Текущие академические и индустриальные проекты, посвященные устойчивому машинному обучению

Д.Е Намиот, Е.А. Ильюшин, И.В. Чижов

***Аннотация*— С ростом применения систем на базе машинного обучения, которые, на сегодняшний день, с практической точки зрения, рассматриваются как системы искусственного интеллекта, растет и внимание к вопросам надежности (устойчивости) такого рода систем и решений. Естественно, что для критических применений, например, систем, принимающих решения в реальном времени вопросы устойчивости являются самыми главными с точки зрения практического использования систем машинного обучения. Собственно говоря, именно оценка устойчивости определяет саму возможность использования машинного обучения в таких системах. Это, естественным образом, отражается в большом количестве работ, посвященных вопросам оценки устойчивости систем машинного обучения, архитектуре таких систем и защите систем машинного обучения от зловредных действий, которые могут повлиять на их работу. При этом необходимо понимать, что проблемы с устойчивостью могут возникать как естественным образом, в силу разного распределения данных на этапах обучения и практического применения (на этапе обучения модели мы используем только часть данных из генеральной совокупности), так и в результате целенаправленных действий (атак на системы машинного обучения). Атаки при этом могут быть направлены как на данные, так и на сами модели.**

***Ключевые слова*—robust machine learning, adversarial machine learning.**

I. Введение

Системы на основе машинного обучения приобрели большую популярность в последнее время. Реалии сегодняшнего времени таковы, что машинное обучение используется в любых случаях отсутствия аналитических моделей и алгоритмов для прямого вычисления. При этом машинное обучение (глубинное обучение) на сегодняшний день является практическим синонимом понятия искусственный интеллект. Естественно, что в таких условиях системы машинного обучения начали применяться и для критических операций. Это не обязательно связано именно с военным (специальным) применением. Системы управления, автономные транспортные средства, медицинские применения – есть уже масса примеров использования ML/DL (машинное обучение/глубинное обучение) систем в критических приложениях.

Проблемы, которые возникли с применением систем машинного обучения, связаны с надежностью (устойчивостью) работы таких систем. Несмотря на впечатляющую производительность алгоритмов DL, многие недавние исследования вызывают опасения по поводу безопасности и надежности моделей машинного обучения [1]. Каким образом можно, например, гарантировать работу некоторого классификатора, основанного на нейронной сети? Принципиальным моментом для систем машинного обучения является то, что система обучается на одних данных, а в практическом использовании будет работать с другими. Вообще говоря, соответствие тренировочных данных генеральной совокупности совсем не гарантировано. Реальные (тестовые) примеры могут отрабатываться совсем неверно. Если же какой-то стороной предпринимаются специальные действия (например, специальная подготовка данных) для неверной работы систем на основе машинного обучения, то говорят об атаках на системы машинного обучения.

Например, Szegedy и другие впервые продемонстрировали, что модели DL строго уязвимы для тщательно созданных состязательных (состязательный здесь и далее – опровергающий) примеров [2]. Точно так же различные типы атак (построения состязательных примеров) с отравлением (специальной модификацией) данных и моделей были предложены против систем DL [3], а в литературе были предложены различные способы защиты от таких стратегий [4]. Однако надежность методов защиты также сомнительна, и различные исследования показали, что большинство методов защиты неэффективны против конкретной атаки. Именно обнаружение того факта, что модели DL не являются ни безопасными, ни надежными, значительно препятствует их практическому развертыванию в критически существенных для безопасности приложениях, таких, например, как прогнозы в здравоохранении, что, естественно, жизненно важно.

Статья получена 9 сентября 2021. Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект»

Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
Е.А. Ильюшин - МГУ имени М.В. Ломоносова (email: john.ilyushin@gmail.com)
И.В. Чижов - МГУ имени М.В. Ломоносова (email: ichizhov@cs.msu.ru).

На рисунке 1 показано количество публикаций, посвященных состязательным примерам [5].

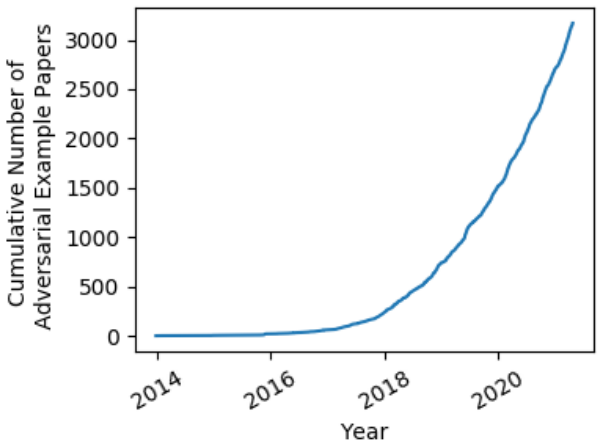

Рис.1. Публикации, посвященные состязательным примерам [5]

Как видно, резкий рост начался после 2018 года. Естественно, при таком довольно большом количестве публикаций есть уже и много работ, которые классифицируют имеющиеся проблемы с устойчивостью (надежностью), равно как и представляют существующие решения. Вместе с тем, текущее состояние этой проблемы таково, что не существует никакого общего решения, которое гарантировало бы работу произвольной системы машинного обучения при всех исходных данных. Гарантировало бы в том смысле, как это делается, например, для программного обеспечения в авионике и других подобных критических применениях. Подробнее об этом идет речь в разделе II.

