DOI 10.15276/imms.v12.no1-2.53
УДК 004.056.5

Informatics and Mathematical Methods in Simulation
Vol. 12 (2022), No. 1-2, pp. 53-64

ВИЯВЛЕННЯ ФОТОПІДРОБОК, ВИКОНАНИХ ЗАСОБАМИ НЕЙРОННИХ МЕРЕЖ

Є.В. Тимофеєв, В.В. Зоріло

1, пр.Шевченка, Національний університет «Одеська політехніка», 65044, Одеса
vikazorilo@gmail.com, timofeevzhenya2107@gmail.com

Глибоке навчання є ефективною та корисною технікою, яка широко застосовується в різних областях, включаючи комп'ютерний зір, машинний зір та обробку природної мови. Deepfake використовує технологію глибокого навчання, щоб маніпулювати зображеннями та відео людини, які люди не можуть відрізнити від реальних. Deepfake зазвичай створюють двома способами. Перший спосіб – за допомогою генеративної змагальної мережі, або GAN, яка використовує дві окремі нейронні мережі, які працюють разом, навчаючись вивчати характеристики реальних зображень, щоб вони могли створювати переконливі фотопідробки. Другий спосіб – за допомогою алгоритму штучного інтелекту (ШІ), який має назву інкодер і працює шляхом запуску знімків облич двох людей через інкодер для знаходження схожості між цими знімками та запуску декодера, який отримує зображення обличчя та міняє їх місцями. Досягнення в області Deepfake однаково вражають і турбують. У чужих руках ця технологія може бути використана для поширення дезінформації та підриву суспільної довіри майже як науково-фантастичний тип крадіжки особистих даних, коли ви можете змусити будь-кого сказати що завгодно. У зв'язку з цим можна констатувати, що задача детектування Deepfake є дуже важливою та потребує застосування нових для цієї галузі досліджень, розробки нових алгоритмів детектування Deepfake та вдосконалення вже існуючих методів. Отже, метою цієї роботи є створення системи виявлення порушень цілісності медіа файлів, виконаних за технологією Deepfake. В роботі виконано вибір нейронної мережі та її навчання на спеціально створеній базі, що дало можливість виявляти Deepfake як у цифрових зображеннях, так і в цифрових відеозаписах.
Ключові слова: deepfake, виявлення фальсифікацій, цифрова криміналістика, нейронна мережа.

Вступ

На перший погляд технологія Deepfake може здаватися веселим, навіть корисним інструментом. Навіть високобюджетні фільми використовують технологію Deepfake, щоб повернути молодші версії улюблених персонажів. Але переваги цієї технології мінімальні в порівнянні з потенційною небезпекою, яку вона несе. Сьогодні ця технологія використовується для створення порнографії, фейкових новин, містифікацій, знущань, фінансового шахрайства тощо. У березні 2022 року діпфейк президента України Володимира Зеленського, який закликає своїх солдатів скласти зброю, потрапив на зламаний український новинний веб-сайт, а потім став вірусним, перш ніж було визнано, що він є фейком. Хоча точні збитки від цього інциденту залишаються невідомими, він продемонстрував одне з найнебезпечніших застосувань цієї технології. Якби це відео не було швидко позначено як фейк, воно могло б призвести до втрат і змінити хід конфлікту. Цю ж технологію можна використовувати як форму фішингу на робочому місці. Хоча хакери зазвичай використовують текстову переписку, щоб, наприклад, видавати себе за вашого боса чи колегу, тепер хакери можуть зателефонувати вам у Zoom, видавати себе за вашого боса та змусити вас придбати предмети або передати конфіденційну інформацію.

Опис технології Deepfake

Термін «Deepfake» походить від «Deep Learning» (Глибоке навчання) і «Fake» (Фейк), і він описує конкретні фотореалістичні відео або зображення, створені за підтримки глибокого навчання. Це слово було названо на честь анонімного користувача Reddit наприкінці 2017 року, який застосував методи глибокого навчання, щоб замінити обличчя людини в порнографічних відео, використовуючи обличчя іншої людини, і

створив фотореалістичні підроблені відео. Для створення таких підроблених відео використовувалися дві нейронні мережі: генеративна мережа та дискримінаційна мережа з технікою FaceSwap. Генеративна мережа створює підроблені зображення за допомогою кодера та декодера. Дискримінаційна мережа визначає автентичність знову створених зображень. Комбінація цих двох мереж називається Генеративною змагальною мережею (GAN).

Генеративні змагальні мережі (GAN) – це алгоритмічні архітектури, які використовують дві нейронні мережі, налаштовуючи одну проти іншої (тому «змагальні»), щоб генерувати нові синтетичні екземпляри даних, які можуть передаватись як реальні дані. Вони широко використовуються при генерації зображень, відео та голосу.

GAN були представлені в статті Яна Гудфеллоу [1] та інших дослідників з Університету Монреаля, включаючи Йошуа Бенджіо, у 2014 році. Згадуючи GAN, керівник дослідження ШІ Facebook Янн Лекун назвав змагальне навчання «найцікавішою ідеєю за останні 10 років у Машинному Вченні».

Потенціал GAN як для добра, так і для зла величезний, оскільки вони можуть навчитися імітувати будь-який розподіл даних. Тобто GAN можна навчити створювати світи, моторошно схожі на наш, у будь-якій області: зображення, музика, мова, проза. У певному сенсі вони художники-роботи. Але їх також можна використовувати для створення підробленого медіа-контенту, і вони є технологією, що лежить в основі Deepfake.

