DOI:**10.33099/2311-7249/2023-47-2-139-146**

УДК: 623.773.4, 004.89

Цибуля Сергій Анатолійович *(кандидат технічних наук, старший дослідник)* [1]
Волокита Артем Миколайович *(кандидат технічних наук, доцент)* [2]

[1] ***Національний університет оборони України, Київ, Україна***
[2] ***Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна***

СПОСОБИ МАСКУВАННЯ ВІЙСЬКОВИХ ОБ'ЄКТІВ ВІД ВИЯВЛЕННЯ СИСТЕМАМИ ШТУЧНОГО ІНТЕЛЕКТУ

У роботі розглянуті наявні підходи впливу на роботу алгоритмів штучного інтелекту, зокрема машинного навчання, що застосовуються в системах комп'ютерного зору для виявлення, класифікації та ідентифікації об'єктів. На даний час найпопулярнішою та найперспективнішою технологією розпізнавання образів є штучні нейронні мережі. Комп'ютерний зір застосовується у військовій справі для виявлення візуальних об'єктів певних класів: людей, озброєння та військової техніки, військових об'єктів тощо. Вхідними даними для аналізу можуть бути: фотографії, відеокадри чи відео потік реального часу, що отримані з космічних, повітряних або наземних засобів розвідки. Для боротьби з системами автоматичного виявлення об'єктів можливо застосовувати підходи, що здатні впливати на моделі машинного навчання, які використовуються у цих системах. Атака на моделі машинного навчання – це спеціальні дії щодо впливу на її елементи з метою досягти бажаної поведінки системи або перешкодити її коректній роботі. За результатами аналізу досліджень різних авторів визначено, що майже кожен алгоритм машинного навчання має певні вразливості. Під час виконання завдань інженерної підтримки військ щодо маскування військових об'єктів, найбільш доступними способами впливу на системи комп'ютерного зору, для введення їх в оману, є зміна фізичних властивостей об'єкта, що маскується, шляхом нанесення на його поверхню спеціальних покриттів і матеріалів. У якості покриттів можливо використовувати згенеровані змагальні патч-зображення, шляхом накладання або наклеювання їх на об'єкт та які здатні вносити завади в роботу алгоритмів засобу розвідки, прицілювання або наведення. Це особливо важливо в перспективі створення автономних систем зброї, які здатні виявляти, ідентифікувати цілі та самостійно приймати рішення на їх ураження.

Ключові слова: *штучний інтелект, машинне навчання, нейронна мережа, комп'ютерний зір, виявлення, ідентифікація, класифікація, інженерна підтримка, маскування військових об'єктів, атака ухилення, патч-зображення, змагальний приклад, отруєння набору даних.*

Вступ

Починаючи з 2017 року у світі розпочалася гонитва за світове лідерство між провідними державами у сфері розвитку штучного інтелекту. Протягом 2017–2019 років понад 30 країн світу (наприклад, Канада, Китайська народна республіка, Федеративна республіка Німеччина, Французька Республіка, російська федерація тощо), розробивши відповідні національні стратегії, визначили розвиток технологій штучного інтелекту одним із важливих пріоритетів державної політики. Так, уряд КНР поставив перед країною амбіційні плани щодо досягнення світового лідерства в області штучного інтелекту до 2030 року. За останнє десятиліття КНР більш ніж утричі збільшила свої інвестиції у наукові дослідження за цим напрямом [1]. Ураховуючи зростаючу роль штучного інтелекту в сферах загальнодержавного значення, розпорядженням Кабінету Міністрів України від 02 грудня 2020 року № 1556-р була схвалена Концепція розвитку штучного інтелекту в Україні [2].

У травні 2020 року було опубліковано доповідь Організації НАТО з науки та технологій «Тенденції у науці й технологіях: 2020–2040», в якій окреслені тенденції розвитку технологій протягом наступних 20 років [3]. Документ базується на аналізі відкритих наукових джерел і досліджень та певних національних науково-дослідних програм, а також ґрунтується на висновках багатьох провідних вчених, інженерів та аналітиків. Цим документом визначаються новітні напрями розвитку науки і технологій, які можуть якісно змінити види озброєння і матимуть вплив на розвиток збройних сил, колективної безпеки та оборони країн членів НАТО [4]. До цього переліку входять технології штучного інтелекту, аналізу неструктурованих даних, автономних транспортних засобів та робототехніки.

Постановка проблеми. Поняття «штучний інтелект» (англ. Artificial intelligence (далі – AI)) його алгоритми і математичні моделі набуло широкого використання у повсякденному житті людства, застосовується в багатьох галузях:

медицині, банківській діяльності, біржовій торгівлі, військовій справі, наукових дослідженнях тощо. Нині новини, що містять такі поняття як «машинне навчання» (англ. Machine learning (далі – ML)), «штучні нейронні мережі» (англ. Artificial neural network (далі – ANN)), «згорткові нейронні мережі» (англ. Convolutional neural network (далі – CNN)), «мережі глибокого навчання» (англ. Deep neural network (далі – DNN)), «генеративні змагальні мережі» (англ. Generative adversarial network (далі – GAN)), «великі дані» (англ. Big data), «комп'ютерний зір» (англ. Computer vision) перебувають на передовиці всіх засобів масової інформації.

