 1	Таб. 2	Таб. 3	Таб. 4

Об актуальности этой задачи свидетельствует достаточное количество публикаций по разработанным решениям для автоматического обнаружения DeepFake.

В качестве публично доступных датасетов для тетсировани механизмов обнаружения атак на биометрическое предъявления с помощью инструмента атаки класса DeepFake доступны следующие датасеты первого и второго поколений [1].

Таблица 1. **Подмена лица**

Table 1. **DeepFake – Identity Swap**

Первое поколение		
Название датасета	Подлинных видео	Фальсифицированных видео
UADFV (2018) [2]	49 (YouTube)	49 (FakeApp)
DeepfakeTIMIT (2018) [3]	–	620 (faceswap-GAN)
FaceForensics++ (2019) [4]	1,000 (YouTube)	1,000 (FaceSwap) 1,000 (DeepFake)
Второе поколение		
DeepFakeDetection[1] (2019)	363 (Actors)	3,068 (DeepFake)
Celeb-DF (2019) [5]	890 (YouTube)	5,639 (DeepFake)
DFDC Preview (2019) [6]	1,131 (Actors)	4,119 (Unknown)

Класс атаки – обнаружение манипуляций с подменой лица получил самое широкое распространение ввиду большого внимания к таким методам изменения изображений, что вызвало и соответствующий интерес к системам обнаружения подделок. В таблице 2 приведены результаты работы известных систем обнаружения DeepFake, основанные на различных технологиях [5].

Таблица 2. **Обнаружение манипуляций с подменой лица**

Table 2. **Face Swap Manipulation Detection**

Источник	**Метод**	**Классификаторы**	**Точность**	**Датасет**
Korshunov and Marcel (2018) [3]	Audio-Visual Features	PCA+RNN PCA+LDA, SVM	EER = 3.3% EER = 8.9%	DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ)

[1] Dufour N., Gully A. Contributing Data to Deepfake Detection Research [Электронный ресурс] // Google AI. 24.09.1019. URL: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html (дата обращения: 03.09.2022).

Источник	Метод	Классификаторы	Точность	Датасет
Matern et al. (2019) [7]	Visual Features Logistic Regression	MLP	AUC = 85.1% AUC = 70.2% AUC = 77.0% AUC = 77.3% AUC = 78.0% AUC = 66.2% AUC = 55.1%	Own UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD DFDC Preview Celeb-DF
Yang et al. (2019) [8]	Head Pose Features	SVM	AUC = 89.0% AUC = 55.1% AUC = 53.2% AUC = 47.3% AUC = 55.9% AUC = 54.6%	UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD DFDC Preview Celeb-DF
Agarwal and Farid (2019) [9]	Head Pose and Facial Features	SVM	AUC = 96.3%	Own (FaceSwap, HQ)
Jung et al. (2020) [10]	Eye Blinking	Distance	Acc. = 87.5%	Own
Li et al. (2019) [5], [11]	Face Warping Features	CNN	AUC = 97.7% AUC = 99.9% AUC = 99.7% AUC = 93.0% AUC = 75.5% AUC = 64.6%	UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD DFDC Preview Celeb-DF
Afchar et al. (2018) [12]	Mesoscopic Features	CNN	Acc. = 98.4% AUC = 84.3% AUC = 87.8% AUC = 68.4% Acc. ‘ 90.0% Acc. ‘ 94.0% Acc. ‘ 98.0% Acc. ‘ 83.0% Acc. ‘ 93.0% Acc. ‘ 96.0% AUC = 75.3% AUC = 54.8%	Own UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ (DeepFake, LQ) FF++ (DeepFake, HQ) FF++ (DeepFake, RAW) FF++ (FaceSwap, LQ) FF++ (FaceSwap, HQ) FF++ (FaceSwap, RAW) DFDC Preview Celeb-DF
Zhou et al. (2018) [13]	Steganalysis Features + Deep Learning Features	CNN SVM	AUC = 85.1% AUC = 83.5% AUC = 73.5% AUC = 70.1% AUC = 61.4% AUC = 53.8%	UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD DFDC Preview Celeb-DF
Rössler et al. (2019) [4]	Mesoscopic Features Steganalysis Features Deep Learning Features	CNN	Acc. ‘ 94.0% Acc. ‘ 98.0% Acc. ‘ 100.0% Acc. ‘ 93.0% Acc. ‘ 97.0% Acc. ‘ 99.0%	FF++ (DeepFake, LQ) FF++ (DeepFake, HQ) FF++ (DeepFake, RAW) FF++ (FaceSwap, LQ) FF++ (FaceSwap, HQ) FF++ (FaceSwap, RAW)
Nguyen et al. (2019) [14]	Deep Learning Features	AE + Multi-Task Learning	AUC = 65.8% AUC = 62.2% AUC = 55.3% AUC = 76.3% EER = 15.1% AUC = 53.6% AUC = 54.3%	UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD FF++ (FaceSwap, HQ) DFDC Preview Celeb-DF
Nguyen et al. (2019) [15]	Deep Learning Features	Capsule Networks	AUC = 61.3% AUC = 78.4% AUC = 74.4% AUC = 96.6% AUC = 53.3% AUC = 57.5%	UADFV DeepfakeTIMIT (LQ) DeepfakeTIMIT (HQ) FF++ / DFD DFDC Preview Celeb-DF