GAN працює наступним чином. Одна нейронна мережа, яка називається генератором, генерує екземпляри даних, а інша (дискримінатор) оцінює їх на достовірність; тобто дискримінатор вирішує, чи входить кожен екземпляр даних, який він переглядає, до фактичного набору даних для навчання, чи ні.

Наприклад, ми збираємось генерувати рукописні цифри, подібні до тих, що містяться в наборі даних MNIST [2], який береться з реального світу. Метою дискримінатора, коли він показує екземпляр із справжнього набору даних MNIST, є розпізнавання тих, які є автентичними. Тим часом генератор створює нові синтетичні образи, які передає дискримінатор. Це робиться в надії, що вони будуть визнані справжніми, навіть якщо вони підроблені. Мета генератора – генерувати прохідні рукописні цифри: брехати, не будучи спійманим. Мета дискримінатора – ідентифікувати зображення, яке надходить від генератора, як підробку. Ось кроки, які виконують GAN (рис.1).

1. Генератор приймає випадкові числа і повертає зображення.
2. Це згенероване зображення подається в дискримінатор разом із зображеннями, взятими з фактичного набору даних.
3. Дискримінатор бере як реальні, так і підроблені зображення та повертає ймовірність, число від 0 до 1, де 1 означає, що фото достовірне, а 0 вказує на підробку.

Отже, маємо подвійний цикл зворотного зв’язку: дискримінатор знаходиться в циклі зворотного зв’язку з реальним зображенням. Генератор знаходиться в петлі зворотного зв'язку з дискримінатором.

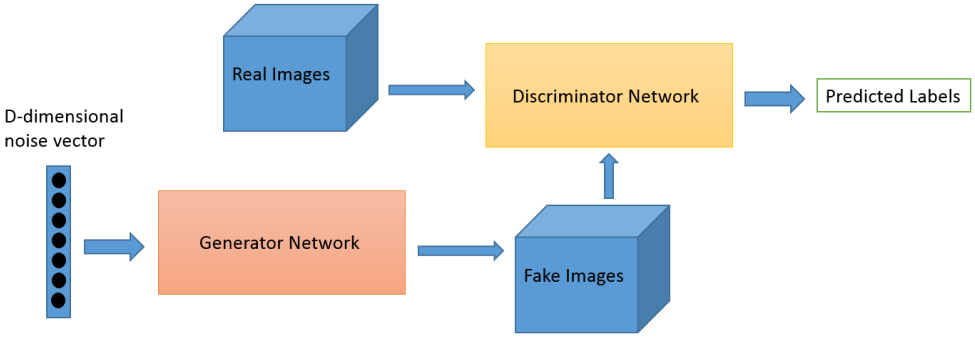

Рис. 1. Схема GAN

Буде корисним порівняти генеративні змагальні мережі з іншими нейронними мережами, такими як автокодери та варіаційні автокодери.

Автокодери кодують вхідні дані як вектори [3]. Вони створюють приховане або стиснене представлення необроблених даних. Вони корисні для зменшення розмірності; тобто вектор стискає вихідні дані в меншу кількість помітних вимірів. Автокодери можна об'єднати з так званим декодером, який дозволяє відновлювати вхідні дані на основі їх прихованого представлення.

Варіаційні автокодери – це генеративний алгоритм, який додає додаткове обмеження для кодування вхідних даних. Варіаційні автокодери здатні як стискати дані, як автокодер, так і синтезувати дані, як GAN. Однак, хоча GAN генерує дані з дрібними деталями, зображення, створені варіаційними автокодерами, мають тенденцію бути більш розмитими.

Методи детектування Deepfake

Протягом останніх років методи глибокого навчання успішно застосовуються для виявлення підроблених зображень. Однак сучасні методи глибокого навчання зображення не можуть бути безпосередньо застосовані для виявлення підроблених відео через наявність значної втрати інформації кадру після стиснення відео. Методи виявлення діпфейків можна поділити на дві основні категорії: ті, що працюють завдяки біологічному аналізу поодиноких осіб та ті, що працюють завдяки аналізу просторових і часових особливостей відео.

Біологічний аналіз поодиноких осіб. Юезен Лі [4] розробив метод, заснований на нейронній мережі для виявлення діпфейку у відео. Цей метод передбачає моргання очей для виявлення підроблених відео. Він використовує згорткову нейронну мережу з рекурсивною нейронною мережею для виявлення фізіологічних сигналів, таких як рух очей і моргання. Потім модель використовує двійковий класифікатор для визначення стану закритих і відкритих очей. Цей підхід перевірено за допомогою набору даних під назвою «eye-blinking» (моргання очей), який сканується з Інтернету. Результати експерименту демонструють ефективність запропонованого підходу у виявленні підроблених зображень.

