DDoS атак на кінцевого користувача // Кібербезпека: освіта, наука, техніка. – 2024. – № 2 (26). – С. 291–304. URL: https://doi.org/10.28925/2663-4023.2024.26.695

10. Аверічев І.М., Роженко А.С., Кихтенко Є.М. Інноваційні підходи до підвищення рівня кібербезпеки корпоративних мереж при використанні хмарних технологій // Кібербезпека: освіта, наука, техніка. – 2025. – № 1 (29). – С. 732–747. https://doi.org/10.28925/2663-4023.2025.29.934

Вплив характеристик навчального набору на коректність виявлення дипфейкових зображень моделлю ResNetSECBAM

УДК 004.056.5:004.93

Дмитро Азарний[1], Анатолій Давиденко[2], Олена Висоцька[3]

[1]Київський національний університет імені Тараса Шевченка, dazarny@gmail.com,
[2]Державний університет інформаційно-комунікаційних технологій, davidenkoan@gmail.com,
[3]Державний університет «Київський авіаційний інститут», Lek_Vys@ukr.net

Стрімкий розвиток генеративних моделей призвів до поширення дипфейкових зображень, що становлять загрозу інформаційній безпеці, цифровій ідентичності та довірі до мультимедійного контенту. Практика показує, що навіть високоточні детектори можуть істотно знижувати ймовірність коректної класифікації при переході до даних, відмінних від навчального набору. Тому актуальним є дослідження міждатасетного переносу моделей виявлення дипфейків для реального застосування систем комп'ютерного зору.

Метою роботи є дослідження впливу характеристик навчального набору даних на коректність виявлення дипфейкових зображень моделлю ResNetSECBAM. Наукова новизна полягає у двосторонньому порівнянні переносу між DFFD [1] та HiDF [2], а також у перевірці впливу їх об'єднання. Модель побудовано на базі ResNet-50 з механізмом уваги CBAM [3], що посилює інформативні канальні та просторові ознаки.

Було проведено три експерименти: навчання на DFFD з тестуванням на HiDF, навчання на HiDF з тестуванням на DFFD та навчання на об'єднаному наборі DFFD+HiDF. За підсумковим розбиттям використано 62 260 зображень DFFD (50 638 навчальних, 5 626 валідаційних, 5 996 тестових) та 69 828 зображень HiDF (48 879 навчальних, 6 982 валідаційних, 13 967 тестових), тобто 132 088 зображень загалом. У третьому експерименті навчальна вибірка становила 99 517 зображень (50 638 DFFD і 48 879 HiDF), валідаційна — 12 608 (5 626 DFFD і 6 982 HiDF), а тестова — 19 963 (5 996 DFFD і 13 967 HiDF); усі тестові зображення були відкладеними і не використовувалися під час навчання. Оцінювання проводилося за метриками Accuracy, Precision, Recall, F1-міра та AUC.

Таблиця 1

Підсумкові результати тестування ResNetSECBAM

Показник класифікації	Навчання DFFD, тестування HiDF	Навчання HiDF, тестування DFFD	Навчання DFFD+HiDF, тестування DFFD+HiDF
Точність (Accuracy)	0.7428	0.2218	0.9845
Влучність (Precision)	0.7969	0.1761	0.9805
Повнота (Recall)	0.7171	0.9980	0.9842
F1-міра	0.7549	0.2994	0.9823
Площа під ROC-кривою (AUC)	0.8320	0.7538	0.9968

Результати експериментів відображено в табл. 1 та на рис. 1.

