Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерні науки
(повна назва)

Кафедра Системотехніки
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий(магістерський)

«Дослідження та розробка методу виявлення підробок зображень та відеоконтенту»
(тема)

Виконав:
студент 2 курсу, групи СПРм-20-1
Захаров Р.С.
(прізвище, ініціали)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми Освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Системне проектування
(повна назва освітньої програми)
Керівник професор Калита Н.І.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри СТ _____ Гребеннік І.В.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Системотехніки

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми Освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне проектування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)

«__ » _____2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Захарову Роману Сергійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження та розробка методу виявлення підробок зображень та відеоконтенту»
затверджена наказом по університету від " 08 " листопада 2021р. № 1664 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 13 грудня 2021 р.
3. Вихідні дані до роботи Розробити метод виявлення підробок зображень та відеоконтенту. Перелік використовуваних програмних засобів: ОС mac OS, ОС Microsoft Windows 10, Python 3, JupyterLab 3.24, редактор початкового коду Xcode, OpenCV 4.5.2, PyTorch 1.10.0, Keras 2.4.0.
4. Перелік питань, що потрібно опрацювати у роботі 4.1 Аналіз предметної області. 4.1.1 Огляд та роль зображень і відеоконтенту у сучасному світі. 4.1.2 Аналіз та огляд зображень. 4.1.3 Аналіз та огляд відео технологій. 4.2 Огляд методів і технологій, що застосовуються в предметній області. 4.2.1 Опис нейронних мереж. 4.2.2 Детальний опис компонентів та принципу роботи штучних нейронних мереж. 4.2.3 Опис згорткових нейронних мереж. 4.2.4 Принцип роботи згорткових нейронних мереж. 4.2.5 Автокодувальник 4.2.6 Опис генеративно-змагальних мереж. 4.2.7 Постановка задачі. 4.3 Дослідження технологій підробок зображень і відео та математичний опис розроблюваного методу для їх виявлення. 4.3.1 Опис та принцип роботи технології DeepFake. 4.3.3 Існуючі підходи для виявлення підроблених зображень та відеоконтенту. 4.3.3 Принцип роботи та математичний опис методу виявлення підробок зображень та відео. 4.3.4 Отримані практичні результати розроблюваного методу виявлення підробок зображень та відео 4.4 Висновки 4.5 Перелік джерел посилань. 4.6 Додатки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) 5.1 Оригінальна архітектура ГЗМ (1 аркуш формату А4). 5.2 Створення DeepFake з використанням кодувальників та декодувальників (1 аркуш формату А4). 5.3 Архітектура FSGAN (1 аркуш формату А4). 5.4 Приклад зображень, із зміненими характеристиками, які не були виявлені методами виявленнями підробок з використанням ЗНМ (1 аркуш формату А4). 5.5 Архітектура технології та алгоритм розробленого методу виявлення підробок медіа-файлів (1 аркуш формату А4). 5.6 Отримані результати роботи мережі P-Net при виявленні обличчя на зображенні (1 аркуш формату А4). 5.7 Процес отримання ознак комп'ютерного зору (1 аркуш формату А4). 5.8 Знаходження обличчя на одному з кадрів підробленого відео за допомогою використання мережі MTCNN (1 аркуш формату А4). 5.9 Кадри з найбільшою розбіжністю кожної ознаки з одного підробленого відео (1 аркуш формату А4)

Календарний план

№	Назва етапів роботи	Терміни виконання етапів	Примітка
1.	Вивчення предметної області з теми кваліфікаційної роботи	08.11. – 10.11.21	виконано
2.	Підбір та вивчення джерел інформації з теми дослідження	11.11.– 13.11.21	виконано
3.	Аналіз та опис методів і технологій, що використовуються у предметній області	14.11. – 17.11.21	виконано
4.	Постановка задачі дослідження та розробки методу виявлення сфальсифікованих зображень і відео	18.11. – 19.11.21	виконано
5.	Опис та принцип роботи існуючих технологій підробок медіа-контенту та методів для їх виявлення	20.11. – 22.11.21	виконано
6.	Математичний опис розроблюваного методу виявлення підробок зображень та відео	23.11. – 25.11.21	виконано
7.	Розробка методу та отримання практичних результатів	26.11 – 03.12.21	виконано
8.	Оформлення пояснювальної записки та програмної документації	04.12 – 11.12.21	виконано
9.	Оформлення графічної частини та презентаційних матеріалів комп'ютерного захисту	12.12.21	виконано
10.	Представлення на рецензування	13.12.21	виконано
11.	Представлення кваліфікаційної роботи до ЕК	15.12.21	виконано

Дата видачі завдання 08 листопада 2021р.

Студент _____ Захаров Р.С.
(підпис)

Керівник роботи _____ проф. Калита Н.І.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка до магістерської кваліфікаційної роботи: 87 с., 4 табл., 40 рис., 2 додатки, 29 джерел інформації

ЗОБРАЖЕННЯ, ВІДЕО, ГЛИБИННА ПІДРОБКА, НЕЙРОННІ МЕРЕЖІ, ЗГОРТКОВІ НЕЙРОННІ МЕРЕЖІ, ОСОБЛИВОСТІ КОМПЮТЕРНОГО ЗОРУ, ПОПЕРЕЛНЯ ОБРОБКА ДАНИХ, КЛАСИФІКАЦІЯ

Об’єктом дослідження є процес підробки зображень та відео.

Предметом дослідження є методи перевірки зображень та відео на оригінальність чи підробку.

Метою даної кваліфікаційної роботи являється дослідження та розробка методу виявлення підробок зображень та відеоконтенту.

Методи дослідження: системний аналіз, комп’ютерний зір, штучні нейронні мережі, методи обробки зображень, згортки та класифікації.

Наукова новизна: запропоновано метод виявлення підробок зображень та відео, в якому на першому етапі виконується попередня обробка зображень і відео на основі аналізу ознак комп’ютерного зору, на другому етапі зображення класифікуються на основі розбіжності певних ознак між кадрами.

Отримані результати дослідження можна використати у якості даних для розробки програмного застосунку, який може знайти своє застосування певними компаніями або правоохоронними органами у кібербезпечних цілях для виявлення підробок задля спростування дезінформації та шантажу.

ABSTRACT

Master's Thesis: 87 pages, 4 tables, 40 figures., 2 appendices, 29 title.

IMAGE, VIDEO, DEEPFAKE, NEURAL NETWORKS, CONVOLUTIONAL NETWORK, FEATURES OF COMPUTER VISION, PREPROCESSING, CLASSIFICATION

The object of the research is the process of image and video tampering.

The subject of the research is methods of checking images and videos for originality or forgery.

The aim of this master's thesis is to research and develop a method for detecting tampered images and video content.

Research methods: systems analysis, computer vision, artificial neural networks, image cropping, convolution and classification methods.

Scientific novelty: a method for detecting fakes of images and videos is proposed, in which at the first stage preliminary processing of images and videos is performed based on the analysis of signs of computer vision, at the second stage the images are classified based on the discrepancy of certain features between frames.

The findings can be used as data to develop a software application that can find its application by certain companies or law enforcement agencies for cybersecurity purposes to detect fakes to refute disinformation and blackmail.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ГЗМ – генеративно-змагальна мережа;

ЗНМ – згорткова нейронна мережа;

НМ – нейронна мережа;

ШНМ – штучна нейронна мережа;

BEGAN – Boundary Equilibrium Generative Adversarian Network (згорткова мережа граничної рівноваги);

CGAN – Conditional Generative Adversarian Network (умовна згорткова мережа);

CycleGAN – Cycle Generative Adversarian Network (циклічна згорткова мережа);

DCGAN – Deep Convolutional Generative Adversarian Network (глибока згорткова мережа);

DDCD - Deepfake Detection Challenge Dataset (набір даних Deepfake Detection Challenge);

InfoGAN – Informational Generative Adversarian Network (інформаційна згорткова мережа);

FSGAN – Face Swapping Generative Adversarian Network (згорткова мережа заміни обличчя);

LSTM – Long short-term memory (довгий ланцюг елементів короткострокової пам'яті);

MTCNN – Multi-Task Cascaded Convolutional Network (багатозадачна каскадна згорткова мережа);

OpenCV – Open-Source Computer Vision Library (бібліотека комп’ютерного зору з відкритим кодом);

ProGAN – Progressive Generative Adversarian Network (прогресуюча згорткова мережа).

ВСТУП

На сьогоднішній день із шаленим прискоренням відбувається розвиток різноманітних технологій, науки та техніки. У повсякденному житті дійсно велику роль відіграють мультимедійні системи, оскільки переважна більшість інформації подається за допомогою відео, звуку та зображень, які використовуються в професійних, навчальних та розважальних цілях. При цьому все частіше розроблюються сучасні системи, які обладнані штучним інтелектом, для побудови та відтворенні медіа-контенту в індустрії бізнесу, реклами, ігор, тощо.

Але окрім позитивних отриманих результатів розвитку цього напряму також з’явились і негативні наслідки, оскільки зараз існує велика кількість відмінних способів для обману чи шантажу людей, якими зазвичай користуються шахраї, або з іншої сторони лише звичайні люди заради жарту та розваги.

Одним із таких способів, який базується на методах штучного інтелекту, є підробка зображень і відео. Даний метод має назву Deepfake, що включає у собі такі поняття та випливає з них як: «глибинне навчання» та безпосередньо «підробка». Deepfake використовують технологію глибокого навчання для маніпулювання зображеннями та відео людини, щоб вона не змогла відрізнити їх від справжніх. Ця методика дозволяє, шляхом накладання обличчя у вигляді зображення, змінити обличчя іншою людини у певному відеоконтенті та створити унікальне відео, де людина буде повторювати усі дії та відтворювати мову особи із похідного відеоряду.

Задля створення насправді чіткої та схожої на реальність підробку потребується достатньо багато фотокарток або відеоданих. Через те, що у зірок шоу-бізнесу, спортсменів, політиків та у інших відомих людей дуже багато фото та відео матеріалів знаходяться у відкритому доступі в Інтернеті, вони становляться легкою мішенню, застосовуючи техніку Deepfake.

Тому виникає потреба щодо знаходження та виявлення таких підробок, щоб у подальшому спростувати поширення дезінформації та інших побічних ефектів.

В останні роки було проведено безліч досліджень, щоб зрозуміти, яким чином працює ця технологія та було представлено багато підходів, що засновані на глибокому навчанні, для виявлення підроблених зображень та відео.

Але розвиток Deepfake також з кожним роком стає все більш потужним, тому потрібно знаходити більш точні та надійні методи для виявлення сфальсифікованих зображень та відео.

Підводячи підсумок, із впевненістю можна сказати, що аналіз Deepfake, пошук та розробка методу для виявлення підробок зображень та відеозаписів є справді актуальною темою, що безперечно буде мати розвиток у подальшому майбутньому.

Метою даної кваліфікаційної роботи являється дослідження та розробка методу виявлення підробок зображень та відеоконтенту.

Об’єктом дослідження є процес підробки зображень та відео.

Предметом дослідження є методи перевірки зображень та відео на оригінальність чи підробку.

Методи дослідження: системний аналіз, комп’ютерний зір, штучні нейронні мережі, методи обробки зображень, згортки та класифікації.

Наукова новизна: запропоновано метод виявлення підробок зображень та відео, в якому на першому етапі виконується попередня обробка зображень і відео на основі аналізу ознак комп’ютерного зору, на другому етапі зображення класифікуються на основі розбіжності певних ознак між кадрами.

Отримані результати дослідження можна використати у якості даних для розробки програмного застосунку, який може знайти своє застосування певними компаніями або правоохоронними органами у кібербезпечних цілях для виявлення підробок задля спростування дезінформації та шантажу.

Дана робота подана на опублікування в науковий журнал університету ХНУРЕ.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Огляд та роль зображень і відеоконтенту у сучасному світі

Сучасний світ із плином часу перейшов на такий етап життя, де у якості головної ролі виступає інформація та економіка, яка будується на ній. Розвиток інформаційного суспільства цілком пов’язан із необхідністю збирання, оброблення та безумовно передачі дуже великих об’ємів інформації. На даний момент інформація являється одним із найцінніших ресурсів, якими користується суспільство. Існує багато способів для зберігання та розповсюдження інформації. Але зараз одними із найкращих для цього методів є використання зображень та відео. Нещодавно було проведено опитування серед людей, що проживають у розвинених країнах, для того, щоб дізнатися, якому вигляду споживання інформації вони надають перевагу. На рисунку 1.1 представлено результати цього опитування [1].

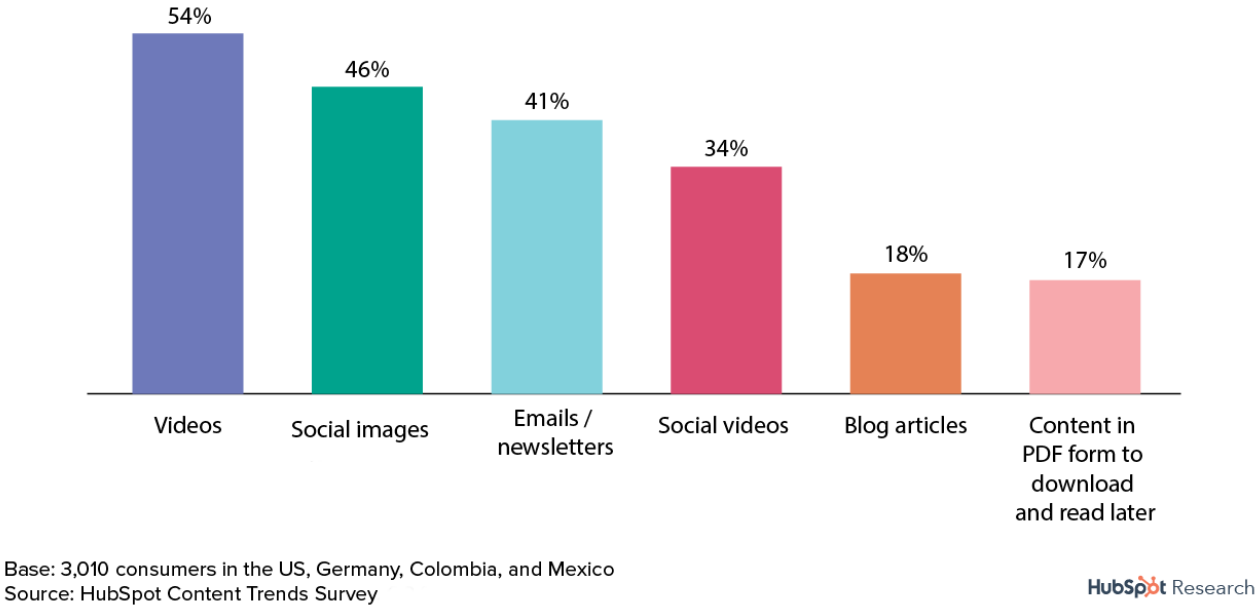

Рисунок 1.1 - Результати опитування щодо переважного способу споживання інформації

Спираючись на дане опитування можна зробити висновок, що відео та зображення являються найпоширенішими та більш приваблюваними шляхами для отримання певної інформації.

Фото та відеоконтент застосовують майже у всіх сферах життя в навчальних та розважальних цілях, для реклами, бізнесу, політики, новин, тощо.

Кожного року прогресують та з'являються все більше соціальних мереж в Інтернеті, у яких міститься велика кількість медіа-контенту. Найбільш популярним ресурсом для розміщення фотографій являється соціальна мережа Instagram. Станом на 2021 рік даним застосунком користується вже більше одного мільярда людей, розміщуючи мільйони зображень кожного дня. На рисунку 1.2 зображено статистику збільшення користувачів даної соціальної мережі [2].

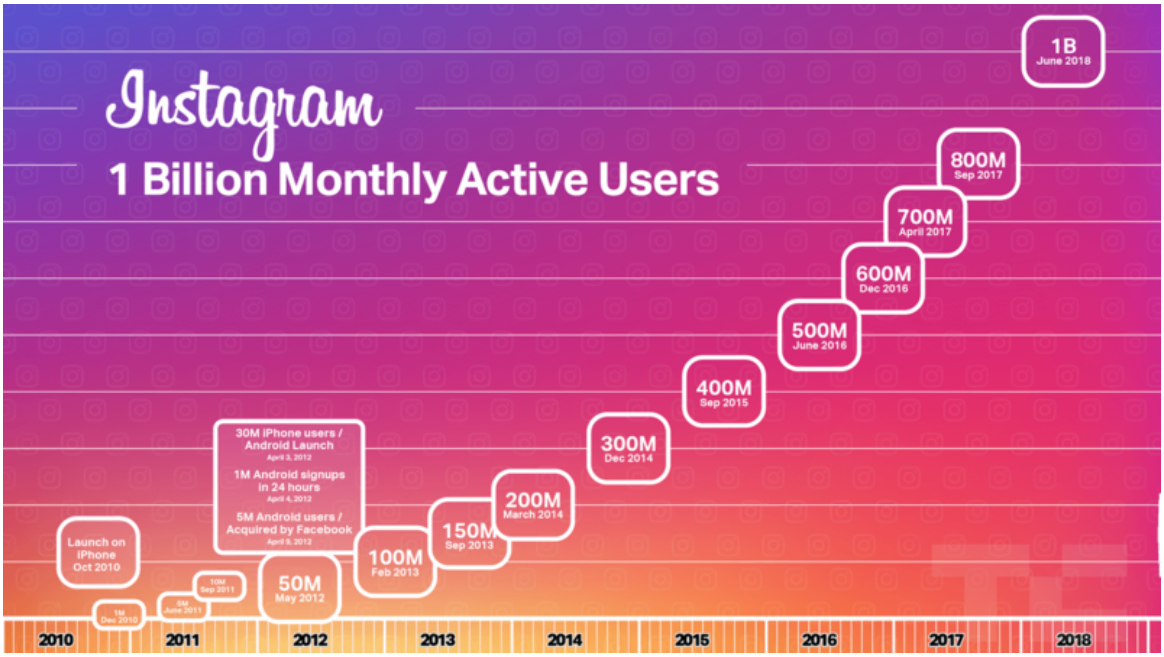

Рисунок 1.2 - Статистика активних користувачів Instagram

Для завантаження та перегляду виключно відеоконтенту на даний момент існують два великі та широко відомі сервіси, як YouTube та TikTok. Не дивлячись на те, що TikTok створили відносно нещодавно, він вже є одним із найбільш використовуваним застосунком, особливо серед молоді. На рисунку 1.3 представлено шалені темпи росту популярності цієї соціальної мережі, де кожного дня користувачі завантажують безліч нових відео [3].

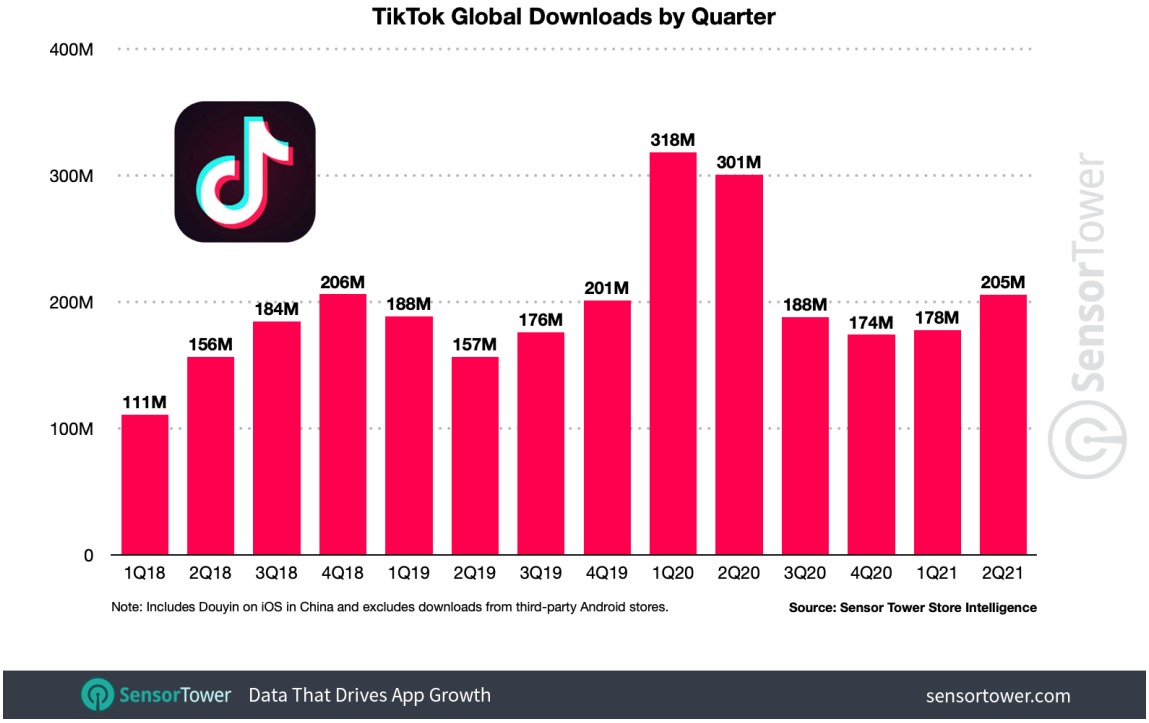

Рисунок 1.3 - Статистика росту популярності TikTok

1.2 Аналіз та огляд зображень

Зображенням називають відтворення форми, виду та кольору предмета світловими променями, що пройшли через оптичну систему з центрованих сферичних поверхонь та які мають одну загальну оптичну вісь. На рисунку 1.4 зображено оптичну вісь, що збігається з червоним променем та симетричні щодо неї промені зеленого та синього кольору, які проходять через різні лінзи [4].