Інші біологічні сигнали, такі як серцебиття, як було показано, є надійним провісником для реального відео. Ціфці та інші розробив модель на основі генеративної змагальної мережі (GAN), яка може виявляти джерело діпфейку, аналізуючи його сигнали [5]. Запропонована модель починається з кількох мереж детекторів, де вхідним сигналом для цієї моделі є реальне відео. Потім пара реалістичного відео та підробленого відео призначається іншому шару, який називається реєстрацією, який витягує цікаві ділянки обличчя і біологічні сигнали для створення клітин фото плетизмографії (ФПГ). Тут осередки ФПГ є просторово-часовими вікнами, які містять кілька облич, витягнутих за допомогою детектора облич. Останній шар відповідає за класифікацію відео як підробленого чи справжнього. Автори використали кілька загальнодоступних наборів даних для перевірки своєї моделі. Результат показує, що моделі активують точність 97,3% у виявленні діпфейків.

Аналіз просторових і часових ознак. Більшість сучасних методів виявлення діпфейків використовують лише один відеокадр [6]. Фактично, маніпуляції з відео можна виконувати з кількома функціями на рівні кадру. Останнім часом багато досліджень показали, що аналіз тимчасової послідовності між кадрами може успішно допомогти відрізнити реальне відео від підробленого. У цій роботі автори представили тимчасову модель для виявлення фейкових відео. У моделі спочатку використовується згорткова нейронна мережа для виділення ознак кадру. Згодом ці ознаки передаються на шар ДКЧП (довгої короткочасної пам'яті) для аналізу тимчасової послідовності зміни обличчя між Нарешті, функція softmax використовується для класифікації відео як реального чи підробленого. Для оцінки було зібрано колекцію з 600 відео з кількох веб-сайтів. Результати експерименту свідчать про ефективність цієї моделі для виявлення діпфейків.

На основі попередньої версії Cycle-GAN [7] було представлено новий підхід під назвою Recycle-GAN [8], який використовує умовні генеративні змагальні мережі для злиття просторових і часових даних. Результати оцінки показують, що поєднання просторових і часових обмежень може дати ефективний результат. Крім того, також було запропоновано новий підхід, заснований на рекурентній згортковій мережі [9]. Підхід складається з двох етапів аналізу: етап обробки обличчя з подальшим виявленням діпфейку. Під час обробки, кадрування та вирівнювання обличчя витягується за допомогою мережі просторового трансформатора. Потім вихідні дані з попередніх етапів передаються для виявлення діпфейку за допомогою рекурентної згорткової мережі, де аналізується тимчасова інформація по кадрах.

Вибір та навчання нейронної мережі

Глибокі нейромережі є областю машинного навчання, яке невпинно розвивається. Розвиток цього напряму призвів до появи великої кількості алгоритмів, побудованих на глибокому навчанні. Нейромережі використовують усі основні алгоритми для генерації підробних зображень.

Нейронні мережі відображають поведінку людського мозку, дозволяючи комп'ютерним програмам розпізнавати закономірності та вирішувати загальні проблеми в області ШІ, машинного навчання та глибокого навчання.

Нейронні мережі, також відомі як штучні нейронні мережі (ШНМ) або змодельовані нейронні мережі (ЗНМ), вони лежать в основі алгоритмів глибокого навчання. Штучні нейронні мережі (ШНМ) складаються з шарів вузлів, що містять вхідний шар, один або кілька прихованих шарів і вихідний шар (рис.2).

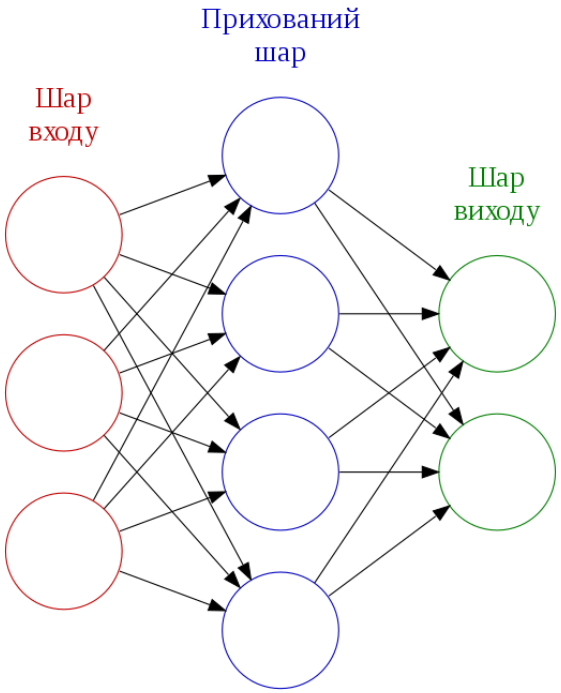

Рис.2. Структура нейронної мережі

Кожен вузол або штучний нейрон з'єднується з іншим і має відповідну вагу та поріг. Якщо вихід будь-якого окремого вузла перевищує вказане порогове значення, цей вузол активується, надсилаючи дані на наступний рівень мережі. В іншому випадку дані не передаються на наступний рівень мережі.

Нейронні мережі покладаються на навчальні дані, щоб з часом покращувати свою точність. Однак, як тільки ці алгоритми навчання будуть максимально налаштовані на точність, вони стануть потужними інструментами в інформатиці та штучному інтелекті,

що дозволить нам класифікувати дані з високою швидкістю. Однією з найвідоміших нейронних мереж є пошуковий алгоритм Google.