Источник	Метод	Классификаторы	Точность	Датасет
Dang et al. (2019) [16]	Deep Learning Features	CNN + Attention Mechanism	AUC = 99.4% EER = 3.1%	DFFD
Dolhansky et al. (2019) [6]	Deep Learning Features	CNN	Precision = 93.0% Recall = 8.4%	DFDC Preview
Wang and Dantcheva (2020) [17]	Deep Learning Features	3DCNN	TCR = 95.13% TCR = 92.25%	FF++ (DeepFake, LQ) FF++ (FaceSwap, LQ)
Güera and Delp (2018) [18]	Image + Temporal Features	CNN + RNN	Acc. = 97.1%	Own
Sabir et al. (2019) [19]	Image + Temporal Features	CNN + RNN	AUC = 96.9% AUC = 96.3%	FF++ (DeepFake, LQ) FF++ (FaceSwap, LQ)
Tolosana et al. [20]	Facial Regions Features	CNN	AUC = 100.0% AUC = 99.4% AUC = 91.0% AUC = 83.6%	UADFV FF++ (FaceSwap, HQ) DFDC Preview Celeb-DF

Использованные показатели качества:
EER – Equal Error Rate
AUC – Area Under Curve
TCR – True Confirmation Rate

Публично доступных датасетов для атак - манипуляция атрибутами лица обнаружить не удалось. Возможно, это связано с тем, что создание такой базы при общедоступных решениях GAN не имеет большого смысла. В таблице 3 приведены результаты работы систем обнаружения манипуляций данного типа [1].

Таблица 3. **Манипуляция атрибутами лица**
Table 3. **Manipulation with face attributes**

Источник	Метод	Классификаторы	Точность	Датасет
Wang et al. (2019) [21]	GAN-Pipeline Features	SVM	Acc. = 84.7%	Own (InterFaceGAN/ StyleGAN)
Nataraj et al. (2019) [22]	Steganalysis Features	CNN	Acc. = 99.4%	Own (StarGAN/ CycleGAN)
Bharati et al. (2016) [23]	Deep Learning Features	(Face Patches) RBM	Overall Acc. = 96.2% Overall Acc. = 87.1%	Own (Celebrity Retouching, ND-IIITD Retouching)
Jain et al. (2019) [24]	Deep Learning Features (Face Patches)	CNN + SVM	Overall Acc. = 99.6% Overall Acc. = 99.7%	Own (ND-IIITD Retouching, StarGAN)
Tariq et al. (2018) [25]	Deep Learning Features	CNN	AUC = 99.9% AUC = 74.9%	Own (ProGAN, Adobe Photoshop)
Dang et al. (2019) [16]	Deep Learning Features	CNN + Attention Mechanism	AUC = 99.9% EER = 1.0%	DFFD (FaceApp/ StarGAN)
Wang et al. (2019) [26]	Deep Learning Features	DRN	AP = 99.8%	Own (Adobe Photoshop)
Marra et al. (2019) [27]	Deep Learning Features	CNN + Incremental Learning	Acc. = 99.3%	Own (Glow/StarGAN)
Zhang et al. (2019) [28]	Spectrum Domain Features	GAN Discriminator	Acc. = 100%	Own (StarGAN/ CycleGAN)
Rathgeb et al. (2020) [29]	PRNU Features	Score-Level Fusion	EER = 13.7%	Own (5 Public Apps)

Несмотря на большую популярность решений Face2Face и Neural Texture, в настоящее время появились технологии, позволяющие создать анимированное изображение любого портрета, используя видеоизображение другого лица. То есть для анимации не требуется большого объема видеоматериала конкретного лица, что было необходимо для более ранних технологий. Более того, новые технологии не требуют использования вычислительных систем с GPU, а вполне работоспособны на мобильных устройствах.