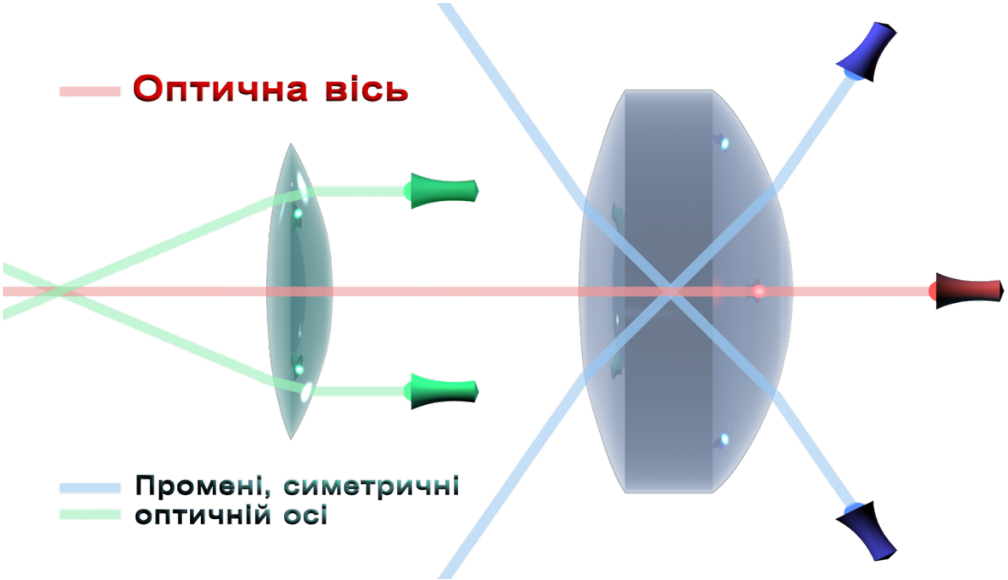

Рисунок 1.4 - Оптична вісь

За умови, коли зображення предмета утворено перетином самих променів, то воно є дійсним, а якщо їх продовженням, то уявним.

Існує декілька типів зображень, а саме:

– оптичне зображення (монохромне, напівтонове, повнокольорове);
– цифрове зображення (растрове, векторне);
– ортоскопічне зображення;

Оптичні зображення використовуються в наукових цілях, а саме в космічній та в аерофотозйомці. Ортоскопічні зображення використовуються в 3-D кінематографі, а цифрові зображення застосовується в інформаційних технологіях.

Цифрове зображення являє собою двовимірне зображення, представлене в цифровому вигляді. Цифровим зображенням можна назвати створення закодованого в цифровій формі уявлення візуальних характеристик об'єкта, таких як фізична сцена або внутрішня структура об'єкта. Цифрове зображення включає у собі обробку, стиснення, зберігання, друк і відображення таких зображень. Ключовою перевагою цифрового зображення в порівнянні з аналоговим зображенням, таким як фотографія з плівки, є можливість робити копії та копії копій в цифровому вигляді на невизначений термін без будь-яких втрат якості зображення. Залежно від способу опису, зображення може бути растровим або векторним [4].

Растрове зображення представляється як прямокутний двовимірний масив чисел, при цьому кожне число відповідає одному елементу зображення або пікселя. Зазвичай ці масиви передаються та зберігаються у стиснутому вигляді.

Важливими характеристиками зображення є:

– розмір зображення в пікселях;
– кількість використовуваних кольорів або глибина кольору;
– кольорова модель;
– дозвіл зображення;

Глибина кольору має наступну залежність:

$$\mathrm{N} = 2^k, \tag{1.1}$$

де N - кількість кольорів;

k - глибина кольору.

Растрові зображення створюють за допомогою різних пристроїв, таких як цифрові фотоапарати, цифрові відеокамери, сканери, тощо. Або вони синтезуються штучно засобами машинної графіки. Залежно від типу стиснення може бути можливо або неможливо відновити зображення в точності таким, яким воно було до стиснення. Найпопулярніший формат, що стискує без втрат являється PNG. До найбільш відомого формату з втратами відноситься формат JPEG. Стиснення використовує розбиття зображення на блоки, квантування просторових спектральних компонент в кожному блоці зображення з подальшим їх ентропійним кодуванням. При детальному розгляді сильно стисненого зображення помітно розмиття різких кордонів і характерний муар поблизу них. При невисоких ступенях стиснення відтворене зображення візуально не відрізняються від вихідного. Дані зображення зберігаються у стиснутому вигляді [4].

До переваг растрових зображень можна віднести:

– висока швидкість обробки складних зображень;

– поширеність;

– растрове представлення зображення природно для більшості пристроїв введення-виведення графічної інформації.

Але вони також мають свої недоліки:

– великий розмір файлів у простих зображень з великої кількості точок;

– неможливість ідеального масштабування;

– неможливість виведення на друк на векторний графічний пристрій.

Векторні зображення отримуються шляхом математичного опису елементарних геометричних об’єктів, що називаються примітивами, такі як крапки, лінії, сплайни та окружності.

Перевагами цього типу зображень є можливість описувати будь-які великі об’єкти файлом мінімального розміру, параметри яких можна легко змінити.

До недоліків відносяться специфікації векторних зображень, які набагато складніші ніж у растрових та те, що не кожна графічна сцена може бути легко зображена у векторному вигляді.

Завдяки цифровим зображенням був здійснений значний прорив у сфері освіти та в області візуалізації медицини. В області технологій цифрова обробка зображень стала більш корисною, ніж обробка аналогових зображень, оскільки з’явилась можливість відтворювати старі зображення та підвищувати різкість, розпізнавати обличчя людей, оброблювання кольорів та інші [5].

Використання цифрових зображень має свої переваги та недоліки. Безперечно вони забезпечують легкий доступ до фотографій та текстових документів, цифрова обробка зображень знижує необхідність фізичного контакту з вихідними зображеннями, а також можливість розширення цифрової фотографії до камери телефонів. Ми можемо брати фотоапарати з собою куди завгодно, а також миттєво відправляти фотографії іншим. Але до негативних наслідків відноситься можливість маніпулювання зображеннями шляхом їх видозміни та підробки.

1.3 Аналіз та огляд відео технологій

Відео являє собою електронну технологію формування, запису, обробки, передачі, зберігання та відтворення рухомого зображення, що засноване на принципах телебачення, а також аудіовізуальний твір, записане на фізичному носії [6].

Існують два види відео - аналогові та цифрові, які можуть передаватися на різних носіях, включаючи радіомовлення, магнітну стрічку, оптичні диски, комп'ютерні файли і мережеві потоки.

Відео технологія була вперше розроблена для механічних телевізійних систем, які були швидко замінені телевізійними системами з катодно-променевою трубкою (ЕПТ), але з тих пір було винайдено кілька нових технологій для пристроїв відображення відео. З самого початку відео було виключно технологію прямого ефіру. Чарльз Гинсбург очолював дослідницьку групу компанії Ampex, що розробила один з перших практичних відеомагнітофонів (VTR). У 1951 році перший VTR захопив живе зображення з телевізійних камер шляхом запису електричного сигналу камери на магнітну відеострічками [7].

Використання цифрових технологій в відео дозволило створити цифрове відео. Спочатку воно не могло конкурувати з аналоговим відео, оскільки раннє цифрове нестиснене відео вимагало непрактично високих бітрейтів, тобто кількість біт, що використовуються для передачі / обробки даних в одиницю часу. Бітрейт прийнято використовувати при вимірюванні ефективної швидкості передачі потоку даних по каналу, а саме мінімального розміру каналу, який зможе пропустити цей потік без затримок [7].

Практичне цифрове відео стало можливим завдяки кодуванню з дискретного косинусного перетворення (DCT), процесу стиснення з втратами, розробленим на початку 1970-х років. DCT-кодування було адаптовано до стиснення відео з компенсацією руху DCT в кінці 1980-х років, починаючи з H.261, першого практичного стандарту кодування цифрового відео [7].

Згодом цифрове відео стало володіти більш високою якістю і, в кінцевому підсумку, набагато більш низькою вартістю, ніж ранні аналогові технології. Після винаходу DVD в 1997 році, а потім Blu-ray Disc в 2006 році, продажу відеокасет і записуючого обладнання різко впали. Розвиток комп'ютерних технологій дозволяє навіть недорогим персональним комп'ютерам і смартфонам записувати, зберігати, редагувати і передавати цифрове відео, що ще більше знижує вартість виробництва відео, дозволяючи творцям програм і віщальним компаніям перейти на безстрічкове виробництво. Поява цифрового мовлення і подальший перехід на цифрове телебачення в більшості країн світу відтісняють аналогове відео на другий план. Станом на 2015 рік, із зростанням використання відеокамер високої роздільної здатності з поліпшеним динамічним діапазоном і кольоровою гамою, а також цифрових форматів проміжних даних з високим динамічним діапазоном і поліпшеною глибиною кольору, сучасні цифрові відеотехнологій зближуються з цифровими кінотехнологіями [7].

До характеристик відеосигналу відносять:

– частота кадрів;
– стандарт розкладання;
– роздільна здатність;
– співвідношення сторін екрану;
– композитність та компонентність;

Кількість (частота) кадрів в секунду - це число нерухомих зображень, що змінюють один одного при показі 1 секунди відеозапису і створюють ефект руху об'єктів на екрані. Чим більше частота кадрів, тим більше плавним і природним буде здаватися рух. Мінімальний показник, при якому рух буде сприйматися однорідним - приблизно 16 кадрів в секунду, але дане значення індивідуальне для кожної людини. Комп'ютерне відео хорошої якості, як правило, використовує частоту 30 кадрів в секунду. Верхня гранична частота мерехтіння, що сприймається людським мозком, в середньому становить 39-42 Гц та зазвичай залежить від умов спостерігання [7].

Стандарт розкладання визначає параметри телевізійної розгортки, яка застосовується для перетворення двовимірного зображення в одновимірний відеосигнал або потік даних. В кінцевому рахунку від стандарту розкладання залежить кількість елементів зображення і кадрова частота.

Розгортка може бути прогресивною або черезрядковою. При прогресивній розгортці всі горизонтальні лінії (рядки) зображення відображаються по черзі одна за одною. При черезрядковій розгортці кожен кадр розбивається на два поля (пів-кадра), кожне з яких містить парні або непарні рядки. За час одного кадру передаються два поля, збільшуючи частоту мерехтіння кінескопа вище фізіологічного порога помітності. Черезрядкова розгортка була компромісом, щоб мати можливість передачі по каналу з обмеженою пропускною здатністю зображення з досить великою роздільною здатністю [7].

Нові цифрові стандарти телебачення, наприклад HDTV, передбачають прогресивну розгортку. Новітні технології дозволяють імітувати прогресивну розгортку при показі відео з чергуванням. Останню зазвичай позначають символом «i» після вказівки вертикального дозволу, наприклад 720 × 576i × 50. Прогресивну розгортку позначають символом «p», наприклад 720p (означає відео з роздільною здатністю 1280 × 720 з прогресивною розгорткою). Також для відмінності частоти кадрів або полів може позначатися такими ж символами кадрова частота, наприклад 24p, 50i, 50p.

До настання цифрової ери відео, горизонтальна роздільна здатність аналогової системи відеозапису вимірювалася в вертикальних телевізійних лініях за допомогою спеціальних телевізійних випробувальних таблиць і позначала кількість елементів в рядку відеозображення, залежне від частотних характеристик пристрою запису. Вертикальна роздільна здатність в зображенні закладена в стандарті розкладання і визначається кількістю рядків.

Співвідношення ширини та висоти кадра є дуже важливим параметром будь-якого відеозапису. Цифрове телебачення стандартної чіткості в основному орієнтується на співвідношення 16:9, застосовуючи цифрове анаморфірування, тобто технологію передачі і запису широкоформатних фільмів цифрового телебачення за допомогою стандартів розкладання, спочатку розрахованих на класичне співвідношення сторін екрану 4:3. При цьому інформаційна ємність такого кадру використовується найбільш ефективно за рахунок трансформації співвідношення сторін пікселя [7].

Кольоровий відеосигнал може передаватися і записуватися двома різними способами: без поділу кольорового і монохромного складових і окремо. Історично першим з'явилося композитне відео, зване Повним кольоровим телевізійним сигналом і містить чорно-білий відеосигнал, колірну піднесну та сигнали синхронізації. Однак такий спосіб зберігання і передачі пов'язаний з неминучим накопиченням перехресних перешкод між сигналами яскравості і кольоровості, тому в найбільш досконалих пристроях ці складові відео передаються і записуються окремо.

Основна відмінність від аналогового відеозапису в тому, що замість аналогового відеосигналу записуються цифрові дані. Цифрове відео може поширюватися на різних відеоносіях, за допомогою цифрових відеоінтерфейсів як потоковий вміст або файлів.

Цифрове зображення, як правило, складається з пікселів. Чим їх більше, тим якісніше і важче це зображення. Піксель являє собою квадратний елемент на екрані комп'ютера. Кожен піксель має певний колір. Формат кодування кольору в пікселі при виведенні його на екран визначається поточними установками відеоадаптера комп'ютера [7].

При виведенні інформації на екран комп'ютера зазвичай використовується формат RGB. По суті, це комбінація трьох кольорів: червоного (Red), зеленого (Green) і синього (Blue) з яких будуються всі інші кольори і відтінки. Результуючий колір варіюється в залежності від процентного вмісту кожного з трьох складових кольорів. Наприклад: R-0 + G-255 + B-0 дасть нам зелене (G = 255), а R-255 + G-255 + B-255 дадуть білий, відповідно, якщо всі три кольори поставити в нуль, то вийде чорний колір.

Режими відеокарти зазвичай перемикаються за допомогою стандартних засобів операційної системи. Крім кольору задається дозвіл екрана, тобто «Розмір» зображення на екрані в пікселях по горизонталі і по вертикалі (найбільш часто використовувані стандартні дозволу: 640 × 480, 800 × 600, 1024 × 768, 1280 × 1024).

Для виводу зображення на екран існують декілька режимів, а саме:

– режим 8 bit (палітровий);
– режим 16 bit (HiColor);
– режим 24 bit (TrueColor);
– режим 32 bit (TrueColor з підтримкою технології MMX).

Палітровий режим відображає 256 кольорів. Колір пікселя кодується у вигляді номера в таблиці (палітрі) з 256 елементів. Для кожного елемента в палітрі зберігається його RGB зміст. Відеокарта бере значення пікселя (номер кольору в палітрі, від 0 до 255) і за значенням отримує значення його RGB складових, які і виводяться на екран [7].

Режим HiColor відображає 65536 кольорів, що забезпечує досить реалістичне зображення. При графічних операціях піксель (крапка) кодується всього двома байтами, що прискорює обробку зображення. Цей режим звичайно є найкращим для роботи з нестислим виглядом відеозображенням [7].

Режим TrueColor являється найбільш реалістичним, оскільки він має можливість відображувати 16581375 кольорів.

Для зменшення займаного обсягу при зберіганні та передачі відео виконується стиснення графічної інформації. Існує ряд способів стиснення цифрового зображення. Один з найпоширеніших полягає в об'єднанні кількох дрібних сусідніх елементів зображення в один. У цьому випадку досягається хороша ступінь стиснення, але знижується якість зображення. І чим більшою буде ступінь стиснення, тим менше докладним і більше нечітким буде результуюче зображення. Корисними наслідками результату операції стиснення є менші тимчасові та ресурсні витрати на наступні маніпуляції з таким зображенням. Файли зі стисненим зображенням займають менше місця і передаються по каналах зв'язку швидше, ніж файли з оригінальним, нестиснутим зображенням [7].

2 ОГЛЯД МЕТОДІВ І ТЕХНОЛОГІЙ, ЩО ЗАСТОСОВУЮТЬСЯ В ПРЕДМЕТНІЙ ОБЛАСТІ

2.1 Опис нейронних мереж

Одною із найбільш розвинених областей машинного навчання являються глибокі нейронні мережі. Цей розвиток вплинув на зростання все більшої кількості різних алгоритмів, що були побудовані саме на глибокому навчанні. Для того, щоб згенерувати підробку зображень дані алгоритми користуються нейромережами.

Спрощеною моделлю біологічної нейронної мережі, що представляє собою сукупність штучних нейронів, які взаємодіють між собою вважають штучну нейронну мережу (ШНМ). З точки зори машинного будування мережа, що розглядається являє собою окремий випадок розпізнавання образів, методів кластеризації, дискримінантного аналізу, тощо [8].

Нейронні мережі не програмуються у звичному розумінні цього слова, оскільки вони мають можливість навчатись. Завдяки цьому у них є одна із найважніших переваг перед традиційними та звичайними алгоритмами. Тобто дані системи повинні обучатися виконувати поставлені завдання, з кожним поступовим кроком покращувати продуктивність та точність на них під час розглядання певних прикладів не маючи у собі особливого програмування для вирішення цих завдань [8].

Приводячи приклад, нейронні мережі під час аналізу та розпізнавання зображень, що містять у собі кішок, мають змогу ідентифікувати їх та розподілити на 2 групи: на зображенні знаходиться «кішка» або «не кішка». Після цього отримані результати будуть використовуватися у подальших ідентифікаціях графічних даних. Таким чином вони вдосконалюють свій особистий комплект характеристик із навчальних даних та матеріалів, що оброблюються.

ШНМ складається з штучних нейронів, кожен з яких представляє собою спрощену модель біологічного нейрона. Штучний нейрон приймає сигнали з багатьох входів, обробляє їх єдиним чином і передає результат на багато інших штучних нейронів, тобто робить те ж саме, що й біологічний нейронів. Біологічні нейрони у свою чергу пов'язані між собою аксонами, місця стиків називаються синапсами. У синапсах відбувається посилення або ослаблення електрохімічного сигналу. Зв'язки між штучними нейронами називаються синаптичними, або

просто синапсами. У синапсу є один параметр - ваговий коефіцієнт, залежно від його значення відбувається ту чи іншу зміну інформації, коли вона передається від одного нейрона до іншого. Саме завдяки цьому вхідна інформація обробляється і перетворюється в результат, а навчання нейронної мережі засноване на експериментальному підборі такого вагового коефіцієнта для кожного синапсу, який і призводить до отримання необхідного результату [9].

На рисунку 2.1 зображена структура простої нейронної мережі.

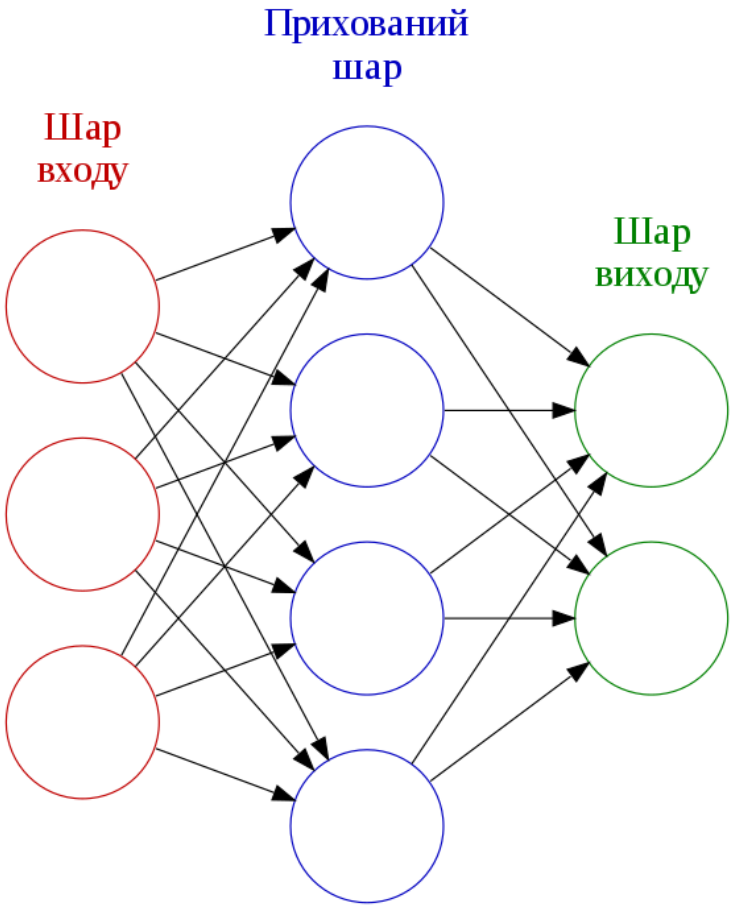

Рисунок 2.1 - Структура штучної простої нейронної мережі

На даному рисунку зображені кругові вузли, шо являють у собі штучні нейрони та стрілки, які показають з’єднання виходу штучного нейрона зі входом іншого. Червоним кольором позначені нейрони вхідного шару, синім - нейрони прихованого шару, а зеленим - шар виходу.

Нейрони вхідного шару отримують дані ззовні, наприклад, від сенсорів системи, що розпізнає обличчя, та після їх детальної обробки передає сигнали, використовуючи синапси до нейронів наступного шару.

Прихований шар отримав свою назву завдяки тому, що він не пов’язан безпосередньо ні з входом, ні з виходом штучної нейронної мережі. Нейрони

цього шару оброблюють отримані сигнали та відправляють їх до нейронів шару виходу. Через те, що мова йде про імітацію нейронів, тому кожен процесор вхідного рівня пов'язаний з декількома процесорами прихованого рівня, кожен з яких, в свою чергу, пов'язаний з декількома процесорами рівня виходу [9].

Ця система здатна до навчання та має змогу знаходити прості взаємозв'язки в даних. Але вона також може відшукувати взаємозв’язки один із одним, що створюються більш складну структуру. У неї можуть міститись більша кількість прихованих слоїв, що перемежовуються шарами, які виконують складні логічні перетворення. Саме такі мережі спроможні до глибокого навчання. В наслідок цього компанія Google отримала можливість значно покращити якість свого програмного продукту Google Translator.

Станом на 2021 рік існують вже більш 20 різних типів нейронних мереж. Для їх систематизації вони були класифіковані за [10]:

– вхідними даними;
– характером навчання;
– характером налаштування синапсів;
– часом передачі сигналу;
– характером зв’язків.

За вхідними даними вони поділяються на аналогові, двоїчні або образні мережі.

Навчання може виконуватись із вчителем, тобто коли вихідний простір рішень вже відомий для системи. Навчання без вчителя відбувається у тому разі, якщо простір рішень починає формуватися лише на основі впливів на вході. Також існує навчання з підкріпленням де використовується система заохочень та штрафів, отриманих у результаті взаємодії ШНМ із середою.

