УДК 004.032

DOI 10.31471/1993-9981-2024-1(52)-121-128

ДОСЛІДЖЕННЯ СТІЙКОСТІ НЕЙРОННИХ МЕРЕЖ, НАВЧЕНИХ ІЗ ВИКОРИСТАННЯМ МОДЕЛЕЙ ЗМАГАЛЬНИХ АТАК

О. В. Мойсеєнко

Івано-Франківський національний технічний університет нафти і газу,
вул. Карпатська, 15, м. Івано-Франківськ, Україна, e-mail: olena.moiseienko@nung.edu.ua

У статті проведено детальний аналіз ефективності змагального навчання для підвищення стійкості нейронних мереж до атак зловмисників у завданнях розпізнавання зображень. Розглянуто питання уразливості нейронних мереж, зокрема їхню схильність до помилкової класифікації під впливом змагальних прикладів, створених спеціально для обману моделі. Дослідження спрямоване на розробку методів навчання, які підвищують стійкість моделей до різних типів атак, зберігаючи при цьому високу якість класифікації чистих зразків. У роботі встановлено, що традиційні підходи до навчання мереж, орієнтовані на протидію лише одному типу атаки, є недостатніми для забезпечення загальної стійкості моделі. Для досягнення комплексного захисту було запропоновано використання декількох типів змагальних прикладів (FGSM, JSMA, C&W). Це дозволяє моделі формувати більш стійкі до атак уявлення даних. З метою оцінювання ефективності запропонованого підходу було проведено серію експериментів з використанням набору даних MNIST, який містить 60000 тренувальних і 10000 тестових зображень у градаціях сірого. Результати досліджень показали, що змагальне навчання значно покращує стійкість моделі до атак. Зокрема, середня точність класифікації для різних типів атак підвищується до 97,48%–97,95%, а застосування розширення даних додатково збільшує точність до 99,42%. Водночас незахищені моделі без доповнення даних демонструють вищу точність лише для окремих атак, але загальна їхня стійкість залишається низькою. Запропонований підхід також дозволяє знизити середню ефективність атак на 29,2%, при цьому зберігаючи високу точність класифікації (98,9%) для чистих зразків. Для оцінювання впливу змагального навчання була використана комбінація метрик, що враховують не лише точність класифікації, але й стійкість моделі до атак. Виявлено, що змагальне навчання сприяє покращенню узагальнюючих властивостей моделі, що дозволяє зменшити вразливість до різноманітних зловмисних введень, створених за допомогою різних атакувальних алгоритмів.

Ключові слова: Змагальні атаки, змагальний тренінг, нейронна мережа, машинне навчання, FGSM, C&W, JSMA, атака методом чорної скриньки.

The article provides a detailed analysis of the effectiveness of adversarial learning to increase the resistance of neural networks to attacks by attackers in image recognition tasks. The issue of vulnerability of neural networks is considered, in particular, their tendency to misclassification under the influence of adversarial examples created specifically to deceive the model. The research is aimed at developing training methods that increase the resistance of models to various types of attacks, while maintaining high quality classification of clean samples. The work found that traditional approaches to network training, focused on countering only one type of attack, are insufficient to ensure the overall stability of the model. To achieve comprehensive protection, the use of several types of adversarial examples (FGSM, JSMA, C&W) was proposed. This allows the model to form more resistant representations of data to attacks. In order to evaluate the effectiveness of the proposed approach, a series of experiments were conducted using the MNIST dataset, which contains 60,000 training and 10,000 test images in grayscale. The research results showed that adversarial learning significantly improves the model's resistance to attacks. In particular, the average classification accuracy for different types of attacks increases to 97.48%–97.95%, and the use of data augmentation further increases the accuracy to 99.42%. At the same time, unprotected models without data augmentation demonstrate higher accuracy only for individual attacks, but their overall robustness remains low. The proposed approach also allows to reduce the average attack efficiency by 29.2%, while maintaining high classification accuracy (98.9%) for pure samples. To assess the impact of adversarial learning, a combination of metrics was used that take into account not only classification accuracy, but also the model's robustness to attacks. It was found that adversarial learning contributes to improving the generalization properties of the model, which allows to reduce vulnerability to various malicious inputs created using different attack algorithms.

Keywords: Competitive attacks, competitive training, neural network, machine learning, FGSM, C&W, JSMA, black box attack.

Вступ

Нейронні мережі активно застосовуються дослідниками завдяки їх численним перевагам у різноманітних додатках. Вони ефективно працюють із складними багатовимірними даними, такими як зображення, відео, звукові сигнали та інші.

