УДК 004.75

DOi 10.30837/ bi.2020.2(95).05

В.В. Білоцерковський[1],С.Г. Удовенко[2], Л.Е. Чала [3]

[1]ХНУРЕ, м. Харків, Україна,
vladyslav.bilotserkovskyi@nure.ua, ORCID iD: 0000-0003-4001-0660

[2]ХНЕУ ім. С. Кузнеця, м. Харків, Україна,
serhiy.udovenko@hneu.net, ORCID iD: 0000-0001-5945-8647

[1]ХНУРЕ, м. Харків, Україна,
larysa.chala@nure.ua, ORCID iD: 0000-0002-9890-4790

МЕТОД НЕЙРОМЕРЕЖЕВОГО РОЗПІЗНАВАННЯ ФАЛЬСИФІКОВАНИХ ЗОБРАЖЕНЬ

Розглянуто методи генерації зображень, фальсифікованих за допомогою технологій Deepfake, і методи їх виявлення. Пропонується метод виявлення фальсифікованих зображень, який оснований на спільному використанні ансамблю згорткових нейронних моделей, механізму Attention та стратегії сіамського навчання мережі. Ансамблі моделей формувалися різними способами (з використанням двох, трьох або більшої кількості складових). Результат обчислювався як середнє значення показників AUC і LogLoss з усіх моделей, що входять в ансамбль. Такий підхід дозволяє покращити точність різних нейромережевих класифікаторів для виявлення статичних та динамічних зображень, створених за технологіями Deepfake.

ТЕХНОЛОГІЯ DEEPFAKE, РОЗПІЗНАВАННЯ ФАЛЬСИФІКОВАНИХ ЗОБРАЖЕНЬ, ЗГОРТКОВА НЕЙРОННА МЕРЕЖА, ATTENTION МЕХАНІЗМ, АНСАМБЛЬ МОДЕЛЕЙ

Белоцерковский В.В., Удовенко С.Г., Чалая Л.Э. Метод нейросетевого распознавания фальсифицированных изображений. Рассмотрены методы генерации изображений, фальсифицированных с помощью технологий Deepfake, и методы их выявления. Предлагается метод выявления фальсифицированных изображений, основанный на совместном использовании ансамбля сверточных нейронных моделей, механизма Attention и стратегии сиамского обучения сети. Ансамбли моделей формировались разными способами (с использованием двух, трех или более составляющих). Результат исчислялся как среднее значение показателей AUC и LogLoss из всех моделей, входящих в ансамбль. Такой подход позволяет улучшить точность нейросетевых классификаторов для выявления статических и динамических изображений, созданных по технологиям Deepfake.

ТЕХНОЛОГИЯ DEEPFAKE, РАСПОЗНАВАНИЕ ФАЛЬСИФИЦИРОВАННЫХ ИЗОБРАЖЕНИЙ, СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ, ATTENTION МЕХАНИЗМ, АНСАМБЛЬ МОДЕЛЕЙ

Bilotserkovskyy V.V., Udovenko S.G., Chala L.E. Method of neural network recognition of falsified images. Methods for generating images falsified using Deepfake technologies and methods for detecting them are considered. A method for detecting falsified images is proposed, based on the joint use of an ensemble of convolutional neural models, the Attention mechanism and a Siamese network learning strategy. The ensembles of models were formed in different ways (using two, three or more components). The result was calculated as the average value of the AUC and LogLoss indices from all the models included in the ensemble. This approach improves the accuracy of convolutional neural network classifiers for detecting static and dynamic images created using Deepfake technologies.

DEEPFAKE TECHNOLOGY, FALSE IMAGE RECOGNITION, CONVOLUTIONAL NEURAL NETWORK, ATTENTION MECHANISM, MODEL ENSEMBLE

Вступ

Розвиток інтелектуальних технологій обробки даних призвів до стрімкого поширення підроблених ствтичних та динамічних (відеозаписів) зображень, що часто використовуються з метою маніпуляції соціумом. При цьому відбувається розповсюдження дезінформації, що призводить до значного викривлення інформаційного простору. Такі викривлення можуть мати зазвичай негативні наслідки [1, 2].

Діпфейк (англ. Deepfake) – методика створення зображення чи відео, що основана на технологіях штучного інтелекту. Назва походить від об'єднання термінів deep (англ. deep – глибинний, тобто пов'язаний з використанням технологію глибинного навчання) та fake (англ. fake – підробка) [3].

С кожним роком процес розвитку глибинного навчання стає дедалі ефективнішим завдяки збільшенню потужності техніки та інформаційним збагаченням. Технології Deepfake на сьогоднішній день становлять загрозу для інформаційного суспільства, створюючи умови для генерування провокаційних медіа матеріалів. Питання того, як розпізнавати та боротись з з негативним застосуванням технологій Deepfake охоплює медійну, правову та технологічну складові, що різняться своєю гнучкістю, пошуком варіантів вирішення, а також способами впровадження та подальшої імплементації.

Методи генерації фальсифікованих за допомогою технологій Deepfake зображень та методи їх виявлення спираються на дуже схожі принципи. Більшість таких методів використовують різні варіанти так званих генеративно-змагальних штучних нейронних мереж, що складаються з двох основних елементів – генератора та дискримінатора [4].

Обидва елементи спільно навчаються на основі протилежних цілей. Мета генератора – створити фейкове зображення, яке важко відрізнити від реального, а мета дискримінатора – виявити синтетично створене зображення.

Таким чином, актуальною є проблема відокремлення фальсифікованих статичних або динамічних зображень (зокрема, зображень облич людини), створених за допомогою технологій Deepfake, від оригінальних.

У даній роботі пропонується підхід до виявлення фальсифікованих зображень, який оснований на спільному використанні ансамблю генеративно-змагальних моделей та механізму Attention. Такий підхід може привести до покращення точності різних нейромережевих класифікаторів для виявлення зображень, створених за технологіями Deepfake.

1. Технології генерації фальсифікованих зображень

Генерація цифрових зображень, як статичних, так і динамічних, є однією з найбільш витратних за кількістю операцій задачею. Для генерації зображень (а також для їх фальсифікації) зазвичай використовують спеціальні архітектури нейронних мереж для роботи із зображеннями, без якої будь-яка їх машинна підробка була б неможливою. До таких архітектур належать, насамперед, згорткові нейронні мережі (ЗНМ, або CNN – convolutional neural network), націлені на ефективне за потужностями та точністю розпізнавання зображень, що входять до складу технологій глибокого навчання [5, 6]. Основна особливість згорткових нейронних мереж полягає в чергуванні двох основних шарів (convolution layers і subsampling layers). Структура цієї мережі є односпрямованою та багатошаровою. Для її навчання використовуються стандартні методи, наприклад, метод зворотного поширення помилки [7]. На сьогоднішній день найрозповсюдженішим методом розпізнавання зображень та їх фрагментів (зокрема, обличчя людини) за допомогою ЗНМ є метод з використанням алгоритму, що передбачає побудову біометричних шаблонів (дескрипторів) по зображенню і пошуку заданого шаблону в базі вже обчислених дескрипторів. Класичні архітектури нейронних мереж з повнозв'язними шарами недоцільно застосовувати на практиці для аналізу зображень, оскільки кількість параметрів таких мереж експоненціально зростає з розміром вхідних даних і кількістю шарів.

Для синтезу фейкових зображень за допомогою ЗНМ останнім часом активно використовуються згорткові генеративно-змагальні мережі DCGAN (deep convolutional generative adversarial networks) [8].

