УДК: 004.932:004.89 **DOI 10.33111/mise.103.14**

Позднякович О.Є.,
к.фіз.-мат.н., доцент кафедри системного аналізу та кібербезпеки,
КНЕУ імені Вадима Гетьмана

Pozdniyakovych O. E.,
Candidate of Physical and Mathematical Sciences, Associate Professor
of System analysis and cybersecurity
KNEU named after V. Hetman

ЗМАГАЛЬНІ АТАКИ НА СИСТЕМИ МАШИННОГО ЗОРУ

ADVERSARIAL ATTACKS ON MACHINE VISION SYSTEMS

Анотація. *Застосування глибоких нейронних мереж у комп'ютерному зорі дало змогу досягти значних результатів у вирішенні багатьох задач, таких як класифікація зображень, виявлення об'єктів, семантична сегментація, класифікація відео. Однак глибокі нейронні мережі чутливі до невеликих змін вхідних даних. Стаття присвячена проблемі змагальних атак на системи машинного зору, які можуть негативно впливати на надійність та точність таких систем. Проводиться аналіз останніх досліджень і публікацій у цій галузі, описуються основні види атак та захист від них.*
Ключові слова: *змагальні атаки, системи машинного зору, глибокі нейронні мережі.*
Abstract. *The use of deep neural networks in computer vision has made it possible to achieve significant results in solving many problems, such as image classification, object detection, semantic segmentation, video classification. However, deep neural networks are sensitive to small changes in the input data. The article is devoted to the problem of adversarial attacks on machine vision systems, which can negatively affect the reliability and accuracy of such systems. The article analyzes the latest research and publications in this area, describes the main approaches to attacks and defense against them.*
Keywords: *adversarial attacks, machine vision systems, deep neural networks.*

Постановка проблеми. Системи машинного зору відіграють важливу роль у сучасному житті. Вони застосовуються в різних галузях, включаючи медицину, автомобільну промисловість, безпеку. Ці системи використовуються для розпізнавання, відстеження і класифікації об'єктів. Однак, системи машинного зору є вразливими до атак, які можуть призвести до їх неправильної роботи. Серед таких атак особливу увагу привертають змагальні атаки. Вони можуть бути здійснені шляхом додавання шуму до зображення, що не помітно для людини, але може змінити результат роботи системи машинного зору

Аналіз останніх досліджень і публікацій. Застосування глибоких нейронних мереж у комп'ютерному зорі дало змогу досяг-

ти значних результатів у вирішенні багатьох задач, таких як класифікація зображень, виявлення об'єктів, семантична сегментація, класифікація відео. Однак глибокі нейронні мережі чутливі до невеликих змін вхідних даних. У зв'язку з цим зростає інтерес до вивчення цих змін та їх вплив на роботу нейронних мереж, особливо в контексті розв'язання візуальних задач [1], [2].

Метою статті є надання огляду основних аспектів, пов'язаних з атаками на системи машинного зору у вирішенні задач класифікації зображень.

Виклад основного матеріалу. У роботі [1] було показано вразливість нейронних мереж у задачі класифікації зображень. Виявилося, що сучасні архітектури схильні до змагальних атак у формі незначних і непомітних для людини змін зображень, що призводять до неправильної класифікації збуреного зображення.

Термін «змагальні атаки» означає протидію роботі системи машинного навчання. Коли модель машинного навчання навчається на наборі даних, вона прагне виявляти закономірності які допоможуть їй робити правильні передбачення для нових даних. Змагальні атаки використовують цей механізм проти моделі.

Атаки можуть бути нецільовими або цільовими:

- *нецільові атаки* змінюють зображення, щоб модель глибокої нейронної мережі віднесла його до будь-якої іншої категорії, відмінної від правильної;
- *цільові атаки* модифікують зображення таким чином, щоб модель передбачила потрібну зловмиснику категорію.

Математичні основи генерації шкідливих збурень. Припустимо, що нейронна мережа — класифікатор, приймає вектор вхідних даних і повертає деякий результат:

$$f(\boldsymbol{x}) = y,$$

де f — модель глибокої нейронної мережі, $\boldsymbol{x}$ — вектор дійсних чисел, кожне з яких відображає значення відповідного пікселя, y — окрема категорія, тобто число з діапазону $1 \dots n$, де n — кількість категорій: $y \in (1 \dots n)$.

Атака модифікує вектор $\boldsymbol{x}$:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\delta},$$

де $\tilde{\boldsymbol{x}}$ — вектор шкідливих вхідних даних, $\boldsymbol{x}$ — вектор початкових вхідних даних, $\boldsymbol{\delta}$ — вектор, що являє собою невелике збурення початкових вхідних даних.

