Застосування та методи підвищення стійкості моделей у системах комп'ютерного зору

Сергій Позднякович, студент[1] (ORCID: 0009-0006-7765-8339)
Тетяна Гончаренко, д-р. техн. наук, проф.[1] (ORCID: 0000-0003-2577-6916)

[1] *Київський національний університет будівництва і архітектури, Київ, Україна*

АНОТАЦІЯ

Робота присвячена аналізу впливу змагальних атак на системи комп'ютерного зору та дослідженню методів підвищення стійкості моделей до таких атак. Проаналізовано особливості побудови змагальних атак, їх вплив на роботу нейронних мереж і розглянуто сучасні підходи до підвищення стійкості моделей.

Ключові слова: змагальні атаки, комп'ютерний зір, нейронні мережі.

1. ВСТУП

Системи комп'ютерного зору широко застосовуються у різних сферах — від автономного транспорту та робототехніки до систем безпеки та біометричної ідентифікації. Високі результати, яких досягають глибокі нейронні мережі, зробили ці технології важливою частиною сучасних рішень. Однак водночас було виявлено їхню вразливість: навіть незначні, на перший погляд, зміни у вхідних зображеннях можуть кардинально вплинути на роботу системи.

Такі зміни називають змагальними прикладами (adversarial examples), а методи їх створення — змагальними атаками. Вони становлять серйозну загрозу, адже можуть бути використані для обходу систем розпізнавання облич, введення в оману систем контролю автономних автомобілів. Тому завдання розробки методів підвищення стійкості моделей до подібних атак є надзвичайно актуальним [1].

2. МЕТА РОБОТИ

Метою роботи є дослідження впливу змагальних атак на моделі комп'ютерного зору та оцінка ефективності різних методів захисту.

3. ЗМАГАЛЬНІ АТАКИ

Змагальні атаки — це спеціально створені вхідні дані, які призначені для зловмисного введення моделі машинного навчання в оману. Основна ідея полягає в тому, що навіть невеликі й непомітні для людини зміни у зображенні можуть призвести до суттєвих помилок класифікації. Така властивість є серйозним викликом для систем комп'ютерного зору, особливо у критичних сферах — транспорті, охороні чи біометричній ідентифікації [1].

Змагальні атаки можуть мати різний характер і створюватися за різними принципами:

Атаки з доступом до моделі (white-box attacks).

У цьому випадку зловмисник має повний доступ до архітектури та параметрів моделі, включаючи ваги та градієнти. Це дозволяє будувати найбільш ефективні та точні збурення. Прикладами таких атак є FGSM (Fast Gradient Sign Method) [1] та PGD (Projected Gradient Descent) [2]. FGSM використовує одноразове додавання шуму у напрямку градієнта, тоді як PGD застосовує багатокрокову ітераційну оптимізацію, роблячи атаку значно потужнішою.

Атаки без доступу до моделі (black-box attacks).

Зловмисник не має знання структури нейронної мережі, але може надсилати запити та отримувати відповіді. У цьому випадку часто використовують так званий ефект transferability — здатність змагальних прикладів, створених для однієї моделі, «переноситися» на інші архітектури та залишатися ефективними [3]. Це робить black-box атаки небезпечними навіть тоді, коли сама модель є конфіденційною.

Оптимізаційні атаки.

Серед них виділяється CW-атака (Carlini & Wagner), яка формулюється як задача оптимізації: знайти мінімальне збурення, яке призведе до неправильної класифікації. Цей метод дозволяє отримувати надзвичайно «непомітні» для людини приклади, що ускладнює їхнє виявлення.

Фізичні атаки.

На відміну від цифрових, вони створюються у реальному середовищі. Прикладом може бути роздруковане зображення зі збуреннями або спеціально створені стікери, які наклеюються на об'єкт. Такі атаки особливо небезпечні для систем відеоспостереження чи автономних автомобілів, оскільки працюють у реальних умовах і не потребують прямого втручання у дані.

Цільові та нецільові атаки.

У цільових атаках зловмисник намагається змусити модель класифікувати об'єкт як конкретний інший клас (наприклад, дорожній знак «Стоп» як «Обмеження швидкості»).

У нецільових атаках важливим є лише те, щоб модель зробила помилку, незалежно від того, який саме клас буде вибраний.

Атаки з обмеженнями.

При побудові змагальних прикладів часто накладаються обмеження, щоб збурення було малопомітним. Для цього використовують різні норми (L_0, L_2, L_∞), які визначають «розмір» внесених змін.

4. МЕТОДИ ЗАХИСТУ

У відповідь на загрозу змагальних атак розроблено низку підходів, спрямованих на підвищення стійкості моделей комп'ютерного зору. Вони різняться за складністю реалізації, обчислювальними витратами та ефективністю. Основні напрями можна поділити на кілька груп:

Adversarial Training (змагальне тренування).

Полягає у додаванні до навчальної вибірки змагальних прикладів поряд зі «звичайними» зображеннями. Це

дозволяє моделі навчитися правильно класифікувати навіть спотворені вхідні дані. Недоліком методу є значне зростання часу навчання та обчислювальних витрат.

Defensive Distillation (захисна дистиляція).

