К ВОПРОСУ О ЗАЩИТЕ ИНФОРМАЦИИ В ИНТЕЛЛЕКТУАЛИЗИРОВАННЫХ ОБРАЗЦАХ ВООРУЖЕНИЯ

Грибунин В.Г.[1], Кондаков С.Е.[2]

Цель статьи: *анализ интеллектуализированных образцов вооружения, использующих машинное обучение, с точки зрения защиты информации. Разработка предложений по развертыванию работ в области обеспечения безопасности информации в подобных изделиях. Определение рациональных первоочередных направлений совершенствования данных изделий в части обеспечения безопасности информации.*

Метод исследования: *системный анализ систем машинного обучения как объектов защиты.*

Полученный результат: *представлены новые угрозы безопасности информации, возникающие при использовании образцов вооружения и военной техники с элементами искусственного интеллекта. Системы машинного обучения рассмотрены авторами как объект защиты, что позволило определить защищаемые активы таких систем, их уязвимости, угрозы и возможные атаки на них. В статье проанализированы меры нейтрализации идентифицированных угроз на основе таксономии, предложенной Национальным институтом стандартов и технологий США. Определена недостаточность существующей нормативной методической базы в области защиты информации для обеспечения безопасности систем машинного обучения. Предложен подход, который должен быть использован при разработке и оценке безопасности систем, использующих машинное обучение. Представлены предложения по развертыванию работ в области обеспечения безопасности интеллектуальных образцов вооружения, использующих технологии машинного обучения.*

Ключевые слова: *искусственный интеллект, системы машинного обучения, защита информации, уязвимости, атаки.*

DOI:10.21681/2311-3456-2021-5-5-11

Актуальность защиты информации в интеллектуализированных образцах вооружения, использующих машинное обучение

Искусственный интеллект (далее — ИИ) проникает во многие сферы нашей жизни. По словам Министра обороны России С. Шойгу, в настоящее время в стране ведется большая работа по созданию роботизированных образцов вооружения с ИИ. Назовем такие образцы вооружения интеллектуализированными. Примеры интеллектуализированных образцов вооружения и военной техники (далее — ИОВ) приведены, например, в отчете [1].

Важной составной частью ИОВ являются системы машинного обучения (далее — СМО). СМО реализуются в программном или программно-аппаратном исполнении, являются по классификации ГОСТ 15408 изделиями информационных технологий, и для них характерны все уязвимости, угрозы и атаки, как и для обычных таких изделий.

Вместе с тем СМО имеют характерные только для них уязвимости, что обуславливает наличие многих специфичных угроз безопасности, которые могут быть реализованы посредством специфичных атак.

Традиционно различают два направления в обеспечении безопасности: функциональная безопасность (safety) и информационная безопасность (security). О функциональной безопасности говорят в основном, в отношении таких объектов, которые создают риски для людей и окружающей среды. Она связана с такими характеристиками как резервирование, отказоустойчивость, устойчивость к внешним воздействиям. В понятие информационной безопасности входит обеспечение таких свойств защищенности информации, как конфиденциальность, целостность и доступность.

Конечно, информационная и функциональная безопасности связаны между собой. Например, нарушение целостности информации в компьютерной системе может привести к функциональному сбою в устройстве, которым она управляет. Более того, методы достижения функциональной и информационной безопасности при разработке систем во многом идентичны, особенно это справедливо в отношении разработки программного обеспечения.

Требования к обеспечению функциональной и информационной безопасности изложены в различных группах нормативных документов. Для функциональной безопасности основополагающим является переведен-

1 Грибунин Вадим Геннадьевич, доктор технических наук, главный научный сотрудник МОУ «ИИФ», г. Москва, Россия. E-mail: wavelet2@mail.ru

2 Кондаков Сергей Евгеньевич, кандидат технических наук, сотрудник Восьмого управления ГШ ВС РФ, г. Москва, Россия. E-mail: sergeikondakov@list.ru

ный на русский язык и принятый в России стандарт ГОСТ Р МЭК 61508. Кроме того, существуют стандарты обеспечения функциональной безопасности для атомных электростанций (например, ГОСТ Р МЭК 60880), авиационной отрасли (например, КТ-178С), встроенных систем (ГОСТ Р 51904). В области информационной безопасности основополагающим, пожалуй, является ГОСТ Р ИСО/МЭК 15408. Кроме того, имеется стандарт по разработке безопасного программного обеспечения ГОСТ Р 56939, другие стандарты, а также многочисленные руководящие документы ФСТЭК России, а для разработки криптографических средств — ФСБ России.