Таблица 4. **Подмена выражения лица** [1]
Table 4. **Facial Expression Swapping** [1]

Источник	Метод	Классификаторы	Точность	Датасет
Matern et al. (2019) [7]	Visual Features	Logistic Regression, MLP	AUC = 86.6%	FF++ (Face2Face, RAW)
Afchar et al. (2018) [12]	Mesoscopic Features	CNN	Acc. = 83.2% Acc. = 93.4% Acc. = 96.8% Acc. ‘ 91% Acc. ‘ 75% Acc. ‘ 85% Acc. ‘ 95%	FF++ (Face2Face, LQ) FF++ (Face2Face, HQ) FF++ (Face2Face, RAW) FF++ (Face2Face, LQ) FF++ (NeuralTextures, LQ) FF++ (NeuralTextures, HQ) FF++ (NeuralTextures, RAW)
Rössler et al. (2019) [4]	Mesoscopic Features Steganalysis Features Deep Learning Features	CNN	Acc. ‘ 98% Acc. ‘ 100% Acc. ‘ 81% Acc. ‘ 93% Acc. ‘ 99%	FF++ (Face2Face, HQ) FF++ (Face2Face, RAW) FF++ (NeuralTextures, LQ) FF++ (NeuralTextures, HQ) FF++ (NeuralTextures, RAW)
Nguyen et al. (2019) [14]	Deep Learning Features	Autoencoder	EER = 7.1% EER = 7.8%	FF++ (Face2Face, HQ) FF++ (NeuralTextures, HQ)
Dang et al. (2020) [16]	Deep Learning Features	CNN + Attention Mechanism	AUC =99.4% EER = 3.4%	FF++ (Face2Face, -)
Wang and Dantcheva (2020) [17]	Deep Learning Features	3DCNN	TCR=90.27% TCR =80.5%	FF++ (Face2Face, LQ) FF++ (NeuralTextures, LQ)
Sabir et al. (2019) [19]	Image + Temporal Features	CNN + RNN	Acc. = 94.3	FF++ (Face2Face, LQ)
Amerini et al. (2019) [30]	Image + Temporal Features	CNN + Optical Flow	Acc. = 81.6%	FF++ (Face2Face, -)

В каждой таблице приведены датасеты, на которых проводилось тестирование систем детектирования, и результаты тестирования с указанием использованного датасета. По результатам обзоров можно сделать вывод о возможности практически близкой к 100% вероятности обнаружения подделки. Тем не менее, результаты исследователей сопровождаются упрощенными условиями тестирования – используются публично доступные датасеты или собственные датасеты, которые используются и при обучении системы, и в процессе тестирования. Но если используется неизвестный разработчикам метод манипуляции, то вероятность обнаружения подделки снижается до достаточно скромных 65,18% – результат получен в 2020 году по открытому тестированию DeepFake Detection Challenge 2019.

В 2019 году организаторы объединились с лидерами индустрии и академическими кругами для того, чтобы провести конкурс Deepfake Detection Challenge (DFDC) с целью стимулировать создание новых методов обнаружения видеоподделок класса DeepFake. Создав и поделившись новой уникальной базой видеоклипов (более 100 000 образцов), DFDC дал возможность экспертам всего мира собраться вместе, оценить качество своих методов выявления подделок, испытать новые подходы и узнать об опыте коллег. В целом, такое открытое совместное сотрудничество поможет индустрии и обществу уверенно встретить вызов, которым является технология DeepFake, и обеспечить понимание, является ли видео контент подлинным. Как и в случае с недавно проведенным Hateful Memes Challenge, организаторы считают, что проведение конкурсов и предоставление датасетов являются средством ускорения развития ИИ.

DFDC стартовал в декабре 2019 года, и 2114 участников представили более 35,000 моделей на соревнование. По завершении конкурса в 2020 году были объявлены результаты. Также проводилась работа с победителями с целью помочь им поделиться кодом наилучших решений по детектированию подделок в ходе конференции Computer Vision and Pattern Recognition (CVPR), где организаторы познакомили с деталями дальнейших планов по предоставлению доступа к необработанным видеоданным, которые послужили основой для создания датасета DFDC. Для создания контента были приглашены более 3 500 участников и время сбора данных составило около 38.5 суток. Это позволит разработчикам ИИ создавать новые методы обнаружения подделок. Более того, помимо помощи в разработке методов детектирования подделок, эти наборы данных могут быть использованы и для развития других направлений ИИ.

