DOI 10.15276/imms.v11.no3.191
УДК 004.056.5

Informatics and Mathematical Methods in Simulation
Vol. 11 (2021), No. 3, pp. 191-199

РОЗРОБКА МЕТОДУ ДЛЯ ПІДВИЩЕННЯ СТІЙКОСТІ ДО СТАТИСТИЧНОГО СТЕГАНОАНАЛІЗУ З ВИКОРИСТАННЯМ ГЕНЕРАТИВНО-ЗМАГАЛЬНИХ МЕРЕЖ

Ю.І. Даракчі, Н.І. Кушніренко, О.Є. Плачінда

Національний університет «Одеська Політехніка», просп. Шевченка, 1,
Одеса, 65044, Україна; e-mail: infsec2011@gmail.com

В сучасних умовах постійного розвитку обчислювальної техніки та мережевих технологій загострюється питання захисту інформації від несанкціонованого доступу. Одним із найефективніших інструментів для вирішення цього питання є стеганографія. Застосування стеганографічних методів дозволяє передавати вбудовану в контейнер додаткову інформацію не привертаючи увагу сторонніх спостерігачів. В якості контейнера доцільно використовувати розповсюджені формати, наприклад зображення JPEG, мільйони яких щодня циркулюють у мережі. Існує достатньо велика кількість стеганоалгоритмів для контейнерів формату JPEG, однак з кожним роком з'являються нові методи стеганоаналізу, які їм ефективно протидіють. Тому вдосконалення існуючих та розробка нових стеганоалгоритмів залишається актуальною задачею. Метою даної роботи є розробка методу вбудовування додаткової інформації в зображення з використанням генеративно-змагальних мереж для адаптивної модифікації контейнера з метою підвищення стійкості до статистичного стеганоаналізу. Попереднє підстроювання контейнера дозволить зберегти розподіл кількості пар коефіцієнтів дискретного косинусного перетворення з певними значеннями. В роботі проаналізовані сучасні методи стеганоаналізу та побудовано статистичний класифікатор для стеганографічного алгоритму F5, розроблено генеративно-змагальну мережу для попередньої модифікації зображення-контейнера, досліджена ефективність роботи розробленого статистичного класифікатора без модифікації контейнера та при її наявності, розроблено програмне забезпечення, яке реалізує стеганографічний метод з адаптивною модифікацією зображення-контейнера. Результати експериментів показали, що модифікований за допомогою генеративно-змагальної мережі контейнер є на 36% більш стійким до виявлення додаткової інформації за допомогою статистичного класифікатора.

Ключові слова: цифрове зображення, дискретне косинусне перетворення, вбудовування додаткової інформації, генеративно-змагальні мережі, стеганоаналітичний метод.

Вступ

В епоху бурхливого розвитку інформаційних технологій особливо актуальним залишається питання захисту інформації від несанкціонованого доступу. В останні десятиліття одним з найефективніших інструментів, який забезпечує вирішення цього питання, стала стеганографія. Стеганографічні системи дозволяють приховати факт присутності секретної інформації завдяки формуванню прихованого каналу передачі даних [1]. Повідомлення вбудовується у контейнер, який може передаватися адресату по відкритим каналам зв'язку без ризику виявлення зловмисниками. В якості стеганографічного контейнера доцільно використовувати файли розповсюджених і загальновживаних цифрових форматів, які не привернуть зайвої уваги, наприклад медіафайли, мільйони яких щодня циркулюють у мережі Інтернет. На сьогоднішній день переважна більшість цифрових зображень зберігається у форматі JPEG, тому його використання в якості стеганографічного контейнера не має викликати підозри з боку

сторонніх спостерігачів. В той же час, застосування статичних зображень в якості контейнера може забезпечити більшу ємність у порівнянні, наприклад, з текстовими форматами. Однак, існує велика кількість стеганоаналітичних алгоритмів, які дозволяють виявляти наявність прихованого повідомлення у зображеннях формату JPEG [2]. Зокрема це статистичні методи стеганоаналізу та методи на основі машинного навчання. Тому актуальним завданням є розробка нових стеганографічних методів, стійких до стеганоаналізу, чому і присвячена дана робота.

