

Дослідження ймовірнісних характеристик розпізнавання штучного контенту у цифрових зображеннях та відеозаписах класифікаторами на основі великих мовних моделей у режимі нульового зображення

Сергій Чорний

кандидат технічних наук, доцент, провідний науковий співробітник лабораторії інженерно-технічних досліджень, Національний науковий центр «Інститут судових експертиз ім. Засл. проф. М. С. Бокаріуса» Міністерства юстиції України, м. Харків, Україна,
ORCID: <http://orcid.org/0000-0002-0445-2886>, e-mail: svch1mail@yahoo.com

Ольга Брендель

доктор філософії в галузі права, завідувач сектору лабораторії інженерно-технічних досліджень, Національний науковий центр «Інститут судових експертиз ім. Засл. проф. М. С. Бокаріуса» Міністерства юстиції України, м. Харків, Україна,
ORCID: <https://orcid.org/0000-0002-1723-0203>, e-mail: olgabrenda@ukr.net

Досліджено ймовірності розпізнавання реальних і штучних зображень та відеозаписів за допомогою великих мовних моделей типу «зображення-текст» у режимі нульового зображення.

Ключові слова: класифікація зображень; БВММ; CNN; DeepFake; зображення-текст; нульове зображення.

Research on the Probabilistic Characteristics of Artificial Content Recognition in Digital Images and Videos by Classifiers Based on Large Language Models in Zero-Shot Image Mode

Sergiy Chorny, Olha Brendel

The probabilities of recognizing real and artificial images and videos using large image-to-text language models in the zero-shot image mode were investigated.

Keywords: image classification; MLLM; CNN; DeepFake; image-to-text; zero-shot image.

Використання типових класифікаторів штучного інтелекту (далі — ШІ) на основі конволютивних нейронних мереж (англ. *Convolutional Neural Network, CNN*) потребує значної кількості тестових зображень, їх анотацій і наявності потужних обчислювальних ресурсів для тренування відповідних моделей [1]. У разі підготовки до тренування (навчання) моделей щодо виявлення штучно генерованих

зображень або відеозаписів необхідні відповідні обчислювальні ресурси для генерування такого штучного тестового контенту. Водночас сьогодні є можливість застосовувати попередньо треновані моделі, що є у вільному доступі, як у вигляді моделей конволютивних нейронних мереж, так і багатомодальних великих мовних моделей (далі — БВММ; англ. *Multimodal Large Language Model, MLLM*) [2].

Метою цього дослідження є оцінка можливостей розпізнавання штучно генерованих зображень і відеозаписів, включно з *DeepFake*, за допомогою попередньо тренованих БВММ загального призначення типу «зображення-текст» (англ. *image-to-text*) у режимі «нульового зображення» (англ. *zero-shot image*). У цьому разі на вхід моделі надається дослідне зображення і текстовий запит у вигляді кількох взаємовиключних запитань, які обумовлюють відповідні класи, для яких розраховують імовірну належність оцінюваного зображення.

Для дослідження відібрано БВММ моделі класифікаторів типу зображення-текст із найвищим рейтингом у хмарному середовищі [3] і за рекомендацією від ШІ *ChatGPT* про найкращий класифікатор такого типу [4]. Кількість класів класифікатора БВММ, для яких розраховують імовірність розпізнавання, задано кількістю початкових взаємовиключних запитань. Дослідження проведено для запитань із двох і трьох класів.

Використано такі питання для двох класів запитань / відповідей БВММ: «*is it a real photograph*», «*is it an artificial image*». Питання для трьох класів запитань / відповідей БВММ були

такі: «*is it a real photograph*», «*is it an artificial image*», «*is it DeepFake or edited image*». Пороговий рівень імовірності правильного визначення відповідного класу для двох запитань становить понад 0.5, для трьох запитань — понад 0.3(3)3. Зображення вважають цілком належним до визначеного класу.

У дослідженні використано 1868 кадрів реального відеозапису з відеокamera дрона під час польоту над пошкодженими будівлями, 6055 кадрів відомого *DeepFake* відеозапису [5] і 1735 кадрів відеозаписів та окремих зображень, генерованих авторами за допомогою промпту *Slow flying in urban area over a few partially destroyed buildings* на ШІ-ресурсі [6]. Розрахунки проведено із застосуванням розробленого авторами програмного забезпечення на мові *Python* із включенням до нього попередньо тренованих моделей [3] і [4] (без їх зміни) на локальному комп'ютері з операційною системою *Windows 64*, *CPU i711370H*, *GPU RTX-3070*.

Узагальнені результати дослідження ймовірностей розпізнавання штучного й реального контенту для зображень і кадрів відеозаписів наведено в таблиці «Середні ймовірності у класах зображень».

Таблиця

Середні ймовірності класифікації зображень

Назва моделі	Класи	Тип зображення		
		Реальні	Штучні	Deep Fake
CLIP	2	0.43017	0.6042	0.8226
CLIP	3	0.3890	0.5075	0.5762
BLIP-base	2	0.5007	0.5016	0.5017
BLIP-base	3	0.3362	0.3356	0.3335
BLIP-large	2	0.5102	0.5467	0.4720
BLIP-large	3	0.2891	0.3379	0.3618

У дослідженнях для випадку двох класів запитань зображення *DeepFake* вважаємо належним до класу штучних зображень. Дослідження з відокремлення зображень типу *DeepFake* в окремий клас має сенс, оскільки такі зображення та відеозаписи можуть бути сумішшю реальних і редагованих фрагментів.

Аналіз здобутих результатів дав змогу дійти наведених далі висновків. Найкращі результати в обох досліджених випадках (2-й і 3-й класи) має БВММ *BLIP-base*, яка забезпечила правильне розпізнавання всіх типів зображень для заданих запитань. Моделі *CLIP* і *BLIP-large* добре розпізнають зображення зі штучним контентом, але гірше розрізняють реальні зображення.

Корисною особливістю використаних БВММ і їх перевагою перед *CNN* є можливість динамічного розширення оцінюваних класів шляхом зміни кількості вхідних текстових запитань для досліджуваних зображень, а також можливість уточнення (роз-

ширення) запитань шляхом унесення до них експертом додаткових (особливих для умов дослідження) ознак автентичності.

Розроблене програмне забезпечення дає змогу досліджувати як окремі зображення, так і відеозаписи на ресурсах автономного комп'ютера помірної потужності.

Перелік джерел посилання

1. Convolutional Neural Network. URL: https://en.wikipedia.org/wiki/Convolutional_neural_network (дата звернення: 10.04.2025).
2. Large Language Model. URL: https://en.wikipedia.org/wiki/Large_language_model (дата звернення: 10.04.2025).
3. Hugging Face. URL: <https://huggingface.co/Salesforce/blip-image-captioning-large> (дата звернення: 10.04.2025).
4. Mozuma-clip 1.0.1. URL: <https://pypi.org/project/mozuma-clip/> (дата звернення: 10.04.2025).
5. Dalí Lives – Art Meets Artificial Intelligence. URL: <https://www.youtube.com/watch?v=mPtcU9VmlIE> (дата звернення: 10.04.2025).
6. Genmo.ai. URL: <https://www.genmo.ai/play/creations> (дата звернення: 10.04.2025).