Для ИОВ должны выполняться как требования функциональной, так и требования информационной безопасности. Вместе с тем СМО имеют определенные особенности, которые не позволяют напрямую использовать вышеприведенные документы в области безопасности. К этим особенностям СМО можно отнести то, что алгоритм работы ПО зависит не только от задаваемых разработчиком спецификаций (называемых в данном случае гиперпараметрами), но и от поступающих на вход СМО данных (подход «data-driven»). В ходе обучения на поступающих данных параметры СМО изменяются неподконтрольным для разработчика образом. Обученная СМО представляет собой «черный ящик» для разработчика, что делает в ряде случаев ее поведение непредсказуемым.

Поэтому на задаче обеспечения безопасности СМО сфокусировано внимание многих исследователей. По проблемам безопасного машинного обучения проводятся исследования, результаты которых публикуются на конференциях, описаны в диссертациях и книгах. Управлением перспективных исследовательских проектов министерства обороны США (DARPA) ведется проект по исследованиям в данной области [2], Национальным институтом стандартов и технологий США (NIST) разработан проект таксономии и терминологии [3], американской некоммерческой компанией MITRE Corporation, являющейся центром технологической и национальной безопасности США, предложена матрица атак на данные системы [4].

В России в рамках деятельности ТК 164 принят ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения». Этот стандарт посвящен только функциональной безопасности. Вместе с тем актуальными являются разработка подходов к определению объекта защиты информации, построение модели угроз и нарушителя, определения механизмов защиты для СМО с учетом особенностей ИОВ.

О системах машинного обучения. Защищаемые активы

Для анализа объекта защиты (защищаемых активов) необходимо описать его основные свойства: состав и его функциональные особенности.

Ядром СМО является обучаемая на входных данных модель; как правило, это классификатор. Процесс обучения обычно происходит заблаговременно, после чего модель готова к эксплуатации (тестированию). Существуют и системы с непрерывным обучением («онлайн-обучением»). В ходе обучения оператор может подавать на вход СМО не только данные, но и метки классов, которым они принадлежат («обучение с учителем»). В случае отсутствия меток классов СМО относятся к категории «обучение без учителя».

Модель включает в себя изменяемые в ходе обучения и предварительно выбираемые параметры, которые называются гиперпараметрами. В ходе обучения происходит настройка параметров модели к данным обучающей выборки (далее – ОВ). ОВ можно рассматривать как небольшую часть от всех возможных данных рассматриваемой предметной области, т. е. генеральной совокупности (далее – ГС). Помимо адаптации к ОВ, которое характеризует свойство выразительности модели, модель должна быть способной и к обработке других данных из ГС – быть способной к обобщению. Эти два требования отчасти противоречат друг другу: хорошо адаптированная к ОВ модель может оказаться неспособной к обобщению; это называется переобучением модели. И наоборот: простые, например, линейные модели, хорошо обобщают новые данные, но плохо отражают структуру ОВ (недообучение модели).

На этапе эксплуатации модель выдает метки классов для поступающих данных либо вероятность их принадлежности к классу. На вход модели поступают векторы признаков, которые получены тем или иным образом путем преобразования измерений объектов физического мира.

Обученная модель может быть использована при обучении другой модели (например, в случае недостаточности входных данных). Это называется переносом обучения.

Исходя из проведенного анализа защищаемыми активами СМО могут являться [5]:

- результаты измерений, из которых получены признаки для обучения;
- алгоритмы получения признаков из результатов измерений;
- алгоритмы обучения модели;
- значения гиперпараметров модели;
- значения параметров обученной модели;
- доверительные вероятности принимаемых решений (на выходе классификатора);
- сами принимаемые классификатором решения.