Зважаючи на вищевикладене зазначимо, що останнім часом питанню підвищення стійкості критичних систем, що використовують технології штучного інтелекту, надається все більше уваги. Одним із важливих його елементів є кібербезпека систем машинного навчання, як складової частини галузі штучного інтелекту.

Аналіз останніх досліджень і публікацій. Розпізнавання образів – важливе завдання комп'ютерного зору, яке застосовується для виявлення візуальних об'єктів певних класів (людей, озброєння та військової техніки, військових об'єктів тощо) на таких цифрових зображеннях як фотографії, скріншоти з відео чи відеокадри. Розпізнавання образів набуло широкого розповсюдження в таких сферах, як безпека та відеоспостереження, інтелектуальна відео хірургія, маркетинг та реклама, автономні транспортні засоби, доповнена реальність та пошук зображень. Тому і найбільше теоретичних досліджень щодо порушення роботи алгоритмів штучного інтелекту, зокрема, машинного навчання, присвячено саме питанню атаки на автоматичне розпізнавання образів. Атаки на системи машинного навчання, з метою отримання певного практичного зиску, з'явилися спочатку в контексті протидії статистичним фільтрам спаму, системам виявлення вірусів та детектування зловмисного трафіку в мережі [5; 6]. Але жодна з розглянутих атак не призначена для впливу на штучні нейронні мережі, які на даний час є найпопулярнішою та найперспективнішою технологією розпізнавання образів.

Термін, що описує алгоритми впливу на машинне навчання – «змагальне машинне навчання» (англ. Adversarial machine learning (далі – AML)), а самі дії отримали назву «змагальні атаки». До набуття поширення машинного навчання, дослідження щодо AML мали лише теоретичний характер. Першими звернули увагу на те, що шляхом невеликих змін системи розпізнавання образів можна змусити видавати неправильні результати, співробітники підрозділу Google AI [7]. Надалі з'явилось досить багато досліджень, що розглядають приклади впливу змагальних атак [8].

Аналіз понад 2000 наукових статей, що пов'язані з безпекою в галузі штучного інтелекту, який був проведений фахівцями компанії Adversa AI, показав, що майже кожен алгоритм машинного навчання потенційно вразливий, і має певні проблеми щодо конфіденційності та безпеки [9].

Перше всебічне дослідження змагальних атак на глибоке навчання в системах комп'ютерного зору представлено у статті [10]. У ній проаналізовано значний перелік наукових статей та обґрунтовано, що змагальні атаки є реальною загрозою глибокому навчанню на практиці, особливо в критично важливих програмах безпеки.

Спробу створити таксономію змагальних атак, шляхом аналізу понад 150 літературних джерел, починаючи з 2016 року, зроблено у роботі [11]. Автори розглянули 41 підхід до виконання фізичних змагальних атак у трьох основних завданнях комп'ютерного зору: виявлення, ідентифікація та класифікація.

Свою таксономічну схему для класифікації наявних фізичних змагальних атак на DNN мережі, як популярну технологію в задачах комп'ютерного зору, запропоновано в роботі [12]. Автори розглянули загальні характеристики фізичних змагальних атак та обговорюються проблеми, які необхідно вирішити для запобігання цих атак.

Армією США впроваджується концепція «мереже-центричної війни» (Network-centric warfare) і «мульти-доменної операції» (Multi-Domain Battle), що передбачає проведення військових операцій у різних просторах (морський, повітряний, кібернетичний, інформаційний тощо), і які вимагають від військ високої мобільності та швидкого прийняття рішень. Для реалізації цієї вимоги всі складові елементи збройних сил (особовий склад, озброєння та військова техніка (далі – ОВТ), штаби тощо) мають бути пов'язані єдиною інформаційною мережею для обміну інформацією в ході бойових дій у режимі реального часу. Одним зі способів вирішення цього завдання стало застосування рішень, які отримали назву «інтернет-бойових речей» (англ. Internet of Battle Things (далі – IoBT)) [13]. Тому, з урахуванням розповсюдження та перспективи IoBT, цікавим є комплексний аналіз нападу та захисту у сфері комп'ютерного зору інтернет речей, що проведено у дослідженні [14]. В роботі зазначено, що поєднання штучного інтелекту і периферійних обчислень, забезпечує розгортання алгоритмів глибокого навчання на периферійних пристроях та робить їх об'єктами привабливими для атак.

Метою статті є визначення способів підвищення ефективності маскування озброєння і військової техніки та інших військових об'єктів від їх виявлення, класифікації та ідентифікації системами штучного інтелекту, шляхом впливу на роботу алгоритмів штучних нейронних мереж, які застосовуються в галузі комп'ютерного зору.

Виклад основного матеріалу дослідження

Атака на модель машинного навчання це спеціальні дії впливу на її елементи (тренувальні дані, алгоритм, тестові дані тощо) з метою

досягнення бажаної поведінки системи або перешкодити її коректній роботі. В атласі MITER, який створений спільно з IBM, NVIDIA, Bosch, Microsoft та іншими компаніями, виділено більш ніж 30 методик атак на алгоритми машинного навчання, які розподіляються на три основні типи атак [15]:

на набори даних для навчання;

під час навчання моделі;

під час виконання моделлю завдань за призначенням.