Одной из основных нерешенных проблем, выявленных в ходе конкурса, явилось то, что трудно разработать решение,

распознающее подделку при переходе на незнакомый датасет. Организаторы создавали конкурс, имея это ввиду. Для определения победителей, участники конкурса передавали свой код в «черный ящик». Этот набор видеоклипов не был доступен участникам, так что конкурсантам было необходимо создать модель, работающую даже в непредусмотренных условиях. Датасет «черного ящика» состоял из примерно 10 000 видеосюжетов, которые не были представлены участникам конкурса. В нем имелись как настоящие сюжеты из интернета, так и видео, сформированные с помощью технологий DeepFake для данного конкурса. Организаторы конкурса проверили, что соотношение подлинных видео и подделок соответствовало датасету, который был доступен публично. Точность работы механизмов детектирования для лучших представленных моделей достигает 82,56%. Но при тестировании моделей с использованием датасета «черного ящика» ситуация с ранжированием существенно изменяется. Наилучший показатель продемонстрировала модель, представленная Селимом Сефербековым. Точность детектирования составила 65,18% для набора данных из «черного ящика». При использовании публичного датасета модель заняла четвертое место. Аналогично, другие модели-победители (со второго по пятое место) заняли при тестировании с использованием публичного датасета более низкие места (37-ое, 6-ое, 10-ое и 17-ое соответственно). Это обстоятельство подчеркивает важность общего подхода с учетом неизвестных видео при создании моделей для детектирования подделок. Сам конкурс проходил на сайте Kaggle и победители определялись по результатам лог файлов при тестировании на закрытом датасете. Детали и результаты конкурса доступны на сайте Kaggle, где также имеется возможность познакомиться с данными и победившими моделями.

К сожалению, таких же больших по масштабу проектов по тестированию систем обнаружения DeepFake после DFDC 2019 не проводилось, но в КНР проводится ежегодное мероприятие DeepFake Game Competition (DFGC), которое является состязанием двух команд, одна из которых поставляет фальсифицированный контент, а другая пытается его обнаружить.

Наиболее заметными особенностями проводившихся в 2021 и 2022 году конкурсов DFGC было наличие двух команд, одна из которых поставляла тестовые базы данных, а вторая должна была обнаруживать поддельные изображения. Более того, в процессе создания DeepFake использовались механизмы состязательных атак с целью затруднить выявление поддельного контента и не позволить определить, какой продукт был использован для создания DeepFake. В конкурсе 2021 года принимало участие 180 индивидуальных конкурсантов, каждый из которых мог представить 1 решение в каждой команде. В конкурсе было организовано два этапа для каждой команды (создатели примеров и анализаторы). Результаты детектирования стабильно ухудшаются при переходе к этапу 2, и, как и в случае с конкурсом DFDC 2019 имеется снижение качества детектирования подделок при переходе на датасет из «черного ящика».

В 2022 году в конкурсе DFGC приняло участие 11 конкурсантов, представивших 35 решений. Практически все участники были представлены компаниями из КНР, а призовое вознаграждение составило 1200 долларов США за первое место. В то же время, можно отметить, что точность определения DeepFake несколько возросла – для датасетов, использовавшихся для обучения, возросла до 90%, для черных ящиков точность составила около 70%. Более подробные результаты можно получить в работах [31, 32].

Таким образом можно сделать вывод, что для организаций, которые используют или планируют использовать биометрические системы идентификации и аутентификации, технологии DeepFake создают дополнительные риски. Данные биометрические системы, в отличие от более широко используемых решений, базирующихся на факторах знания и владения, не требуют наличия криптографических секретов. Надежность биометрических систем базируется на подходе, связанным с невозможностью предъявления биометрических характеристик субъекта в его отсутствие. Любые попытки предъявления биометрических характеристик без присутствия их владельца выявляются системой обнаружения атак на биометрическое предъявление (Presentation Attack Detection – PAD) и приводят к блокировке результата работы системы сравнения биометрических признаков. PAD – обязательный компонент биометрической системы, от правильного функционирования которого зависит безопасность биометрической системы аутентификации в целом.

Из вышесказанного очевидно, что для того, чтобы иметь возможность оценить качество биометрической системы в целом, необходимо сопоставление решений как по таким характеристикам как FAR/FRR (False Accept Rate/False Reject Rate – вероятность ложного допуска/вероятность ложного недопуска), так и по аналогичным характеристикам системы PAD – APCER/BPCER (Attack Presentation Classification Error Rate/Bonafide Presentation Classification Error Rate – Уровень Вероятности обнаружения атаки/Уровень вероятности ошибки при предъявлении субъекта).

Использование DeepFake создало новую разновидность атак на биометрическое предъявление, которая не всегда может быть обнаружена стандартными решениями активных и пассивных модулей PAD.