Основна ідея побудови генеративно-змагальних мереж полягає у навчанні пари мереж в процесі їх постійного змагання. У випадку генерації фальсифікованих зображень з використанням DCGAN мережа-генератор (G) намагається створити фальсифіковане зображення, яке важко буде відрізнити від справжнього. Мережа-дискримінатор (D), у свою чергу, намагається розрізнити підробки і справжні зображення. Важливо, що генератор ніколи не отримує реальні дані, на вхід подається тільки випадковий вектор (джерело ентропії, що іноді інтерпретується як простір прихованих змінних, latent-space). Єдиний спосіб навчання для нього – взаємодія з дискримінатором. Дискримінатор при цьому отримує на вході або створені генератором дані, або об'єкт реальної навчальної вибірки. Помилка навчання дискримінатора розраховується з урахуванням того, звідки прийшли дані. В процесі навчання (рис.1) генератор навчається розподілу початкової вибірки і починає створювати дані усе більш близькі до реальних в той час, як дискримінатор стає більш точним в розпізнаванні підробки від оригіналу.

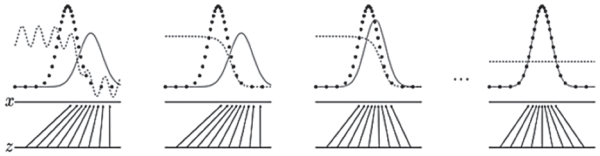

Рис. 1. Ілюстрація до процесу навчання генеративно-змагальної мережі

GAN мережі навчаються одночасно, оновлюючи розподіл дискримінатора (D, переривчаста лінія), який показує ймовірність належності розподілу генератора (G, безперервна лінія) до розподілу реальних даних (чорна точкова лінія). Нижня горизонтальна лінія – це область визначення, з якої вибирається z (в цьому випадку рівноймовірно). Верхня горизонтальна лінія вища – область визначення для x. В процесі навчання відбувається уточнення залежності x=G(z). На останній ітерації (d) дискримінатор вже практично не може відрізнити реальні дані від згенерованих.

Мережі, які лежать в основі генератора і дискримінатора, є багатошаровими мережами, що складаються із згортальних і повнозв'язних шарів. D і G можуть відрізнятися і не обов'язково, щоб вони були повністю дзеркальними. Оскільки завдання генератора полягає в зіставленні простору прихованих змінних в простір даних, то можна це записати у вигляді:

$$G: G(z) \to R^{|x|}, \tag{1}$$

де $z \in R^{|x|}$ — вибірка з простору прихованих змінних; $x \in R^{|x|}$ — справжні дані.

Для дискримінатора загальна функція зіставлення даних у ймовірність їх належності до істинних може бути представлена таким чином:

$$D: D(x) \to (0,1). \tag{2}$$

Крім мереж DCGAN, для генерації фальсифікованих зображень за технологіями Deepfake широко використовуються згорткові мережі типу автоенкодер (Autoencoder – AE). Автоенкодер – це тип багатошарової згорткової нейронної мережі, яка має здійснювати таку автоасоціативну апроксимацію функції, щоб її вихідний сигнал якомога точніше відповідав значенню вхідного сигналу. Таке тотожне співвідношення виявляється не тривіальним, якщо накласти деякі специфічні обмеження на нейронну мережу автоенкодера. Такими обмеженнями можуть бути, перш за все, обмеження кількості нейронів на прихованому шарі і накладення критеріїв розрідженості на активацію цих нейронів. Слід зазначити, що в загальному випадку AE характеризується наступними особливостями: має симетричну структуру; містить непарну кількість шарів; складається з кодера і декодера; містить так зване «пляшкове горлечко», тобто вихідний шар кодера (вхідний шар декодера), в якому використовуються лінійні функції активації; кількість нейронів в «пляшковому горлечку» має бути менше, ніж розмірність вхідних даних, або ж повинна відбуватися лише часткова активація цих нейронів; у вхідному і вихідному шарах AE зазвичай використовуються гаусові функції активації.

Симетричність архітектури значно полегшує завдання настройки параметрів AE, оскільки необхідно визначати лише половину всіх параметрів мережі.

Розглянемо класичний AE з одним прихованим шаром. Позначимо вхідні сигнали мережі як $x(1),x(2),...,x(n)$.

Спочатку AE перетворює (кодує) вхідний сигнал $x \in [0,1]^d$ у деяке внутрішнє представлення $y \in [0,1]^d$ з використанням перетворення настуного вигляду:

$$y = s(Wx + b), \quad (3)$$

де W та b – загальні матриці ваг і зміщень мережі; s – нелінійна функція перетворення (наприклад, гіперболічний тангенс).

Потім внутрішнє представлення (код) вхідного сигналу y перетворюється (декодується) в сигнал z, який є реконструкцією вхідного сигналу і має таку ж розмірність. Дане перетворення можна записати в такий спосіб:

$$z = s(W'y + b'), \quad (4)$$

де W' и b' – настроювані матриці ваг і зміщень мережі декодуючого перетворення.

Параметри моделі (W, W', b, b') налаштовуються таким чином, щоб мінімізувати помилку реконструкції вхідного сигналу, що може бути здійснено з використанням різних функцій втрат, наприклад, квадратичної функції або крос-ентропії.

Вдосконаленим варіантом AE, що може бути використаний для створення фальсифікованих зображень за технологіями Deepfake, є так званий шумопригнічуючий автоенкодер (ШАЕ – DAE – Denoising Autoencoder). ШАЕ є модифікацією AE, в якій входи поєднуються з деякими завадами і система навчається для відновлення даних без перешкод [9]. Таким чином, ШАЕ можна розглядати як стохастичне розширення класичного автокодера, яке змушує модель навчатися відновленню вхідного сигналу при подачі на вхід його зашумленої версії. Схема навчання ШАЕ може використовуватися в технологіях Deepfake для шумопригнічення і є ефективним способом навчання автоенкодера більш значущим прихованим уявленням даних.

Базовий варіант зазальної архітектури ШАЕ представлений на рис. 2.

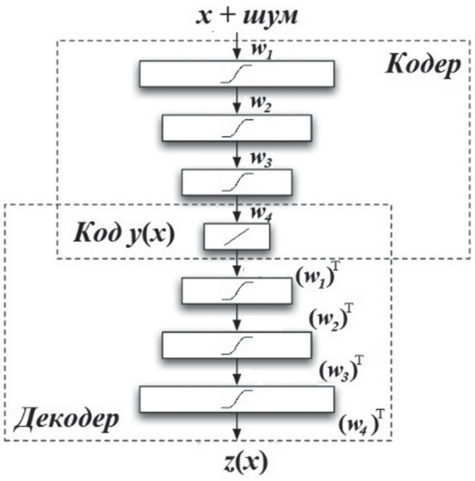

Рис. 2. Загальна структура ШАЕ

Технологія DeepFake використовує зазвичай два AE (або ШАЕ) з однаковими кодерами, але різними декодерами (A та B), що дозволяє мати однаково побудований базовий вектор для різних зображень, що і робить можливою деформацію оброблюваних зображень (зокрема, зображень людських облич) надалі. Упродовж фази навчання, дві мережі відпрацьовують окремо. Декодер A навчається тільки на обличчях першого AE, а декодер B – на обличчях другого AE. В той же час усі базові вектори отримуються з використанням однієї і тієї ж архітектури кодерів, тобто кодер визначає загальні ознаки в обох зображеннях.

Після закінчення процесу навчання базовий вектор суб'єкта A передається в декодер суб'єкта B, який намагається відтворити суб'єкт B за інформацією, отриманою від суб'єкта A. Якщо мережа добре узагальнює особливості зображення, та базовий вектор буде враховувати вирази обличчя та їх розташування.