Однак останні дослідження показали, що нейронні мережі, включаючи найсучасніші моделі машинного навчання, вразливі до так званих «змагальних атак» [1]. Це ситуації, коли мережа робить неправильні висновки щодо даних, які дуже схожі на коректні. Крім того, було виявлено, що різні типи нейронних мереж, навіть із різними архітектурами та навчанням на різних наборах даних, схильні до однакових помилок у таких змагальних прикладах. Це свідчить про наявність фундаментальних недоліків у підходах до навчання цих моделей.

Хоча спочатку причини виникнення змагальних атак були невідомі, існують теорії, які пояснюють їх появу. Серед можливих причин називають нелінійність глибоких нейронних мереж, недоліки моделі усереднення або недостатню регуляризацію під час навчання.

Нейронні мережі, незважаючи на свою високу ефективність, залишаються вразливими до різних видів змагальних атак. Основні слабкі місця обумовлені їхньою залежністю від вхідних даних і архітектури мережі. Зокрема, вони схильні до зловмисних атак із використанням збурень, які залишаються непомітними навіть для людського ока. Це становить серйозну загрозу безпеці та викликає труднощі в реалізації програм на їх основі. У зв'язку з цим розробка ефективних методів і механізмів кіберзахисту є нагальною необхідністю.

Метою дослідження є підвищення стійкості нейронних мереж для розпізнавання зображень до кібератак. Це передбачається досягти шляхом навчання моделей із використанням сценаріїв можливих змагальних атак, а також аналізу впливу структури мережі та стратегії атак на її продуктивність.

Теоретичне обґрунтування

Оскільки вразливість нейронних мереж до атак підтверджена, дослідники активно працюють над розробкою методів, які підвищують їх стійкість до агресивних впливів. Попри численні запропоновані рішення, більшість із них виявилися недостатньо ефективними, оскільки нові методи атак або модифікації існуючих можуть обходити захисні механізми та успішно обманювати мережі [2]. Серед різних підходів найбільш перспективним виявився метод змагального навчання, який демонструє високі результати у підвищенні стійкості моделей.

Наша ідея полягає у використанні змагального навчання для підготовки мережі до протидії різним типам змагальних атак. Оскільки тренування мережі з урахуванням лише одного типу атак може збільшити її вразливість до інших видів змагальних прикладів, ми пропонуємо створити методику, яка забезпечить стійкість до кількох типів атак одночасно.

Метою розробки є створення ефективної моделі змагального навчання, яка зробить нейронну мережу менш уразливою до змагальних атак.

Для кращого моделювання сутності змагальної атаки розглядаємо приклад із задачі навчання з учителем в задачах класифікації. У цій задачі використовуються пари «об'єкт-мітка», за якими мережа навчається передбачати значення для нових об'єктів.

Якщо розглядати задачу класифікації з геометричної точки зору, метою є розділення простору таким чином, щоб новий об’єкт належав до «правильного» класу. За умови доступу до генеральної сукупності даних і роздільності класів ідеальну гіперплощину можна було б провести точно. Однак, оскільки генеральна сукупність зазвичай недоступна, використовуються алгоритми машинного навчання для наближення до цієї «ідеальної» гіперплощини на основі доступних даних.

Будь-яке відхилення цієї гіперплощини від ідеальної створює «зазор», в якому об'єкти можуть бути класифіковані некоректно. Завдання атакуючого полягає в тому, щоб змінити вектор параметрів об’єкта так, щоб він потрапив у цей зазор і був класифікований неправильно.

Змагальне навчання є одним із методів протидії таким атакам. Це підхід, у якому використовується альтернативна цільова функція для забезпечення узагальнення моделі як для звичайних, так і для змагальних даних. У нашому дослідженні ми зосереджуємося на змагальному навчанні як засобі підвищення захисних властивостей і надійності моделей машинного навчання.

У багатьох реальних умовах зловмисники можуть використовувати лише атаки типу «чорної скриньки», оскільки параметри моделі зазвичай залишаються закритими постачальником послуг. Окрім того, постійне тестування моделей може бути витратним як за часом, так і за ресурсами [3]. В таких умовах атакуючий може тренувати власну модель, імітуючи цільову мережу, та використовувати її для створення змагальних прикладів. Ці приклади потім передаються до захищуваної мережі з надією, що вона зробить помилкову класифікацію.