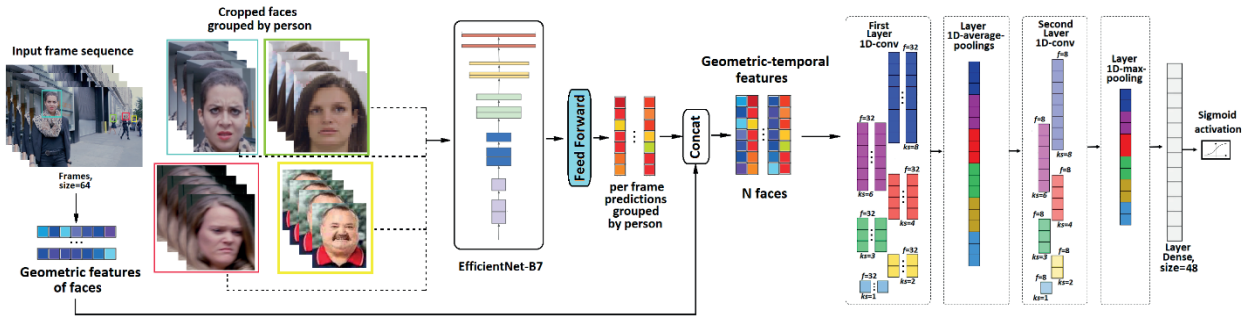
С учетом вышеизложенного, и широким использованием биометрических технологий в работе финансовых организаций, необходимо иметь решение по технологиям обнаружения DeepFake в системе обнаружения атак на биометрическое предъявление и уметь самостоятельно или с участием внешних компаний тестировать защищенность биометрических систем финансовых учреждений от атак на биометрическое предъявление с использованием вышеуказанной технологии.

В соответствии с таким подходом, лаборатория Кибербезопасности Сбербанка разработала собственное решение для обнаружения DeepFake. Изучая реальные и поддельные видео, эксперты лаборатории Кибербезопасности заметили, что есть корреляция между ложными срабатываниями и расположением лица в кадре. Они поняли, что для противодействия обходу системы выявления DeepFakes путем расположения реальных лиц рядом с фейковыми надо изучать предсказания лиц каждого человека по отдельности[2].

[2] В рамках исследования по выявлению DeepFake Сбербанк получил два патента на технологии. Одно из изобретений Роспатент признал патентом недели.

Разработанный алгоритм содержит следующие основные этапы. На первом этапе извлекаются все лица на анализируемых кадрах в видео и рассчитывается для них предсказание – вероятностная оценка, характеризующая наличие признаков характерных для DeepFake. На втором этапе определяются изображения лиц одного человека на анализируемых кадрах. На третьем этапе получаются геометрические характеристики лица человека в кадре. В качестве такой меры выбран размер части кадра, которое заполнено лицом. Другими словами, рассчитывается доля пространства, которое занимает лицо в кадре. Гипотеза состоит в том, что шумовые лица, такие как портреты на стенах, проходящие на заднем фоне люди, будут иметь меньшую долю пространства в кадре, чем основные. При этом нельзя просто исключать такие лица по минимальному размеру. Поскольку реальная съемка может происходить и на некотором расстоянии от людей. При этом, даже при равных размерах лиц, второстепенные люди будут иметь свои паттерны в последовательности предсказаний в кадрах. Это может быть вызвано излишним движением таких людей в кадре или наоборот неподвижностью (как для портретов), нахождением не в фокусе видео и другими факторами. По завершению третьего этапа для лиц каждого человека получается упорядоченная последовательность его характеристики пространственного заполнения. Объединяя последовательность изменений предсказаний DeepFake во времени и последовательность пространственного заполнения, мы получаем матрицу характеристик лиц человека.

Для расчета предсказания на первом этапе эксперты обучали ансамбли нейронных сетей. За основу каждой отдельной сети была использована предобученная модель. Такая модель получена методом глубокого обучения; для разработки ее архитектуры применялся метод составного масштабирования. При этом для выбора начальной архитектуры использовались инструменты автоматического машинного обучения (AutoML). Такие модели способны показать высокую точность при быстрой скорости работы. Используя такую нейронную сеть в качестве начальной инициализации, эксперты лаборатории обучили свои модели разделять дипфейки и реальные лица. Модели научились обнаруживать артефакты, которые характерны для используемого алгоритма генерации поддельных изображений. Архитектура решения представлена на рисунке ниже.



Поскольку по некоторым алгоритмам генерации DeepFake датасеты достаточно малы, был применен подход с инкрементным обучением. На первом шаге выполнялось обучение модели классифицировать различные алгоритмы DeepFake, далее – обучение модели различать конкретный алгоритм DeepFake и реальные видео.

