УДК 004.056.2

ИССЛЕДОВАНИЕ МЕТОДА ДЕТЕКТИРОВАНИЯ АТАК ОТРАВЛЕНИЯ ДАННЫХ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ИЗОБРАЖЕНИЙ

Вавилова А.С. (Университет ИТМО)
Научный руководитель – кандидат технических наук, доцент Волошина Н.В.
(Университет ИТМО)

Введение. Эффективность применения алгоритмов машинного обучения в задачах распознавания изображений основывается на способности моделей обучаться на большом количестве источников данных. Использование открытых источников при формировании датасетов для обучения открывает перед злоумышленником дополнительный возможный вектор проведения атак отравлениях данных [1]. Отравление обучающих данных приводит к снижению точности распознавания и появлению большое количества некорректных прогнозов. Реализация такой атаки на этапе обучения для алгоритма машинного обучения, работающего в условиях реального производственного процесса, может быть критическим с точки зрения нарушения безопасности системы и получения ощутимых убытков [2].

Основная часть.
Основные меры по снижению эффективности атак отправления данных основываются на: выявлении аномальных обучающих данных при возникновении дрейфа [3] (потери способности прогнозирования при поступлении определенных данных на входе модели), отслеживании происхождения данных на входе (анализ меток данных) или разделение датасетов на версии для обучения для отслеживания аномальных результатов [4]. Такие механизмы безопасности неэффективны и ресурсоемки в условиях необходимости обучения модели на большом объеме данных, а также при получении несбалансированных данных (неполного набора данных с точки зрения возможных вариаций, например, исчерпывающий набор ракурсов головы человека и результатов применения световых эффектов на лице) [5].
В условиях необходимости обучения системы распознавания изображений на большом объеме данных предлагается исследование возможности применения механизма проверки корректности работы модели после завершения процесса обучения, до момента применения алгоритма машинного обучения в реальном производственном процессе.
Предлагаемый метод детектирования атак отправления данных основывается на использования дополнительной модели, меток и этапе сравнения векторов меток результатов работы исходного алгоритма и дополнительного.
Одним из заключительных этапов работы является проведение экспериментального сравнения точности распознавания моделей на тестовом наборе данных.

Выводы. В условиях большого количества обучающих данных из открытых источников реализация атаки отравления данных приводит к некорректной работе алгоритма машинного обучения. Исследование методов детектирования атак отравления данных и внедрение соответствующих дополнительных механизмов безопасности является основой по снижению уровня эффективности реализации атак отправления данных.

Список использованных источников:

1. Xue M. et al. Machine learning security: Threats, countermeasures, and evaluations //IEEE Access. – 2020. – Т. 8. – С. 74720-74742.
2. Islam G., Storer T. A case study of agile software development for safety-Critical systems projects //Reliability Engineering & System Safety. – 2020. – Т. 200. – С. 106954.
3. Sethi T. S., Kantardzic M. On the reliable detection of concept drift from streaming unlabeled data //Expert Systems with Applications. – 2017. – Т. 82. – С. 77-99.
4. Barrak A., Eghan E. E., Adams B. On the co-evolution of ml pipelines and source code-

empirical study of dvc projects //2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER). – IEEE, 2021. – С. 422-433.

5. Liu H. et al. Adaptiveface: Adaptive margin and sampling for face recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2019. – С. 11947-11956.

Вавилова А.С. (автор) Подпись

Волошина Н.В. (научный руководитель) Подпись