УДК 004.8

М. Б. Марчук

В. В. Лукічов

АНАЛІЗ АТАК НА СТЕГАНОГРАФІЧНІ АЛГОРИТМИ ТА ЦИФРОВІ ВОДЯНІ ЗНАКИ

Вінницький національний технічний університет

Анотація

У роботі систематизовано та проаналізовано основні класи атак на стеганографічні алгоритми та системи цифрових водяних знаків (watermarking). Розглянуто методи стеганоаналізу — від статистичних підходів до нейромережевих детекторів, а також класифікацію атак на водяні знаки за категоріями видалення, геометричних, криптографічних і протокольних атак. Окрему увагу приділено сучасним загрозам, пов'язаним із застосуванням генеративного штучного інтелекту для видалення водяних знаків із згенерованих зображень.

Ключові слова: цифрові водяні знаки, стеганографія, приховування інформації, стегоаналіз, захист інформації.

Abstract

The paper systematizes and analyzes the main classes of attacks on steganographic algorithms and digital watermarking systems. Steganalysis methods are reviewed — from statistical approaches to neural-network detectors — as well as the classification of watermarking attacks into removal, geometric, cryptographic, and protocol categories. Special attention is paid to modern threats associated with the use of generative artificial intelligence for removing watermarks from generated images.

Keywords: digital watermarking, steganography, concealment of information, steganalysis, information security.

Вступ

Стеганографія та цифрові водяні знаки належать до технологій приховування інформації (information hiding), проте розв'язують різні задачі: стеганографія забезпечує прихований канал зв'язку, тоді як водяні знаки призначені для захисту авторських прав, підтвердження цілісності та встановлення походження цифрового контенту. Надійність обох технологій оцінюється не лише за непомітністю та місткістю вбудовування, а насамперед за стійкістю до атак. Саме аналіз атак (для стеганографії — стеганоаналіз) визначає реальний рівень безпеки системи, оскільки виявляє межі, у яких приховане або захисне маркування лишається непомітним і незруйнованим.

Актуальність теми посилюється стрімким поширенням генеративного штучного інтелекту та сервісів автоматичного маркування контенту (зокрема SynthID і Stable Signature), що зробило вразливості систем приховування інформації предметом активних сучасних досліджень. Метою роботи є систематизація та аналіз основних класів атак на стеганографічні алгоритми та системи цифрових водяних знаків, а також окреслення сучасних викликів, пов'язаних із застосуванням генеративних моделей.

Атаки на стеганографічні алгоритми (стеганоаналіз)

Стеганоаналіз — це сукупність методів виявлення самого факту прихованої передачі даних, а в розширеному сенсі — оцінювання довжини, вилучення чи руйнування прихованого повідомлення. За рівнем інформації, доступної зловмиснику, атаки класифікують на атаку лише за стего-об'єктом (stego-only), атаку з відомим контейнером (known-cover), атаку з відомим повідомленням та атаку з обраним повідомленням (chosen-message).

Історично першими стали статистичні атаки. Класична робота А. Вестфельда та А. Пфіцмана (1999) запропонувала χ^2-атаку (хі-квадрат), яка виявляє характерне вирівнювання частот пар значень яскравості (PoV — pairs of values), що виникає внаслідок вбудовування у молодші біти (LSB), і успішно зламувала поширені на той час інструменти EzStego, Jsteg, Steganos та S-Tools [1]. Подальшого розвитку набули RS-аналіз (Regular-Singular) та метод пар вибірок (sample pair analysis), які дають змогу не лише виявити, а й оцінити довжину вбудованого повідомлення; ці підходи стали еталонними універсальними статистичними атаками [2, 3]. Окремий клас становлять цільові атаки на конкретні алгоритми, зокрема методи злому F5 та OutGuess.

Наступним поколінням стали методи на основі ознак (feature-based), у яких із зображення обчислюється набір статистичних дескрипторів (SPAM, SRM — Spatial Rich Models, maxSRMd2), що подаються на ансамблевий класифікатор. Такий «дворівневий» підхід (ручні ознаки та окремий класифікатор) тривалий час був стандартом стеганоаналізу зображень [2].

Найсучаснішим напрямом є нейромережевий стеганоаналіз на основі згорткових нейронних мереж (CNN), де виділення ознак і класифікація поєднані в єдиній моделі. Послідовність архітектур — GNCNN (Qian, 2015), Xu-Net (2016), Ye-Net (2017), Yedroudj-Net (2018), SRNet (2019), Zhu-Net та GBRAS-Net — поступово підвищувала точність виявлення, перевершивши класичні методи на основі rich-моделей [4]. Ці мережі автоматично навчаються розпізнавати слабкий стего-сигнал у шумовому залишку зображення. Водночас дослідження стійкості показують, що нейромережеві детектори чутливі до елементарних трансформацій зображення — масштабування, стиснення, обрізання та додавання шуму, що знижує їхню узагальнювальну здатність у реальних умовах і відкриває простір для атак ухилення [5].

Атаки на системи цифрових водяних знаків (watermarking)

Атаки на системи watermarking традиційно поділяють на чотири основні категорії: атаки видалення (removal), геометричні (geometric), криптографічні (cryptographic) та протокольні (protocol) [6].