Выдзначимо, що для отримання задовільних результатів необхідно мати від 200 до 2000 зображень. Навчання таких моделей є дуже витратним з погляду на обчислювальну складність (наприклад, щоб згенерувати однохвилинне відео і, треба близько 18 хвилин на одній відеокарті). Архітектуру DeepFake можна оптимізувати багатьма способами, проте найбільш

поширеними з них є комбінування моделей з GAN архітектурою. Генератори поступово вчаться створювати найбільш реалістичне зображення суб'єкта, а дискримінатори — визначати, яка з них згенерована, а яка оригінальна. За рахунок навчання GAN генераторів синтезоване зображення стає все більш реалістичним. Таким чином, Encoder і Decoder відповідають за перенесення зображення, а дискримінатор генеративних мереж – за покращення результату [8]. Цей підхід продемонстрованио на рис. 3.

Ще один підхід передбачає використання архітектури GAN, що складається з декількох генеративно-змагальних мереж, кожна з яких відповідає за свою операцію. Для навчання такої нейромережі потрібен потужний кластер відеокарт. Незважаючи на це, такий підхід є найбільш перспективним, тому що дає кращий результат. Залучивши до цього підходу рекурентні мережі, можна ще більше покращити якість DeepFake. Рекурентні мережі дозволяють отримувати базовий вектор, ґрунтуючись не лише на одному зображенні, а на декількох зображеннях. Базовий вектор генерується на поточному зображенні, а також зображеннях з відео створених за декілька секунд до і після початкового зображення. При цьому переходи з одного стану і положення обличчя до іншого стають набагато реалістичніше.

Одним з останніх варіантів використання згорткових генеративно-змагальних мереж для генерації підроблених статичних та динамічних (відео) зображень є модель FSGAN (Face Swapping GAN). Головна особливість цієї моделі полягає в тому, що вона може застосовуватися до конкретних зображень без необхідності попереднього навчання на них [9]. Наприклад, у моделі FSGAN, що застосовується для зображень людських облич, одна нейромережа вчиться підганяти зображення обличчя під параметри цільового відео (поворот голови, нахил убік або вперед), друга переносить риси обличчя, а третя здійснює операцію image blending (злиття зображень), щоб зображення було більш реалістичним (без ушкоджень або артефактів). Особливістю такої технології DeepFake є наявність рекурентної нейронної мережі для реконструкції зображень осіб, яка коригує як їх позу, так і різні вирази обличчя.

Для відеозображень використовується підхід безперервної інтерполяції представлень обличчя на основі реконструкції, тріангуляції Делоне і баричентричних координат. Закриті ділянки обличчя обробляються окремою мережею, а для об'єднання двох зображень облич застосовується мережа, яка зберігає цільовий колір шкіри і особливості освітлення. На рис. 4. продемонстрований принцип використання моделі FSGAN.

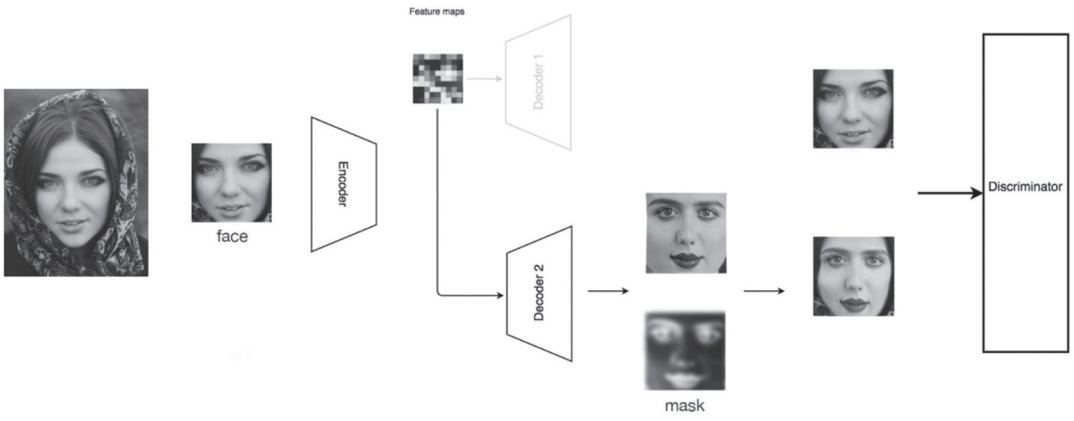

Рис. 3. Застосування GAN для покращення результатів DeepFake

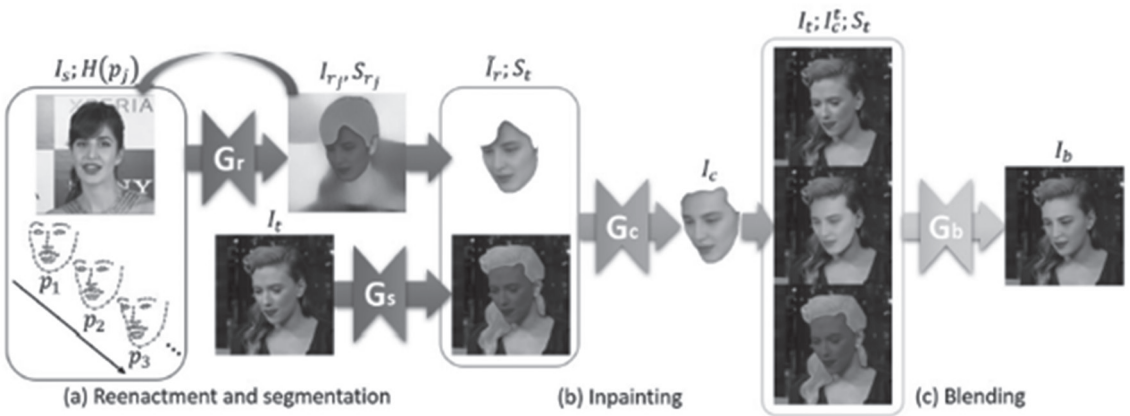

Рис.4. Принцип використання моделі FSGAN

На рис. 4 Is є початковим зображенням обличчя Fs, а It – підсумковим зображенням обличчя Ft. Алгоритм створює нове зображення, використовуючи It, та намагаючись замінити Ft на Fs, зберігаючи його положення і вираз. FSGAN складається з кількох компонентів. По перше, це генератор інсценування Gr і згортальна мережа для сегментації ділянок обличчя Gs. Генератор Gr визначає опорні точки обличчя Ft та формує таке зображення Ir, щоб Fr відображало Fs з таким же положенням і виразом, як у Ft. Він також обчислює маску сегментації Sr обличчя Fr. Мережа Gs сегментує особу і волосся обличчя Ft. На виділеному зображенні Ir можуть бути відсутніми деякі частини обличчя. Далі застосовується мережа Gc, що використовує сегментацію St та виявляє пропущені пікселі. Остання частина моделі об'єднує зображення обличчя Fc з початковим зображенням It для отримання кінцевого результату.

2. Методи виявлення фальсифікації зображень

Для виявлення підроблених зображень найчастіше використовується згорткова рекурентна мережа LSTM (Long Short Term Memory). Ця модель складається з двох основних блоків: згортальної мережі для витягання елементів зображення і блоку аналізу частини послідовності зображень [3]. Архітектура моделі представлена на рис. 5.

Отримуючи послідовність тестових зображень, модель витягає набір функцій для кожного кадру за допомогою згортальної мережі. Після цього функції декількох послідовних кадрів об'єднуються і передаються моделі LSTM для аналізу. Рекурентна мережа видає ймовірність правливості послідовності.

Як вхідні дані використовується послідовність векторів ознак, витягнутих згортальною мережею з поданих на вхід кадрів, і згортальна мережа з двома вузлами для визначення того, чи є зображення згенерованим або оригінальним відео. LSTM мережа є проміжним блоком системи і не вимагає додаткового навчання.

