УДК 004.8

М. Б. Марчук

В. В. Лукічов

МЕТОД ВІДСЛІДКОВУВАННЯ ЗОБРАЖЕНЬ ЩО ЗАСТОСОВУЮТЬСЯ ДЛЯ DEEPFAKE

Вінницький національний технічний університет

Анотація

У роботі було розроблено метод, що дозволяє відслідковувати зображення, які використовувались при створенні deepfake-контенту. Відслідковування відбувається за рахунок вбудовування прихованого повідомлення в зображення, яке не є видимим неозброєному людському оку, але розпізнається алгоритмом. Сам метод створено на базі методів стеганографії та глибинного машинного навчання. У результаті було розроблено архітектуру нейронної мережі, здатної приховувати та розпізнавати повідомлення на зображеннях обличь, що дозволяє відслідкувати оригінальне зображення у випадку використання його при створенні deepfake.

Ключові слова: діпфейк, стеганографія, машинне навчання, походження.

Abstract

The paper presents the method for tracking images that potentially can be utilized for creating deepfake content. The tracking is based on embedded secret messages which are unrecognizable for the human eye but detectable for the algorithm. The method is based on steganography and deep learning. As a result, the neural network architecture was developed, which can embed and retrieve secret messages on face images, which could be utilized for tracking original images that are utilized for deepfakes..

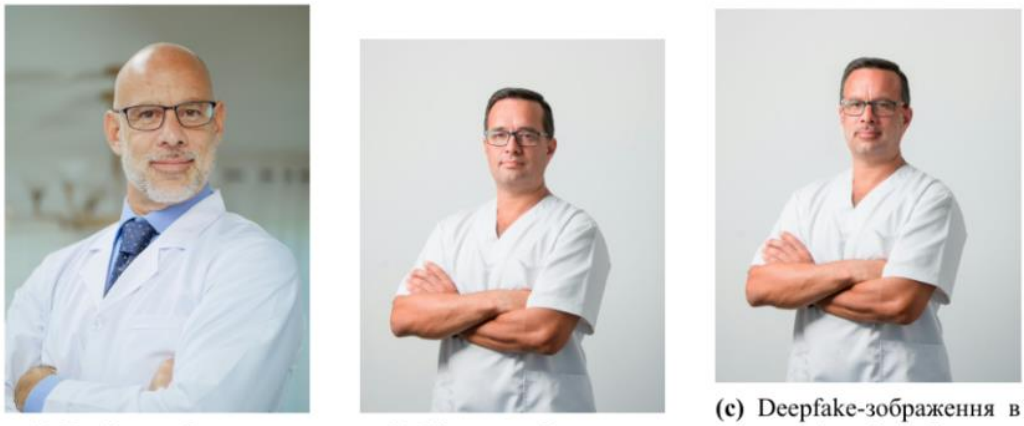
Keywords: deepfake, steganography, machine learning, provenance.

Вступ

На сьогоднішній день стан розвитку штучного інтелекту надає людству недоступні до цього можливості. Одним з способів застосування таких технологій є створення контенту в рази швидше й дешевше за традиційні інструменти, при цьому не поступаючись в якості. Генеративний штучний інтелект дозволяє створювати реалістичні зображення та відео без необхідності знімати справжні сцени на камеру і обробляти їх в графічних редакторах. Але хоч це і надає позитивні ефекти для креативності та здешевлення виготовлення медіа-контенту, такі технології несуть і негативні наслідки, адже завдяки їм спрощується створення дезінформації, організація інформаційних атак та шахрайство. Це підсилюється уставленим поширенням соціальних мереж, які дозволяють поширювати інформацію швидко та без валідації неупереджених осіб.

Одним з поширених способів створення контенту для дезінформації та шахрайства є зміна або заміна людських облич на зображеннях за допомого штучного інтелекту. Такі зображення з зміненими обличчями називаються deepfake [1].

Деякі дослідники працювали над методами розпізнавання deepfake. Зазвичай такі методи грунтуються на пасивних класифікаторах на базі штучних нейронних мереж, що навчались розрізняти автентичні зображення від deepfake [2]. Проблемою таких рішень є те, що оскільки моделі вчаться на згенерованих зображеннях з deepfake-моделей, що існували на момент тренування, модель не генералізується до зображень з deepfake-моделей, що з’явились пізніше (рис. 1).