Для розробки ефективної моделі навчання ми припускаємо, що зловмисник прагне змусити класифікатор неправильно класифікувати вхідні дані, зарахувавши їх до будь-якого класу, окрім правильного. Для цього розглядається слабкий супротивник, який має доступ лише до вихідних даних нейронної мережі.

Зловмисник не володіє інформацією про архітектуру мережі, включаючи кількість, тип і розмір шарів, а також не має доступу до навчальних даних, використаних для налаштування її параметрів. Така ситуація відповідає класичній атаці типу «чорної скриньки», коли зловмиснику не потрібно знати внутрішню структуру системи для того, щоб спотворити її роботу.

Цільова модель. У нашому дослідженні припускається, що метою зловмисника є багатокласовий класифікатор нейронної мережі. Така модель виводить вектори ймовірностей, де кожен компонент вектора представляє ймовірність, що вхідні дані належать до одного з попередньо визначених класів. Завдання класифікатора полягає у виборі класу з найвищою ймовірністю, тоді як зловмисник прагне змінити вхідні дані так, щоб класифікатор зробив неправильний вибір.

Щоб оцінити ефективність змагального навчання у захисті від різних типів змагальних атак, ми тренуємо нейронну мережу за допомогою змагальних прикладів, створених різними відомими методами атак. Після навчання різних моделей, кожна з яких тренувалася на певному наборі змагальних прикладів, проводимо атаки чорної скриньки на ці мережі, використовуючи різні методи атак. Потім аналізуємо, наскільки успішно кожна мережа класифікує спотворені зображення.

Рівень успішності атак визначає, наскільки ефективною є атака в обмані моделі, шляхом кількісної оцінки частки атак, які досягли успіху. Він показує, у скількох випадках змагальним прикладам вдалося змусити модель машинного навчання зробити неправильну класифікацію або викликати несподівану поведінку.

Точність моделі використовується для загальної оцінки її ефективності. Вона визначається як частка правильно класифікованих екземплярів (як позитивних, так і негативних) відносно загальної кількості екземплярів. Це дає змогу зрозуміти, наскільки добре модель справляється зі звичайними й спотвореними даними, що є ключовим показником її надійності в умовах атак.

Вдалі_атаки (%) = (К-ть вдалих атак) / (Загальна к-ть атак)*100%

Точність=(w+x)/(w+x+y+z),

де w, x, y і z вказує на істинно-негативний (TN), істинно-позитивний (TP), хибно-

негативний (FN) і хибно-позитивний (FP) результати кібератаки відповідно.

Методика експерименту

Хоча змагальні приклади можна створювати з нуля, зазвичай вони формуються шляхом модифікації оригінальної вибірки таким чином, щоб цільова мережа неправильно класифікувала змінені дані. У цьому дослідженні для моделювання використано загальнодоступну базу даних MNIST, а також змагальні приклади, згенеровані трьома різними методами атак: FGSM, JSMA та методом Карліні-Вагнера (C&W).

Ці методи були обрані через те, що, хоча всі вони мають однакову мету - створити змагальні приклади, - їхні підходи до генерації даних відрізняються, що забезпечує різноманітність зображень.

«Хороший» змагальний приклад для навчання нейронної мережі - це зображення, яке візуально майже не відрізняється від оригінального, але мережа його класифікує неправильно. Використання таких прикладів у процесі навчання дозволяє моделі краще протистояти спотворенням і підвищує її стійкість до атак.

Змагальний тренінг. Припустімо, що зловмисник намагається створити змагальні приклади для обману класифікатора нейронної мережі зі зловмисною метою, наприклад, для обходу автентифікації. У реальних умовах зловмисник зазвичай не має доступу до параметрів цільової мережі. Крім того, перевірка мережі шляхом випробування великої кількості вхідних даних і спостереження за її результатами може бути обмеженою через технічні чи ресурсні причини.

Тому більш реалістичною атакою є атака чорної скриньки (або сірої скриньки), за якої зловмисник не володіє інформацією про параметри цільової мережі. У моделі атаки, розглянутій у цій роботі, передбачається, що зловмисник знає точну або приблизну архітектуру мережі, але не має доступу до її параметрів.

Для здійснення атаки зловмисник спочатку тренує власну нейронну мережу, яка має таку ж або подібну архітектуру, як і цільова мережа. Після цього на основі своєї моделі він генерує змагальні приклади, використовуючи обраний алгоритм атаки. Ці приклади потім застосовуються до цільової мережі з метою обману її класифікатора. Такий підхід дозволяє зловмиснику здійснювати атаку навіть за відсутності детальної інформації про внутрішні параметри цільової мережі.

