УДК 004.42

ДИПФЕЙКИ – СТВОРЕННЯ ТА БОРОТЬБА З НИМИ

Кудрик К.І., студентка, kostyantyn.kudryk@kname.edu.ua, ХНУМГ
Бредіхін В.М., к. т. н., доцент, bredixinv@gmail.com, ХНУМГ

Вступ. У сучасному світі технології швидко розвиваються, надаючи нові можливості, але з'являються і нові виклики та загрози. Однією з таких загроз є технологія створення Deepfake, яка дозволяє створювати фальшиві відео та аудіо, які надзвичайно важко виявити. Розглянемо основні технології створення Deepfake та методи їх виявлення.

Deepfake — це реалістична маніпуляція аудіо-, фото- та відеоматеріалами за допомогою штучного інтелекту для досягнення максимальної схожості з реальними зображеннями та звуком. Він заснований на генеративних змагальних мережах (GAN), які складаються з двох мереж - генеративної мережі, яка створює зображення, та дискримінаційної мережі, яка розрізняє справжні та підроблені зразки.

До сучасних технологій створення Deepfake відносять: глибоке навчання та штучні нейронні мережі, генеративно-змагальні мережі (GAN), автоенкодери [1].

Країни в усьому світі запроваджують правові норми для вирішення проблем, пов'язаних із технологією deepfake. Ці правила спрямовані на запобігання зловживанню технологією deepfake і захист конфіденційності та безпеки людей. Для боротьби з дипфейками за допомогою технічних рішень використовують: аналіз артефактів, аналіз руху та голосу, використання технологій блокчейн. Співпраця між дослідниками, розробниками та політиками і розробка та вдосконалення алгоритмів виявлення deepfake має вирішальне значення для ефективної боротьби з deepfake.

Основний матеріал. Генерація Deepfake передбачає вилучення кадрів із відео, визначення контурів обличчя та оптимізацію обчислювальних ресурсів.

Може знадобитися ручне втручання, щоб забезпечити точне визначення контурів обличчя як у вихідному, так і в цільовому відео.

Перший крок це підготовка набору даних — вилучення кадрів із відео та визначення обличь за допомогою спеціального програмного забезпечення. Перегляд розпізнаних обличь вручну займає багато часу, але це необхідний для отримання якісних результатів.

Далі потрібно вирівняти та обрізати обличчя для кращих результатів навчання. Цей процес гарантує, що всі грані правильно вирівняні та придатні для введення в нейронну мережу.

Потім вирівняні зображення обличчя використовуються для створення набору даних для навчання нейронних мереж. Для кращої продуктивності цей набір даних розділено на підмножини для навчання та перевірки.

Обрана модель навчається за допомогою підготовленого набору даних, що може тривати кілька днів або тижнів. Тривале навчання зазвичай призводить до кращих результатів, тоді як продуктивність обладнання відіграє вирішальну роль.

Навчена модель використовується для накладання згенерованих обличь на вихідні зображення. Метод Пуассона зазвичай використовується для змішування створених обличь із вихідними зображеннями.

Останнім кроком є зшивання кадрів разом із оригінальним аудіо з однаковою частотою кадрів. Цей процес поєднує оригінальну сцену з накладеними обличчями, в результаті чого створюється глибоке фейкове відео [2].

Для клонування мови генеративні змагальні мережі обирають випадкові речення подібної структури та довжини.

Серед програм для створення Deepfake слід виділити:

- DeepFaceLab — це популярне програмне забезпечення для створення дипфейків із використанням нейронних мереж для заміни обличь у відео;
- FakeApp — це настільна програма, яка дозволяє користувачам легко створювати фотореалістичні відео із заміненими обличчями;
- Neural Voice Puppetry – це система використовує нейронні мережі для аналізу записаної мови та створення високоякісного відео із синхронізованим рухом губ;
- Avatarify — це безкоштовний фільтр Deepfake, який анімує зображення нерухомих обличь у режимі реального часу під час відео дзвінків на таких платформах, як Skype і Zoom. дослідження.

