Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ РОЗПІЗНАВАННЯ DEEPFAKE КОНТЕНТУ

Текстова частина магістерської роботи

за спеціальністю „Комп'ютерні науки" 122

Керівник магістерської роботи
доцент, к. ф-м. н.
Нагірна А.М.

_____

(підпис)
"___" _____ 2023 р.

Виконала студентка
Оксюта І. М.
"___" _____ 2023 р.

Київ 2023

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ
Зав. кафедри інформатики
к.ф-м.н., доц.
Гороховський С.С

_____
(підпис)
“_____”_____2022 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на магістерську роботу

студенту 2 р.н. магістерської програми Комп’ютерні Науки Оксюті Ірині Миколаївні
Розробити Інформаційну систему розпізнавання Deepfake контенту
Зміст текстової частини до магістерської роботи:

Зміст

Анотація

Вступ

1 Предметна область

2 Задача розпізнавання deepfake відео. Огляд існуючих рішень

3 Опис рішення та програмного продукту

Висновки по роботі та рекомендації для подальших досліджень

Список літератури

Дата видачі “___”_____2022 р.

Керівник
А.М. Нагірна, доцент, к.ф-м.н. _____
(підпис)

Завдання отримала
І.М. Оксюта _____
(підпис)

Тема: Інформаційна система розпізнавання Deepfake контенту

Календарний план виконання роботи:

№ п/п	Назва етапу дипломного проекту (роботи)	Термін виконання етапу	Примітка
1.	Отримання завдання на дипломну роботу.	02.11.2022	
2.	Огляд технічної літератури за темою роботи.	16.11.2022	
3.	Виконати аналіз сучасних методів узагальнення розпізнавання deepfake контенту	30.11.2022	
4.	Розробка алгоритму адаптації моделі розпізнавання deepfake	18.12.2022	
5.	Програмування розробленого алгоритму	20.01.2023	
6.	Виконання експериментів та тестування роботи алгоритму	14.03.2023	
7.	Виконання порівняльного аналізу результатів, отриманих за допомогою розробленого алгоритму та існуючими методами узагальнення	05.04.2023	
8.	Написання пояснювальної роботи.	15.04.2023	
9.	Створення слайдів для доповіді та написання доповіді.	27.04.2023	
10.	Аналіз отриманих результатів з керівником, написання доповіді та попередній захист магістерської роботи.	05.05.2022	
11.	Корегування роботи за результатами попереднього захисту.	16.05.2023	
12.	Остаточне оформлення пояснювальної роботи та слайдів.	31.05.2023	
13.	Захист магістерської роботи (проекту)	12.06.2023	

Студентка Оксюта І.М.

Керівник Нагірна А.М. “____”_____ 2022 р.

ЗМІСТ

Анотація

Робота присвячена дослідженню узагальнення та адаптації методів розпізнавання deepfake контенту до роботи в реальних сценаріях, з тестовими даними відмінними від навчальних. Проведено аналіз існуючих підходів до узагальнення моделей машинного навчання. Запропоновано нові підходи до вирішення проблеми адаптації, а саме навчання моделі розпізнавання в парадигмі мета-навчання та використання однокрокового навчання під час тестування. Результатом роботи є система, яка направлена на розпізнавання підроблених облич на відео. Проведені експерименти показують ефективність запропонованого підходу в умовах різниці методів генерації deepfake для відеофайлів у навчальному та тестовому наборах. Отримані результати можуть бути імплементовані в системах перевірки достовірності контенту або в рамках персонального використання.

Ключові слова: deepfake, deep learning, машинне навчання, розпізнавання облич, комп’ютерний зір, нейронні мережі, мета-навчання.

ВСТУП

Актуальність. Технологія deepfake в сучасному світі набула досить великої популярності. Заміну обличчя або голосу людини використовують в розважальних індустріях для досягнення більшої реалістичності та створенні якісних персонажів. Але більш загрозливим є підробка медіаконтенту з метою маніпуляцій, дезінформації та шахрайства. Щодня в інформаційному середовищі генерується величезна кількість новин та публікацій в соціальних мережах. Тому важливо дбати про інформаційну гігієну та мати змогу відрізнити підробку. Цією задачею займаються дослідники, будуючи моделі розпізнавання deepfake методами машинного навчання та комп'ютерного зору. Готові рішення показують достатньо високу точність виявлення та вже зараз можуть бути успішно імплементовані у відповідні системи фільтрації контенту. Але існуючі моделі навчаються на визначеному наборі даних, що обмежує їх здатність виявляти раніше невідомі методи підробки. Саме тому не можна бути певним у ефективності створеного рішення у майбутньому. Ця проблема постійної еволюції методів фальшування вимагає розробки моделей, які можуть адаптуватись до раніше небачених даних, та бути достатньо ефективними без додаткового доопрацювання. Розробка таких узагальнених моделей навчання має велике значення для поліпшення інформаційної безпеки, запобігання шахрайству та інших негативних наслідків використання цієї технології.

Мета дослідження. Створити модель розпізнавання deepfake відео, як узагальнення та адаптацію моделі, яка навчена на одному наборі даних, для ефективної роботи з тестовими зразками, які були сформовані іншими методами підробки.

Завдання дослідження. Проаналізувати існуючі методи створення фальшивих відео та зображень. Вивчити та порівняти існуючі набори даних для задачі розпізнавання deepfake. Оцінити ефективність сучасних моделей

машинного навчання на цих наборах даних. Дослідити методи та підходи для адаптації моделей до потенційно нестабільних тестових даних, які суттєво відрізняються від навчальних. Сформувати підхід та розробити архітектуру, яка дозволить ефективніше розпізнавати підробки в непередбачуваних умовах. Проаналізувати ефективність запропонованого методу у порівнянні з вихідними моделями навчання.

Об’єкт дослідження. Інструменти обробки зображень та відеофайлів. Моделі машинного навчання для розпізнавання deepfake відео. Підходи для покращення ефективності розпізнавання deepfake в умовах появи нових методів підробки.

Предмет дослідження. Можливість побудови нової архітектури узагальнення моделі розпізнавання deepfake, з метою стабілізації роботи на прикладах в реальному світі. Дослідження нової парадигми навчання адаптованих моделей.

Джерела дослідження. Тематичні статті та дослідження в електронному форматі, форуми спільноти CVPR, документація відповідних фреймворків та інструментів обробки даних.

Наукова новизна одержаних результатів полягає в створенні підходу для адаптації моделей розпізнавання deepfake до роботи з тестовими даними, які можуть за своєю природою суттєво відрізнятись від навчальних даних.

Практичне значення одержаних результатів. Запропонований метод адаптації може бути застосований до будь-якої архітектури моделі розпізнавання deepfake відео та вимагає лише зміни процесу навчання та тестування. Використання open-source бібліотек та фреймворків для машинного навчання дозволяє легко впроваджувати описаний підхід.

Робота складається з трьох розділів.

У першому розділі роботи наведено чітке визначення поняття Deepfake, проблематика використання цієї технології. Розглянуто основні риси та види підробок, описано принципи роботи методів генерації підробок, включаючи автокодери та генеративно-змагальні мережі (GAN). Визначено високий потенціал розвитку технології deepfake в майбутньому.

Другий розділ присвячений задачі розпізнавання Deepfake відео. Описана формалізація задачі з точки зору машинного навчання, визначені основні складові системи виявлення підробок. Проведено аналіз існуючих наборів даних, методів розпізнавання облич та архітектур нейронних мереж для вирішення поставленої задачі. Відповідно до наведених існуючих методів узагальнення та адаптації моделей розпізнавання deepfake, зроблені висновки про можливі шляхи побудови рішення у ході цієї роботи.

У третьому розділі описано етапи однокрокового навчання під час тестування та парадигми мета-навчання. Реалізовано програмний продукт відповідно до запропонованого підходу. Для досягнення об’єктивності дослідження, проведено експерименти з тестуванням ефективності розробленого методу у порівнянні з існуючими, показано переваги такого підходу.

Завдяки високому рівню наукової деталізації та комплексному підходу, ця робота здатна значно сприяти розумінню проблематики Deepfake та розвитку методів розпізнавання та захисту від цього виду маніпуляційного медіаконтенту.

РОЗДІЛ 1. ПРЕДМЕТНА ОБЛАСТЬ

1.1. Визначення проблематики Deepfake

Deepfake (від англ. «deep learning» + «fake») — це технологія, яка передбачає використання штучного інтелекту і методів машинного навчання для створення або обробки зображень, відео чи аудіо, щоб вони виглядали справжніми, але насправді є сфабрикованими. Цей підхід дозволяє вносити правки до вихідного контенту, які можуть варіюватися від непомітної зміни виразу обличчя до повної заміни обличчя людини. Ця технологія знайшла різні застосування в індустрії розваг, пропонуючи нові творчі можливості та покращуючи візуальні ефекти. Зокрема зараз поширено використання deepfake інструментів у кіновиробництві, соціальних мережах, комп'ютерних іграх.

Але головне занепокоєння щодо deepfakes полягає в їх використанні з метою обману та маніпулювання окремими особами, організаціями та суспільством. Ось деякі з основних проблем, які виникають в наслідок використання цієї технології:

1. **Дезінформація та фейкові новини**. Deepfakes мають потенціал для поширення дезінформації у великих масштабах. Їх можна використовувати для створення переконливих відео з громадськими діячами, політиками чи знаменитостями, які говорять або роблять те, чого насправді ніколи не робили. Це може призвести до поширення неправдивої інформації та підриву довіри у суспільства.
2. **Репутаційні збитки**. Deepfakes можна використовувати для псування репутації окремих осіб або організацій. Створюючи реалістичні, але сфабриковані відео чи зображення, зловмисники можуть створювати враження, що хтось займається незаконною, аморальною чи компрометуючою діяльністю. Це може мати серйозні наслідки для цільових осіб, у тому числі шкоду для особистих стосунків, кар'єрних

перспектив і психічного благополуччя.

3. **Порушення конфіденційності**. Deepfake може порушувати особисту конфіденційність, накладаючи чиєсь обличчя на відвертий або компрометуючий вміст без їхньої згоди. Таке втручання в приватне життя може спричинити емоційний стрес, занепокоєння та може використовуватися як інструмент для переслідування чи шантажу.
4. **Політичні маніпуляції**. Deepfakes становлять значну загрозу демократичним процесам. Створюючи фальшиві відео чи аудіозаписи політичних кандидатів, підробки можуть впливати на громадську думку, впливати на вибори та підривати цілісність демократичної системи.
5. **Довіра до ЗМІ**. Deepfakes підривають довіру до ЗМІ та концепцію автентичності. Оскільки технологія стає все більш досконалою, стає все важче відрізнити реальний вміст від маніпуляційного. Це може мати далекосяжні наслідки, включаючи скептицизм щодо законних доказів, новин і публічного дискурсу.
6. **Юридичні проблеми.** Deepfakes піднімають складні юридичні та етичні питання. Створення та поширення підробленого контенту без згоди залучених осіб може порушувати права на конфіденційність, інтелектуальну власність і може призвести до судових спорів.

