

АНАЛІЗ МЕТОДІВ ВИКОРИСТАННЯ ТЕХНОЛОГІЇ DEERFAKE ДЛЯ ІНФОРМАЦІЙНИХ АТАК

Вінницький національний технічний університет

Анотація

У статті розглянуто сучасні підходи до створення та використання технології Deepfake у контексті інформаційних атак. Проаналізовано основні методи формування аудіовізуального контенту з використанням інструментів штучного інтелекту. Наведено класифікацію типових технік маніпуляції зображенням та відео, а також охарактеризовано поширені підходи до їх реалізації. В роботі акцентовано увагу на зростаючих ризиках, пов'язаних із використанням Deepfake для поширення дезінформації, підриву довіри до цифрового контенту та втручання в особисту ідентичність. Підкреслено актуальність подальшого розвитку ефективних засобів виявлення підробок і удосконалення методів захисту від подібного роду загроз.

Ключові слова: deepfake, штучний інтелект, машинне навчання, інформаційна атака, захист інформації

Abstract

The article examines modern approaches to the creation and use of Deepfake technology in the context of information attacks. The main methods of forming audiovisual content using artificial intelligence tools are analyzed. A classification of typical image and video manipulation techniques is presented, and common approaches to their implementation are characterized. The paper focuses on the growing risks associated with the use of Deepfake to spread disinformation, undermine trust in digital content, and interfere with personal identity. The author emphasizes the relevance of further development of effective means of detecting fakes and improving methods of protection against such threats.

Keywords: deepfake, artificial intelligence, machine learning, information attack, information protection

Вступ

Одним з трендів, що розвивався в останні роки, є поширення дезінформації в мережі Інтернет. Цьому сприяв розвиток соціальних мереж, які дозволяють поширювати інформацію швидко та без валідації неупереджених осіб. У контексті використання штучного інтелекту в створенні дезінформації зазвичай вгадується термін «Deepfake». Під даним терміном мається на увазі низка методів для створення та/або маніпуляції аудіовізуальним контентом, в результаті чого утворюються відео з зображенням осіб, на яких вони роблять або говорять щось, чого ніколи не траплялось в дійсності. Такі маніпуляції можливі завдяки використанню засобів штучного інтелекту. Незважаючи на те, що відео, створені цим методом, ніколи не знімалися на камеру, для пересічного спостерігача складається враження, що вони автентичні.

Методи генерації Deepfake-контенту можуть використовуватись для різноманітних випадків незаконної діяльності як от шахрайство, поширення дезінформації та імітації особистості без її згоди. Через це існує потреба в методах, що можуть допомогти в виявленні Deepfake-контенту.

Більшість методів розпізнавання Deepfake, що існують на момент написання, є пасивними методами розпізнавання Deepfake-зображень, що використовують алгоритми глибинного навчання для знаходження неточностей в згенерованому контенті, на основі чого і визначається автентичність. Але проблемою таких методів є те, що інструменти для створення Deepfake

постійно розвиваються. Оскільки нові інструменти генерації не мають недоліків попередніх поколінь, інструменти розпізнавання з часом втрачають свою актуальність.

Для того, щоб зрозуміти, як покращити існуючі методи захисту і розпізнавання, необхідно структурно дослідити, як працюють методи створення Deepfake-контенту і які інструменти для цього існують. Метою даного дослідження є аналіз та порівняння методів створення Deepfake-контенту. Для досягнення мети дослідження необхідно дослідити, які моделі машинного навчання використовуються для генерації контенту, які принципи їх дії та які їхні недоліки можна використовувати для розпізнавання Deepfake-контенту.

Загальні відомості

Deepfake - це загальна назва для способів маніпуляції аудіовізуальним контентом, щоб зобразити дії певних осіб, які ті ніколи не вчиняли. Саме слово deepfake утворене комбінацією словосполучення “deep learning”, що позначає один з популярних методів використання штучних нейронних мереж, та “fake”, що в перекладі означає “підробка” [1].

Deepfake може використовуватись для таких злочинів як поширення дезінформації чи шантаж. Наприклад, в 2022 році в мережі поширювалось Deepfake-відео з президентом України Володимиром Зеленським, де він нібито закликав військовослужбовців “скласти зброю”. Відео було створено методами машинного навчання і в реальності зображених на відео подій не відбулося [2].

Техніки створення Deepfake діляться на чотири категорії:

- Заміна обличчя: ця техніка передбачає заміну обличчя однієї людини на інше на зображенні або відео;
- Перенесення жестикуляції: ця техніка передбачає передачу міміки та рухів однієї людини іншій.
- Маніпуляції з атрибутами: Цей метод дозволяє змінювати певні атрибути обличчя, наприклад, змінювати колір очей, додавати зморшки або змінювати вираз обличчя.
- Синхронізація по губах: Синхронізація губ передбачає синхронізацію звуку з рухами губ у відео.