Інформація протікає через нейронну мережу двома способами. Коли вона навчається (тренується) або працює нормально (після навчання), шаблони інформації надходять у мережу через вхідні блоки, які запускають шари прихованих блоків, а ті, у свою чергу, надходять до вихідних блоків. Ця поширена конструкція називається мережею прямого зв'язку. Не всі блоки працюють постійно. Кожен блок отримує вхідні дані від одиниць зліва, і вхідні дані помножуються на вагу з'єднань, по яких вони рухаються. В кожному блоці необхідно знайти суму всіх вхідних даних, які він отримує таким чином, і (у найпростішому типі мережі), якщо сума перевищує певне порогове значення, блок запускає одиниці, до яких він підключений (ті, що праворуч).

Щоб нейронна мережа навчалася, має бути задіяний елемент зворотного зв'язку. Нейронні мережі тренуються за допомогою процесу зворотного зв'язку, який називається зворотним поширенням. Це включає порівняння результату, який виробляє мережа, з результатом, який вона повинна була генерувати, після чого відбувається використання різниці між ними для зміни ваги зв'язків між одиницями в мережі. З часом зворотне розповсюдження змушує мережу вчитися, зменшуючи різницю між фактичним і передбачуваним виходом до точки, де передбачуване і фактичне значення точно збігаються.

Згорткові нейронні мережі (ЗНМ) (рис.3) зазвичай використовуються для розпізнавання зображень, візерунків та/або комп'ютерного зору, тому для детектування Deepfake буде застосовуватися саме цей тип нейронних мереж.

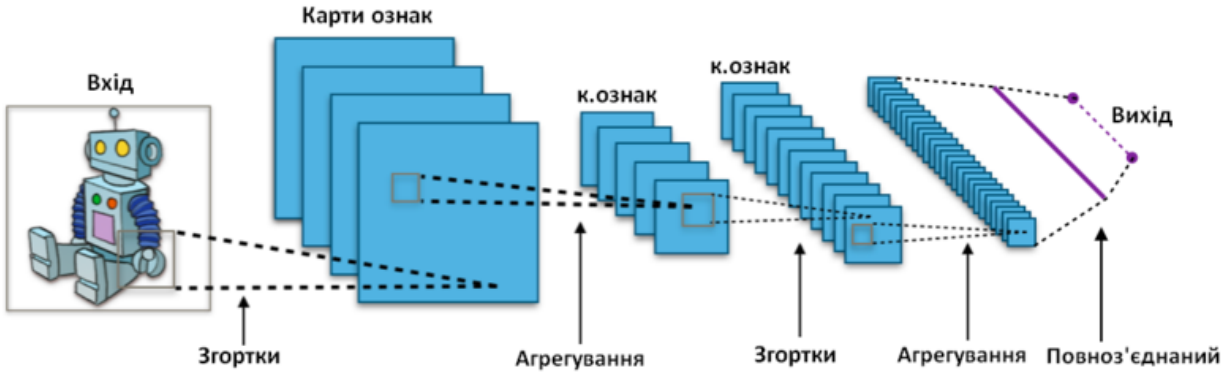

Рис.3. Структура згорткової нейронної мережі

Ці мережі використовують принципи лінійної алгебри, зокрема множення матриці для визначення шаблонів у зображенні. Вони мають три основних типи шарів, а саме:

- згортковий шар;
- об'єднуючий шар;
- повнозв'зний шар.

З кожним шаром ЗНМ збільшується у своїй складності, ідентифікуючи більші частини зображення. Більш ранні шари зосереджені на простих елементах, таких як кольори та краї. Коли дані зображення просуваються через шари ЗНМ, вона починає розпізнавати більші елементи або форми об'єкта, поки нарешті не ідентифікує передбачуваний об'єкт.

Припустимо, що вхідним буде кольорове зображення, яке складається з матриці пікселів у 3D. Це означає, що вхідні дані будуть мати три виміри – висоту, ширину та глибину – які відповідають RGB-зображенню. Згортковий шар має детектор особливостей, також відомий як ядро або фільтр, який переміщається по сприйнятливих полях зображення, знаходячи особливості. Цей процес відомий як згортка.

Детектор особливостей – це двовимірний (2-D) масив ваг, який представляє частину зображення. Хоча вони можуть відрізнятися за розміром, розмір фільтра зазвичай є матрицею 3×3; це також визначає розмір рецептивного поля. Потім фільтр застосовується до області зображення, і між вхідними пікселями та фільтром обчислюється скалярний добуток. Цей скалярний добуток потім подається у вихідний масив. Після цього фільтр зміщується на один крок, повторюючи процес, поки ядро не охопить усе зображення. Остаточний вихід із ряду точкових добутків із входу та фільтра відомий як карта ознак, карта активації або згорнута функція.

Шар об’єднання, також відомий як знижувач дискретизації, зменшує розмірність, зменшуючи кількість параметрів у вхідних даних. Подібно до згорткового шару, шар об’єднання застосовує фільтр по всьому зображенню, але різниця в тому, що цей фільтр не має жодних ваг. Замість цього ядро застосовує функцію агрегації до значень у сприйнятливому полі, заповнюючи вихідний масив.

Назва повнозв'язного шару влучно описує себе. У повнозв’язному шарі кожен вузол вихідного шару підключається безпосередньо до вузла попереднього шару. Цей шар виконує завдання класифікації на основі ознак, витягнутих через попередні шари та їх різні фільтри.

