УДК 004.054

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ СОСТЯЗАТЕЛЬНОГО ОБУЧЕНИЯ В КАЧЕСТВЕ МЕХАНИЗМА ПРОТИВОДЕЙСТВИЯ АТАКАМ НА СЛОЖНЫЕ СИСТЕМЫ

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

Проблема создания методов обеспечения безопасности глубинных нейронных сетей от атак на основе состязательных примеров является значимой во всех областях применения технологии машинного обучения. Одним из перспективных методов противодействия такому виду атак является обучение нейронной сети состязательными примерами, сгенерированными на основе алгоритма атаки. В представленной работе проведен анализ эффективности применения состязательного обучения в качестве механизма противодействия атакам на сложные системы при реализации сценария проведения атаки методом «черного ящика».

Введение. Одной из главных проблем применения глубинных нейронных сетей является уязвимость таких структур к атакам на основе состязательных данных. В задачах классификации в качестве состязательного примера выступает искаженное изображение, при попадании в систему нейронная сеть классифицирует такой кадр некорректно и выдаст неверный прогноз. Состязательно обучение – один из наиболее эффективных механизмов противодействия атакам на основе состязательных примеров. В процессе состязательного обучения в обучающие данные для нейронной сети добавляются образцы, созданные на основе известного алгоритма состязательной атаки. В условиях, когда атакующий использует аналогичный метод, обученная нейронная сеть будет устойчива к действиям злоумышленника. Существует множество исследований о методах противодействия различным состязательным атакам, однако работ об обучении нейронных сетей противодействию нескольким типам атак с оценкой эффективности применения состязательного обучения не найдено.

Основная часть. Основа для проведения анализа эффективности применения состязательного обучения в качестве механизма противодействия атакам на сложные системы заключается в обучении модели нейронной сети на основе состязательных примеров, полученных на основе алгоритмов известных состязательных атак, с целью повышения устойчивости к различным типам таких атак. Значимым результатом исследования является установление связи между обучением модели с использованием одного типа состязательных примеров и соответствующим понижением уровня стойкости к атакам на основе другого типа состязательных данных.

Нейронные сети в эксперименте обучаются с расширением объема данных (вращением, трансляцией) для отслеживания влияния увеличения объема данных на стойкость к состязательным атакам.

В рамках работы ключевым является предположение о невозможности проведения злоумышленником реконструкции рассматриваемой модели нейронной сети путем многократного тестирования с собственным набором данных. Одним из заключительных этапов работы является проведение различных атак методом «черного ящика» для исследования корректности классификации искаженных изображений.

Выводы. В условиях ограниченного набора обучающих данных глубинные нейронный сети не могут корректно классифицировать данные, незначительно модифицированные или измененные. Злоумышленники использует такую ограниченность, создавая образцы, которые кажутся однозначно распознаются человеческим сознанием, но неправильно

классифицируются нейронной сетью. Обучение состязательности – эффективный метод противодействия состязательным атакам при условии использования всеми участниками процесса аналогичных методов для создания состязательных данных. В ходе исследования экспериментально доказано, что состязательное обучение нейронной сети на основе одного метода атаки не приводит к устойчивости модели к другим методам атаки, при этом нейронная сеть, обученная с помощью нескольких типов состязательных данных, остается уязвимой для атак с сценарием «белого ящика».