УДК 004.8+004.056

М.Д. Кренцін
Л.М. Куперштейн

Аналіз атак на моделі машинного навчання

Вінницький національний технічний університет

Анотація

У даній роботі досліджено основні методи атак на моделі машинного навчання. Наведено класифікацію атак та розглянуто їх основні переваги.

Ключові слова: машинне навчання, цільові атаки, нецільові атаки, BlackBox, WhiteBox.

Abstract

In this works main methods of attacks on machine learning models were analyzed. The classification of attacks is given and their main advantages are considered..

Keywords: machine learning, targeted attacks, non targeted attacks, BlackBox, WhiteBox.

Вступ

На сьогоднішній день машинне навчання (МН) активно застосовується в багатьох сферах нашого життя. Такі алгоритми призначені для того, щоб фільтрувати текстові дані, розпізнавати об’єкти на зображеннях, приймати медичні рішення, навіть допомагають торгувати на фондових біржах [1]. Алгоритм МН приймає важливі рішення, тому необхідно бути впевненим, що нічого не може вплинути на правильність результатів його роботи.

Розпізнавання об’єктів на зображеннях є досить актуальною задачею сьогодення, адже воно використовується в багатьох галузях. Виявлення об’єктів на зображеннях використовується у таких задачах, як визначення автомобільного номера автомобіля, що проїхав; визначення чи присутні на зображенні люди; виявлення у реальному часі небажаних інцидентів у навколишньому середовищі, автентифікація за дактилоскопічним відбитком або зображенню обличчя тощо [2-4]. Аналізуючи різні підходи до розпізнавання зображень, зловмисники створюють відповідні атаки, що дозволяють отримати хибно негативні та хибно позитивні результати розпізнавання [5]. Для того, щоб унеможливити неправильний результат роботи алгоритмів, необхідно проаналізувати можливі види атак.

Класифікація атак

Всі атаки можна розділити на 2 класи: WhiteBox (WB) та BlackBox (BB) [6]. У випадку з WB відома вся інформація про навчену модель, тоді як у випадку з BB у є інформація лише про вхід та вихід моделі. Окремим випадком BlackBox є GrayBox, коли невідома інформація про навчену модель, але є інформація про тип алгоритму і його гіперпараметри. Але даний тип не виділяється в окремий клас, так як додаткової інформації недостатньо для переходу до WB.

Атаки також класифікуються на цільові та нецільові. Цільові атаки – це такі, що відбуваються в певному напрямку. Нехай є модель, яка дозволяє класифікувати зображення на N різних класів і певне вхідне зображення X, яке модель відносить до класу Y. Цільовою атакою буде та, що змусить зображення X віднести до визначеного класу Z, а нецільова атака може віднести до будь-якого (головне, щоб це був не клас Y).

Аналіз WhiteBox атак

На сьогоднішній день існує ряд методів атак на моделі машинного навчання. Дослідники виявляють і моделюють складні методи спотворення, руйнування і викрадення моделей і даних. Як відомо, незначні цілеспрямовані зміни в зображенні дозволяють зробити так, що система розпізнавання зображень визначить вхідне зображення зовсім не так, як передбачалось, хоча людина не зможе відрізнити змінене від початкового. Розглянемо нижче основні методи атак.

L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) атака [7]. Постановку методу L-BFGS можна виразити формулою 1.

$$minimizec|r| + loss_f(x + r, l) \quad (1)$$

З неї випливає, що ми хочемо мінімізувати функцію втрат в напрямку цільового класу з обмеженням, що внесені зміни були мінімальними. Використання L-BFGS допоможе знайти оптимальні шкідливі приклади, виходячи з наявних обмежень, але пошук такого прикладу може зайняти тривалий час і, навіть, не дати результат.

FGSM атака. Наступним етапом розвитку став метод FGSM (Fast Sign Gradient Method), який можна показати за допомогою формули 2.

$$X' = X + \varepsilon * sign(\nabla_x J(\theta, X, y)) \quad (2)$$

де X – оригінальне вхідне зображення, ε – множник, θ – параметри моделі, J – функція втрат, y – оригінальна вхідна мітка, ∇ – градієнт, $sign$ – знакова функція.

Даний метод працює набагато швидше L-BFGS. Тут беруться знаки від функції градієнта вихідної функції втрат, помноживши знак на деякий ε, і додаємо до вихідного зображення (рис. 1). До фотографії панди додається шумова карта з рівною 0.007, і виходить, що фотографія панди тепер розпізнається як Гібон з ймовірністю 99,3% [8].