У рамках цієї моделі атаки ми аналізуємо вплив змагального навчання на стійкість до атак. Як базу даних використовуємо MNIST [4], який є простим і широко вживаним набором даних для оцінки продуктивності алгоритмів машинного навчання. Для цільових нейронних мереж обрано дві відомі моделі: LeNet5 і ResNet. Ці ж моделі, найймовірніше, застосовуються зловмисником для створення змагальних прикладів.

Для створення змагальних прикладів атак використовуємо три алгоритми: FGSM, JSMA та C&W, як зазначалося раніше.

Процес навчання:

- для кожної з нейронних мереж проводиться навчання декількох моделей з різними наборами даних;
- одна модель навчається виключно на чистих даних із навчального набору;
- інші моделі проходять змагальне навчання, у межах якого до чистих даних додаються змагальні приклади. Наприклад, при навчанні моделі за допомогою FGSM до кожного чистого зразка з навчального набору створюється відповідний змагальний приклад. У результаті навчальний набір збільшується вдвічі порівняно з початковим.

Параметри атак:

- для FGSM використовується параметр ε, який визначає максимальну зміну кожного пікселя між оригінальним зразком і його змагальним прикладом.

У таблиці 1 наведено детальні параметри, які використовуються для

Таблиця 1 - Параметри для алгоритмів атаки та час для генерації змагальних прикладів

Алгоритм атаки	Параметри	Час генерації 60000 змагальних прикладів	
		LeNet5	ResNet
FGSM (L_∞)	ε = 0,2	48	66
JSMA (L_0)	θ= 1,0, γ= 0,2	2226	9180
C&W (L_2)	conf.: 5	8142	49,542

кожного алгоритму атаки. Ці параметри дозволяють оцінити вплив різних стратегій атак і змагального навчання на продуктивність моделей нейронних мереж.

У методі C&W (Карліні-Вагнера) параметр, який визначає рівень достовірності (confidence), відповідає за ступінь модифікації зображення. Цей параметр контролює, наскільки сильно змагальний приклад змінює оригінальне зображення для того, щоб спричинити неправильну класифікацію.

Високе значення достовірності означає, що змагальний приклад буде значно відрізнятися від оригіналу, що збільшить ймовірність того, що мережа неправильно класифікує зображення. Однак таке сильне спотворення зробить зображення менш схожим на оригінал і більш очевидним для людського спостерігача. Низьке значення достовірності дозволяє мінімізувати зміну зображення, роблячи його більш схожим на оригінал, але ймовірність того, що зображення буде неправильно класифіковане, зменшується.

Тому зловмисник може налаштувати цей параметр для досягнення бажаного балансу між ефективністю атаки (правильна класифікація) і сприйнятливістю зображення до людського ока.

Для обох моделей, LeNet5 і ResNet, використовується оптимізатор Adam із швидкістю навчання 0,0001. Процес навчання триває 50 епох для LeNet5 і 100 епох для ResNet.

Щоб компенсувати недостатню кількість даних, під час навчання часто застосовують методи доповнення даних через випадкові перетворення. Ми також використовуємо такі підходи, як випадкове обертання, зміщення та стирання. Зокрема, у кожній епосі зображення в навчальному наборі можуть випадково обертатися на кут у діапазоні [−20, 20] градусів, а також зміщуватися в будь-якому з чотирьох напрямків максимум на 6 пікселів (близько 20% ширини та висоти зображення). У випадку випадкового стирання видаляється прямокутна область довільного розміру, що охоплює до 30% пікселів.

Під час змагального тренування обертання та зміщення застосовуються лише до чистих зображень, а не до змагальних прикладів..

Стійкість мережі LeNet5

У таблиці 2 представлено результати тестування точності мереж LeNet5, навчених із різними комбінаціями чистих і змагальних прикладів. Для кожного тестового набору наведено середню та максимальну точність, отриману в проміжку між 20-ю та 50-ю епохами.

Наприклад, мережа, навчена лише на чистих зображеннях, досягає максимальної точності 99,42% на чистих тестових зразках. Однак її точність на змагальних прикладах, створених методами FGSM, JSMA та C&W, становить лише 50,23%, 64,07% і 55,28% відповідно. Це свідчить про те, що без врахування змагальних прикладів мережа легко піддається атакам, що значно знижує її точність.