Для детектування Deepfake була обрана модель нейронної мережі (рис.4) ,заснована на статті, опублікованій Даріусом Афчаром та ін. у 2018 році [10]. Це двійковий класифікатор, побудований як відносно неглибока згорткова нейронної мережі, навчена класифікувати зображення на один із двох класів. Один клас відноситься до «реальних» зображень (зображення реальних людей), а інший – до «підроблених» зображень (зображень, створених Deepfake ШІ).

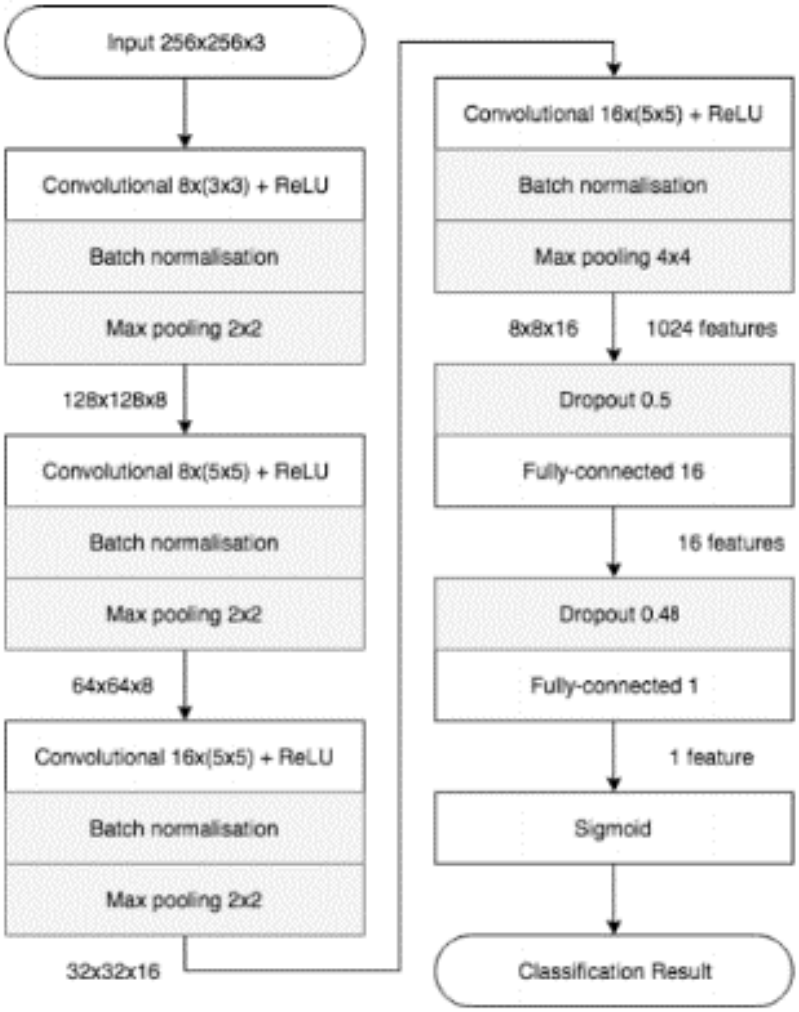

Рис.4. Модель нейронної мережі

− Вхідний шар 3×256×256, який приймає розмір зображення та кількість кольорових каналів.

− Згортковий шар із 8 фільтрами, розміром 3×3 і кроком 1, за яким слідує максимальний об'єднуючий шар розміром 2×2.

− Згортковий шар із 8 фільтрами розміром 5×5 і кроком 1, за яким слідує максимальний об'єднуючий шар розміром 2×2.

− Два згорткових шари з 16 фільтрами розміром 5×5 і кроком 1, за якими слідують максимальні шари об'єднання з вікном об'єднання 2×2.

− Повнозв'язний шар з 16 нейронами.

− Повністю підключений вихідний шар з одним блоком і сигмовидною активацією [11].

Згортковий шар, представлений conv2d, є найважливішою частиною нейронної мережі. Тут встановлюється розмір та кількість фільтрів, які використовуватимуться у згортці. Кожен фільтр є окремим елементом зображення, наприклад, горизонтальна лінія. Під час згортки цей фільтр проходить по зображенню, щоб оцінити, якою мірою певні області цього зображення відповідають фільтру. Після згорткового шару слідує шар пакетної нормалізації. Пакетна нормалізація – це новий метод підвищення швидкості та стабільності нейронних мереж. Він працює шляхом нормалізації вхідних даних для кожного шару мережі, що зменшує взаємозалежність між параметрами даного шару та вхідним розподілом наступного шару. Ця взаємозалежність називається внутрішнім коваріантним зрушенням і надає дестабілізуючу дію на процес навчання. Останній шар наших згорткових блоків – це шар пулу. Саме на рівні пулу значно зменшується розмірність даних, що значно прискорює обчислення. У даній моделі використовується максимальний пул цього шару, що означає зменшення області значень пікселів до максимального значення цієї області. Незважаючи на те, що ця архітектура ретельно дотримується, були проведені експерименти з різними функціями активації. Зокрема, крім використання стандартної активації ReLU також були проведені експерименти з ELU [12] та LeakyReLU [13].

ReLU є функцією активації вибору, оскільки немає очевидного ризику «мертвих нейронів» [14] . Крім того, існує активація LeakyReLU після повністю підключеного шару з 16 одиниць. За цим немає жодної видимої причини, крім того, що використовує газета.