Соответственно, самые главные работы по указанному направлению еще не написаны (или, по крайней мере, не опубликованы публично). Поэтому, на наш взгляд, было бы интересно рассмотреть существующие академические и промышленные проекты в области устойчивого машинного обучения, которые и призваны как-то решить имеющуюся проблему.

Эта статья написана в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по подготовке магистерской программы "Artificial Intelligence in Cybersecurity" [6]. Состояние дел в этой области меняется весьма стремительно. Описаны проекты по состоянию на сентябрь 2021.

Остальная часть статьи структурирована следующим образом. В разделе II кратко приводится обзор текущего состояния исследований в области устойчивых моделей машинного обучения. В разделе III рассматриваются промышленные и академические проекты в этой области. В разделе IV представлены индустриальные проекты и стартапы. И, наконец, раздел V представляет собой заключение.

II. Текущее состояние

Независимо от используемых моделей и методов получения независимых параметров (признаков), выбора анализируемых переменных и т.д., любые модели машинного обучения всегда пытаются распространить полученные при обучении результаты на всю генеральную совокупность данных. В общем случае, вообще говоря, нет (или может не быть) оснований для этого. Это и есть основная проблема. Даже полное понимание принципов работы системы на обучающих примерах не поможет, если окажется, что модель не работает на реальных данных. Соответственно, проблема устойчивости (надежности) заключается в проверке (подтверждении) того, что сконструированная система может работать с данными, которые отличаются от тех, на которых она была обучена.

Устойчивые (надежные) и безопасные системы машинного обучения - это системы, поведение которых в процессе эксплуатации не отличается от заявленного на этапе тестирования и обучения.

Формально, например, для системы классификации это можно определить следующим образом:

Некоторый классификатор C является δ-устойчивым в точке $\vec{X}$ только и если

$$\left\|\vec{X}-\vec{X}_0\right\|_\infty \le \delta \Rightarrow C(\vec{X}) = C(\vec{X}_0) \qquad (1)$$

Интуитивное определение, которое говорит, что если разница между исходными данными в пространстве признаков не превышает δ, то такие объекты должны классифицироваться схожим же образом.

Важно, что устойчивость относится только к реакции на изменение данных и ничего не определяет через характер этих изменений. Нарушение указанного условия может быть вызвано, например, тем, что тренировочный набор данных отличается от генеральной совокупности, может быть неправильный выбор признаков, ошибки в алгоритме, а также намеренное внесение искажений в реальные данные. В последнем случае говорят об атаках на системы машинного обучения. Естественно предположить, что системы, задействованные в критических применениях, могут чаще становится объектом атаки.

Собственно говоря, формула (1) и определяет текущее состояние решения проблемы устойчивости. Да, во множестве работ показывается, что модификации данных могут нарушать условие схожести классификатора. Но в полном соответствии с определением, ищутся некоторые минимальные модификации. Типичное описание атак для систем машинного обучения классифицирующих изображения – “незаметные человеческому глазу изменения позволяют обойти ограничения …”. Очевидно, что такое описание явно предполагает присутствие человека в контуре принятия решения. Но для автоматизированных систем, очевидно, размер искажений не играет никакой роли. Соответственно, атаки с неограниченным бюджетом на изменения могут всегда быть успешны.

Также в формуле (1) речь идет об изменениях исходных данных модели, но эти исходные данные не всегда (не во всех моделях) будут некоторыми прямыми характеристиками объектов (как, например, отдельные точки в изображениях). Во многих случаях параметры (features) моделей – это искусственные характеристики (например, свертки при работе со звуком). Изменения (возмущения) таких параметров может быть не просто связать с изменениями реальных характеристик. Сама идея того, что мы рассматриваем сеть как черный ящик, исключает какие-либо доказательства ее свойств, включая устойчивость. А требование объяснимости может вступать в противоречие с тем, что для повышения точности нам нужно все больше параметров. Итогом является отсутствие на сегодняшний день решений, которые гарантировали бы устойчивость для произвольной сети на любых данных.

Кратко, текущее состояние проблемы создания устойчивых моделей машинного обучения можно описать выражением “есть понимание”. Да, признается, что устойчивость (надежность) составляет на сегодня, пожалуй, основную проблему для применения систем на базе машинного обучения в критических областях. Это находит свое отражение, как в академических статьях, так и в академических и промышленных проектах, посвященных устойчивому машинному обучению.