Аналіз досліджень та публікацій

Формат JPEG передбачає стиснення із втратами, одним з кроків алгоритму є дискретне косинусне перетворення (ДКП). Наявність втрат дозволяє зменшити помітність викривлень, викликаних вбудовуванням додаткової інформації у контейнер. Існує велика кількість стеганографічних алгоритмів, які призначені для приховування інформації у частотній області зображень. Деякі з них, наприклад, використовують різницю коефіцієнтів ДКП для кодування бітів стеганографічного повідомлення, наприклад алгоритм Коха-Жао [1]. Інші виконують модифікацію найменших значущих бітів коефіцієнтів ДКП, наприклад алгоритм F5 [2]. У свою чергу, постійно розробляються нові методи стеганоаналізу, спрямовані на контейнери JPEG формату. В [3] запропоновано швидкодіючий алгоритм для виявляння прихованого повідомлення при його вбудовуванні у найменші значущі біти коефіцієнтів ДКП контейнера на основі аналізу різниці розбалансу парних та непарних коефіцієнтів ДКП. Однак, при частковому заповненні контейнера, алгоритм не забезпечує надійного виявлення. В [4] запропоновано стеганографічний алгоритм для аналізу зображень JPEG на основі Rich-моделі, який забезпечує виявлення прихованого повідомлення навіть за умов часткового заповнення контейнера. Це можливо за рахунок зміни розподілу кількості пар коефіцієнтів ДКП з різними значеннями. Але якщо розподіл пар коефіцієнтів ДКП не змінюється при вбудовуванні повідомлення, надійне виявлення неможливе. У роботі [5] розглянуто можливість виявлення факту модифікації контейнера на основі аналізу окремих блоків із використанням сингулярних чисел. Даний алгоритм забезпечує надійне виявлення прихованого повідомлення у випадку часткового заповнення контейнера, але із збільшенням коефіцієнту заповнення контейнера ефективність виявлення може зменшуватись.

Останнім часом все більшого розвитку набувають методи стеганоаналізу засновані на машинному навчанні [6]. Даний підхід полягає у поєднанні побудови складних статистичних моделей для зображень з можливостями методів машинного навчання, які використовують статистичні показники розподілу для знаходження стійких закономірностей та класифікації. Наприклад, в [7] запропоновано метод стеганографічного аналізу з використанням машинного навчання, що є альтернативою використанню Rich-моделі. Він більш ефективно виявляє приховані повідомлення, хоча і має меншу швидкодію. За допомогою методів машинного навчання вдалося побудувати ефективні класифікатори для багатьох поширених стеганографічних алгоритмів: YASS [8], F5 [6] та інших.

Оскільки стеганоаналітичні методи спираються на аналіз статистичних змін значень коефіцієнтів ДКП, доцільним є дослідження можливості мінімізації цих змін. Одним із способів зменшити викривлення розподілу значень коефіцієнтів ДКП може бути приховування інформації шляхом додавання шумоподібного повідомлення, яке імітує шум фотографічного сенсора, як запропоновано в [9]. Разом з цим, таке рішення обмежує вибір контейнера фотографічними зображеннями. Актуальною задачею є розробка більш універсальних стеганографічних методів, які б забезпечили би попередню модифікацію контейнера таким чином, щоб після вбудовування

повідомлення розподіл значень коефіцієнтів ДКП був наближений до розподілу вихідного контейнера. У цьому відношенні дуже перспективно виглядає застосування генеративно-змагальних мереж (ГЗМ), які добре зарекомендували себе в галузях комп'ютерного зору та обробки природної мови (генерація текстів і зображень), а також все частіше знаходять застосування в задачах стеганографії [10].

Мета статті та постановка завдань

Метою даної роботи є розробка методу вбудовування додаткової інформації в зображення з використанням генеративно-змагальних мереж для адаптивної модифікації контейнера з метою підвищення стійкості до статистичного стеганоаналізу. Попереднє підстроювання контейнера дозволить зберегти розподіл кількості пар коефіцієнтів ДКП з певними значеннями.

