

МАЗУРЕЦЬ ОЛЕКСАНДР

Хмельницький національний університет

<https://orcid.org/0000-0002-8900-0650>e-mail: exe.chong@gmail.com**ЖАРНОВСЬКИЙ ОЛЕКСАНДР**

Хмельницький національний університет

e-mail: stalker2v2@gmail.com**ГЛАДУН ОЛЕКСАНДР**

Хмельницький національний університет

e-mail: olexandrgladun@gmail.com**СОБКО ОЛЕНА**

Хмельницький національний університет

<https://orcid.org/0000-0001-5371-5788>e-mail: olena.sobko.ua@gmail.com

НЕЙРОМЕРЕЖЕВЕ ВИЯВЛЕННЯ ФЕЙКОВИХ ЗОБРАЖЕНЬ ЛЮДЕЙ

Запропоновано метод ідентифікації згенерованих штучним інтелектом фейкових зображень людей засобами машинного навчання. Метод ґрунтується на використанні комбінації двох згорткових нейронних мереж для автоматизованого аналізу зображень з метою ідентифікації автентичності та походження. Особливістю запропонованого методу є те, що він дозволяє не тільки ідентифікувати згенеровані штучним інтелектом зображення, але і метод їх генерації, що може значно покращити ідентифікацію згенерованих штучним інтелектом зображень. Особливістю запропонованого методу є те, що він дозволяє не тільки ідентифікувати згенеровані штучним інтелектом фейкові зображення, але і метод їх генерації, що може значно покращити ідентифікацію згенерованих штучним інтелектом зображень. Метод працює на основі перетворення вхідних даних у вигляді зображення, моделі для ідентифікації зображення, моделі для знаходження походження у вихідні дані у вигляді відсоткової оцінки автентичності зображення та його походження. На першому кроці зображення проходить через попередню обробку, що включає в себе перетворення, зміна розміру та створення тензору. Другим кроком потрібно завантажити натреновану мережу для ідентифікації зображення як згенерованого мережею чи реального, та проаналізувати завантажене зображення; в результаті отримуються відсоткові значення, що потрібно проаналізувати та вивести. Якщо вхідне зображення класифіковано як згенероване, то виконується третій крок і потрібно завантажити мережу для ідентифікації походження та додатково проаналізувати зображення. Архітектура нейронної мережі використовує розширену кількість вихідних шарів відповідно до методів генерації.

Для навчання мережі було підготовлено 17000 зображень людей та розроблений відповідний програмний додаток для дослідження ефективності. Встановлено, що використання тренувальної вибірки у відношенні класів згенерованих та реальних 1 до 1, збільшення розмірів груп, застосування випадкової аугментації та нормалізації сприяло значному покращенню кінцевих показників при менших витратах у часі. Це дозволило покращити максимальну точність класифікаторів до 92% для ідентифікації автентичності та 95% для встановлення методу генерації. В той же час збільшення коефіцієнту навчання дало негативний результат, а збільшення кількості епох не дало видимих результатів при значних збільшеннях витрат у часі. Одержані результати свідчать про спроможність запропонованого методу ефективно ідентифікувати згенеровані штучним інтелектом фейкові зображення засобами машинного навчання.

Ключові слова: згорткові нейронні мережі, класифікація зображень, фейкові зображення, штучний інтелект.