Главным защищаемым активом СМО является полученная в ходе обучения граница принятия решений (гиперплоскость в n-мерном пространстве). Большинство атак нарушителя направлено на то, чтобы как можно более точно оценить эту границу.

Уязвимости, угрозы и атаки на СМО

Уязвимость – это «недостаток (слабость) программного (программно-технического) средства или информационной системы в целом, который (которая) может быть использована для реализации угроз безопасности информации».

Перечислим слабости (уязвимости) СМО:

- ОВ принципиально не может отражать всех данных ГС;
- процесс обучения модели и дальнейшего принятия ею решений скрыт от разработчика и оператора системы;

УДК 004.491

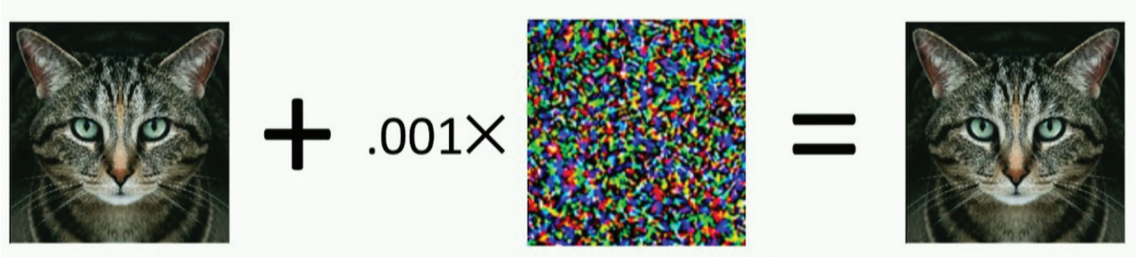

Рис.1. Добавление незначительного специально рассчитанного шума приводит к тому, что изображение справа классифицируется как собака

- параметры обученной модели отражают информацию о данных ОВ, что делает потенциально возможным их получение;
- сравнительно легко получить «теневую», или суррогатную модель за счет изучения защищаемой модели, как «черного ящика».

При моделировании можно выделить категории нарушителя в соответствии с его знанием о СМО («белый ящик» – «черный ящик», иногда выделяют промежуточный «серый ящик»), с его возможностями по совершению атак, по доступности для него обучающих данных, входов/выходов модели при эксплуатации и т.д.

В зависимости от своей категории нарушитель может совершать те или иные атаки. Все атаки могут быть разделены на два больших класса: атаки, выполняемые на этапе обучения и атаки этапа эксплуатации.

В отношении ИОВ первый класс атак означает, что нарушитель может вмешиваться в разработку изделия (если речь не идет об изделии с онлайн-обучением). Атаки этапа обучения называют еще атаками отравления, так как в них производится та или иная манипуляция с ОВ. Как правило, эта манипуляция заключается в незначительном изменении векторов ОВ, например, входных изображений, с целью получения контрпримера (adversarial example), на котором модель выдает некорректный результат.

На рис.1 представлен пример того, как добавление шума к ранее верно классифицированному изображению кошки приводит к неправильному решению – изображение справа классифицируется как собака. Любопытно, что существуют подходы к созданию «универсального» шума, добавление которого ко всем изображениям выборки приводит к некорректной классификации ее экземпляров [6].

Другой тип манипуляции заключается в добавлении к ОВ своих контрпримеров, также с целью некорректного обучения модели. Например, так можно создать «потайной люк». Контрпримеры в известных атаках нарушитель получает в результате решения той или иной задачи оптимизации.

В атаках этапа эксплуатации нарушитель изучает пары «вход/выход» модели, создавая свою теневую (суррогатную) модель. После того, как она будет создана, он сможет изучить границу принимаемых ею решений и подобрать контрпримеры. Подобные атаки называют еще атаками «обхода» или «уклонения» [7]. В отношении ИОВ подобную атаку нарушитель сможет провести, например, после захвата изделия на поле боя.

Создание суррогатной модели позволяет нарушителю проводить градиентную атаку даже без знания параметров атакуемой модели. Контрпример для суррогатной модели с большой долей вероятности будет контрпримером для атакуемой модели (это называется переносом контрпримера).