Для демонстрації точності моделі в роботі [3] було розглянуто 300 відеозображень, згенерованих за допомогою DeepFake. Автори моделі перевіряли точність, використовуючи 20, 40 і 80 послідовних кадрів. Результати продемонстровано в табл. 1.

Таблиця 1

Оцінка точності генерації DeepFake-зображень з використанням рекурентних моделей

Модель	Training acc. (%)	Validationacc. (%)	Test acc. (%)
Conv-LSTM, 20 fraimes	99,5	96,9	96,7
Conv-LSTM, 40 fraimes	99,3	97,1	97,1
Conv-LSTM, 80 fraimes	99,7	97,2	97,1

Модель демонструє непогану точність, але вона може працювати тільки з послідовностями підроблених відеозаписів.

При використання цього підходу, було зроблено спостереження, що хоча GAN-подібні моделі здатні генерувати фотореалістичні візуальні і геометричні сигнали, які виходять за рамки можливостей виявити їх людським оком, біологічні сигнали, приховані природою, як і раніше складно відтворити. Біологічні сигнали є невід'ємною частиною відео із зображенням осіб, яку у свою чергу є сферою застосування Deepfake технології.

Ще одна технологія виявлення Deepfake зображень використовує компактну нейромережу MesoNet. [10]. Ця технологія дозволяє ефективно розпізнавати підробні відеозаписи, згенеровані за допомогою алгоритмів DeepFake або Face2Face. Дослідження показали, що створити одну узагальнену модель для ефективного виявлення підробок, створених цими алгоритмами одночасно, за допомогою цієї архітектури неможливо. Тому доцільним є застосування для виявлення деяких типів підробок зображень різних варіантів нейромережі MesoNet з невеликою кількістю шарів, зокрема мереж Meso-4 та MesoInception4.

Meso-4 використовує чотири блоки згортки та пулинг з одним прихованим шаром. Згорткові шари використовують активаційну функцію ReLU, щоб запобігти ефекту розмитого градієнту. Для регулярізації і підвищення надійності в Meso-4 використовується Dropout шар.

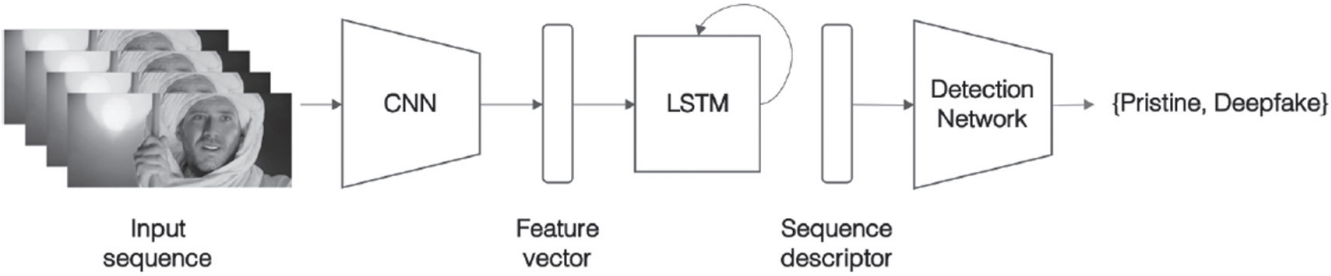

Рис. 5. Архітектура рекурентної моделі LSTM

Альтернативна модель MesoInception4 відрізняється від Meso-4 тим, що замість перших двох згорткових блоків використовується Inception модуль. Ідея цього модуля полягає в об'єднанні кількох згорткових шарів різної розмірності для збільшення розмірності простору в якому модель оптимізується. Основною перевагою цієї моделі є її компактність. У моделі відносно мало параметрів, завдяки чому навчання відбувається швидко.

Розглянемо можливість виявлення зображень, згенерованих за допомогою технологій Deepfake, за методом парного навчання, що використовує модифіковану структуру згорткової мережі, яка зветься мережею із загальними підробними функціями (CFFN – Cascade Feed Forward Networks). Ця структура використовує детектор глибокої підробки (DeepFD) для ідентифікації підробного обличчя і початкових зображень, а також двоетапний метод навчання, який поєднує в собі функцію CFF на основі стратегії парного навчання і навчання класифікатора (рис. 6).

Представлення стратегії контрольованого навчання у виявленні підробних зображень обличчя пов'язані з проблемами пошуку навчальної вибірки, яка б містила зображення, згенеровані різними варіантами GAN. Щоб здолати ці проблеми, підробні і реальні зображення об'єднуються в пару і слідують одне за одним, використовуючи парну інформацію для побудови порівняльної функції втрат, щоб упізнати відмітну загальну фальсифіковану особливість (CFF) за пропонованим CFFN. Як тільки фунція розпізнавання CFF навчена, класифікатор використовує дискримінаційну функцію CFF, щоб ідентифікувати, чи є зображення реальним або підробним.

Чимало варіантів мереж CNN (наприклад, DenseNet, ResNet і Xception) можуть бути використані для виявлення ознак фальшування з тренувального набору та для створення класифікатора. Проте, більшість з цих навчаються контрольованим чином, тому ефективність класифікації залежить від системи навчання.

Відомо, що CNN ефективно використовуються тільки в наданні ознак високого рівня, щоб ідентифікувати, чи є зображення підробним. Проте CFF підробних зображень можуть існувати не лише в представленні високого рівня, але також і в представленні об'єктів середнього рівня. Тому міжрівневі елементи можуть бути інтегровані в класифікаційний шар для підвищення продуктивності розпізнавання підробних зображень. Для навчання відповідної моделі доцільно використовувати дві функції втрат для оптимізації мережі CFFN і класифікатора одночасно. Проте зазвичай важко визначити вагові значення для двох функцій. Оскільки основною метою пропонованого CFFN є навчання дискримінаційних ознак, це дозволяє навчати CFF шляхом мінімізації контрастнихї втрат в першу чергу. Після цього будь-який класифікатор можна використовувати для розпізнавання підробного зображення обличчя на основі навченої CFF. Отже, каскадна мережа CFFN спочатку навчається на основі контрастних втрат, а потім оптимізується мережа класифікаторів шляхом мінімізації втрат від перехресної ентропії.

Основним недоліком контрольованого навчання за розглянутим підходом є те, що важко визначити фрагмент, який виключений з етапу навчання. Щоб підвищити продуктивність пропонованого методу, вводяться контрастні втрати при вивченні CFF шляхом парного навчання. Варіант відповідної архітектури CFFN наведено на рис. 7.

Основною проблемою цього методу є те, що модель CFFN не здатна класифікувати всі можливі підробки, які згенеровані за допомогою базових алгоритмів DeepFake. Зображення, створенні за допомогою деяких сучасних модифікацій згорткових