Для боротьби з Deepfake слід виділити наступні інструменти:

- Semantic Forensics (SemaFor) — програма експертизи змісту (семантичного аналізу) мультимедійних матеріалів використовується для пошуку фальсифікованих медійних матеріалів (текстів, аудіо, зображень, відео) для захисту від великомасштабних дезінформаційних атак у режимі реального часу;
- Assembler програма яка допомагає у перевірці автентичності зображень. Як зазначено в описі вона поєднує кілька методів виявлення маніпуляцій над зображенням, включаючи детектор, який визначає дипфейки, створені за допомогою нейромережі StyleGAN. Assembler оснащений сьома детекторами, кожен із яких розроблено визначення конкретної маніпуляції із зображенням;
- компанія Adobe випустила White paper функції, що дозволить маркувати зображення, оброблені у фоторедакторі. Система Content Authenticity Initiative (CAI) додаватиме до зображень теги, які допоможуть відстежити всю історію фото аж до того, якою камерою було знято. За допомогою тегів, захищених криптографією, буде фіксуватися факт обробки зображення.

Сучасні методи базуються виявлення дипфейкових відео на основі біометричної поведінки обличчя та голосу, а не артефактів, створених системами синтезу осіб, дорогими рішеннями для нанесення водяних знаків або іншими підходами.

Так, наприклад, контрастне навчання є основою підходу POI-Forensics засновано на переміщеннях та звукових сигналах, унікальних для реальної людини, яку дипфальсифікують [3]. Вектори, отримані з вихідного матеріалу для кожного випадку, порівнюються з тими ж векторами потенційно фальшивого відео, з аспектами і ознаками, отриманими як з відео, так і з аудіокомпонентів потенційно фальшивого відео, рис. 1.

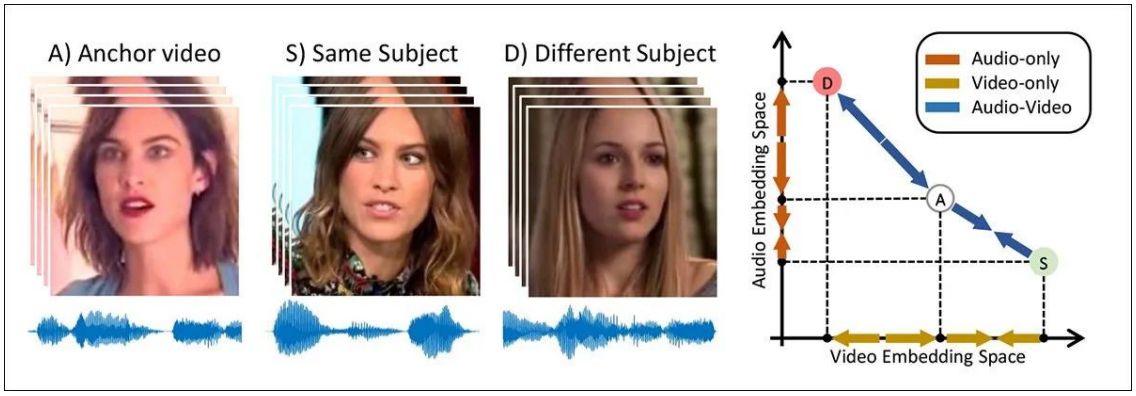

Рисунок 2 – Виявлення діпфейкового відео [4]

POI-Forensics використовує мультимодальний підхід до перевірки особистості, використовуючи програмну біометрію на основі візуальних та звукових сигналів. Фреймворк включає окремі аудіо- і відеомережі, які в кінцевому підсумку отримують характеристичні векторні дані, які можна порівняти з тими ж витягнутими функціями в досліджуваному відео потенційного дипфейка, рис. 2.