Під час навчання моделей, атаці (спеціальній модифікації, отруєнню) можуть бути піддані тренувальні або тестові дані. Створення наборів даних для навчання (image's data set), а, особливо, таких специфічних даних як зображення озброєння та військової техніки, є часозатратним процесом. Тому привабливим варіантом для розробників є не створення власних наборів, а застосування загально доступних. Найбільший відомий набір даних комп'ютерного зору є ImageNet (image-net.org). Це проєкт зі створення великої бази даних розмічених зображень призначених для тестування методів розпізнавання образів та машинного зору, містить в собі зразки військової техніки. Він має певні обмеження для використання, але знаходиться у вільному доступі на torrent серверах. На платформі Kaggle, для змагань з аналітики та передбачувального моделювання, можливо завантажити готовий набір зображень танків різних країн та поколінь «Military tanks dataset (images) a collections of various military tank image». На сайті images.cv також можна завантажити набір зображень танків у розмірах (16×16, 32×32 тощо) з розподілом набору на тренувальні та перевірочні частини у необхідних пропорціях. Без попередньої перевірки не можливо зрозуміти, чи всі ці загальнодоступні набори даних є коректними, чи «отруєними».

Проблема ускладнюється тим, що на роботу моделі машинного навчання може вплинути навіть невеликий обсяг «шкідливих» даних. Так, дослідження оцінювання впливу на алгоритм дозування ліків для пацієнтів, що працює на основі нейронної мережі, продемонструвало, що додавши 8% шкідливих даних до навчального набору, було отримано 75% змін від дозування, запропонованого алгоритмом, що навчений на незараженому навчальному наборі [16]. Тому, без впевненості про чистоту загальнодоступних навчальних наборів даних, не можна їх використовувати у системах військового призначення. Принциповим моментом для систем машинного навчання є те, що система навчається та перевіряється на основі одних даних, а практично працює з іншими.

Атаки на саму модель є найбільш ефективними. Проте зазначені атаки в реальному житті зустрічаються рідко, адже для їх проведення необхідно мати інформацію про алгоритм моделі або доступ до її навчальних даних. Водночас, враховуючи те, що створення власної моделі машинного навчання потребує певних теоретичних знань та часу, розробники систем комп'ютерного зору часто використовують такі популярні моделі як YOLO, Single-Shot Detector (SSD), Mask Region-based Convolutional Network (Mask R-CNN) тощо. Також, з метою економії часу та коштів, а деякі складні архітектури нейронних мереж не можливо навчити на звичайному комп'ютері, розробники виконують навчання моделей в хмарних сервісах або використовують вже попередньо навчені моделі нейронних мереж шляхом заміни останніх прошарків мережі під необхідне завдання. За таких умов, основні параметри та вагові коефіцієнти залишаються без змін [17]. Відповідно, якщо початкова модель зазнала втручання, то підсумкова модель частково успадкує видачу неправильних результатів [18].

Здебільшого атаками, що можна практично здійснити, є атаки, які виконуються на етапі застосування нейронних мереж – це, так звані, «атаки ухилення» (evasion attack). Їх метою є примушення мережі видавати неправильні відповіді у певних ситуаціях. Як правило, для атак ухилення використовуються «змагальні приклади» (adversarial examples), суть яких полягає у зміні вхідних даних так, що модель не може їх правильно інтерпретувати.

Атаки ухилення (рис. 1) ділять на різні категорії або групи за:

бажаною відповіддю;

доступністю моделі;

способом підбору завад.

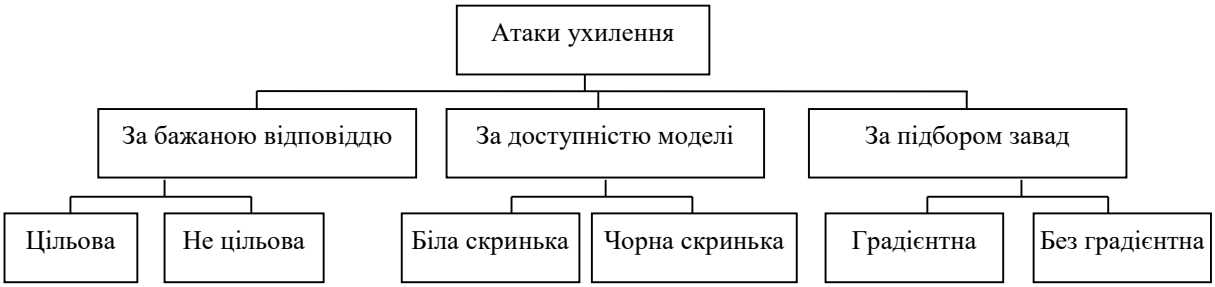

Рисунок 1 – Класифікація атак ухилення

Ця класифікація є лише однією із багатьох способів поділу атак ухилення на нейронні мережі. Класифікація атак може змінюватися з появою нових методів та відкриттям нових вразливостей нейронних мереж.

Атаки «за бажаною відповіддю» спрямовані на отримання певної відповіді від системи, та поділяються на два види:

нецільові (non-targeted), що виконуються з метою викликати будь-яку помилку або невірну класифікацію вхідних даних;

цільові (targeted), їх метою є отримання певних конкретних помилкових відповідей.

