УДК 004.056

Влияние атак FGSM, CW, BIM, PGD, SMP на предсказательную способность моделей машинного обучения

Сивков Д.И. (ИТМО)
Научный руководитель – к.т.н., доцент ФБИТ, Воробьева А.А. (ИТМО)

Введение. Тема, связанная с безопасностью систем, основанных на методах искусственного интеллекта с каждым годом, очевидно, становится все более важной. Особое внимание необходимо уделять устойчивости моделей машинного обучения к атакам отравления, где злоумышленник внедряет зловредные (состязательные) примеры в обучающий набор данных и атакам уклонения, где происходит манипулирование входными данными во время тестирования обученной модели машинного обучения.

В данной работе рассматривается влияние наиболее известных состязательных атак, таких как Fast Gradient Signed Method (FGSM), Carlini & Wagner (CW), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), и Saliency Map Method (SMP), на предсказательную способность моделей машинного обучения.

Основная часть. Рассматриваются особенности таких состязательных атак как:

1) Fast Gradient Signed Method (FGSM) – метод атаки, который использует градиенты модели машинного обучения для создания искаженных входных данных.
2) Carlini & Wagner (CW) – более сложный метод состязательной атаки, направленный на минимизацию различия между оригинальным и искаженным изображениями при одновременном обеспечении классификации искаженного образца в неверный класс.
3) Basic Iterative Method (BIM) – усовершенствованная версия FGSM, которая применяет множественные малые изменения вместо одного большого изменения.
4) Projected Gradient Descent (PGD) – считается одной из самых мощных техник состязательной атаки.
5) Saliency Map Method (SMP) – метод, использующий карты значимости для определения наиболее влиятельных пикселей изображения для изменения классификации.

Выводы.

Состязательные атаки представляют угрозу для систем машинного обучения, способную нарушить их надежность и эффективность. Исследование влияния атак FGSM, CW, BIM, PGD, и SMP на предсказательную способность моделей показывает, что даже современные и продвинутые алгоритмы машинного обучения уязвимы перед подобранными искажениями входных данных.

Список использованных источников:

1) Bader Rasheed, Adil Khan IMPROVING ROBUSTNESS OF DEEP NETWORKS USING CLUSTER-BASED ADVERSARIAL TRAINING // Russian Law Journal. 2023. №9S. URL: https://cyberleninka.ru/article/n/improving-robustness-of-deep-networks-using-cluster-based-adversarial-training (дата обращения: 15.02.2024).
2) Li Huayu, Namiot Dmitry A SURVEY OF ADVERSARIAL ATTACKS AND DEFENSES FOR IMAGE DATA ON DEEP LEARNING // International Journal of Open Information Technologies. 2022. №5. URL: https://cyberleninka.ru/article/n/a-survey-of-adversarial-attacks-and-defenses-for-image-data-on-deep-learning (дата обращения: 15.02.2024).
3) Ауси Рим Мохаммед Худхейр, Заргарян Елена Валерьевна, Заргарян Юрий Артурович ГЛУБОКОЕ ОБУЧЕНИЕ МЕТОДАМ ЗАЩИТЫ ОТ АТАК // Известия ЮФУ. Технические науки. 2023. №2 (232). URL: https://cyberleninka.ru/article/n/glubokoe-obuchenie-metodam-zaschity-ot-atak (дата обращения: 15.02.2024).

Сивков Д.И. (автор)
Воробьева А.А. (научный руководитель)