Как и в любой другой научной области, все начинается с таксономии. Для естественных причин расхождения реальных данных с теми, на которых модель тренировалась, принято говорить о смещении распределения (distribution shift) [7]. В указанной работе [7] различают сдвиг обобщения предметной области и смещение субпопуляции При сдвиге обобщения предметной области обучающие и тестовые распределения содержат данные из связанных, но разных доменов. Например, записи о пациентах, но полученные из разных больниц. Изображения, снятые разными камерами и т.д. В смещении субпопуляции тренировочные и тестовые данные есть разные подмножества одного и того же распределения. Одним из наиболее часто встречающихся (и наиболее изученным) является так называемый ковариативный сдвиг [8]. Здесь мы предполагаем, что, хотя распределение входных данных может меняться со временем, функция маркировки, то есть условное распределение $P(y|x)$, не изменяется. То есть проблема возникает из-за сдвига в распределении признаков (covariates). Сдвиг метки – это обратная ситуация, вероятности для меток (заключений) P(y) меняются, а условные вероятности P(x|y) остаются постоянными. Например, медицинская диагностическая система, где вероятность встретить диагноз *y* уменьшается со временем, а классификатор для выработки диагноза $P(x|y)$ не изменяется. И, наконец, сдвиг концепции – это когда сами определения меток (заключений) меняются. Для того же примера диагностической системы – это смена критериев выработки диагнозов (заключений).

Искусственно созданные проблемы для систем машинного обучения принято называть атаками. Здесь также есть своя классификация (и далеко не одна).
Например, следующая таблица приводит пример классификации атак (и методов защиты) в зависимости от атакуемой компоненты системы машинного обучения

Таблица 1 Атаки на системы ML

Атака	**Место атаки**	**Затрагиваемые параметры**	**Методы противодействия**
Adversarial Attack	использование	Входные данные	Gradient Masking, Pre-Processing Filters, Adversarial Retraining
Backdoor Attack	тренировка	Параметры сети	Pruning, Fine Tuning
Data poisoning	тренировка, использование	Входные данные	Encryption, Local Training
IP stealing	использование	Отклик системы	Obfuscation, Encryption
Neural-level trojan	тренировка	Отклик системы	Data filtering
Side-channel Attack	использование	Отклик системы	Randomness

И это, естественно, только некоторые примеры нескольких сот существующих атак, приводимые исключительно в целях иллюстрации. Все развитие в этой области идет по следующей схеме: описание новой атаки (опровергающих примеров) для модели машинного обучения – построение защиты (чаще – ограничения для атакующих) – новая атака и т.д.

При этом проблемы, обозначенные в классификации атак, не стоит считать такими уж редкими. На самом деле, те же атаки с отравлением случаются гораздо чаще, чем об этом можно было бы предположить. Простой пример – социальные сети могут использовать (используют, на самом деле) дообучение своих рекомендательных систем на основе реального поведения пользователей. А “нужное” поведение может быть легко смоделировано. Другой пример, связанный с отравлением датасетов – это просто ошибки в разметке. Многие известные наборы данных, которые используются в разных системах, а также являются базой для предварительно обученных сетей, просто содержат множество ошибок. Пример – работы MIT [9].

Текущее состояние так называемого adversarial machine

learning может быть компактно проиллюстрировано следующим рисунком.