Атаки «за доступністю моделі» відображають рівень доступності до системи або моделі, що атакується. Під час атаки на білу скриньку (white-box) особа, що проводить атаку, має повний доступ до архітектури, параметрів та інформації про внутрішні процеси моделі. Під час атак на чорну скриньку (black-box) є обмежений або взагалі відсутній доступ до цієї інформації, вона може базуватися лише на зовнішніх спостереженнях та обміні даними з моделлю.

Атаки за способом «підбору завад» відображають методи, що використовуються для знаходження оптимальних завад або змін у вхідних даних. Атаки на основі градієнта (gradient-based) базуються на властивості моделей машинного навчання, що невеликі зміни вхідних даних можуть призводити до значних змін у вихідних результатах моделі. Градієнти використовуються для підрахунку змін, які максимізують або мінімізують функцію втрат, що призводить до помилкової класифікації. Безградієнтні (non-gradient-based) атаки використовують такі методи, як оптимізація, еволюційні алгоритми або генеративні моделі для пошуку оптимальних змін.

У реаліях військової справи, противник під час атаки на системи штучного інтелекту, не має доступу до цифрових вхідних даних (фотографій, відео) з безпілотних літальних апаратів, камер спостереження, головних частин самонаведення ракет і автономних боєприпасів, літаків розвідників або супутників. Також йому, зазвичай, не відомі моделі машинного навчання та їхня структура. Результати атаки можна оцінити лише за непрямими ознаками впливу або наступних дій противника. Тому, єдиним можливим способом перешкоджанню роботі систем комп'ютерного зору є атаки з фізичним впливом (physical attacks).

Атаки з фізичним впливом на нейромережі можна класифікувати за наступними напрямами:

вплив на середовище;

маніпуляція з вхідними даними;

вплив на систему обробки даних.

Під час атаки «впливу на середовище» змінюються або спотворюються дані, що надходять на сенсори системи. Це може бути фізична зміна навколишнього середовища (зміна освітлення, розміщення перешкод тощо) або зміна фізичних властивостей об'єкта (додавання спеціального покриття або матеріалу).

Під час «маніпуляції зі вхідними даними» змінюються дані, що надходять до системи. Це може бути фізична зміна зображення шляхом накладання спеціальних маркувань або спотворень.

Під час «впливу на систему обробки даних» відбувається прямий вплив на саму систему або її компоненти. Це може бути фізичне пошкодження, шляхом завдання вогневого удару по системі, або зміна її апаратної частини, вплив на сенсори (зміна поля зору або чутливості), зміна робочих параметрів, порушення роботи алгоритмів обробки даних або фізичне втручання у роботу нейромережі шляхом проведення кібератаки, маніпуляції з електричними сигналами, впровадження шуму та завад усередині системи за допомогою засобів РЕБ.

Узагалі, класифікація атак фізичного впливу на нейромережі є гнучкою і залежить від контексту, системи та цілей під час атаки. Ці типи атак можуть бути комбіновані або використовуватися у різних комбінаціях для досягнення бажаного ефекту.

Під час виконання завдань інженерної підтримки військ щодо маскування військових об'єктів найбільш доступними способами впливу на системи комп'ютерного зору є зміна фізичних властивостей об'єкта, що маскується, шляхом нанесення на його поверхню спеціальних покриттів і матеріалів. Як запропоновано в роботі [19], з метою маскування, на поверхню об'єкта нанести змагальне покриття з аерогелю. Щоб спростити та здешевити фізичну реалізацію, автори оптимізували змагальну текстуру покриття, створивши її подібною до QR-коду. Під час тестування на об'єкті з нанесеною текстурою, продуктивність детектора на основі моделі YOLOv3 знизилась на 64,6%.

Фізична зміна зовнішнього виду об'єкта можлива шляхом накладання на нього покриттів зі спеціально сформованих зображень, так званих змагальних патчей (англ. patch клаптик) (рис. 2).

Змагальний патч – це спеціально згенероване зображення з певними візерунками, яке можна прикріпити на поверхню цільового об'єкта. Його перевагою є те, що вони мають просте використання. Таке зображення можна роздрукувати на принтері, а потім наклеїти/повісити на необхідний об'єкт. Цей вид атак називається патч-атаки (patch attack).

Вперше патч-атаку було описано у роботі [20], де шляхом заміни локальної області зображення на оптимізовану текстуру вдалося досягти зниження ефективності роботи DNN мережі. Практичне застосування патч-зображень на фізичні об'єкти, було досліджено у роботі [21]. Автори, шляхом накладання наліпок на автомобіль Toyota Camry, домоглися порушити роботу нейромережі, що призвело до неможливості виявлення автомобіля системою розпізнавання. Найбільш відомий приклад патч-атаки пов'язаний із розпізнаванням дорожніх знаків, процес розпізнавання яких вдалося порушити за допомогою декількох наліпок патч-зображень зроблених на знак [22]. Результати експерименту свідчать, що оптимізоване патч-зображення може значно знизити (до 47%) продуктивність детектора системи розпізнавання [23].