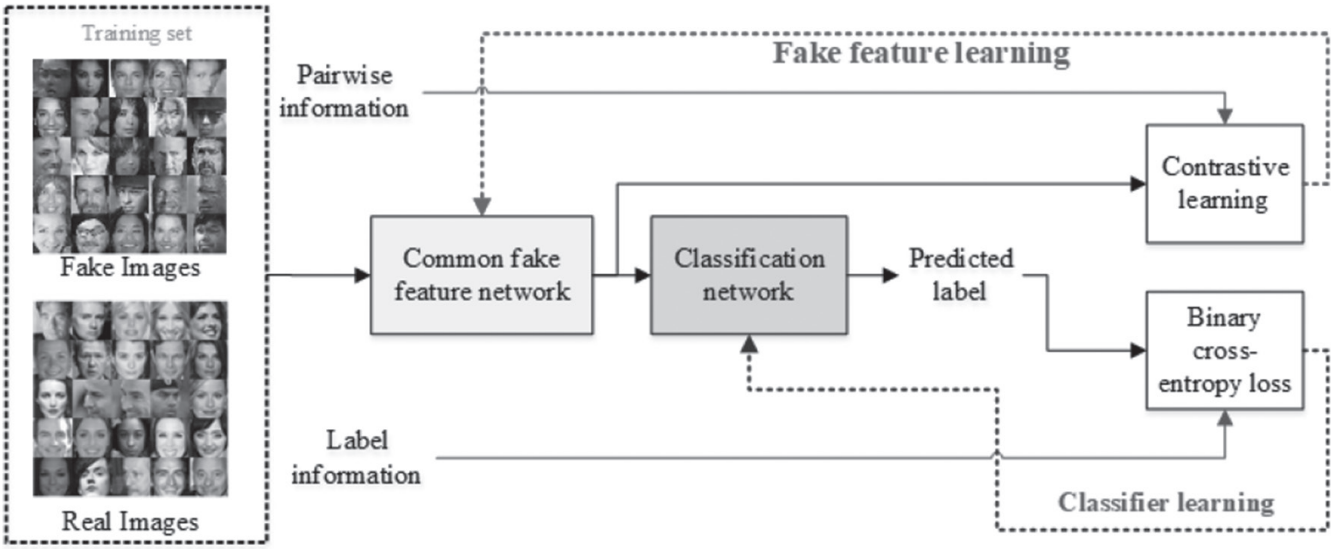

Рис. 6. Метод на основі CFF і стратегії парного навчання

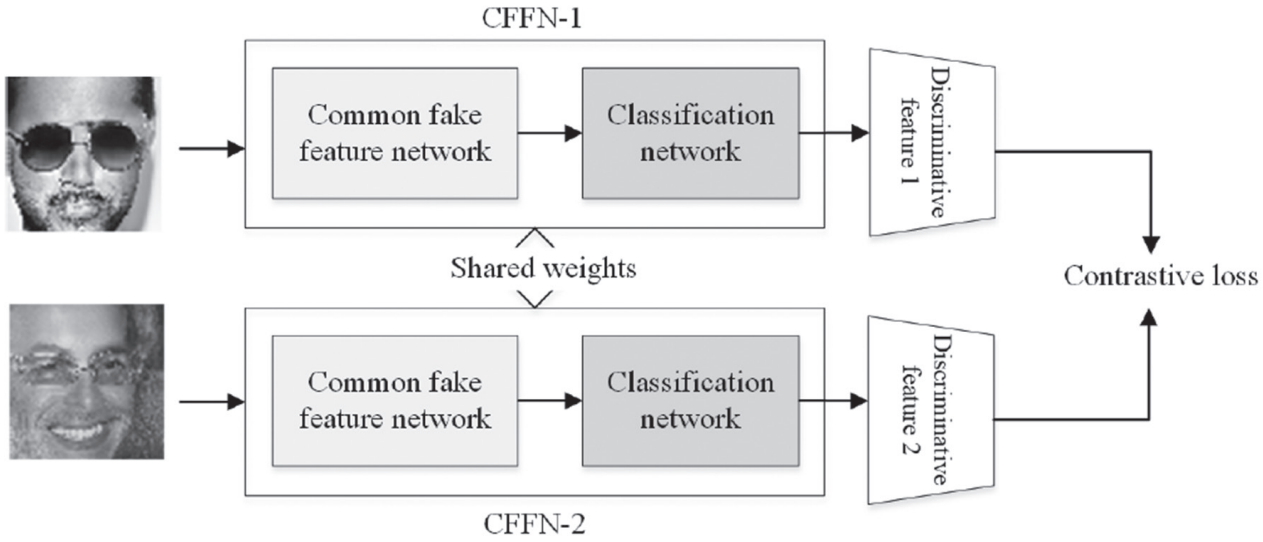

Рис. 7. Архітектура CFFN

нейронних мереж, таких як FSGAN, в основному класифікуються як початкові.

Існуючі методи класифікації фальсифікованих зображень можуть показувати непогані результати на невеликих наборах даних. Проте внаслідок того, що ці моделі мають невеликі розміри нейромереж, то на великих наборах даних результати суттєво погіршуються. Тому доцільним є створення узагальненого класифікатору на основі ансамблю з описаних згортальних моделей, що використовуються для виявлення Deepfake зображень.

3. Пропонований метод виявлення фальсифікованих зображень

Пропонований метод ґрунтується на концепції виявлення підроблених зображень за допомогою ансамблю згорткових нейронних мереж.

Відомо, що ансамбль моделей може привести до значного покращення точності різних класифікаторів на основі згорткових мереж для виявлення різних високорівневих семантичних особливостей, які доповнюють одна одну, позитивно впливаючи на точність ансамблю для цієї конкретної проблеми.

Для цього в якості відправної точки було обрано сімейство моделей EfficientNet, які перспективними у багатьох задачах обробки зображень [11]. Цей набір архітектури досягає більшої точності та ефективності в порівнянні з іншими сучасними CNN. Приклад цієї архітектури продемонстрований на рис. 8.

Враховуючи особливості архітектури EfficientNet, пропонується здійснити дві модифікації, щоб зробити модель вигідною для ансамблювання. З одного боку, пропонується додати до моделі Attention механізм, який реалізує аналітичний метод, що дозволяє визначити, яка частина досліджуваного відео є більш інформативною для процесу класифікації. З іншого боку, до мережі пропонується додати сіамські стратегії навчання для екстраполяції додаткової інформації про дані.

Наведемо опис архітектури EfficientNet з пропонованим Attention механізмом і стратегіями навчання мережі. Серед сімейства моделей EfficientNet

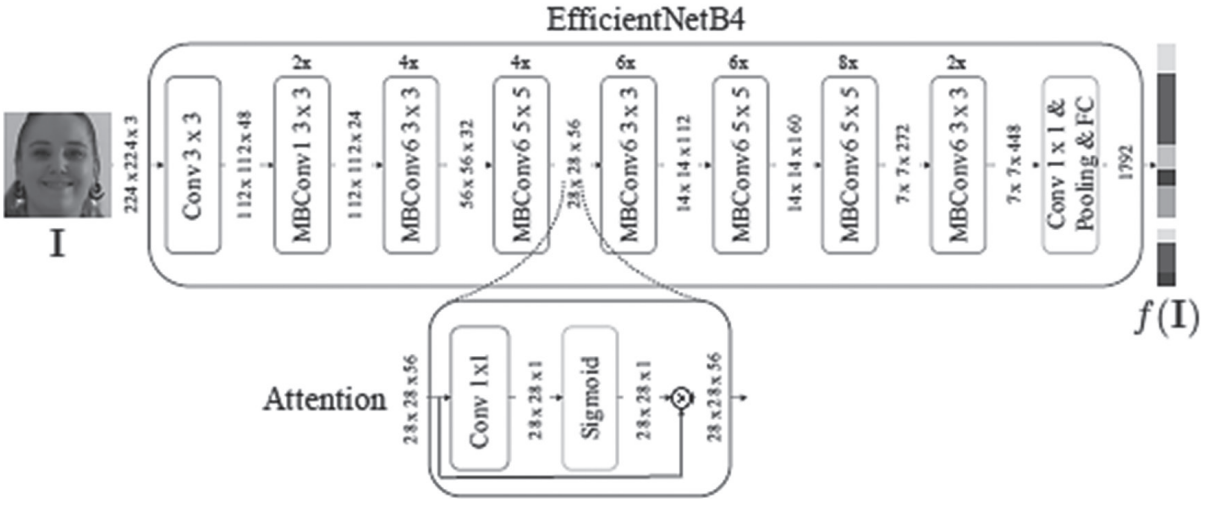

Рис. 8. Архітектура EfficientNetB4

була вибрана EfficientNetB4 в якості базового рівня для цих моделей. Цей вибір мотивований можливістю компромісу, пропонованого цією архітектурою, з точки зору вимірів (тобто кількості параметрів), часу виконання (тобто вартості FLOPS) і класифікацією продуктивності. З 19 мільйонами параметрів і 4,2 мільярдами FLOPS, EfficientNetB4 досягає 83,8% точності, що є кращим результатом, на наборі даних ImageNet. У тому ж наборі даних XceptionNet, використовуваний в якості базового методу виявлення маніпуляцій з обличчям, досягає 79% точність для 23 мільйонів параметрів і 8,4 мільярдів FLOPS.