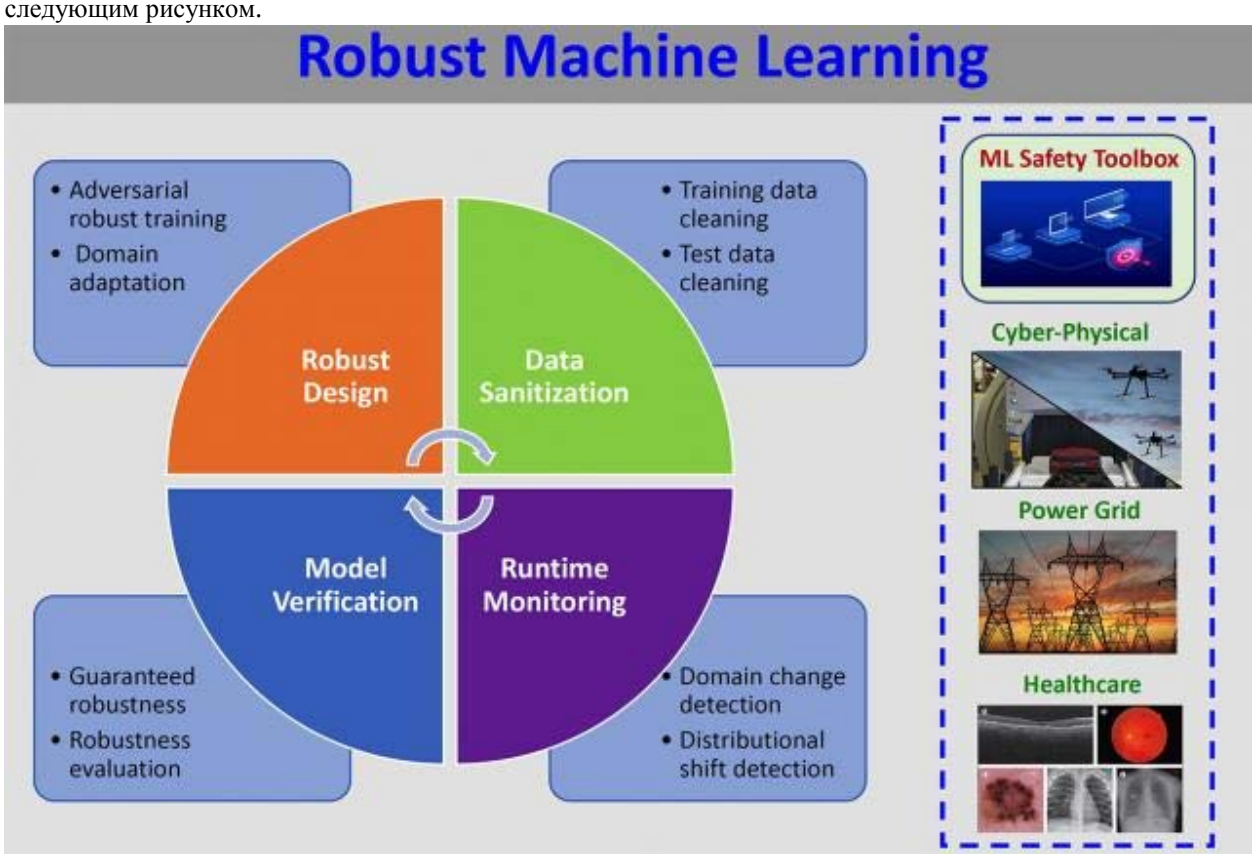

Рис. 2. Robust machine learning [10]

При этом реальным подтверждением устойчивости является только верификация моделей. Определение сдвига – это именно лишь только определение изменения характера данных. Очистка данных и фильтрация – это лишь потенциальная возможность ограничить вредоносные данные там, где это возможно. Например, при работе с голосовыми данными мы можем попробовать ограничить возможный частотный диапазон. Состязательные тренировки, кратко, представляют собой включение возможных атакующих данных в тренировочные наборы.

Верификация, естественно, привлекает к себе большое внимание, но на сегодняшний день нет примеров использования этого подхода для реальных (больших) нейронных сетей [11, 12].

Существующие обзоры по данному направлению, например, [13] описывают, естественно, текущее состояние проблемы и существующие (опубликованные) работы. А идея данной статьи – это оценка направления будущих работ.

III Академические и промышленные проекты, посвященные устойчивому машинному обучению

Государственный заказ на такого рода работы очевиден. Объединенный центр искусственного интеллекта, созданный Пентагоном для помощи военным США в использовании ИИ, недавно сформировал подразделение для сбора, проверки и распространения моделей машинного обучения с открытым исходным кодом и отраслевых моделей среди групп в Министерстве обороны. Часть этих усилий указывает на ключевую проблему с использованием ИИ в военных целях. «Red team» машинного обучения, известная как Группа тестирования и оценки, будет исследовать предварительно обученные модели на предмет слабых сторон. Другая группа по кибербезопасности исследует код и данные ИИ на предмет скрытых уязвимостей [14].

Обоснование по дословной цитате: “Мы не знаем, как создавать системы, полностью устойчивые к состязательным атакам”. Содержание работ, конечно, неизвестно, но, например, есть оценка состояния проблемы и план работ в отчете Центра безопасности и новых технологий Джорджтаунского университета, который участвует в этих работах. В частности, в отчете говорится, что «отравление данных» в системах Искусственного интеллекта (ИИ) может представлять серьезную угрозу национальной безопасности. Это будет включать проникновение в процесс, используемый для обучения модели ИИ, возможно, с помощью агента-добровольца для маркировки изображений, подаваемых в алгоритм, или путем размещения изображений в сети, которые собираются и передаются в модель ИИ [15].

Отчет касается организации работ по защите разделяемых (переиспользуемых) ресурсов в системах ИИ, к которым относятся наборы данных, предварительно натренированные модели и средства разработки. Автор отчета Andrew Lohn, сотрудник RAND Corporation, согласно его профайлу в Google Scholar. Из свежих работ есть краткий отчет по состязательным атакам и связанным рискам [16] с достаточно неутешительным выводом: “Предлагаемые

методы защиты могут дать только краткосрочное преимущество. Противостояние атакующий-защищающийся в системах машинного обучения напоминает игру кошки-мышки. При этом обороняющиеся проигрывают, их методы защиты обходятся (преодолеваются), и они пока не успевают за атакующими. Тем не менее, защитные меры могут повысить затраты злоумышленников в некоторых узких случаях, а правильное понимание уязвимостей машинного обучения может помочь защитникам снизить риск. Можно ожидать, что эффективность защитных стратегий и тактик будет меняться в течение многих лет, но по-прежнему не сможет противостоять более сложным атакам".