В настоящее время разработано много различных атак [7] на СМО. Например, в табл. 1 работы [8] приведено 33 алгоритма атаки только против СМО, используемых в области компьютерного зрения. Также активно ведется разработка механизмов защиты от атак нарушителя.

В работе [9] было показано, что практически все классификаторы, основанные на сетях глубокого обучения, не являются зависимыми к небольшим изменениям входных данных. Соответствующая атака, получившая название DeepFool, является эффективной против всех таких классификаторов.

Защита от атак на СМО

Таксономия (классификация) методов защиты от атак представлена в проекте документа [3] и приведена на рис. 2.

Несколько иная классификация приведена в [10]. Методы защиты от атак на СМО можно разделить на два больших класса: проактивные и реактивные.

В проактивных методах защиты исследователи стремятся достичь робастности (устойчивости) моделей СМО к контрпримерам. В реактивных методах исследователи пытаются обнаружить контрпримеры и предотвратить их попадание на вход модели (отфильтровать).

С точки зрения используемых подходов методы защиты могут быть разбиты на следующие классы [10]:

- маскирование (обфускация) градиента;
- дополнительные детекторы;
- статистические методы;
- методы препроцессинга;
- ансамбли классификаторов;
- измерение близости.

Дадим краткую характеристику каждого класса методов защиты от атак на СМО.

Маскирование градиента затрудняет нарушителю определение направление модификации вектора для получения контрпримера. Существуют различные спо-