Класифікація за характером налаштування синапсів поділяє їх на фіксовані зв’язки та динамічні. У першому випадку вагові коефіцієнти обираються одразу, виходячи з поставленої задачі, а у другому у процесі навчання відбувається їх налаштування.

За часом передачі сигналу вони поділяються на синхронні та асинхронні мережі. А за характером зв’язків відносять такі типи, як мережі прямого поширення, у яких усі зв’язки мають чіткий напрямок від входу до виходу, рекурентні мережі, де сигнал нейронів виходу може частково відправитися знову на вхід та інші [10].

Зазвичай штучні нейронні мережі експлуатують для вирішення складних питань, які потребують аналітичних розрахунків. Найпоширенішими задачами є [10]:

– розпізнавання образів;
– кластеризація;
– класифікація;
– прийняття рішень та управління;
– прогнозування;
– апроксимація;
– стиснення даних і асоціативна пам'ять.

Також вони знайшли своє застосування в аналізі даних, вирішенню оптимізаційних задач, знаходження патернів у великих обсягах даних і т.д.

Нейронні мережі мають багато переваг перед традиційними обчислюваними методами, а саме [10]:

– рішення задач в умовах неоднозначності (завдяки можливості НМ до навчання вона дозволяє вирішувати задачі з невідомими закономірностями);

– стійкість до шумів вихідних даних (нейронна мережа має змогу виявляти непотрібні та неінформативні параметри самостійно, тому відпадає необхідність у попередньому аналізу даних);

– гнучкість структури нейронних мереж (нейрони та зв’язки між ними можна комбінувати різними методами;

– висока швидкодія (вхідні дані обробляються багатьма нейронами одночасно, завдяки чому нейронні мережі вирішують завдання швидше, ніж більшість інших алгоритмів);

– адаптація до змін навколишнього середовища (дані мережі, що навчаються за допомогою даних, здатні прилаштовуватися до певних змін навколишнього середовища).

Але у нейронних мереж є також свої недоліки [10]:

– відповіді, що видає ШНМ завжди є приблизними, оскільки вони не здатні давати чіткі та однозначні результати;
– нездатність вирішувати обчислювальні задачі;
– трудомісткість і тривалість навчання;
– нездатність прийняття рішень в кілька етапів.

2.2 Детальний опис компонентів та принципу роботи штучних нейронних мереж

Штучна нейронна мережа включає у свій склад прості елементи, які називаються нейронами. З точки зору математики, штучний нейрон відображують як нелінійну функцію від єдиного аргументу, а саме лінійної комбінації сукупності сигналів входу. Ця функція має назву активація. Після цього отриманий результат відправляється на вихід [11].

На рисунку 2.2 наведено схему штучного нейрону.

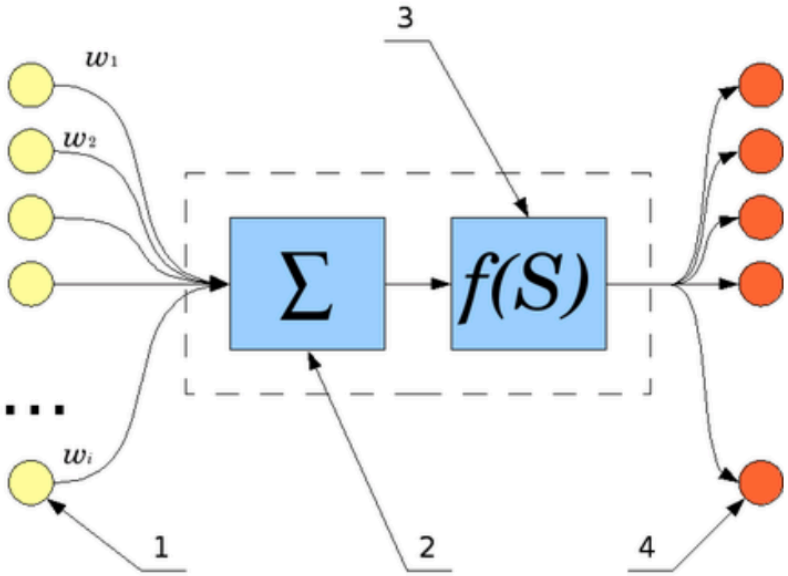

Рисунок 2.2 - Схема штучного нейрону

Цифрою 1 позначені нейрони, у яких вихідні сигнали поступають на вхід. Цифрою 2 відмічений суматор усіх вхідних сигналів. f(S) - являється обчислювачем функції, що передається. Цифрою 4 помічені нейрони, на входи яких подається сигнал певного нейрону. А ω_i представляють собою ваги вхідних сигналів [11].

Також до компонентів ШНМ відносяться:

– $a_i\,(t)$ - активація, що безпосередньо має залежність від параметру часу;

– θ_i - поріг, який зберігає константну поведінку, за умови, що він не буде видозмінений функцією навчання;

– f - функція активації, що розраховує нову активацію в установлений час;

– f_{out} - функція виходу, яка потрібна для підрахунку виходу з активації.

Нейрони, які подаються на вхід, використовуються у якості інтерфейсу входу для мережі та не мають своїх попередників. За аналогією, у нейронів виходу нема наступників.

Система сполучена з'єднаннями, які відправляють вихід нейрона i до входу нейрона j, де i являється попередником j, а j - наступником i. Всі сполучення мають свою вагу - w_{ij}.

До нейронних мереж також належить функція поширення, що розраховує вхід $p_j(t)$ до нейрону j з виходів $o_i(t)$ нейронів попередників. Ця функція виглядає таким чином:

$$p_j(t) = \sum_i o_i(t) w_{ij}. \tag{2.1}$$

Останньої складовою нейронної мережі є правило навчання. Воно виступає у якості алгоритму, котрий потрібен для зміни параметрів ШНМ, для того, щоб установлений вхід у мережу видав відповідний вихід. Як правило, навчання цілком залежить від порогів змінних та змінювання ваг у мережі.

Модель нейронної мережі можна записати у математичному вигляді та визначити функцією $f : X \to Y$. Інколи існують такі випадки, коли моделі близько поєднують до деякого правила навчання.

Визначаючи функцію нейронної мережі $f(x)$, у якості набору відмінних функцій $g_i(x)$, у подальшому їх можна розкласти на нові функції. Властивість цієї функції слугує зручним способом задля представлення мережевої структури, у якої стрілки представляють залежність між цими функціями. Найбільш використовуваним методом компонування представляє собою нелінійна зважена сума. ЇЇ вираз виглядає наступним чином:

$$f(x) = \text{K}(\textstyle\sum_i w_i g_i(x)), \tag{2.2}$$

де К - функція активації або функція, що визначена заздалегідь.

Одна із найважливіших характеристик функції 2.2 гарантує плавний перехід за умови, коли відбувається змінювання значень входу, а саме незначна зміна входу є причиною невеликого видозмінення на виході [11].

На рисунку 2.3 зображено граф залежностей штучної нейронної мережі.

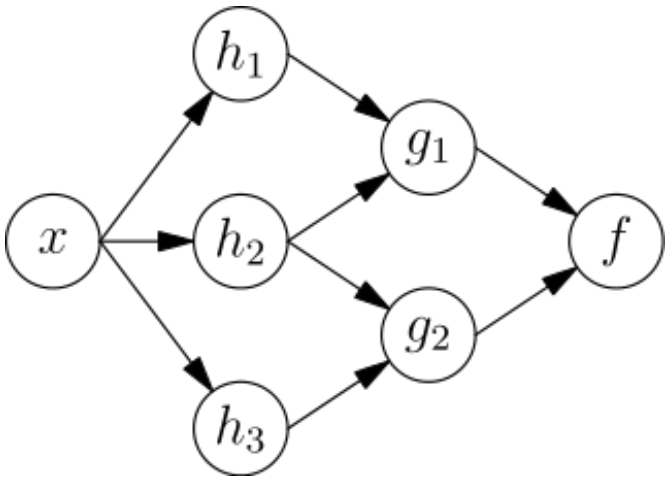

Рисунок 2.3 - Граф залежностей ШНМ

Набір функцій g_i розглядається як вектор $g = (g_1 + g_2, \ldots, g_n)$. На рисунку 2.3 схема демонструє розклад f із залежностями між змінними, показаними стрілками. Їх може бути інтерпретовано двома способами:

– функціональний;

– ймовірнісний.

Перший спосіб являється функціональним. Вхід, у якості змінної *x,* було перетворено на вектор *h,* що є тривимірним. Далі відбувається його перетворення на вектор *g,* який являється двомірним. Останньою ітерацією виступає перетворення цього вектору на *f.* Даний метод зазвичай використовується при оптимізації.

Іншим спосіб - ймовірнісний. Величина, можливими значеннями якої є результати випробувань чи спостережень явищ або процесів, що носять випадковий характер $F = f(G)$ повністю залежить від другої змінної $G = g(H)$, що у свою чергу має залежність від $H = h(X)$, а вона від величини X. Цей метод знайшов своє застосування у роботі з графовими моделями.

Розглянуті два способи є рівноцінними, тобто у будь-якому з випадків для певної архітектури складові різних шарів незалежні. Тому складові g не мають залежності між собою при установленому їх входу *h.* Це допомагає вдосконалити реалізацію паралелізму [12].

Навчання нейронних мереж є одною з найцікавіших та найбільш важливих з її характеристик. Як було розглянуто вище, навчання поділяють на три парадигми, в основі яких покладено особливості машинного навчання.

Навчання з вчителем передбачає, що для кожного вхідного вектору x_i існує такий вектор вихідних значень d_i. Сукупність цих векторів називається

навчальною парою (x_i, d_i), а множини даних пар - навчальною вибіркою. Процес навчання полягає у почерговому подавання на вхід ШНМ навчальних пар, розрахунок похибки між дійсними та бажаними значеннями та корегування параметрів мережі в сторону зменшення похибки.

Навчання без вчителя являється одним із способів машинного навчання, де випробувана система при вирішенні задач навчається виконувати задані завдання спонтанним чином. Експериментатор у цьому випадку не втручається в процес. Даний підхід використовується тільки для тих задач, де опис множини об'єктів або навчальна вибірка вже відома та є необхідність визначити внутрішні зв'язки, закономірності та залежності, що існують між ними. Цей метод зазвичай протиставлять навчанню з вчителем, у випадку коли для кожного компонента, що проходить навчання, примусово задається правильна відповідь, тому необхідно знайти якусь залежність між реакціями та стимулами мережі.

Навчання з підкріпленням виступає проміжним варіантом двох попередніх методів. Замість вчителя до схеми навчання включають блок під назвою «критика», який потрібен для відстеження реакції середовища на вхідний сигнал. Далі він спираючись на неї ідентифікує евристичну похибку, яку була влаштована в процес навчання системи. Ці парадигми базуються на відповідних правилах навчання, що визначають основні особливості їх застосування [13].

Теорія навчання розглядає три фундаментальні властивості, а саме:

– ємність;
– обчислювана складність;
– складність зразків.

Ємність включає у собі кількість зразків, що може запам'ятати мережа та які функції чи межі ухвалення рішень можуть бути сформульовані.

Складність зразків необхідна для визначення числа навчальних прикладів, які потребуються для досягнення здатності мережі до узагальнення. Мала кількість прикладів має змогу викликати перенавчення мережі, за умови відповідного функціонування на прикладах навчальної вибірки, але недоречно на тестових прикладах, що підлягають теж самому статистичному розподілу [13].

Існують чотири базові типи навчання:

– корекція за помилкою;
– дельта правило;
– навчання Больцмана;

– навчання методом змагання.

Правило корекції за помилкою використовується при навчанні із вчителем, де для кожного вхідного прикладу заданий бажаний вихід *d*. Дійсний вихід мережі *y* може не сходитись із бажаним. При цьому методі навчання вага зв'язку не змінюється до того моменту, коли поточна реакція перцептрона (математична або комп'ютерна модель сприйняття інформації) залишається правильною. При виникненні помилкової реакції вага змінюється на одиницю, а знак тим часом визначається протилежним від знаку помилки. Навчання відбувається лише тоді, коли помиляється перцептрон [13].

Дельта правило можна розглянути у вигляді математичного запису. Нехай вектор $X = x_1, x_2, \ldots . x_r, \ldots \ldots x_m$ - вектор вхідних сигналів, а вектор $D = d_1, d_2, \ldots . d_k, \ldots \ldots d_n$ - вектор сигналів, які повинні бути отримані вед перцептрона під впливом вхідного вектору. Вхідні сигнали, що поступають на входи перцептрона необхідно зважити та зробити підсумок. В результаті даних дій було отримано вектор $Y = y_1, y_2, \ldots . y_r, \ldots \ldots y_m$ вихідних значень. Після цього можна отримати вектор помилки $E = e_1, e_2, \ldots . e_r, \ldots \ldots e_m$. Його розмірність збігається з розмірність вектору вихідних сигналів. Компоненти вектору помилок визначаються як різниця між очікуваними та реальними значеннями вихідного сигналу нейрона:

$$E = D - Y \tag{2.3}$$

Формула для коригування *j*-ії ваги *i*-го нейрона можна представити наступним чином:

$$w_j(t+1) = w_j(t) + e_i x_i \, , \tag{2.4}$$

де t - номер поточної ітерації навчання;

w - вага.

Номер сигналу *j* змінюється в межах віл одиниці до розмірності вхідного вектору *m.* Номер нейрону *i* змінюється в межах від одиниці до кількості нейронів *n.* вага вхідного сигналу нейрона змінюється у бік зменшення помилки пропорційно величині сумарної помилки нейрона. Часто вводять коефіцієнт пропорційності η, на який множиться величина помилки. Цей коефіцієнт

називають швидкістю навчання. Отже, підсумкова формула для коректування ваг зв'язків перцептрону має наступний вигляд:

$$w_j(t+1) = w_j(t) + e_i \eta x_j \tag{2.5}$$

Правило навчання Больцмана є стохастичним правилом навчання, яке виходить з інформаційних теоретичних і термодинамічних принципів. Метою навчання Больцмана є таке налаштування вагових коефіцієнтів, при якому стани видимих нейронів задовольняють бажаний розподіл вірогідності. Навчання Больцмана може розглядатися як спеціальний випадок корекції за помилкою, в якому під помилкою розуміється розбіжність кореляцій полягань в двох режимах [13].

Основні ознаки навчання:

– використовують каскад багатьох шарів вузлів нелінійної обробки для виділення ознак та подальших перетворень. Кожен наступний шар використовує вихід із попереднього як вхід. Алгоритми можуть бути з керованим або спонтанними навчаннями, а застосування включають у собі розпізнавання образів та класифікацію;

– ґрунтуються на спонтанному навчанні декількох шарів ознак або представлень даних;

– є частиною ширшої області машинного навчання представлень даних;

– навчаються кількома рівням представлень, що відповідають різним рівням абстракції. Дані рівні формують ієрархію понять.

При навчанні змагання вихідні нейрони змагаються між собою за активізацію. Це навчання має місце в біологічних нейронних мережах. Навчання за допомогою змагання дозволяє кластеризувати вхідні дані: подібні приклади групуються мережею відповідно до кореляцій і подаються одним елементом. При навчанні модифікуються тільки ваги нейрона, що «переміг». Ефект цього правила досягається за рахунок такої зміни збереженого в мережі зразка (вектору вагів зв’язків нейрона, що «переміг»), при якому він стає трохи ближчим до вхідного прикладу [13].

2.3 Опис згорткових нейронних мереж

На даний момент такий процес як аналіз зображень являється одним із тих, що потребує витрачати дуже велику кількість операцій для вирішення поставленої задачі. Оскільки при подаванні, навіть невеликого розміром зображення, необхідна така нейронна мережа, на вході якої були б потрібні сотні тисяч нейронів. Задля вирішення даної проблеми створено особливу архітектуру НМ, що працюють із зображеннями, а саме - згорткову нейронну мережу. Але у той самий час вона також дала можливість для підроблювання зображень.

Згорткова нейронна мережа (ЗНМ) являється класом глибинних ШНМ прямого поширення. Цей інструмент використовують для аналізу зображень, тобто для класифікації та розпізнавання об’єктів або обличчя на фото й для розпізнавання мови. Ця мережа використовує різновид багатошарових перцептронів, які розроблені таким чином, щоб використовувалась мінімальна обробка за вимогою [14].

За рахунок застосування спеціальної операції згортки ЗНМ дозволяє одночасно виконати зменшення кількості інформації, котра зберігається в пам'яті інформації. Це дає змогу набагато краще впоратись із фотографіями, що мають достатньо високу роздільну здатність. Також вона здатна виділити опорні ознаки зображення, такі як грані, ребра чи контури.

Наступний рівень обробки проаналізованих граней та ребер може виявити елементи текстур, які вже повторювались. А далі ці фрагменти можуть скластись у повне та цільне зображення.

ЗНМ використовують невелику кількість попередньої обробки, порівнюючи з іншими існуючими алгоритмами для класифікації зображень. Тому це означає, що вона навчається використовуючи допоміжні фільтри. А в традиційних алгоритмах необхідно виконувати дані операції вручну. Незалежність ЗНМ у конструюванні ознак та роботи без людських зусиль дає їм велику перевагу серед аналогів [14].

Фактично кожен шар нейронної мережі використовує власне перетворення. На перших шарах мережа працює з такими елементами, як ребра або грані, то на подальших рівнях вже використовуються такі речі, як текстура або певні частини об’єктів.

В результаті аналізу опрацювання можна коректно класифікувати картинку або виділити необхідний об’єкт зображення на кінцевому етапі.

Згорткові нейронні мережі складаються з трьох шарів:

– шари входу;

– шари виходу;

– приховані шари.

Зазвичай приховані шари містять у собі шари згортки, нормалізації та агрегувальні шари. Згорткові шари застосовують до входу операцію згортки, передаючи результат до наступного шару. Згортка імітує реакцію окремого нейрону на зоровий стимул.

Застосування згорткових мереж знайшли своє призначення у багатьох та різних сферах життя. Одними із перших та найбільш тривіальних задач, що були вирішені завдяки нейронним мережам стали класифікації зображень [15]. На рисунку 2.4 зображена різноманітна вибірка картинок за певними класами.

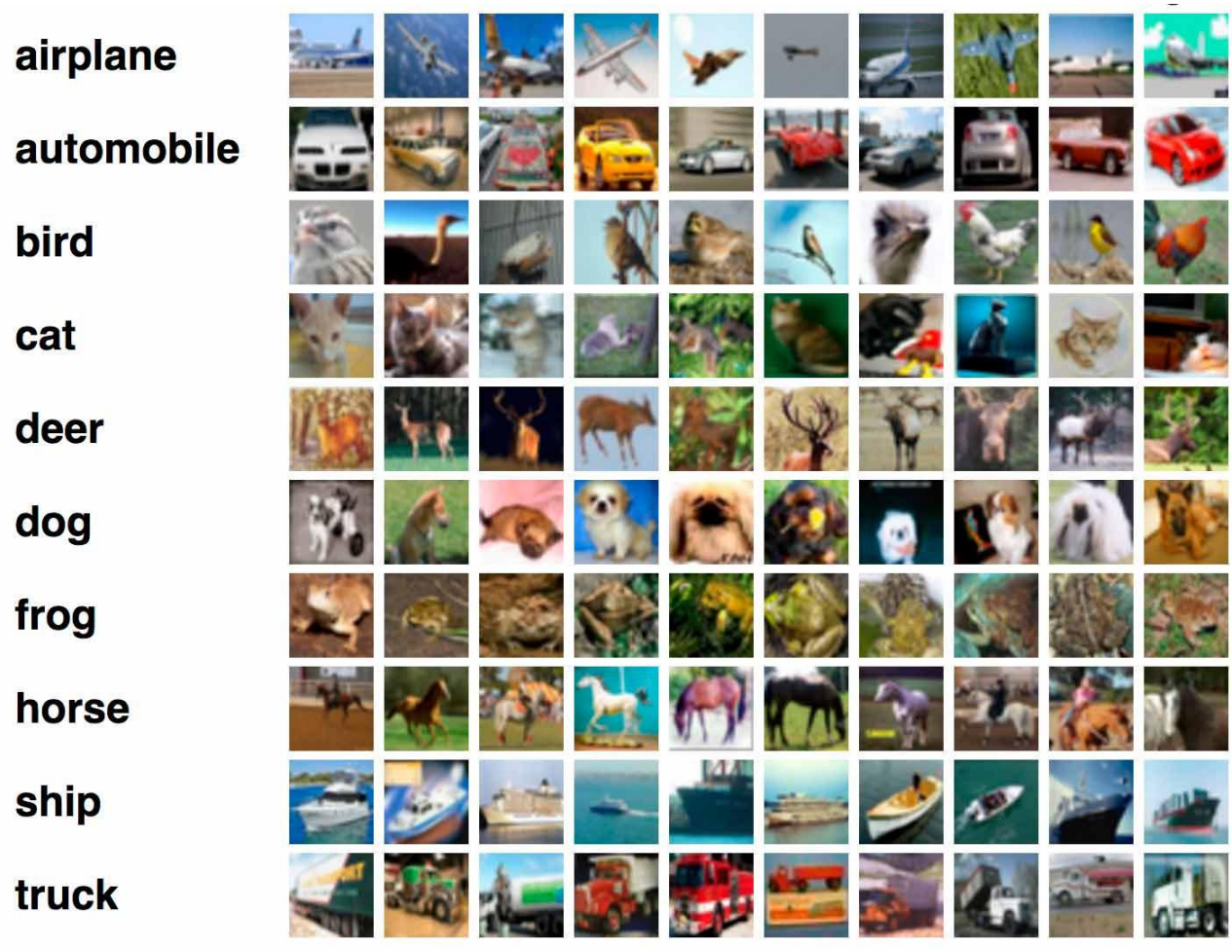

Рисунок 2.4 - Приклад класифікації зображень

Вищезгадана здатність ЗНМ також знайшла активне застосування у медицині, оскільки нейронну мережу можна навчити класифікувати різні

симптоми чи хвороби. Також методику розпізнавання зображень використовують в агробізнесі для того, щоб мати можливість прогнозувати очікуваного врожаю конкретних земель.

Із швидким розвитком безпілотного транспорту система розпізнавання об’єктів стала її невід’ємною частиною, що допомагає ідентифікувати знаки дорожнього руху, потенційну небезпеку, тощо. На рисунку 2.5 зображено систему, що використовує згорткові нейронні мережі, які розпізнають об’єкти дорожнього руху.