Для нецільової атаки результат моделі (категорія) має відрізнятися від результату, що видається для початкових даних:

$$f(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x}).$$

У разі цільової атаки:

$$f(\tilde{\boldsymbol{x}}) = \tilde{y},$$

де $\tilde{y}$ — цільова категорія шкідливої атаки.

Для успішної атаки необхідно мінімізувати величину шкідливого збурення, щоб зробити його непомітним для людини.

При нецільовій атаці:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \underset{\boldsymbol{\delta}}{argmin}\{\|\boldsymbol{\delta}\|_p : f(\tilde{\boldsymbol{x}}) \neq f(\boldsymbol{x})\}.$$

У разі цільової атаки:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \underset{\boldsymbol{\delta}}{argmin}\{\|\boldsymbol{\delta}\|_p : f(\tilde{\boldsymbol{x}}) = \tilde{y}\}.$$

Величина числа p залежить від того, який спосіб визначення відстані використовується для оцінювання шкідливого збурення:

• *Збурення за нормою* l_0. Норма l_0 збурення дорівнює кількості пікселів зображення, які змінюються під час атаки. При цьому діапазон зміни кожного пікселя не обмежений. Існують варіанти атаки де змінюється мінімальна кількість пікселів, розташованих довільно відносно один одного ([2], [3], [4]);

• *Збурення за нормою* l_1. $\|\boldsymbol{\delta}\|_1 = \sum_{i=1}^{n} |\delta_i|$.

• *Збурення за нормою* l_2. Евклідова норма внесеного збурення $\|\boldsymbol{\delta}\|_2 = \sqrt{\sum_{i=1}^{n} \delta_i^2}$.

• *Збурення за нормою* l_∞. Максимальне значення внесеного збурення для окремого пікселя $\|\boldsymbol{\delta}\|_\infty = \max\{|\delta_i|\}_{i=1}^{n}$.

Грунтуючись на інформації про внутрішній устрій моделі глибокої нейронної мережі, її прогнозах, атаки можна класифікувати як атаки в режимі білої та чорної скриньки:

• *Атаки в режимі білої скриньки.* У цій моделі загроз зловмисник володіє значною інформацією про архітектуру моделі, значення її параметрів і користується нею для безпосереднього створення збурень.

• *Атаки в режимі чорної скриньки*. У таких атаках не передбачається прямого доступу до архітектури та параметрів моделі. Однак припускається взаємодія з моделлю і можливість спостерігати її вихід залежно від входу, який їй надсилається.

Найперша запропонована атака, L-BFGS [1], належить до класу атак у режимі білої скриньки. Щоб ввести в оману класифікатор, використовуються непомітні збурення, додані до нормалізо-

ваного входу $\boldsymbol{x} \in [0,1]$. Атака L-BFGS розв'язує задачу оптимізації з обмеженнями, метою якої є знаходження $\tilde{\boldsymbol{x}} \in [0,1]$ з обмеженням за нормою l_2:

$$\min_{\boldsymbol{\delta}} \; c\|\boldsymbol{\delta}\|_2 + \; L(\tilde{\boldsymbol{x}}, \tilde{y}),$$

де c — гіперпараметр, $L(\tilde{\boldsymbol{x}}, \tilde{y})$ — втрата перехресної ентропії яка вимірює різницю між категорією y, передбаченою цільовим класифікатором на основі збурених вхідних даних, і категорією цільової помилкової класифікації $\tilde{y}$. Ця втрата спонукає генерувати $\tilde{\boldsymbol{x}}$, який здатний ввести в оману цільовий класифікатор.

Атака Карліні-Вагнера [2] генерує збурення, що обмежене l_p-нормою, шляхом розв'язання задачі оптимізації з обмеженнями, аналогічної L-BFGS. Ця атака мінімізує втрати в рамках обмежень l_p-норми, які вимірюють різницю між логіт-значенням $\tilde{z}_y$ зразка $\tilde{\boldsymbol{x}}$, що належить тієї ж категорії y передбачення за $\boldsymbol{x}$, і максимальним логіт-значенням серед усіх інших категорій:

$$\min_{\boldsymbol{\delta}} \left(\|\boldsymbol{\delta}\|_p + \; c\left(\max_{i=1,\dots n} \{\tilde{z}_i; i \neq y\} - \; \tilde{z}_y \right) \right),$$

де n — загальна кількість категорій, $p \in \{0, 2, \infty\}$, а $c > 0$ — константа, яка визначається за допомогою лінійного пошуку для знаходження оптимального значення.