MAZURETS OLEKSANDR, ZHARNOVSKYI OLEKSANDR, HLADUN OLEKSANDR, SOBKO OLENA

Khmelnytskyi National University

NEURAL NETWORK DETECTION OF FAKE IMAGES OF PEOPLE

This work proposes method for identifying artificial intelligence-generated images of people using machine. The method is based on a combination of two convolutional neural networks for automated image analysis to identify authenticity and origin. The peculiarity of the proposed method is that it allows not only to identify artificial intelligence-generated images, but also the method of their generation, which can significantly improve the identification of artificial intelligence-generated images. The peculiarity of the proposed method is that it allows not only to identify artificial intelligence-generated images, but also the method of their generation, which can significantly improve the identification of artificial intelligence-generated images. The method works on the basis of converting input data in the form of an image, a model for image identification, a model for finding the origin into output data in the form of a percentage assessment of the authenticity of the image and its origin. In the first step, the image undergoes preprocessing, which includes transformation, resizing and tensor creation. In the second step, it is necessary to load a trained network to identify the image as generated by the network or real, and analyze the loaded image; as a result, percentage values are obtained that need to be analyzed and output. If the input image is classified as generated, then the third step is performed and it is necessary to load a network for identifying the origin and additionally analyze the image. The neural network architecture uses an extended number of output layers according to the generation methods.

Dataset that consist of 17,000 images of people was prepared for training the network and a corresponding software application was developed to study its effectiveness. It was established that the use of a training sample in the ratio of generated and real classes 1 to 1, increasing the size of groups, applying random augmentation and normalization contributed to a significant improvement in the final metrics with less time. This allowed improving the maximum accuracy of classifiers to 92% for identifying authenticity and 95% for establishing the generation method. At the same time, increasing the learning rate gave a negative result, and increasing the number of epochs did not give visible improvements with significant increases in time. The results obtained indicates the ability of the proposed method to effectively identify images generated by artificial intelligence using machine learning.

Аналіз предметної області та постановка задачі

Технологія генерації штучних зображень має широкий спектр застосувань, що робить її корисною в багатьох сферах людської діяльності для покращення креативності, продуктивності та часткової автоматизації [1], але в той же час, використання штучного інтелекту для генерації зображень дозволяє людям зловживати цією технологією для особистої вигоди.

Створення дипфейків є однією із головних проблем простого та загальнодоступного ШІ для генерації якісних зображень. Згенеровані штучним інтелектом зображення та неправдиві твердження можуть швидко поширитися у соціальних мережах, охопивши тисячі людей. Цю проблему ускладнюють алгоритми, які віддають перевагу сильному емоціональному вмісту, що є часто суперечливим або сенсаційним матеріалом, що призводить до дезінформації [2].

У статті пропонується метод ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання що ґрунтується на використанні комбінації двох згорткових нейронних мереж для автоматизованого аналізу зображень з метою ідентифікації автентичності та походження.

Останні публікації

Класифікація згенерованих штучним інтелектом зображень є досить складною задачею через комплексні методи генерації та варіативність. Щоб вирішити цю проблему у [3] розглядають покращення існуючих методів ідентифікації згенерованих штучним інтелектом зображень. Метод використовує розроблену згорткову нейронну мережу MobileNet-v2 для ідентифікації зображень обличчя людей. В результаті тестування було отримано точність у 72% та зазначено що модель має значні труднощі із зображеннями що містять аксесуари. Для покращення результатів були використані випадкові трансформації, зміна яскравості та контрасту, а також зміна параметрів мережі що дозволили покращити результат до 76%.

Метод розглянутий у [4] зосереджується на аналізі згенерованих штучним інтелектом робіт мистецтва. Розроблений метод використовує існуючі згорткові нейронні мережі VGG-19, ResNet-50 та ViT та створений датасет із зображень людей. В результаті тестування ефективності моделей отримані результати демонструють найкращий результат у мережі ViT з точністю в 97% на 50000 зображень розміром 200x200.

Інший підхід було розглянуто у [5], що використовують по-піксельний аналіз зображень використовуючи PRNN та ELA. Розроблений метод використовує власну архітектуру згорткової нейронної мережі та створений датасету із реальних зображень, та зображень згенерованих Dall E, Stable Diffusion та Open Art. Ефективність розробленої моделі визначали за допомогою метрик Accuracy, Precision, Recall, F1. Зазначається що використання ELA мали незначні відмінності метрик в порівнянні з PRNN, але значно швидше тренування на епоху.

Метою роботи є підвищення точності ідентифікації згенерованих штучним інтелектом зображень людей методами машинного навчання.