Входом в мережу EfficientNetB4 є квадратне кольорове зображення, тобто в експериментах використовується обличчя, виділене з відеокадру. Насправді, зазвичай рекомендується відстежувати інформацію про обличчя, а не використовувати повний кадр в якості вхідних даних для мережі для підвищення точності класифікації. Крім того, обличчя можуть бути легко використані з зображень за допомогою будь-якого з існуючих детекторів зображень облич. Вихід мережі є вектором ознак з заданого числа елементів, визначуваним як $f(I)$. Остаточний результат, пов'язаний з цією функцією, характеризує якість класифікації.

Алгоритми архітектур Transformer або Residual Attention Networks дозволяють нейронній мережі визначити, яка частина її вхідних даних, частина зображення або послідовності слів є актуальною для виконання поставленого завдання. У контексті виявлення зображень, згенерованих за допомогою технології DeepFake, дуже корисно знати, яка саме частина входу дає нейронній мережі найбільшу частину інформації для ухвалення рішення. Механізм Attention реалізується шаром нейромережі, який має дві основні особливості:

– на вхід подається карта ознак, витягнута за допомогою EfficientNetB4 до певного шару. Вона обирається так, щоб ці ознаки мали достатню інформацію про вхідний кадр, не будучи занадто детальними або, навпаки, занадто невизначеними;

– ознаки обробляються тільки одним згортальним шаром розмірністю 1, після чого передаються в активаційну сигмоїдальну функцію для отримання єдиної Attention карти.

З одного боку, цей простий механізм дозволяє мережі зосередитися тільки на найбільш вагомих частинах ознак об'єктів, з іншого боку, він дає глибше розуміння того, які частини вхідного слою мережа вважає найбільш інформативними. Дійсно, отримана Attention карта може бути легко зіставлена з вхідним зразком, підкреслюючи, яким елементам було приділено більше уваги нейронною мережею. Результат Attention блоку остаточно обробляється блоками EfficientNetB4, що залишилися. Отримана модель була названа EfficientNetB4Att.

Навчання моделей відбувається за стратегіями наскрізного та сіамського навчання. Перша стратегія є більш класичною і використовує метрику оцінки на наборі даних DFDC (Deepfake Detection Challenge). Інша стратегія спрямована на використання можливостей узагальнень, пропонованих нейронними мережами, для отримання дескриптора функції, який дає перевагу подібності між вибірками, що належать до одного й того ж класа. Кінцева мета полягає в тому, щоб навчити представлення в просторі декодованих шарів мережі, які добре розділяють вибірку зображень (здебільшого зображень обличчя) реального і фальсифікованого класів.

Під час наскрізного навчання мережа отримує зразки зображення обличчя і формує оцінку $\widehat{y_i}$., яка ще не пройшла через сигмоїдальну функцію активації. Оновлення ваг здійснюється за допомогою функції LogLoss:

$$L_L = -\frac{1}{N}\sum_{i=1}^{N}\Big[\, y_i \log(S(\widehat{y_i}) + (1-y_i)\log(1-S(\widehat{y_i}))\Big], \quad (5)$$

де $\widehat{y_i}$ – змінна, що представляє i-у оцінку обличчя; N — загальна кількість зображень облич, використовуваних для навчання; $S\left(\widehat{y_i}\right)$ — активаційна сигмоїдальна функція.

Візначимо, що $\widehat{y_i} \in \{0,1\}$ та відповідає міткам обличчя. Зокрема, мітка 0 пов'язана з зображеннями облич, що виходять із справжніх початкових відеозаписів, а мітка 1 пов'язана з підробними зображеннями.

Процедура сіамського навчання (triplet margin loss) була адаптована насупним чином:

$$L_T = \max(0, \mu + \delta_+ - \delta_-), \quad (6)$$
$$\delta_+ = \left|\left| f(Ia) - f(Ip) \right|\right|_2,$$
$$\delta_- = \left|\left| f(Ia) - f(In) \right|\right|_2,$$

де δ_+ та μ — позитивні значення; *Ia* — реальне зображення; *Ip* — позитивний зразок, що належить до того ж класу, що й *Ia* (інше реальне зображення); *In* – зображення, згенероване за допомогою DeepFake.

Навчання завершується настроюванням шару класифікації поверх мережі, наслідуючи описаний вище наскрізний підхід.

4. Експериментальні дослідження

Експерименти проводилися на двох наборах даних: FF++ та DFDC [12]. FF ++ — це великий набір даних по маніпуляціях з обличчям, створений з використанням state-of-the- art методів редагування відеозаписів. Цей набір даних містить два класичних підходи маніпуляції обличчями,а саме Face2Face і FaceSwap, разом з двома стратегіями, ґрунтованими на навчанні (DeepFake і NeuralTextures). Кожен метод застосовувався до 1000 високоякісних відеозаписів, завантажених з YouTube, щоб показувати зображення без перешкод і зайвих об'єктів. Усі послідовності

містили не менше 280 кадрів. Для імітації реалістичних налаштувань відеозаписи було стиснено з використанням кодека H.264. Відеозаписи високої та низької якості генерувалися з використанням параметра квантування з постійною швидкістю, рівною 23 і 40 відповідно.

DFDC – відкритий набір фальсифікованих відеозаписів, зібраний спільно компаніями Microsoft, Facebook і Amazon. Він складається із понад 119000 відеофрагментів, створених спеціально для задачі класифікації відеозаписів, згенерованих за допомогою DeepFake, що представляють як реальні, так і підробні відеоролики. Справжні відеоролики — це послідовності невеликих відео з акторами (або іншими персонажами), обраними по декількох чинниках (стать, шкіра, вік і тому подібне).

Підробні відеоролики створюються, паралельно із справжніми, застосовуючи різні методи DeepFake, наприклад, різні алгоритми зміни особи. Точні алгоритми, використовувані для створення підроблених відео не відомі.

Для створення ансамблю були обрані наступні моделі архітектури: XceptionNet (найбільш ефективна модель для виявлення маніпуляцій на зображеннях з обличчями); EfficentNetB4 (архітектура state-of-the- art у багатьох завданнях, пов'язаних з обробкою зображень); EfficentNetB4Att (архітектура, що здатна виділяти найбільш релевантні області зображень).

Кожна модель навчалася і тестувалася незалежно від інших. В наборі FF++ розглядалися тільки відеозаписи, згенеровані при квантуванні з постійною швидкістю. Модель XceptionNet навчалася з використанням одного і того ж підходу для навчання, в той час як дві моделі EfficientNet навчалися як наскрізним, так і сіамським способами.

В результаті були досліджені чотири навчених моделі: EfficientNetB4 і EfficientNetB4Att, які навчалися за допомогою класичного наскрізного підходу, а також EfficientNetB4ST і EfficientNetB4AttST, навчені з використанням сіамської стратегії. Усі ці моделі є похідними від EfficientNetB4 та можуть сприяти остаточному об'єднанню.