(a) Вихідне зображення **(b)** Цільове зображення **(c)** Deepfake-зображення в результаті заміни обличчя

Рис 1. Приклади вихідного і цільового зображень та результату створення deepfake за допомогою них

Таким чином існує необхідність в розробці іншого виду захисту зображень, що дозволяє не залежати від прив'язки до конкретних методів генерації. Одним з таких видів захисту є вбудовування прихованих стегінографічних повідомлень в зображення. У випадку використання таких зображень для створення deepfake в згенерованому зображенні можна знайти приховане повідомлення і відслідкувати, яке зображення є оригінальним.

Результати дослідження

В рамках даної роботи було розроблено метод, який дозволяє приховувати повідомлення в вихідне зображенні обличчя. Метод вбудовує приховане стеганографічне повідомлення в зображення, яке можна відслідкувати. Повідомлення заховане в зображенні таким чином, що для людини його неможливо розрізнити візуально, але можна розпізнати алгоритмом. Для виконання даної задачі було розроблено архітектуру штучної нейронної мережі, що складається з декількох модулів.

Основна ідея даної архітектурної мережі полягає в тому, щоб видобувати репрезентації обличь з зображень та ховати повідомлення в них. Таким чином, якщо обличчя перенесеться з вихідного зображення на deepfake, на останньому репрезентація обличчя буде та ж сама, що і на оригінальному зображенні, що допоможе зберегтись повідомленю.

На вхід подається вихідні зображення та повідомлення. Припускається, що дане зображення може бути використане зловмисниками, щоб створити deepfake. Для даного методу в якості повідомлення використовується бінарна послідовність розміром 128 біт.

На початку повідомлення подається на вхід модуля Message Encoder. Даний модуль перетворює вхідне повідомлення з формату 128-бітної бінарної послідовності у внутрішнє представлення, що використовується іншими модулями.

Після цього внутрішнє представлення повідомлення та вихідне зображення подаються на вхід на Identity Encoder. Метою даного модуля є видобути репрезентацію обличчя з вихідного зображення, приховати повідомлення в цій репрезентації та повернути нову репрезентацію обличчя з прихованим повідомленням. Для видобування репрезентації обличчя використовується метод для розпізнавання обличь ArcFace [3]. Оскільки репрезентація обличчя є ненульовим 128-вимірним вектором, для оптимізації в якості функції втрат використовувався косинус подібності між репрезентацію вихідного зображення F_s та репрезентацією з прихованим повідомленням F_w:

$$cos\theta = \frac{F_S \cdot F_W}{|F_S||F_w|} \quad (1)$$

Отримана репрезентація з прихованим повідомленням разом з вихідним зображенням подається на модуль Watermark Encoder, метою якого є модифікувати вихідне зображення таким чином, щоб репрезентація обличчя на цьому зображенні відповідала репрезентації з прихованим повідомленням. Таким чином на виході отримується зображення з стеганографічно прихованим повідомленням (рисю 2).

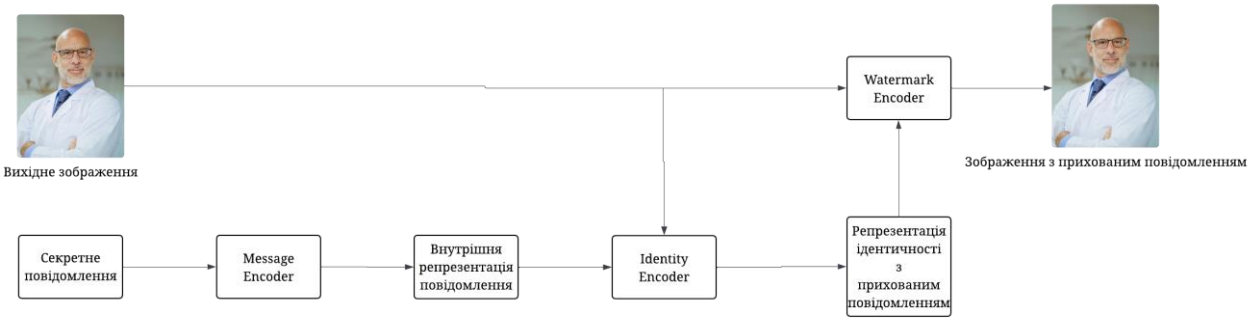


Рис 2. Архітектура штучної нейронної мережі для приховування повідомлення в зображенні

В подальшому отримане зображення з прихованим повідомленням замінює вихідне зображення для випадків публічного поширення (наприклад, в соціальних мережах). Припускається, що в такому середовищі зловмисники можуть взяти це зображення для створення з нього deepfake. Якщо це станеться, то з такого зображення можна буде витягнути приховане повідомлення, щоб довести, таке зображення є результатом маніпуляцій з іншим зображенням та знайти оригінал.