Для досягнення мети в роботі розв’язуються наступні задачі:

– проаналізувати сучасні методи стеганоаналізу та побудувати статистичний класифікатор для стеганографічного алгоритму F5;

– побудувати генеративно-змагальну мережу для попередньої модифікації зображення-контейнера;

– дослідити ефективність роботи розробленого статистичного класифікатора без модифікації контейнера та при її наявності;

– розробити програмне забезпечення, яке реалізує стеганографічний метод з адаптивною модифікацією зображення-контейнера.

Основна частина

Для проведення експериментів в роботі обрано стеганографічний алгоритм F5, стеганоаналіз якого залишається актуальним напрямом завдяки наступним його особливостям [6]:

– забезпечує велику пропускну спроможність;

– висока ефективність (вбудовує більше бітів за одну зміну) завдяки матричному кодуванню;

– добре протидіє візуальним атакам;

– має високу стійкість до статистичних атак;

– використовує розповсюджений формат для контейнерів (JPEG);

– має відкритий код.

При вбудовуванні бітів стеганографічного повідомлення у коефіцієнти ДКП алгоритм F5 здійснює декремент значень цих коефіцієнтів. Хоча візуально такі зміни можуть бути непомітними завдяки псевдовипадковому вибору коефіцієнта ДКП для вбудовування, відносна кількість коефіцієнтів із певним значенням буде змінюватися. Кількість коефіцієнтів із значеннями 1, –1 буде зростати, а коефіцієнтів із більшими абсолютними значеннями зменшуватися. Тому можливим є виявлення факту вбудовування повідомлення за допомогою аналізу зміни статистики зображення.

Для експериментів в якості контейнерів було використано набір даних «Linnaeus 5 256X256» [11], зображення формату JPEG. Розмір кожного із зображень 256×256 пікселів. У якості повідомлення було використано псевдовипадкову бітову послідовність із рівномірним розподілом.

Нижче наведено кроки, необхідні для визначення розподілу значень пар коефіцієнтів ДКП та побудови класифікатора з використанням машинного навчання. На даному етапі наявні 8000 цифрових зображень, у які була вбудована додаткова інформація за алгоритмом F5 та така ж кількість зображень, які не підлягали стеганографічному перетворенню.

Масив розподілу значень пар коефіцієнтів ДКП формується за наступним алгоритмом:

– Із набору даних обирається монохромне цифрове зображення $I_{(n,m)}$, де n,m — його ширина та висота.

– Виконується ДКП пікселів зображення, у результаті якого отримуємо масив коефіцієнтів тієї ж розмірності. Його перший елемент (DC коефіцієнт) пропорційний до середньої яскравості блоку і зазвичай набагато більше інших елементів. Інші 63 елементи (AC коефіцієнти) пропорційні інтенсивності колірних переходів у блоці.

– Після виконання квантування коефіцієнтів отримуємо цілі значення.

– Формується масив зсувів з 10 елементів типу (i', j'), де i', j' — координатний зсув для обрання парного коефіцієнта.

– Будується нульова матриця A, розмірність якої дорівнює подвоєному максимальному коефіцієнту ДКП після квантування (без урахування DC коефіцієнтів).

– Для кожного коефіцієнта K з координатами i, j, який не є нульовим чи DC коефіцієнтом, визначаємо парний коефіцієнт $K_{(i+i', j+j')}$. Далі будемо позначати пари як (p,q).

– Виконується інкремент значень $A_{(p,q)}$ для кожної пари коефіцієнтів.

– Виконуємо нормалізацію відносно кількості блоків ДКП (використовуємо блоки 8×8):

$$A_н = A / \left(\frac{n}{8} + \frac{m}{8} \right),$$

– Повторюємо кроки 5-8 для кожного з 10 зсувів.

– Об’єднуємо отримані матриці в одну, яка й буде характеризувати зображення.