Основна частина

Метод ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання призначений для автоматизованої ідентифікації зображень людей згенерованих штучним інтелектом. Особливістю запропонованого методу є те, що він дозволяє не тільки ідентифікувати згенеровані штучним інтелектом зображення, але і метод їх генерації, що може значно покращити ідентифікацію згенерованих штучним інтелектом зображень.

Схема методу наведена на рис.1. Метод працює на основі перетворення вхідних даних у вигляді зображення, моделі для ідентифікації зображення, моделі для знаходження походження у вихідні дані у вигляді відсоткової оцінки автентичності зображення та його походження

На першому кроці зображення проходить через попередню обробку, що включає в себе перетворення, зміна розміру та створення тензору. Попередньо оброблене зображення буде використане навченою нейронною мережею для оцінки.

Другим кроком потрібно завантажити натреновану мережу для ідентифікації зображення як згенерованого мережею чи реального, та проаналізувати завантажене зображення. В результаті отримуються відсоткові значення що потрібно проаналізувати та вивести. Використана нейромережа для аналізу автентичності зображена схематично на рис.2. Архітектура мережі складається із вхідного шару, трьох шарів згортки, шару втрат та двох повноз'єднаних шарів. Використання шарів згортки добре підходить для задач класифікації зображень, проте значно збільшує витрати у часі.

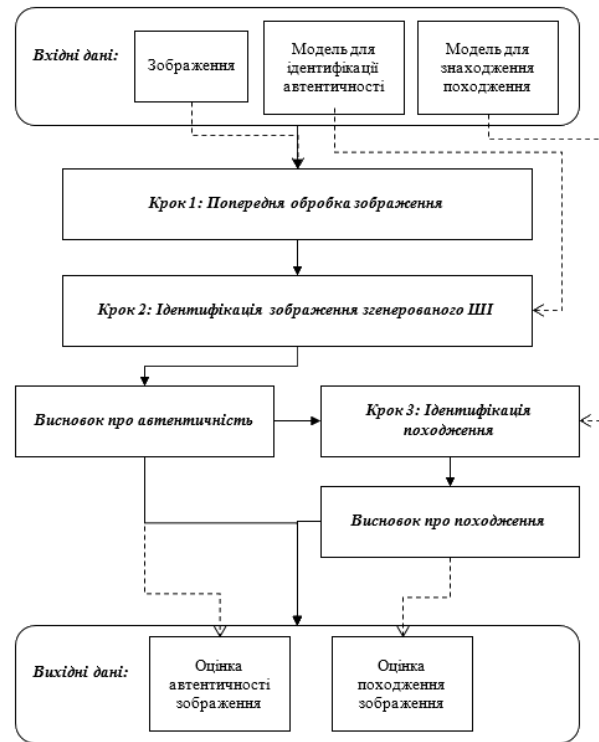


Рис. 1. Метод ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання

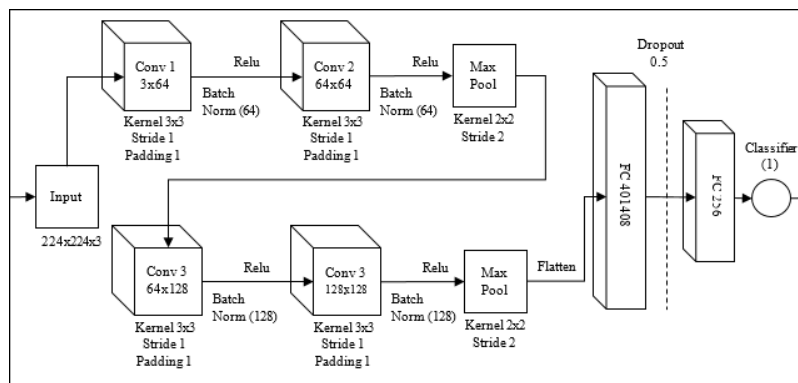


Рис. 2. Архітектура згорткової нейронної мережі для ідентифікації зображень людей

В залежності від того чи зображення реальне чи згенероване, виконується третій крок. Якщо вхідне зображення класифіковано як згенероване потрібно завантажити мережу для ідентифікації походження та додатково проаналізувати зображення. Архітектура мережі наведена на рис.3. Архітектура мережі використовує розширену кількість вихідних шарів відповідно до методів генерації.