Здесь можно также назвать программу DARPA Guaranteeing AI Robustness Against Deception (GARD) [17]. GARD стремится создать теоретические основы системы машинного обучения для выявления уязвимостей системы, характеристики свойств, которые повысят надежность системы, и поощрения создания эффективных средств защиты. В настоящее время защита от машинного обучения, как правило, очень специфична и эффективна только против определенных атак. GARD стремится разработать средства защиты, способные противостоять широким категориям атак. Кроме того, текущие парадигмы оценки устойчивости ИИ часто фокусируются на упрощенных мерах, которые могут не иметь отношения к безопасности. Чтобы проверить актуальность для безопасности и широкую применимость, средства защиты, созданные в рамках GARD, будут измеряться на новом испытательном стенде с использованием оценок на основе сценариев.

Последнее представляется особенно интересным в плане работ. Это уже отмечалось выше, что в настоящее время при построении систем защиты (при оценке устойчивости) используются достаточно искусственные возмущения данных, которые исходят из малозаметности изменений для человека. При этом очевидно, что человек совсем не обязательно присутствует в цепочке принятия решений для всех приложений. А для критических применений – так и вовсе отсутствует. Возможно, что сценарии будут связаны с более реалистичными изменениями.

Для выполнения работ выбраны 17 организаций, включая университеты MIT, Carnegie Mellon, а также компании Intel и IBM [18].

Федеральное правительство США финансирует национальную программу по искусственному интеллекту. В рамках этой программы по разным направлениям выбрано 7 институтов, которые получают федеральное финансирование для исследования различных аспектов искусственного интеллекта. Основаниями (базовыми элементами) систем машинного обучения в рамках этой программы занимается специально созданный Институт машинного обучения в университете Техаса [19].

Устойчивое машинное обучение входит в список его основных научных направлений: "Обучение моделей машинного обучения требует больших вычислительных ресурсов, и решение о том, как устанавливать параметры, - это больше искусство, чем наука. Нет никаких хороших математических оснований для того, как устанавливать шкалы и рычаги. Второе направление - устойчивость как к ошибкам в данных, так и к преднамеренным злоумышленным манипуляциям. Текущая модель, основанная на огромном количестве качественных (достоверных) обучающих данных, непригодна для большинства приложений. По мере того как люди пытаются использовать модели машинного обучения для приложений с высокими ставками, эта проблема будет усугубляться попытками взлома. Именно разработка надежных моделей машинного обучения станет ключом к их широкому внедрению" [20].

Отмечается, что очень важной частью разработки крупномасштабных систем искусственного интеллекта или машинного обучения является количественная оценка и оценка неопределенности, потому что без этого можно, фактически, принимать катастрофические решения в эпоху больших данных

Вместе с тем, мы пока не нашли большого количества публикаций от этой группы исследователей по заявленным темам. Либо результатов пока нет, либо пока не все публикуется.

Information Science and Technology Institute (ISTI) – это национальная лаборатория в Лос Аламосе и ее Образовательный центр национальной безопасности [21] объявили первым приоритетом для своих исследований интерпретируемость и объяснимость моделей машинного обучения: "В сфере национальной безопасности существует острая потребность в интерпретируемых моделях машинного обучения, особенно в приложениях для критических применениях (связанных с высоким риском). К сожалению, большинство общепринятых методов интерпретируемости ориентированы именно на обработку изображений и модели классификации, обученные на размеченных данных. Между тем, многие приложения национального значения в Лос-Аламосе и других местах помимо изображений используют временные ряды, текстовые или числовые данные. При этом они требуют использования моделей без учителя для таких задач, как извлечение знаний или обнаружение аномалий. Летом 2021 года проекты в рамках этой целевой области будут включать разработку и / или оценку методов интерпретируемости для моделей, которые используют текст и / или временные ряды для приложений национальной безопасности" [22].

В другой национальной лаборатории (Ливермор) Центр прикладных вычислений имеет два проекта в интересующей нас области – устойчивое машинное обучение [23] и объяснимые системы искусственного интеллекта [24].

Следующие изображение из постера Центра иллюстрируют интересующие нас направления:

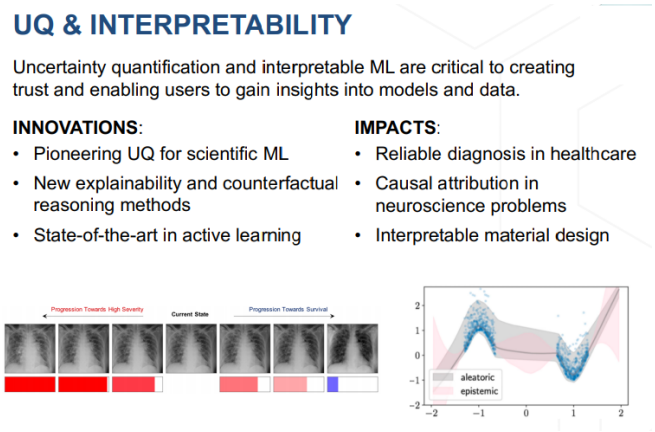

Рис. 2. Оценка неопределенности [25]

