УДК 004.8+004.056.5

Л. М. Куперштейн[1]
Н. В. Людва[1]
С.О. Прокопенко[2]

АНАЛІЗ МОЖЛИВОСТЕЙ НЕЙРОНИХ МЕРЕЖ ДЛЯ ВИЯВЛЕННЯ МУЛЬТИМЕДІЙНИХ ФЕЙКІВ

[1] Вінницький національний технічний університет;
[2]Харківський національний економічний університет ім. С. Кузнеця

Анотація

Дослідження розглядає застосування передових технологій, зокрема згорткових нейронних мереж (CNN) і генеративних змагальних мереж (GAN), для виявлення мультимедійних фейків. Використання CNN дозволяє ефективно виявляти невідповідності у зображеннях, а GAN є інструментом для створення синтетичного контенту, такого як deepfakes, а також основою для побудови методів виявлення фейків, що має важливе значення для протидії маніпуляціям та дезінформації.

Ключові слова: мультимедійні фейки, дезінформація, згорткові нейронні мереж, генеративні змагальні мережі, виявлення фейків.

Abstract

The study examines the use of advanced technologies, such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), to detect multimedia fakes. The use of CNNs allows for the effective detection of inconsistencies in images, while GANs are becoming a tool for creating synthetic content, such as deepfakes, with fake detection methods, which is important for countering manipulation and disinformation.

Keywords: multimedia fakes, disinformation, convolutional neural networks, generative adversarial networks, fake detection.

Вступ

Виявлення мультимедійних фейків, особливо у сфері політичного контенту, стає все більш важливим у сучасну цифрову епоху. З поширенням високотехнологічних інструментів редагування зображень та відео розповсюдження маніпульованого медіа стало потужним інструментом дезінформації та пропаганди [1]. У зв'язку з цим дослідники і технологи все частіше звертаються до передових інтелектуальних технологій, зокрема згорткових нейронних мереж і генеративних змагальних мереж, щоб розробити методи виявлення сфабрикованого мультимедійного контенту [2].

CNN – клас глибоких нейронних мереж, особливо пристосованих до аналізу візуальних образів, продемонстрували чудові можливості в різних завданнях комп'ютерного зору, включаючи класифікацію зображень, виявлення об'єктів і сегментацію [3]. Використовуючи свою ієрархічну архітектуру та здатність виокремлювати складні ознаки із зображень, CNN пропонують багатообіцяючий шлях для виявлення невідповідностей та нерівностей, що вказують на маніпуляції в медіа.

Доповнюючи CNN, GAN стали потужним інструментом для створення синтетичного мультимедійного контенту, зокрема зображень і відео. Однак сама природа GAN – коли мережа-генератор створює реалістичні результати, а мережа-дискримінатор розрізняє справжні та фальшиві зразки – дає можливість використовувати навчання в умовах суперництва для виявлення фальшивих медіа [3]. Навчивши дискримінантні мережі розрізняти справжній і сфабрикований контент, дослідники можуть використовувати змагальну динаміку GAN для розробки надійних механізмів виявлення.

Це дослідження має на меті проаналізувати можливості CNN і GAN у виявленні мультимедійних фейків з особливим акцентом на політичному контенті. З огляду на потенційний вплив маніпуляцій у політичних медіа на суспільний дискурс, виборчі процеси та демократичні інститути, розробка надійних методів виявлення фейків має важливе суспільне значення.

Результати дослідження

Суть діпфейків полягає у створенні реалістичних на вигляд зображень або відео, що можуть бути використані для обману та маніпуляції аудиторією. В контексті маніпуляції населенням, діпфейки можуть бути застосовані для поширення дезінформації, впливу на громадську думку або дискредитації публічних осіб, що підвищує необхідність у розробці надійних методів їх виявлення та протидії [4].

Приклади використання діпфейків для України під час війни включають створення фальшивих відео із заявами публічних осіб, які насправді ніколи не робили. Наприклад, були створені діпфейк-відео із зображенням президента України, які поширювали дезінформацію про капітуляцію чи здачу позицій [5]. Такі фальшиві матеріали можуть дестабілізувати ситуацію, викликати паніку серед населення та деморалізувати військових.