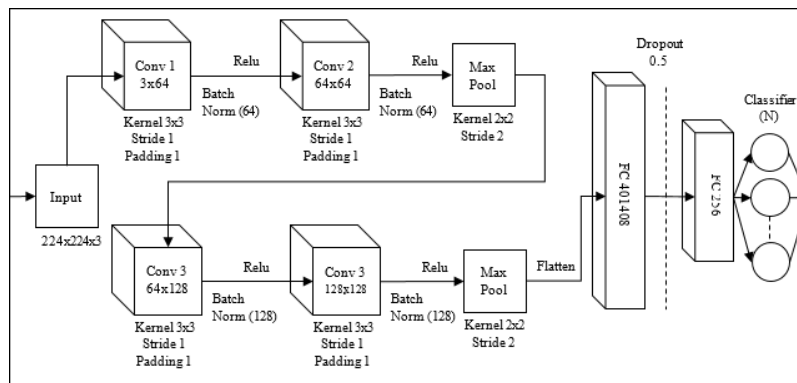


Рис. 3. Архітектура згорткової нейронної мережі для виявлення походження

Результатом аналізу зображення є відсоткова оцінка та віднесення його до одного з двох класів «реальне» чи «згенероване» для мережі по ідентифікації штучних зображень людей, та «Stable Diffusion», «StyleGan», «PhotoshopExperts» для мережі виявлення походження. Мережа може бути додатково налаштована для використання нових методів генерації.

Отже, запропонований метод ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання дозволяє виявляти штучні зображення, ґрунтуючись на використанні комбінації двох згорткових нейронних мереж для автоматизованого аналізу зображень з метою ідентифікації автентичності та походження

Дані дослідження

Для навчання розроблених нейромереж було сформовано набір даних із 17000 зображень. Зображення були взяті із Flickr-Faces-HQ Dataset (Nvidia) [6], 1 Million Fake Faces [7], Face Dataset Using Stable Diffusion v.1.4 [8], Real vs Fake Faces dataset. [9]. Сформований набір розподілений на два датасети.

Датасет для ідентифікації зображення, розподілений на «реальні» та «згенеровані» та складається із 14000 зображень, по 5000 на клас, а також 2000 для валідації та 2000 для тестування.

Датасет для виявлення походження розподілений на три класи «Stable Diffusion», «StyleGan», «PhotoshopExperts» та складається із 3000 зображень, по 800 зображень на клас для тренування, 300 для валідації та 300 для тестування. Зображення стандартизовано до розміру 224x224, нормалізовано та перетворено у тензор.

Таким чином був описаний набір даних, що буде використаний для навчання та тестування розроблених нейромереж.

Дослідження ефективності методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання

Для оцінки ефективності розробленого методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання був розроблений програмний додаток що складається із 4 модулів для взаємодії із користувачем написаних на мові C#, та двох функціональних модулів мови Python. Приклад роботи застосунку наведений на рис. 4.

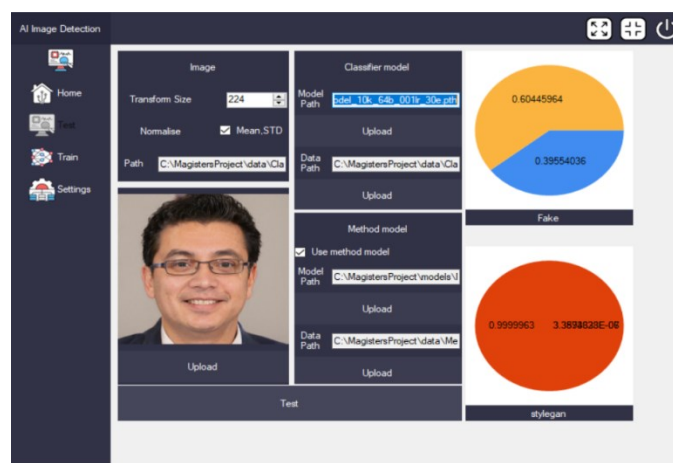


Рис. 4. Програмна реалізація методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання

В подальшому використовуючи функціональні модулі Python було виконано дослідження ефективності розробленого методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання.