Атаки видалення спрямовані на знищення або послаблення вбудованого знака без помітного погіршення якості контенту. До них належать лінійна та нелінійна (зокрема медіанна) фільтрація, шумозаглушення (denoising), додавання шуму (гаусового, типу «сіль-перець»), стиснення з втратами (JPEG) та різноманітні операції згладжування й підвищення різкості. Колузійні атаки (collusion) є окремим різновидом, що використовує кілька копій контенту з різними знаками для відновлення «чистої» копії; вони становлять серйозну загрозу для систем цифрового відбитка (fingerprinting).

Геометричні атаки не видаляють знак безпосередньо, а порушують синхронізацію детектора з вбудованою інформацією шляхом геометричних перетворень — поворотів, масштабування, зсувів, обрізання та зрізування. Стандартним інструментом оцінювання стійкості до таких спотворень слугує бенчмарк StirMark.

Криптографічні атаки спрямовані на пошук секретного ключа (атака повного перебору, brute force) або використовують детектор як оракул (oracle attack), послідовно змінюючи контент доти, доки знак не перестане виявлятися. Протокольні атаки спрямовані проти самої концепції застосування: атака інвертування (invertibility) ґрунтується на тому, що зловмисник «віднімає» власний фіктивний знак і оголошує себе власником, звідки випливає вимога невіднімності (non-invertibility) знака, а атака копіювання (copy attack) переносить оцінений знак з одного контенту на інший без знання ключа [6].

Сучасні виклики: атаки на нейромережеві знаки та контент генеративного ШІ

Поширення генеративних моделей (дифузійних моделей та генеративно-змагальних мереж) спричинило появу як нових схем watermarking (HiDDeN, Stable Signature, Tree-Ring, SynthID, Gaussian Shading), так і потужних атак на них. Окремий клас становлять регенераційні атаки (regeneration attacks): зображення спершу зашумлюється, а потім відновлюється генеративною моделлю. Показано, що невидимі піксельні водяні знаки можуть бути доказово видалені за допомогою генеративного штучного інтелекту, що ставить під сумнів надійність значної частини наявних схем [7]. Адаптивні атаки з оптимізацією, чорноскриньові атаки та методи підробки (forging) знака додатково підривають стійкість сучасних систем [9].

Важливим є й теоретичний результат: робота «Watermarks in the Sand» доводить неможливість «сильного» watermarking за наявності у зловмисника оракулів якості та збурення, що окреслює

фундаментальні межі стійкості будь-яких схем маркування для генеративних моделей [8]. Для систематичного й відтворюваного оцінювання надійності розроблено спеціалізовані бенчмарки (StirMark, WAVES, W-Bench), які дають змогу порівнювати схеми за єдиними наборами атак [9].

Висновки

Систематизовано основні класи атак на стеганографічні алгоритми та системи цифрових водяних знаків. Показано, що стеганоаналіз пройшов еволюцію від статистичних χ^2- та RS-методів через ознакові підходи (SRM) до нейромережевих детекторів, які наразі забезпечують найвищу точність, але лишаються вразливими до елементарних трансформацій зображення. Атаки на watermarking систематизовано за чотирма категоріями (видалення, геометричні, криптографічні, протокольні), доповненими колузійними та сучасними регенераційними атаками на основі генеративного штучного інтелекту.

Зроблено висновок, що з поширенням генеративного штучного інтелекту проблема стійкості водяних знаків набуває як прикладної, так і фундаментальної значущості, а проєктування захищених схем потребує врахування всього спектра розглянутих атак та використання стандартизованих бенчмарків для об'єктивного оцінювання якості.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Westfeld A., Pfitzmann A. Attacks on Steganographic Systems: Breaking the Steganographic Utilities EzStego, Jsteg, Steganos, and S-Tools — and Some Lessons Learned // Information Hiding. Lecture Notes in Computer Science, vol. 1768. Berlin, Heidelberg : Springer, 2000. P. 61–76.
2. Li B., He J., Huang J., Shi Y. Q. A Survey on Image Steganography and Steganalysis // Journal of Information Hiding and Multimedia Signal Processing. 2011. Vol. 2, No. 2. P. 142–172. Режим доступу: https://bit.kuas.edu.tw/~jihmsp/2011/vol2/JIH-MSP-2011-03-005.pdf
3. Image steganography techniques for resisting statistical steganalysis attacks: A systematic literature review [Електронний ресурс] // PLOS ONE. 2024. Режим доступу: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0308807
4. Sensitivity of deep learning applied to spatial image steganalysis [Електронний ресурс] // PeerJ Computer Science. 2021. Режим доступу: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8444093/
5. Deep learning for steganalysis: evaluating model robustness against image transformations [Електронний ресурс] // Frontiers in Artificial Intelligence. 2025. Режим доступу: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1532895/full
6. A Comprehensive Review on Digital Image Watermarking [Електронний ресурс]. arXiv:2207.06909. 2022. Режим доступу: https://arxiv.org/abs/2207.06909
7. Zhao X., Zhang K., Su Z. et al. Invisible Image Watermarks Are Provably Removable Using Generative AI // Advances in Neural Information Processing Systems (NeurIPS). 2024. Режим доступу: https://arxiv.org/abs/2306.01953
8. Zhang H., Edelman B. L., Francati D. et al. Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models // International Conference on Machine Learning (ICML). 2024. Режим доступу: https://arxiv.org/abs/2311.04378
9. Secure and Robust Watermarking for AI-generated Images: A Comprehensive Survey [Електронний ресурс]. arXiv:2510.02384. 2025. Режим доступу: https://arxiv.org/abs/2510.02384

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.