Застосування змагальних атак має широке коло використання, як для маскування об'єктів від автоматичного виявлення на аерофотознімках, так

і на зображеннях дистанційного зондування Землі з космосу. Практичні дослідження щодо приховування від автоматичного розпізнавання літаків на аерофотознімках, шляхом нанесення змагального зображення на ділянки поверхні землі, де розміщений об'єкт (рис. 3), проведені у роботі [24]. Автори перевірили вплив цих змагальних зображень на роботу 16 популярних моделей класифікаторів (YOLO, SSD, Faster R-CNN, Swin Transformer, TOOD тощо). Результати дослідження засвідчили, що під час використання змагальних зображень ймовірність виявлення об'єктів деякими класифікаторами впала до 0.

(a) Camouflage net (b) Patch camouflage

Рисунок 2 – Маскування за допомогою маскувальної сітки (a) та накладанням патч-зображення (b) [23]

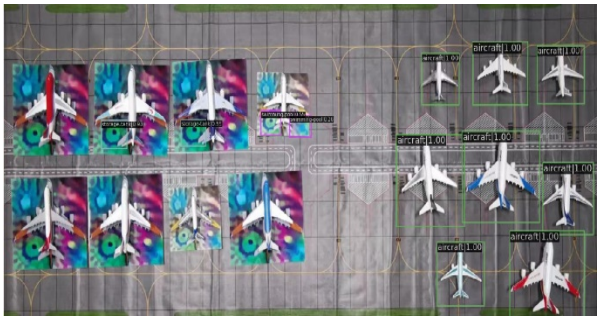

Рисунок 3 – Зліва розміщені літаки, для приховування яких використовуються змагальні патч-зображення [24]

Результати експериментів проведених щодо створення змагальних прикладів для тестувальних систем з аналізу зображень дистанційного зондування, показали вразливість CNN, ймовірність видачі неправильного результату якими досягала понад 80% [25]. Незважаючи на те, що застосування змагальних зображень показує задовільну ефективність атак на нейронні мережі є недоліки, що стримують ефективність такого типу атак:

лише певна область зображення є вирішальною для впливу на прийняття рішення системою розпізнавання, а зміна неспецифічних областей може мати зворотний ефект;

патч-зображення являють собою мозаїчний візерунок, який виглядає неприродно та помітні для людини-спостерігача, що шкодить скритності об'єкта.

Враховуючи наведене, одним з напрямів розвитку таких атак є адаптивне коригування форми патчів. Наприклад, використання генеративних змагальних мереж для автоматичного створення зображень, малюнок яких буде найбільш наближеним до природних зображень.

На ефективність маскування впливають різні демаскуючі ознаки. Тінь від об'єкта, її контрастність, розмір, кольорова гамма тощо – можуть зробити об'єкт помітним або незвичним порівняно з навколишнім середовищем. Щоб розв'язати цю проблему в роботі [26] автори провели оптимізацію 2D-зображення об'єкта і трансформували його текстуру для наближення ефекту тіні та нанесли її на масштабовану модель Tesla Model 3, яку надалі роздрукували за допомогою 3D-принтера. Результати проведення експерименту свідчать про те, що середнє зниження продуктивності двох детекторів на базі моделей EfficientDetD0 і YOLOv4 становить 47,5%. Також результати підтвердилися моделюванням виявлення об'єктів у середовищі CARLA Simulator.

Одним із завдань маскування є імітація військових об'єктів [27]. Для цього застосовуються, як плоскі 2D горизонтальні та вертикальні макети, так макети у вигляді 3D-об'єктів, що мають вигляд реальних об'єктів ОВТ. Питання введення в оману та порушення розпізнавання нейронною мережею шляхом побудови 3D-об'єктів розглянуто у роботі [28]. Авторами запропонований алгоритм Expectation Over Transformation, який дозволяє побудувати змагальні 3D приклади через процес 3D-рендерінга. Результати цих досліджень дозволили генерувати змагальні приклади, які призводять до невірної роботи нейромереж з класифікації об'єктів. На рис. 4 показано приклад надрукованої на 3D-принтері черепахи, яка розпізнається класифікатором TensorFlow InceptionV3, як гвинтівка (де, зелена рамка –

черепаха, червона – гвинтівка, чорна – не черепаха). Результати цієї роботи підтверджують, що змагальні атаки можуть застосовуватися для імітації 3D об'єктів.

Рисунок 4 – Результати роботи класифікатора InceptionV3 [27]

Таким чином, можна стверджувати, що під час використання змагальних атак на системи виявлення та класифікації об'єктів, які працюють на основі алгоритмів машинного навчання таких, як нейронні мережі, можна досягти успішних результатів цих атак:

1. Змагальне (маскувальне) зображення знижує ймовірність виявлення об'єкта, а ділянка розміщення об'єкта не пропонується системою або пропонується лише частково як можливий кандидат для класифікації.

2. Ділянка перебування об'єкта визначена, але невірно класифікована, або оцінка класифікації надто низька, щоб подолати порогове значення показника критерія прийняття рішення про виявлення.

3. Ділянка перебування об'єкта успішно визначена та класифікована, але сам об'єкт не має чіткої класифікації.

