УДК 004.056.2

ПОДХОДЫ К ОБНАРУЖЕНИЮ АТАК НА ОСНОВЕ ВРЕДОНОСНЫХ ВОЗУМУЩЕНИЙ НА НЕЙРОННЫЕ СЕТИ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

Бучаев А.Я. (Университет ИТМО), **Есипов Д.А.** (Университет ИТМО), **Роговой В.** (Университет ИТМО)

Научный руководитель – доцент, кандидат технических наук Попов И.Ю. (Университет ИТМО)

Введение. В современном мире системы искусственного интеллекта все чаще используются для решения задач обработки изображений, в том числе классификации и детектирования объектов. Существует ряд атак [1], предполагающих внесение вредоносных возмущений в обрабатываемые данные для провоцирования некорректного отклика модели или встраивания в нее бэкдора. В противодействие таким атакам и в целях обеспечения целостности и верификации моделей в некоторых случаях используют другие нейронные сети, однако при таком подходе возможно сохранение уязвимости ввиду ее связи с используемой технологией. В текущей работе предлагается использование статистических методов для обнаружения вредоносных возмущений во входных данных модели и верификации моделей машинного обучения.

Основная часть. В текущей работе для обнаружения вредоносных возмущений предполагается статистический анализ компонент изображения. Следует отметить, что конкретный подход к обнаружению зависит от типа вносимого возмущения.

В данной работе предполагается анализ изображений в том числе обучающей выборки статистическими и вероятностными методами. Модификация изображений посредством внесения шума существенно искажает статистическое распределение шумовой компоненты исследуемого изображения.

Предполагаемый алгоритм обнаружения вредоносных возмущений включает в себя следующие шаги:

1) Снятие шумовой компоненты исследуемого изображения.
2) Применение преобразования Фурье для дальнейшей обработки.
3) Полученный результат преобразования Фурье предположительно имеет нормальное распределение, которое может быть оценено при помощи нечеткого байесовского классификатора [2] для выявления аномалий и отклонений, вызванных наложением стороннего шума.
4) В случае отсутствия преобразования предполагается анализировать шумовую компоненту, характеризующуюся равномерным распределением, оценка которого может быть получена посредством анализа относительной и абсолютной однородности фреймов шумовой маски.

Также возможно применение описанного подхода к выходным весам нейронной [3] сети для обнаружения бэкдора системах искусственного интеллекта обработки изображений. Кроме того, возможна адаптация предложенного подхода к другим видам атак, в том числе к однопиксельным атакам и атакам уклонения [4].

Выводы. В текущей работе представлен подход по обнаружению атак на основе внесения вредоносных возмущений в изображения. В дальнейшей работе предполагается построение обобщенной зависимости, охватывающей различные случаи наложения вредоносных возмущений, а также написание программных инструментов, позволяющих провести анализ сформулированных гипотез.

Список использованных источников:

1. Huang X. et al. A survey of safety and trustworthiness of deep neural networks:

Verification, testing, adversarial attack and defence, and interpretability //Computer Science Review. – 2020. – Т. 37. – С. 100270.

2. T. C. Glenn, A. Zare and P. D. Gader, "Bayesian Fuzzy Clustering," in IEEE Transactions on Fuzzy Systems, vol. 23, no. 5, pp. 1545-1561, Oct. 2015, doi: 10.1109/TFUZZ.2014.2370676.

3. Owen Shen, "Interpretability in ML: A Broad Overview" [Электронный ресурс] The Gradient, 2020. URL: https://thegradient.pub/interpretability-in-ml-a-broad-overview. (Дата обращения: 14.02.2023).

4. Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2020. An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 218–228. https://doi.org/10.1145/3394486.3403064

Бучаев А.Я.	Подпись
Есипов Д.А.	Подпись
Роговой В.	Подпись
Попов И.Ю. (научный руководитель)	Подпись