УДК 004.03

Исследование методов противодействия атакам, на системы детекции объектов, основанных на сверточных нейронных сетях с использованием метода FPN

Ерыпалов К.И (Университет ИТМО**)**
Роговой Виталий, ассистент, факультет безопасности информационных технологий (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 атаки [2], с целью исследования устойчивости модели сверточной сети с использованием метода пирамидальной сети признаков, после чего были применены следующие методы противодействия:

1. Медианное сглаживание [3], нелинейный фильтр, представляющий собой технологию нелинейной обработки сигналов, которая может эффективно подавлять шумы на основе статистической теории сортировки.

2. Использование пространственного контекста [4], при котором необходимо взять два случайных изображения из исходного датасета и целевой объект одного из них вставить в то же место на втором изображении.

3. Adversarial Pixel Masking, при данном противодействии используются методы преобразования изображений, такие как преобразование Хаара, получая новое изображение, на котором обнаруживается патч и замещается черными пикселями после чего полученное изображение складывается с начальным атакованным изображением.

4. Сверточные слои Габора, В этом методе изображения сначала разлагаются на собственные каналы RGB. Затем они попадают в банк фильтров Габора. Благодаря

своей высокой способности извлекать низкоуровневые функции изображения фильтры Габора могут повысить надежность сети на этом этапе.

.

На основе полученных данных (подсчет ошибок первого и второго рода, метрик precision, recall и AP) проведен анализ и были выявлены наиболее результативные методы противодействия системы детекции объектов, основанных на сверточных нейронных сетях .

Выводы.

Проведены теоретические и экспериментальные исследования возможных методов противодействия на системы детекции объектов с использованием методов пирамидальной сети признаков. Проанализированы полученные результаты (количество ошибок первого и второго рода, метрики качества), с помощью которых были выявлены наиболее эффективные методы противодействия с точки зрения сохранения точности работы модели.

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. Cohen Jeremy, Rosenfeld Elan, Kolter Zico. Certified adversarial robustness via randomized smoothing // international conference on machine learning / PMLR. — 2019. — P. 1310–1320.
4. Xiang Chong, Mittal Prateek. Detectorguard: Provably securing object detectors against localized patch hiding attacks // Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. — 2021. — P. 3177–3196.