Перед тренуванням нейронної мережі дані розбиваються на тренувальну та тестову вибірки. Після завершення процесу навчання на основі тренувальній множині виконується перевірка результату на тестових даних, встановлюється порогове значення. Для оцінки ефективності запропонованого алгоритму на тренувальних та тестових даних використані ймовірності помилок першого та другого роду. Чим менше значення помилок, тим вище ефективність отриманої мережі. У якості вихідної гіпотези H_0 виступає припущення, що контейнер порожній, а H_1, відповідно, що у контейнері є приховане повідомлення. Ймовірності вибору та вірності цих гіпотез наведено у табл. 1.

Таблиця 1

Ефективність класифікатора

Гіпотеза, що була обрана мережею	Правильна гіпотеза	
	H_0	H_1
H_0	0,949	0,051
H_1	0,051	0,949

З наведених результатів видно, що ймовірність успішного виявлення наявності додаткової інформації складає 96%, а детектування порожнього контейнера майже 95%. Відповідно помилок першого роду — 5% та другого 4%. Такі значення дозволяють мережі надійно виявляти факт вбудовування додаткової інформації.

Для протидії стеганоаналізу на основі класифікатора необхідно вирішити проблему зміни статистичних показників контейнера. Можливим шляхом у даному випадку може бути штучна модифікація зображення-контейнера за умови збереження візуальної подібності до його початкового стану. Для розв'язання такої задачі доцільним є використання генеративно-змагальних мереж, які знайшли широке застосування в галузі генерації зображень. Генеративно-змагальні мережі складаються з двох нейронних мереж: одна навчена генерувати дані, а інша – відрізняти підроблені дані від реальних даних. Хоча ідея структури для генерації даних не нова, мережі ГЗМ дають вражаючі результати по створенню штучних зображень та відео. Можливості ГЗМ у стеганографії можна розглядати з різних боків: змагальна гра, генератор або функція відображення. Вони узгоджуються із класифікацією основних стратегій у стеганографії, тобто модифікації, синтезу та селекції [10].

Модифікація зображення, заснована на ГЗМ, фокусується на змагальній грі між стеганографом і стеганалізатором. Даний підхід використовує генератор, навчений для побудови різних ключових елементів, що дозволяє створювати більш захищене від стеганоаналізу стегоповідомлення, яке може передавати по відкритому каналу зв'язку. Фактично генеративно-змагальна мережа складається з трьох мереж: генератор G, дискримінатор D та стеганоаналізатор S (рис. 1).