Для кожного набору даних були застосовані різні способи розділення даних. Набір DFDC був розбитий по структурі папок (перші 35 папок для навчання, папки від 36 до 40 для перевірки і останні 10 папок для тестування). Для набору FF ++ використовувалося таке розділення: 720 відеозаписів для навчання, 140 для перевірки і 140 для перевірки з пулу оригінальних послідовностей, отриманих з YouTube. Відповідні фальсифіковані відеозаписи присвоюються одному і тому ж розділенню.

В експериментах була розглянута обмежена кількість кадрів для кожного відеозапису. На етапі навчання цей вибір мотивується двома основними міркуваннями: при використанні дійсно невеликої кількості кадрів на відеозапису спостерігається сильна тенденція до перенавчання; збільшення кількості кадрів не покращує продуктивність виправданим чином. Це випливає з рис. 9, де представлені втрати при навчанні і перевірці залежно від ітерацій навчання, а також при виборі змінної кількості кадрів відеозапису.

Слід відзначити, що мінімальні втрати при перевірці не покращують вибір 15 кадрів на відео замість 32, проте вибір 32 кадрів на відео допомагає запобігти перенавчанню. Для тестування також треба брати до

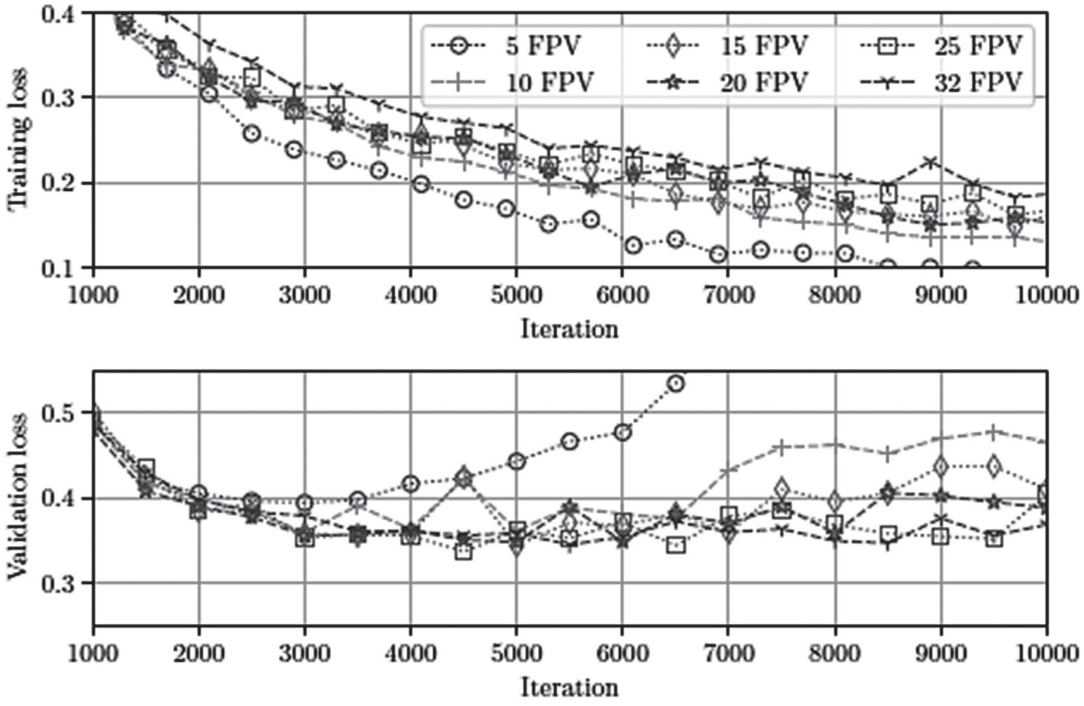

Рис. 9. Графік взаємозв'язку кількості кадрів з помилкою

уваги апаратні і часові обмеження, що накладаються завданням набору DFDC. Згідно з цим було обмежено кількість аналізованих кадрів з кожної послідовності до 32 для етапів навчання і тестування.

Обсяг даних можна скоротити, адже не вся інформація про кадри корисна для процесу глибокого виявлення. Дійсно, можна в основному сфокусувати аналіз на фрагменті, де знаходиться зображення обличчя суб'єкта. Отже, на кроці попередньої обробки витягається обличчя з кожного кадру сцени за допомогою екстрактора BlazeFace, який в експериментах виявився більш швидший, ніж детектор MTCNN. У разі виявлення більше однієї особи зберігається обличчя з найкращою оцінкою достовірності. В результаті вхідними даними для мереж вибрано квадратне кольорове зображення розміром 224×224 пікселі.

Під час навчання і перевірки, щоб зробити моделі надійнішими, виконувалися операції доповнення даних на вхідних гранях. Зокрема, випадковим чином застосовувалися масштабування, горизонтальне відображення, випадковий контраст яскравості, насиченість відтінку, додавання шуму і стискання JPEG. Зокрема, використовувалася бібліотека Albumentation в якості бібліотеки доповнення даних, тоді як Pytorch використовувався як фреймворк Deep Learning. Для навчання використовувався оптимізатор Адама з гіперпараметрами, що становлять $\beta_1 = 0,9; \beta_2 = 0,999; \epsilon = 10^{-8}$, а початкова швидкість навчання становила 10^5.

Для наскрізного навчання задавався максимум ітерацій (20000) обробки пакету з 32 прикладів (16 реальних, 16 підроблених), вибраних випадковим чином і рівномірно з усіх відеозображень навчальної вибірки, або відбувалося переривання навчання при досягненні плато на функції втрат. Валідація моделі в цьому контексті виконувалася кожні 500 ітерацій навчання на 6000 прикладах, також вибраних випадковим чином і рівномірно з усіх відеозображень навчального набору даних. Початкова швидкість навчання зменшується з коефіцієнтом 0,1, якщо втрати перевірки не зменшуються після 10 процедур перевірки (5000 ітерацій), і навчання припиняється, коли досягнута мінімальна швидкість навчання.

Для сіамського навчання процедура витягання особливостей здійснювалася з тією ж кількістю ітерацій, валідацією і швидкістю навчання, як і при наскрізному навчанні. Основна відмінність цих двох підходів полягала в застосуванні різних функцій втрати і у розмірі пакету, який для сіамського навчання складав дванадцять триплетів прикладів (шість початкових, шість підроблених і шість з наявністю як початкових так і підроблених ознак). Пакети обиралися пропорційно їх навчальній вибірці. Точне налаштування рівня класифікації виконувалося послідовно, як і для процедури наскрізного навчання з параметрами, наведеними вище.

Навчання моделей відбувалося з використанням NVIDIA 2070 RTX Super та пакету Apex, а також із застосуванням техніки Accumulation Gradient. Ці підходи дозволили тренувати глибокі моделі на цій відеокарті, проте уповільнили навчання і показали результати гірше, ніж без них.

Після навчання був проведений аналіз отриманих результатів. При застосуванні EfficientNetB4Att моделі, щоб показати ефективність Attention механізму при витяганні найбільш інформативного вмісту облич, була оцінена Attention карта, розрахована для декількох облич з набору FF++. Як вихідний шар був вибраний сигмоїд у блоці Attention, який є 2d-картою з розміром 28×28. Потім він був збільшений до вхідного розміру обличчя (224×224) і накладений на вхідне обличчя. Результати наведено на рис. 10.

Слід відзначити, що цей простий механізм уваги дозволяє виділити найбільш деталізовану частину обличчя, наприклад очі, рот, ніс і вуха. Навпаки, плоскі області (де градієнти малі) не є інформативними для мережі. Було показано, що артефакти методів глибокої генерації в основному локалізовані навколо рис обличчя. Наприклад, грубо змодельовані очі і зуби з надмірно білими областями, як і раніше, є основними ознаками методів підробки зображень. Щоб визначити, чи є ознаки, що створюються шарами кодування мережі при навчанні сіамським способом, дискримінаційними для завдання, було обраховано проекцію по зменшеному простору, використовуючи відомий алгоритм t-SNE [13].

