УДК 004.03

Исследование атак на Convolutional Neural Network, выполняющую задачу детекции объектов, на основе уязвимостей метода Feature Pyramid Network.

Ерыпалов К.И (Национальный Исследовательский Университет ИТМО)
Воробьева Алиса Андреевна, кандидат технических наук, факультет информационных технологий, доцент
(Национальный Исследовательский Университет ИТМО)

Введение.

Одним из методов, позволяющих улучшить точность работы сверточных сетей, является построение пирамиды сети признаков (FPN) [1]. Однако, в связи с тем, что Искусственный интеллект проник в огромное количество сфер нашей жизни, увеличилось и количество атак на данные системы с целью нарушения целостности, конфиденциальности или доступности. В частности, они были направлены на системы, реализованные на основе сверточных нейронных сетей (CNN), которые используются для детекции объектов. Поэтому специалистам по информационной безопасности важно знать природу атак и методов противодействия им.

Основная часть.

Исследование заключается в поиске оптимальных методов защиты от атак на модель детекции объектов с использованием метода пирамидальной сети признаков. В работе была рассмотрена архитектура построения FPN, проанализированы аналоги в работе с изображениями в сверточных сетях (стандартное решение в CNN - single feature map, пирамидальная иерархия объектов, структурирования пирамида изображений). На основе этого анализа были выявлены преимущества пирамидальной сети признаков и доказана необходимость дальнейшего исследования уязвимостей данного метода.

На модели детекции объектов были проведены adversarial [2], L-BFGS, FGSM [3], Jacobian saliency map, One pixel атаки [4], а также атаки с помощью генеративно-состязательной сети с целью исследования устойчивости модели сверточной сети с использованием метода пирамидальной сети признаков. Также было проверено, имеется ли возможность управлять данными атаками (можем ли мы заставить модель выдать принадлежность нужному нам неверному классу). На основе полученных данных (подсчет ошибок первого и второго рода, метрик precision, recall и F1-score)

Выводы.

Проведены теоретические и экспериментальные исследования возможных атак на системы детекции объектов с использованием методов пирамидальной сети признаков. Проанализированы полученные результаты и сформирован вектор дальнейшего исследования по противодействию существующим атакам.

Список использованных источников:

1. Lin T. Y. et al. Feature pyramid networks for object detection //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2017. – С. 2117-2125.
2. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.
3. Milton M. A. A. Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2-FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system //arXiv preprint arXiv:1806.08970. – 2018.
4. Su J., Vargas D. V., Sakurai K. One pixel attack for fooling deep neural networks //IEEE Transactions on Evolutionary Computation. – 2019. – Т. 23. – №. 5. – С. 828-841.