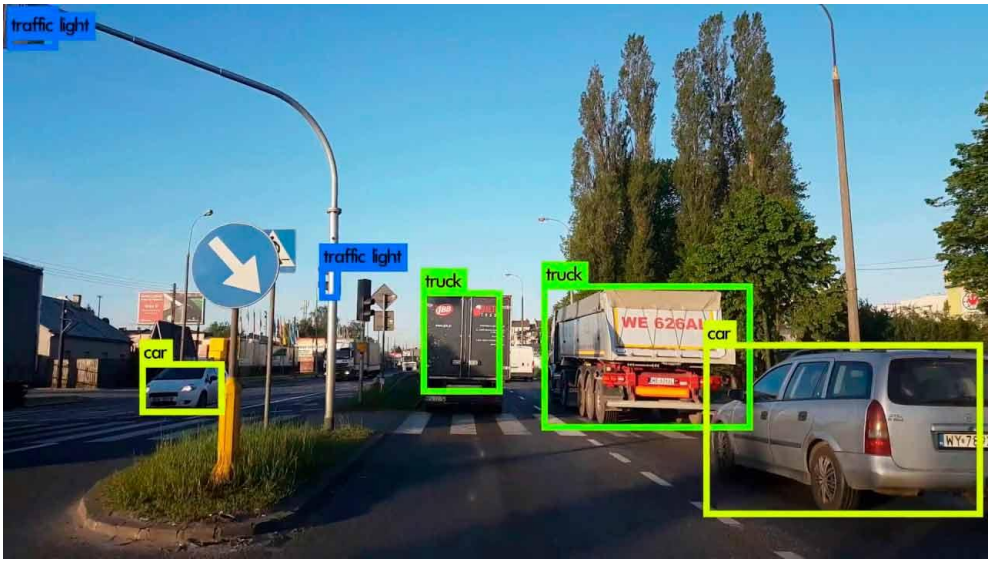

Рисунок 2.5 - Розпізнавання об’єктів дорожнього руху

Подібним чином високою популярністю нейронні мережі стали користуватися для розпізнавання обличчя та людей. Спочатку виділяється обличчя на зображенні, а потім за рахунок використання ЗНМ можна розпізнати обличчя конкретної людини. Цією технологією користується значна кількість просунутих країн для забезпечення небезпеки та знаходження злочинців та інших порушників закону [15].

2.4 Принцип роботи згорткових нейронних мереж

Під згорткою можна мати на увазі упорядковану процедуру змішування двох джерел інформації. У випадку, коли згортка застосовується до зображень, вона діє у двох вимірах, а саме по ширині та висоті та також ділиться на дві складові. Перша - це вхідне зображення з трьома матрицями пікселей, по одній на червоний, синій та зелений колір. Піксель у свою чергу складається з цілого

чисельного значення від 0 до 255 для кожного колірного каналу. Друга складова - це ядро згортки або матриця дійсних чисел, властивості якої можна розглядати як правило «перемішування» вхідного зображення та ядра. У результаті чого отримується змінене зображення, яке зазвичай називається як карта ознак. Для кожного кольорового каналу належить по одній карті [16].

На рисунку 2.6 продемонстровано приклад вхідного зображення, ядра згортки та отриманої карти ознак.

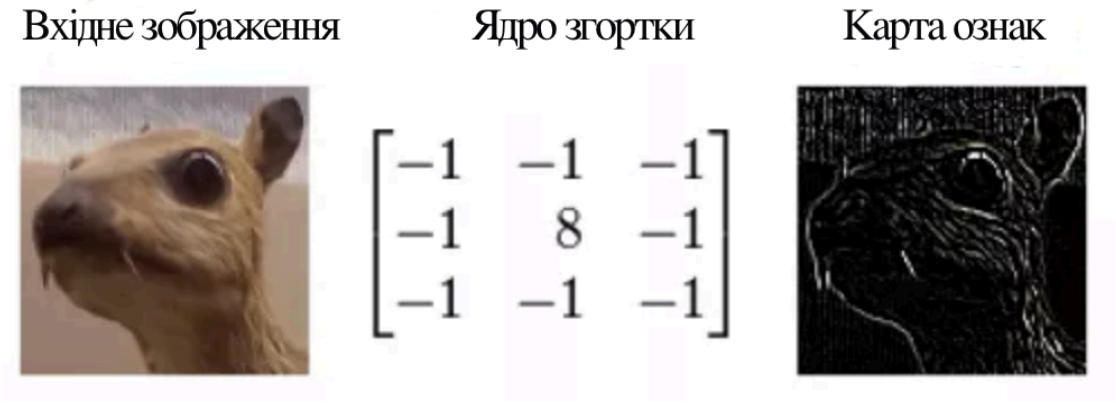

Рисунок 2.6 - Приклад складових операції згортки

Сама двовимірна згортка є нескладною операцією та працює наступним чином: фільтр, що являю собою колекцію ядер, переміщується уздовж зображення та визначає на присутність деяку шукану характеристику в певній його частині. Для того, щоб отримати дану відповідь необхідно здійснити операцію згортки, що є сумою добутку елементів фільтра та матриці вхідних сигналів [16].

При цьому також може виконуватись процедура нормалізації, коли вихідне значення нормалізується за розміром ядра, щоб загальна інтенсивність зображення та карти ознак залишалася незмінною.

Ядро, повторюючи описану вище операцію для кожної позиції, що він займає, перетворює початкову двовимірну матрицю характеристик у дещо іншу матрицю, яка також є двовимірною. Характеристики, що були отримані є сумою зважених вхідних параметрів. Вагами називають елементи матриці ядра [16].

Для побудови згорткових слоїв використовують дві базові техніки:

– паддінг;

– страйд.

Іноді виникає така проблема, коли в процесі переміщення ядра краї початкової матриці не потрапляють у центр ядра. Це може призвести до помилки, оскільки розміри матриць повинні бути однаковими. Але даний недолік можна вирішити за допомогою паддінгу, що надає додатковий простір («хибні пікселі, значення яких дорівнює нулю»). Тому завдяки цієї можливості ядро має змогу обійти пікселі, що знаходяться з краю. При цьому розміри обох матриць збігаються [16].

Другий спосіб застосовується при виникненні потреби зменшити розмірність матриці вхідних сигналів. Це є основою ЗНМ, де імпульси, які поступають до системи, зменшуються зі швидкістю прямо пропорційно кількості каналів. Для вирішення даної задачі можна використовувати пулінговий шар, який може відсіювати певне число імпульсів. Інший спосіб полягає у використанні страйду. Основний принцип його роботи полягає у пропуску ядром деякого числа кроків під час його переміщення. В залежності від значення страйду (наприклад: значення 2) розмір отриманої матриці зменшиться приблизно в два рази [16].

Але вищерозглянуті підходи застосовуються лише для зображень з одним вхідним каналом, оскільки зараз переважна більшість зображень мають три канали (червоний, синій та зелений). Більш глибинні нейронні мережі здатні обробити більш велику кількість каналів [15].

На рисунку 2.7 наведено зображення з трьома каналами RGB.

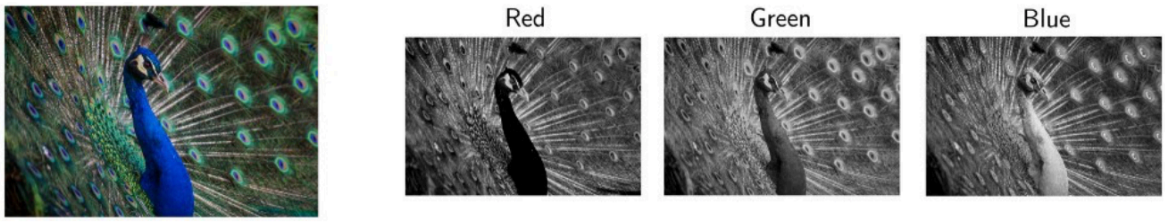

Рисунок 2.7 - Зображення з RGB каналами

У поліканальній версії нейронної мережі поняття «фільтр» та «ядро» набувають інше значення. Фільтр у даному випадку виступає в якості колекції ядер. На кожен окремий канал припадає по одному ядру, де вони є повністю унікальними. Також фільтр породжує лише єдиний вихідний канал.

Фільтр працює наступним чином: ядра переміщуються вздовж відповідних їм каналам, генеруючи їх оброблену версію.

Варто зазначити, що бувають випадки, коли ядра мають більшу вагу, ніж інші. Тоді нейронна мережа реагуватиме на певний колір більш виражено [17].

На рисунку 2.8 зображено приклад роботи згортки зображення з трьома каналами.

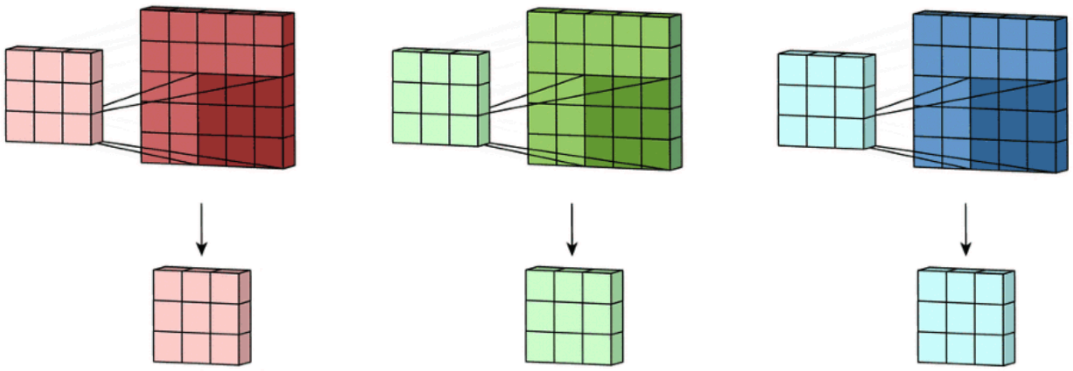

Рисунок 2.8 - Приклад роботи згортки зображення з трьома каналами

Після цього отримані результати роботи виконують операцію суми для того, щоб отримати єдиний канал. Ядра створюють власні версії каналів, а це випливає в результуючий вихідний канал. На рисунку 2.9 відображено отримання нового єдиного каналу.

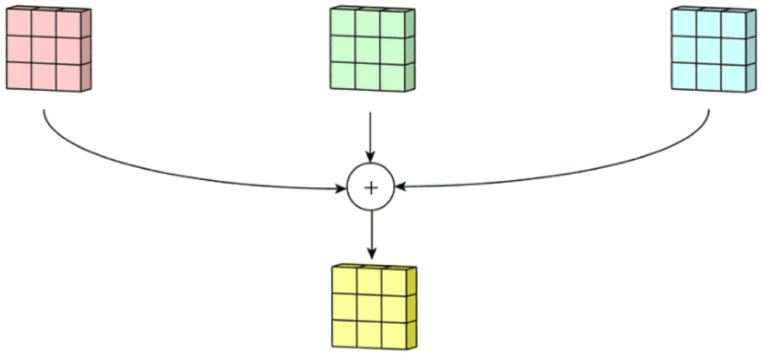

Рисунок 2.9 - Процес отримання єдиного каналу

Для того, щоб завершити процес необхідно до результату, що було отримано, використати прийом зміщення. У ЗНМ зміщення застосовується до кожного вихідного фільтру. Ця операція являється простою та містить у собі додавання зміщення з кожним із сигналів отриманого результату каналу.

Розглядувана послідовність дій являється повністю подібною до будь-якої кількості фільтрів, тобто кожен фільтр повинен виконувати незалежну обробку зображення, що подається на вхід. Також він застосовує власну систему ядер та зміщення, створюючи деяку відповідь, унаслідок чого вони додаються та дають вже глобальну відповідь. Потрібно зауважити, що кількість каналів даної відповіді еквівалентно кількості фільтрів [17].

Далі отриманий результат має можливість бути пропущеним через нелінійний перетворювач та переданим до наступного згорткового шару для того, щоб процес повторився.

Структура роботи згорткових шарів не є складною, оскільки як і у простих нейронних мережах використовуються лінійні перетворення, однак вони все ж таки мають свої відмінності.

Розглянемо матрицю, що подається на вхід розміром 4×4 та завдання, коли потрібно її перетворити на іншу, що має розмір 2×2. У випадку використання НМ, яка має прямі зв’язки було б необхідно зробити розгортання вхідною матриці до вектору, довжина якого становила б 16 одиниць (сигналів). Наступним кроком являвся б його пропуск крізь систему з входами та виходами кількістю в 16 та 4 відповідно [17].

На рисунку 2.10 зображена вагова матриця між шарами нейронної мережі, що була описана вище.

$$\begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} & w_{1,6} & w_{1,7} & w_{1,8} & w_{1,9} & w_{1,10} & w_{1,11} & w_{1,12} & w_{1,13} & w_{1,14} & w_{1,15} & w_{1,16} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} & w_{2,5} & w_{2,6} & w_{2,7} & w_{2,8} & w_{2,9} & w_{2,10} & w_{2,11} & w_{2,12} & w_{2,13} & w_{2,14} & w_{2,15} & w_{2,16} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} & w_{3,5} & w_{3,6} & w_{3,7} & w_{3,8} & w_{3,9} & w_{3,10} & w_{3,11} & w_{3,12} & w_{3,13} & w_{3,14} & w_{3,15} & w_{3,16} \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} & w_{4,5} & w_{4,6} & w_{4,7} & w_{4,8} & w_{4,9} & w_{4,10} & w_{4,11} & w_{4,12} & w_{4,13} & w_{4,14} & w_{4,15} & w_{4,16} \end{bmatrix}$$

Рисунок 2.10 - Вагова матриця між шарами входу та виходу

Але для вирішення даного завдання можна використати деяке ядро k, що має розмір 3×3. Після його використання вагова матриця перетвориться до наступного вигляду, що відображено на рисунку 2.11.

$$\begin{bmatrix} k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 & 0 & 0 & 0 & 0 \\ 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} & 0 \\ 0 & 0 & 0 & 0 & 0 & k_{1,1} & k_{1,2} & k_{1,3} & 0 & k_{2,1} & k_{2,2} & k_{2,3} & 0 & k_{3,1} & k_{3,2} & k_{3,3} \end{bmatrix}$$

Рисунок 2.11 - Матриця сформована переміщення ядра

Незважаючи на те, що розмірність даних матриць є однаковою, створені вони все ж таки за допомогою зовсім різних методів. Характерною рисою у першому прикладі є множинні зв’язки між однорідними шарами. У другому прикладі - робота ядра в двомірному просторі. Тому для матриці, яка має у собі 64 елементів міститься лише 9 параметрів, що будуть згодом повторно використовуватися [17]. Як можна побачити на рисунку 2.18 множинні зв’язки зруйновані нульовими вагами.

2.5 Автокодувальник

Автокодувальник (Autoencoder) являється важливою спеціальною архітектурою ШНМ, яка дозволяє застосовувати навчання без вчителя при використанні зворотного виникнення помилки. Звичайна та проста архітектура автокодувальника представляє собою мережу прямого поширення, що не має зворотних зв’язків. Він своєю структурою має подібні риси перцептрона та також містить у собі вхідний, проміжний та вихідний шар. Але відрізняється від перцептрона тим, що вихідний шар автокодувальника повинен містити ідентичну кількість нейронів, що й вхідний шар [18].

Основним принципом роботи та навчанням мережі автокодувальника полягає в стисненні початкових вхідних даних для того, щоб відобразити в прихованому просторі, а наступним кроком виконати відновлення даних із цього простору в якості вихідних даних. Метою цих перевтілень є отримання відгуку на шарі виходу, що являється найбільш схожим та близьким до вхідного шару [18].

На рисунку 2.12 представлено архітектуру автокодувальника та принцип його роботи.

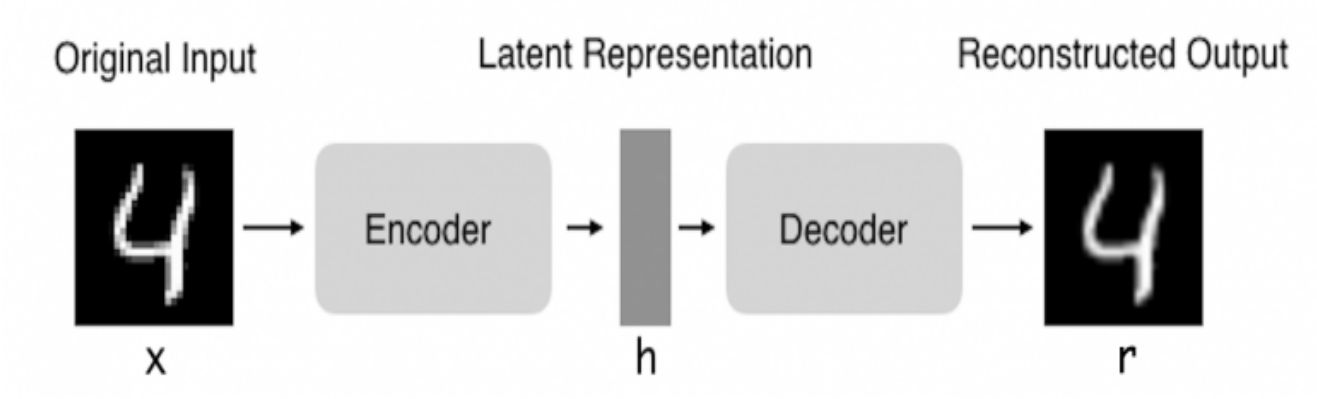

Рисунок 2.12- Архітектура та принцип роботи автокодувальника

Автокодувальник поділяється на 2 складові, а саме:

– кодувальник;

– декодувальник.

Кодувальник потрібен для того, щоб стискати вхідні дані, які потім відправляються в приховане середовище. Він представлений функцією кодування:

$$h = f(X) \tag{2.6}$$

Декодувальник необхідно використовувати для того, щоб відновити з прихованого простору вже вихідні дані. Його представленням є аналогічна функція декодування.

Автокодувальник у цілому можна описати функцією:

$$g(f(x)) = r \tag{2.7}$$

де значення r повинне співпадати із значенням x, що подається на вході.

Для недопущення отримання тривіального результату, на проміжний шар накладаються такі обмеження як:

– розмір проміжного шару повинен бути меншим ніж розмір вхідних та вихідних шарів;

– штучне обмеження кількості одночасно активних нейронів, що називається розрідженою активацією.

Завдяки цим обмеженням штучна мережа починає пошук узагальнення та кореляції в даних, що поступають на вхід та виконувати їх стиснення. Тобто НМ

автоматично навчається виділяти загальні ознаки з вхідних даних, які потім кодуються в значеннях ваг штучної нейронної мережі. Тому, наприклад, при навчанні цієї мережі на наборі вхідних зображень, нейронна мережа має можливість власноруч навчитися розпізнавати елементи під різними кутами [18].

На рисунку 2.13 зображено приклад результату роботи автокодувальника.

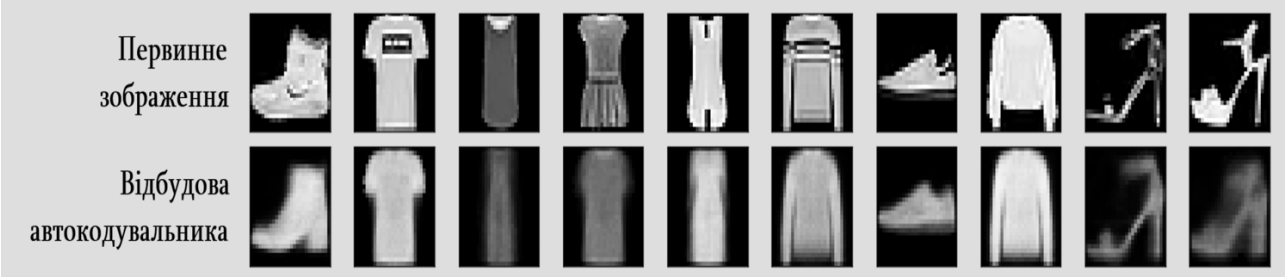

Рисунок 2.13 - Результати роботи автокодувальника

Основними застосуваннями автокодувальників є зниження розмірності та інформаційний пошук. Але також вони знайшли своє застосування у таких задачах як:

– виявлення аномалій;
– обробка зображень;
– пошук ліків;
– передбачення популярності;
– машинний переклад.

2.6 Опис генеративно-змагальних мереж

Генеративно-змагальна мережа (ГЗМ) представляє собою один із алгоритмів машинного навчання, що побудований на комбінації, який містить у собі дві нейронні мережі. Одна з них являється генеративною моделлю або генератором (мережа G), яка потрібна для генерації зразків заданої категорії. Інша називається дискримінантною моделлю або дискримінатором (мережа D), що намагається розпізнати створений зразок [19].

Оскільки мережі G та D мають зовсім протилежну мету та задачі, що потрібно вирішувати, а саме створювати зразки та відповідно їх відсіювати, між ними починає відбуватися антагоністична гра. Наприклад, генератор може створювати зображення, на яких відтворено схожі на людські обличчя. У свою

чергу дискримінатор прагне визначите чи являється на зображенні справжнє обличчя [19].

Експерти відзначають, що процес навчання відбувається без вчителя та є досить швидким, тому через це генератор починає відтворювати обличчя, які важко відрізнити від реальних людей. Директор Facebook з досліджень штучного інтелекту Ян Лекун оцінив змагальне тренування мереж як найперспективніший вектор розвитку машинного навчання.

Мережа дискримінаторів є стандартною згортковою мережею, що класифікує зображення, які подаються до неї, використовуючи біномний класифікатор, який потрібен для розпізнавання зображення на реальні або підроблені. Генератор в певному сенсі являється зворотною мережею згортки.

Незважаючи на те, що стандартний згортковий класифікатор приймає зображення і зменшує його роздільну здатність, для того, щоб отримати ймовірність, генератор у свою чергу приймає вектор випадкового шуму та відповідно перетворює його в зображення. Перший відсіює дані за допомогою методів зниження дискретизації, таких як maxpooling, а другий генерує нові дані [19].

