Міністерство освіти і науки, молоді та спорту України
Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики
Кафедра обчислювальної математики

Курсова робота

на тему:

"Розробка алгоритмів захисту від атак на глибокі нейронні мережі"

Виконав:
студент IV курсу групи ПМп-41
напрямку підготовки (спеціальності)
113 – "Прикладна математика"
Бугрій Б.О.

Науковий керівник:
доц. Музичук Ю.М.

Львів - 2021

Зміст

Вступ

Сьогодні розумні системи штучного інтелекту є невід'ємною частиною життєдіяльності суспільства. Завдяки таким системам людству вдалося досягти значних результатів у багатьох сферах та галузях, таких як, наприклад, медицина, програмна інженерія, машинобудування та робототехніка.

Зараз найбільш поширеним типом алгоритмів машинного навчання є глибокі нейронні мережі, які здатні знаходити закономірності у великих масивах даних, а результати їх роботи часто перевершують людей. Проте, як показали нещодавні дослідження [1], такі алгоритми часто використовують антиінтуїтивні, в порівнянні з смисловим значенням, закономірності стосовно певних характеристик. Через це нейронні мережі є вразливими до різного роду зловмисних втручань, які можуть призвести до невірних результатів.

Моделі, які показують відмінні результати на звичайних даних, можуть бути легко ошуканими зразками, які лише трохи відрізняються від правильно класифікованих прикладів. Це розкриває фундаментальні недоліки в алгоритмах машинного навчання, які можуть бути використані зловмисниками для заподіяння шкоди, що може вплинути на безпеку технологічних процесів людської життєдіяльності і призвести до катастроф великого маштабу.

... TODO

Щоб ефективно захиститись від можливих загроз, потрібно знати слабкі місця алгоритмів та стратегії нападу, що використовують зловмисники. Саме тому ми розглянемо деякі підходи до атак на нейронні мережі, знання яких дозволить покращити системи машинного навчання, зробити їх більш стійким до зловмисних втручань.

... TODO

1 Опис проблеми

1.1 Постановка задачі

Нехай система машинного навчання M на основі зразків $x \in S$ робить передбачення y. Тут S – множина всеможливих зразків-зображень з предметної області, які допустимі для використання моделлю M.

Означення 1.1 *Зразок* $x^{adv} = x + \tau$, $x^{adv} \in S$ *називається ошукуючим якщо для достатньо малих збурень* τ *виконуються такі умови:*

$$M(x) = y_{true} \tag{1.1}$$

$$M(x^{adv}) \neq y_{true} \tag{1.2}$$

де y_{true} *- правильне передбачення.*

Нашою метою є побудова максимально ефектвної моделі машинного навчання, яка буде менш вразливою до такого типу ошукуючих зразків.

Задачу захисту моделі від ошукуючої атаки можна формалізувати наступним чином. Нехай $S^{adv}(M) \subset S$ – множина ошукуючих зразків для моделі M. Необхідно знайти модель M' яка є, в певному сенсі, модифікацією оригінальної моделі, таку, що

$$S^{adv}(M') = \emptyset. \tag{1.3}$$

Такий "ідеальний" випадок є практично неможливим, тому ми будемо використовувати пом'якшений варіант, в якому задовільнятиметься умова

$$n(S^{adv}(M')) < n(S^{adv}(M)) \tag{1.4}$$

де $n(S)$ –кількість елементів в множині S.

1.2 Альтернативна постановка задачі

1.3 Пов'язані поняття

2 Огляд атак на моделі машинного навчання

Майже всі атаки грунтуються на додаванні до оригінальних зразків (зображень) шуму τ, після чого отримуємо ошукуючий зразок x^{adv} такий, що

$$\begin{aligned} M(x^{adv}) &\neq y_{true}, \\ x^{adv} &= x + \tau, \end{aligned} \tag{1}$$

де M - модель машинного навчання, y_{true} - правильне передбачення. Таким чином ми можемо ошукати нейронну мережу. В залежності від того, який результат ми бажаємо отримати, чи якими даними ми володіємо, атаки можна поділити на умовні групи, як, наприклад, атаки з ціллю та без цілі, або атаки на чорну та білу скриньки [6].

Нейронні мережі є особливо чутливими до атак на білу скриньку. Отримавши доступ до всіх параметрів, у зловмисників з'являється можливість обчислити точні похідні для функцій (у випадку диференційовності моделі), за допомогою яких виконується класифікація. Тоді під час атаки можна застосувати різні чисельні методи для вирішення конкретної проблеми, що дозволяє досягти дуже високої ефективності згенерованих ошукуючих зразків.