Методи deepfake стають дедалі доступнішими для звичайних користувачів та дозволяють отримати високий рівень реалістичності, що ускладнює визначення автентичності вмісту. Яскравим прикладом можна назвати українську компанію Reface, яка створює програмні продукти та мобільні застосунки в цій сфері. У 2020 році компанія залучила 5,5 мільйонів доларів США та суттєво виросла в зв’язку з обставинами пандемії коронавірусу. Їхній додатки незмінно входять до числа найкращих безкоштовних розважальних програм і отримали мільйони завантажень по всьому світу.

Деяка статистика використання технології deepfake:

1. Згідно з дослідженням компанії Deeptrace, яка займається кібербезпекою, кількість підроблених відео, доступних в Інтернеті, зросла на 330% з 2018 по 2019 рік.
2. Глобальне опитування, проведене компанією Ipsos у 2019 році, показало, що 71% респондентів знають про deepfakes, а 61% занепокоєні їхнім потенційним впливом.
3. У звіті Оксфордського інституту Інтернету за 2020 рік підкреслюється, що використання deepfake для політичних маніпуляцій зросло в усьому світі, причому такі випадки були виявлені щонайменше у 28 країнах.
4. У 2019 році Європейська комісія опублікувала керівні принципи боротьби з дезінформацією, які включають положення про deepfakes.
5. У 2020 році український парламент ухвалив законодавство, що передбачає покарання за створення та поширення deepfakes з метою введення в оману.
6. Банківський та фінансовий сектор також є вразливим до подібного шахрайства. У 2020 році повідомлялося про випадок, коли британська енергетична компанія була ошукана на 220 000 євро після того, як шахраї використали підроблення аудіо, щоб видати себе за генерального директора компанії.

Саме тому вирішення проблем, пов’язаних із deepfake, дуже важливе з етичної точки зору та вимагає багатогранного підходу. Це передбачає розробку надійних технологій виявлення підробок, впровадження суворіших правил і політики щодо створення та розповсюдження фальсифікованих медіа, сприяння медіа грамотності для інформування громадськості про ризики та проблеми, пов’язані з deepfakes.

1.2. Види та методи створення Deepfake контенту

Deepfake бувають різних форм і можуть бути класифіковані на різні типи залежно від їхнього застосування та використовуваних методів. До

найпоширеніших типів підробок можна віднести:

1. **Face Swap**: заміна обличчя однієї людини обличчям іншої на зображеннях або відео. Ці підробки використовують алгоритми штучного інтелекту для аналізу та зіставлення рис обличчя цільової особи з обличчям вихідної особи, створюючи переконливу і безшовну підміну.
2. **Voice Synthesis**: маніпуляції та створення синтетичного звуку, який імітує голос конкретної людини. Тренуючись на великих наборах аудіоданих, алгоритми штучного інтелекту можуть генерувати мову, яка звучить надзвичайно схоже на голос цільової людини. Такі глибокі підробки мають наслідки для імітації голосу і можуть бути використані для створення фальшивих аудіозаписів.
3. **Puppeteering**: маріонетковий deepfake, що передбачає маніпуляції з виразом обличчя та рухами людини на відео. Використовуючи алгоритми штучного інтелекту для аналізу та відстеження рухів обличчя у вихідному відео, фейк може змінювати вираз обличчя та дії людини в режимі реального часу або на етапі постобробки.
4. **Lip-Sync**: синхронізація рухів губами людини на відео з іншим джерелом звуку. Накладаючи форму звукової хвилі на рухи рота цільової людини, ці підробки створюють ілюзію природної синхронізації губ, що корисно для цілей дубляжу та локалізації.

Ця робота зосереджується на проблемі підробки облич людей на відео та зображеннях. Тому необхідно визначити методи, якими відповідні deepfakes генеруються та потенціал розвитку таких технологій. Найчастіше створення підробленого контенту відбувається з використанням алгоритмів машинного навчання. Системі подається велика кількість даних, які вона потім використовує, щоб навчитися створювати свої власні, нові дані. В основному такі рішення базуються на нейронних мережах, а саме Autoencoders або GAN.

1.2.1 Autoencoders

Autoencoders - це нейромережеві архітектури, які широко використовуються в глибинному навчанні для різних завдань. Це неконтрольовані моделі навчання, які мають на меті реконструювати вхідні дані зі стисненого представлення. У контексті deepfake автокодери можна використовувати для вивчення прихованих представлень облич і генерування синтетичних зображень облич [1]. Складові такої системи можна описати наступним чином:

1. Encoder: Кодерний компонент автокодера бере вхідне зображення, наприклад, обличчя, і стискає його в більш низькорозмірне представлення, яке часто називають latent code, в прихованому просторі. Цей процес кодування фіксує основні характеристики вхідного зображення.
2. Decoder: Компонент декодера приймає latent code, згенерований кодером, і намагається реконструювати оригінальне вхідне зображення. Він вчиться декодувати приховане представлення назад у вихідний простір даних.
3. Training: Під час навчання автокодер прагне мінімізувати різницю між вхідним зображенням і реконструйованим виходом. Параметри кодера та декодера оновлюються за допомогою backpropagation для оптимізації втрат при реконструкції.
4. Latent Space Manipulation: Після навчання, створеним автокодером прихованим простором можна маніпулювати для генерації нових синтетичних зображень обличчя. Модифікуючи latent code, наприклад, змінюючи певні атрибути або риси, декодер може створювати різні варіації вхідного обличчя.

Схему роботи автокодера під час навчання та під час генерації deepfake показано на рисунку 1.1 [2].

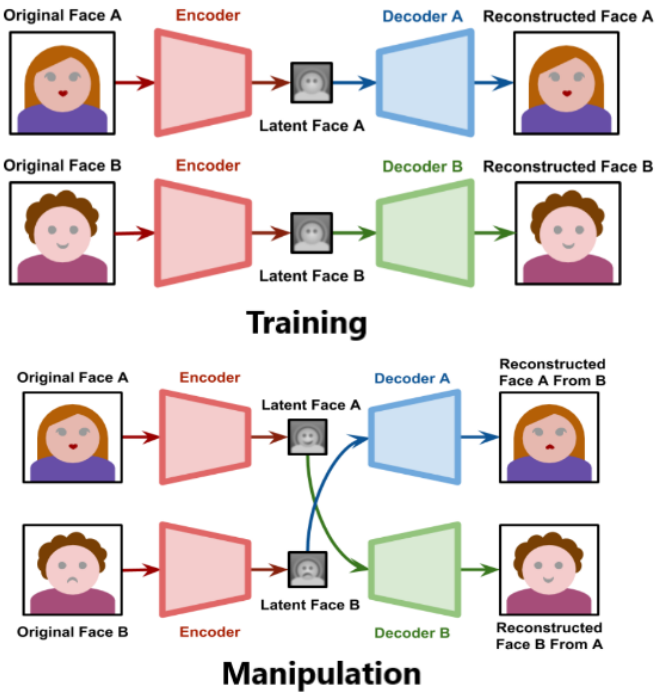


Рисунок 1.1. Схема роботи автокодера.

Окрім стандартного автокодера поширеними є наступні типи:

1. Variational Autoencoder (VAE): VAE розширюють можливості звичайного автокодера за рахунок імовірнісного моделювання. Прихований простір у VAE не є фіксованим представленням, а скоріше розподілом. Це дозволяє генерувати різноманітні результати шляхом вибірки з латентного простору. VAE широко використовуються для створення нових зразків даних і вивчення структурованих представлень.
2. Sparse Autoencoder: розріджені автокодери вводять обмеження розрідженості на латентне представлення. Заохочуючи активність лише підмножини нейронів у прихованому шарі, розріджені автокодери можуть вивчати більш стійкі та значущі представлення. Вони корисні для завдань із виділення ознак і виявлення аномалій.
3. Denoising Autoencoder: автокодувальники з усуненням шумів навчені відновлювати чисті дані з пошкоджених або зашумлених вхідних даних. Додаючи шум до вхідних даних і навчаючи автокодер відновлювати вихідні дані, автокодери з шумозаглушенням навчаються фіксувати базову структуру та видаляти шум.
4. Contractive Autoencoder: контрактивні автокодери включають термін

регуляризації, який заохочує модель вивчати більш стабільні та надійні представлення. Ця регулярізація знижує чутливість моделі до невеликих змін у вхідних даних.

Автокодери також можна використовувати в поєднанні з іншими методами, такими як заміна обличчя або передача стилю, для створення deepfake зображень або відео.

1.2.2 GAN

Generative Adversarial Network (GAN) були представлені в статті Яна Гудфеллоу у 2014 році [3]. Посилаючись на GAN, директор з досліджень штучного інтелекту Facebook Янн ЛеКун назвав змагальне навчання «найцікавішою ідеєю за останні 10 років у ML».

У архітектурі GAN використовуються дві основні компоненти. Генератор - це нейронна мережа, яка створює нові приклади даних, а дискримінатор намагається визначити, наскільки ці приклади є правдоподібними. Дискримінатор вирішує, чи є приклад даних справжнім (тобто ймовірно належить до початкового набору даних) чи фальшивим. Генератор намагається обманути дискримінатор і навчається на більшому обсязі даних, щоб отримати реалістичні результати. Ця архітектура має змагальний характер, оскільки генератор і дискримінатор працюють один проти одного, маючи протилежні цілі: одна модель намагається імітувати реальність, а інша - виявити підробки. Ці два компоненти тренуються одночасно, покращуючи свої можливості з часом. Вони можуть навчитися ідентифікувати та відтворювати складні навчальні дані, такі як зображення, аудіо та відео. В контексті deepfake складові GAN можна описати наступним чином:

1. Навчальні дані: GAN потребують великого набору реальних даних для навчання. Це може бути колекція зображень або відео цільової особи (чиє обличчя буде замінено) та особи-джерела (чиє обличчя

буде вставлено).

2. Мережа генератор: генератор відповідає за створення синтетичних зображень. Ця мережа генерує зображення або відео, які нагадують зовнішність особи-мішені, але при цьому включають міміку і рухи особи-джерела.
3. Мережа дискримінатор: дискримінатора намагається відрізнити реальні медіа від створених. Вона тренується на реальних і прикладах створених генератором і вчиться точно їх класифікувати.
4. Змагальне навчання: під час навчання генератор і дискримінатора протистоять одна одній. Генератор прагне створювати все більш реалістичні підробки, щоб обдурити дискримінатор, тоді як дискримінатор намагається стати більш ефективним у розрізненні справжніх і фейкових даних.

Цей змагальний процес навчання триває доти, доки генератор не зможе створювати deepfake, які дискримінаторній мережі буде важко відрізнити від справжніх зображень. Загальна архітектура GAN представлена на рисунку 1.2 [4].

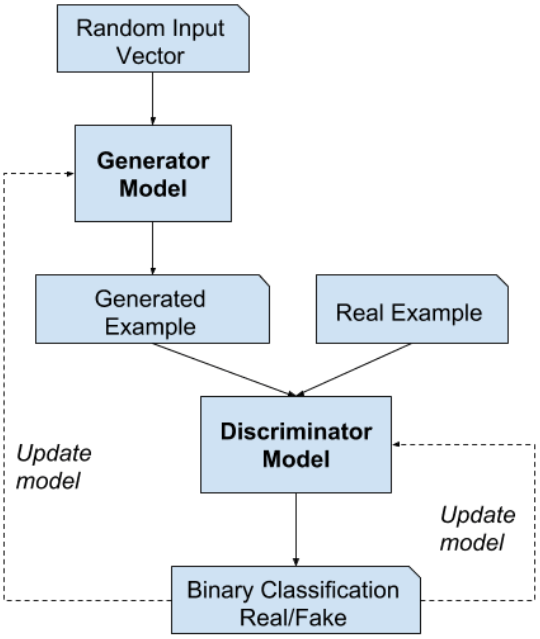


Рисунок 1.2. Архітектура GAN.