Методы количественной оценки неопределенности (UQ) играют ключевую роль в снижении влияния неопределенностей, как во время процессов оптимизации, так и в процессе принятия решений. Они применялись для решения множества реальных проблем в науке и технике. Байесовская аппроксимация и методы ансамблевого обучения - это два широко используемых типа методов количественной оценки неопределенности (UQ). Различные методы UQ используются в таких приложениях, как компьютерное зрение (например, беспилотные автомобили и обнаружение объектов), обработка изображений (например, восстановление изображений), анализ медицинских изображений (например, классификация и сегментация медицинских изображений), обработка естественного языка (например, классификация текстов, тексты в социальных сетях и оценка риска рецидивизма), биоинформатика и т. д. [26]

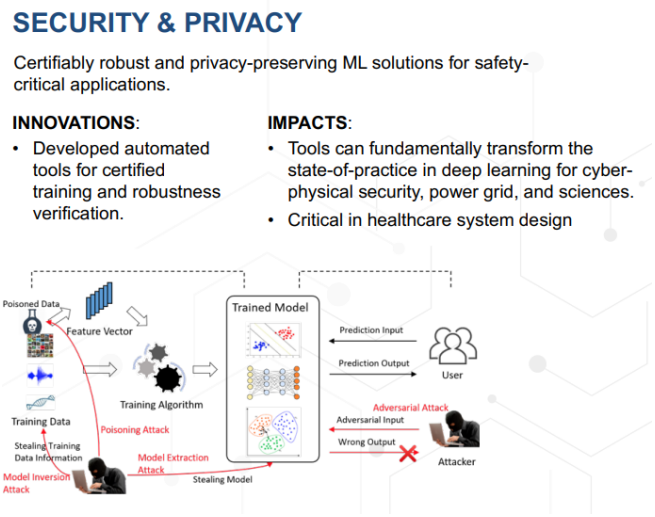

Рис. 3 Гарантированная устойчивость [25]

Сертифицированная тренировка моделей – это гарантии отсутствия троянов и закладок. А важность именно верификации моделей отмечалась ранее.

The Intelligence Advanced Research Projects Activity (IARPA) [27] инвестирует в высокорисковые и высокооплачиваемые исследовательские программы для решения некоторых из наиболее сложных задач, стоящих перед агентствами и дисциплинами в разведывательном сообществе. Среди списка проектов есть, например, исследовательский проект по закладкам в системах машинного обучения [28]. По данному проекту доступны публикации [29], работы выполнялись в Университете Беркли.

На европейском уровне можно отметить проект ELLIS - Европейская лаборатория машинного обучения [30]. Среди его программ [31] есть направление устойчивого машинного обучения [32], а также интерпретируемого машинного обучения [33]

Вместе с тем, Google Scholar видит только небольшое количество публикаций, упоминающих ellis.eu. В статьях обсуждается, почему европейское сообщество отстает от Китая в вопросах искусственного интеллекта [34].

Из академических и промышленных проектов, в первую очередь мы бы отметили группу в MIT CSAIL [35], которую возглавляет Alexander Madry [36] и Center for Deployable Machine Learning (CDML) [37]. Ссылки на работы есть на сайте лаборатории [38] , включай исходные коды Python библиотеки для проверки робастности [39]. В частности, этой группой создан учебник по Adversarial machine learning [40].

По направлению работ, в целом, это можно описать как adversarial training. Вместе с тем, учитывая большой состав групп, они охватывают довольно широкий спектр проблем и имеют впечатляющее количество публикаций, а также очень интересные PhD диссертации.

Сотрудники MIT стоят также за открытым репозиторием защит Robust ML [41]. На сайте публикуются защиты от атак типа белого ящика на системы классификации изображений (здесь авторы следуют общему тренду, когда все, что касается робастности относится, в подавляющем большинстве случаев, к работе с изображениями). Здесь же публикуются и опровержения защит (то есть, атаки, которые их преодолевают). При этом нужно понимать, что защита в данном случае – это, например, гарантированные границы для функции потерь при заданном ограничении на модификацию [42]. То есть это ни в коей мере не является гарантированным подтверждением работоспособности, как это принято понимать, например, для программного обеспечения критических систем.

На таком же уровне значимости, если не выше, следует рассматривать Google (Deepmind). Их собственный манифест относительно устойчивого машинного обучения [43] отмечает, что с точки зрения программиста, ошибка - это любое поведение, несовместимое со спецификацией, то есть предполагаемой функциональностью системы. Deepmind проводит исследования методов оценки соответствия систем машинного обучения не только обучающему и тестирующему набору, но и списку спецификаций, описывающих желаемые свойства системы. Такие свойства могут включать устойчивость к

достаточно небольшим возмущениям на входе, ограничения безопасности, позволяющие избежать катастрофических отказов, или создание прогнозов, согласующихся с законами физики.