На рисунку 2.14 представлено оригінальну архітектуру ГЗМ.

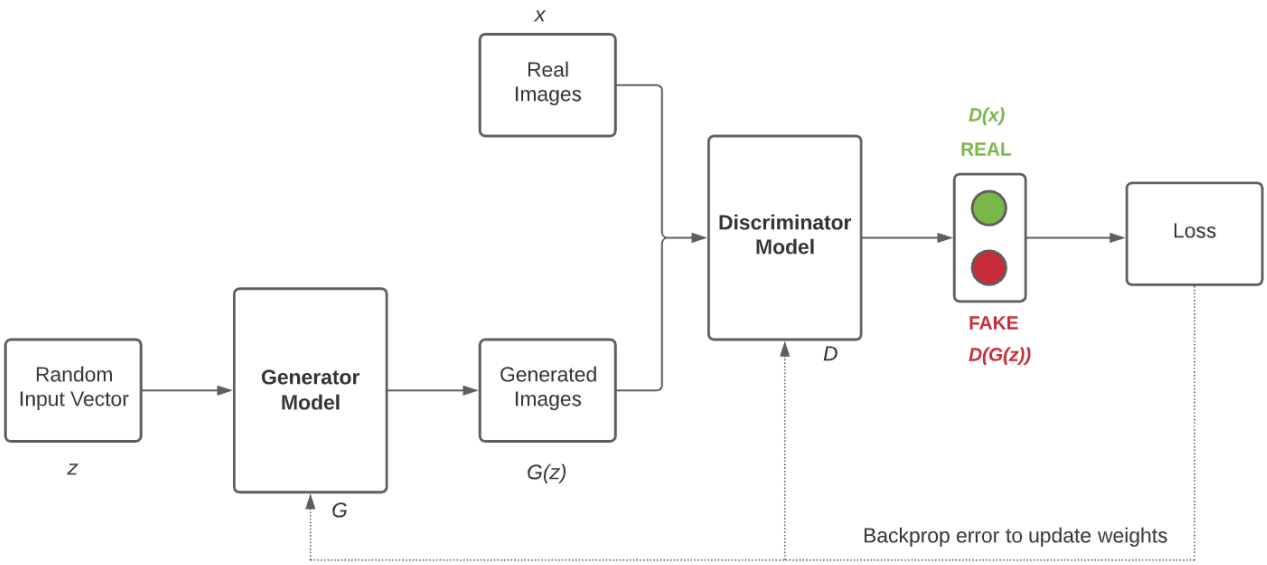

Рисунок 2.14 - Оригінальна архітектура ГЗМ

Архітектура генеративно-змагальної мережі складається з таких компонентів як [18]:

– G - генератор;

– D - дискримінатор;

– z - випадковий шум (вхідний вектор фіксованого розміру);

– x - реальне зображення;

– $G(z)$ - зображення, створене генератором (підробне зображення);

– $\mathrm{D}(\mathrm{G}(\mathrm{z}))$ - виведення дискримінатору, коли на вході подане підроблене зображення;

– $\mathrm{D}(\mathrm{x})$ - виведення дискримінатору, коли на вході подане реальне зображення.

Дискримінатор прагне максимізувати функцію втрат $V(D,G)$ правильно класифікуючи справжні та підроблені зображення. Тобто він намагається зробити таку умову, коли $\mathrm{D}(\mathrm{x})$ якомога ближче до 1, а саме правильно класифікувати реальні зображення як реальні. Інша умовою для нього буде прагнення зробити $\mathrm{D}(\mathrm{G}(\mathrm{z}))$ якомога ближче до 0, відповідно правильно класифікувати зображення як фальшиві. Таким чином дискримінатор намагається максимізувати обидві умови [19].

Зі свого боку генератор хоче мінімізувати функцію втрат $V(D,G)$, щоб обдурити дискримінатор. Тому він прагне зробити $\mathrm{D}(\mathrm{G}(\mathrm{z}))$ ближче до 1, після чого дискримінатор зазнає невдачі та неправильно класифікує отримані дані.

Модель генератора G приймає випадковий вхідний вектор z у якості вхідного сигналу та далі генерує зображення $G(z)$. Отримані зображення разом із реальними зображеннями x з навчальних даних потім подаються на модель дискримінатор D. Вона виконує класифікуванння зображень на їх справжність.

Наступним кроком необхідно виміряти втрати та вони повинні бути передані назад для того, щоб оновити ваги генератора та дискримінатора.

Під час навчання дискримінатору потрібно заморозити генератор та поширити помилки назад. Це необхідно робити, щоб оновлювався лише дискримінатор. Під час навчання генератору відбуваються аналогічно-протилежні дії. Таким чином обидві моделі стають з кожною новою ітерацією дедалі кращими [20].

Навчання потрібно припинити за умови досягнення рівноваги Неша або коли $D(x) = 0.5$ для всіх значень x. Тобто, коли згенеровані зображення мають вигляд майже реальних.

Але під час навчання генеративно-змагальної мережі можуть виникнути наступні проблеми [20]:

– колапс мод (генератор виходить із ладу, тобто видає обмежену кількість різних зразків);

– проблема стабільності навчання (параметри моделі дестабілізуються та не сходяться);

– зникаючий градієнт (дискримінатор стає надто сильним, а градієнт генератора зникає і навчання не відбувається);

– проблема заплутування (виявлення кореляції в ознаках, які не пов'язані або слабко пов'язані у реальному світі);

– висока чутливість до гіперпараметрів.

Не дивлячись на те, що не існує універсального підходу для вирішення більшості проблем є деякі допоміжні властивості, які допомагають при навчання ГЗМ, а саме [20]:

– нормалізація даних;

– семплювання із багатовимірного нормального розподілу замість рівномірного;

– використовувати нормалізаційні шари в G, D;

– використовувати мітки для даних, якщо вони є, тобто навчати дискримінатор ще й класифікувати зразки.

На теперішній час колапс мод являється одною з головних перешкод у навчання ГЗМ. Потенційно можливі вирішення цієї проблеми є використання метрики Вассерштейна всередині функції помилки. Тим самим це дозволяє дискримінатору швидше навчатися виявляти виходи, що повторюються, на яких стабілізується генератор. Або можна скористуватися розгорнутою ГЗМ. При її використанні для генератора використовується функція втрат, яка не тільки як поточний дискримінатор оцінює виходи генератора, а й від виходів майбутніх версій дискримінатора [20].

За декілька років звичайна генеративна-змагальна мережа еволюціонувала та набула покращень. Зараз існують такі види, як [20]:

– глибока згорткова мережа (DCGAN);

– умовна мережа (CGAN);

– інформаційна мережа (InfoGAN);

– мережа Вассерштейна;

– мережа граничної рівноваги (BEGAN);

– прогресуюча мережа (ProGAN);

– циклічна мережа (CycleGAN).

Першою та головною покращеною версією звичайної ГЗМ стала мережа DCGAN. Вона являється більш стійкою для навчання та генерує високі за якістю результати. Наступним кроком, щоб дозволити моделі краще генерувати зображення з високою роздільною здатністю, що є одним із слабких місць ГЗМ, автори цієї мережі запропонували наступні техніки:

– зіставлення об'єктів;
– усереднення;
– одностороннє згладжування міток;
– віртуальна пакетна нормалізація.

Тобто використання покращеної версії DCGAN необхідно для того, щоб отримати зображення високої якості.

Умовна мережа CGAN використовує додаткову інформацію, наприклад, результат у вищій якості, а також є можливість контролювати певною мірою як виглядатимуть згенеровані зображення.

Умовна CGAN є розширенням GAN. У нас є умовна інформація Y, яка описує деякі аспекти інформації. Наприклад, якщо ми маємо справу з обличчями, Y може описати опис таких атрибутів, як стать людини або колір волосся. Потім цю інформацію поміщають і в генератор, і в дискримінатор [21].

На рисунку 2.15 представлена схема умовної мережі CGAN з інформацією обличчя людини.

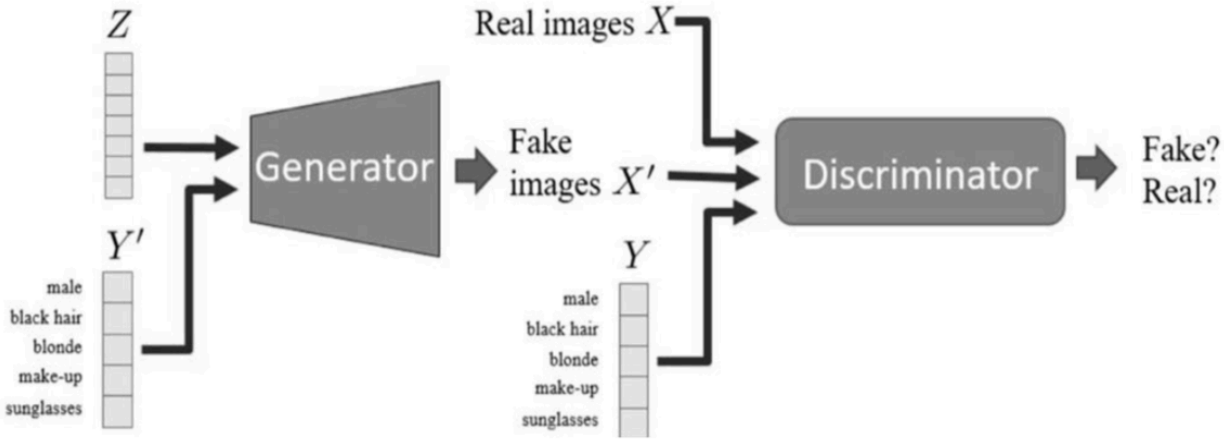

Рисунок 2.15 - Схема умовної мережі CGAN з інформацією про обличчя

Ця мережа має декілька важливих та корисних властивостей, а саме:

– по мірі навчання та задавання інформації моделі, ГЗМ вчиться використовувати її та навіть має можливість генерувати кращі результати;

– є декілька способі для управління представлень зображень.

Інформаційна мережа здатна кодувати значущі об'єкти зображення у частині вектора шуму Z неконтрольованим чином, наприклад, кодувати поворот числа. Але використовувати цю мережу рекомендується, коли набір даних є не дуже складним [20].

У ГЗМ існує проблема зі сходженням, у результаті чого, невідомо коли треба завершити навчання. Функція втрат не корелює з якістю зображення, що є значною перешкодою, бо потрібно безперервно дивитись на на зразки, щоб сказати, чи правильно ви навчаєте мережу чи ні. Також нема числового значення, яке може вказати наскільки коректно налаштовані параметри. Мережа Вассерштейна вирішує дані проблеми, оскільки має у собі функцію втрат [21].

Отримані результати показують, що чим менша втрата, тим краща якість зображення та доводять, що мережа Вассерштайна має функцію втрат, яка корелює з якістю зображення та забезпечує сходження. Тому використання цієї мережі гарантує високу стабільність навчання.

Метод, який використовується у мережах BEGAN врівноважує генератор та дискримінатор під час навчання. Крім того, він забезпечує нову приблизну міру збіжності, швидке та стабільне навчання та високу візуальну якість [21].

На рисунку 2.16 надано приклад згенерованих облич людей за допомогою використання BEGAN.

Рисунок 2.16 - Згенеровані обличчя людей за допомогою BEGAN

Виходячи з отриманих результатів, можна впевнено сказати, що ця мережа демонструє високу реалістичність при генерації медіа-контенту, а також має стабільне навчання та досить просту архітектуру.

Створення зображень високої роздільної здатності є великою проблемою. Чим більше зображення, тим легше мережі можна зробити помилку, тому що їй потрібно вчитися генерувати складніші та делікатні деталі.

Прогресуюча мережа, яка була створена на основі покращеної мережі Вассерштейна, пропонує розумний шлях, поступово додаючи нові шари високої роздільної здатності під час навчання, що створює неймовірно реалістичні зображення. Кожен із цих шарів збільшує роздільну здатність зображень як для дискримінатора, так і для генератора [21].

На першому етапі виконується навчання генератора та дискримінатору зображень, які мають низьку роздільну здатність. Під час моменту сходження потрібно підвищити роздільну здатність за допомогою згладжування.

Внаслідок цього прогресивного навчання генеровані зображення в ProGAN мають вищу якість і час навчання знижується. Причина цього полягає в тому, що цій мережі не потрібно вивчати всі великомасштабні та дрібномасштабні уявлення відразу. У прогресуючій мережі спочатку вивчаються дрібномасштабні шари, тобто коли шари з низькою роздільною здатністю сходяться, а потім модель може зосередитися на очищенні великомасштабних структур.

Незважаючи на те, що для навчання прогресуючої мережі потрібно багату часу, можна отримати зображення, які будуть набагато краще, ніж реальні.

На поточний час циклічна мережа є найбільш передовою ГЗМ для генерації зображень. Ці мережі не вимагають парних наборів даних для навчання перекладу між доменами, що добре, тому що такі дані дуже важко отримати. Тим не менш, циклічні мережі все ще повинні бути навчені з даними двох різних доменів X і Y (наприклад, X: коні, Y: зебри). Щоб обмежити переклад з одного домену до іншого, вони використовують те, що називається «послідовна втрата циклу» [21].

Всі мережі, які були описані, є розширенням стандартної ГЗМ та безперечно кожна з них має свої переваги та недоліки, тому потрібно використовувати їх відповідно до поставлених задач, щоб користуватися їх сильними сторонами.

2.7 Постановка задачі

Після аналізу та визначення актуальності реальної проблеми обраною задачею, яку необхідно розв’язати, є виявлення підроблених зображень та відео. Для вирішення цієї задачі необхідно виконати дослідження принципу роботи DeepFake, існуючих аналогів та розробити метод ідентифікації медіа-матеріалів як оригіналу або як підробки.

У якості початкових даних виступають зображення та відео надані для ідентифікації.

Зображення мають відповідати наступним вимогам:

– дозволеними форматами зображень являються JPG або PNG, оскільки вони є найбільш популярними та гнучкими, бо стискають дані без втрат;

– мінімально-дозволеним розміром зображень являється 160 × 160 пікселів (для кращої якості бажано ще вище);

– на зображенні повинне бути чітко розміщено обличчя людини, щоб отримати точні та коректні результати після обробки.

Відео-файли мають відповідати таким вимогам:

– дозволеними форматами відео є MPEG, MOV та AVI, оскільки вони підтримується сучасними відеокодеками та являються найпоширенішими форматами запису більшості камер;

– роздільна здатність відео повинна бути не менш ніж 1280 × 720 пікселів;

– на відео повинне бути чітко зображено обличчя людини на середній або близький відстані для отримання найбільш точних та правильних результатів.

Необхідно перевірити, чи надане зображення є дійсним, або ж воно є підробленим. Для того, щоб відповісти на це запитання потрібно розробити метод ідентифікації, що дозволяє відрізнити чи є медіа-контент підробкою або реальним.

У результаті роботи цього методу користувач може отримати такі відповіді як:

– зображення або відео являється підробленим, якщо метод розпізнав DeepFake;

– зображення або відео являється оригінальним, якщо метод не виявив на них DeepFake;

– результат не може бути отриманим, оскільки зображення або відео не відповідають висунутим вимогам до вхідних даних.

У системах, де буде застосовуватися метод вимогами до безпеки є:

– забезпечення надійності та конфіденційності під час завантаження зображень та відео;

– здійснення регулярної перевірки на наявність вразливостей для запобігання витоку даних та збоїв системи.

3 ДОСЛІДЖЕННЯ ТЕХНОЛОГІЙ ПІДРОБОК ЗОБРАЖЕНЬ І ВІДЕО ТА МАТЕМАТИЧНИЙ ОПИС РОЗРОБЛЮВАНОГО МЕТОДУ ДЛЯ ЇХ ВИЯВЛЕННЯ

3.1 Опис та принцип роботи технології DeepFake

DeepFake походить від двох слів: «глибинне навчання» й «підробка» та представляє собою методику синтезу зображення людини, що базується на штучному інтелекті. Ця технологія використовує початкові зображення або відео, які можна знайти у вільному просторі в Інтернеті та проводить маніпуляцію з ними, використовуючи комбінацію певних штучних мереж.

Вперше світ побачив DeepFake у 2017 році після того, як анонімний користувач на форумі завантажив десятки відео-матеріалів, які скомпрометували американських зірок шоу-бізнесу та політиків. Він використав технологію DeepFake, замінивши акторів відео для дорослих на обличчя відомих людей, а також опублікував записи з політичними діячами, змінюючи сенс їх висловлювань, де начебто вони ображають своїх конкурентів.

Незважаючи на те, що опубліковані відео за короткий час були спростовані, суспільство занепокоїлось можливим розвитком цієї технології, яка у майбутньому може надати більш суттєві удари по репутації відомих людей. Через те, що відкритий код одного з методів DeepFake потравив до мережі Інтернет, майже кожна людина, який має пристрій з потужною відеокартою та зібраними медіа-матеріалами певної людини, має змогу зробити підробку фото та відео власноруч.

Підроблені зображення та відео-ролики зараз легко знайти в популярних соціальних мережах та відеохостингах, таких як YouTube, TikTok, Vimeo та інших. У 2020 році у сервісі для створення та перегляду коротких відео TikTok користувач завантажив підроблене відео від особи популярної зірки кіно Тома Круза.

На рисунку 3.1 зображено фрагмент цього відео, де майже неймовірно відрізнити на перший погляд чи є це відео справжнім, чи ні.

Рисунок 3.1 - Фрагмент підробленого відео з використанням обличчя Тома Круза

Окрім шантажу, руйнування репутації та глузування з відомих людей технологія DeepFake також може застосовуватись з позитивною метою. Так, наприклад, зірка футболу Девід Бекхем взяв участь у зйомці соціального ролику про небезпеку малярії на 10 різних мовах світу. Замість того, щоб вчити велику кількість різномовних реплік, носії цих мов записали їх, а штучний інтелект підлаштовував артикуляцію футболіста під речення цих мов [22].

На рисунку 3.2 наведено приклад згенерованого обличчя Девіда Бекхема та його артикуляцію, що підлаштована під аудіозапис іспанською мовою.

Рисунок 3.2 - Приклад згенерованого обличчя та артикуляції Девіда Бекхема під час запису соціального ролика

Але все ж таки, як показала практика, DeepFake несе більш загрозливий характер та з кожним роком стає більш небезпечним та важким для його розпізнавання. На рисунку 3.3 відображено хронологію еволюції DeepFake.

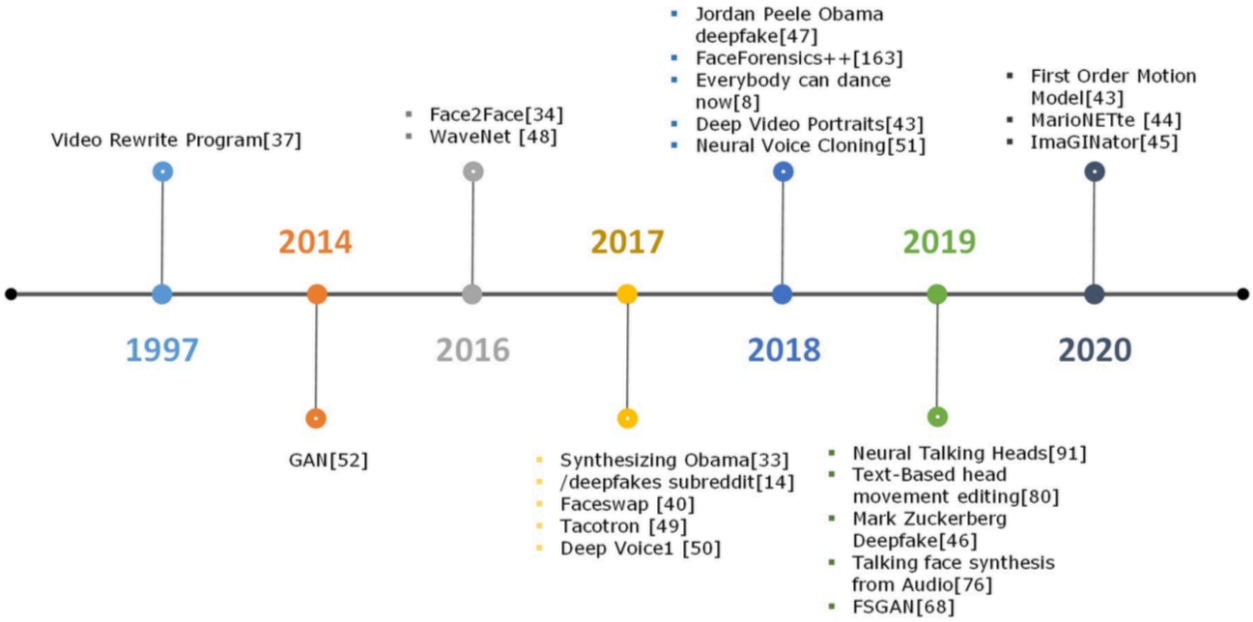

Рисунок 3.3 - Хронологія еволюції технології Deepfake

Створення підроблених зображень та відео поділяються на такі категорії як [23]:

– заміна обличчя (face-swap);

– синхронізація губ (lip-syncing);

– реконструкція обличчя або метод маріонетки (puppet-mastery);

– синтез обличчя та маніпулювання атрибутами (face synthesis and attribute manipulation);

– аудіо глибинні підробки (audio Deepfake).

При заміні обличчя людини у вихідному відео автоматично замінюється обличчям у цільовому відео, що показано на рисунку 3.4.