Генеративні змагальні мережі продемонстрували значний потенціал у

широкому діапазоні сфер і застосувань. Їхня здатність генерувати реалістичні та різноманітні синтетичні дані відкрила нові можливості для розробників. Одним із відомих застосувань GAN є StyleGAN від компанії NVIDIA. Технологія може генерувати правдоподібні зображення облич, тварин чи інших об’єктів. За допомогою додаткових інструментів мережа також може маніпулювати цими зображеннями. Остання версія продукту StyleGAN 3 була випущена у 2021 році та суттєво удосконалювала результати роботи не тільки на зображеннях, а й на відеофайлах.

1.3. Висновки до розділу 1

Технологія deepfake стає все більш популярною у сучасному світі та вимагає відповідального використання. Окрім застосування в сфері розваг та загальному прогресу у питаннях генерації контенту, дуже важливо розуміти, що цей інструмент використовується також в шахрайських цілях.

Зловмисники можуть використовувати його для поширення дезінформації, маніпулювання людьми, впливу на політику, тощо.

Проаналізувавши найпоширеніші методи створення deepfake, можна сказати, що ці інструменти мають дуже великий потенціал, та вже зараз активно удосконалюються. Відповідально використовуючи ці технології, ми можемо використати їх потенціал для створення позитивного впливу на суспільство. Щоб зменшити потенційні ризики та запобігти зловмисному використанню технології deepfake, мають існувати відповідні етичні міркування, гарантії конфіденційності та правові рамки. Поточні дослідження, розвиток методів протидії цьому явищу безсумнівно відіграватимуть вирішальну роль у забезпеченні надійнішого та безпечнішого цифрового середовища.

РОЗДІЛ 2: ЗАДАЧА РОЗПІЗНАВАННЯ DEEPFAKE ВІДЕО. ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

2.1 Набори даних

Перш за все, для побудови будь-яких рішень з протидії deepfake необхідно зібрати достатньо великий та репрезентативний набір прикладів. Безпосередньо для задачі розпізнавання deepfake за допомогою методів машинного навчання, датасет для навчання та оцінювання відповідних моделей має містити багато варіантів підробок та бути приведеним до єдиного формату даних. Кожен екземпляр набору даних має містити мітку, яка вказує на те, чи є зразок справжнім чи підробленим. Різноманітність таких датасетів гарантує, що алгоритм виявлення deepfake піддається широкому спектру варіацій, включаючи різних людей, умови освітлення, ракурси камери та методи маніпулювання. Ці умови підвищують стійкість алгоритму та можливості узагальнення. Також, використовуючи стандартизовані відомі набори даних, дослідники можуть порівнювати ефективність різних моделей і визначати напрямки для вдосконалення.

У наступних пунктах наведено опис основних наборів даних, які використовують для розробки методів боротьби з deepfake, та їхня специфіка.

2.1.1 DFDC

Набір даних DFDC є одним із найбільш об’ємних і широко використовуваних ресурсів для роботи по виявленню deepfakes. Він був створений у рамках змагання DeepFake Detection Challenge, організованого Facebook у співпраці з академічними установами та галузевими партнерами. Набір даних являє собою різноманітну колекцію відео, що включає як справжні зразки так і підроблені з різними сценаріями [5]. Також цей набір даних поділений на навчальний та тестовий, що дає змогу дослідникам коректно порівнювати різні підходи.

Ключові характеристики набору даних DFDC:

1. Розмір і різноманітність: Набір даних DFDC складається з 5000 відео на яких

присутні актори різної статі, віку, етнічної приналежності. Маніпуляції проводилися з урахуванням двох підходів до заміни облич: метод А, який створює зображення вищої якості з обличчями, розташованими ближче до камери, враховуючи джерело та замінені обличчя в однакових пропорціях, і метод В, який має нижчу якість заміни. Зрештою, набір даних складається з 4464 зразків для навчання і 780 для тестування, кожен тривалістю 15 секунд і з різною роздільною здатністю. Приклад кадрів з відеофайлів набору показано на рисунку 2.1.

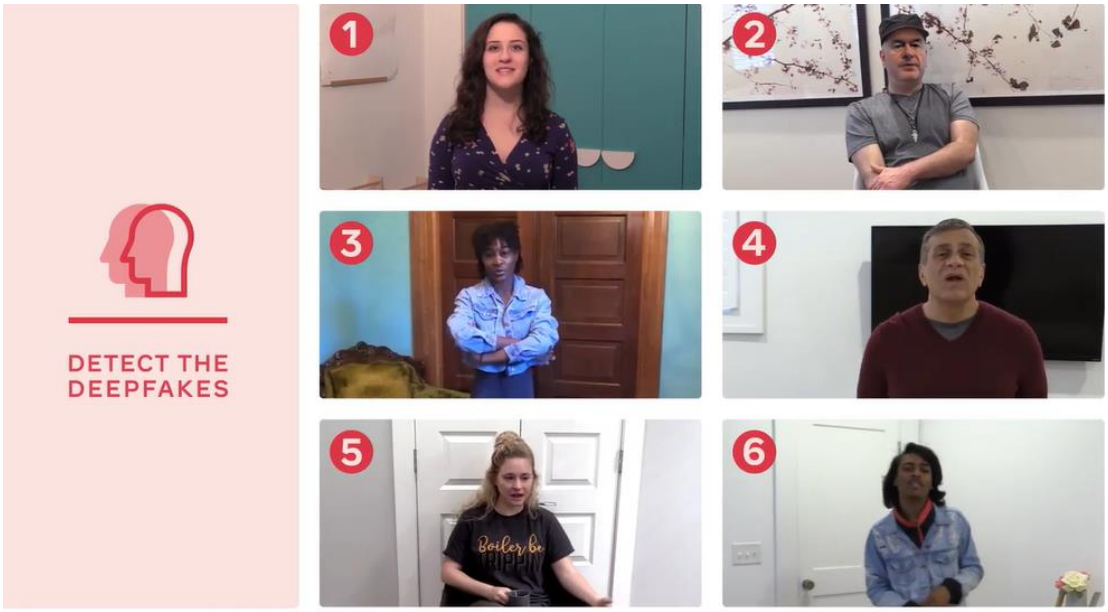


Рисунок 2.1. Приклад даних DFDC.

2. Анотації та маркування: кожне відео ретельно анотовано, надаючи дослідникам базові мітки, які вказують на те, чи є відео справжнім чи підробленим. Цей процес анотації дає змогу контролювати навчання та оцінювати алгоритми розпізнавання deepfake за допомогою стандартних метрик якості.
3. Різні методи підробки: набір даних включає deepfake, створені за допомогою різноманітних методів, таких як заміна обличчя, відтворення обличчя та синтез голосу. Автори також використовували інші техніки для досягнення різноманіття даних, а саме: покращення якості зображення, розмиття, зміна частоти кадрів.

Набір даних DFDC відіграв ключову роль у стимулюванні досліджень і розробок у сфері виявлення deepfake. Надаючи велику та різноманітну колекцію анотованих відео, він дав змогу дослідникам тренувати та оцінювати моделі та алгоритми, порівнювати їх продуктивність і досліджувати нові підходи. Оскільки технологія deepfake продовжує розвиватися, набір даних DFDC залишається цінним ресурсом для дослідників. Датасет дозволяє розробляти нові підходи та оцінювати їхню ефективність в порівнянні з існуючими методами.

2.1.2. FaceForensics ++

Ще одним відомим набором даних для розробки алгоритмів виявлення deepfakes є FaceForensics++. Це комплексний набір даних, розроблений спеціально для вдосконалення методів виявлення глибоких фейків [6]. Він був створений як розширення оригінального набору даних FaceForensics, усунувши деякі його обмеження та включивши більш складні маніпуляції з обличчям. Набір даних FaceForensics++ будується на 1000 непідроблених відео. Ці відео оброблюються з використанням різних технік: DeepFakes, FaceSwap, Face2Face, NeuralTextures.Таким чином дослідникам фактично доступні 5 версій одного й того ж відео: 1 реальне і 4 підробки. Набір даних розроблений таким чином, щоб бути різноманітним і складним, із відео з різними акторами, фоном, умовами освітлення та рівнями якості. Приклад різних типів обробки вхідного відео показаний на рисунку 2.2.

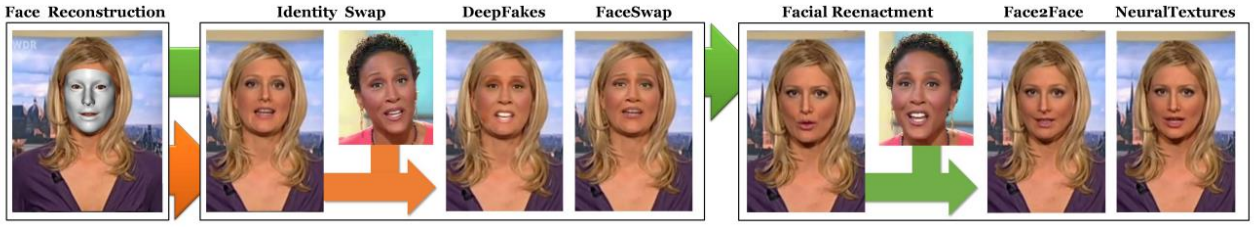


Рисунок 2.2. Обробка відео в датасеті Face Forencis++.

Техніки маніпуляції можна описати наступним чином:

1. **FaceSwap** — це графічний підхід до перенесення області обличчя з вихідного відео в цільове відео. На основі виявлених міток обличчя виділяється область обличчя. Використовуючи ці орієнтири, метод створює 3D-модель шаблону

за допомогою blendshapes. Ця модель проектується на цільове зображення шляхом мінімізації різниці між спроектованою формою та локалізованими орієнтирами за допомогою текстур вхідного зображення. Зрештою, візуалізована модель змішується із зображенням і застосовується корекція кольору.

2. **DeepFakes** — це , окрім терміну для підроблених зображень, також назва конкретного методу маніпуляції. Обличчя в цільовому відео замінюється обличчям, яке спостерігалося у вихідному відео чи колекції зображень. Метод заснований на двох автокодерах зі спільним кодувальником, які навчені реконструювати навчальні зображення вихідного та цільового обличчя відповідно. Для обрізання та вирівнювання зображень використовується детектор обличчя. Для створення підробленого зображення необхідний навчений кодер і декодер застосовуються до цільового обличчя. Потім результат автокодера змішується з рештою зображення за допомогою пуассонівського редагування зображень.
3. **Face2Face** — це система відтворення обличчя, яка передає вирази обличчя з вихідного відео на цільове відео, зберігаючи ідентичність цільової особи. Оригінальна реалізація заснована на двох відеовходах потоки з ручним вибором ключових кадрів. Ці кадри використовуються для створення щільної реконструкції обличчя, яка може бути використана для повторного синтезу обличчя при різному освітленні та виразах.
4. **NeuralTextures** — це метод, що використовує оригінальні відеодані для вивчення нейронної текстури цільової людини, включаючи мережу рендерингу. Модель тренується з photometric reconstruction loss в поєднанні з adversarial loss. Цей підхід дозволяє змінювати міміку обличчя людини.