Dropout – випадання, додається після повністю підключеного шару з 16 нейронами для боротьби з перенавчанням. Перенавчання – одна з проблем глибоких нейронних мереж (Deep Neural Networks, DNN), яка полягає в наступному: модель добре пояснює лише приклади з навчальної вибірки, адаптуючись до навчальних прикладів, замість того, щоб вчитися класифікувати приклади, що не брали участь у навчанні (втрата здатності до узагальнення). За останні роки було запропоновано безліч рішень проблеми перенавчання, але одне з них перевершило всі інші, завдяки своїй простоті та чудовим практичним результатам; це рішення – Dropout. Головна ідея Dropout – замість навчання однієї DNN навчити ансамбль кількох DNN, а потім усереднити отримані результати. Нейромережі для навчання створюються за допомогою виключення з нейронної мережі (dropping out) нейронів із ймовірністю p, таким чином, ймовірність того, що нейрон залишиться в мережі, становить q=1-p. «Виключення» нейрона означає, що при будь-яких вхідних даних або параметрах він повертає 0. Виключені нейрони не роблять свій внесок у процес навчання на жодному з етапів алгоритму зворотного поширення помилки (backpropagation); тому вимкнення хоча б одного нейрона рівносильне навчанню нової нейронної мережі.

Flatten (згладжування) – це перетворення даних в одновимірний масив для введення їх у наступний шар. Вирівнюються вихідні дані згорткових шарів, щоб створити один довгий вектор ознак. І це пов'язано з остаточною моделлю класифікації, яка

називається повністю зв’язаним шаром. Іншими словами, відбувається переміщення всіх піксельних даних в один рядок і з’єднання з останнім шаром.

Dense (повнозв’язний шар) – це простий шар нейронів, в якому кожен нейрон отримує вхідні дані від усіх нейронів попереднього шару, тому його називають щільним. Повнозв’язний шар використовується для класифікації зображень на основі вихідних даних із згорткових шарів. Робота одного нейрона. Шар містить кілька таких нейронів.

Для навчання нейронної мережі був використаний набір даних діпфейків, який складався з зображень, які були вилучені зі 175 відео, знятих з популярних платформ діпфейків, та модифікований, а саме з кожного зображення було вилучено обличчя і таким чином був створений новий набір даних для навчання нейронної мережі. Для вилучення облич використовувався метод HOG (гістограми орієнтованих градієнтів), який є дескриптором ознак і використовується в технологіях комп’ютерного зору та обробці зображень з метою виявлення об’єктів. Ця техніка підраховує кількість напрямків градієнта в локальних областях зображення. Дескриптор HOG фокусується на структурі або формі об’єкта. Це краще, ніж будь-який інший дескриптор ознак, оскільки він використовує величину градієнта, а також кут градієнта для обчислення ознак. Для областей зображення він генерує гістограми, використовуючи величину та орієнтацію градієнта. Градієнти розраховуються в межах розподілення зображення на блоки. Блок розглядається як піксельна сітка, в якій градієнти складаються з величини і напрямку зміни інтенсивності пікселя всередині блоку.

Модифікація набору даних була зроблена для збільшення якості навчання нейронної мережі. Фінальний набір даних складає 11780 зображень, які діляться на реальні зображення та діпфейки. Вхідними даними є зображення розміром 256×256×3, де 256×256 – висота і ширина зображення в пікселях, 3 – кількість кольорових каналів. Приклад оригінального зображення та зображення після модифікації можна подивитися на рис.5.

а) оригінальне зображення

б) модифіковане зображення

Рис. 5. Приклад зображень

Для навчання був використаний оптимайзер Adam – один із найефективніших алгоритмів оптимізації у навчанні нейронних мереж. Він поєднує в собі ідеї RMSProp та оптимізатора імпульсу [15] .

Після аналізу великої кількості тестувань на меншому наборі даних, який складав 1000 зображень, при швидкості навчання рівній 0.001 та чотирьох епохах було ідентифіковано, що найкращі результати у тренуванні та тестуванні показала модель нейронної мережі, де шар Dropout змінений з 0,5 до 0,48. Архітектура модифікованої моделі знаходиться на рис.6.