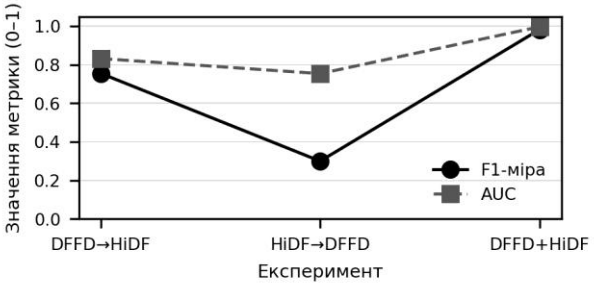

Рис. 1. Порівняння F1-міри та AUC у трьох експериментах

На основі аналізу результатів проведених експериментів можна зробити наступні висновки:

1) навчання лише на DFFD забезпечило помірний перенос на HiDF: F1-міра становила 0.7549, AUC — 0.8320, а кількість хибних спрацювань і пропущених підробок дорівнювала 1410 та 2183 відповідно;
2) у зворотному напрямі перенос виявився слабшим: F1-міра дорівнювала 0.2994; модель пропустила лише 2 підробки, проте сформувала 4664 хибні спрацювання, тобто масово відносила реальні зображення до класу fake;
3) найкращий результат отримано у третьому експерименті: F1-міра становила 0.9823, AUC — 0.9968, кількість хибних спрацювань — 171, а пропущених підробок — 138.

Отримані результати свідчать, що різноманітність навчальних прикладів є важливим чинником формування узагальнених ознак. Імовірно, кращий перенос моделі, навченої на DFFD [1], частково пов’язаний з кількома типами генерації у цьому наборі. Водночас на результат впливають і джерела реальних зображень, тобто походження оригінальних фото, умови їх отримання,

роздільна здатність і попередня обробка, які можуть змінювати статистику класу real.

Таким чином, високі валідаційні метрики всередині одного набору можуть створювати хибне уявлення про практичну придатність детектора. Об'єднання різнорідних джерел даних істотно підвищує коректність виявлення дипфейкових зображень моделлю ResNetSECBAM і може бути основою для подальших досліджень стійких систем детекції.

1. Dang H., Liu F., Stehouwer J., Liu X., Jain A. K. DFFD: Diverse Fake Face Dataset. URL: https://cvlab.cse.msu.edu/dffd-diverse-fake-face-dataset.html.
2. Kang C., Jeong S., Lee J. та ін. HiDF: A Human-Indistinguishable Deepfake Dataset // Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. – 2025. – P. 5527–5538. – DOI: 10.1145/3711896.3737399.
3. Woo S., Park J., Lee J.-Y., Kweon I. S. CBAM: Convolutional Block Attention Module // Proceedings of the European Conference on Computer Vision (ECCV). – 2018. – P. 3–19. – DOI: 10.1007/978-3-030-01234-2_1.

Модель Claude Mythos та кібербезпека: загрози та виклики

УДК 004.8:004.056

Олег Ясній[1], Любов Цимбалюк[2], Анна Турчманович[3]

Тернопільський національний технічний університет імені Івана Пулюя ,
[1]oleh.yasniy@gmail.com, [2]lubovtsymbaliuk@gmail.com,
[3]turchmanovich101@gmail.com

Claude Mythos — одна з найпотужніших сучасних моделей ШІ від Anthropic, що має значний вплив на кібербезпеку [1, 2, 3]. Однак вона не єдина загроза: інші передові моделі, зокрема GPT-5.4 Cyber (OpenAI) та Big Sleep (Google), також володіють подібними можливостями. Ера атак із використанням ШІ настала, і організації не можуть залишатися лише реактивними [1].

Мета даної роботи – проаналізувати систему ШІ Claude Mythos та виявити основні загрози та виклики, які виникають перед бізнесом.

Багато компаній роками недофінансовували кібербезпеку, оскільки ради директорів і керівництво не надавали їй пріоритету. Це створило приховані слабкі місця, які ШІ-інструменти швидко виявляють; для частини бізнесів наслідки можуть бути критичними. Особливо вразливі галузі з розгалуженими операційними технологіями — енергетика, комунальні послуги, виробництво, водопостачання, транспорт — де багато систем працюють десятиліттями й не підлягають ефективному патчингу. Усунення інвестиційного розриву вимагатиме значного нарощення витрат, тоді як більшість організацій планує лише помірне зростання бюджетів [1].

Mythos створювався як інструмент для роботи з великими кодовими базами й автоматизації розробки, але саме ці можливості роблять його потенційно небезпечним: модель може виявляти приховані недоліки, поєднувати дрібні вразливості в складні атаки, відновлювати вихідний код із розгорнутого ПЗ і,