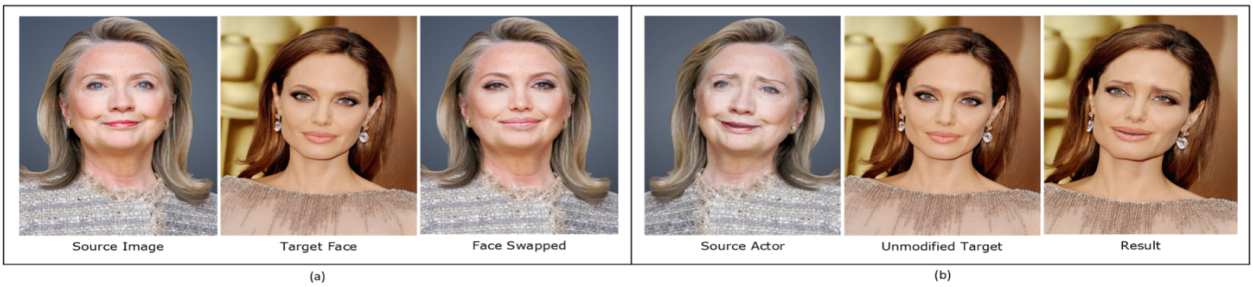

Рисунок 3.4 - Приклад роботи технології заміни обличчя (Face-swap)

Традиційні підходи, які замінюють обличчя, зазвичай проходять три кроки для виконання цієї операції. Спочатку виявляється обличчя на вхідних зображеннях, а потім з бібліотеки облич людей обирається зображення з кандидатом, який має схожий зовнішній вигляд та позу з людиною на вхідній медіа-записі. Далі метод починає замінювати очі, ніс, рот обличчя та коригувати освітлення та колір зображення обличчя кандидата, щоб воно відповідало вхідному та їх змішування відбулось плавним чином. На останньому етапі виконується ранжування кандидатів на заміну шляхом обчислення відстані збігу в області перекриття [23].

Ці підходи можуть дати хороші результати за певних умов, але мають два основні обмеження, а саме:

– відбувається повна заміна обличчя вхідної особи на цільову, тим самим втрачаючи вираз вхідного зображення людини;

– синтетичний результат дуже жорсткий та замінена особа виглядає неприродно, оскільки для отримання кращих результатів потрібна висока точність та відповідність пози людини.

Переважна більшість підходів для заміни обличчя використовують пари кодувальників та декодувальників. Кодувальник потрібен для отримання прихованих особливостей обличчя із зображення. А після цього декодувальник починає відновлювати обличчя. Щоб змінити місце обличчя на вхідному та цільовому зображеннях, потрібно мати дві пари кодувальників та декодувальників, причому кожен кодувальник спочатку навчається на вхідному, а потім на цільовому зображенні. Після закінчення навчання декодувальники змінюються місцями, тому оригінальний кодувальник вхідного зображення та декодувальник цільового зображення використовуються для регенерації цільового зображення з урахуванням особливостей вихідного зображення. У результаті, обличчя вхідного зображення накладається на цільове, при цьому зберігає його вираз обличчя [23].

На рисунку 3.5 продемонстровано створення підробленого зображення, використовуючи кодувальник та декодувальник.

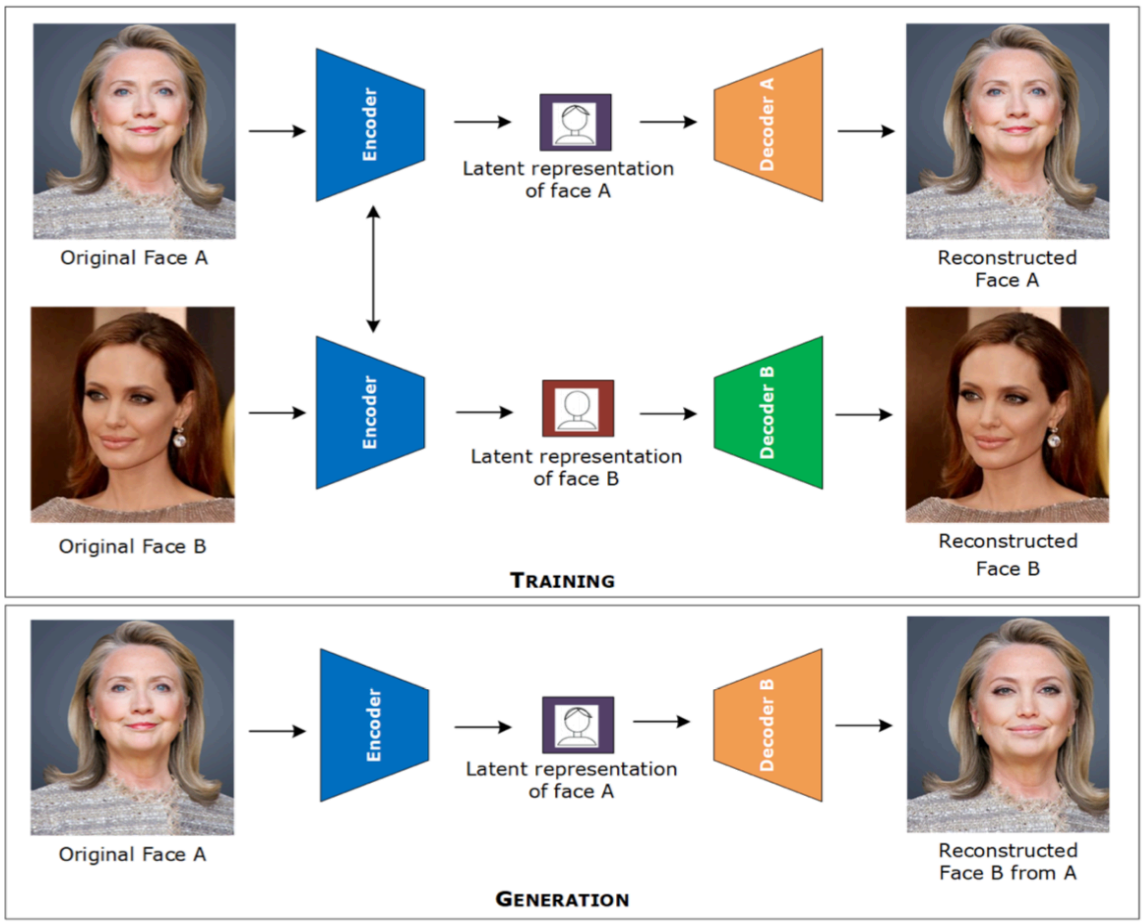

Рисунок 3.5 - Створення DeepFake з використанням кодувальників та декодувальників

Для оптимізації цієї моделі її скомбінували з генеративно-змагальною мережею. На сьогоднішній день існує така ГЗМ, яка безпосередньо базується на заміні обличчя та має назву Face-Swapping Gan (FSGAN). Ця технологія являється одною з найбільш точних та нових серед існуючих аналогів. Вона може використовуватися для генерації як і зображень, так і відео-контенту. Головна особливість цієї моделі полягає у тому, що FSGAN не потрібно навчати на тих особах, до яких буде застосовуватися цей підхід [23].

Модель Face-Swapping Gan складається з трьох нейронних мереж, кожна з яких виконує свої відповідні функції, а саме [23]:

– перша рекурентна НМ намагається підігнати обличчя вхідного зображення до особливих характеристик цільового медіа-запису (зміна положення голови ліворуч або праворуч, зміна кута нахилу голови назад або вперед);

– друга НМ відтворює перенос із вхідного зображення рис обличчя людини на відео;

– третя НМ відповідає за функцію плавного змішування зображень для того, щоб вихідна картинка не мала ніяких пошкоджень з артефактами та відповідно була більш реалістичною та збалансованою.

Цю модель також можна вдосконалити, додавши ще одну нейронну мережу, яка буде виконувати обробку тих ділянок обличчя людини, що були закриті під час запису відео.

На рисунку 3.6 зображено архітектуру підходу FSGAN.

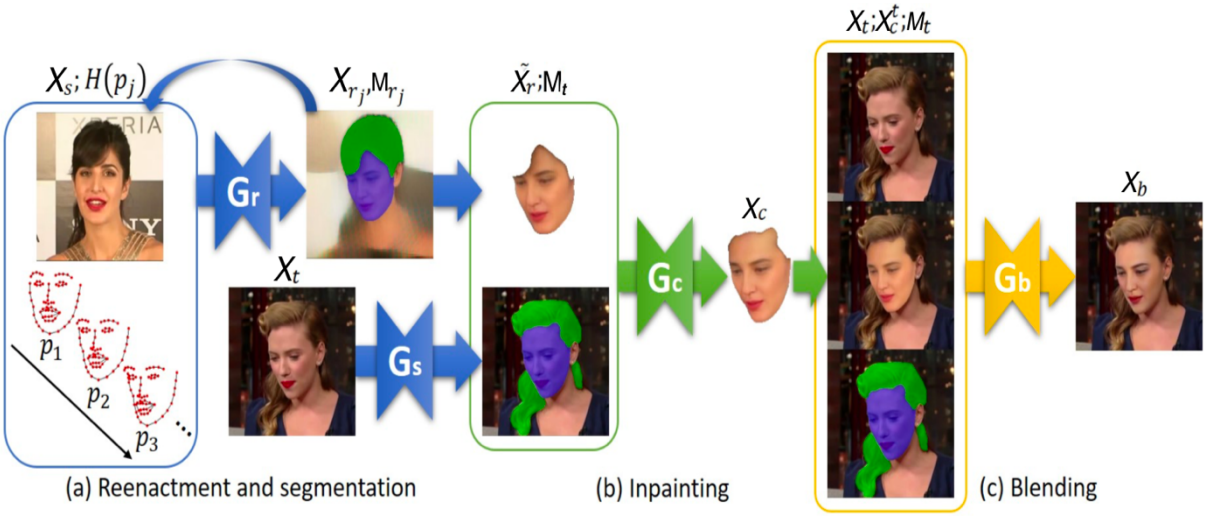

Рисунок 3.6 - Архітектура FSGAN

Нехай архітектура FSGAN складається з таких компонентів, як [23]:

– G_r - рекурентний генератор реконструкції;
– G_s - генератор сегментації;
– G_c - генератор зафарбування;
– G_b - генератор змішування;
– X_s - вхідне зображення;
– F_s - обличчя людини з вхідного зображення;
– X_t - цільове зображення;
– F_t - обличчя людини цільового зображення;
– X_r - згенероване зображення обличчя людини;
– M_r - сегментація згенерованого зображення;
– M_t - сегментаційна маска;
– X_b - фінальний результат підробленого зображення.

Якщо X_s вхідне зображення обличчя людини F_s, а X_t являється цільовим зображенням обличчя людини F_t, то ця модель починає створювати зображення

X_r. При цьому алгоритм намагається зробити заміну обличчя F_t на F_s, але при умові, що буде збережено його вираз та місце положення. Генератор G_r виконує функцію реконструкції, а саме робить витягування опорних точок F_t, генеруючи нове зображення X_r. Варто зазначити, що обличчя F_r представляє F_s з ідентичним виразом та позою обличчя F_t. Також G_r виконує обчислення маски M_r, а генератор G_s вже виконує сегментацію особи людини F_t та її волос. Після цього генератор G_c починає зафарбовувати ті місця обличчя, що були відсутні на зображенні X_r. Він використовує M_t, яка допомагає знаходити втрачені пікселі. На останньому етапі генератор G_b виконує операцію змішування, тим самим поєднуючи обличчя F_c з вхідним зображенням X_t. На виході отримуємо фінальний результат підробленого зображення X_b [23].

Окрім генерування та заміни обличчя людей, DeepFake також можна використовувати для генерації аудіо та видозміни артикуляції. Цей підхід має назву синхронізації губ (lip-syncing) та передбачає синтез відеозапису цільової особи таким чином, щоб область рота в відео, що маніпулюється, відповідала довільному аудіовходу.

Ключовим аспектом синтезу візуального мовлення є рух і зовнішній вигляд нижньої частини рота та прилеглої до нього області. Щоб передати повідомлення більш ефективно та природно, важливо генерувати правильні рухи губ разом із виразом обличчя.

З наукової точки зору, синхронізація губ має безліч застосувань в індустрії розваг, таких як створення фотореалістичних цифрових персонажів у фільмах чи іграх, голосових ботів та дубляж фільмів іноземними мовами. Більш того, він може допомогти людям, які погано чують, зрозуміти сценарій, читаючи по губах відео, створене з використанням справжнього звуку [23].

Принцип роботи цього підходу полягає у використанні рекурентної мережі, яка виконує трансформацію обраного аудіозапису a у деяку послідовність зміни опорних точок b, що розміщені на губах людини. Новітня рекурентна мережа LSTM сканує повністю відеозапис та виконує пошук ротової області. Після цього вона замінює рух губ вхідного запису на нові згенеровані відповідно до наданого аудіо. Для знешкодження розмитості на фінальному етапі відбувається згладжування медіа-запису, що підвищує реальність вихідного відео.

На рисунку 3.7 наведено приклад роботи алгоритму синхронізації губ та класифікації частини обличчя та тіла людини.

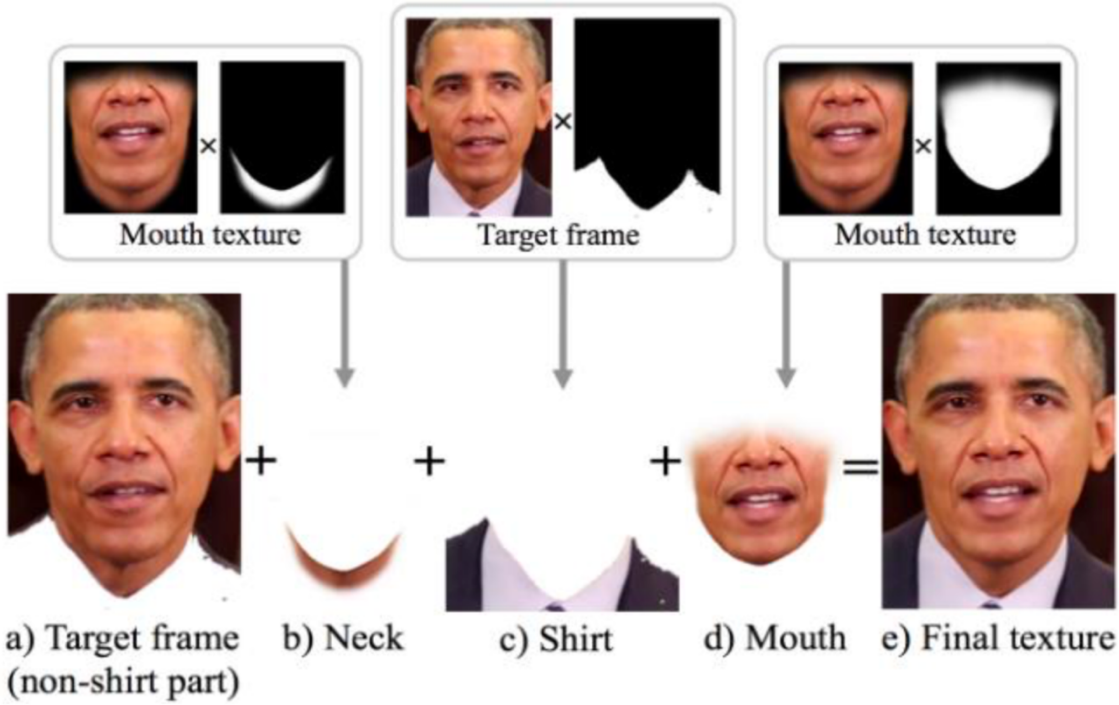

Рисунок 3.7 - Приклад роботи алгоритму синхронізації губ та класифікації частини обличчя та тіла людини

Інший підхід синтезу обличчя (face synthesis and attribute manipulation) вивчається вже кілька десятиліть. Він широко застосовується в мистецтві, анімації та індустрії розваг. Однак останнім часом його почали використовувати також й для створення глибинних підробок. Маніпуляції з особою можна розділити на дві категорії: генерація обличчя та редагування його атрибутів.

Створення обличчя включає синтез фотореалістичних зображень людини, якої не існує в реальному житті. На відміну від цього, редагування атрибутів особи включає зміну її зовнішнього вигляду існуючого зразка шляхом зміни області, специфічної для даного атрибуту, при збереженні незмінними нерелевантних областей [23].

Редагування атрибутів обличчя включає такі аспекти, як:

– зняття/одягання окулярів;

– зміна погляду;

– ретушування шкіри (наприклад, вирівнювання шкіри, видалення шрамів, мінімізація зморшок);

– модифікації вищого рівня, такі як вік, стать і т.д.

На рисунку 3.8 представлено покращення якості генерації синтетичних зображень облич людини з кожним наступним роком.

Рисунок 3.8 - Демонстрація покращення якості генерації синтетичних зображень облич людини

Нещодавно було запропоновано кілька підходів на основі GAN для редагування атрибутів обличчя цієї особи. У цій маніпуляції GAN приймає вихідне зображення обличчя як вхідні дані і генерує відредаговане зображення обличчя із заданим атрибутом.

Наприклад, генеративна мережа ICGAN використовує кодувальник, який відображає вхідне зображення обличчя у сховане середовище та вектор маніпуляції атрибутами, а ГЗМ відновлює зображення обличчя з новими атрибутами, враховуючи змінений вектор атрибутів як умову. Це призводить до втрати інформації та зміни оригінальної ідентичності особи у синтезованому зображенні [23].

Однак такий підхід має несподівані спотворення та розмитість, а тому не дозволяє зберегти вихідні дрібні деталі у згенерованому зображенні.

На рисунку 3.9 зображено приклади різних маніпуляцій з обличчям з використанням новітніх ГЗМ.

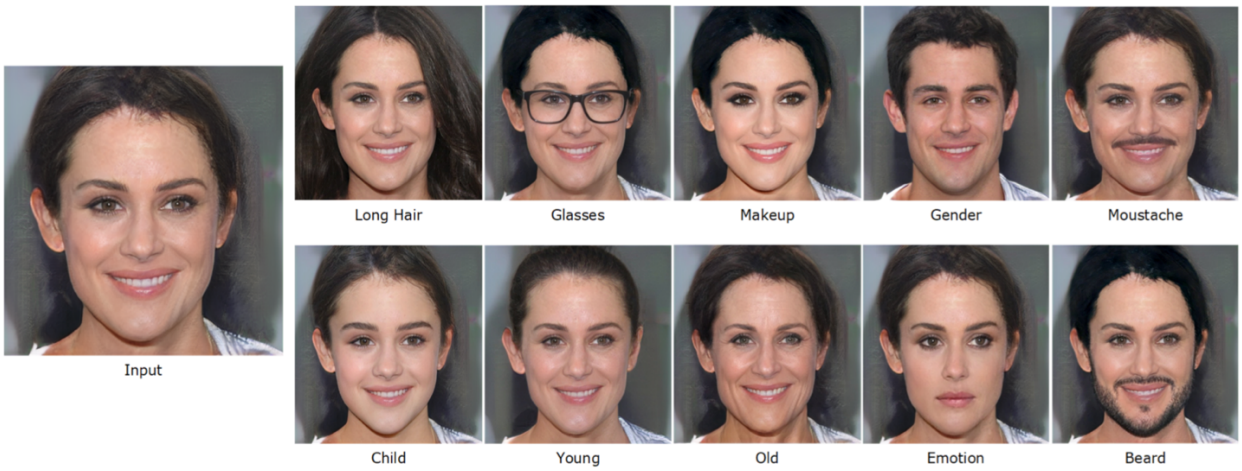

Рисунок 3.9 - Приклади різних маніпуляцій з обличчям з використанням новітніх ГЗМ

Не дивлячись на те, що йде безперервне покращення візуальної якості технології DeepFake, все ще існує низка невирішених проблем. До них відносяться [23]:

– узагальнення (генеративні моделі управляються даними, тому вони відображають вивчені в процесі навчання особливості вихідних даних. Для створення високоякісних підробок необхідно мати великий обсяг даних для навчання. Більш того, сам процес навчання потребує кількох десятків годин для створення переконливого аудіовізуального контенту. Крім того, перенавчання моделі для кожної конкретної цільової особи являється достатньо складним обчислюваним процесом. У зв'язку з цим потрібна узагальнена модель, що дозволяє використовувати навчену модель для кількох цільових осіб, які не помічені під час навчання, або з невеликою кількістю доступних навчальних зразків);

– проблема ідентичності (збереження ідентичності цільової особи є значною проблемою, особливо у завданнях відтворення обличчя людини, де його вирази керуються деякою вихідною ідентичністю. Дані обличчя особи, що рухається, частково переносяться на згенеровану людину. Це відбувається, коли навчання проводиться на одній або кількох особистостях, але сполучення даних виконується для однієї і тієї ж особи);

– парне навчання (навчена модель під наглядом може генерувати високоякісний результат, але за рахунок поєднання даних. Поєднання даних полягає у створенні бажаного результату шляхом виявлення схожих вхідних прикладів з навчальних даних. Цей процес трудомісткий і незастосовний у тих

сценаріях, коли на етапі навчання задіяні різні моделі поведінки особи та кілька осіб);

– варіювання пози та відстань від камери (Існуючі методи технології підробки зображень та відео генерують хороші результати для фронтального виду обличчя. Однак якість контенту, що маніпулюється, значно погіршується в сценаріях, коли людина на далекій відстані від камери, оскільки збільшення відстані від пристроїв захоплення призводить до низькоякісного синтезу обличчя та можливих артефактів);

– умови освітленості (існуючі підходи для створення DeepFake роблять підроблену інформацію в контрольованому середовищі з постійними умовами освітлення. Однак різка зміна умов освітлення, наприклад, у сценах у приміщенні або на вулиці, призводить до невідповідності кольорів та дивних артефактів в одержуваних відео);

– оклюзії (однією з основних проблем при створенні DeepFake є виникнення оклюзії, яка виникає, коли область вхідної особи та жертви закрита рукою, волоссям, окулярами або будь-якими іншими предметами. Більш того, оклюзія може бути результатом прихованої особи або частини ока, що в кінцевому підсумку призводить до невідповідності рис обличчя людини в контенті, що маніпулюється);

– тимчасова когерентність (ще одним недоліком згенерованих підробок є наявність явних артефактів, таких як мерехтіння та тремтіння між кадрами. Ці ефекти виникають через те, що механізми генерації DeepFake працюють над кожним кадром, не беручи до уваги тимчасову когерентність. Щоб подолати це обмеження, деякі методи або надають цей контекст генератору, або дискримінатору, або враховують втрати тимчасової зв'язності, або використовують рекурентні мережі, або комбінацію всіх цих підходів);

– відсутність реалізму в синтетичному аудіо (хоча якість, безумовно, стає набагато кращою, все ще існує потреба у вдосконаленні. Основними проблемами аудіопідробок є відсутність природних емоцій, пауз, дихання та темпу, в якому говорить об'єкт).