Для генерації контенту використовуються три основні види моделей машинного навчання:

- GAN (від англ. Generative Adversarial Networks) складається з двох нейронних мереж - генератора та дискримінатора, які працюють у тандемі для створення та вдосконалення фейкового контенту. Генератор створює синтетичні зображення, а дискримінатор оцінює їхню автентичність, спонукаючи генератор постійно вдосконалювати свої результати. Результатом цього змагального процесу є високореалістичні глибокі підробки, які можуть переконливо відтворювати вираз обличчя, освітлення та інші складні деталі [3].
- Варіаційні автоенкодери (VAE): ще одна популярна модель генерації глибоких фейків. Вони складаються з кодера та декодера, які працюють разом для стиснення та реконструкції даних. VAE відомі своєю здатністю генерувати різноманітний синтетичний контент [4].
- Дифузійні моделі останнім часом привертають увагу своєю здатністю генерувати високоякісний синтетичний контент. Ці моделі працюють шляхом поступового додавання шуму до даних, а потім навчаються повертати процес шуму назад, щоб генерувати нові дані [5].

Найпоширеніші методи створення Deepfake:

- StyleGAN (Style-based Generative Adversarial Network) - це клас генеративних моделей, представлений компанією NVIDIA, призначений для синтезу фотореалістичних зображень високої роздільної здатності. На відміну від традиційних GAN, StyleGAN представляє нову архітектуру, де генератор отримує вхідні дані з окремого латентного простору (так званого W-простору) через мережу відображення, і застосовує отримані вектори стилю на різних рівнях мережі синтезу за допомогою адаптивної нормалізації екземплярів (AdaIN, від англ. adaptive instance normalization). Це дозволяє тонко керувати різними атрибутами зображення, такими як структура обличчя, текстура та освітлення.

Механізм багаторівневого управління StyleGAN дозволяє розділити візуальні особливості, що призводить до безпрецедентної якості та реалістичності у створенні синтетичних облич. Він широко використовується для створення фейкових зображень і аватарів завдяки своїй здатності генерувати нерозрізнені синтетичні ідентичності або змінювати реальний вміст обличчя. Доступність і висока точність роблять його центральним інструментом у створенні зловмисних підробок, що створює значні проблеми для систем верифікації особистості та автентичності контенту.

- DeepFaceLab - це набір інструментів для створення глибоких фейків з відкритим вихідним кодом, який широко використовується для заміни облич у відео. Він надає повний конвеєр для заміни облич за допомогою методів глибокого навчання, таких як автокодери та згорткові нейронні мережі. Основний робочий процес включає вилучення облич з вихідного та цільового відео, навчання моделі для вивчення рис обличчя, а потім накладання згенерованого обличчя на цільові кадри за допомогою таких методів, як уточнення маски та безшовне накладання зображень.

DeepFaceLab підтримує різні архітектури моделей (наприклад, SAE, H128, DF) і дозволяє тонко налаштовувати вираз обличчя, вирівнювання та освітлення для підвищення реалістичності. Простота використання і здатність генерувати високоякісні заміни облич зробили його одним з найпопулярніших інструментів для створення відео з глибоким фейком. Завдяки своїй практичній ефективності та широкій доступності DeepFaceLab відіграє важливу роль у поширенні фейкових загроз, включаючи маніпуляції з особистими даними, дезінформацію та атаки з використанням чужих осіб.

Висновки

У результаті проведеного дослідження було проаналізовано ключові методи створення Deepfake-контенту, їхні технічні особливості та можливості використання в шкідливих цілях. Поширення Deepfake-технологій створює суттєві загрози для інформаційної безпеки, приватності та довіри до цифрового контенту. Основними інструментами генерації фейкових зображень і відео є моделі глибинного навчання, зокрема GAN, VAE та дифузійні моделі, які дозволяють досягати високого рівня реалістичності створеного контенту. Особливу увагу слід звернути на поширеність інструментів, таких як StyleGAN та DeepFaceLab, що забезпечують доступність створення фейків навіть для користувачів без глибоких технічних знань.

Незважаючи на наявність методів виявлення Deepfake, постійне вдосконалення генеративних моделей знижує ефективність традиційних підходів. Тому актуальним завданням є розробка нових, більш стійких до еволюції атак технологій розпізнавання, зокрема активних методів захисту, що враховують особливості генеративних алгоритмів. Подальші дослідження мають бути спрямовані на інтеграцію методів цифрової перевірки достовірності, удосконалення алгоритмів детекції та створення систем превентивного контролю поширення шкідливого контенту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mirsky. Y., Lee. W The Creation and Detection of Deepfakes: A Survey : ACM Computing Surveys 54(1):1-41, 2021, 1 ст.
2. Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn. URL: <https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia> (дата звернення: 21.03.2025)
3. Mishra. A., Bharwaj. A., Batra. K. Deepfakes - Generating Synthetic Images, and Detecting Artificially Generated Fake Visuals Using Deep Learning : 14th International Conference on Cloud Computing, Data Science & Engineering, 2024
4. Battula Thirumaleshwari Devi, et al. A Comprehensive Survey on Deepfake Methods: Generation, Detection, and Applications . International Journal on Recent and Innovation Trends in Computing and Communication, 2024, 11(9), 654–678.

5. Croitoru. F., Hiji. A., Hondru. V., Ristea. N. Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook, 2024

Марчук Михайло Борисович - аспірант кафедри захисту інформації, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця, e-mail: dzgamech@gmail.com.

Науковий керівник: Лукічов Віталій Володимирович - доцент кафедри захисту інформації, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, Вінниця, e-mail: lukichov.vitalyi@vntu.edu.ua.

Marchuk Mykhalo M. - Postgraduate Student of the Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: dzgamech@gmail.com.

Scientific Supervisor: Lukichov Vitalyi V. - associate professor of the Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: lukichov.vitalyi@vntu.edu.ua.