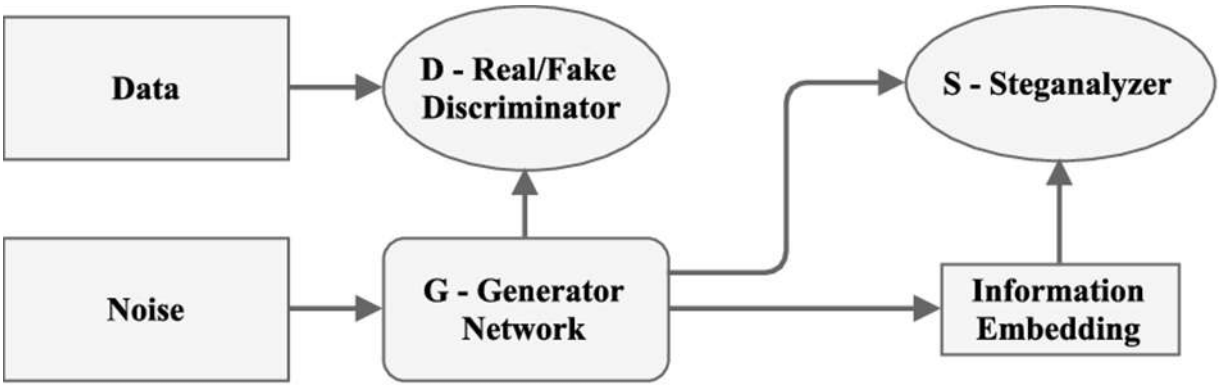

Рис. 1. Структурна схема ГЗМ для модифікації контейнера

На рис. 2 зображена запропонована схема навчання ГЗМ для модифікації контейнера з вбудовуванням на основі стеганографічного методу F5. Модифікація контейнера перед вбудовуванням прихованого повідомлення проводиться таким чином, щоб після нього за статистичними та візуальними характеристиками контейнер був максимально подібним до порожнього. На вхід генератора подаються порожній та заповнений контейнери. Мережа дискримінатора є стандартною згортковою мережею, яка може класифікувати зображення. Генератор є зворотною мережею згортки. Створені таким чином модифіковані контейнери після вбудовування секретного повідомлення будуть більш наближені до порожнього контейнера, але це можливо лише при заздалегідь відомому повідомленні.

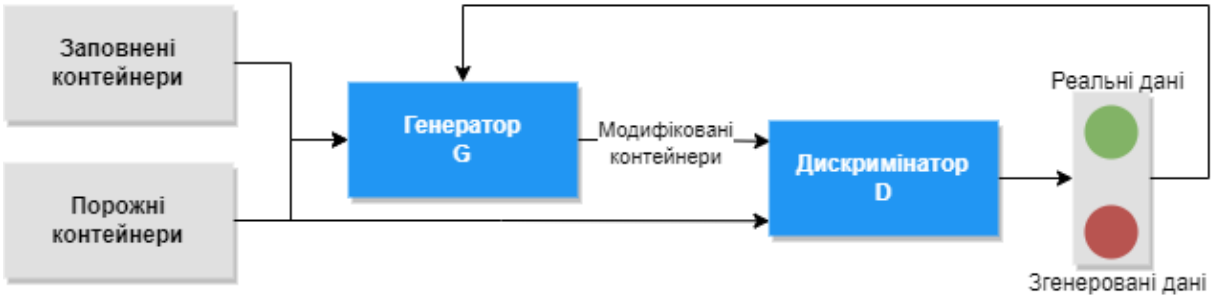

Рис. 2. Схема навчання запропонованої ГЗМ

На рис. 3 наведена спрощена схема роботи стеганографічної системи з застосуванням методу модифікації контейнера. Після вбудовування повідомлення в контейнер статистичний класифікатор (стеганоаналізатор) перевіряє контейнер з повідомленням. Якщо на виході він не виявляє факт вбудовування додаткової інформації, то вважаємо, що повідомлення вбудовано успішно та готово для подальшої передачі по відкритому каналу зв'язку. Якщо класифікатор виявляє повідомлення, то застосовуємо ГЗМ, якій передаємо оригінальний та заповнений контейнери для генерації модифікованого контейнера. Після чого здійснюємо повторне вбудовування, але вже у модифікований контейнер.

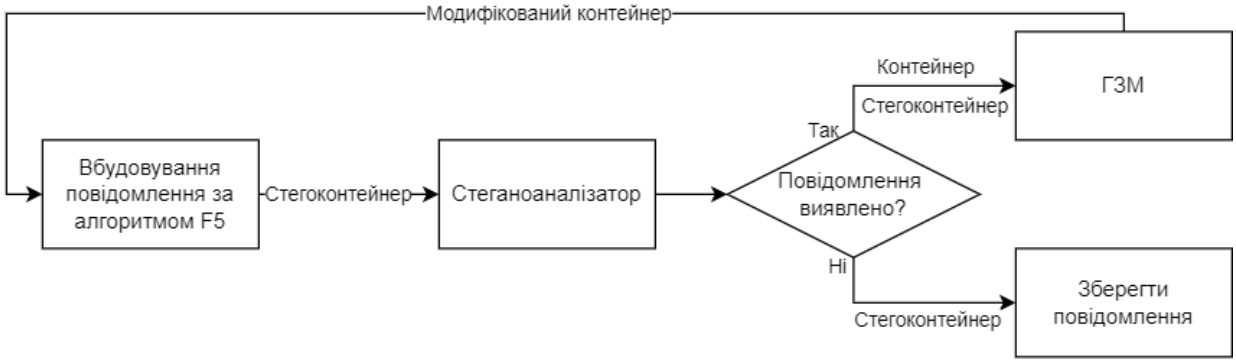

Рис. 3. Структурна схема запропонованої стеганографічної системи

Тренування ГЗМ виконується наступним чином:

1. Для тренування генератора використовується набір зображень «Linnaeus 5 256X256», який розбивається на 3 множини: тренувальну, тестову та валідаційну. На тестову вибірку відводиться 20% від тренувальної – 1600 зображень. На валідаційну – 20% від різниці між кількістю зображень тренувальної та тестовою вибірок, тобто 1280.
2. Будується генеративно-змагальна мережа.
3. Під час її тренування за допомогою алгоритму F5 проводиться вбудовування повідомлення у кожний контейнер. Отримані стеганоповідомлення разом із порожніми контейнерами подаються на вхід генеративної мережі.
4. Дискримінатор приймає як реальні, так і модифіковані зображення і повертає ймовірності від 0 до 1, причому 1 являє собою справжнє зображення (оригінальний контейнер) і 0 представляє фейкове (створене нейронною мережею).

Проведено дослідження роботи розробленого метода модифікації контейнера за умов застосування запропонованого в статті статистичного класифікатора. Для оцінки ефективності запропонованого методу використані ймовірності помилок першого та другого роду. Помилки першого роду демонструють, з якою ймовірністю класифікатор під час аналізу порожнього контейнера покаже, що він заповнений. Помилки другого

роду демонструють, з якою ймовірністю класифікатор покаже, що цей контейнер є порожнім. Чим ближче значення помилок до 0.5, що відповідає випадковому вгадуванню, тим менш дієвим буде стеганоаналіз, і тим більшою буде ефективність запропонованої ГЗМ.

Таблиця 2

Ефективність класифікатора за умови модифікації контейнера

Гіпотеза, що була обрана мережею	Правильна гіпотеза	
	H_0	H_1
H_0	0,599	0,401
H_1	0,401	0,599

З отриманих результатів видно, що ймовірність успішного виявлення значно зменшилась у порівняння із вбудовуванням без попередньої модифікації (табл. 1). Ймовірність помилок першого та другого роду збільшилась на 36%. Такі значення більше не дозволяють класифікатору надійно виявляти факт вбудовування секретного повідомлення.

Висновки

У даній роботі було запропоновано метод, що базується на застосуванні генеративно-змагальних мереж та здатен значно підвищити стійкість до статистичного стеганоаналізу. Під час його розробки було проаналізовано сучасні методи стеганоаналізу та побудовано статистичний класифікатор для стеганографічного алгоритму F5 засобами машинного навчання.

Для аналізу контейнера та детектування наявності вкладеного повідомлення класифікатор розглядає характеристичну матрицю зображення, сформовану на основі аналізу розподілу пар АС коефіцієнтів ДКП. Для його тренування у якості секретного повідомлення було використано псевдовипадкову послідовність бітів з рівномірним розподілом для наближення його статистичних властивостей до повідомлення, яке пройшло процес шифрування.

Тренування генеративно-змагальної мережі та перевірка ефективності її роботи проводилися на окремних частинах набору даних «Linnaeus 5 256X256». Загальний розмір вибірки становить 8000 зображень формату JPEG.

Оцінка ефективності роботи генеративно-змагальної мережі проводилась, виходячи з максимізації ймовірностей помилок першого та другого роду класифікатора. У результаті проведеного моделювання ймовірність помилки під час виявлення вбудованого повідомлення до модифікації контейнера становить 4,5%. Після модифікації цей показних зростає до 40%. З отриманих результатів видно, що застосування генеративно-змагальної помітно зменшує ймовірність виявлення повідомлення, прихованого за алгоритмом F5.

Також було розроблено програмне забезпечення для реалізації стеганографічного метода з адаптивною модифікацією контейнера. Для тренування мереж було застосовано бібліотеку для машинного навчання від Google — TensorFlow, а також хмарний сервіс Google Colab для прискорення процесу навчання. Графічний інтерфейс було реалізовано за допомогою мови програмування Python та бібліотеки PySide6.