Висновки й перспективи подальших досліджень

У роботі розглянуті наявні підходи впливу на моделі машинного навчання, що застосовуються для виявлення та ідентифікації об'єктів системами комп'ютерного зору. За результатами проведеного аналізу, можна констатувати, що майже кожен алгоритм машинного навчання принципово вразливий і має проблеми з безпекою. Тому, кожний елемент системи штучного інтелекту військового призначення (математичні моделі, алгоритми машинного навчання та набори вхідних даних, що використовуються для навчання й тестування) спрямований на підвищення обороноздатності держави і зобов'язаний мати певні ступені обмеження розповсюдження та конфіденційності. Важливість цього питання зазначається у керівних документах як відомчого, так і загальнодержавного рівня [29; 30], де до інформації з обмеженим доступом відносяться: відомості про несекретне програмне забезпечення, що використовується під час виконання розвідувальних завдань; напрями, науково-технічні ідеї, результати, можливість застосування (реалізації) фундаментальних, пошукових прикладних наукових досліджень у системах або їх складових з метою підвищення ефективності технічної розвідки, засобів прицілювання або наведення, їх можливостей з виявлення об'єктів та цілей на фоні місцевості; покращення ефективності протидії засобам прицілювання або наведення зброї противника, зменшення можливостей з виявлення об'єктів та цілей.

Це є особливо важливим в перспективі створення автономних систем зброї, які будуть здатні виявляти, ідентифікувати та самостійно приймати рішення на ураження цілей.

Список бібліографічних посилань

1. America's eroding technological advantage: nds rdt&e priorities in an era of great-power competition with China. URL: https://govini.com/wp-content/uploads/2021/04/Govini_NDS-Priorities-RDTE.pdf (дата звернення: 26.05.2023). **2. Про схвалення** Концепції розвитку штучного інтелекту в Україні : Розпорядження Кабінету Міністрів України № 1556-р від 02.12.2020. URL: https://zakon.rada.gov.ua/laws/show/1556-2020-р (дата звернення: 26.05.2023). **3. Reding D.F., Eaton J.** Science & Technology Trends: 2020-2040. Exploring the S&T Edge. *NATO Science & Technology Organization*. Brussels. Belgium. 2020. P. 160. URL: https://www.sto.nato.int/pages/tech-trends.aspx. (дата звернення: 26.05.2023). **4. Войтовський К. Є.** Глобальні тренди розвитку науки і технологій: нові виклики і можливості. Київ : Національний інститут стратегічних досліджень, 2020. 6 с. URL: https://niss.gov.ua/doslidzhennya/nacionalna-bezpeka/globalni-trendi-rozvitku-nauki-i-tekhnologiy-novi-vikliki-i (дата звернення: 30.10.2022). **5. Lowd D., Meek C.** Good word attacks on statistical spam filters. *Proceedings of the second conference on email and anti-spam*, 2005. P. 1–8. **6. Biggio B., Nelson B.and Laskov P.** Poisoning attacks against support vector machines. Proceedings of 29th Int. Conf. Mach. Learn, 2012. P. 1467–1474. DOI: 10.48550/arXiv.1206.6389. **7. Szegedy C. et al.** Intriguing properties of neural networks. CoRR, abs/1312.6199. 2013. P. 10. DOI: 10.48550/arXiv.1312.6199. **8. Carlin N.** A Complete list of all (arXiv) adversarial example papers. URL: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. (дата звернення: 26.05.2023). **9. Adversa AI.** URL: https://adversa.ai/report-secure-and-trusted-ai/ (дата звернення: 26.05.2023). **10. Akhtar N., Mian A.** Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*. 2018. Vol. 6. P. 14410–14430. DOI: 10.1109/ACCESS.2018.2807385. **11. Wei Hui et al.** Physical Adversarial Attack meets Computer Vision: *A Decade Survey*. 2022. P. 32. DOI: 10.48550/arXiv.2209.15179. **12. Wang D., Yao W., Jiang T., Tang G. and Chen X.** A Survey on Physical Adversarial Attack in Computer Vision. *ArXiv abs/2209.14262*. 2022. P. 26. DOI: 10.48550/arXiv.2209.14262. **13. Niranjan S. et al**. Analyzing the applicability of Internet of Things to the battlefield environment. International *Conference on Military Communications and Information Systems*. Brussels, Belgium, 23 May, 2016. P. 1–8. DOI: 10.1109/ICMCIS.2016.7496574. **14. Liang H., He E., Zhao Y., Jia Z., Li H.** Adversarial Attack and Defense: A Survey. *Electronics*. 2022. № 11(8). P. 1283. DOI:

10.3390/electronics11081283. **15. MITRE ATLAS™** (Adversarial Threat Landscape for Artificial-Intelligence Systems). URL: https://atlas.mitre.org/ (дата звернення: 26.05.2023). **16. Jagielski M. et al.** Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, 2018. P. 19–35. DOI: 10.1109/SP.2018.00057. **17. Гонтаренко Я. Д., Красношлик Н. О.** Використання нейронних мереж для розпізнавання дій людини по відео. *Вісник Черкаського національного університету імені Б. Хмельницького. Серія «Прикладна математика. Інформатика».* 2019. № 2. С. 59–72. DOI: 10.31651/2076-5886-2019-2-59-72. **18. Gu T., Liu K., Dolan-Gavitt B. and Garg S.** BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*. 2019. Vol. 7. P. 47230-47244. DOI: 10.1109/ACCESS.2019.2909068. **19. Zhu X., Hu Z., Huang S., Li J. and Hu X.** Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. P. 13317-13326. **20. Brown T., Mané D., Roy A., Abadi M. and Gilmer J.** Adversarial patch. 2017. *ArXiv:1712.09665*. **21. Zhang Y., Foroosh H., David P. and Gong B.** CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. International Conference on Learning Representations. URL: https://openreview.net/pdf?id=SJgEl3A5tm. (дата звернення: 26.05.2023). **22. Eykholt K. et al.** Robust Physical-World Attacks on Deep Learning Visual Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. P. 1625-1634. DOI: 10.48550/arXiv.1707.08945. **23. Hollander R. den et al.** Adversarial patch camouflage against aerial detection. *Artificial Intelligence and Machine Learning in Defense Applications II.* Vol. 11543. SPIE, 2020, P. 77–86. DOI: 10.1117/12.2575907. **24. Lian J., Wang X., Su Y., Ma M. and Mei S.** CBA: Contextual Background Attack Against Optical Aerial Detection in the Physical World. *IEEE Transactions on Geoscience and Remote Sensing*. 2023. Art no. 5606616. Vol. 61. P. 1–16. DOI: 10.1109/TGRS.2023.3264839. **25. Chen L. et al.** Attack Selectivity of Adversarial Examples in Remote Sensing Image Scene Classification. *IEEE Access.* 2020. Vol. 8. P. 137477–137489. DOI: 10.1109/ACCESS.2020.3011639. **26. Suryanto N. et al.** Dta: Physical camouflage attacks using differentiable transformation network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022. P. 15305–15-14. DOI: 10.48550/arXiv.2203.09831. **27. Керівництво** з виконання інженерних заходів маскування військ та об'єктів : наказ начальника Головного управління оперативного забезпечення Збройних сил України від 06.12.2017 № 90. Київ. 138 с. **28. Athalye A., Engstrom L., Ilyas A. and Kwok R.** Synthesizing Robust Adversarial Examples. 2017. arXiv:1707.07397. **29. Про затвердження** Переліку відомостей Міністерства оборони України, які містять службову інформацію (ПСІ – 2016) (зі змінами) : Наказ Міністерства оборони України від 27.12.2016 № 720. 31 с. **30. Про затвердження** Зводу відомостей, що становлять державну таємницю : Наказ Служби безпеки України від 23.12.2020 № 383. 121 с..

WAYS TO MASK MILITARY OBJECTS FROM DETECTION BY ARTIFICIAL INTELLIGENCE SYSTEMS

[1] Tsybulia Serhii ***(Candidate of Technical Sciences, Senior Researcher)***
[2] Volokyta Artem ***(Candidate of Technical Sciences, Associate Professor)***

[1] The National Defence University of Ukraine, Kyiv, Ukraine
[2] National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine

The paper examines available approaches to influence the work of artificial intelligence algorithms, in particular machine learning, used in computer vision systems for object detection, identification, and classification. Currently, the most popular and most promising pattern recognition technology is artificial neural networks. Computer vision is used in military affairs to detect visual objects of certain classes: people, weapons and military equipment, military objects, etc. The input data for the analysis can be: photographs, video frames or real-time video stream obtained from space, air or ground reconnaissance means. To combat automatic object detection systems, it is possible to apply approaches capable of influencing the machine learning models used in these systems. An attack on a machine learning model is a special action to influence its elements in order to achieve the desired behavior of the system or prevent its correct operation. Based on the results of the analysis of research by various authors, it was determined that almost every machine learning algorithm has certain vulnerabilities. During the execution of tasks of engineering support of the troops regarding the camouflage of military objects, the most accessible ways of influencing computer vision systems, in order to mislead them, is to change the physical properties of the masked object by applying special coatings to its surface and materials As coatings, it is possible to use generated adversarial patch images, by superimposing or pasting them on the object, and which are capable of interfering with the work of the reconnaissance, aiming or guidance algorithms. This is especially important in the perspective of creating autonomous weapon systems capable of detecting, identifying targets and independently making decisions to destroy them.

Keywords: *artificial intelligence; machine learning; artificial neural networks; computer vision; detect; identify; classify; engineering support; camouflage of military objects; evasion attack; adversarial patch images; adversarial examples; data poisoning.*