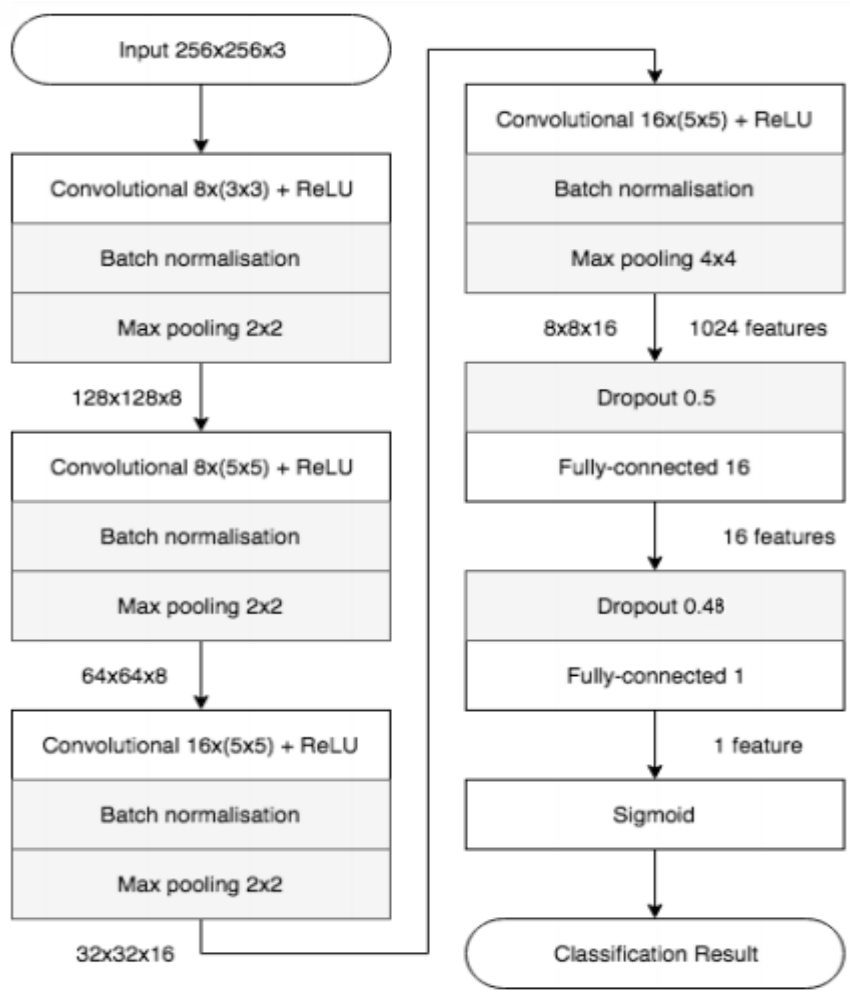

Рис.6. Змінена модель нейронної мережі

А саме після другої епохи була встановлена максимальна точність підтвердження (validation accuracy) – 98,57% та мінімальна втрата підтвердження (validation loss) – 1,26%. В той час на моделі нейронної мережі [1]: максимальна точність підтвердження (validation accuracy) – 97,58% та мінімальна втрата підтвердження (validation loss) – 1,93%. Графік навчання моделі [1] знаходиться на рис.7, а модифікованої моделі на рис.8.

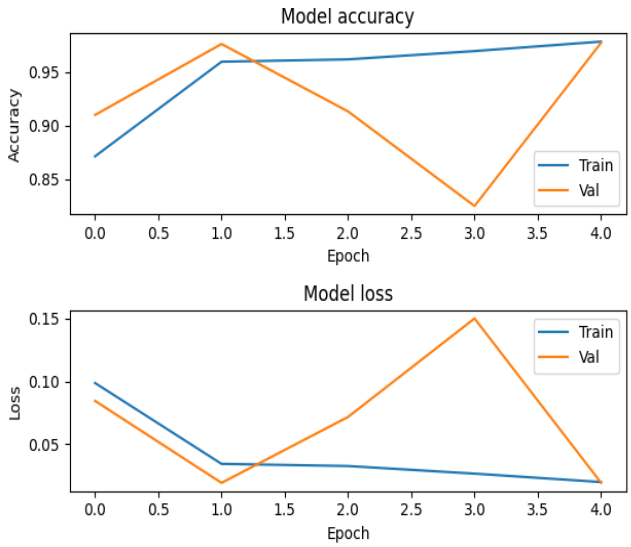

Рис. 7. Графік тестового навчання моделі нейронної мережі [1]

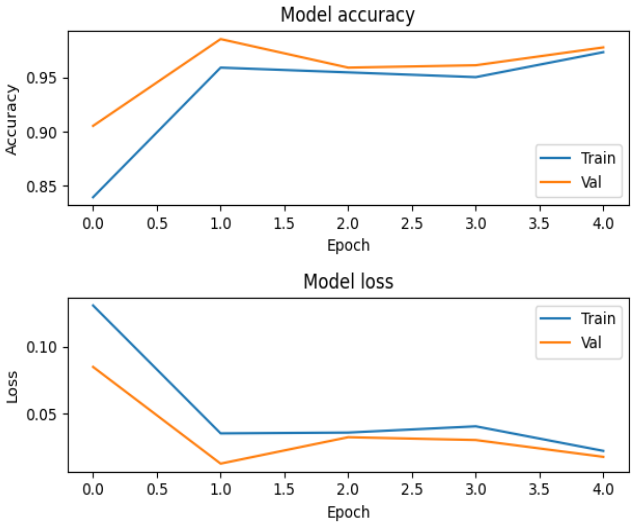

Рис.8. Графік тестового навчання модифікованої моделі нейронної мережі

При більшому розмірі набору даних, а саме – 11780 зображень, при швидкості навчання рівної 0.001 та 30 епохах у модифікованій моделі найкращий результат було отримано на 28-й епосі. Результат навчання моделі можна побачити у таблиці 1, а графік навчання на рис 9.

