УДК 004.896

Подход к верификации устойчивости модели компьютерного зрения к атакам состязательными патчами

Черняков А.А. (ВКА)

Научный руководитель – доктор технических наук, старший преподаватель Менисов А.Б.

(ВКА)

Введение. Современный технологический прогресс в области искусственного интеллекта и компьютерного зрения открывает новые возможности в здравоохранении, безопасности, автоматизации транспорта и других областях. Однако уязвимости таких систем, вызванные состязательными атаками, представляют серьезные риски. Исследование и разработка механизмов защиты от таких атак необходимы для обеспечения безопасности и надежности применения данных технологий [1, 2]. В рамках исследования разработан метод верификации устойчивости моделей компьютерного зрения к атакам с использованием состязательных патчей.

Основная часть.

Основной подход заключается в генерации патчей, которые изменяют прогнозы моделей, и последующем анализе их влияния на метрики производительности. Работа включает четыре этапа:

1. Генерация состязательных патчей с помощью оптимизации по градиенту.
2. Подготовка моделей и данных (использовались предобученные модели, такие как ResNet101 [3], и набор данных ImageNet).
3. Анализ изменения метрик, включая precision и recall, при различных параметрах патчей.
4. Расчёт целевой функции, которая нормирует влияние патчей по метрикам и параметрам.

Эксперименты проводились на моделях ResNet101, Inception v3 и GoogLeNet. Были варьированы размер и эпохи генерации патчей. Результаты показали, что наибольшее влияние патчи оказывают при площади покрытия от 20% до 80%. ResNet101 продемонстрировала наибольшую устойчивость, а GoogLeNet — наименьшую.

Выводы. В результате исследования разработана функция оценки устойчивости моделей, которая возвращает значение от 0 до 1 в зависимости от воздействия патча. Результаты применимы в анализе защищенности систем компьютерного зрения и проектировании методов защиты, а перспективы развития лежат в оптимизации вычислений и расширении области применения подхода.

Список использованных источников:

1. Whig P. et al., Sustainable Development through Machine Learning, AI and IoT. Springer, 2023.
2. Xu K. et al., Adversarial t-shirt! ECCV, 2020.
3. He K. et al. Deep residual learning for image recognition //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – С. 770-778..