Аналіз проекції, отриманої за допомогою EfficientNetB4Att, починаючи з 20 відеороликів набору даних FF++, показав, що кадри одного і того ж відеозапису об'єднуються в невеликі субрегіони, причому усі

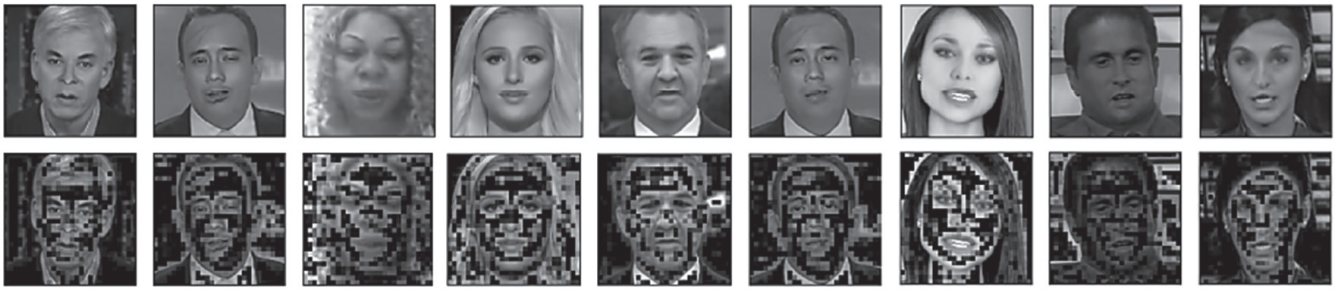

Рис. 10. Ефект застосування Attention механізму

початкові зразки згруповані у верхній області діаграми, тоді як підробні зразки знаходяться в нижній області. Кадри з тих же відеокластерів об'єднуються в менші субрегіони. Це підтверджує доцільність використання механізму Attention на додаток до класичного наскрізного підходу. З отриманих моделей формувалися ансамблі різними способами (з використанням двох, трьох або більшої кількості моделей. Результат обчислювався як середнє значення з усіх моделей, що входять в ансамбль. В табл.2 наведені результати обчислення показників AUC і LogLoss, отримані в експериментах.

Таблиця 2

Показники точності та якості для різних ансамблів моделей

Xception Net	EfficientNet B4	B4ST	B4Att	B4AttST	AUC FF++	AUC DFDC	LogLoss FF++	LogLoss DFDC
✓					*0.9273*	*0.8784*	*0.3844*	*0.4897*
	✓				0.9382	0.8766	0.3777	0.4819
		✓			0.9337	0.8658	0.3439	0.5075
			✓		0.9360	0.8642	0.3873	0.5133
				✓	0.9293	0.8360	0.3597	0.5507
	✓	✓			0.9413	**0.8800**	0.3411	0.4687
	✓		✓		0.9428	**0.8785**	0.3566	0.4731
	✓			✓	0.9421	0.8729	0.3370	0.4739
		✓	✓		0.9423	0.8760	0.3371	0.4770
		✓		✓	0.9393	0.8642	**0.3289**	0.4977
			✓	✓	0.9390	0.8625	0.3515	0.4997
	✓	✓	✓		**0.9441**	**0.8813**	0.3371	**0.4640**
	✓	✓		✓	0.9432	0.8769	**0.3269**	**0.4684**
	✓		✓	✓	**0.9433**	0.8751	0.3399	0.4717
		✓	✓	✓	0.9426	0.8719	0.3304	0.4800
	✓	✓	✓	✓	**0.9444**	0.8782	**0.3294**	**0.4658**

Аналізуючи отримані результати, слід відзначити, що стратегія ансамблювання моделей зазвичай обирається з точки зору характеристик. Кращі результати завжди досягаються комбінацією з двох або більше мереж, тобто об'єднання мереж допомагає збільшити як точність глибокого виявлення (оцінену за допомогою міри AUC), так і якість виявлення (оцінену за допомогою міри LogLoss).

Висновки

Запропонований метод сприяє виявленню маніпуляцій з обличчям у підробних відеопослідовностях, згенерованих за допомогою глибинного навчання та технологій DeepFake. Метод використовує набір моделей EfficientNet, навчених з використанням двох основних концепцій: Attention механізму, який дозволяє нейронній мережі визначити, яка частина її вхідних даних є актуальною для виконання поставленого завдання, та стратегії сіамського навчання по сіамському триплету, яка спрямована на використання можливостей узагальнень, пропонованих нейронними мережами, для отримання дескриптора функції, який дає перевагу подібності між вибірками, що належать до одного й того ж класа. Дослідження показали, що застосування сімейства згорткових моделей EfficientNet разом з Attention механізмом, дає відмінний результат. Об'єднання декількох таких моделей в ансамбль сприяє підвищенню якості виявлення фальсифікацій.

Перспективним продовженням досліджень є розширення ансамблю моделей архітектури EfficientNet (зокрема, за рахунок використання моделей EfficientNetB6 і EfficientNet7). Крім того, доцільно розглянути можливість застосування більш складних функцій втрат для процедур навчання нейромережевого класифікатора.

Список літератури:

[1] *Gonzalez R.C.* Digital Image Processing / R.C. Gonzalez, R.E. Woods // 4th edition Pearson/Prentice Hall, 2018. – 1168p.

[2] *Norvig P.* Artificial Intelligence: A Modern Approach / P. Norvig, S. Russell // Global Edition. – Pearson Education Limited, 2016. – 1152p.

[3] *Güera D.* Deepfake Video Detection Using Recurrent Neural Networks / D. Güera, E. Delp // Proc. of 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand. – 2018, P. 1-6.

[4] *McCann M.* Convolutional neural networks for inverse problems in imaging / M. McCann, K. Jin., M Unser //: A review //IEEE Signal Processing Magazine. – 2017. – V. 34. – №. 6. – P. 85-95.

[5] *Sikorskiy O.* Convolutional neural networks in image classification / O. Sikorskiy // Information Innovative Technologies – 2017. – №1.– P. 397-401.

[6] *Schmidhuber J.* Deep learning in neural networks: An overview / J. Schmidhuber / /Neural networks. – 2015. – V. 61. – P. 85-117.

[7] *Goodfellow Ian J.* Generative Adversarial Networks / Ian J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio // Proc. of 27th International Conference on Neural Information Processing Systems. – 2014. – P. 2672-2680.

[8] *Bezsonov O.* Processing of noisy digital images with use of evolving autoencoders / O. Bezsonov, O. Rudenko, S. Udovenko, O. Dudinova // Eastern-European Journal of Enterprise Technologies. – V. 6/9(90). – 2017. – P. 63-69.

[9] *Nirkin Y.* FSGAN: Subject Agnostic Face Swapping and Reenactment / Y. Nirkin, Y. Keller, T. Hassner //. – arXiv:1908.05932v1 – 2014. – P. 2672-2680.

[10] *Afchar D.* MesoNet:: a Compact Facial Video Forgery Detection Network / D. Afchar, V. Nozick, J. Yamagishi, I. Echizen // eprint arXiv:1809.00888. – 2019. – 8 p.

[11] *Tan M. Le.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. V. Le // – 2019. – Arxiv link: https://arxiv.org/abs/1905.11946

[12] *Dolhansky B. et al.* The deepfake detection challenge datase / 2020. – URL https://arxiv.org/abs/2006.07397

[13] *Schubert E.* Intrinsic t-Stochastic Neighbor Embedding for Visualization and Outlier Detection / E. Schubert, M. Gertz // Proc. of 10th International Conference on Similarity Search and Applications. – 2017. – P. 188–203.

Надійшла до редколегії 28.10.2020