Результати експерименту з дослідження ефективності розробленого методу

Експеримент було проведено використовуючи розроблені нейромережі та отриманий набір даних [10, 11]. Для оцінки ефективності методу були натреновані нейромережі з параметрами наведеними на таблиці 1.

Таблиця 1

		Параметри навчання нейромереж				
		Parameters				
ImageClassifier	Model_1_0	10	0.001	32	0.5	None
	Model_1_1	10	0.001	64	0.5	Mean, STD
	Model_1_2	10	0.005	64	0.5	Mean, STD
	Model_1_3	30	0.001	64	0.3	Mean, STD
MethodClassifier	Model_2_0	10	0.001	32	0.5	None
	Model_2_1	10	0.005	64	0.5	Mean, STD
	Model_2_2	20	0.001	64	0.5	Mean, STD

Для оцінки під час тренування та валідації були використані метрики Accuracy та Loss [12, 13]. Завдяки збільшенню розмірів групи, використанню нормалізації та відношенню кількості зображень класів 1 до 1, вдалося значно покращити показники точності з 60% до 70% (рис. 5). В той же час збільшення коефіцієнту навчання дало негативний результат в 53%. При збільшенні кількості епох до 30 для мережі ідентифікації зображення та до 20 для знаходження походження результат покращився до 95%. При послідовному збільшенні кількості епох не було отримано значних покращень при значних витратах у часі.

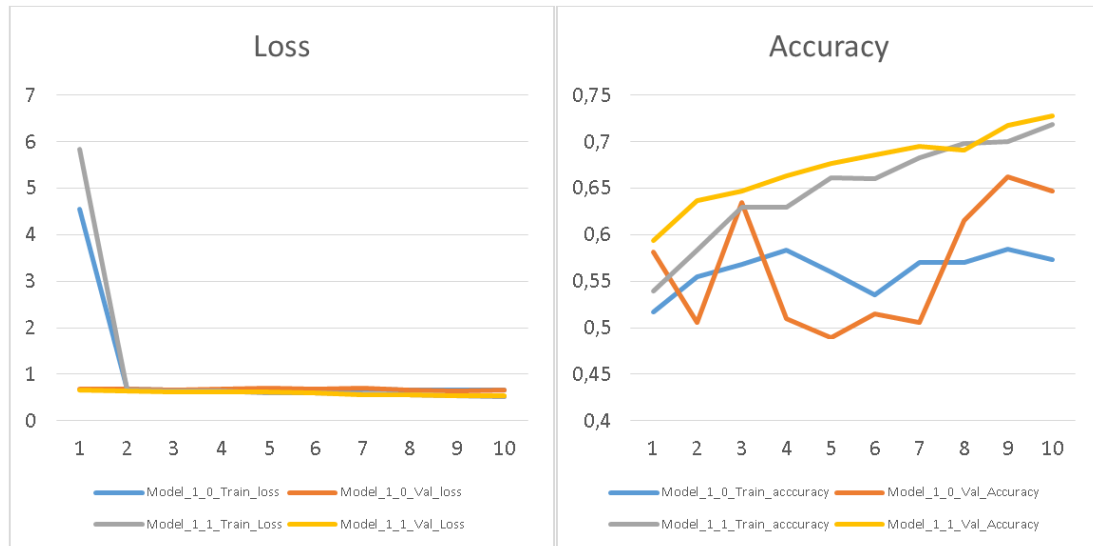


Рис. 5. Порівняння метрик мереж під час тренування

Для оцінки під час тестування були використані метрики Accuracy, Precision, Recall, F1. Результати тестування наведені на рис. 6.

