УДК 004.032.36

ИССЛЕДОВАНИЕ УЯЗВИМОСТЕЙ НЕЙРОННЫХ СЕТЕЙ, ОСНОВАННЫХ НА ИСКАЖЕНИИ ПАРАМЕТРОВ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

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

Противодействие атакам злоумышленников на системы на основе машинного обучения является важным направления в обеспечении безопасности систем, построенных на основе технологии искусственного интеллекта. Исследование параметров нейронных сетей с целью выявления слабых характеристик модели является основным способом обнаружения потенциальных уязвимостей нейронных сетей, связанных с искажением. В представленной работе проведен анализ работы моделей нейронных сетей в рамках установленных сценариев с целью выявления критических недостатков в системе на основе машинного обучения, при использовании которых возникает угроза некорректной работы модели нейронной сети, а также предложен метод повышения уровня безопасности на основе состязательного обучения, устойчивого к искажениям.

Введение. Многочисленные исследования глубинных нейронных сетей выявили отсутствие стойкости таких сетей к состязательным атакам, реализация которых приводит к некорректным результатам работы системы на основе машинного обучения. Высокая производительность и перспективы повсеместного применения технологии глубокого обучения исключают возможности отказа от применения глубинных нейронных сетей и являются основанием для поиска методов противодействия подобным атакам. Вектора состязательных атак не ограничиваются набором различных входных данных, источником уязвимостей нейронных сетей равнозначно становятся и параметры модели нейронной сети. Атаки, ориентированные на искажение параметров, в зависимости от способа реализации нейронной сети (программного или аппаратного) могут использовать различные механизмы воздействия на модель, например, отправление обучающих данных, квантование, сжатие, добавление фонового шума.

Основная часть. В рамках исследования уязвимостей нейронных сетей, ориентированных на искажение, используется метод оценки надежности параметров модели, основанный на измерении объема потерь, связанных с возникновением искусственного искажения характеристик нейронной сети в рамках заданного сценария. Традиционный анализ функции потерь приводит к заданию параметров модели сети на основе минимального объема потерь, что приводит к выбору более уязвимых характеристик нейронной сети к преднамеренным искажениям.

В рамках работы основным тезисом исследования является задание параметров модели нетрадиционным способом, ориентированным на увеличении функции потерь, гарантирующее рост объема искажений, требуемых для достижения некорректной работы системы на основе машинного обучения.

Значимым результатом исследования является экспериментальное применение алгоритмов состязательного обучения для улучшения стойкости нейронной сети к искажению параметров модели.

Выводы. Результатом проведенного исследования уязвимостей нейронных сетей, основанных на искажении параметров модели машинного обучения, является теоретическое и экспериментальное обоснование отличия стойкости нейронной сети к случайным изменениям и к искажениям на основе градиентного метода. В качестве решения обозначенной проблемы наличия уязвимостей нейронных сетей к атакам на основе искажения параметров модели предложено состязательное обучения на примерах, стойких к искажениям.