Якщо мережа навчена змагальності за допомогою одного набору змагальних прикладів, вона ефективно класифікує зразки, створені цим самим методом, але не може впоратися з іншими типами атак. Наприклад, навчання на прикладах FGSM дозволяє мережі протистояти атакам FGSM, але не атакам JSMA або C&W.

Таблиця 2 - Точність тестування LeNet5 (20 - 50 епох, мережа навчається за допомогою доповнення даних, включаючи випадкове обертання, випадкове зміщення і випадкове стирання)

Навчена модель	Максимум				Медіана			
	чистий	FGSM	JSMA	C&W	чистий	FGSM	JSMA	C&W
чистий	99,42	50.23	64.07	55.28	99,23	42.01	61.06	46.07
FGSM	99,40	99,31	63,90	75,67	99,24	99,27	60,34	65.08
JSMA	99,42	50,35	96,75	81.20	99.30	48,46	96,35	75.02
C&W	99,37	74.12	69,00	99,57	99.22	61.17	63.30	99,31
FGSM + JSMA	99,40	99,32	96,29	90.05	99,28	99,26	95.15	84,94
FGSM + C&W	99,43	99,32	65,27	99,65	99,24	99,26	62,96	99,42
JSMA + C&W	99,43	73,98	96,70	98,99	99,35	70,71	95,83	98,58
FGSM + JSMA + C&W	99,36	99.30	95,82	98,81	99,27	99.21	94,92	98,39

Таблиця 3 – Точність тестування ResNet18
(40 - 100 епох (мережа навчається за допомогою доповнення даних, включаючи випадкове обертання, випадкове зміщення і випадкове стирання)

Навчена модель	Максимум				Медіана			
	чистий	FGSM	JSMA	C&W	чистий	FGSM	JSMA	C&W
чистий	99,68	70,69	80,70	93,27	99,50	65.30	76,82	91,54
FGSM	99,65	95,92	78,40	97,51	99,56	92.24	73,53	96,69
JSMA	99,66	78,71	97,78	96.13	99,56	73,85	96,80	95,42
C&W	99,62	88,99	81,37	98,60	99,52	85,71	76,29	98.11
FGSM + JSMA	99,65	95,34	96,34	98.05	99,54	93,80	95,34	97,53
FGSM + C&W	99,67	97,54	77,87	98,79	99,57	96,59	75.14	98,38
JSMA + C&W	99,69	89,80	97,00	98,63	99,56	86,57	95,57	98,25
FGSM + JSMA + C&W	99,63	97,57	96,68	98,84	99,51	96,58	95,37	98,44

Навчання на двох наборах змагальних прикладів, таких як FGSM і C&W, робить мережу стійкою до атак, створених цими методами, але вона залишається вразливою до змагальних прикладів, згенерованих іншими методами (наприклад, JSMA).

Нарешті, мережа, навчена з використанням чистих зразків разом із прикладами FGSM, JSMA та C&W, демонструє високу точність на всіх трьох типах змагальних прикладів, зберігаючи при цьому високу точність класифікації чистих зразків. Завдяки додаванню більшої кількості змагальних прикладів і розширенню навчального набору мережа може стати ще надійнішою.

Продуктивність мережі ResNet18

Оскільки результати можуть залежати від архітектури нейронної мережі, ми повторили експеримент для моделі ResNet18. На відміну від LeNet5, яка має 648,226 параметрів, ResNet18 значно більша за розміром із 11,175,370 параметрами. Загальновідомо, що ResNet18 краще справляється з навчанням складних функцій на зображеннях. У цьому експерименті як захисник, так і зловмисник використовували власні моделі ResNet18 для створення змагальних прикладів. Стратегія збільшення даних залишалася такою ж, як і для LeNet5.

Оскільки ResNet18 потребує більше епох для зближення, її навчання тривало 100 епох, а результати точності тестування оцінювалися в період між 40-ю і 100-ю епохами. Результати наведено в таблиці 3.

Як і у випадку з LeNet5, незахищена ResNet18 демонструє значно нижчу точність на змагальних прикладах (FGSM, JSMA та C&W) порівняно з чистими тестовими зображеннями. При навчанні на змагальних прикладах, створених певним методом, мережа стає стійкою до атак цього ж методу.

Змагальне навчання з використанням кількох наборів змагальних прикладів забезпечує надійність мережі, зберігаючи високу точність для чистих тестових даних. У порівнянні з LeNet5, ResNet18 демонструє вищу надійність, навіть коли не використовується змагальне навчання. Наприклад, точність для FGSM-зразків у LeNet5 становить приблизно 50%, тоді як у ResNet18 вона досягає 70%. Аналогічне підвищення точності спостерігається для інших методів атак.