Знаковий метод швидкого градієнта (FGSM) [5] ґрунтується на градієнті та визначає напрямок збурення таким чином, що втрати в цільовій моделі збільшуються. FGSM оцінює збурення, обчислюючи градієнт функції втрат, $L(\boldsymbol{x}, y)$ щодо відношення до заданого входу $\boldsymbol{x}$, з невеликим ε для створення непомітних шкідливих збурень:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \varepsilon \, \text{sign}(\nabla_x L(\boldsymbol{x}, y)).$$

Збурення генеруються в напрямку $\nabla_x L(\boldsymbol{x}, y)$, що вводить цільову модель в оману таким чином, щоб вона уникла вихідної категорії y, що призведе до нецільової атаки. Цей метод також можна використовувати для цільової атаки, зменшивши втрати цільової моделі по відношенню до цільової категорії $\tilde{y}$:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \, sign\big(\nabla_x L(\boldsymbol{x}, \tilde{y})\big),$$

що створює збурення, яке вводить класифікатор в оману і змушує його обрати цільову категорію $\tilde{y}$.

Атаки в режимі чорної скриньки передбачають наявність обмеженої інформації або відсутність інформації про цільову модель. Таку атаку можна виконати за допомогою навчання замісної моделі, яка здатна апроксимувати цільову модель (*апроксимація межі рішення*).

Атака замісної чорної скриньки (substitute blackbox attack, SBA) [6] навчає модель-замінник, яка імітує оригінальну модель, а потім використовує атаки в режимі білої скриньки на навченій замісній моделі для створення обманних збурень.

Атаки шляхом оцінки градієнта. Оптимізація нульового порядку (Zeroth Order Optimization, ZOO) [7] генерує збурення, спостерігаючи за змінами ймовірностей, що дозволяє апроксимувати градієнти, які можна використовувати для стохастичного спуску по координатам при виконанні змагальної атаки.

Однією із проблем атаки у режимі чорної скриньки є велика кількість запитів, які потрібні для прогнозування невідомих цільових моделей. Атака з обмеженням запитів [8] створює збурення з обмеженою кількістю запитів.

Методи випадкового пошуку створюють випадкові збурення поки отримане зображення не введе цільову модель в оману. SemanticAdv [9] змінює вхідне зображення у колірному просторі HSV. Відтінок і насиченість вхідного зображення спотворюються випадковим збуренням, що рівномірно обирається з діапазону допустимих значень. SemanticAdv вносить нові збурення до тих пір, доки класифікатор не буде введений в оману або не буде досягнуто максимальна кількість спроб.

Захист від змагальних атак. Різні способи захисту від атак спрямовані на те, щоб відрізнити збурене зображення і відхилити його.

Аналіз головних компонент (principal component analysis, PCA) [10], ґрунтується на тому, що головні компоненти збурених зображень мають більшу дисперсію, ніж у незбурених зображень.

Статистичний тест [11] використовує той факт, що розподіли даних збурених і незбурених зображень різні. За допомогою статистичного тесту на максимальну середню невідповідність [12], перевіряється, є група даних оманливою чи ні.

Допоміжні моделі можна навчити відрізняти збурені зображення від чистих. Бінарний класифікатор [13] приймає зображення як вхідні дані і генерує двійкову мітку, яка вказує, чи є вхідні дані збуреними чи ні.

Більшість атак використовують градієнт функції втрат цільової моделі. Маскування градієнта спрямоване на те, щоб зробити

інформацію про градієнт цільової моделі непридатною для створення збурень. Прикладами маскування є недиференційований градієнт, градієнт, що зникає/вибухає, і стохастичний градієнт.

Трансформація зображень (image transformation) [14] пропонує кілька підходів до обробки зображень, включаючи кадрування, масштабування, зменшення бітової глибини та стиснення у форматі JPEG. Мінімізація загальної дисперсії також може використовуватися для впорядкування невеликого набору окремих пікселів на зображенні. Зшивання зображень, коли зображення складається з невеликих ділянок, є ще одним підходом, який замінює локальні ділянки на збуреному зображенні чистими ділянками на його найближчому сусідньому зображенні. Ці недиференційовані перетворення ускладнюють атакуючій стороні виведення градієнта функції втрат із цільової моделі.

Градієнт, що зникає/вибухає, під час навчання захисної моделі робить градієнти функції втрат дуже маленькими або великими. Такий підхід відволікає противника від атаки на вхідні зображення. Захисна дистиляція [15] додає гнучкості процесу класифікації, тому модель менш впевнена у своєму прогнозі.