References

1. America's eroding technological advantage: nds rdt&e priorities in an era of great-power competition with China. Avaible at: <https://govini.com/wp-content/uploads/2021/04/Govini_NDS-Priorities-RDTE.pdf> [Accessed 26 May 2023]. **2. Pro skhvalennja** Koncepciji rozvytku shtuchnogho intelektu v Ukrajini [On the approval of the Concept of the development of artificial intelligence in Ukraine]. Avaible at: <https://zakon.rada.gov.ua/laws/show/1556-2020-p> [Accessed 26 May 2023]. **3. Reding, D.F., Eaton, J.** 2020. Science & Technology Trends: 2020-2040. Exploring the S&T Edge. Avaible at: <https://www.sto.nato.int/pages/tech-trends.aspx> [Accessed 26 May 2023]. **4. Voitovsky, K. E.** 2020. Ghlobaljni trendy rozvytku nauky i tekhnologhij: novi vyklyky i mozhlyvosti [Global trends in the development of science and technology: new challenges and opportunities]. National Institute of Strategic Studies. Avaible at: <https://niss.gov.ua/doslidzhennya/nacionalna-bezpeka/globalni-trendi-rozvitku-nauki-i-tekhnologiy-novi-vikliki-i> [Accessed 26 October 2022]. **5. Lowd, D., Meek, C.** 2005. Good word attacks on statistical spam filters. Proceedings of the second conference on email and anti-spam (CEAS), pp. 1-8. **6. Biggio, B., Nelson, B. and Laskov, P.** 2012. Poisoning attacks against support vector machines. Proceeding 29th Int. Conf. Int. Conf. Mach. Learn, pp. 1467-1474. doi: 10.48550/arXiv.1206.6389. **7. Szegedy, C. et al.** 2013. Intriguing properties of neural networks. CoRR, abs/1312.6199. 10 p. doi: 10.48550/arXiv.1312.6199. **8. Carlin, N.A.** 2019. Complete list of all (arXiv) adversarial example papers. Avaible at: <https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html> [Accessed 26 May 2023]. **9. Adversa**, AI. Avaible at: <https://adversa.ai/report-secure-and-trusted-ai/> [Accessed 26 May 2023]. **10. Akhtar, N., Mian, A.** 2018. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, vol. 6, pp. 14410-14430. doi: 10.1109/ACCESS.2018.2807385. **11. Wei Hui et al.** 2022. Physical Adversarial Attack meets Computer Vision: A Decade Survey. doi: 10.48550/arXiv.2209.15179. **12. Wang, D., Yao, W., Jiang, T., Tang, G. and Chen, X.** 2022. A Survey on Physical Adversarial Attack in Computer Vision. ArXiv abs/2209.14262. doi: 10.48550/arXiv.2209.14262. **13. Suri, N. et al.** 2016. Analyzing the applicability of Internet of Things to the battlefield environment. International Conference on Military Communications and Information Systems (ICMCIS), Brussels, Belgium, May 23, pp. 1-8, doi: 10.1109/ICMCIS.2016.7496574. **14. Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.** 2022. Adversarial Attack and Defense: A Survey. Electronics, no. 11(8): 1283. doi: 10.3390/electronics11081283. **15. MITRE** ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems). Avaible at: <https://atlas.mitre.org/> [Accessed 26 May 2023]. **16. Jagielski, M. et al.** 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. IEEE Symposium on Security and Privacy, San Francisco, USA. pp. 19-35. doi: 10.1109/SP.2018.00057. **17. Ghontarenko, Ja.D., Krasnoshlyk, N.O.** 2019. Vykorystannja nejronnykh merezh dlja rozpiznavannja dij ljudyny po video [Using neural networks to recognize human actions on video.] Visnyk Cherkasjkogho nacionaljnogho universytetu imeni B. Khmeljnycjkogh, no № 2, pp. 59-72. doi: 10.31651/2076-5886-2019-2-59-72. **18. Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S.** 2019. BadNets: Evaluating backdooring attacks on deep neural networks. IEEE Access, vol. 7, pp. 47230-47244. doi: 10.1109/ACCESS.2019.2909068. **19. Zhu, X., Hu, Z., Huang, S., Li, J. and Hu, X.** 2022. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13317-13326. **20. Brown, T., Mané, D., Roy, A., Abadi, M. and Gilmer, J.** 2017. Adversarial patch. arXiv:1712.09665. **21. Zhang, Y., Foroosh, H., David, P. and Gong, B.** 2018. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. International Conference on Learning Representations. Avaible at: <https://openreview.net/pdf?id=SJgEl3A5tm> [Accessed 26 May 2023]. **22. Eykholt, K. et al.** 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1625-1634. doi: 10.48550/arXiv.1707.08945. **20. Hollander, R. et al.** 2020. Adversarial patch camouflage against aerial detection. Artificial Intelligence and Machine Learning in Defense Applications II, vol.11543. SPIE, pp. 77-86. doi: 10.1117/12.2575907. **24. Lian, J., Wang, X., Su, Y., Ma, M. and Mei, S.** 2023. CBA: Contextual Background Attack Against Optical Aerial Detection in the Physical World. IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-16, Art no. 5606616. doi: 10.1109/TGRS.2023.3264839. **25. Chen, L. and al.** 2020. Attack Selectivity of Adversarial Examples in Remote Sensing Image Scene Classification 2020. IEEE Access, vol. 8, pp. 137477-137489. doi: 10.1109/ACCESS.2020.3011639. **26. Suryanto, N. et al.** 2022. Physical camouflage attacks using differentiable transformation network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15305–15314. doi: 10.48550/arXiv.2203.09831. **27. Guidelines** for the implementation of engineering measures for the camouflage of troops and objects. 2018. Kyiv: Main Department of Operational Support of the Armed Forces of Ukraine. 6 Dec. № 90. **28. Athalye, A., Engstrom, L., Ilyas A. and Kwok, R.** 2017. Synthesizing Robust Adversarial Examples. arXiv:1707.07397. **29. On approval** of the List of information of the Ministry of Defense of Ukraine, which contains official information. 2016. Kyiv: Ministry of Defense of Ukraine. 27 Dec. № 720. **30. On approval** of the Compendium of information constituting a state secret. 2020. Kyiv: Security Service of Ukraine. 23 Dec. № 383.