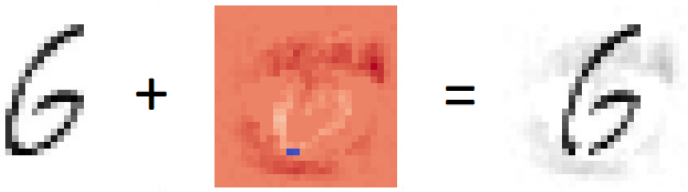

Рисунок 1. Початкове зображення класифікується нейронною мережею як "шість". Після додавання до нього певного шуму НМ класифікує його як "п'ять".

Метод швидкого градієнтного спуску є одним з найпростіших методів генерації ошукуючих прикладів. Його застосовують для атак на різні системи машинного навчання, зокрема й на нейронні мережі [2].

Основною метою методу є максимізація функції витрат J, розглядаючи її як $J(x, y_{true})$, зафіксувавши ваги нейронної мережі. В такому випадку потрібно знайти x^{adv} таке, що виконується нерівність:

$$J\left(x^{adv}, y_{true}\right) > J\left(x, y_{true}\right) \tag{2}$$

Тоді, згідно з [3], ми зможемо знайти розв'язок поставленої задачі. Пошук ошукуючих зразків здійснюємо за формулою:

$$x^{adv} = x + \epsilon sign\left(\nabla_x J\left(x, y_{true}\right)\right) \tag{3}$$

де ϵ - додатній параметр, який називатимемо *розміром кроку*. Напрямок $sign\left(\nabla_x J\left(x, y_{true}\right)\right)$ є *напрямком зростання* для функції J.

Варто зауважити, що метод не гарантує знаходження ошукуючого зразка, бо нерівність (2) свідчить лише про збільшення функції правдоподібності. Тому в загальному випадку доцільно використовувати ітераційну форму методу. Також можна застосовувати інші методи, такі як [6], для того, щоб зменшити розмір шуму τ.

Більшість реальних систем, які використовують нейронні мережі, не розголошують своїх конфігурацій (як от структура мережі і її ваги), тому атаки, накшталт швидкого градієнтного спуску не так просто застосувати на практиці. В деяких сценаріях зловмисники мають можливість використовувати модель для отримання передбачень на основі власних зразків (як от, наприклад, при використанні програмних інтерфейсів від Google AI чи Amazon). Запити до мережі дозволено повторювати необмежену кількість разів і на основі результатів покращувати ошукуючі зразки.

Тоді є можливість обчислити градієнт наближено, для чого достатньо лише результату класифікації $f(x)$ та відповідного вхідного параметра x, як показано в [5]. Якщо обчислимо часткові похідні наближено, за формулою

$$\frac{\partial J(\mathbf{x})}{\partial \mathbf{x}_i} \approx \lim_{h \to 0} \frac{J\left(\mathbf{x} + he_i\right) - J\left(\mathbf{x} - he_i\right)}{2h}, \tag{4}$$

де e_i - вектор, в якому i-тий елемент дорівнює одиниці, а всі решта – нулю, то зможемо застосувати для атаки на чорну скриньку алгоритми, що використовують похідні для розв'язання задачі оптимізації, такі як швидкий градієнтний спуск.

При використанні наближених значень похідних в методі швидкого градієнту (3) вдається досягти результативності, близької до тої, яку отримуємо під час атак на білу скриньку. Це зумовлено тим, що для ефективної роботи методів достатньо лише знаку градієнту. Те, що значення обчислені з певною похибкою, значним чином не впливає на результат. Недоліком такого сценарію є необхідність виконувати класифікацію значну кількість разів, що займає багато часу та ресурсів і може бути легко виявлено.

TODO розписати детальніше, можливо поділити на розділи

2.1 Атаки на "відкриті" нейронні мережі

2.2 Атаки на "закриті" нейронні мережі

3 Принципи захисту від ошукуючих атак

3.1 Боротьба з штучно створеним шумом

3.2 Надмірне тренування як причина вразливості

4 Захисне випаровування та його модифікації

4.1 Ідея методу та процес тренування

4.2 Оцінка захисту

5 Методи з гарантованою стійкістю

5.1 Захист PixelDP

5.2 Оцінка захисту

6 Висновок

Література

[1] *Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus* / Intriguing properties of neural networks / arXiv preprint arXiv:1312.6199 (2014)

[2] *Ian Goodfellow, Jonathon Shlens, Christian Szegedy* / Explaining and Harnessing Adversarial Examples / arXiv preprint arXiv:1412.6572 (2014)

[3] *Alfio Quarteroni, Riccardo Sacco, Fausto Saleri* / Numerical Mathematics / – Springer, 2000. –300 p.

[4] *Nicolas Papernot, Patrick McDaniel, Ian Goodfellow* / Transferability in machine learning: from phenomena to black-box attacks using adversarial samples / arXiv preprint arXiv:1605.07277 (2016)

[5] *Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh* / ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models / arXiv preprint arXiv:1708.03999 (2017)

[6] *Nicholas Carlini, David Wagner* / Towards Evaluating the Robustness of Neural Networks / arXiv preprint arXiv:1608.04644 (2017)