За методом PixelDefend [16] генеративну мережу навчають на чистих даних, щоб апроксимувати їхній розподіл, а отже, спонукають мережу слідувати початковому розподілу, навіть коли вхідними даними є збурені зображення. Через велику кількість параметрів у генеративній мережі градієнти функції втрат можуть стати дуже маленькими або великими, що не дає зробити висновки про необхідні збурення.

Стохастичний градієнт вводить противника в оману щодо цільової моделі, яку атакують, додаючи випадкові операції до вхідних даних або мережі.

Стохастичне відсікання [17] відкидає випадкову підмножину вихідних даних прихованого шару (або активації) і збільшує решту для компенсації. Цей підхід зберігає вузли з ймовірностями, пропорційними їхнім значенням активації, і масштабує вузли, що залишилися, щоб зберегти динамічний діапазон активацій у кожному шарі. Такий підхід дозволяє попередньо навченим моделям бути більш стійкими до збурених зображень без тонкого налаштування.

Висновки та перспективи подальшого дослідження. Глибокі нейронні мережі, як правило, більш вразливі до атак зловмисників, ніж традиційні моделі машинного навчання, частково через те, що наскрізна архітектура навчання глибокої нейронної мережі спрощує атаку, а частково через те, що деякі властивості глибоких нейронних мереж все ще недостатньо вивчені.

Аналіз атак дає можливість визначити стратегії захисту глибоких нейронних мереж. Засоби захисту можуть визначити, чи є вхідні дані збуреними, або запобігти атаці за допомогою градієнтного маскування, де градієнт походить із функції втрат.

Взаємодія між атаками і захистом є важливим напрямом майбутніх досліджень, які допоможуть зрозуміти обмеження моделей глибоких нейронних мереж і розробити більш надійні глибокі нейромережі.

Іншою важливою областю дослідження є фізичні атаки, які змінюють зовнішній вигляд реальних об'єктів або поміщають об'єкти супротивника в навколишнє середовище, і відповідні засоби захисту від таких атак.

Подолання вразливості глибоких нейронних мереж до зловмисних маніпуляцій із зображеннями як у цифровому просторі, так і у фізичному світі є ключовим для впровадження глибоких нейронних мереж у реальні сфери застосування, таких як комерція, безпека та аутентифікація.

Бібліографічні посилання

1. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations, 2014.

2. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy, 2017.

3. The limitations of deep learning in adversarial settings / Nicolas Papernot, Patrick McDaniel, Somesh Jha et al. // 2016 IEEE European symposium on security and privacy (EuroS&P) / IEEE. — 2016. — P. 372–387.

4. Su Jiawei, Vargas Danilo Vasconcellos, Sakurai Kouichi. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. — 2019. — Vol. 23, no. 5. — P. 828–841.

5. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations, 2015

6. Papernot N., McDaniel P., Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv: 1605.07277, 2016.

7. Chen P. Y., Zhang H., Sharma Y., Yi J., Hsieh C. J. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitutemodels. In: Proceedings of the 10th ACMWorkshop on Artificial Intelligence and Security, 2017.

8. Ilyas A., Engstrom L., Athalye A., Lin J. Black-box adversarial attacks with limited queries and information. In: Proceedings of the International Conference on Machine Learning, 2018.
9. Hosseini H., Poovendran R. Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018.
10. Hendrycks D., Gimpel K. Early methods for detecting adversarial images. arXiv:1608.00530, 2016.
11. Grosse K., Manoharan P., Papernot N., Backes M., McDaniel P. On the (statistical) detection of adversarial examples. arXiv:17, 2017.
12. Gretton A., Borgwardt K. M., Rasch M. J., Schölkopf B., Smola A.A kernel two sample test. The Journal of Machine Learning Research, 2012. 13. P. 723–773.
13. Gong Z., Wang W., Ku W. S., 2017. Adversarial and clean data are not twins. arXiv: 1704.04960.
14. Guo C., Gardner J., You Y., Wilson A. G., Weinberge K., 2019. Simple black-box adversarial attacks. In: Proceedings of the International Conference on Machine Learning.
15. Papernot N., McDaniel P., Wu X., Jha S., Swami A., 2016c. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy.
16. Song Y., Kim T., Nowozin S., Ermon S., Kushman N., 2018. Pixeldefend: leveraging generative models to understand and defend against adversarial examples. In: Proceedings of the International Conference on Learning Representations.
17. Dhillon G. S., Azizzadenesheli K., Lipton Z. C., Bernstein J. D., Kossaifi J., Khanna A., Anandkumar A., 2018. Stochastic activation pruning for robust adversarial defense. In: International Conference on Learning Representations.

Статтю подано до редакції: 15.11.2023