Для каждого алгоритма генерации эксперты из лаборатории обучали свой ансамбль сетей. Были созданы ансамбли для таких алгоритмов как DeepFakes, DeepFakesDetection, DeepNostalgia, FaceShifter, Face2Face, NeuralTextures, FaceSwap и других. Объединяя предсказания ансамблей в один вектор, выполнялся расчет итоговой оценки на основе интегрального классификатора. Такой подход позволяет выявлять скрытые взаимосвязи между предсказаниями моделей для различных алгоритмов генерации синтетических изменений. Это обеспечивает достижение сверхаддитивного эффекта (синергетического) и повышает качество выявления видео с присутствием синтетических изменений изображений лиц. В еще одном варианте интегральный классификатор формирует не только интегральную оценку наличия синтетических изменений лиц людей на видео, но и наиболее вероятный алгоритм, с помощью которого были созданы данные синтетические изменения.

Заключение

В рамках данной работы для каждого типа манипуляций рассмотрены следующие характеристики: набор используемых технологий манипуляций, существующие общедоступные датасеты для тестирования и методы обнаружения факта изменения изображения, включая основные показатели вероятности обнаружения таких вмешательств в изображение.

Патент № 2768797 C1 Российская Федерация, МПК G06V 10/70, G06T 1/40, G06N 3/02. Способ и система для определения синтетически измененных изображений лиц на видео : № 2021130421 : заявл. 19.10.2021 : опубл. 24.03.2022 / К. Е. Вышегородцев, А. В. Балашов, Г. А. Вельможин, В. В. Сысоев ; заявитель Публичное акционерное общество «Сбербанк России». URL: https://www.elibrary.ru/item.asp?id=48376871 (дата обращения: 03.09.2022).

Патент № 2774624 C1 Российская Федерация, МПК G06K 9/62. Способ и система определения синтетических изменений лиц в видео : № 2021135267 : заявл. 01.12.2021 : опубл. 21.06.2022 / И. А. Оболенский, В. В. Сысоев, А. В. Балашов ; заявитель Публичное акционерное общество «Сбербанк России». URL: https://www.elibrary.ru/item.asp?id=48775354 (дата обращения: 03.09.2022).

Традиционные методы обнаружения подделок, разработанные до эры DeepFake, обычно базируются на «отпечатке» самой камеры, основанном на характеристике сенсора камеры, объектива, используемых цветовых фильтров, программном обеспечении камеры, «отпечатке» используемого для редактирования программного обеспечения, изменении скорости съемки для видеоизображений.

Использование классических технологий в современных условиях затрудняется тем обстоятельством, что при публикации фото/видео контента в сети Интернет происходят процессы изменения размеров изображения, применения различных технологий сжатия.

По результатам исследования можно сделать следующие выводы о возможности детектирования подделок:

Подмена лица: несмотря на наличие большого количества подходов к обнаружению DeepFake, достаточно сложно выделить наиболее удачный механизм обнаружения для всех датасетов. Наиболее целесообразным может оказаться применение нескольких технологий с последующим получением обобщенного результата.

Подмена Атрибутов: детектируется почти со 100% вероятностью (используется «отпечаток» системы генерации GAN, характерный для каждой системы). Но если удалить «отпечаток» или замаскировать его белым шумом, то удастся ли получить те же результаты?

Подмена выражения лица: поскольку имеется всего одна база изображений данного типа, весьма низкого качества, то результаты обнаружения манипуляций близки к 100%. Но при появлении новых баз более высокого качества ситуация может измениться.

Уделяемое мировыми разработчиками внимание к проблеме обнаружения DeepFake создает уверенность в том, что использование биометрических технологий в приложениях финансовых учреждений не будет скомпрометировано мошенниками, взявшими на вооружение новые технологии.

References

[1] Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*. 2020;64:131-148. doi: https://doi.org/10.1016/j.inffus.2020.06.014

[2] Li Y., Chang M.-C., Lyu S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE Computer Society; 2018. p. 1-7. doi: https://doi.org/10.1109/WIFS.2018.8630787

[3] Korshunov P., Marcel S. Deepfakes: a New Threat to Face Recognition? Assessment and Detection. arXiv:1812.08685. 2018. doi: https://doi.org/10.48550/arXiv.1812.08685

[4] Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niessner M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE Computer Society; 2019. p. 1-11. doi: https://doi.org/10.1109/ICCV.2019.00009

[5] Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE Computer Society; 2020. p. 3204-3213. doi: https://doi.org/10.1109/CVPR42600.2020.00327

[6] Dolhansky B., Howes R., Pflaum B., Baram N., Ferrer C.C. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:1910.08854. 2019. doi: https://doi.org/10.48550/arXiv.1910.08854

[7] Matern F., Riess C., Stamminger M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). Waikoloa, HI, USA: IEEE Computer Society; 2019. p. 83-92. doi: https://doi.org/10.1109/WACVW.2019.00020

[8] Yang X., Li Y., Lyu S. Exposing Deep Fakes Using Inconsistent Head Poses. In: ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE Computer Society; 2019. p. 8261-8265. doi: https://doi.org/10.1109/ICASSP.2019.8683164

[9] Agarwal S., Farid H. Protecting World Leaders Against Deep Fakes. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA; 2019. p. 1-8. Available at: https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf (accessed 03.09.2022).