Таким чином, можна зробити висновок, що без використання змагальності структура моделі та кількість її параметрів відіграють важливу роль у надійності мережі. Проте, при застосуванні змагального навчання точність класифікації змагальних прикладів у LeNet5 та ResNet18 стає приблизно однаковою.

Висновки

Запропонована модель навчання значно підвищує стійкість нейронних мереж до атак. Результати показують, що мережа, навчена протистояти певному методу атаки, здебільшого ефективна лише проти цього конкретного методу, але залишається вразливою до інших типів атак.

Доведено, що змагальне навчання з кількома типами змагальних прикладів дозволяє досягти високої точності для кожного з методів атак. Однак інші фактори, наприклад, різні стратегії захисту або атаки, можуть впливати на загальну надійність моделі. Наприклад, якщо захисник використовує доповнення даних у процесі навчання, а зловмисник ні, то точність мережі при атаках FGSM може суттєво знижуватися.

Загалом змагальне навчання не гарантує 100% захисту від усіх типів атак, зокрема атак «чорної скриньки», де доступ до цільової мережі обмежений. Експерименти показують, що мережа, навчена на двох типах змагальних прикладів (наприклад, FGSM і C&W), ефективно протистоїть цим методам, але залишається вразливою до інших атак, таких як JSMA. Мережа, навчена з використанням чистих даних, FGSM, JSMA та C&W, демонструє високу точність проти всіх трьох методів атак, зберігаючи при цьому високу точність класифікації чистих даних.

При створенні більшої кількості змагальних прикладів і використанні збільшеного набору даних мережа досягає більшої надійності. Наприклад, точність виявлення атак для мережі, навченої трьома методами (FGSM, JSMA, C&W), становить у середньому 97,48% і 97,95% для різних типів мереж. Застосування методів розширення даних може підвищити точність до 99,42%.

Однак мережа, яка не захищена і навчена без доповнення даних, показує вищу точність для змагальних прикладів усіх трьох методів атак. Модель змагального навчання загалом знижує рівень ефективності атак у середньому на 29,2% для різних зловмисних вводів, при цьому точність класифікації чистих зразків у базі даних MNIST залишається високою — 98,9%.

Список використаних джерел / References

1. Moosavi-Dezfooli S.M., Fawzi A., Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 27–30 June 2016, pp. 2574–2582.

2. Papernot N., McDaniel P., Jha S., Fredrikson, M., Celik Z.B., Swami, A. The limitations of deep learning in adversarial settings. *In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbrucken, Germany, 21–24 March 2016, pp. 372–387.

3. Madry A., Makelov A., Schmidt L., Tsipras, D., Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv 2018, arXiv:1706.06083.

4. Schott L., Rauber J., Bethge M., Brendel W. Towards the first adversarially robust neural network model on MNIST . arXiv 2019, arXiv:1805.09190/

5. Schmidt L., Santurkar S., Tsipras D., Talwar K., Madry A. Adversarially robust generalization requires more data. Adv. *Neural Inf. Process. Syst*. 2018, 31, 5014–5026.

1. Tramèr F., Boneh D. Adversarial training and robustness for multiple perturbations. *Adv. Neural Inf. Process. Syst.* 2019, 32, 5866–5876.

7 Guo C., Gardner, J.R., You Y., Wilson A.G., Weinberger K.Q. Simple Black-box Adversarial Attacks. arXiv 2019, arXiv:1905.07121.

9 LeCun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Proc. IEEE* 1998, 86, 2278–2324.

10 Chen P.Y., Sharma Y., Zhang H., Yi J., Hsieh C.J. EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. *In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18),* New Orleans, LA, USA, 2–7 February 2018.

11 Kurakin A., Goodfellow I., Bengio S. Adversarial examples in the physical world. arXiv 2017, arXiv:1607.02533.

12 Dong Y., Liao F., Pang T., Su H., Zhu, J., Hu X., Li J. Boosting Adversarial Attacks with Momentum. *In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018, pp. 9185–9193.

RESEARCH ON THE STABILITY OF NEURAL NETWORKS TRAINED USING COMPETITIVE ATTACK MODELS

O. V. Moiseienko

Ivano-Frankivsk National Technical University of Oil and Gas; Ivano-Frankivsk, Karpatska Str., 15, 76019, Ukraine, e-mail: *olena.moiseienko@nung.edu.ua*