В рамках такого исследования, Deepmind выделяет три важные технические задачи, которые, по мнению компании, предстоит решить сообществу машинного обучения (то есть, задачи носят абсолютно общий характер):

1) Эффективное тестирование соответствия спецификациям. Deepmind исследует эффективные способы проверки соответствия систем машинного обучения свойствам (таким как инвариантность или надежность), желаемым разработчиком и пользователями системы. Один из подходов к выявлению случаев, когда модель может не соответствовать желаемому поведению, заключается в систематическом поиске наихудших результатов во время оценки.

2) Тренировка (обучение) моделей машинного обучения в соответствии со спецификациями. Даже при большом количестве обучающих данных стандартные алгоритмы машинного обучения могут создавать прогнозные модели, которые делают прогнозы несовместимыми с желательными спецификациями, такими как надежность или справедливость. Соответственно, это требует пересмотра алгоритмов обучения, так, чтобы можно было создавать модели, которые не только хорошо подходят для обучающих данных, но и согласуются с заданными спецификациями.

3) Формальное доказательство соответствия моделей машинного обучения спецификациям. Существует потребность в алгоритмах, которые могут проверить, что предсказания модели доказуемо согласуются со спецификацией, представляющей интерес для всех возможных входных данных. Хотя в области формальной проверки такие алгоритмы изучаются в течение нескольких десятилетий, эти подходы трудно масштабировать до современных систем глубокого обучения.

В качестве конкретных направлений заявлено следующее:

a) Обучение состязательной оценке и проверке: по мере масштабирования и усложнения систем ИИ будет все труднее разрабатывать алгоритмы состязательной оценки и проверки, хорошо адаптированные к модели ИИ. Если удастся использовать возможности ИИ для облегчения оценки и проверки, этот процесс можно будет масштабировать.

b) Разработка общедоступных инструментов для состязательной оценки и проверки: важно предоставить инженерам и практикам ИИ простые в использовании инструменты, которые предиктивно (до наступления негативных последствий) оценивают возможные виды отказов системы ИИ. Это потребует стандартизации алгоритмов состязательной оценки и проверки. Упоминавшийся выше проект RobustML [41], кстати, также можно рассматривать как попытку предложить некоторый стандарт для состязательных атак и проверок.

c) Расширение диапазона состязательных примеров: на сегодняшний день большая часть работы над состязательными примерами сосредоточена на инвариантности модели к небольшим возмущениям, как правило, изображений. Это стало отличной испытательной площадкой для разработки подходов к состязательной оценке, надежному обучению и проверке, но нужны и альтернативные спецификации, имеющие непосредственное отношение к реальному миру.

Этот пункт хотелось бы выделить особо. Ограниченность модификаций и повсеместная ориентация на классификацию изображений являются одними из самых серьезных ограничений существующих подходов.

При этом в Deepmind отмечают, что “ручное” создание спецификаций для систем ИИ будет затруднено. Соответственно, необходимы системы, которые могут использовать частичные “человеческие” спецификации и изучать дополнительные спецификации на основе оценочной обратной связи. Отсюда – внимание к системам, использующим обучение с подкреплением [44].

Очень интересным является манифест Университета Карнеги Меллон [73], который описывает проблемы с устойчивостью систем машинного обучения с позиций необходимых программных систем. Работа выпущена в рамках исследовательского проекта по инженерии систем Искусственного интеллекта [74]. Эти работы финансируются офисом национальной разведки - U.S. Office of the Director of National Intelligence (ODNI)

Из других проектов можно отметить проект Fairness & Robustness in Machine Learning института математики университета Тулузы [45]. Представленные результаты касаются больше справедливости и интерпретируемости результатов.

ETH Zurich поддерживает проект Safe Artificial Intelligence [46]. Судя по представленным публикациям, большое внимание уделяется сертификации устойчивости систем машинного обучения. Сотрудниками, участвующими в проекте, создана промышленная компания Latticeflow [47], которая позиционирует себя как первая в мире доверительная платформа искусственного интеллекта, позволяющая организациям создавать и развертывать надежные модели искусственного интеллект. Среди клиентов Швейцарские федеральные железные дороги (SBB) и Siemens, а также правительственные учреждения, такие как армия США и Федеральное управление информационной безопасности Германии (BSI). Компания не приводит в открытом виде описания продукта (впрочем, это характерно для многих программных продуктов, касающихся безопасности), но

представление о назначении и схеме работы можно понять из имеющегося описания проекта по оценке устойчивости системы распознавания железнодорожных знаков [48]. Система LatticeFlow использовалась для тестирования, когда к существующим данных (изображениям) применялись некоторые стандартные преобразования (повороты, изменение цвета и яркости, фона и т.д.), после чего оценивалось качество распознавания (классификации) уже с новыми данными.

Далее можно отметить проект Института Алана Тьюринга по устойчивому машинному обучению [49]. Безопасный и надежный ИИ является приоритетной областью в дорожной карте правительства Великобритании в области ИИ [50]. Работы по проекту не публикуются на сайте, но есть, по крайней мере, еще одна программа, посвященная adversarial machine learning [51]. Этот проект касается также классификации изображений, относится к направлению Защиты и безопасности. Последние работы участников группы, согласно Google Scholar, посвящены как раз сдвигу распределения [52] (совместная работа с сотрудниками Яндекс). База для этих проектов – исследовательская группа в Оксфорде - Oxford Applied and Theoretical Machine Learning Group [53]. Ее публикации по теме состязательного обучения выделены в отдельную группу [54].