Включаючи в набір даних різноманітні методи маніпулювання deepfake, FaceForensics++ надає дослідникам широкий і складний набір відео для навчання й оцінки алгоритмів розпізнавання. Включення візуальних артефактів ще більше підвищує автентичність набору даних, забезпечуючи більш об’єктивне

представлення проблем, з якими стикаються під час виявлення підробок в реальних умовах.

2.1.3 Інші датасети

Набори даних FaceForensics++ і DFDC, привернули значну увагу і широко цитувалися та використовувалися дослідниками в цій галузі. Вони слугували еталонними наборами даних для навчання, оцінки та порівняння алгоритмів виявлення deepfake. Численні дослідницькі статті та публікації посилаються на ці набори даних, що свідчить про їхній вплив і використання в науковому співтоваристві. Але окрім цих датасетів досить популярними для використання вважаються наступні [7]:

1. **DeeperForensics-1.0** — це великомасштабний набір даних, який включає як реальні, так і оброблені відео. Він складається з понад 11000 фейкових відео згенерованих методом DFVAE(з використанням автокодерів).
2. **DFD (DeepFake Detection)**. Набір даних був опублікований Google Research у 2019 році. Датасет містить 3068 оброблених та 363 реальних відео, створених із застосуванням покращеного методу Deepfake. Дані зібрані з різних вебплатформ, щоб відобразити різноманіття та виклики, пов’язані з реалістичністю підробок.
3. **CelebDF** – цей набір містить 590 реальних та 5639 синтезованих відео зі знаменитостями, згенерованих вдосконаленим методом DF. Мета CelebDF — надати еталонний набір даних, який конкретно стосується виявлення deepfake за участю відомих громадських діячів.
4. **Deepfake-TIMIT** містить 640 підроблених відео, отриманих із 10 послідовностей зображень 32 людей, вилучених із набору даних VidTIMIT 11, згенерованих за допомогою алгоритму зміни обличчя на основі GAN. Автори вручну відібрали 16 пар людей, які мали певну візуальну схожість, і поміняли їхні обличчя. Відео розділені на дві основні категорії: 320 відео низької якості, з приблизно 200 кадрами 64 × 64 пікселів кожен і 320 відео

високої якості, із приблизно 400 кадрами розміром 128 × 128 пікселів.

5. **UADFV** — це синтетичний набір даних, наданий University of Albany з метою допомогти виявити підроблені відео з обличчям за допомогою фізіологічних сигналів. Наприклад, моргання очей, властивість, яку автори класифікують як недостатньо добре представлену в синтезованих відео. Набір даних складається з 49 підроблених відео, згенерованих через мобільний додаток FakeApp 10, де оригінальні обличчя людини замінюються обличчям Ніколаса Кейджа. Кожна послідовність має роздільну здатність 294 × 500 пікселів і в середньому тривалість 11,14 секунди.

Кожен з описаних вище наборів даних містить різні техніки маніпуляції та створення deepfake і може бути використаний як бенчмарк для поточного дослідження. Але головною проблемою залишається обмеженість та відображення конкретного періоду часу в еволюції методів deepfake, оскільки ця сфера постійно розвивається.

2.2 Моделі розпізнавання deepfake

Загалом задачу розпізнавання deepfake відео можна представити як бінарну класифікацію зображень облич з кадру відеофайлу. Алгоритми зазвичай включають кілька етапів або компонентів для аналізу та ідентифікації маніпуляційного вмісту. Хоча конкретні реалізації можуть відрізнятися, існують деякі загальні кроки, які можна виокремити в алгоритмах виявлення deepfake:

1. Preprocessing: вхідні відео попередньо обробляються для нормалізації даних і підготовки їх для наступного аналізу. Це може включати зміну розміру, обрізання або застосування корекції кольору для забезпечення узгодженості та оптимізації даних для моделі.
2. Feature Extraction: алгоритми перетворює попередньо оброблені вхідні в дані у вектор ознак. Цей крок спрямований на виявлення шаблонів підробок і характеристик, які відрізняють реальний вміст від маніпуляційного. Зазвичай використовувані методи включають згорткові шари, операції об’єднання та

інші методи виділення ознак, такі як аналіз оптичного потоку або аналіз частотної області.

3. Навчання моделі: отримані вектори ознак використовуються для навчання моделі глибокого навчання, наприклад згорткової нейронної мережі (CNN) або рекурентної нейронної мережі (RNN). Модель вчиться класифікувати та розрізняти реальний і підроблений контент на основі вилучених ознак. Навчання зазвичай виконується з використанням великого набору даних, який містить як справжні, так і підроблені зразки.
4. Оцінка моделі: навчена модель оцінюється за допомогою окремого тестового набору даних для оцінки її продуктивності. Використовують такі показники оцінки, як точність, precision, recall, і F1 score. Оцінка допомагає дослідникам точно налаштувати модель і порівняти її з іншими алгоритмами виявлення.
5. Post-processing: деякі алгоритми виявлення deepfake можуть включати кроки постобробки для уточнення прогнозів моделі або пом’якшення впливу помилкових результатів. Це може включати додатковий аналіз, методи фільтрації або confidence thresholding для підвищення точності та надійності виявлення.
6. Ensembling: ансамблеві підходи можуть бути використані для підвищення ефективності виявлення. Кілька моделей з різними архітектурами або стратегіями навчаються незалежно, а їхні прогнози об’єднуються або агрегуються для прийняття остаточного рішення. Ансамблеві методи можуть підвищити загальну точність і надійність виявлення.

Важливо зауважити, що кроки, описані вище, забезпечують загальну структуру для алгоритмів виявлення deepfake, але конкретні деталі та методи, які використовуються на кожному кроці, можуть відрізнятися залежно від дослідження чи реалізації. У наступних пунктах описано 2 основні кроки алгоритму: виокремлення облич на відео, та класифікація отриманих зображень.

2.2.1 Методи розпізнавання облич

Важливим кроком під час препроцесингу вхідного відео є ідентифікація обличчя людини в кадрі. Наразі існує досить багато методів вирішення цієї задачі, але найпопулярнішими можна назвати інструменти Dlib і MTCNN.

Dlib [8] — це потужна бібліотека з відкритим вихідним кодом, написана мовою C++, відома широким спектром функцій комп'ютерного зору. Детектор обличчя Dlib заснований на поєднанні функцій Histogram of Oriented Gradients (HOG) і класифікатора Support vector machine (SVM). Ось деякі ключові особливості Dlib:

1. Точність: детектор овідомий своєю точністю визначення облич на зображеннях і відео. Він добре працює за різних умов освітлення, орієнтацій і масштабів.
2. Швидкість: реалізація в Dlib оптимізована для швидкодії та може досягти виявлення обличчя в реальному часі на сучасному обладнанні.
3. Надійність: детектор розроблений для обробки складних сценаріїв, таких як оклюзії, частково приховані обличчя та нефронтальні пози.
4. Розпізнавання орієнтирів обличчя: на додаток до визначення обличчя Dlib надає модель визначення орієнтирів, яка може точно визначити місцезнаходження ключових точок обличчя, таких як очі, ніс і рот.

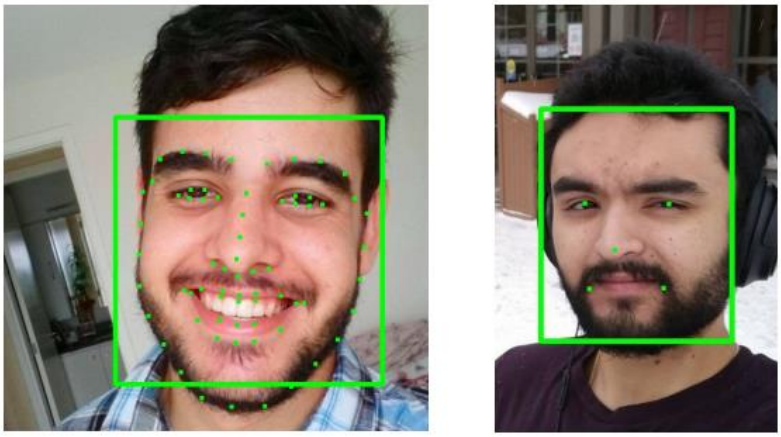
MTCNN(Multi-Task Cascaded Convolutional Neural Networks) [9] — це алгоритм розпізнавання на основі глибокого навчання, спеціально розроблений для виявлення облич на зображеннях і відео. Він складається з каскадної архітектури з трьома етапами згорткових нейронних мереж (CNN). Ось деякі ключові аспекти MTCNN:

1. Кілька етапів: MTCNN використовує багатоступеневу архітектуру для поступового вдосконалення процесу виявлення обличчя. На першому етапі пропонуються потенційні області обличчя, на другому етапі уточнюються пропозиції, а на третьому етапі виконується локалізація орієнтирів обличчя.
2. Точність: MTCNN продемонструвала вражаючу точність виявлення облич

у складних умовах, включаючи різні розміри облич, пози та оклюзії.

3. Локалізація орієнтирів обличчя: подібно до Dlib, MTCNN також забезпечує локалізацію орієнтирів обличчя на додаток до визначення обличчя. Ця функція дозволяє точно визначити риси обличчя.

І Dlib, і MTCNN пропонують надійні можливості виявлення облич, кожна з яких має свої сильні сторони та сфери застосування. Вибір між ними залежить від таких факторів, як вимоги до продуктивності, доступні ресурси та потреби конкретної програми. Дослідники та розробники часто враховують ці фактори, коли обирають найбільш підходящий метод визначення обличчя для своїх проектів. Приклад роботи детекторів облич зображено на рисунку 2.3 [11].



(а) (б)

Рисунок 2.3. Приклад роботи детекторів обличчя: (а) - Dlib, (б) - MTCNN

2.2.2 Алгоритми для класифікації deepfake.

Як було згадано раніше, методи виявлення deepfake можна розглядати як процес бінарної класифікації. На підставі особливостей методи виявлення таких підробок можна розділити на чотири типи:

1. Виявлення на основі візуальних функцій. Візуальний метод виявлення deepfake використовує особливості зображення, які можна побачити неозброєним оком, наприклад моргання, положення голови, а також відмінності у формах рис обличчя.
2. Виявлення на основі локальних функцій. Підхід використовує метод виділення ознак на основі пікселів, який виділяє особливості кожної

області зображення.

3. Глибоке навчання для виявлення deepfake на основі функцій. Подібно до підходу, заснованого на локальних функціях, цей метод також виконує процес вилучення ознак на рівні пікселів. Різниця полягає в тому, що процес вилучення об'єктів є багаторівневим, тому він може отримувати складніші об'єкти та пояснювати глибші зв'язки, ніж локальний підхід.
4. Виявлення глибоких фейків на основі тимчасових функцій. На відміну від інших трьох підходів, метод вилучає ознаки з кількох послідовних кадрів, щоб отримати часові особливості відео. Цей метод виявлення можна використовувати лише з відео.