DOI:10.21681/2311-3456-2021-5-5-11

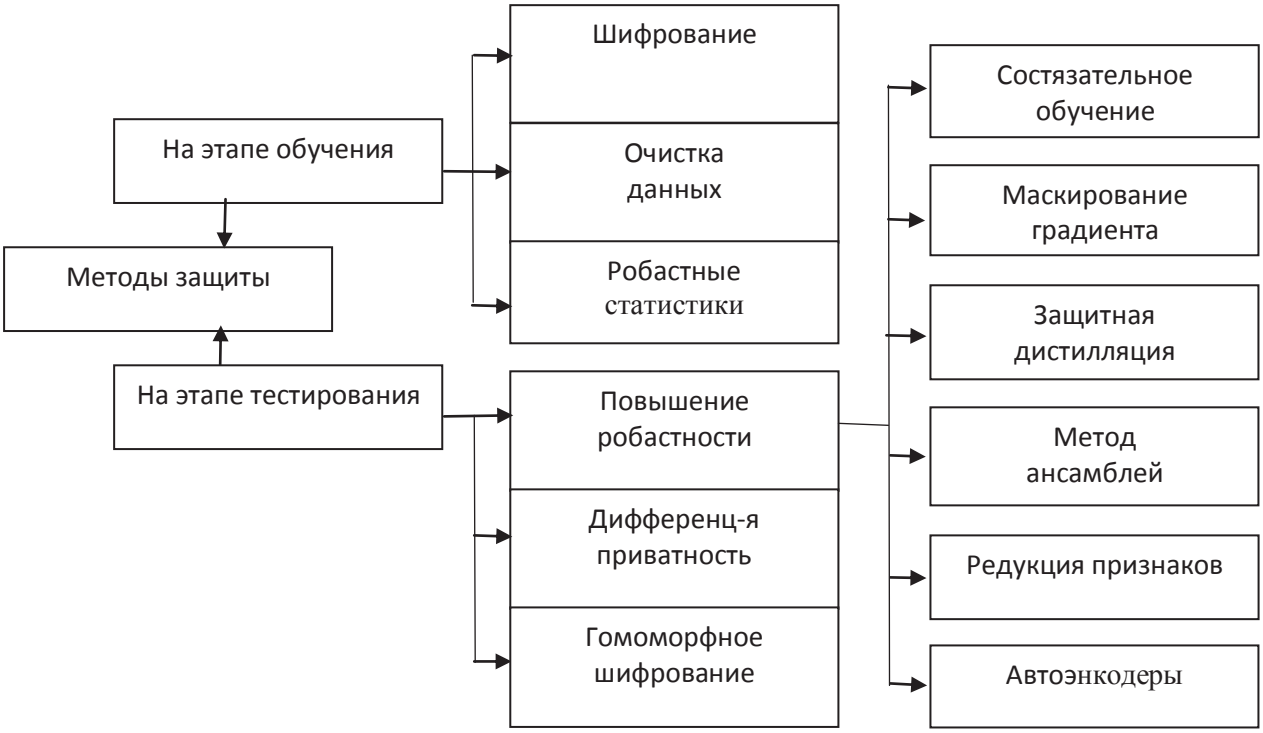

Рис.2. Таксономия методов защиты СМО от атак [3]

собы маскирования градиента [10]. Наиболее интересными представляются способ, основанный на состязательном обучении, и способ, использующий дистилляцию знаний.

В методе *состязательного обучения* [11] «защитник» обучает модель, затем добавляет к ОВ рассчитанные им самим контрпримеры для получившейся обученной модели и выполняет повторное обучение уже на расширенной ОВ. Недостатком является громоздкость метода и тот факт, что для эффективной защиты контрпримеры нужны для самых разных потенциально возможных атак.

Дистилляция знаний – это технология, разработанная для обучения относительно простых сетей на базе ОВ без ручной разметки, но с использованием предварительно обученной более мощной модели [12]. В защитной дистилляции используются две нейросети одной архитектуры. Первая нейросеть (сеть-учитель) обучается на размеченном наборе ОВ. В результате обучения получаются вероятности принадлежности к классам данных ОВ. Эти вероятности используются в качестве меток классов для обучения второй нейросети (сеть-ученик), на выходе которой также получаются вероятности принадлежности к классам, которые и используются далее.

Использование *дополнительных детекторов* – это метод реактивной защиты, суть которого заключается в следующем. Защитник использует состязательное обучение для получения модели, которая работает как фильтр для контрпримеров. Далее эти контрпримеры могут быть отброшены [13] либо выделены в особый класс для классификатора [14].

В *статистических методах* защита основана на поиске различия между статистическим распределением примеров и контрпримеров [15].

В *методах препроцессинга* вектора ОВ подвергаются предварительной обработке перед подачей на вход классификатора. Например, для изображений это может быть обрезание, масштабирование, поворот, очистка от шумов, снижение размерности и т.д. [16]. В основе методов лежит предположение о том, что после обработки контрпримеры «утратят свои свойства» и перестанут быть таковыми. В этих методах защита основана на том, что к векторам добавляется шум и, тем самым, маскируется градиент. Положительным качеством методов является то, что они универсальны по отношению к используемым атакам и моделям классификатора.

В основе методов *ансамблей классификаторов* лежит предположение о том, что контрпример, актуальный против одного классификатора, не будет работоспособным против других классификаторов [17].

В методах *измерения близости* изучаются решения, принимаемые скрытыми слоями глубокой нейросети. При этом предполагается, что для векторов одного класса расстояние от центроида класса будет невелико. Вводятся соответствующие метрики и определяются вероятные контрпримеры [18].

В работе [3] выделены также такие меры защиты, как шифрование данных, в том числе, гомоморфное шифрование, реализация дифференциальной приватности.

Стоит отметить, что существуют программные библиотеки с открытым исходным кодом, моделирующие атаки на СМО и механизмы защиты от них. К ним относятся, например, Foolbox [19], CleverHans [20].

В поддерживаемой крупнейшей компанией в мире производителей и поставщиков аппаратного и программного обеспечения IBM библиотеке ART (Adversarial Robustness Toolbox) [21] реализованы как атаки, так и меры защиты. В ней смоделированы свыше 40 различных атак, сгруппированных по категориям: «отравление», «уклонение», «экстракция» параметров модели и атаки «вывода», более 30 защитных механизмов, 3 метрики робастности, а также метрики сертификации и верификации.

Данная библиотека позволяет исследовать безопасность самых различных архитектур СМО: сетей глубокого обучения, градиентного бустинга, деревьев решений, машин опорных векторов, случайного леса, логистической регрессии и многих других. Она позволяет использовать практически все популярные программные фреймворки в области СМО, такие как Keras, PyTorch, TensofFlow, Scikitlearn и другие.