Діпфеки створюються за допомогою типу штучного інтелекту, який називається генеративними змагальними мережами (GAN). GAN складаються з двох нейронних мереж, генератора і дискримінатора, які навчаються одночасно в змагальному процесі [2].

Типова архітектура GAN для створення діпфейків зображена на рис. 1.

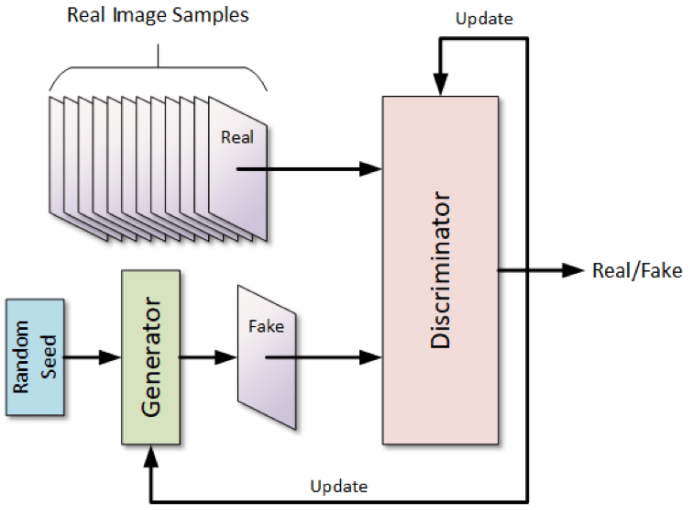

Рис. 1. Архітектура GAN

Ось покрокове пояснення того, як GAN використовують для створення діпфейків [6].

Мережа-генератор: Роль генератора полягає у створенні реалістичних на вигляд зображень або відео. У контексті діпфейк генератор навчається створювати зображення, схожі на обличчя цільової особи. Спочатку генератор створює випадкові зображення, які не схожі на обличчя жертви. Однак у міру навчання він вчиться генерувати зображення, які дискримінатору стає дедалі важче відрізнити від справжніх.

Мережа дискримінатора: Роль дискримінатора полягає в тому, щоб розрізняти реальний і згенерований контент. Його навчають на наборі даних, що містить як реальні, так і синтетичні зображення.

У процесі навчання дискримінатор покращує свою здатність відрізняти реальні зображення від створених генератором.

Змагальне навчання: Генератор і дискримінатор тренуються в змагальній манері. Генератор прагне створювати зображення, які неможливо відрізнити від реальних, тоді як дискримінатор прагне правильно класифікувати, чи є зображення реальним або згенерованим.

Цей процес призводить до безперервного циклу вдосконалення: генератор стає кращим у створенні реалістичних зображень, а дискримінатор – у відрізненні справжніх зображень від фейкових.

Для виявлення зображень, згенерованих GAN, за допомогою глибинних мереж було запропоновано різні методи. Одним з таких методів є метод на основі нейронних мереж для виявлення фальшивих відео GAN [7]. Цей метод використовує методи попередньої обробки для аналізу статистичних особливостей зображення і покращує виявлення фальшивих зображень обличчя, створених людиною. Існують різні підходи на основі глибокої згорткової нейронної мережі для виявлення фальшивих зображень, згенерованих GAN. Модель спочатку використовує мережу глибокого навчання для вилучення рис обличчя на основі мереж розпізнавання облич. Потім використовується етап точного налаштування, щоб зробити риси обличчя придатними для виявлення справжніх/підроблених зображень. Ці методи показують хороші результати в тестах на різних наборах даних, підтверджуючи їх ефективність у виявленні підроблених зображень.

На додаток до традиційних моделей глибокого виявлення підробок, для ефективного виявлення фальшивих зображень було впроваджено гібридний підхід, а саме двопотокову мережу для виявлення фальсифікацій облич [8]. Потік класифікації облич використовується в GoogleNet для навчання моделі на підроблених і справжніх зображеннях. Потім, потік патч-триплетів використовується для аналізу ознак за допомогою екстрактора ознак стеганоаналізу і фіксує низькорівневі характеристики камери та залишки локального шуму. Експериментальні результати показують, що цей підхід може навчатися як на фальшивих, так і на справжніх зображеннях.