Список літератури

1. Конахович Г.Ф., Прогонов Д.О., Пузиренко О.Ю. Комп'ютерна стеганографічна обробка й аналіз мультимедійних даних. К.: Alex Print Centre, 2018. 558 с.
2. Denemark T, Fridrich J. Steganography With MultipleJPEG Images of the Same Scene. *IEEE Transactions on Information Forensics and Security*. 2017. Vol. 12. Is. 10. P. 2308–2319.
3. Калашніков М.В. Яковенко О.О., Кушніренко Н.І., Чечельницький В.Я., Статистичне виявлення стеганографічних повідомлень у зображеннях формату JPEG. *Електротехнічні та комп'ютерні системи*. 2017. № 25(101). С.310–316.
4. Fridrich J., Kodovsky J. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security*, 2012. No 7(3), P. 868–882.
5. Kobozeva A.A., Bobok I.I. Method for detecting digital image integrity violations due to its block processing. *Radiotekhnika*. 2019. No 4 (199), P. 130–141.
6. Fridrich J., Sedighi V. Histogram Layer, Moving Convolutional Neural Networks Towards Feature Based Steganalysis. *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2017*. San Francisco, CA, 2017. URL: DOI:10.2352/ISSN.2470-1173.2017.7.MWSF-325
7. Chen M., Boroumand, M., Fridrich, J. Deep Learning Regressors for Quantitative Steganalysis. *Electronic Imaging,* 2018. No 7, 160-1–160-7.
8. Fridrich J., Kodovsky J., Pevný T. Modern Steganalysis Can Detect YASS. *Proc. SPIE, Electronic Imaging, Media Forensics and Security XII.* San Jose, CA, 2010. P. 02-01 - 02-11.
9. Denemark T., Bas P., Fridrich J. Natural Steganography in JPEG Compressed Images. *Electronic Imaging*. 2018. No 7. P. 316-1–316-10.
10. Zhang Y. Zhang W., Chen K., Liu J., Liu Y., Yu N. Adversarial Examples Against Deep Neural Network based Steganalysis. *Acm Workshop on Information Hiding & Multimedia Security.* Insbrik, Austria, 2018. P. 67-72.
11. Chaladze G. Linnaeus 5 dataset. URL: http://chaladze.com/l5/

DEVELOPMENT OF A METHOD TO INCREASE RESISTANCE TO STATISTICAL STEGANALYSIS USING GENERATIVE ADVERSARIAL NETWORK

Yu.I. Darakchi, N.I. Kushnirenko, O.E. Plachinda

National Odessa Polytechnic University,
1, Shevchenko Ave., Odesa, 65044, Ukraine; e-mail: infsec2011@gmail.com

In the current conditions of constant development of computer technology and networking, the issue of protection of information from unauthorized access is becoming more acute. One of the most effective tools for solving this problem is steganography. The use of steganographic methods allows the transfer of additional information embedded into the container without attracting the attention of outside observers. As a container, it is advisable to use common formats, such as JPEG images, millions of which are circulating on the network every day. There are quite a few steganographic algorithms for JPEG containers, but every year developed new methods of steganalysis that effectively counteract them. Therefore, the improvement of existing and development of new steganographic algorithms remains an urgent task. The aim of this work is to develop a method of embedding additional information in the image using generative adversarial networks for adaptive modification of the container in order to increase resistance to statistical steganalysis. Pre-tuning the container will preserve the distribution of the number of pairs of discrete cosine transform coefficients with certain values. The modern methods of steganalysis are analyzed and the statistical classifier for steganographic algorithm F5 is constructed, the generative adversarial network for preliminary modification of the image-container is developed, efficiency of work of the developed statistical classifier without and with modification of the container is investigated. The results of the experiments showed that the container modified by the generative-competitive network is 36% more resistant to the detection of additional information using the statistical classifier.

Keywords: digital image, discrete cosine transform, additional information, generative adversarial network, steganalysis method.