В межах даної роботи цю задачу вирішує модуль Watermark Decoder, який витягує повідомлення з зображення. Під час етапу тренування функцією втрат є бінарна перехресна ентропія між оригінальним повідомленням та результатом виконання Watermark Decoder (рис. 3).

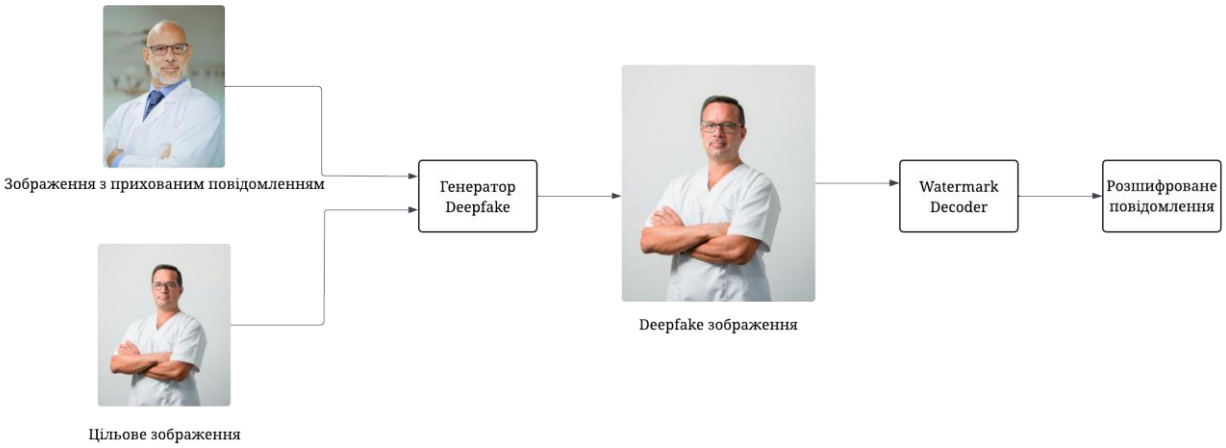


Рис 3. Архітектура штучної нейронної мережі для видобуваннння повідомлення в зображенні

Висновки

В результаті виконання даної роботи було розроблено метод, який дозволяє проактивно захищати зображення від зловмисного використання deepfake та відслідковувати їх у такому випадку. Розроблений алгоритм дозволяє вбудувати повідомлення у зображення з мінімальними візуальними змінами та з високою точністю його розпізнавання. Точність алгоритму може зменшуватись в умовах, якщо зображення зазнає різних модифікацій, як от компресія, розмиття чи зміна розміру. В цілому за результатами роботи підтверджено, що даний метод може застосовуватись для задач інформаційної безпеки, протидії дезінформації та захисту особистих медіаданих.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Марчук М. АНАЛІЗ МЕТОДІВ ТА ЗАСОБІВ АКТИВНОГО ЗАХИСТУ ВІД DEEPFAKE [Електронний ресурс] / М.Б. Марчук, В.В. Лукічов // Вісник Вінницького політехнічного інституту. – 2025. – Т. 3, Черв. 2025. – С. 126–132. – Режим доступу: https://doi.org/10.31649/1997-9266-2025-180-3-126-132. – Назва з екрана.
2. Mirsky Y. The Creation and Detection of Deepfakes: A Survey [Electronic resource] / Yisroel Mirsky, Wenke Lee // ACM Computing Surveys. – 2021. – Vol. 54. – P. 1–41. – Mode of access: https://doi.org/10.1145/3425780. – Title from screen.
3. ArcFace: Additive Angular Margin Loss for Deep Face Recognition [Electronic resource] / Jiankang Deng [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2021. – P. 1. – Mode of access: https://doi.org/10.1109/tpami.2021.3087709 (date of access: 21.03.2026). – Title from screen.

Марчук Михайло Борисович - аспірант кафедри захисту інформації, Вінницький національний технічний університет, Вінниця, email: 00-23-049.stud@vntu.vn.ua.

Лукічов Віталій Володимирович - доцент кафедри захисту інформації, Вінницький національний технічний університет, Вінниця, email: lukichov.vitalyi@vntu.edu.ua.

Mykhailo Marchuk – PhD student at Faculty of Information Security, Vinnytsia National Technical University, email - 00-23-049.stud@vntu.vn.ua.

Vitalii Luckichov – Associate Profesor at Faculty of Information Security, Vinnytsia National Technical University, email - lukichov.vitalyi@vntu.edu.ua.