Візуальні трансформери є ще одним потужним інструментом для створення та виявлення діпфейків. Вони використовують архітектуру трансформерів для обробки зображень, що дозволяє їм ефективно враховувати контекст і деталі на різних рівнях абстракції. Основні переваги візуальних трансформерів перед GAN включають можливість кращого розпізнавання контексту, вищу точність та стабільність у процесі навчання [9]. Візуальні трансформери можуть бути інтегровані в існуючі системи виявлення фальсифікацій, що робить їх більш універсальними та ефективними. Використання трансформерів дозволяє зменшити кількість помилкових спрацювань та покращити загальну ефективність виявлення фальсифікацій.

SORA (Style-based Optimized Reconstruction and Adaptation) є ще одним потужним інструментом для створення діпфейків [10]. Це передовий метод, який використовує стилізацію та адаптацію для генерації високоякісних зображень і відео. SORA дозволяє створювати надзвичайно реалістичні зображення шляхом точного налаштування стилю цільової особи, що забезпечує високу точність та деталізацію.

SORA може використовуватися для виготовлення реалістичних діпфейків завдяки своїй здатності точно відтворювати особливості та стиль обличчя цільової особи [11]. Цей метод забезпечує високу ступінь контролю над кінцевим результатом, дозволяючи створювати зображення, які важко відрізнити від справжніх. SORA також може бути інтегрована з іншими моделями та методами для підвищення точності та реалістичності діпфейків, що робить її ефективним інструментом у сфері генеративного дизайну.

Щоб вирішити питання зловживання, такі як створення deepfake відео, OpenAI розробляє інструмент для виявлення відео, згенерованих за допомогою Sora. Ця ініціатива спрямована на зменшення потенційних етичних і юридичних проблем, пов’язаних із використанням генеративних технологій штучного інтелекту. OpenAI також співпрацює з експертами для тестування моделі та забезпечення відповідності етичним стандартам перед її публічним випуском [12].

Для виявлення діпфейків існує кілька додаткових методів, що базуються на різних підходах та технологіях. Першим підходом є аналіз метаданих, який дозволяє виявити невідповідності у технічних деталях файлів, таких як інформація про камеру, дата створення та інші параметри, що можуть бути змінені або відсутні у діпфейках. Спектральний аналіз використовує частотні компоненти зображень або відео для виявлення аномалій, які можуть свідчити про підробку [13]. Цей метод допомагає виявити зміни, внесені під час створення діпфейків.

Аналіз морфологічних ознак досліджує форму, розміри та відстані між об'єктами на зображенні, що може виявити аномалії, характерні для підроблених медіа. Наприклад, неприродні пропорції обличчя можуть вказувати на діпфейк. Моделі, що вивчають час, такі як рекурентні нейронні мережі (RNN) або довга короткострокова пам'ять (LSTM), можуть бути використані для аналізу відео з метою виявлення невідповідностей у рухах об'єктів або змінах кадрів [14].

Інтеграція мультимодальних даних включає комбінацію аналізу зображень, аудіо та тексту для досягнення більш точних результатів. Наприклад, синхронізація рухів губ з аудіо може бути використана для виявлення фальшивих відео.

Висновки

Діпфейки представляють серйозну загрозу, оскільки можуть бути використані для поширення дезінформації, маніпуляції громадською думкою та дискредитації публічних осіб. У контексті війни в Україні, діпфейки використовувалися для створення фальшивих заяв публічних осіб, що могло дестабілізувати ситуацію та викликати паніку серед населення.

Діпфейки створюються за допомогою генеративних змагальних мереж (GAN), які складаються з генератора та дискримінатора. Генератор створює реалістичні зображення, тоді як дискримінатор намагається їх розрізняти. Цей змагальний процес покращує якість створених зображень.

Для виявлення діпфейків існують різні методи, включаючи використання глибоких нейронних мереж, гібридних підходів, таких як двопотокові мережі, візуальні трансформери. Останні використовуються для обробки зображень і мають переваги в точності та стабільності навчання.