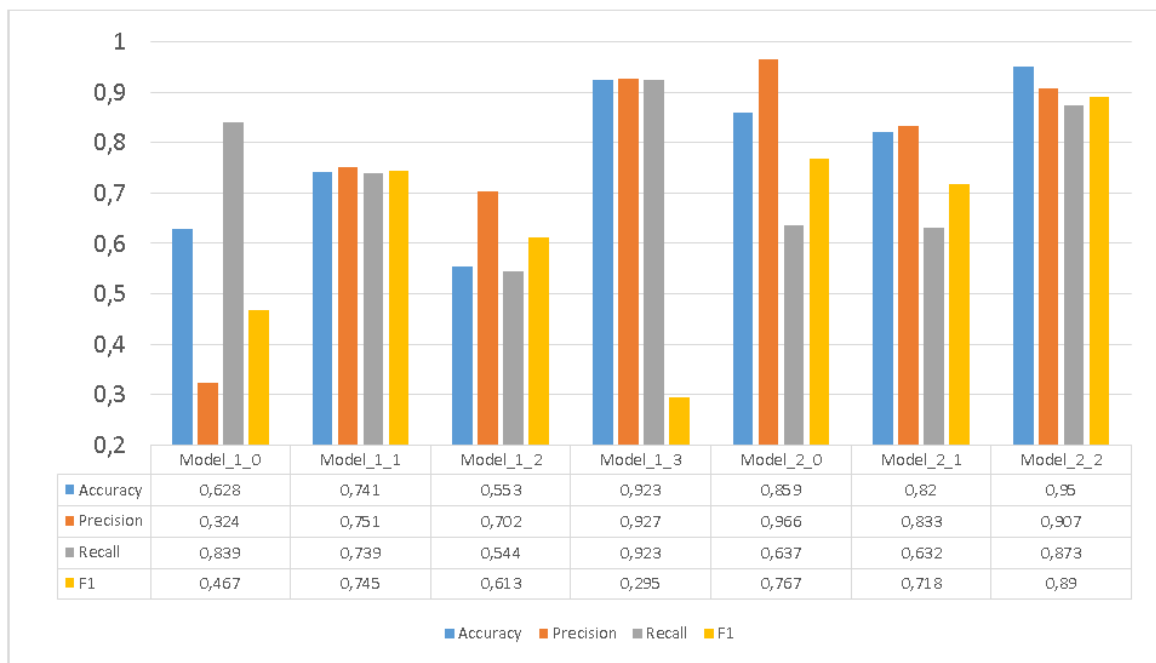


Рис. 6. Результати тестування

Як видно із рис. 6, збільшення кількості групи, застосування трансформацій, нормалізації та збільшення кількості епох мали позитивний ефект на результати під час тестування, а в свою чергу збільшення коефіцієнту навчання негативний. Найкращий отриманий параметр точності в 92% для мережі по ідентифікації зображення та 95% для виявлення походження.

Одержані результати свідчать про спроможність запропонованого методу ефективно ідентифікувати згенеровані штучним інтелектом зображення засобами машинного навчання. Ітерація параметрів навчання мережі дозволили значно покращити ефективність методу. Подальші дослідження будуть спрямовані на покращення архітектури нейромережі, а також на алгоритми тренування та тестування.

Висновки

У статті проведено аналіз сучасного стану сфери ідентифікації зображень людей згенерованих штучним інтелектом, виконаний аналіз та окреслено задачі для покращення та запропонований метод ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання, що ґрунтується на використанні комбінації двох згорткових нейронних мереж для автоматизованого аналізу зображень з метою ідентифікації автентичності та походження та дозволяє не тільки ідентифікувати згенеровані штучним інтелектом зображення, але і метод їх генерації, що може значно покращити ідентифікацію згенерованих штучним інтелектом зображень

Для навчання розроблених нейромереж було сформовано набір даних із 17000 зображень. Сформований набір складається із двох датасетів, для ідентифікації зображення, розподілених на «реальні» та «згенеровані», та датасет для виявлення походження розподілених на три класи «Stable Diffusion», «StyleGan», «PhotoshopExperts».