Кожен з цих підходів необхідно розглянути більш детально [10]:

1) Виявлення Deepfake на основі візуальних функцій

У 2018 році вперше було запропоновано метод виявлення глибоких фейків на основі візуальних функцій. Припущення, яке використовується в цьому підході, полягає в тому, що існує різниця в блиманні очей між підробленим та оригінальним відео. Інший алгоритм візуальних особливостей звертає увагу на невідповідність поз голови. Цей підхід розраховує невідповідність між розташуванням обличчя та частинами тіла за межами обличчя, наприклад, шиєю та плечима. Також існує метод, який враховує деякі візуальні особливості в deepfake обличчях, які називаються візуальними артефактами (рисунок 2.4). Ці характеристики включають різницю в кольорі лівого та правого ока, непропорційні тіні, деталі відбиття світла, які не з'являються, і геометрію, яка не Для виділення цих візуальних особливостей використовується геометричний і кольоровий підхід вилучення з певних ділянок обличчя, таких як очі, ніс, брови, губи та зуби.

Рисунок 2.4. Візуальні артефакти deepfake.

Однак із розвитком методів генерації deepfake візуальні функції стає все важче виявляти, тому цей метод розпізнавання стає менш надійним.

2) Виявлення Deepfake на основі локальних функцій

Метод виявлення на основі локальних функцій використовує методи вилучення функцій, які виділяють особливості зображення на рівні пікселів. Дослідження, проведене в 2017 році, використовувало комбінацію двох функцій згортки та функції стегоналізу зображення для виявлення змінених ділянок обличчя на зображенні. Цей метод є основою методів виявлення на основі локальних і глибинних ознак. Ще один метод виділення ознак, який використовується в процесі виявлення глибоких фейків — це Scale Invariant Feature Transform (SIFT), який виявляє піксель ключової точки на зображенні та виділяє її ознаки з ключової точки. Методи розпізнавання підробок на основі локальних функцій мають досить хорошу ефективність у деяких відеоданих. Однак із розвитком алгоритму deepfake отримані зображення та відео стають більш реалістичними, тому задача вимагає більш складного аналізу зображень.

3) Deep Learning

Метод виявлення deepfake засобами машинного навчання зазвичай вимагає використання нейронних мереж глибокого рівня, які можуть виокремити ознаки, складніші за ознаки, отримані методами вилучення локальних ознак.

Дослідження, проведене в 2018 році, порівнювало кілька архітектур CNN, таких як InceptionNet, DenseNet і XceptionNet, у виявленні зображень підробок. Модель Xception [12], розширення архітектури Inception, продемонструвала значні

переваги для задач розпізнавання deepfakes. Ключові переваги використання підтверджені статистикою та дослідженнями:

1. Висока точність. Модель Xception досягла вражаючої точності в тестах. Наприклад, у DFDC найефективніші моделі, включаючи варіанти архітектури Xception, досягли показників AUC що перевищують 0,99.
2. Потужне визначення ознак. Згорткові шари моделі, які можна розділити по глибині, дозволяють отримувати дрібні деталі та складні візерунки на зображеннях.
3. Transfer learning. Модель має переваги від попереднього навчання на великомасштабних наборах даних, таких як ImageNet, який містить мільйони позначених зображень. Це попереднє навчання дозволяє моделі отримувати загальні візуальні знання, які можна точно налаштувати на наборах даних виявлення deepfake.
4. Ефективність обчислень. Модель забезпечує баланс між продуктивністю та ефективністю обчислень. Завдяки згортковим шарам, які можна розділити по глибині, модель зменшує кількість параметрів і операцій порівняно з традиційними архітектурами CNN. Ця ефективність є перевагою для програм виявлення глибоких фейків у реальному часі або систем з обмеженими обчислювальними ресурсами.
5. Можливість узагальнення. Навчання моделі Xception на різноманітних наборах даних покращує її здатність до узагальнення.

   Архітектура XceptionNet наведена на рисунку 2.5.

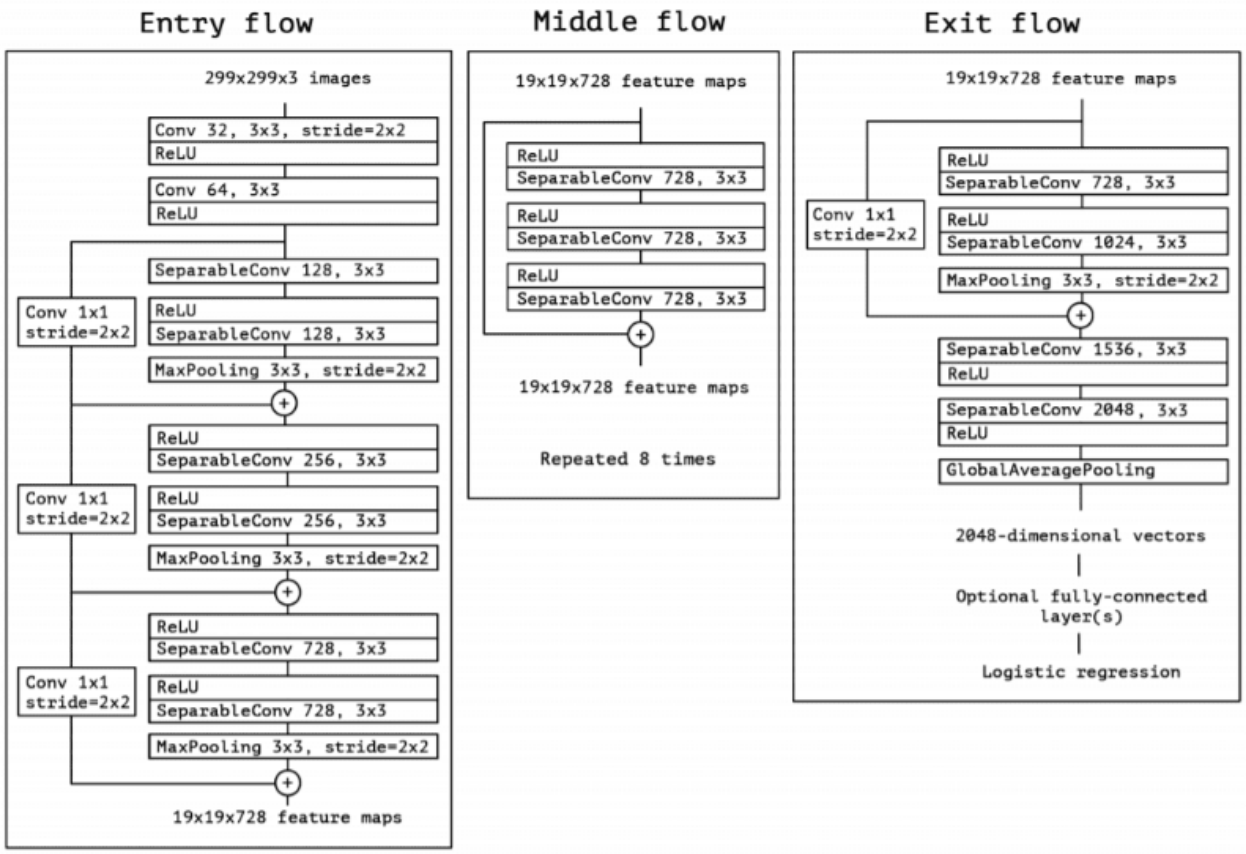


Рисунок 2.5. Архітектура мережі XceptionNet.

Іншою моделлю CNN, розробленою для виявлення deepfake, є MesoNet [13], яка використовує початковий модуль Inception як основу своєї архітектури. Ця модель здатна виявляти підроблені стиснуті відео, оскільки найчастіше такі матеріали поширюються соціальними мережами, які обмежують їхні розміри та якість. Однак дослідження з використанням архітектури Capsule Network здатне зрівнятись з продуктивністю MesoNet. Наприклад, CapsuleGAN, були досліджені для завдань виявлення deepfake. Ці моделі використовують шари капсули для захоплення просторових зв’язків і мають потенціал до підвищення надійності моделей проти змагальних атак.

4) Виявлення Deepfake на основі тимчасових функцій.

Це процес розпізнавання відео з використанням часових характеристик, отриманих із серії послідовних кадрів із відео. По суті, ряд кадрів можна розглядати як послідовні набори даних. Добре відомим методом обробки послідовних даних є рекурентні нейронні мережі (RNN). Досить непогано показали себе архітектури ResNet50 і DenseNet.

Загальне порівняння описаних методів на різних датасетах наведено на рисунку 2.6.

Metode	UADFV	DF-TIMIT		FF-DF	DFD	DFDC	Celeb-DF
		LQ	HQ				
Two-stream	85,1	83,5	73,5	70,1	**52,8**	**61,4**	**53,8**
Meso4	84,3	87,8	**68,4**	84,7	76	75,3	**54,8**
Face Warping *Artifact*	97,4	99,9	93,2	80,1	74,3	72,7	**56,9**
Head Pose	89	**55,1**	**53,2**	**47,3**	**56,1**	**55,9**	**54,6**
Visual Artifact	70,2	**61,4**	**62,1**	**66,4**	**69,1**	**61,9**	**55**
Xception	91,2	95,9	94,4	99,7	85,9	72,2	**65,3**
Multi-task	**65,8**	**62,2**	**55,3**	76,3	**54,1**	**53,6**	**54,3**
Capsule Network	**61,3**	78,4	74,4	96,6	**64**	**53,3**	**57,5**

Рисунок 2.6.Порівняння точності розпізнавання.

Кожен із методів виявлення глибоких фейків має переваги та недоліки. Підхід до виявлення deepfake із візуальними функціями може в деяких випадках виявляти підробки у зображеннях і відео. Ще одна перевага цього методі полягає в тому, що його легко інтерпретувати та пояснити людям. Однак із розвитком алгоритмів генерації deepfake контенту цей підхід стає неактуальним. Підхід із локальними функціями має кращу надійність, порівняно з візуальними функціями, оскільки він виділяє всі ділянки обличчя, а отримані риси є внутрішніми характеристиками зображення, які не видно візуально. Різниця між двома алгоритмами до 38% EER(Equal Error Rate) у наборі даних DF-TIMIT.

Підхід з глибоким навчанням, наразі має найвищий рівень точності, а саме архітектура XceptionNet, яка досягла рівня точності до 99,7% на наборі даних FaceForensics-DF. Проблема цього методу полягає в його низькому рівні інтерпретованості, оскільки він не може пояснити, яка область пікселів є частиною deepfake. Тим часом метод виявлення на основі тимчасових ознак має можливість захоплення часових особливостей, які неможливо зафіксувати за допомогою інших

підходів. Однак, згідно з результатами проведених досліджень, цей метод не зміг перевершити результативність методу на основі нейронних мереж [14].

2.3 Методи узагальнення

Незважаючи на велике різноманіття наборів даних для вивчення та розробки моделей розпізнавання deepfake і наявності суттєвих успіхів у цій сфері, все одно поточні рішення є обмеженими з точки зору актуальності цих рішень у майбутньому. Еволюція методів генерації підробленого контенту та можливість протиставляти будь-якій моделі розпізнавання зразки, які не будуть схожі на навчальні, призводитиме до поганих результатів. Тому постає питання розробки методів узагальнення та адаптації рішень до небачених тестових даних для забезпечення надійнішої роботи в процесі експлуатації.