Використовує ідею навчання моделі на «м'яких» вихідних значеннях (soft labels), отриманих від попередньо тренованої мережі. Такий підхід зменшує чутливість до малих збурень, але може втрачати ефективність проти сильніших атак, зокрема ітеративних.

Регуляризація та згладжування.

До цього напряму належать методи, що обмежують надмірну чутливість моделі: використання dropout, weight decay, а також label smoothing. Такі методи зменшують переобучення та роблять межі прийняття рішень більш «плавними».

Методи попередньої обробки даних.

Застосування згорткових фільтрів, обтинання високочастотних складових або випадкових перетворень зображення (наприклад, випадкове масштабування чи поворот) може ускладнити атаку. Проте такі методи часто лише частково ефективні, адже потужні алгоритми атак здатні врахувати навіть ці зміни.

Гібридні та сучасні підходи.

Новітні дослідження зосереджуються на поєднанні кількох захисних стратегій. Наприклад, TRADES (TRadeoff-inspired Adversarial DEfense) балансує між точністю на «чистих» і «змагальних» прикладах, тоді як методи на основі сертифікованої стійкості намагаються математично гарантувати певний рівень захисту.

5. АНАЛІЗ АТАК

У системах комп'ютерного зору змагальні атаки є однією з ключових проблем, які значно знижують надійність та практичну цінність сучасних моделей. Висока розмірність простору ознак і нелінійність внутрішніх представлень роблять нейронні мережі вразливими до навіть незначних змін у вхідних даних, що залишаються непомітними для людини [1]. У результаті класифікатор або детектор може видавати некоректний результат, який атакувальник може використати у власних цілях.

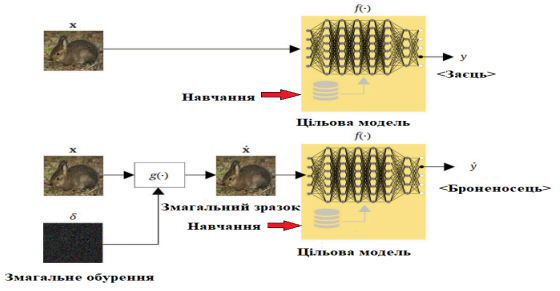

Рис.1 Наочний приклад змагальної атаки на класифікатор зображень. Класифікатор атакованої моделі присвоює вихідному зображенню позначення «заєць» (вгорі), а обманному зображенню, отриманому за допомогою атаки базового ітеративного методу та ж цільова модель присвоює позначення «броненосець» (внизу). Цільова модель являє собою ілюстративне та абстрактне представлення класифікатора Inception V3, навченого в ImageNet. Варто відзначити, що з метою наочності показане тут спрямоване збурення в 20 разів більше, ніж реальне

Теоретично доведено, що абсолютний захист від змагальних атак практично неможливий, оскільки будь-яка модель у високовимірному просторі матиме «слабкі місця». Тому завдання полягає не у створенні повністю стійкої системи, а у зменшенні ймовірності успішної атаки та ускладненні процесу її реалізації. Для цього застосовуються різні методи. Найбільш відомим підходом є змагальне навчання, коли до тренувального набору додають приклади з атаками [2]. Це дозволяє підвищити стійкість, однак лише щодо обмеженого класу атак, подібних до тих, що використовувались у процесі навчання. Іншим напрямом є регуляризація та використання різних методів нормалізації, які знижують чутливість моделі до локальних змін у даних. Також активно досліджуються механізми виявлення атак, які передбачають додаткові модулі, здатні розпізнати підозрілі або змінені вхідні приклади ще до того, як вони потраплять у модель.

У перспективі найбільш значущим напрямом розвитку є створення сертифікованих моделей, які здатні гарантувати стійкість у межах заданого збурення. Іншим важливим вектором є розробка гібридних стратегій, що поєднують кілька методів одночасно, а також формування єдиних стандартів оцінки безпеки, які дозволять системно порівнювати різні рішення. Таким чином, проблема змагальних атак у комп'ютерному зорі залишається відкритою та вимагає комплексного підходу, що поєднує алгоритмічні, математичні й концептуальні засоби протидії.

6. ВИСНОВКИ

1) У роботі було розглянуто проблему змагальних атак у системах комп'ютерного зору та проведено аналіз існуючих підходів до їх реалізації й протидії. Теоретичний огляд підтвердив, що нейронні мережі залишаються вразливими до навіть незначних змін у вхідних даних, що робить їх ціллю для атакуючих [1].
2) Було описано ключові методи захисту, серед яких змагальне навчання, регуляризація та методи виявлення збурень. Показано, що жоден з них не є універсальним, а ефективність кожного залежить від типу атаки та умов застосування.
3) Проведений аналіз дозволив зробити висновок, що між стійкістю та точністю існує компроміс: підвищення захисту часто супроводжується зниженням якості класифікації на «чистих» даних [2]. Це обмежує практичну придатність багатьох сучасних методів.
4) Перспективними напрямами подальших досліджень є розробка сертифікованих моделей з формальними гарантіями стійкості, створення гібридних стратегій захисту та уніфікація стандартів оцінки безпеки . Такі підходи здатні забезпечити більш системний та надійний захист від широкого спектра атак.

Список літератури

1) Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations, 2015.

2) Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations, 2018.

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