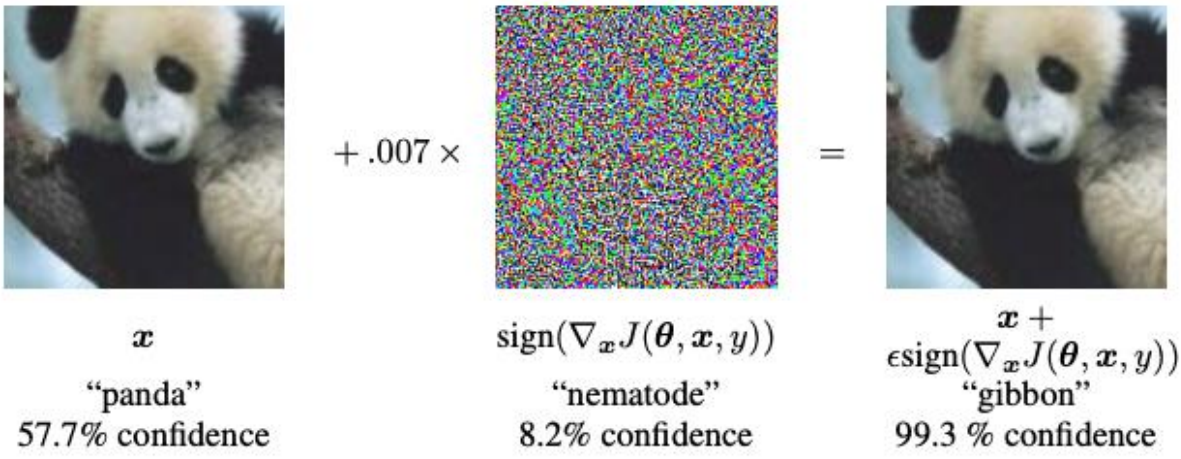

Рисунок 1 – Зашумлення зображення методом FGSM

DeepFool нецільова атака. Вона основана на відміну від попередніх в тому, що намагається зробити мінімальну шумову карту, яка «обдурить» алгоритм [9]. Метод не дозволяє зробити з одного класу якийсь конкретний, а робить будь-який інший, який найближче до вихідного зображення.

JSMA (Jacobian based saliency map) атака, в якій обчислюється пряма похідна, на підставі чого будується карта градієнтів [10]. На карті кожному параметру об'єкта за фактом відповідає внесок даного параметра в зміну кінцевого результату роботи алгоритму. Тим самим, метод дозволяє змінити якомога менше параметрів в об'єкті.

One pixel атака, суть якої полягає у тому, що знаходиться лише один піксель, який треба змінити певним чином і зображення не зможе бути розпізнане належним чином [11].

Аналіз BlackBox атак

Методи засновані на узагальненні BlackBox моделі. Маючи доступ до відправки даних в BlackBox моделі (Teacher) і доступ до виходу даної моделі, ми можемо сформувати датасет, на якому можливо навчити власну модель (Student), тим самим узагальнивши Teacher модель. Після цього можна застосувати WhiteBox атаку на Student модель, і з великою ймовірністю дана атака пройде і на Teacher моделі [12]. Імовірність такої атаки тим вище, чим більше знань про Teacher модель.

GAN-based методи. Наступним етапом розвитку BlackBox атак стали атаки, засновані на вбудовування BlackBox моделі в архітектуру генеративно-змагальної мережі (GAN), яка дозволяє генерувати нові об'єкти, які згодом будуть передані BlackBox .

Даний метод дозволяє згенерувати шкідливі приклади практично для будь-якої архітектури [13]. Для його роботи також потрібен доступ до входу і виходу моделі, що атакується.

Висновки

Було розглянуто основні методи атак на моделі машинного навчання. Проведено класифікацію методів атак за двома критеріями. Для кожного методу виявлено його особливості, переваги та недоліки. Досліджені методи є такими, що призначені для атак на моделі розпізнавання зображень. WhiteBox методи базуються на тому, що певним чином додається шум на вхідне зображення. BlackBox методи передбачають створення допоміжної моделі з аналогічними параметрами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Васюра А. С., Мартинюк Т.Б., Куперштейн Л.М. Методи та засоби нейроподібної обробки даних для систем керування: монографія. – Вінниця : УНІВЕРСУМ-Вінниця, 2008. 175 с.
2. Lotufo R.A., Morgan A.D., and Johnson AS., 1990, Automatic Number-Plate Recognition, Proceedings of the IEE Colloquium on Image analysis for Transport Applications, V01.035, pp.6/1-6/6, February 16, 1990
3. Propp M, Samal A (1992) Artificial neural network architectures for human face detection. In: Proceeding ofartificial neural networks in engineering, vol 2, pp 535–540
4. Кренцін М.Д., Куперштейн Л.М., Штокал А.С., Восьмушко О.В. Система підтримки роботи ситуаційного центру на основі інтелектуальних хмарних технологій. Збірник матеріалів міжнародної науково-технічної конференції молодих вчених, аспірантів та студентів CSYSC-2018. м. Івано-Франківськ, 2018. с.119-120
5. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, «Multi-pie,» Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.
6. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, «Practical black-box attacks against machine learning,» in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, 2017, pp. 506–519
7. Morales, J. L.; Nocedal, J. (2011). «Remark on «algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization»». ACM Transactions on Mathematical Software. 38: 1–4. doi:10.1145/2049662.2049669. S2CID 16742561
8. Adversarial example using FGSM URL: https://www.tensorflow.org/tutorials/generative/adversarial_fgsm
9. S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, «Deepfool: a simple and accurate method to fool deep neural networks» in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582
10. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. IEEE, 2016, pp. 372–387
11. J. Su, D. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. arXiv preprint arXiv:1710.08864, 2017
12. Practical Black-Box Attacks against Machine Learning URL: https://arxiv.org/pdf/1602.02697.pdf
13. UPSET and ANGRI : Breaking High Performance Image Classifiers URL: https://arxiv.org/pdf/1707.01159.pdf

Кренцін Михайло Дмитрович, аспірант кафедри захисту інформації, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: mishatron98@gmail.com

Куперштейн Леонід Михайлович, доцент кафедри захисту інформації, Вінницький національний технічний університет, Вінниця, e-mail: kupershtein.lm@gmail.com

Mykhailo Krentsin, PhD student of the Department of Information Protection, Faculty for Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: mishatron98@gmail.com

Leonid Kupershtein, Associate Professor of the Department of Information Protection, Vinnytsia National Technical University, Vinnytsia, e-mail: kupershtein.lm@gmail.com