Загального підходу до узагальнення як такого немає. Останнім часом розв’язанню проблеми узагальнення було присвячено декілька цікавих робіт. Наприклад, дослідження Face X-ray [15] припускає, що генерація deepfake відбувається шляхом змішування основного зображення і чужорідного обличчя, тому навчившись визначати межу змішування, можна класифікувати: для підробок її можна виокремити, а для реальних відео вона відсутня. Приклад наведено на рисунку 2.7.

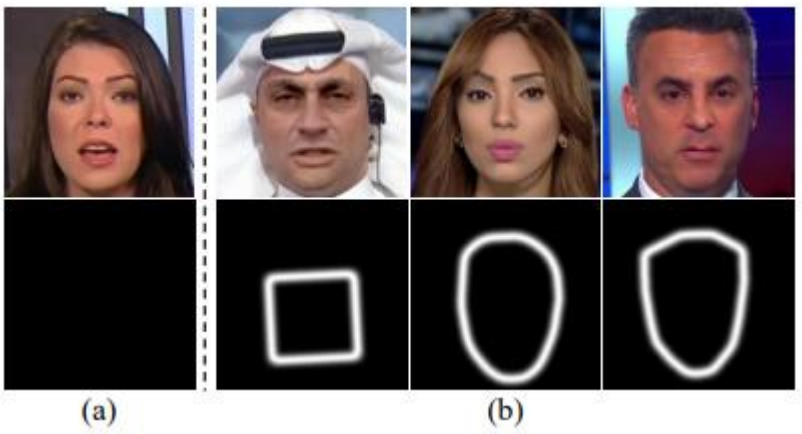


Рисунок 2.7. Визначення межі змішування в Face X-ray.

Масштабні експерименти показують, що цей підхід залишається

ефективним, якщо застосовувати його до підробок, створених небаченими під час навчання методами маніпулювання обличчями.

У статті [16] вчені посилаються на явище фазового спектру зображень. Новий метод просторово-фазового неглибокого навчання (SPSL), який поєднує просторове зображення та фазовий спектр для захоплення артефактів підвищення дискретизації підробки обличчя для покращення можливості передачі та розпізнавання deepfake.

У статті [17] дослідники виявляють, що поточні детектори на основі CNN мають тенденцію переналаштовуватися на специфічні для методу колірні текстури і, таким чином, не можуть узагальнити. Зважаючи на те, що шуми зображення видаляють колірні текстури та виявляють розбіжності між автентичними та підробленими областями, пропонується використовувати високочастотні шуми для виявлення підробки обличчя. Були розроблені 3 модулі, щоб повною мірою використовувати переваги високочастотних функцій. Приклад їхньої роботи показаний на рисунку 2.8.

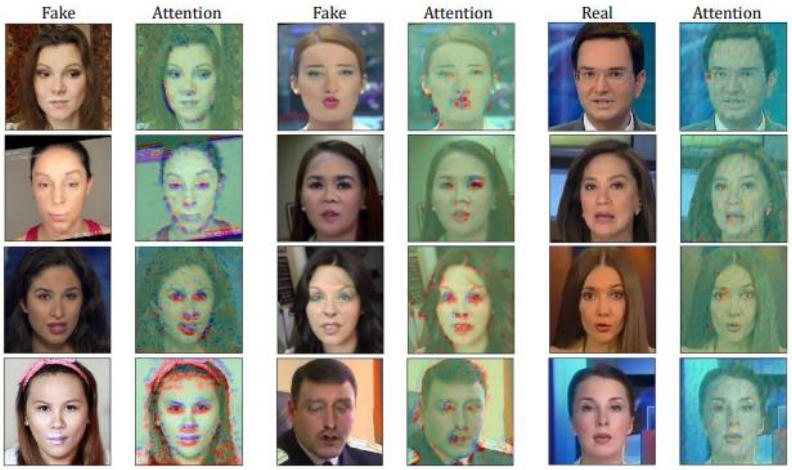


Рисунок 2.8. Робота різних модулів визначення шумів.

Хоча ці методи є ефективними в багатьох випадках, низькорівневі артефакти, на які вони спираються, чутливі до етапів постобробки, які відрізняються в різних наборах даних, що ставить під загрозу їх узагальнення. У деяких інших роботах пропонується запозичити ознаки з інших завдань, таких як читання з губ, декомпозиція зображення обличчя і геометричні орієнтири, що можуть натякати на ненормальність підробок. Хоча ці функції можуть принести певні покращення для

існуючих моделей розпізнавання, є велика ймовірність того, що майбутні алгоритми створення deepfake будуть розроблені на основі цих досліджень, щоб синтезувати більш реалістичні підробки, створюючи ще більші загрози безпеці цифрового середовища. Тому спиратись на особливості зображення для досягнення узагальнення може бути не найефективнішим підходом.

2.4. Висновки до розділу 2

На сьогодні задача розпізнавання deepfake контенту має досить багато розв’язань. Створено окремі набори даних для розробки алгоритмів, деякі з них вже встигли стати канонічними та дозволяють дослідникам об’єктивно оцінювати свої методи на спільних даних, порівнюючи ефективність. Фактично задачу виявлення підробки обличчя можна розділити на дві частини: розпізнавання облич та класифікація deepfake це чи ні. Якщо для проблеми розпізнавання обличчя на відео уже є сталі ефективні рішення (Dlib, MTCNN), то питання безпосередньо класифікації має ще пройти шлях вдосконалення. Готові рішення, такі як XceptionNet показують блискучі результати при навчанні та тестуванні на визначеному наборі даних.

Але в умовах реального світу і роботи з небаченими типами підробок якість таких рішень суттєво знижується. Тому необхідно працювати над створенням алгоритмів узагальнення та адаптації моделей класифікації. Дослідження в цьому напрямку вже ведуться, деякі з них показують хороші результати. Проте, в основному, вони зосереджуються на відмінностях в характеристиках зображень, реальних та підроблених. Такий підхід може виявитись неефективним в умовах доступності таких досліджень для вчених, які розробляють методи генерації deepfake, вони можуть адаптуватись та мінімізувати різницю між вихідним та обробленим контентом.

РОЗДІЛ 3: ОПИС РІШЕННЯ ТА ПРОГРАМНОГО ПРОДУКТУ

3.1. Навчання під час тестування

З попереднього аналізу існуючих методів узагальнення стає зрозуміло, що більшість направлені на пошук відмінностей на рівні пікселів і насправді залишають можливість створювати підробки які зможуть мінімізувати такі візуальні артефакти. Тому для адаптації моделі розпізнавання до небаченого раніше методу deepfake можна припустити що найефективнішим рішенням буде «показати» моделі такий зразок та дозволити пристосуватись. Ключовим в цій парадигмі є запобігання перенавчанню під час тесту, яке, очевидно, дуже ймовірне в умовах навчання на тестових зразках і подальшій класифікації цих зразків.

Концепція Test-Time Training (TTT) була вперше описана в [18] . Автори підкреслюють складність розгортання моделей глибокого навчання в реальних сценаріях, де розподіл даних може відрізнятися від розподілу, який спостерігається під час навчання. Цей зсув розподілу може призвести до погіршення продуктивності моделі, оскільки модель намагається узагальнити невидимі дані. Щоб подолати цю проблему, дослідники пропонують підхід до навчання під час тестування, який використовує самоконтроль. Самонагляд передбачає навчання моделі прогнозуванню певних властивостей або перетворень даних, не покладаючись на зовнішні мітки. Ключова ідея полягає у використанні окремої самоконтрольованої моделі, яка називається «модель вчителя», для генерації псевдоміток для даних тесту. Потім ці псевдомітки використовуються для навчання основної моделі, яку називають «моделлю студента», під час тестування. Цей процес дозволяє моделі студента вдосконалювати свої представлення та адаптуватися до зміни розподілу, підвищуючи ефективність узагальнення. Автори оцінюють свій підхід до навчання під час тестування на різних еталонних наборах даних, включаючи CIFAR-10, CIFAR-100 та ImageNet. Вони порівнюють свій метод з іншими базовими підходами та демонструють значні покращення в точності моделі за умов змін розподілу.

Однак, незважаючи на деякі обнадійливі результати, поточні методи TTT спрямовані на вибір емпіричних самоконтрольованих завдань, що має високий ризик погіршення продуктивності, якщо завдання не вибрано належним чином.

Ця робота адаптує концепцію TTT спеціально для завдання виявлення deepfake. Підробку можна легко синтезувати шляхом змішування двох різних зображень облич, тому доцільно буде використати її як псевдонавчальний зразок для точного налаштування попередньо навченої моделі під час прогнозування класу.

На цьому етапі ми припускаємо, що модель розпізнавання вже навченна, відповідно ваги моделі дорівнюють θ. Алгоритм навчання під час тестування оновлює ваги, що дозволяє моделі адаптуватись до кожного окремого тестового зображення. Це досягається шляхом, спочатку, генерації псевдонавчального зразка, а потім використання його для формування міні-навчального набору і оновлення параметрів моделі за допомогою одноразового навчання. Розглянемо кожен крок окремо.

1. Генерація псевдонавчального зразка.

Створення псевдонавчальних зразків: як показано на рисунку 3.1, для кожного тестового зразка спочатку випадковим чином обирається шаблонне зображення із навчального набору даних, вони вирівнюються на основі орієнтирів обличчя. Потім виділяється опукла оболонка, використовуючи орієнтири тестового зразка, отримуємо деформовану остаточну маску, застосовуючи випадкову деформацію та розмиття опуклої оболонки. За допомогою деформованої маски ми можемо обрізати відповідний вміст обличчя з корекцією кольору з шаблонного зразка. Нарешті, використовуючи тестове зображення, деформовану маску та обрізане шаблонне обличчя, за допомогою методів альфа та пуассонівського змішування, техніки face swap отримуємо сукупність псевдонавчальних зразків. Далі випадковим чином обирається одна підробка з набору. Всі згенеровані зразки псевдонавчання позначені як негативні (тобто підроблені). Важливо, що хоча ми не знаємо, чи є тестове зображення справжнім чи підробленим ми впевнені, що синтезоване псевдозображення є підробкою.

Рисунок 3.1. Генерація псевдонавчальних зразків

2. Однокрокове навчання.