Для оцінки ефективності запропонованого методу було створено програмну реалізацію, яка складається із набору ноутбуків, реалізованих у хмарному сервісі «Google Colab» (для навчання нейромережевої моделі гібридної архітектури та для розширення отриманого набору даних методом аугментації тексту), та застосунку з графічним інтерфейсом користувача на мові Python.

Для оцінки ефективності розробленого методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання був розроблений програмний додаток що складається із 4 C# модулів для взаємодії із користувачем та двох функціональних модулів Python.

Із проведеного дослідження видно, що збільшення кількості групи, застосування трансформацій, нормалізації та збільшення кількості епох мали позитивний ефект на результати під час тестування, а в свою чергу збільшення коефіцієнту навчання негативний. Найкращий отриманий параметр точності в 92% для мережі по ідентифікації зображення та 95% для виявлення походження. Отримані результати свідчать про спроможність запропонованого методу ефективно ідентифікувати згенеровані штучним інтелектом зображення засобами машинного навчання. Ітерація параметрів навчання мережі дозволили значно покращити ефективність методу.

Література

1. How AI Image Generators Can Benefit You [Електронний ресурс] – 2024 – Режим доступу до ресурсу: <https://www.linkedin.com/pulse/how-ai-image-generators-can-benefit-you-imran-butt-frzff>.
2. Mapping the misuse of generative AI [Електронний ресурс] – 2024 – режим доступу до ресурсу: <https://deepmind.google/discover/blog/mapping-the-misuse-of-generative-ai/>.
3. Analyzing and Improving Existing Neural Network-Based Approaches to Identify AI Generated Images [Електронний ресурс] – 2024 – Режим доступу до ресурсу: <https://www.jsr.org/hs/index.php/path/article/view/5931>.
4. Identifying AI-Generated Art with Deep Learning [Електронний ресурс] – 2023 – Режим доступу до ресурсу: https://www.researchgate.net/publication/375116319_Identifying_AI-Generated_Art_with_Deep_Learning.
5. Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks [Електронний ресурс] – 2023 – Режим доступу до ресурсу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10674908/>.
6. Flickr-Faces-HQ Dataset (Nvidia) - Part 1 [Електронний ресурс] – Режим доступу до ресурсу: <https://www.kaggle.com/datasets/xhlulu/flickrfaceshq-dataset-nvidia-part-1>
7. 1 Million Fake Faces – 1 dataset [Електронний ресурс] – режим доступу до ресурсу: <https://www.kaggle.com/datasets/tunguz/1-million-fake-faces>
8. Face Dataset Using Stable Diffusion v.1.4 [Електронний ресурс] – Режим доступу до ресурсу: <https://www.kaggle.com/datasets/bwandowando/faces-dataset-using-stable-diffusion-v14>
9. Real vs Fake Faces dataset [Електронний ресурс] – Режим доступу до ресурсу: <https://www.kaggle.com/datasets/uditsharma72/real-vs-fake-faces>
10. Zharnovskiy O. Approach to Identification of Artificial Intelligence-Generated People Images by Means of Machine Learning / O. Zharnovskiy, O. Mazurets, O. Sobko // Key Aspects of the Development of Scientific Research in Modern Conditions: Proceedings of the XLV International Scientific and Practical Conference. – 2024. – Constanta, Romania. – Pp. 69–73.
11. Жарновський О.В. Практична реалізація методу ідентифікації згенерованих штучним інтелектом зображень людей засобами машинного навчання / О.В. Жарновський, Я.М. Казмірчук, О.В. Собко, О.В. Мазурець // Збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2024». – 2024. – Хмельницький. – С. 198–204.
12. Pokhytun A. Method for Neural Network Detecting Changed Images of People's Faces Using CNN / A. Pokhytun, O. Mazurets, M. Molchanova, O. Tyschenko // New Horizons in Scientific Research:

Challenges and Solutions: Proceedings of the 1st International Scientific and Practical Conference. – 2024. – Marseille, France. – Pp. 35–40.