Таким чином, цими обмеженнями та вразливостями DeepFake можна скористатись при розробці методу для виявлення підробок зображень та відеоконтенту.

3.2 Існуючі підходи для виявлення підроблених зображень та відео

Існуючі підходи щодо виявлення маніпуляцій із зображеннями та відео спрямовані на використанні просторових або тимчасових артефактів, які були залишені під час створення підробок, чи на класифікації на основі даних.

Просторові артефакти включають у себе певні невідповідності, аномалії фону картинки або сліди використання ГЗМ. Тимчасові артефакти пов'язані з виявленням різних варіацій у поведінці людини, фізіологічних сигналів, когерентності та синхронізації відеокадрів.

Інший підхід не базується на конкретному артефакті, а повністю орієнтується на дані та виявляються маніпуляції з медіа-контентом шляхом використання класифікації.

На рисунку 3.10 наведено загальний технологічний процес виявлення DeepFake, використовуючи як ручні методи, що ґрунтуються на певних ознаках, так і методи на основі глибокого навчання.

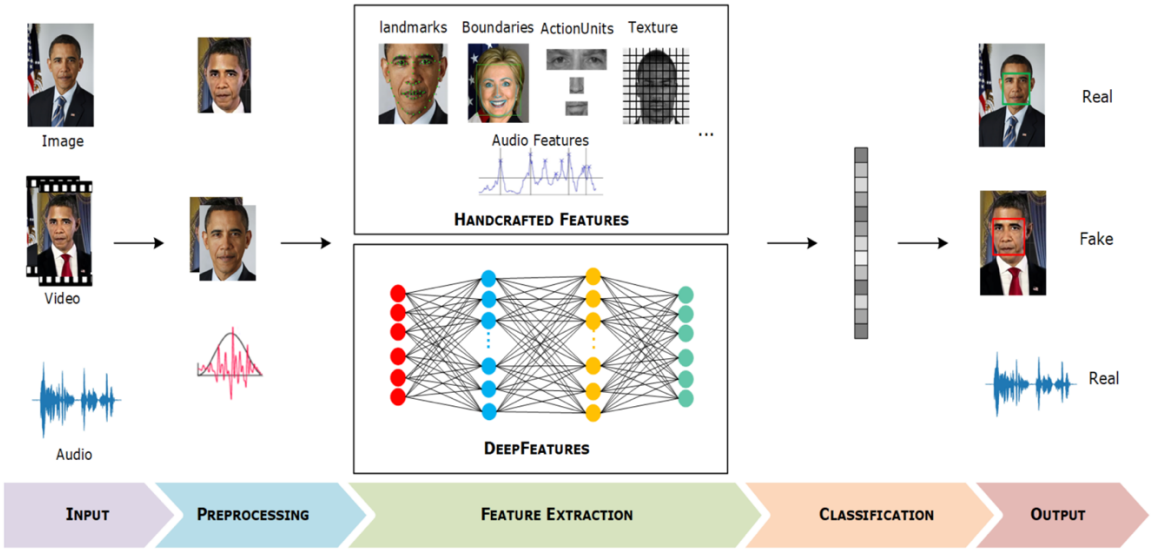

Рисунок 3.10 - Технологічний процес виявлення DeepFake

У своїй більшості, ручний підхід полягає у пошуку високорівневих візуальних артефактів в обличчі особи. Методи, що наслідують цей підхід, намагаються виділити конкретні збої в процесі генерації, який не відтворює ідеально всі деталі реальної особи. Наприклад, в обличчях, згенерованих ГЗМ, може виникнути невідповідність між кольором лівого та правого ока, а також інші форми асиметрії, такі як сережки тільки з одного боку або вуха з помітно різними характеристиками. Глибокі підробки, навпаки, часто генерують

непереконливі дзеркальні відображення в очах, які або відсутні, або представлені у вигляді білих плям, або грубо змодельовані зуби, які виглядають як одна біла пляма [24].

На рисунку 3.11 представлено приклад артефактів згенерованих підробок.

Рисунок 3.11 - Приклад артефактів згенерованих підробок

Цими можливими артефактами користується автор методу виявлення DeepFake Ф. Метерн, де для їх уловлювання будуються прості ознаки. Схожий метод автора Лі Чанга використовує один фактор, а саме моргання очей, яке має певну частоту та тривалість у людей, що не відтворюється у глибоко підроблених відео. Його рішення засновується на довгостроковій рекурентній мережі, що працює лише з послідовностями очей, призначене для уловлювання таких тимчасових невідповідностей [24].

Також до методів, що використовують біологічні чинники людини, відносяться биття серця, артефакти спотворення обличчя або невідповідність пози голови. У методі Х. Янга було зроблено спостереження, що алгоритми синтезу особи на основі ГЗМ здатні генерувати особу з високим рівнем реалізму та з великою кількістю деталей, але не мають явних обмежень на розташування цих деталей в обличчі людини. Отже, розташування знакових точок обличчя, таких як кінчики очей, ніс і рот, можна використовувати як дискримінаційні ознаки для перевірки справжності зображень ГЗМ. Ця ж проблема є у відеороликах DeepFake і може бути виявлена за допомогою 3D-оцінки положення голови [24].

Метод автора М. Олбрайта полягає у використанні певної моделі, задачі якої є обробка відео на вході, зчитування секторів відеозапису з різними частинами обличчя особи. Після цих операцій починається дослідження отриманих біологічних сигналів на їх природність. Можливі артефакти на обличчі людини можуть бути знайдені на щоках, носі чи інших ділянках [24].

На рисунку 3.12 відображено процес роботи методу, що базується на аналізі біологічних сигналів людини.

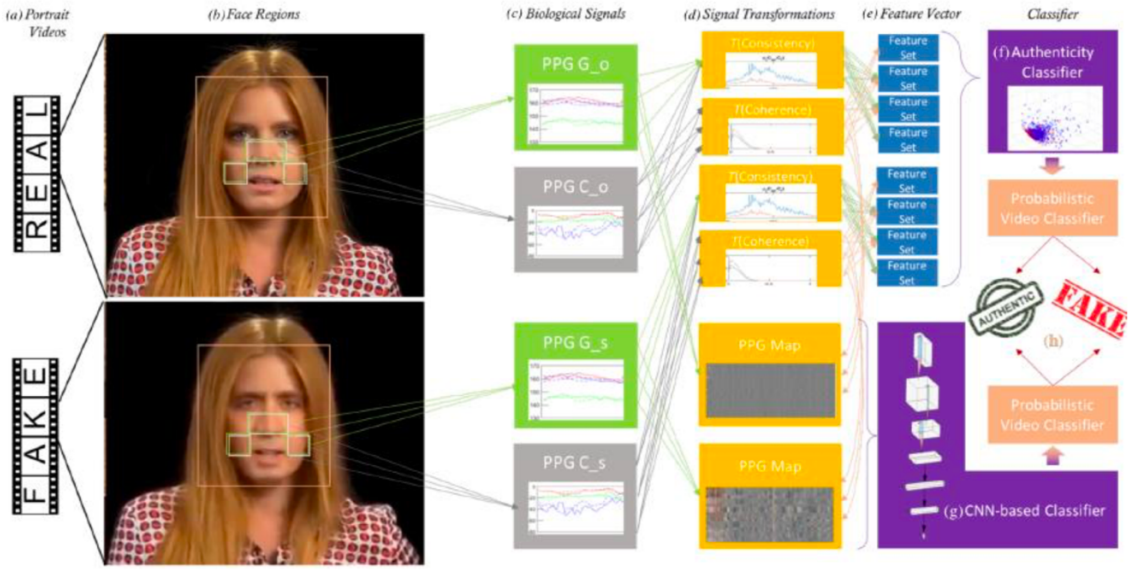

Рисунок 3.12 - Процес роботи методу, що базується на аналізі біологічних сигналів людини

Точність цього методу варіюється від 72 до 75%. До недоліків цього підходу можна віднести складність виконання, не досить високу точність та багатогодинне навчання цієї моделі.

Очевидною перевагою всіх цих методів ручного підходу є те, що візуальні артефакти не схильні до зміни розміру та стиску. Виявлення глибоких підробок цими методами може бути неефективним з таких причин [24]:

– часові характеристики відео змінюються від кадру до кадру;

– втрачається дуже важлива візуальна інформація, оскільки методи стиснення застосовуються зазвичай після зміни інформації у відео.

Також слід зазначити, що DeepFake розвивається, синтезуються більш досконалі підробки, та їх буде практично неможливо виявити ручним способом

Тому для подолання проблем методів, заснованих на ручній обробці ознак, активно вивчається підхід глибокого навчання (deep learning).

Один із таких методів являє собою застосування архітектури MiddleNet, яка складається з сукупності двох НМ, оскільки один із авторів цієї технології дійшов до висновку, що вирішення проблеми виявлення Deepfake та його аналогів не можливе з використанням лише однієї унікальної мережі.

Даний метод було розміщено між макроскопічним та мікроскопічним рівнем. Тобто це проміжна технологія, яка використовує ГНМ, що має у собі декілька шарів. Було запропоновано використовувати саме цю техніку, оскільки, ґрунтуючись на шумі картинки, неможливо застосовувати мікроаналіз у контексті відеозапису стисненого вигляду. Аналогічним чином, око людини важко розрізняє глибинні підробки на високих семантичних рівнях [25].

Ці архітектури засновані на мережах, які показають високі результати точності для класифікації зображень, в яких відбувається чергування шарів згорток.

Архітектура Middle-4 складається з 4 послідовних шарів згортки та об’єднання, а також слід за ними є щільна мережа, що містить прихований шар у собі. Для покращення узагальненості, згорткові шари використовують функції активації (Rectified Linear Unit), які потрібні для того, щоб внести нелінійність, та пакетну нормалізацію (Batch Normalization) для регулювання їх виходу, та запобігання ефекту градієнта, який зникає, а повністю підключені шари використовують метод виключення (Drop Out) для регулювання та підвищення їх стійкості [25].

У свою чергу архітектура MiddleIncept-4 полягає у заміні перших 2 шарів Middle-4 на початковий модуль (Module Inception). Початковий модуль являється блоком моделі зображення, метою якого є апроксимація оптимальної локальної розрідженої структури в ЗГМ. Тобто, він дозволяє використовувати кілька типів розміру фільтра, замість того, щоб бути обмеженим одним розміром фільтра, в одному блоці зображення, яке потім об’єднується та передається на наступний шар [25].

Ідея модуля полягає в тому, щоб скласти вихід кількох згорткових шарів різними формами ядра і таким чином збільшити простір функцій, в якому оптимізується модель. Замість згорток 5×5 оригінального модуля пропонується використовувати розширені згортки 3×3, щоб уникнути високої семантики [25].

У таблиці 3.1 наведено результати класифікації відео на двох наборах даних з використанням агрегації зображень.

Таблиця 3.1 - Результати класифікації відео на двох наборах даних

Мережа	Результати	
Набір даних	DeepFake	Face2Face
Middle-4	0.969	0.953
MiddleIncept-4	0.984	0.953

До переваг архітектури MiddleNet відносяться:

– невелика кількість параметрів, що забезпечують компактність моделі;

– висока швидкість навчання;

– точні результати на відео з високою роздільною здатністю.

Але у відеозаписах низької якості точність відчутно спадає, а також зараз немає звітів та результатів щодо точності виявлення підробок найновітніших алгоритмів DeepFake.

Інший існуючий метод виявлення підробок медіа-контенту базується на роботі ансамблю ЗНМ [26]. Використання ансамблю моделей призводить до суттєвого покращення точності результатів. Для реалізації цього методу автором обрано сімейство моделей EficientNet, оскільки вони є одними з найперспективніших для виконання такої задачі, як обробка зображень та відео [26]. На рисунку 3.13 наведено архітектуру цієї моделі.

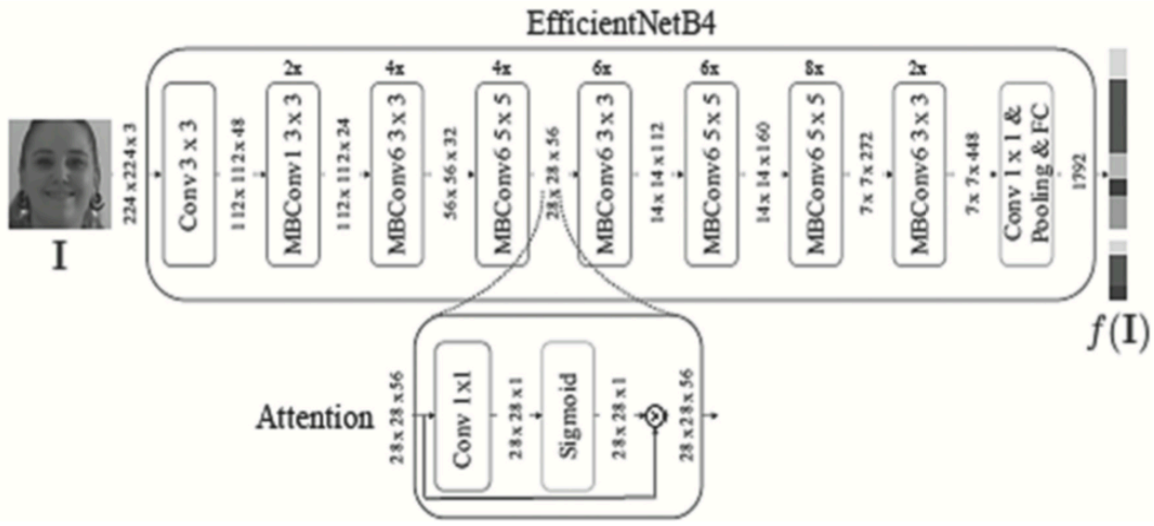

Рисунок 3.13 - Архітектура EficientNetB4

Зважаючи на особливості цієї архітектури, автор запропонував виконати декілька модифікацій, а саме [26]:

– додавання до моделі механізм Attention, який виконує функцію реалізації аналітичного метода та може визначити найбільш інформативну частину відео для виконання класифікації;

– додавання сіамських стратегій навчання для того, щоб екстраполювати додаткову інформацію щодо даних.

Вхідними даними до мережі являються зображення з обличчями, які оброблюються механізмом Attention, аналізуючи певні особливості, після чого вони передаються в активаційну сигмоїдальну функцію [26].

Навчання мережі відбувається з використанням двох стратегій, а саме:

– наскрізне навчання;

– сіамське навчання.

Перша стратегія використовує у якості даних метрику оцінки з певного набору зображень та відео, а інша направлена на використання можливостей узагальнення [26]. У ході дослідження виявлено, що артефакти, утворені під час генерації підробки, зазвичай були у вигляді неприродних згенерованих очей, зубів та інших частин обличчя. Навчання відбувалось на двох наборах даних FF+ та DFDC та показники точності становили 94% та 87% відповідно [26].

У своїй більшості, існуючі методи для виявлення підробок користуються ЗНМ, оскільки ця модель демонструє високу продуктивність, особливо у розпізнаванні зображень, серед технологій штучного інтелекту. Проте ефективність роботи залежить від кількох чинників зображення. Тому швидкість виявлення ЗНМ значно падає при зміні таких параметрів як:

– розмитість;

– яскравість;

– контрастність;

– шум;

– кут нахилу голови.

Таким чином, зловмисники мають можливість користуватись цією вразливістю, використовуючи фільтри для зміни певних характеристик зображень або відео.

На рисунку 3.14 надано приклад зображень, із зміненими характеристиками, які не були виявлені методами з використанням ЗНМ.

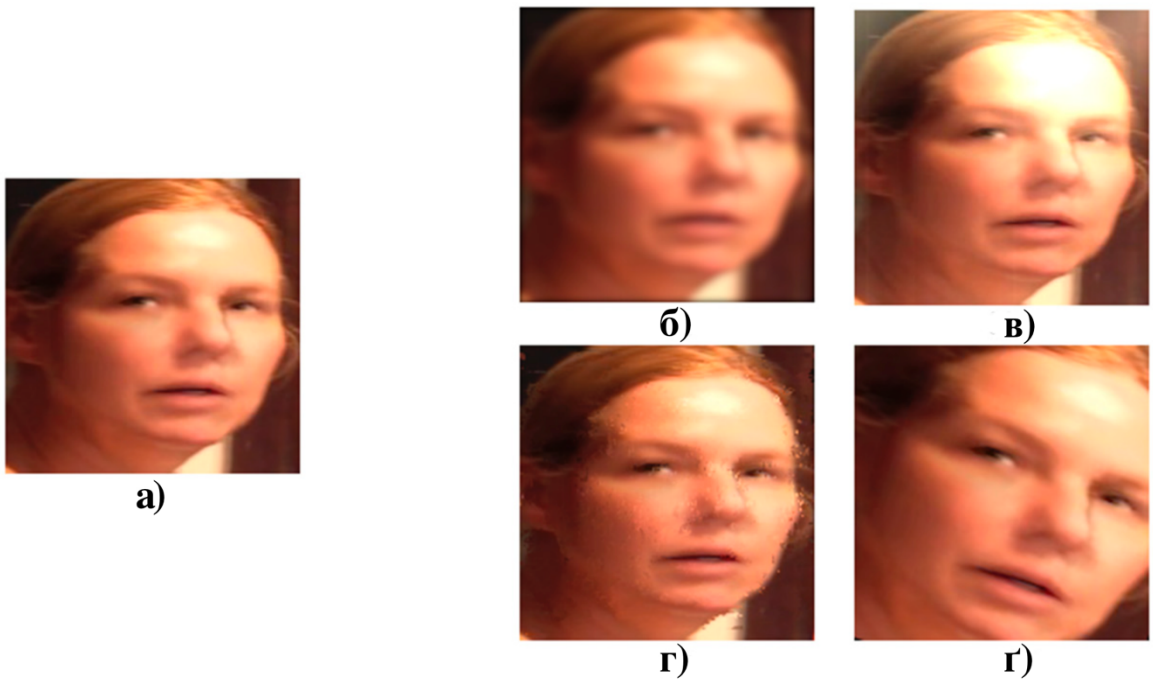

Рисунок 3.14 - Приклад зображень, із зміненими характеристиками, які не були виявлені методами з використанням ЗНМ

Зображення «а» являється тим, яке може бути виявленим як кадр загального маніпульованого відео. Але, наприклад, зображення «б» виявити традиційними методами неможливо через накладання Гаусового шуму. Розпізнати на підробку зображення «в» не має можливості через змінювання яскравості. Зображення «г» та «ґ» є невиявленими через шум чорних та білих часток та зміни куту нахилу голови людини відповідно.

3.3 Принцип роботи та математичний опис методу виявлення підробок зображень та відеоконтенту

Потенційні вразливості щодо змін певних характеристик зображень та відео необхідно прийняти до уваги під час розробки методу виявлення підробок. З цієї причини метод повинен надавати можливість аналізувати особливості комп’ютерного зору (feature of computer vision), тобто такий вимірюваний фрагмент даних зображення, який є унікальним для цього конкретного об’єкта. Це може бути окремий колір на зображенні або певна форма, наприклад лінія, край або сегмент.

На рисунку 3.15 представлена запропонована архітектура розроблюваного методу для виявлення підробок медіа-файлів.

Рисунок 3.15 - Архітектура технології та алгоритм виявлення підробок медіа-файлів

Розроблюваний метод поділяється на два головних процеси, а саме:

– попередня обробка даних (Preprocessing);

– класифікація (Classification).

На етапі попередньої обробки даних першим кроком є поділ обраного відео на кадри. Далі отримані кадри обробляються багатозадачною каскадною згортковою нейронною мережею (MTCNN). Принципи її роботи розглянемо більш детально.

MTCNN являє собою систему, що була розроблена для виявлення та вирівнювання облич людини з зображень. Процес її роботи складається з трьох етапів згорткових мереж, які здатні розпізнавати обличчя та місцезнаходження орієнтирів, таких як очі, ніс та рот [27].

Спочатку після отримання зображення виконується зміна його розміру у різних масштабах для побудови піраміди зображення, яка є входом наступної триступеневої каскадної мережі.

Після створення піраміди для кожної масштабованої копії зображення за допомогою ядра розміром 12×12, яке проходить через усі частини картинки, відбувається сканування на наявність облич. Воно починається з лівого верхнього кута ділянки зображення від координати (0,0) до (12,12). Ця частина

зображення передається до мережі P-Net, яка повертає координати обмежувальної рамки, якщо помічає обличчя. Потім цей процес повторюється з ділянками (0+2a, 0+2b) до (12+2a, 12+2b), зміщуючи ядро 12×12 на 2 пікселі вправо або вниз за раз. Зсув на 2 пікселя називається кроком, або скільки пікселів ядро переміщається щоразу [27].

Зміщення на 2 пікселя допомагає знизити складність обчислень без істотного зниження точності, оскільки обличчя на більшості зображень значно більше двох пікселів, малоймовірно, що ядро пропустить обличчя лише тому, що воно зсунулось на 2 пікселі. У той же час на комп'ютері буде виконуватися на чверть менше операцій, що дозволить програмі працювати швидше та займати менше пам'яті.

Кожне ядро буде меншим по відношенню до великого зображення, тому воно зможе знайти менші обличчя за розміром на збільшеному зображенні. Аналогічно, ядро буде більшим щодо зображення меншого розміру, тому воно зможе знайти більші обличчя на зображенні меншого масштабу [27].

На рисунку 3.16 наведено приклад виводу результатів роботи мережі P-Net.

Kernel Coordinates	Bounding box: x1	Bounding box: y1	Bounding box: x2	Bounding box: y2	Confidence
(0,5)	0.50	0.25	0.80	0.58	0.98
(0,6)	0.45	0.17	0.75	0.49	0.96
(0,7)	0.52	0.08	0.73	0.41	0.94
(1,4)	0.42	0.34	0.67	0.66	0.92
(1,8)	0.40	0.01	0.66	0.32	0.96
(3,5)	0.32	0.22	0.49	0.59	0.88
(3,6)	0.25	0.20	0.52	0.53	0.91
(6,6)	0.01	0.18	0.25	0.50	0.89
(6,8)	0.02	0.00	0.22	0.33	0.90

Рисунок 3.16 - Зразок виводу результатів P-Net

Виходячи з результатів, на деяких границях мережа має більшу впевненість, що там знаходиться обличчя ніж в інших. Таким чином потрібно

ретельно перевірити результати роботи мережі P-Net та видалити поля з найменшою впевненістю.