Популярность библиотеки ART может быть подчеркнута, например, тем фактом, что за последний год на GitHub было сделано более 500 её форков.

Приведённое краткое описание методов атак и методов защиты систем машинного обучения показывает, как многогранна данная предметная область. Исследования пока находятся в начальной стадии, тем не менее, векторы развития определены. При разработке интеллектуализированных образцов вооружения нужно не только учитывать вопросы безопасности их применения, но и защиты обрабатываемой в них информации, а также защиты используемых для обучения данных и всей остальной информации, влияющей на эффективность применения ИОВ. Все ИОВ должны проходить оценку соответствия, в том числе, и требованиям безопасности информации.

Выводы

Проведенный анализ рассмотренных выше уязвимостей, атак и механизмов защиты СМО позволяет сделать следующие выводы:

1. особенности СМО обуславливают существование по отношению к ним серьезных угроз безопасности, парирование которых является важной задачей;
2. существующие методические документы по формированию угроз безопасности, модели нарушителя не учитывают особенностей СМО;
3. требования по защите СМО от специфических угроз отсутствуют;
4. в рамках существующей системы оценки соответствия (сертификации) изделий информационных технологий соответствующие виды испытаний не предусмотрены.

Все вышесказанное позволяет предложить, что для эффективного и безопасного использования интеллектуализированных образцов вооружения в Минобороны России необходимы следующие первоочередные меры:

- создание типовых моделей угроз СМО для различных классов ИОВ;
- разработка показателей и критериев безопасности СМО и требований по безопасности информации;
- разработка типовых методик оценки соответствия ИОВ, использующих СМО, требованиям по безопасности информации;
- разработка инструментальных средств оценки безопасности СМО.

Литература

1. Artificial Intelligence and Autonomy in Russia. — CNA Report. – May, 2021. — Режим доступа: https://www.cna.org/CNA_files/centers/CNA/sppp/fsp/russia-ai/Russia-Artificial-Intelligence-Autonomy-Putin-Military.pdf
2. Guaranteeing AI Robustness Against Deception (GARD). Режим доступа: https://darpa.mil/program/guaranteeing-ai-robustness-against-deception.
3. A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269, October 2019. Режим доступа: https://doi.org/10.6028/NIST.IR.8269-draft
4. Adversarial ML Threat Matrix. December 2020. Режим доступа: https://github.com/mitre/advmlthreatmatrix.
5. Грибунин, В.Г. Безопасность систем машинного обучения. Защищаемые активы, уязвимости, модель нарушителя и угроз, таксономия атак. / В.Г. Грибунин, В.Г., Р.Л. Гришаненко, А.П. Лабазников, А.А. Тимонов // Известия института инженерной физики. — Серпухов. – 2021 г. — №3. — С.65-71.
6. Vadillo J., Santana R. Universal adversarial examples in speech command classification. — Preprint. – 2021. — Режим доступа: https://arxiv.org/pdf/1911.10182.pdf
7. Chakraborty A. A survey on adversarial attacks and defences. / A.Chakraborty, M.Alam, V.Dey, A.Chattopadhyay, D.Mukhopadhyay // CAAI Transactions on Intelligence Technology. – 2021. — v.6. — P.25-45.
8. Machado G.R. Adversarial Machine Learning in Image Classification: A Survey Towards the Defender's Perspective. / G. R. Machado, E. Silva, R. R. Goldschmidt // Preprint — September, 2020. Режим доступа: https://arxiv.org/pdf/2009.0372v1.
9. Moosavi-Dezfooli S. DeepFool: a simple and accurate method to fool deep neural networks / S.Moosavi-Dezfooli, A.Fawzi, P.Frossard. // Preprint — July, 2016. Режим доступа: https://arxiv.org/pdf/1511.04599v3.
10. Machado G. Adversarial Machine Learning in Image Classification: A Survey Towards the Defender's Perspective. / G. Machado, E. Silva, R. Goldschmidt. Preprint — September, 2020. — Режим доступа: https://arxiv.org/pdf/2009.0372v1

