УДК 004.896

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ НАДЕЖНОСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ОТ ВОЗДЕЙСТВИЯ ФИЗИЧЕСКИМИ ШАБЛОНАМИ УКЛОНЕНИЯ

Вандышев К.А., Житихин А.Е. , Менисов А.Б. (Военно-космическая академия им. А.Ф.Можайского)

Научный руководитель – к.т.н., Менисов А.Б.

(Военно-космическая академия им. А.Ф.Можайского)

Введение. В настоящее время существуют прикладные системы в различных направлениях, использующие технологию искусственного интеллекта (далее ИИ). К примеру, распознаватели кодов товарных номенклатур, детекторы негабаритных рудных объектов на конвейерных лентах, интеллектуальные химические лаборатории по составлению формул сложных веществ, автоматизированные системы охраны и многие другие. В условиях общего роста применения технологии ИИ, необратимо повышается количество компьютерных инцидентов в таких системах. Для решения вопроса безопасности интеллектуальных систем необходимо создать средство тестирования построенных моделей для дальнейшего их дообучения и повышения устойчивости к атакам уклонения, в частности.

Основная часть. Существует несколько основных классов атак на развернутые модели ИИ:

1. Атаки одним пикселем [1];
2. Зашумление исходной фотографии [2];
3. Физические атаки уклонения[5];
4. Backdoor-атаки[6];
5. Отравление данных;
6. Реверс-инжиниринг моделей ИИ.

Помимо этого, традиционно рассматриваются две модели совершения атаки[3]:

1. Атаки в условиях, когда структура атакуемой системы неизвестна, но известны входные и выходные данные (Black box);
2. Атаки в условиях, когда известна определённая часть структуры и параметров целевой системы (Gray box);
3. Атаки в условиях, когда структура известна и на каждом этапе совершения атаки наблюдаемы параметры целевой системы (White box).

В данной работе на практике рассматривается тестирование разработанной системы распознавания стрелкового оружия, в основе которой лежит модель нейронной сети семейства YOLO (You Look Only Once), обученная распознаванию оружия. Обучающая выборка не предполагает наличия состязательных данных. Для проведения эксперимента также были разработаны средства тестирования вышеописанными методиками.

Ценность работы заключается в возможности повышения стойкости моделей нейронных сетей и моделей машинного обучения состязательными данными[4], полученными с помощью разработанных средств тестирования целевых систем. Важным свойством полученных средств тестирования является универсальность, то есть средства тестирования не зависят от аппаратных и программных средств функционирующей системы.

Выводы. Проведен анализ существующей проблемы, составлен математический аппарат программного решения, разработан прототип функционирующей системы и средства тестирования. Полученные результаты позволяют судить о высокой применимости данного программного комплекса в области безопасности технологии ИИ.

Список использованных источников:

1. Jiawei Su, Danilo Vasconcellos Vargas and Kouichi Sakurai. One Pixel Attack for Fooling Deep Neural Networks // [Электронный ресурс] https://arxiv.org/pdf/1710.08864.pdf;
2. Alessandro Erba, Riccardo Taormina, Stefano Galelli, Marcello Pogliani, Michele Carminati, Stefano Zanero, Nils Ole, Politecnico di Milano. Real-time Evasion Attacks with Physical Constraints on Deep Learning-based Anomaly Detectors in Industrial Control Systems // [Электронный ресурс] https://d1wqtxts1xzle7.cloudfront.net/96160878/;
3. Костюмов В.В. Обзор и систематизация атак уклонением на модели компьютерного зрения // [Электронный ресурс] https://cyberleninka.ru/article/n/obzor-i-sistematizatsiya-atak-ukloneniem-na-modeli-kompyuternogo-zreniya/viewer/;
4. Hanjie Chen, Yangfeng Ji. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation // [Электронный ресурс] https://arxiv.org/abs/2203.12709/.
5. Yan Zhang1† , Yi Zhu2† , Zihao Liu1 , Chenglin Miao1∗ , Foad Hajiaghajani2 , Lu Su3 , Chunming Qiao2 https://www.acsu.buffalo.edu/~yzhu39/Yi_Zhu_homepage_files/papers/SenSys22.pdf
6. Wang B. et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks //2019 IEEE Symposium on Security and Privacy (SP). – IEEE, 2019. – С. 707-723.