13. Zharnovskiy O. Neural Network Method for Detection of Fake Document Images for Personality Identification Systems / O. Zharnovskiy, O. Mazurets, O. Sobko // Black Sea Science 2024: Proceedings of the International Competition of Student Scientific Works. – 2024. – Odesa: ONUT. – Pp. 434–448.

References

1. How AI Image Generators Can Benefit You [Elektronnyi resurs] – 2024 – Rezhym dostupu do resursu: <https://www.linkedin.com/pulse/how-ai-image-generators-can-benefit-you-imran-butt-frzff>.
2. Mapping the misuse of generative AI [Elektronnyi resurs] – 2024 – rezhym dostupu do resursu: <https://deepmind.google/discover/blog/mapping-the-misuse-of-generative-ai/>.
3. Analyzing and Improving Existing Neural Network-Based Approaches to Identify AI Generated Images [Elektronnyi resurs] – 2024 – Rezhym dostupu do resursu: <https://www.jsr.org/hs/index.php/path/article/view/5931>.
4. Identifying AI-Generated Art with Deep Learning [Elektronnyi resurs] – 2023 – Rezhym dostupu do resursu: https://www.researchgate.net/publication/375116319_Identifying_AI-Generated_Art_with_Deep_Learning.
5. Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks [Elektronnyi resurs] – 2023 – Rezhym dostupu do resursu: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10674908/>.
6. Flickr-Faces-HQ Dataset (Nvidia) - Part 1 [Elektronnyi resurs] – Rezhym dostupu do resursu: <https://www.kaggle.com/datasets/xhlulu/flickrfaceshq-dataset-nvidia-part-1>
- 1 Million Fake Faces – 1 dataset [Elektronnyi resurs] – rezhym dostupu do resursu: <https://www.kaggle.com/datasets/tunguz/1-million-fake-faces>
7. Face Dataset Using Stable Diffusion v.1.4 [Elektronnyi resurs] – Rezhym dostupu do resursu: <https://www.kaggle.com/datasets/bwandowando/faces-dataset-using-stable-diffusion-v14>
8. Real vs Fake Faces dataset [Elektronnyi resurs] – Rezhym dostupu do resursu: <https://www.kaggle.com/datasets/uditsharma72/real-vs-fake-faces>
9. Zharnovskiy O. Approach to Identification of Artificial Intelligence-Generated People Images by Means of Machine Learning / O. Zharnovskiy, O. Mazurets, O. Sobko // Key Aspects of the Development of Scientific Research in Modern Conditions: Proceedings of the XLV International Scientific and Practical Conference. – 2024. – Constanta, Romania. – Pp. 69–73.
10. Zharnovskiy O.V. Praktychna realizatsiia metodu identyfikatsii zghenerovanykh shtuchnym intelektom zobrazen liudei zasobamy mashynnoho navchannia / O.V. Zharnovskiy, Ya.M. Kazmirchuk, O.V. Sobko, O.V. Mazurets // Zbirnyk naukovykh prats za materialamy XVI Vseukrainskoi naukovo-praktychnoi konferentsii «Aktualni problemy kompiuternykh nauk APKN-2024». – 2024. – Khmelnytskyi. – S. 198–204.
11. Pokhytun A. Method for Neural Network Detecting Changed Images of Peoples Faces Using CNN / A. Pokhytun, O. Mazurets, M. Molchanova, O. Tyschenko // New Horizons in Scientific Research: Challenges and Solutions: Proceedings of the 1st International Scientific and Practical Conference. – 2024. – Marseille, France. – Pp. 35–40.
12. Zharnovskiy O. Neural Network Method for Detection of Fake Document Images for Personality Identification Systems / O. Zharnovskiy, O. Mazurets, O. Sobko // Black Sea Science 2024: Proceedings of the International Competition of Student Scientific Works. – 2024. – Odesa: ONUT. – Pp. 434–448.