11. Kannan H., Kurakin A., Goodfellow I. 2018. Adversarial Logit Pairing. // Preprint. – 2018. – Режим доступа: https://arxiv.org/pdf/1803.06373.
12. Бирюкова В.А. Технология дистилляции знаний для обучения нейронных сетей на примере задачи бинарной классификации. // Интеллектуальные системы. Теория и приложения. № 24(2) – 2020. С.23-52.
13. Chen J., Meng Z., C. Sun, W. Tang, Y. Zhu. ReabsNet: Detecting and Revising Adversarial Examples. // Preprint. — 2017. — Режим доступа: https://arxiv.org/pdf/1712.08250.
14. Z. Gong, W. Wang, W. Ku. Adversarial and clean data are not twins. // Preprint. — 2017. — Режим доступа: https://arxiv.org/pdf/1704.04960.
15. R. Feinman, R. Curtin, S. Shintre, A. Gardner. Detecting Adversarial Samples from Artifacts. Preprint. — 2017. — Режим доступа: https://arxiv.org/abs/1703.00410.
16. C. Guo, M. Rana, M. Cisse, L. Van Der Maaten. Countering adversarial images using input transformations. – Preprint. — 2017. — Режим доступа: https://arxiv.org/pdf/1711.00117.
17. F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, P. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. – Preprint. — 2017. — Режим доступа: https://arxiv.org/pdf/1705.07204.
18. X. Cao, N. Gong. Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification. // In Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017). ACM, NewYork, NY, USA, 278–287. Режим доступа: https://doi.org/10.1145/3134600.3134606.
19. https://www.foolbox.readtehedocs.io.
20. https://www.cleverhans.io.
21. https://www.ibm.com/blogs/research/tag/adversarial-robustness-toolbox-art.
22. Гарбук С.В. Задачи нормативно-технического регулирования интеллектуальных систем информационной безопасности // Вопросы кибербезопасности. 2021, № 3 (43)- 68-83 с. DOI: 10.21681/2311-3456-2021-3-68-83

TOWARD TO INFORMATION SECURITY OF AI-ENHANCED WEAPONS

Gribunin V.G.[3], Kondakov S.E.[4]

Abstract

Purpose of the article: *Analysis of intellectualized weapons using machine learning from the point of view of information security. Development of proposals for the deployment of work in the field of information security in similar products.*

Research method: *System analysis of machine learning systems as objects of protection. Determination on the basis of the analysis of rational priority directions for improving these systems in terms of ensuring information security.*

Obtained result: *New threats to information security arising from the use of weapons and military equipment with elements of artificial intelligence are presented. Machine learning systems are considered by the authors as an object of protection, which made it possible to determine the protected assets of such systems, their vulnerabilities, threats and possible attacks on them. The article analyzes the measures to neutralize the identified threats based on the taxonomy proposed by the US National Institute of Standards and Technology. The insufficiency of the existing regulatory methodological framework in the field of information protection to ensure the security of machine learning systems has been determined. An approach is proposed that should be used in the development and security assessment of systems using machine learning. Proposals for the deployment of work in the field of ensuring the security of intelligent weapons using machine learning technologies are presented.*

Keywords: *artificial intelligence, machine learning, information security, vulnerabilities, attacks.*

References

1. Artificial Intelligence and Autonomy in Russia. — CNA Report. – May, 2021. — Rezhim dostupa: https://www.cna.org/CNA_files/centers/CNA/sppp /fsp/russia-ai/Russia-Artificial-Intelligence-Autonomy-Putin-Military.pdf
2. Guaranteeing AI Robustness Against Deception (GARD). Rezhim dostupa: https://darpa.mil/program/guaranteeing-ai-robustness-against-deception.