[10] Jung T., Kim S., Kim K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access.* 2020;8:83144-83154. doi: https://doi.org/10.1109/ACCESS.2020.2988660

[11] Li Y., Lyu S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. arXiv:1811.00656. 2019. doi: https://doi.org/10.48550/arXiv.1811.00656

[12] Afchar D., Nozick V., Yamagishi J., Echizen I. MesoNet: a Compact Facial Video Forgery Detection Network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE Computer Society; 2018. p. 1-7. doi: https://doi.org/10.1109/WIFS.2018.8630761

[13] Zhou P., Han X., Morariu V.I., Davis L.S. Two-Stream Neural Networks for Tampered Face Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA: IEEE Computer Society; 2017. p. 1831-1839. doi: https://doi.org/10.1109/CVPRW.2017.229

[14] Nguyen H.Y., Fang F., Yamagishi J., Echizen I. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). Tampa, FL, USA: IEEE Computer Society; 2019. p. 1-8. doi: https://doi.org/10.1109/BTAS46853.2019.9185974

[15] Nguyen H.H., Yamagishi J., Echizen I. Use of a Capsule Network to Detect Fake Images and Videos. arXiv:1910.12467. 2019. doi: https://doi.org/10.48550/arXiv.1910.12467

[16] Dang H., Liu F., Stehouwer J., Liu X., Jain A. On the Detection of Digital Face Manipulation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE Computer Society; 2020. p. 5780-5789. doi: https://doi.org/10.1109/CVPR42600.2020.00582

[17] Wang Y., Dantcheva A. A Video is Worth More than 1000 Lies. Comparing 3DCNN Approaches for Detecting Deepfakes. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). Buenos Aires, Argentina: IEEE Computer Society; 2020. p. 515-519. doi: https://doi.org/10.1109/FG47880.2020.00089

[18] Güera D., Delp E. Deepfake Video Detection Using Recurrent Neural Networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand: IEEE Computer Society; 2018. p. 1-6. doi: https://doi.org/10.1109/AVSS.2018.8639163

[19] Sabir E., Cheng J., Jaiswal A., AbdAlmageed W., Masi I., Natarajan P. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops'2019). Long Beach, CA, USA: Computer Vision Foundation; 2019. p. 80-87. doi: https://doi.org/10.48550/arXiv.1905.00582

[20] Tolosana R., Romero-Tapiador S., Fierrez J., Vera-Rodriguez R. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. In: Del Bimbo A., et al. Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science. Vol. 12665. Cham: Springer; 2021. p. 442-456. doi: https://doi.org/10.1007/978-3-030-68821-9_38

[21] Wang R., Juefei-Xu F., Ma L., Xie X., Huang Y., Wang J., Liu Y. FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20). Yokohama, Yokohama, Japan; 2021. Article number: 476. p. 3444-3451. Available at: https://dl.acm.org/doi/pdf/10.5555/3491440.3491916 (accessed 03.09.2022).

[22] Nataraj L., Mohammed T.M., Manjunath B.S., Chandrasekaran S., Flenner A., Bappy J.H., Roy-Chowdhury A.K. Detecting GAN generated Fake Images using Co-occurrence Matrices. In: Proc. IS&T Int'l. Symp. on Electronic Imaging: Media Watermarking, Security, and Forensics. IS&T; 2019. p. 532. doi: https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532

[23] Bharati A., Singh R., Vatsa M., Bowyer K.W. Detecting Facial Retouching Using Supervised Deep Learning. *IEEE Transactions on Information Forensics and Security*. 2016;11(9):1903-1913. doi: https://doi.org/10.1109/TIFS.2016.2561898

[24] Jain A., Singh R., Vatsa M. On Detecting GANs and Retouching based Synthetic Alterations. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). Redondo Beach, CA, USA: IEEE Computer Society; 2018. p. 1-7. doi: https://doi.org/10.1109/BTAS.2018.8698545

[25] Tariq S., Lee S., Kim H., Shin Y., Woo S.S. Detecting Both Machine and Human Created Fake Face Images In the Wild. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security (MPS'18). New York, NY, USA: Association for Computing Machinery; 2018. p. 81-87. doi: https://doi.org/10.1145/3267357.3267367