З навчального набору обирається зображення з відомою міткою(не важливо підробка це чи ні), для тестового зразка генерується псевдонавчальний зразок відмічений завжди як підробка, з цих зображень формується міні-тренінговий набір. Потім виконується операція однокрокового градієнтного спуску для оновлення параметрів моделі. Алгоритм можна описати наступним чином:

1) Дано: навчена модель $f(x, \theta)$, learning rate γ, тестовий зразок x_{test} , навчальний датасет D.
2) Випадково обираємо x_{temp} з D і синтезуємо підробку, змішуючи x_{temp} і x_{test}.
3) Обчислюємо функцію втрат:

$$\sum_{x_i \in \{x_{test}, x_{temp}\}} L(f(x_i, \theta), y_i)$$

4) Оновлюємо параметри моделі через градієнтний спуск:

$$\theta^* = \theta - \gamma \sum_{x_i \in \{x_{test}, x_{temp}\}} \nabla_\theta L(f(x_i, \theta), y_i)$$

Після етапу оновлення ми можемо отримати остаточний результат класифікації, застосувавши оновлений детектор $f(x, \theta^*)$ до тестового зразка. Запропонований алгоритм навчання під час тестування також можна розуміти як метод адаптації домену. Зокрема, в такому алгоритмі навчання кожне зображення

розглядається як певний домен, визначений його вмістом, і невідоме тестове зображення матиме різницю домену відносно навчальних даних. Псевдонавчальна вибірка, створена за допомогою описаної вище процедури, більш релевантна тестовому зображенню, ніж навчальна, оскільки синтезується на основі тестового зображення. Таким чином, навчання під час навчання на псевдозразку може змусити модель краще адаптуватися до тестових зображень.

3.2. Мета-навчання MAML

У деяких нещодавніх роботах структура TTT також використовувалася в парадигмі Model-Agnostic Meta-Learning (MAML) [19], яка забезпечує швидку адаптацію глибоких нейронних мереж до нових завдань з обмеженими даними. Автори вирішують проблему навчання моделей глибокого навчання на завданнях із невеликими наборами даних, де традиційні методи намагаються досягти задовільного результату. Вони пропонують підхід до метанавчання, який спрямований на вивчення ініціалізації моделі, яка може швидко адаптуватися до нових завдань лише на кількох прикладах.

Ключова ідея полягає в тому, щоб навчити «учня», який називається мета-модель, виконувати різноманітні пов’язані завдання. Мета-модель вчиться фіксувати знання, що не залежать від завдань, і узагальнювати різні завдання. Під час фази адаптації мета-модель налаштовується з використанням невеликої кількості даних, що стосуються конкретного завдання, що дозволяє їй швидко адаптуватися до нового завдання. Також, автори представляють алгоритм оптимізації, який називають алгоритмом «reptile», який оптимізує ініціалізацію мета-моделі та сприяє ефективній адаптації до нових завдань. Алгоритм ітеративно оновлює параметри моделі, обчислюючи градієнти втрат, що стосуються конкретного завдання, і застосовуючи ці градієнти в напрямку, що покращує продуктивність завдань.

Ефективність запропонованого підходу оцінюється на різних еталонних наборах даних і завданнях, включаючи задачі класифікації зображень, навчання з підкріпленням і регресії. Результати демонструють, що навчені у такий спосіб

моделі можуть досягти високих результатів з набагато меншою кількістю навчальних даних порівняно з традиційними методами навчання.

Запропонований підхід MAML під час навчання моделі розпізнавання покликаний забезпечити хорошу ініціалізацію для кроку навчання під час тестування. В рамках цієї роботи алгоритм навчання моделі розпізнавання deepfake можна описати наступним чином:

1) Дано: розмір батчу T, learning rates λ та γ навчальний датасет D
2) Цикл навчання:
3) Цикл для кожного t з T:
4) Обираємо x_r^1 та x_r^2 з датасету і синтезуємо підробки x_f^1 та x_f^2, різними операціями змішування, скажімо, один з альфа-змішування та один із змішуванням Пуассона.
5) (x_r^1, x_f^1) використовуємо для навчання, (x_r^2, x_f^2) - для обчислення втрат.
6) Обчислюємо втрати

$$L_t(f(x_i, \theta), y_i), \qquad x_i \in (x_r^1, x_f^1)$$

7) Обчислюємо адаптовані параметри моделі:

$$\theta_t^* = \theta - \gamma \nabla_\theta L_t(f(x_i, \theta), y_i), \qquad x_i \in (x_r^1, x_f^1)$$

8) Кінець циклу по T
9) Обчислюємо мета втрати під час навчання:

$$\sum_t L_t(f(x_i, \theta_t^*), y_i), \quad x_i \in (x_r^2, x_f^2)$$

10) Оновлюємо параметри моделі

$$\theta \leftarrow \theta - \lambda \nabla_\theta \frac{1}{T} \sum_t L_t(f(x_i, \theta_t^*), y_i), \quad x_i \in (x_r^2, x_f^2)$$

11) Кінець навчання

У наведеному вище алгоритмі x_r^1 і x_f^1 схожі на support set, а x_r^2 і x_f^2 схожі на query set оригінальної статті. Важливо, що у query set включаються як x_r^2, так і x_f^2. Вони представляють два різні сценарії: x_r^2 — це зображення з початкового навчального домену, а x_f^2 — зображення з нового (синтезованого випадковим чином) домену. Включення обох зображень заохочує модель добре працювати для

існуючого домену та мати можливість узагальнювати на новий домен.

3.3. Опис програмного продукту

Запропонований метод складається з етапу мета-навчання та кроку оновлення параметрів моделі під час тестування. На етапі навчання під час тесту спочатку буде створено псевдонавчальний зразок для кожного тестового зображення, а потім буде виконано одне оновлення ваг градієнтним спуском, щоб отримати параметри моделі, специфічні для зразка. Мета-навчання імітує операцію навчання під час тесту шляхом вибірки епізодів із даних навчання.

В першу чергу, для подальшого аналізу ефективності та можливості порівняння результатів з іншими дослідженнями, необхідно обрати еталонний набір даних для навчання. Проаналізувавши можливі варіанти, для навчання було обрано набір даних FaceForensics++, описаний у пункті 2.1.2., оскільки цей набір містить відео, оброблені різними техніками deepfake та дозволяє порівнювати результат роботи детектора на різних доменних областях.

В якості baseline моделі розпізнавання deepfake, яку потрібно додатково навчити адаптації до тестових зразків, було вирішено використовувати архітектуру XceptionNet (пункт 2.2.2), попередньо навчену на датасеті ImageNet. Аспекти реалізації та попередньо сформовані параметри моделі доступні в мережі та можуть бути легко імплементовані. В ході експериментів з навчанням нейронної мережі, було визначено наступні гіперпараметри, які використовуються в цьому дослідженні: для оптимізатора Adam $\beta_1 = 0.9, \beta_2 = 0.99$; розмір батчу для кроку мета-навчання – 20.

Для вилучення облич з відеофайлів під час навчання та тестування використовується бібліотека Dlib(пункт 2.2.1). Вона є досить ефективною та також дозволяє окрім обличчя, вилучати також необхідні landmarks для генерації псевдонавчальних зразків. Згідно з налаштуваннями FF++, з кожного відеофайлу вилучається 270 кадрів. Потім виявлені в процесі розкадровування обличчя, вирізаються з зображення, масштабуються до розмірів 256x256 пікселів, а після нормалізуються до [-1;1].Також для навчання моделі не використовується жодних

додаткових аугментацій навчального датасету.

Використовуючи результати дослідження пов’язаного з обробкою відео контенту [20], для цієї роботи були застосовані аналогічні інструменти. Для реалізації цього проекту була обрана мова програмування Python 3.9. Для побудови та розгортання нейронної мережі був використаний фреймворк PyTorch. В умовах обмежених ресурсів локальної машини, навчання моделей виконувалось на платформі Google Colab, з можливістю підключення до GPU.

Перелік основних бібліотек для реалізації запропонованого методу:

1. Decord. Бібліотека для швидкої обробки відеоряду. Працює швидше ніж стандартні інструменти зчитування.
2. OpenCV, skimage. Класичні бібліотеки для розв’язання задач комп’ютерного зору. Мають великий набір інструментів та вбудованих алгоритмів.
3. Dlib. Бібліотека для вилучення облич та landmarks з зображень. Використовувалась pretrained версія детектора.
4. Для створення псевдонавчальних зразків використовували open-source реалізації таких алгоритмів: faceswap, VAE-autoencoder, face-blend.
5. Numpy, Scipy, Matplotlib. Стандартні бібліотеки для обчислень та побудови графіків.

Для демонстрації реалізованого рішення та практичного використання було вирішено використовувати платформу **Streamlit.io**. Ця платформа прямо інтегрується з Python і дозволяє швидко створити користувацький інтерфейс для завдань машинного навчання. Крім того, вона надає інструменти для швидкого розгортання веб-додатків. Графічний інтерфейс показано на рисунку 3.2.

Рисунок. 3.2 Графічний інтерфейс програми.

Користувач може завантажити відео у форматі MP4. Після натискання кнопки «Detect» алгоритм починає свою роботу. У вікні з'являється відеопрогравач, який відтворює вхідне відео. Відбувається оновлення параметрів моделі відповідно до тестового зразка та передбачення класу для кадрів відео. Обробка займає певний час, після завершення роботи алгоритму у вікні з’являється відео з результатами виявлення deepfake. Користувач також може завантажити результати розпізнавання.

На боковій панелі (рисунок 3.3.) користувач може вибрати частоту розпізнавання кадрів, щоб прискорити роботу алгоритму. Для більш детального аналізу рекомендується оцінювати кожен кадр, але цей процес може зайняти більше часу.

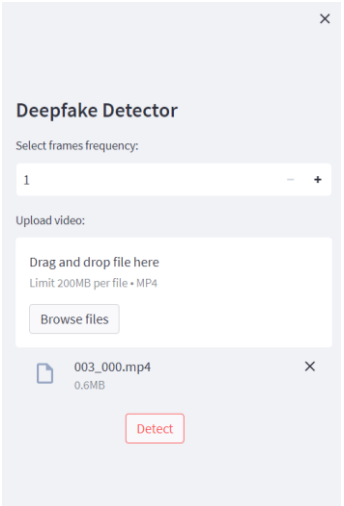


Рисунок. 3.3 Бічна панель інтерфейсу користувача.

3.4. Аналіз отриманих результатів

Пропонуючи метод узагальнення конкретної моделі розпізнавання deepfake неможливо не порівняти результати роботи з baseline моделлю. Також, вважаючи, що існують інші дослідження методів адаптації, описані у пункті 2.3., релевантно буде порівняти результати роботи відносно них. Вибір для навчання саме датасету FaceForesics++ дає можливість порівняти роботи методу Face X-Ray () та метод SRM (), оскільки ці дослідження також беруть за основу архітектуру XceptionNet.

У реальних сценаріях підроблені зображення, ймовірно, створюватимуться невидимими методами з невидимих джерел, тому можливість узагальнення для різних наборів даних буде надзвичайно важливою. Щоб оцінити можливість узагальнення, були проведені наступні експерименти. Навчаючи метод на кожному з чотирьох типів підроблених даних у FF++ (тобто DeepFake, Face2Face, FaceSwap і NeuralTextures), перевірка виконується на контрольних наборах даних, включаючи еталонні датасети DFD, DFDC. Таке налаштування дозволяє змоделювати реальні умови, оскільки всі методи deepfake, які синтезують підробки в тестових даних, не прослідковуються під час навчання. Значення метрики AUC під час тестування наведені в Таблиці 1.