Через те, що на зображенні може знаходись більш одного обличчя, неможливо обрати рамку з найбільшим рівнем достовірності та видалити всі інші. Тому наступним кроком потрібно зменшити кількість обмежуючих рамок. Це можна зробити з використанням техніки, що має назву придушення немаксимумів (NMS). Використання цієї техніки здійснюється шляхом сортування обмежувальних рамок за рівнем достовірності, обчислюється площа кожного з ядер, а також площа перекриття між кожним ядром та ядром з найбільшою оцінкою. Ядра, які сильно перетинаються з ядром із високою оцінкою, видаляються. Нарешті, NMS повертає список граничних осередків, що "вижили" [27].

NMS виконує свою роботу один раз для кожного масштабованого зображення, а потім ще раз з усіма ядрами, що вижили, з кожного масштабу. Це дозволяє позбавити зайві рамки, що обмежують та звузити пошук до однієї точної рамки для кожного обличчя.

Далі координати обмежувальних рамок перетворюються на координати реального зображення, шляхом їх добутку на його фактичну ширину та висоту.

На рисунку 3.17 зображено приклад виявлення обличчя з його точними координатами.

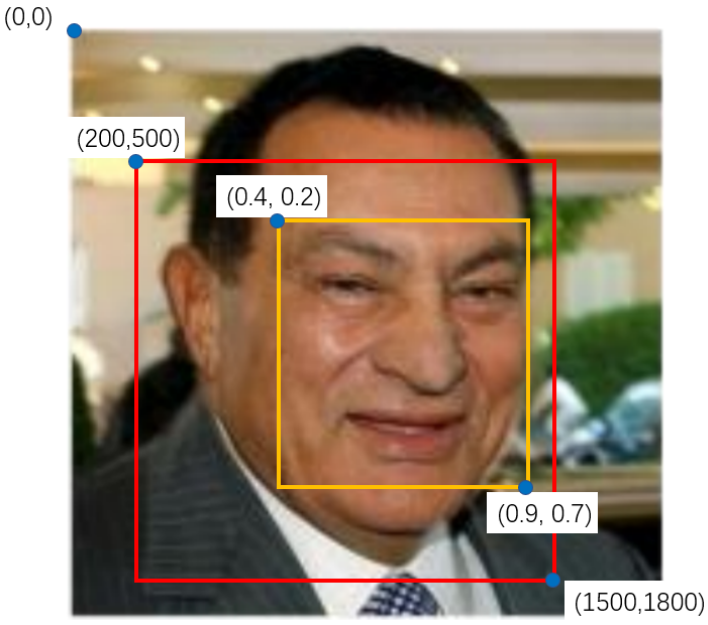

Рисунок 3.17 - Приклад виявлення обличчя з його точними координатами

Ця технологія показує високу точність, оскільки має у своїй наявності три нейронні мережі, котрі мають декілька шарів та кожна мережа коректує результати роботи попередньої. За допомогою використання NMS та мереж R-Net та O-Net відсіюється велика кількість помилкових обмежувальних блоків, що значно зменшує об’єм даних [27].

Після того, як виявлені кадри з наявністю обличчя впорядковані, з них вилучаються різні характеристики комп’ютерного зору. Вектор ознак був створений шляхом отримання ознак комп'ютерного зору з вирівняних зображень обличчя за допомогою таких операцій, як:

– обчислення;
– кластеризація;
– фільтрація.

У таблиці 3.2 наведено вилучені характеристики комп’ютерного зору із зображень, що містять обличчя.

Таблиця 3.2 - Вилучені характеристики комп’ютерного зору

Назва атрибута	Пояснення
mse (mean squared error)	Середня квадратична різниця між розрахунковими та фактичними значеннями
psnr (the peak signal-to-noise ratio)	Відношення між максимально можливою потужністю сигналу і потужністю спотворюючого шуму
ssim (structural similarity index measure)	Сприймана якість цифрових телевізійних та кінематографічних зображень
rgb (red, green, blue)	Відсотковий вміст кожного червоного, зеленого та синього кольору у зображенні
hsv (hue, saturation, value)	Відсоткове співвідношення кожного відтінку, насиченості та значення зображення
histogram	Гістограма, що показує кількість пікселів у зображенні з певним значенням яскравості чи тону
luminance	Середнє значення загальної яскравості зображення
variance	Дисперсія зображення
edge_density	Відношення кількості крайових пікселів до загальної кількості пікселів у зображенні

dct (discrete cosine transform)	Зміщення дискретного косинусного перетворення зображення

Середньоквадратична помилка (mse) визначає подібність зображень, використовуючи різницю інтенсивності пікселів між двома зображеннями. Пікове відношення сигналу до шума (psnr) оцінює інформацію про втрати якості зображення. Psnr фокусується на числових відмінностях, а не на візуальних відмінностях людини, оскільки psnr розраховується за допомогою mse, коли mse дорівнює 0, psnr також встановлюється на 0. Міра індексу структурної подібності (ssim) оцінює тимчасові відмінності, що відчуваються людиною щодо яскравості, контрасту та структурних аспектів. RGB та HSV представляють колірний простір зображення. Гістограма є розподілом відтінків у зображеннях, а яскравість являє собою середнє загальне значення.

Оскільки метод створення DeepFake синтезує цільове зображення кожного кадру, це може призвести до неприродних змін різних характеристик комп'ютерного зору. Крім того, при створенні підробок цільове зображення виходить з обмеженою роздільною здатністю, а його розмір змінюється у вигляді матриць перетворення для припасування до вихідного зображення. Тому різкість часто виявляється гіршою. Крім того, виникають спотворення та розмитість. Тому обрані особливості значно впливають на процес створення підробок зображень та відео.

На рисунку 3.18 відображено процес отримання ознак комп'ютерного зору.

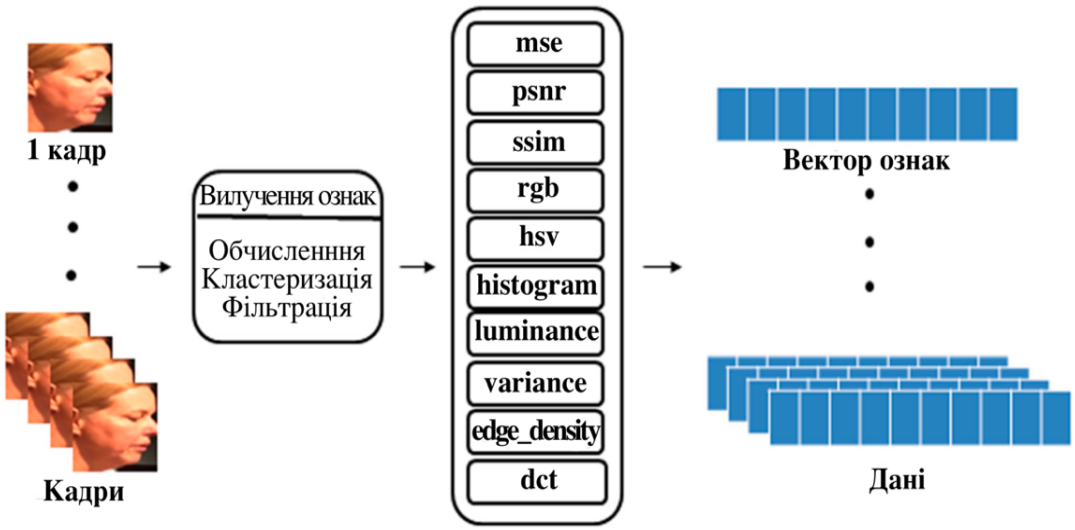

Рисунок 3.18 - Процес отримання ознак комп'ютерного зору

Кожна з ознак зображення розраховується за формулою:

$$x_i = \frac{abs(x_{i+1} - x_i)}{mean(x)}, \quad (3.1)$$

де x_i – значення ознаки для i-го кадру;

x_{i+1} – значення ознаки для $i + 1$-го кадру;

$mean(x)$ – середнє значення ознаки всіх кадрів, отриманих з одного відео.

Далі виконується підрахунок розбіжності. Для кожної ознаки вона розраховується шляхом угруповання швидкості зміни між певними витягнутими номерами кадрів. Розрахована розбіжність кожної ознаки використовуються у якості даних для навчання глибинної нейронної мережі. Після цього додається залежна змінна, що вказує чи дані відео є підробкою, або підробки не виявлено. Глибока нейронна мережа навчається на цих даних та згодом використовує їх для виявлення підробок.

3.4 Отримані практичні результати розроблюваного методу виявлення підробок зображення та відео

Розбиття та вилучення кадрів із відео виконується за допомогою використання мови програмування Python 3 та OpenCV, що являє собою бібліотеку комп'ютерного зору, яка призначена для аналізу, класифікації та обробки зображень. Для експерименту обрано відео тривалістю 63 секунди, що містить у собі 1512 кадрів.

За допомогою мережі MTCNN оброблюються отримані кадрі, шукаючи обличчя людини. У ході дослідження підраховано, що один кадр оброблюється приблизно одну секунду. Але можна досягти значного прискорення, проводячи обробку на графічному процесорі, використовуючи програмно-апаратну архітектуру паралельних обчислень CUDA від компанії Nvidea та фреймворк машинного навчання Pytorch для мови програмування Python. Ці компоненти дають можливість оброблювати приблизно від 60 до 100 кадрів за одну секунду. Тому, наприклад, відео, що триває 5 хвилин та має у собі 7200 кадрів обробиться приблизно за 72 секунди, враховуючи максимальну продуктивність.

На рисунку 3.19 зображено приклад знаходження обличчя мережею MTCNN на одному з кадрів відео.

Рисунок 3.19 - Приклад знаходження обличчя мережею MTCNN на одному з кадрів відео

Далі виконано обробку характеристик комп'ютерного зору отриманих кадрів з обличчями та розраховано кожну з них за формулою (3.1). На рисунку 3.20 представлено кадри з найбільшою розбіжністю кожної ознаки з одного підробленого відео.

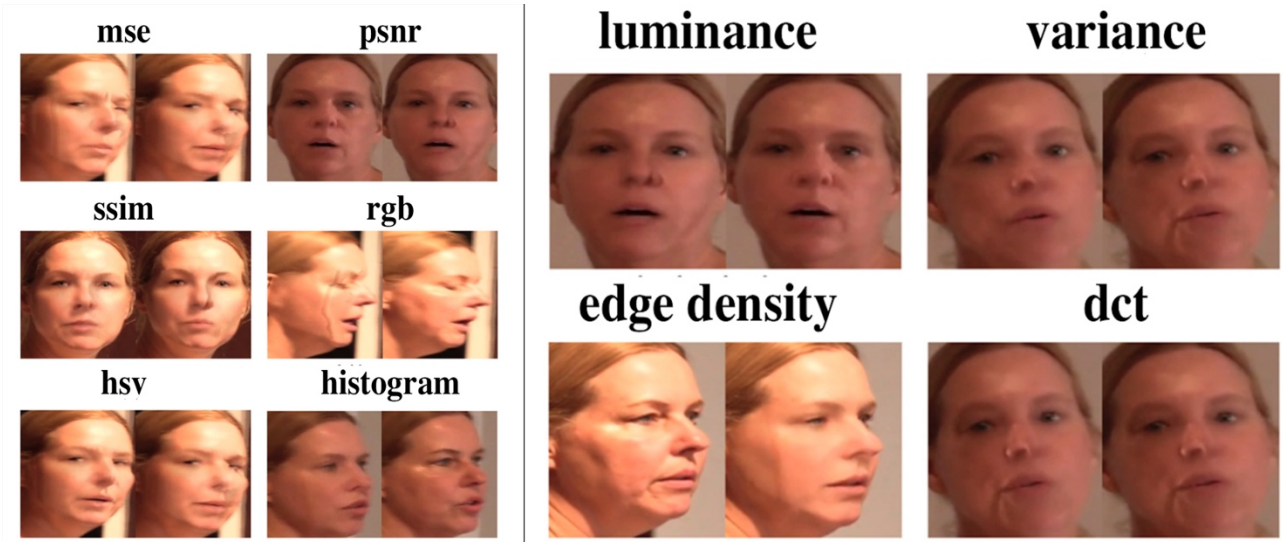

Рисунок 3.20 - Кадри з найбільшою розбіжністю кожної ознаки з одного підробленого відео

На другому етапі класифікації зображень, тобто віднесення до підроблених або реальних зображень виконується навчання глибинної нейронної мережі для якої потрібно два набора даних, а саме:

– FaceForensics++;
– DFDC.

Перший набор складається з 1000 оригінальних відеопослідовностей, які були оброблені чотирма автоматизованими методами маніпулювання: Deepfakes, Face2Face, FaceSwap та NeuralTextures. Дані були отримані з 977 відеороликів YouTube, і всі відео містять, у своїй більшості фронтальні обличчя, які легко відстежити.

Система організації конкурсів з дослідження даних, а також соціальна мережа фахівців з обробки даних та машинного навчання Kaggle, яка належить корпорації Google створила набір даних DFDC, що містить у собі більше ста тисяч відео, які важать приблизно 470 Gb [28].

Для навчання розробленої моделі машинного навчання можна використовувати Keras. Це відкрита бібліотека, написана мовою Python, яка забезпечує взаємодію зі штучними нейронними мережами. Вона є надбудовою над фреймворком TensorFlow [29].

Оскільки під час обробки певних ознак комп'ютерного зору отриманих з кадрів підроблених відео та подальшому навчанні оперується дуже велика кількість даних та обчислювальних процесів з ними, потребується обчислювана машина з ультра характеристиками для того, щоб реалізувати розроблений метод, зменшити час на його навчання та виконання.

У таблиці 3.3 наведено мінімальні системні вимоги для проведення експерименту.

Таблиця 3.3 - Мінімальні системні вимоги для проведення експерименту

Системний ресурс	Мінімальні вимоги
CPU	AMD Ryzen 9 5950X 16-core Processor
RAM	32 GB DDR4
GPU	Nvidia GeForce RTX 3070 Ti
VRAM	8 GB GDDR6X

У таблиці 3.4 наведено порівняння точності виявлення глибинних підробок двох існуючих моделей MiddleNet та SVM на двох наборах даних.

Таблиця 3.4 - Порівняння точності виявлення глибинних підробок двох існуючих моделей на двох наборах даних

Назва методу	Назва набору даних	
	FaceForensics++	DFDC
MiddleNet	93,21%	77,71%
SVM	64,24%	62,91%

Метод SVM використовує застарілі технології та не надає можливості виявляти підробки новітніх моделей, тому він має відносно низькі показники точності. У свою чергу метод MiddleNet справляється з виявленням більшості сучасних підробок та показує вищі показники точності знаходження підробленого медіа-контенту.

Виходячи з того, що модель MiddleNet навчається так само як і розроблювана модель, використовуючи мову програмування Python та бібліотеку Keras на двох існуючих наборах даних, можна сказати, що очікувана точність розробленого методу буде не менше 93,21% та 77,71% включно, тому що у ній враховано недоліки порівняної моделі та задіяні новітні та більш потужні технології.

ВИСНОВКИ

У процесі виконання кваліфікаційної роботи проаналізовано предметну область обраної теми, а саме детальний опис понять «зображення» та «відео», принцип їх створення, вдосконалення з плином часу, роботи з ними та їх роль у сучасному світі, пояснення існуючої проблеми з приводу фальсифікації медіа-контенту та важливість її вирішення за допомогою методу, що потрібно розробити для виявлення підробок із використанням найновітніших нейронних мереж та інших механізмів глибинного навчання.

Спираючись на аналіз, визначено основні вимоги для дослідження й розробки методу, який перевірятиме фото або відео на справжність.

Проведено аналіз щодо видів нейронних мереж, виявлено їх область застосування, а також досконально описано принцип їх роботи. Більш глибоко розглянуто згорткові нейронні мережі, які безпосередньо потрібні для роботи з зображеннями й відео та представлено їх застосування у різних сферах життя.

Для вирішення поставленої задачі здійснено дослідження, яким саме чином відбувається підроблення медіа-матеріалів, розглянуто моделі для підробок та проаналізовано їх сильні сторони та можливі вразливості. Після цього проведено аналіз методів виявлення підробок зображень та відео, що вже існують, їх архітектури, використаних технологій та отриманих результатів точності.

На основі отриманої інформації, а саме переваг та недоліків моделей підробок та методів їх виявлення, розроблено метод, який дозволяє витягнути з медіа-матеріалів певні характеристики комп’ютерного зору, знайти найбільшу розбіжність між кадрами та використовувати цю інформацію для навчання глибинної нейронної мережі, яка буде перевіряти на оригінальність зображення та відео. У ході розробки визначено мінімальні системні вимоги для запуску методу та навчання нейронної мережі на двох наборах даних, а також зроблено припущення щодо точності виявлення підробок, на основі результатів існуючих методів та застосованих ними технологіями.

При доопрацюванні та підтримці цей проект може знайти своє застосування як і у власних цілях, так і для деяких компаній або правоохоронних органів у кібербезпечних цілях для виявлення підробок задля спростування дезінформації, шантажу або інших чинників.

Розроблений метод можна вдосконалювати, шляхом застосування більш сучасного апаратного забезпечення та новітніх технологій, а також за допомогою оновлення набору даних для навчання моделі, що сприятимуть підвищенню результатів точності.

Зважаючи на те, що розроблений метод відповідає вимогам, котрі були висунуті у поставці задачі, можна підвести підсумок, що кваліфікаційна робота виконана у повному обсязі.

За результатами досліджень підготовлена до публікації стаття «Двоетапний метод розпізнавання фальсифікованих зображень та відео» у науково-технічний журнал «Біоніка інтелекту» ХНУРЕ [30].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Глобальне опитування щодо способів задля споживання контенту. URL: https://blog.hubspot.com/preferences/ (дата звернення: 01.10.2021)

2.Статистика активних користувачів Instagram. URL: https://techcrunch.com/instagram/ (дата звернення: 01.10.2021)

3.Статистика завантажень застосунку TikTok. URL: https://sensortower.com/tiktok-downloads/ (дата звернення: 02.10.2021)

4. М. О. Романюк, Крочук, А. С., Пашук І. П. Зображення : ЛНУ ім. Івана Франка, 2012. 564 с.

5.Зображення URL: https://uk.wikipedia.org/wiki/Зображення(дата звернення: 07.10.2021)

6. Джаконія В.Є. Запис та відтворення інформації. 2002. 311–316 с.

7.Відео URL: https://www.wikiwand.com/ru/Відео(дата звернення: 11.10.2021)

8. Bhadeshia H. K. D. H. Neural Networks in Materials Science: ISIJ International 39 (10). 1999. 966–979 с.

9. Gurney K. An introduction to neural networks. 1997. 123 с.

10. Штучна нейронна мережа URL: https://www.wikiwand.com/uk/Штучна_нейронна_мережа/Література(дата звернення: 14.10.2021)

11. Міркес Є. М. Нейроінформатика:Навч. посіб. Новосибірськ. 2003. 55 с.

12. Zhernova, P.Y. Adaptive Kernel Data Streams Clustering Based on Neural Networks Ensembles in Conditions of Uncertainty about Amount and Shapes of Clusters / Zhernova, P.Y., Deineko, A.O., Bodyanskiy, Y.V., Riepin, V.O. // Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018. 2018. 7–12 с.

13. Бодянський Є.В. Штучні нейронні мережі: архітектури, навчання, застосування. [Текст] / Є.В. Бодянский, О.Г. Руденко — Харків: ТЕЛЕТЕХ, 2004. — 369с

14. Хайкін. С. Штучні мережи. М.: Вільямс, 2006. 1104 с.

15.Згорткові нейронні мережі. URL: https://evergreens.com.ua/ua/articles/cnn.html(дата звернення: 18.10.2021)

16. Згортка в Deep Learning. URL:https://www.reg.ru/blog/svyortka-v-deep-learning-prostymi-slovami/(дата звернення: 20.10.2021)

17. Глибинне навчання. URL: https://medium.com/@balovbohdan/глубокое-обучение-разбираемся-со-свертками-6e47bfc27792(дата звернення: 22.10.2021)

18. Cheng-Yuan Liou. Autoencoder for words. 2014. 84-96 с.

19. Goodfellow Ian J. Generative Adversarial Networks. 2014. 3 с.

20. A Detailed Explanation of GAN with Implementation Using Tensorflow and Keras. URL: https://www.analyticsvidhya.com/blog/2021/06/a-detailed-explanation-of-gan-with-implementation-using-tensorflow-and-keras/(дата звернення: 29.10.2021)

21. Айрапетов А.Е. Коваленко А.А. Види генеративно-змагальних мереж. 2018. 7-12 с.

22. DeepFake of David Beckham. URL: https://techcrunch.com/2019/04/25/the-startup(дата звернення: 05.11.2021)

23. Nirkin Y. DeepFake methods and FSGAN as subject agnostic face swapping and reenactment. 2019. 1-8 с.

24. Masood M. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. 2019. 1-21 с.

25. Afchar D. MiddleNet: a Compact Facial Video Forgery Detection Network. 2018 .5-6 с.

26. Білоцерковський В.В., Удовенко С.Г., Чала Л.Е. Метод нейромежевого розпізнавання фальсифікованих зображень. Науково-технічний журнал «Біоніка інтелекту» 2020. №1. С. 32-42.

27. Multi-task Cascaded Convolutional Networks (MTCNN) for Face Detection and Facial Landmark Alignment URL:https://medium.com/@iselagradilla94/multi-task-cascaded-convolutional-networks-mtcnn(дата звернення: 26.11.2021)

28. DeepFake Dataset URL:https://www.kaggle.com/c/deepfake-detection-challenge/data(дата звернення: 01.12.2021)

29. Шолле Ф. Deep Learning with Python. 2018. 400 с.

30. Н.І. Калита, Р.С. Захаров. Двоетапний метод розпізнавання фальсифікованих зображень та відео. Науково-технічний журнал «Біоніка інтелекту» 2021. №2. С. 00-00.