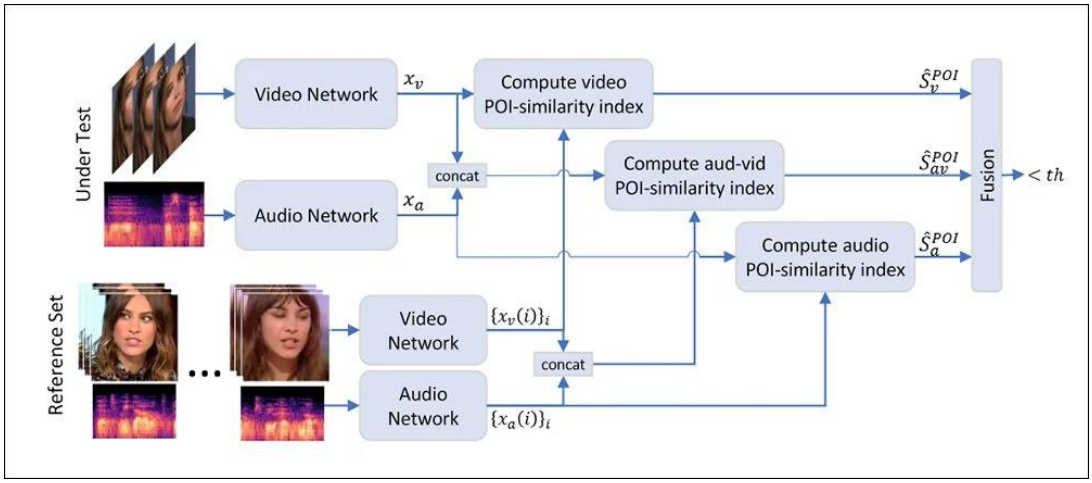

Рисунок 2 – Концептуальна архітектура POI-Forensics [4]

Навчання проводилося на тренувальному наборі даних, який у результаті містив 2250 фото з дипфейками та 1825 фото з реальними людьми. Датасет DeepfakeTIMIT [5] складається з відео, що належать тільки до класу «дипфейки», датасет DFDC [6] складається з відео, що належать до обох класів, і має файл з метаданими, в якому містяться мітки класу для кожного з відео та приналежність до набору даних: тренувального, валідаційного або тестовому. Тестова та валідаційна вибірки містили у собі по 200 фото дипфейків та реальних людей.

Переваги даного методу полягають у двох аспектах:

- такі артефакти можуть бути змодельовані безпосередньо за допомогою простих операцій обробки зображень, щоб зробити його негативним прикладом. Оскільки навчання моделі глибокої підробки для створення негативних прикладів вимагає багато часу і ресурсів, метод, що описує, економить багато часу і ресурсів при зборі навчальних даних;

- оскільки такі артефакти зазвичай є у відео deepfake з різних джерел, даний метод більш надійний порівняно з іншими. Він оцінюється на двох наборах даних відео deepfake для оцінки його ефективності на практиці.

Висновок. Однак слід вказати, що цей метод навряд чи виявляться успішними у довгостроковій перспективі. Зрештою згадані дослідження підказують творцям дипфейків, як покращувати дискримінативні мережі, що, у свою чергу, призводить до більш ретельного навчання генеративних мереж і, як наслідок, підвищує якість підробок.

Нейромережі технології машинного навчання стрімко вдосконалюються, на підставі чого можна дійти висновку, що створення абсолютно достовірних дипфейк-відео, які буде неможливо відрізнити від реальних зробивши експертизу, але ж методи боротьби з ними теж будуть вдосконалюватись..

Список використаних джерел

1. Declaration for the Future of Internet 2022. URL: https://digitalstrategy.ec.europa.eu/en/library/declaration-future-internet
2. Вальорска М.А. Діпфейк та дезінформація: практ. посіб. / Аґнєшка М. Вальорска ; пер. з нім. В. Олійника. Київ: Академія української преси; Центр Вільної Преси, 2020. 36 с.
3. Exposing DeepFake Videos By Detecting Face Warping Artifacts. URL: https://arxiv.org/abs/1811.00656
4. Deepfake Detection Based on Original Human Biometric Traits URL: https://www.unite.ai/deepfake-detection-based-on-original-human-biometric-traits/
5. P. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” arXiv preprint arXiv: 1812.08685, 2018.B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. C. Ferrer, “The deepfake detection challenge (DFDC) preview dataset” arXiv preprint arXiv: 1910.08854, 2019.