[26] Wang S., Wang O., Owens A., Zhang R., Efros A. Detecting Photoshopped Faces by Scripting Photoshop. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE Computer Society; 2019. p. 10071-10080. doi: https://doi.org/10.1109/ICCV.2019.01017

[27] Marra F., Saltori C., Boato G., Verdoliva L. Incremental learning for the detection and classification of GAN-generated images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, Netherlands: IEEE Computer Society; 2019. p. 1-6. doi: https://doi.org/10.1109/WIFS47025.2019.9035099

[28] Zhang X., Karaman S., Chang S.-F. Detecting and Simulating Artifacts in GAN Fake Images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS). Delft, Netherlands: IEEE Computer Society; 2019. p. 1-6. doi: https://doi.org/10.1109/WIFS47025.2019.9035107

[29] Rathgeb C., Botaljov A., Stockhardt F., Isadskiy S., Debiasi L., Uhl A., Busch C. PRNU-based Detection of Facial Retouching. *IET Biometrics*. 2020;9(4):154-164. doi: https://doi.org/10.1049/iet-bmt.2019.0196

[30] Amerini I., Galteri L., Caldelli R., Del Bimbo A. Deepfake Video Detection through Optical Flow Based CNN. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE Computer Society; 2019. p. 1205-1207. doi: https://doi.org/10.1109/ICCVW.2019.00152

[31] Peng B. et al. DFGC 2021: A DeepFake Game Competition. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). Shenzhen, China: IEEE Computer Society; 2021. p. 1-8. doi: https://doi.org/10.1109/IJCB52358.2021.9484387

[32] Peng B. et al. DFGC 2022: The Second DeepFake Game Competition. In: 2022 IEEE International Joint Conference on Biometrics (IJCB). Abu Dhabi, United Arab Emirates: IEEE Computer Society; 2022. p. 1-10. doi: https://doi.org/10.1109/IJCB54206.2022.10007991

Поступила 03.09.2022; одобрена после рецензирования 05.10.2022; принята к публикации 10.10.2022.
Submitted 03.09.2022; approved after reviewing 05.10.2022; accepted for publication 10.10.2022.

Об авторах:

Балашов Александр Викторович, исполнительный директор, Лаборатория кибербезопасности, ПАО «Сбербанк России» (117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0002-6490-7547,** Balashov.A.Vik@sberbank.ru

Вышегородцев Кирилл Евгеньевич, руководитель направления, Лаборатория кибербезопасности, ПАО «Сбербанк России» (117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19), кандидат технических наук, **ORCID: https://orcid.org/0000-0002-8792-8594,** KEVyshegorodtsev@sberbank.ru
Свичкарь Денис Анатольевич, руководитель направления отдела оценки защищенности технологий И\А, Управление криптографии, аутентификации и идентификации, Департамент кибербезопасности, ПАО «Сбербанк России» (117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0002-0158-1682,** DASvichkar@sberbank.ru
Хенкин Петр Владимирович, исполнительный директор – начальник отдела, Управление криптографии, аутентификации и идентификации, Департамент кибербезопасности, ПАО «Сбербанк России» (117312, Российская Федерация, г. Москва, ул. Вавилова, д. 19), **ORCID: https://orcid.org/0000-0001-6141-9970,** PVKhenkin@sberbank.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Aleksandr V. Balashov, Executive Director, Cybersecurity Laboratory, PJSC “Sberbank of Russia” (19 Vavilov St., Moscow 117312, Russian Federation), **ORCID: https://orcid.org/0000-0002-6490-7547,** Balashov.A.Vik@sberbank.ru
Kirill E. Vyshegorodtsev, Associate Director of the Cybersecurity Lab, PJSC “Sberbank of Russia” (19 Vavilov St., Moscow 117312, Russian Federation), Cand.Sci. (Eng.), **ORCID: https://orcid.org/0000-0002-8792-8594,** KEVyshegorodtsev@sberbank.ru
Denis A. Svichkar, Associate Director, Cybersecurity Department, PJSC “Sberbank of Russia” (19 Vavilov St., Moscow 117312, Russian Federation), **ORCID: https://orcid.org/0000-0002-0158-1682,** DASvichkar@sberbank.ru
Petr V. Khenkin, Executive Director, Cybersecurity Department, PJSC “Sberbank of Russia” (19 Vavilov St., Moscow 117312, Russian Federation), **ORCID: https://orcid.org/0000-0001-6141-9970,** PVKhenkin@sberbank.ru

All authors have read and approved the final manuscript.