Таблиця 1

Результат навчання моделі модифікованої моделі нейронної мережі

loss	2,44%
acc	96,89%
val_loss	7,34%
val_acc	91,17%

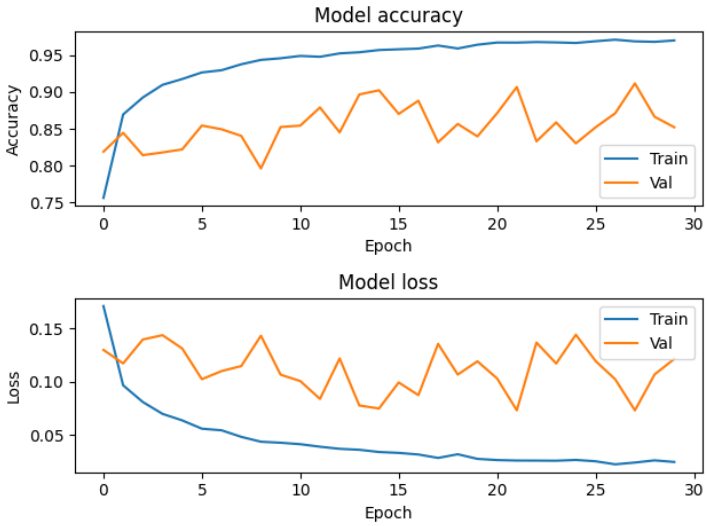

Рис.9. Графік тестового навчання модифікованої моделі нейронної мережі

Точність класифікації створеної моделі є 91,17%, що є досить великим показником, але крім того, завдяки модифікації набору даних вдалося підвищити точність класифікації зображень, які не входять до використаного для навчання моделі набору даних.

Висновки. В роботі представлено алгоритм детектування Deepfake за допомогою згорткової нейронної мережі. Точність класифікації створеної моделі є 91,17%, що є досить великим показником, але крім того, завдяки модифікації набору даних вдалося підвищити точність класифікації зображень, які не входять до використаного для

тестування моделі набору даних. Предметом подальшого розвитку даної роботи є вдосконалення нейромережі за рахунок збільшення набору даних для навчання.

Список літератури

1. Generative Adversarial Networks. URL: https://arxiv.org/abs/1406.2661
2. MNIST URL: https://paperswithcode.com/dataset/mnist.
3. Holländer B. (2020) Autoencoders: Overview of Research and Applications. URL: https://towardsdatascience.com/autoencoders-overview-of-research-and-applications-86135f7c0d35 ?gi=1b7b6132aae
4. Li Y., Chang M.C., Lyu S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *IEEE International Workshop on Information Forensics and Security (WIFS)*. 2018. P. 1-7.
5. Ciftci U.A., Demir I., Yin, L. How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. *IEEE International Joint Conference on Biometrics (IJCB)*, 2020, P.1-10.
6. Lima O., Franklin S., Basu S., Karwoski B., George A. Deepfake Detection Using Spatiotemporal Convolutional Networks. 2020. URL: https://www.researchgate.net/publication/342520585_Deepfake_Detection_using_Spatiotemporal_Convolutional_Networks
7. Zhu J.Y., Park T., Isola P., Efros A.A. (2017) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *IEEE International Conference on Computer Vision*. 2017, P. 2223-2232.
8. Bansal A., Ma S., Ramanan D., Sheikh Y. (2018) Recycle-GAN: Unsupervised Video Retargeting. *European Conference on Computer Vision*. 2018. P.119-135.
9. Sabir E., Cheng J., Jaiswal A., AbdAlmageed W., Masi I., Natarajan P. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *CVPR Workshops*. 2019
10. Afchar D et al. Mesonet: a compact facial video forgery detection network. URL: https://arxiv.org/abs/1809.00888
11. Функції активації нейромережі URL: https://neurohive.io/ru/osnovy-data-science/activation-functions/
12. Clevert D.A, Unterthiner T., Hochreiter S. (). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 2015. URL: https://doi.org/10.48550/arXiv.1511.07289
13. Maas A.L. Rectifier Nonlinearities Improve Neural Network Acoustic Models. 2013. URL: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf
14. Оссовский С. Нейронные сети для обработки информации. М.: Финансы и статистика, 2002. 344 с.
15. Реалізація та порівняння оптимізаторів моделей у глибокому навчанні URL: https://habr.com/ru/company/skillfactory/blog/525214/

DETECTION OF PHOTO FORGERY PERFORMED BY TOOLS OF NEURAL NETWORKS

V.V. Zorilo, E.V. Timofeev

1, Shevchenko Ave., National Odessa Polytechnic University, 65044, Odesa Ukraine
vikazorilo@gmail.com, timofeevzhenya2107@gmail.com

Deep learning is an effective and useful technique that is widely used in a variety of fields, including computer vision, machine vision, and natural language processing. Deepfake uses deep learning technology to manipulate images and videos of a person that people cannot distinguish from real ones. Deepfake is usually created in two ways. The first way is through a generative competition network, or GAN, which uses two separate neural networks that work together to learn how to study the characteristics of real images so that they can create compelling photo fakes. The second way is with an artificial intelligence (AI) algorithm called an encoder, which works by running pictures of two people's faces through an encoder to find similarities between those pictures and running a decoder that takes a face image and swaps them. The achievements in the field of Deepfake are equally impressive and disturbing. In someone else's hands, this technology can be used to spread misinformation and undermine public confidence almost as a sci-fi type of identity theft, when you can get anyone to say anything. In this regard, it can be stated that the task of Deepfake detection is very important and requires the application of new research in this area, the development of new algorithms for Deepfake detection and improvement of existing methods. Therefore, the purpose of this work is to create a system for detecting violations of the integrity of media files made by Deepfake technology. The work selected the neural network and its training on a specially created basis, which made it possible to detect Deepfake in both digital images and digital videos.
Keywords: deepfake, detection of falsifications, digital criminal science, neural network.