Таблиця 1. Результати порівняння

Навчальні дані	Алгоритми	DFD	DFDC
FF++ (DF)	XceptionNet	0.808	0.654
	Face X-ray	0.811	0.609
	SRM	0.855	0.679
	Запропонований підхід	**0.867**	**0.755**
FF++ (F2F)	XceptionNet	0.695	0.708
	Face X-ray	0.679	0.633
	SRM	0.801	0.687
	Запропонований підхід	**0.882**	**0.798**
FF++ (FS)	XceptionNet	0.657	0.708
	Face X-ray	0.625	0.646

	SRM	0.705	0.679
	Запропонований підхід	**0.821**	**0.803**
FF++ (NT)	XceptionNet	0.724	0.646
	Face X-ray	0.645	0.613
	SRM	0.791	0.656
	Запропонований підхід	**0.837**	**0.756**

Розглянуті дослідження покладаються на непомітні візуальні артефакти для узагальнення, але ці артефакти можуть виглядати по-різному у підробках з різних наборів даних, що обмежує їхню адаптацію. Для порівняння, запропонований метод пропонує точне налаштування попередньо навченого детектора до тестових даних, і це забезпечує адаптацію до зразків перед оцінкою, таким чином покращуючи узагальнення. Водночас можна спостерігати, що метод перевершує інші моделі в розглянутому тестуванні. Середнє значення метрики AUC досягає 0.814. Алгоритм SRM, який є другим за показниками, створений підхід перевершує на 10% за загальною ефективністю. Ці результати наочно демонструють життєздатність та успішність запропонованої стратегії для узагальнення методів розпізнавання deepfake.

Також слід враховувати, не ідеальність відпрацьовування модуля Dlib для виявлення обличчя, що може вплинути на ефективність роботи в реальних умовах. Окремо варто зауважити, що незважаючи на оптимізацію складових алгоритму для швидкодії, етап навчання під час тестування вимагає більше часу для класифікації, ніж стандартні нейронні мережі. Це може впливати на показники продуктивності під час практичного застосування. Вирішенням цієї проблеми може бути використання GPU для алгоритму та паралелізація процесів.

3.5. Висновки до розділу 3

Для узагальнення моделі розпізнавання deepfake запропоновано алгоритм з використанням парадигм Test-Time Training та Model-Agnostic Meta-Learning. В рамках дослідження було створено програмний продукт з користувацьким

інтерфейсом та проведені експерименти для оцінки ефективності. Порівняно з існуючими детекторами, переваги методу полягають у наступному: використання структури навчання під час тесту та мета-навчання, робить можливою швидку адаптацію навченої моделі до тестових даних, що можуть надходити з домену, відмінного від навчальних даних; навчання під час тестування не покладається на створені вручну візуальні чи інші артефакти зображень, що залишає менше можливостей для вдосконалення у відповідь алгоритмів deepfake.

Висновки по роботі та рекомендації для подальших досліджень

У першому розділі даної роботи була описана проблематика та особливості використання технології Deepfake. Окрім застосування в індустрії розваг, цей спосіб обробки медіаданих стає все більш популярним для шахрайських цілей, введення в оману, маніпуляцій людьми, впливом на політику держав. Для усвідомлення потенціалу розвитку алгоритмів генерації deepfake, були проаналізовані відповідні методи: автокодери та генеративно-змагальні мережі (GAN). На сьогоднішній день ці інструменти активно вдосконалюються, досягають високої реалістичності контенту. Застосунки для створення підробок є у вільному доступі та користуються популярністю. Саме тому, важливо відповідально ставитись до питання цифрової гігієни, щоб зменшити потенційні ризики та запобігти зловмисному використанню технології deepfake. Активну участь у підтримці безпеки інформаційного середовища також відіграють поточні дослідження методів виявлення підробок.

У другому розділі було формалізовано задачу розпізнавання deepfake відео методами машинного навчання, описано основні складові для дослідження. Було розглянуто різні набори даних, такі як DFDC та FaceForensics ++, які використовуються для навчання та оцінки алгоритмів розпізнавання deepfake облич людей на відео. Також були проаналізовані підходи до розпізнавання підробок, включаючи методи вилучення облич з зображень та алгоритми для класифікації deepfake. Враховуючи ціль поточного дослідження, а саме створення узагальненого алгоритму виявлення deepfake, було вивчено існуючі методи вирішення цього питання. Наведені дослідження показують кращу ефективність роботи, порівняно з класичними моделями машинного навчання без адаптацій. Проте, в більшості, вони зосереджуються на пошуку візуальних артефактів на зображеннях, які залишаються в процесі маніпуляцій контенту. Це залишає можливість для вдосконалення технологій deepfake, щоб мати змогу не лишати слідів, що робить такі алгоритми виявлення менш ефективними в майбутньому.

У третьому розділі було описано запропоноване рішення для узагальнення

моделі розпізнавання підробок обличчя на відео, з метою адаптації її до тестових даних, створених новими методами deepfake, які можуть суттєво відрізнятись від навчальних даних. У якості базової моделі було обрано XceptionNet, яка показує хороші результати в тестуваннях на різних датасетах. Також, на цю архітектуру спирається більшість досіджень по узагальненню, що дає можливість порівняти з ними запропонований метод. Було розглянуто підхід навчання під час тестування та метод мета-навчання MAML, що дозволяє швидше адаптувати моделі до даних з різних доменів. Реалізовано програмний продукт, створено користувацький інтерфейс для демонстрації роботи алгоритму. Також були проведені експерименти та аналіз результатів роботи алгоритму в порівнянні з іншими методами узагальнення. Результати підкреслюють ефективність у запропонованого підходу для розпізнавання deepfake відео, що були згенеровані у спосіб відмінний від навчальних даних.

Продовжуючи дослідження в області CVPR, початі під час роботи над алгоритмом вилучення ключових фрагментів зображень у системах відеопошуку [20], було визначено характер та мету поточної роботи. Попередні напрацювання дозволили швидше реалізувати запропнований алгоритм.

Загалом, дана робота зосереджується на проблематиці deepfake контенту, оглядає існуючі рішення та надає власне рішення у вигляді алгоритму та програмного продукту для розпізнавання підробленого відеоконтенту. Перевагою отриманої в результаті архітектури є адаптація стандартних моделей машинного навчання до роботи в реальних умовах. В роботі реалізовано алгоритм за допомогою мережі XceptionNet, але запропонований підхід може бути застосований відносно моделі будь-якої архітектури. Створений програмний продукт можна імплементовувати у відповідні системи перевірки контенту або використовувати як окремий інструмент. Розробка з використанням open-source бібліотек та фреймворків робить таке рішення простим у реалізації. Єдиним обмеженням, як у випадку з будь-якими системами обробки візуального контенту, є потужність процесорів. Рекомендується використовувати GPU для швидкодії.

В рамках подальших кроків та досліджень, необхідно провести тестування

при різних рівнях стиснення і спотворення зображень, щоб оцінити ефективність алгоритму в реальних умовах. Крім того, важливо додавати нові методи генерації deepfake контенту для створення псевдонавчальних зразків, що дозволить покращити роботу моделі та забезпечити більш точне розпізнавання. Також важливо оптимізовувати час відпрацювання алгоритму і вдосконалювати baseline модель, можливо замінюючи її на більш нові та ефективні архітектури, у випадку їх появи. Ці напрямки досліджень сприятимуть подальшому розвитку і вдосконаленню методів боротьби з deepfake.

Список літератури

1. M. Masood, M. Nawaz, K. Malik, A. Javed, і A. Irtaza, *Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward*. [Електронний ресурс] – 2021 – Режим доступу: https://www.researchgate.net/publication/349703826_Deepfakes_generation_and_detection_state-of-the-art_open_challenges_countermeasures_and_way_forward

2. A. Zucconi, *Understanding the Technology Behind DeepFakes*. [Електронний ресурс] – 2018 Режим доступу: https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/

3. I. J. Goodfellow, *Generative Adversarial Networks* [Електронний ресурс] – 2014– Режим доступу: 10.48550/arXiv.1406.2661.

4. J. Brownlee, *A Gentle Introduction to Generative Adversarial Networks (GANs)* [Електронний ресурс] – 2019 – Режим доступу: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

5. B. Dolhansky, *The DeepFake Detection Challenge (DFDC) Dataset* [Електронний ресурс] – 2020 – Режим доступу: 10.48550/arXiv.2006.07397.

6. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, і M. Niessner, *FaceForensics++: Learning to Detect Manipulated Facial Images* [Електронний ресурс] – 2019 – Режим доступу: https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf

7. L. A. Passos, D. Jodas, K. A. P. da Costa, L. A. S. Júnior, D. Colombo, і J. P. Papa, *A Review of Deep Learning-based Approaches for Deepfake Content Detection* [Електронний ресурс] – 2022 – Режим доступу: http://arxiv.org/abs/2202.06095

8. Офіційний сайт Dlib http://dlib.net/

9. K. Zhang, Z. Zhang, Z. Li, і Y. Qiao, *Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks*, [Електронний ресурс] – 2016 – Режим доступу: https://arxiv.org/abs/1604.02878

10. K. N. Ramadhani і R. Munir, *A Comparative Study of Deepfake Video Detection Method* [Електронний ресурс] – 2020 – Режим доступу:

https://informatika.stei.itb.ac.id/~rinaldi.munir/Penelitian/Makalah-ICOIACT-2020.pdf

11. G. A. Bezerra і R. B. Gomes, *Recognition of Occluded and Lateral Faces Using MTCNN, DLIB and Homographies* [Електронний ресурс] – 2018 – Режим доступу: http://sibgrapi.sid.inpe.br/col/sid.inpe.br/sibgrapi/2018/10.16.17.36/doc/Recognition_of_Occluded_and_Lateral_Faces.pdf

12. F. Chollet, *Xception: Deep Learning with Depthwise Separable Convolutions* [Електронний ресурс] – 2017 – Режим доступу: https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf

13. D. Afchar, V. Nozick, J. Yamagishi, і I. Echizen, *MesoNet: a Compact Facial Video Forgery Detection Network* [Електронний ресурс] – 2018 – Режим доступу: https://www.semanticscholar.org/reader/4258fd74a8bbb03b0997e76ba8ab6684b161f9d3

14. U. A. Ciftci і I. Demir, *FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals* [Електронний ресурс] – 2020 – Режим доступу: https://arxiv.org/pdf/1901.02212.pdf

15. L. Li, *Face X-ray for More General Face Forgery Detection* [Електронний ресурс] – 2020 – Режим доступу: http://arxiv.org/abs/1912.13458

16. H. Liu, *Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain* [Електронний ресурс] – 2021 – Режим доступу: http://arxiv.org/abs/2103.01856

17. Y. Luo, Y. Zhang, J. Yan, і W. Liu, *Generalizing Face Forgery Detection with High-frequency Features* [Електронний ресурс] – 2021 – Режим доступу: http://arxiv.org/abs/2103.12376

18. Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, і M. Hardt, *Test-Time Training with Self-Supervision for Generalization under Distribution Shifts* [Електронний ресурс] – 2020 – Режим доступу: http://arxiv.org/abs/1909.13231

19. C. Finn, P. Abbeel, і S. Levine, *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks* [Електронний ресурс] – 2017 – Режим доступу: http://arxiv.org/abs/1703.03400

20. А.Афонін, І.Оксюта *Алгоритм вилучення ключових фрагментів зображень у системах відеопошуку* [Електронний ресурс] –2022 – Режим доступу: http://nrpcomp.ukma.edu.ua/article/view/275215/270319