УДК 004.056

РАЗРАБОТКА АЛГОРИТМА ЗАЩИТЫ СИСТЕМ РАСПОЗНАВАНИЯ ОБЪЕКТОВ НА ИЗОБРАЖЕНИЯХ НА ОСНОВЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

Якимов Я. Д. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н., ассистент Коржук В.М.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе представлен анализ существующих соревновательных атак и защит на системы распознавания объектов с использованием машинного обучения. Доклад также включает в себя разработанный алгоритм по защите от соревновательных атак и подготовленную реализацию алгоритма на системе распознавания дорожных знаков.

Введение. Благодаря цифровизации, созданные системы распознавания объектов на основе машинного обучения используются в различных сферах повседневной жизни. Их внедряют в системы городского видеонаблюдения, в системы контроля и управления доступом (СКУД), а также в беспилотные транспортные средства (БТС), с целью проведения дополнительных мер контроля и улучшения показателей безопасности. Перечисленные системы активно улучшаются, и ошибка из-за человеческого фактора влияет на правильное функционирование систем. Поэтому в дополнение или даже на смену приходят системы компьютерного зрения. Они показывают высокие показатели обнаружения, распознавания и классификации объектов на множестве различных данных.

Однако недавние работы в исследовании проблем использования нейронных сетей и машинного обучения в целом демонстрируют, что системы распознавания с использованием нейронных сетей уязвимы к специально сгенерированным искажениям на изображениях или видеоматериалах – соревновательным атакам.

Большинство алгоритмов защиты представляют собой решения для конкретных видов нейронных сетей или же направлены против конкретных соревновательных атак. Например, алгоритм «adversarial training» или же соревновательного обучения – обучение нейронных сетей с использованием заранее сгенерированных по конкретной методике вредоносных образцов. Еще одним алгоритмом защиты является защитная дистилляция – это метод обучения классификации с учителем, когда при обучении, одна модель нейронной сети обучается предсказывать выходные вероятности другой модели, обученной на базовом стандарте, что позволяет внести элемент случайности в систему распознавания и улучшить показатель точности выданного решения.

Соответственно, целью работы будет разработка алгоритма, объединяющего методы детектирования соревновательных атак, методы очистки входных изображений, а также саму модель нейронной сети, классифицирующую объекты на изображениях.

Основная часть. Одним из критериев разработки алгоритма является проектирование без зависимости от конкретных целей или моделей нейронной сети.

Модуль детектора данного алгоритма позволит избежать переобучения моделей нейронных сетей из-за использования соревновательного обучения и допускать к модели распознавания только неискаженные изображения. Такой подход уменьшает влияние вредоносных входных данных на способность модели распознавать обычные данные. Если модуль детектора определит, что входные изображения поступают от злоумышленника, то они будут переданы в модуль очистки для лучшего распознавания центральной моделью.

Очистка входных изображений необходима для полноценной работы алгоритма, так как простого детектирования недостаточно для точного определения объекта на изображении и противодействия атакам. Данный модуль позволяет избавиться от зашумлений или же так

называемых «заплаток», которые могут применяться в реальном мире в виде наклеек или носимых объектов. Зашумления и «заплатки» могут привести к неожиданному и потенциально опасному поведению систем, которыми управляет компьютерное зрение. Модуль очистки обучается отдельно от центральной модели и никак не влияет на её работу. Далее, центральная модель нейронной сети обрабатывает входные данные и выдает результат классификации, который может использоваться в соответствии с целью всей системы, где используется данный ряд модулей.

Выводы.

Представленный в докладе алгоритм тестировался на датасете «German Traffic Sign Recognition Benchmark (GTSRB)», содержащем изображения дорожных знаков. Также для тестирования были подготовлены изображения с вредоносными искажениями типа «заплаток» на основе того же датасета. На чистых изображениях подготовленная сверточная нейронная сеть показывает результат более 98% правильно определенных знаков на тестовом наборе изображений. При подаче на вход данных, содержащих «заплатки», которые покрывают около 25% исходного изображения - точность распознавания падает ниже 60%.

Реализация представленного алгоритма включает в себя модуль детектора и модуль очистки изображений. Реализация показала улучшение результатов до более чем 90% точности распознавания. Для чистоты эксперимента тестирование представленного алгоритма проводилось на том же наборе данных.

Таким образом, в перспективе данный алгоритм может применяться в системах распознавания дорожных знаков или аналогичных, где решается задача классификации. При использовании в программно-аппаратных комплексах беспилотных транспортных средств (БТС) и системах автопилотирования автомобилей данный алгоритм может помочь снизить количество дорожно-транспортных происшествий с участием БТС и снизить вероятные жертвы среди участников дорожного движения.

Якимов Я.Д. (автор) _____

Коржук В.М. (научный руководитель) _____