МЕТОДИ АВТОМАТИЧНОГО ВИЯВЛЕННЯ НЕЗАКОННОГО ВІДЕОКОНТЕНТУ

Кібірєв Д.О., Федорченко В.М.
Харківський національний університет радіоелектроніки Харків, Україна

У сучасну цифрову епоху відеоконтент став одним із найбільш поширених способів передачі інформації. Однак з розвитком технологій стрімко зростає і кількість випадків незаконного розповсюдження відео - від порушення авторських прав до публікації заборонених або шкідливих матеріалів [1]. Це створює потребу у впровадженні автоматизованих методів виявлення незаконного контенту, які працюють в режимі реального часу, здатні аналізувати великі обсяги даних і інтегруватися в системи модерації та безпеки.

Під незаконним відеоконтентом розуміється будь-який відеоматеріал, розповсюдження або зберігання якого заборонене законом чи порушує інтелектуальні, етичні або безпекові норми.

Метою доповіді є аналіз сучасних методів автоматичного виявлення незаконного відеоконтенту.

Метод перцептивного гешування (Perceptual Hashing) дозволяє створити унікальний "відбиток" відео на основі його візуальних характеристик. Він ефективний навіть після стиснення або часткових змін відео. Найбільш відомі алгоритми pHash, aHash, dHash. Переваги даного методу: стійкість до редагування відео; висока швидкість обробки; простота інтеграції в існуючі системи. В якості недоліків можна виділити те, що не завжди метод точний при великих модифікаціях (наприклад, поворот, інверсія кольорів), потребує великої бази гешів для високої ефективності та може давати хибнопозитивні/хибнонегативні результати без оптимізації.

Інший метод оснований на аналізі цифрових водяних знаків (ЦВЗ). Суть його в тому, що правовласники вбудовують у відео непомітні цифрові позначення [2]. Системи автоматично виявляють такі знаки, ідентифікуючи джерело нелегального копіювання. Переваги використання ЦВЗ: висока стійкість до типових атак: масштабування, обтинання, стиснення; підтримка трасування витоків (відстеження по джерелу); можливість юридичного підтвердження прав на контент. Метод може застосовуватись в поєднанні з ШІ для масової автоматичної перевірки та дає змогу автоматично блокувати або сповіщати про незаконний контент. Недоліками є те, що не всі методи водяного знака є достатньо стійкими, вимагає впровадження на етапі створення або розповсюдження контенту, не завжди можливо виявити водяний знак, якщо відео зазнало глибокого перекодування, а також потребує спеціального програмного забезпечення для вбудовування/зчитування.

Метод машинного навчання (ML) і глибокого навчання (DL) є потужними інструментами для автоматичного виявлення незаконного відеоконтенту, особливо в умовах великого обсягу даних, які щоденно з'являються в Інтернеті. Глибоке навчання використовує нейронні мережі, які можуть

автоматично виділяти ознаки з відеоданих без ручної обробки. Ці підходи дозволяють будувати адаптивні системи, здатні навчатися з даних, виявляти складні шаблони, розпізнавати відео з порушенням авторських прав або неприйнятним вмістом. На даний час нейронні мережі навчаються класифікувати контент за допомогою аналізу зображення, звуку, тексту (субтитри) тощо. Застосовують мережі CNN для аналізу кадрів, RNN або Transformer для аналізу звуку/мовлення, Multimodal networks для аналізу одразу кількох типів даних. Переваги методів ML/DL: висока точність в умовах великої кількості відео, стійкість до обфускацій, змін формату, редагування, можливість навчати моделі під нові загрози, підтримка реального часу та масштабування. В якості недоліки можна вказати необхідність великих обсягів навчальних даних, високі обчислювальні витрати, чутливість до перенавчання або недостатнього навчання та ускладненість пояснення результатів DL-моделей (black-box).

Метод семантичного аналізу змісту оснований на використанні NLP та комп'ютерного зору для виявлення заборонених фраз, символіки, об'єктів (зброя, насильство) в кадрі.

Також можуть використовуватись системи розпізнавання обличчя на основі автоматичного виявлення конкретних осіб, які фігурують у заборонених матеріалах або порушують конфіденційність.

На даний час основними проблемами в застосуванні методів автоматичного виявлення незаконного відеоконтенту є:

– обфускація контенту, тобто змінення формату, фільтрів, розміру для обходу автоматичного виявлення;

– похибки класифікації, які призводять до хибного спрацьовування або пропуску шкідливого відео;

– морально-правова дилема, тобто необхідність знаходження балансу між контролем і свободою слова;

– обчислювальна складність методів, що потребує потужних ресурсів для обробки великих відеопотоків.

Пропонується використовувати гібрідний метод, заснований на аналізі цифрових водяних знаків () з використанням методів машинного (ML) і глибокого навчання (DL).

Методи автоматичного виявлення незаконного відеоконтенту є критично важливими для захисту цифрового простору. Поєднання штучного інтелекту, машинного навчання та цифрового маркування (ЦВЗ) дозволяє створити надійні механізми боротьби з нелегальним відео.

Список літератури

1. Кібірєв, Д.О., Федорченко В.М. Аналіз методів захисту відеоконтенту від несанкціонованого копіювання. ХНУРЕ, 2023.

2. Martovytskyi V., Ruban I., Bolohova N., Sievierinov O., Zhurylo O., Permiakov O., Nosyk A., Nepokrytov D., Krylenko I. (2021). Development of Methods for Generation of Digital Watermarks Resistant to Distortion. Eastern-European Journal of Enterprise Technologies, 6(2), 114.