Для боротьби з маніпуляціями та дезінформацією важливо використовувати різноманітні методи виявлення підробок, включаючи аналіз метаданих, спектральний аналіз, аналіз морфологічних ознак. Інтеграція мультимодальних даних також може підвищити точність виявлення фальшивих медіа. Постійний стрімкий розвиток технологій штучного інтелекту вимагає подальших досліджень та розробки нових ефективних технологій для виявлення діпфейків для протидії їх зловживанням та забезпечення достовірності мультимедійного контенту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Finger L. Overview of how to create deepfakes - it's scarily simple. Forbes. URL: https://www.forbes.com/sites/lutzfinger/2022/09/08/overview-of-how-to-create-deepfakesits-scarily-simple/?sh=28fd12972bf1 (date of access: 10.05.2024).
2. Lu Y., Ebrahimi T. Assessment framework for deepfake detection in real-world situations. EURASIP journal on image and video processing. 2024. Vol. 2024, no. 1.
3. A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends / A. Younesi et al. IEEE Access. 2024. P. 1.
4. Економічна правда. Епоха "глибоких" підробок: що таке deepfake та як від нього захиститися. Економічна правда. URL: https://www.epravda.com.ua/publications/2020/08/14/664022/ (дата звернення: 12.05.2024).
5. Baig R. The deepfakes in the disinformation war – DW – 03/18/2022. dw.com. URL: https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433 (дата звернення: 13.05.2024).
6. Amalraj Victoire D. T., Abishek A., Ajay Rakesh T. A. M. A Chat Application for Disabled using Convolutional Neural Network Deep Learning Algorithm. Quing: International Journal of Innovative Research in Science and Engineering. 2023. Т. 2, № 2. С. 128–140.
7. Karandikar A. Deepfake Video Detection Using Convolutional Neural Network. International Journal of Advanced Trends in Computer Science and Engineering. 2020. Т. 9, № 2. С. 1311–1315.
8. Stanciu D.-C., Ionescu B. Deepfake Video Detection with Facial Features and Long-Short Term Memory Deep Networks. 2021 International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 15–16 jul. 2021 p. 2021.
9. End-to-End object detection with transformers / N. Carion et al. Computer vision – ECCV 2020. Cham, 2020. P. 213–229.
10. Технологічний ритм. Які можливості має Sora AI ?. Друкарня. URL: https://drukarnia.com.ua/articles/yaki-mozhlivosti-maye-sora-ai-V3o3k (дата звернення: 19.05.2024).
11. Deepfakes are about to become a lot worse, openai's sora demonstrates. linnk. URL: https://www.spiceworks.com/tech/artificial-intelligence/guest-article/deepfakes-are-about-to-become-a-lot-worse-openais-sora-demonstrates/ (дата звернення: 15.05.2024).
12. OpenAI: we'll help you detect videos made with sora genai tool. Technology News For IT Channel Partners and Solution Providers | CRN. URL: https://www.crn.com/news/security/2024/openai-we-ll-help-you-detect-videos-made-with-sora-genai-tool (дата звернення: 16.05.2024).
13. Lyu S. DeepFake detection. Multimedia forensics. Singapore, 2022. P. 313–331.
14. MCW: a generalizable deepfake detection method for few-shot learning / L. Guan et al. Sensors. 2023. Vol. 23, no. 21. P. 8763.

Куперштейн Леонід Михайлович — к. т. н., доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця, email: kupershtein@vntu.edu.ua.

Прокопенко Сергій Олександрович **–** аспірант, Харківський національний економічний університет ім. С. Кузнеця, email: prokopenko.serhii@gmail.com

Людва Назарій Вікторович — студент групи 1БС-20б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, email: nazarliudva@gmail.com.

Kupershtein Leonid — PhD (eng), associated professor of information protection department, Vinnytsia National Technical University, Vinnytsia, email: nazarliudva@gmail.com.

Prokopenko Serhii **–** PhD student, Semen Kuznets Kharkiv National University of Economics, email: prokopenko.serhii@gmail.com

Liudva Nazariy — student of group 1БС-20б, Faculty of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: kupershtein@vntu.edu.ua.