3 Vadim Gribunin, Dc.Sc., Chief Researcher of the IIF, Moscow, Russia. E-mail: wavelet2@mail. ru

4 Sergey Kondakov, Ph.D., Russian Defense Ministry employee, Moscow, Russia. E-mail: sergeikondakov@list.ru

3. A Taxonomy and Terminology of Adversarial Machine Learning. Draft NISTIR 8269, October 2019. Rezhim dostupa: https://doi.org/10.6028/NIST.IR.8269-draft
4. Adversarial ML Threat Matrix. December 2020. Rezhim dostupa: https://github.com/mitre/advmlthreatmatrix.
5. Gribunin V.G. Bezopasnost` sistem mashinnogo obucheniia. Zashchishchaemy`e aktivy`, uiazvimosti, model` narushitelia i ugroz, taksonomiia atak. / V.G. Gribunin, V.G., R.L. Grishanenko, A.P. Labaznikov, A.A. Timonov // Izvestiia instituta inzhenernoi` fiziki. — 2021 g., № 3, s.65-71.
6. Vadillo J., Santana R. Universal adversarial examples in speech command classification. — Preprint. – 2021. — Rezhim dostupa: https://arxiv.org/pdf/1911.10182.pdf
7. Chakraborty A. A survey on adversarial attacks and defences. / A.Chakraborty, M.Alam, V.Dey, A.Chattopadhyay, D.Mukhopadhyay // CAAI Transactions on Intelligence Technology. – 2021. — v.6. — P.25-45.
8. Machado G.R. Adversarial Machine Learning in Image Classification: A Survey Towards the Defender's Perspective. / G. R. Machado, E. Silva, R. R. Goldschmidt // Preprint — September, 2020. Rezhim dostupa: https://arxiv.org/pdf/2009.0372v1.
9. Moosavi-Dezfooli S. DeepFool: a simple and accurate method to fool deep neural networks / S.Moosavi-Dezfooli, A.Fawzi, P.Frossard. // Preprint — July, 2016. Rezhim dostupa: https://arxiv.org/pdf/1511.04599v3.
10. Machado G. Adversarial Machine Learning in Image Classification: A Survey Towards the Defender's Perspective. / G. Machado, E. Silva, R. Goldschmidt. Preprint — September, 2020. — Rezhim dostupa: https://arxiv.org/pdf/2009.0372v1
11. Kannan H., Kurakin A., Goodfellow I. 2018. Adversarial Logit Pairing. // Preprint. – 2018. – Rezhim dostupa: https://arxiv.org/pdf/1803.06373.
12. Biriukova V.A. Tekhnologiia distilliatcii znanii` dlia obucheniia nei`ronny`kh setei` na primere zadachi binarnoi` classifikatcii. // Intellektual`ny`e sistemy`. Teoriia i prilozheniia. № 24(2) – 2020. S.23-52.
13. Chen J., Meng Z., C. Sun, W. Tang, Y. Zhu. ReabsNet: Detecting and Revising Adversarial Examples. // Preprint. — 2017. — Rezhim dostupa: https://arxiv.org/pdf/1712.08250.
14. Z. Gong, W. Wang, W. Ku. Adversarial and clean data are not twins. // Preprint. — 2017. — Rezhim dostupa: https://arxiv.org/pdf/1704.04960.
15. R. Feinman, R. Curtin, S. Shintre, A. Gardner. Detecting Adversarial Samples from Artifacts. Preprint. — 2017. — Rezhim dostupa: https://arxiv.org/abs/1703.00410.
16. C. Guo, M. Rana, M. Cisse, L. Van Der Maaten. Countering adversarial images using input transformations. – Preprint. — 2017. — Rezhim dostupa: https://arxiv.org/pdf/1711.00117.
17. F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, P. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. – Preprint. — 2017. — Rezhim dostupa: https://arxiv.org/pdf/1705.07204.
18. X. Cao, N. Gong. Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification. // In Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017). ACM, New-York, NY, USA, 278–287. Rezhim dostupa: https://doi.org/10.1145/3134600.3134606.
19. https://www.foolbox.readtehedocs.io.
20. https://www.cleverhans.io.
21. https://www.ibm.com/blogs/research/tag/adversarial-robustness-toolbox-art.
22. Garbuk S.V. Zadachi normativno-tekhnicheskogo regulirovaniia intellektual`ny`kh sistem informatcionnoi` bezopasnosti // Voprosy` kiberbezopasnosti. 2021, № 3 (43)- 68-83 s. DOI: 10.21681/2311-3456-2021-3-68-83.