Allen Institute for AI (Paul Allen, Microsoft co-founder) [55] также поддерживает исследования в области состязательных моделей. Пример – семинары в университете Вашингтона [56].

В университете Калифорнии, Беркли есть группа, которая занимается развитием верифицируемого ИИ [57]. Исследовательская группа, которая стоит за этим направлением [58] как занимается формальными методами верификации, равно как и разработкой соответствующих приложений (SMT solvers). С учетом важности формальных методов верификации для систем машинного обучения, это один из наиболее интересных проектов. При этом же университете существует Центр проблем кибербезопасности [68]. Устойчивое машинное обучение является одним из его направлений [69].

IV Индустриальные проекты и стартапы

Из промышленных R&D центров можно назвать, например, Центр компании Bosch по искусственному интеллекту [59]. Состязательные атаки рассматриваются группой, работающей в области объяснимых моделей глубинного обучения [60].

Компания Форд имеет свою группу Core-Artificial Intelligence/Machine Learning (AI/ML). Пример - объявление о найме, касающееся работ по устойчивому машинному обучению [61].

Яндекс объявил конкурс по поиску решений эффективной работы со сдвигом распределения [62]. Цель задачи - повысить осведомленность о сдвигах в распределении реальных данных. Целью участников будет разработка моделей, устойчивых к сдвигу в распределении, и выявление такого сдвига с помощью мер неопределенности в их прогнозах. Участники могут принять участие в трех отдельных треках, для которых Яндекс предоставил наборы данных в трех областях: прогноз погоды, машинный перевод и прогноз движения транспортных средств.

Стартап Adversa [63] предлагает услуги состязательного тестирования ИИ систем. Это соответствует тому, что ИТ компании делают, например, для сторонних веб-сервисов – тестирование проникновения [64]. Компания опубликовала довольно интересный отчет по атакам [65]. К этой же группе можно отнести проект по сбору информации относительно проблем (инцидентов), связанных с искусственным интеллектом [72]. Единый подход к описанию важен с точки зрения преодоления силоса данных [70, 71].

Естественно, проекты открываются и в области понимания (интерпретации результатов) систем машинного обучения. Или даже в более широком смысле, как, например, Robust AI [66] - когнитивная платформа промышленного уровня, которая дает роботам возможность рассуждать о здравом смысле. Здесь исследования касаются уже не только (и не столько) глубинного обучения, но и символического ИИ.

V Заключение

Как видно из приведенного обзора, явно заявили о работах в области верификации систем машинного обучения, то есть – формального доказательства их работы с заданными данными, что только и является “настоящим” подтверждением устойчивости (настоящим по сравнению с другими методами) в Google Deepmind и университете Беркли. Вместе с тем необходимо отметить, что публикаций по заявленным проектам пока мало. Это может быть связано как с тем, что пока еще нет значимых результатов, так и с ограничениями на публикацию. Возможно, для последующих версий обзора необходимо добавить обзор патентных заявок, которые могут предшествовать публикациям результатов. В заявленных проектах по-прежнему доминирует классификация изображений. Вместе с тем, необходимо отметить, что наличие состязательных примеров – это фундаментальная характеристика текущей архитектуры систем машинного обучения, когда данные разделены, и тренировочные данные, вообще говоря, могут сколь угодно отличаться от генеральной совокупности. Соответственно, необходимость работы со сдвигом распределения, необходимость построения обобщений “на лету”, используя малое число примеров должно стать частью новой модели (моделей) систем глубинного обучения, о чем говорится в свежей статье Yoshua Bengio, Yann Lecun и Geoffrey Hinton (тех, кто практически стоял у истоков текущих моделей) [67].

Благодарности

Мы благодарны сотрудникам кафедры Информационной безопасности факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова за ценные обсуждения данной работы.

Библиография

[1] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.
[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199 , 2013
[3] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, “Poison frogs! targeted clean-label poisoning attacks on neural networks,” in Advances in Neural Information Processing Systems, 2018, pp. 6103–6113.
[4] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” IEEE transactions on neural networks and learning systems, 2019.
[5] A Complete List of All (arXiv) Adversarial Example Papers https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. Retrieved: Aug, 2021
[6] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: Aug, 2021
[7] Koh, Pang Wei, et al. "Wilds: A benchmark of in-the-wild distribution shifts." International Conference on Machine Learning. PMLR, 2021.
[8] Nair, Nimisha G., Pallavi Satpathy, and Jabez Christopher. "Covariate shift: A review and analysis on classifiers." 2019 Global Conference for Advancement in Technology (GCAT). IEEE, 2019.
[9] Major ML datasets have tens of thousands of errors https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors Retrieved: Aug, 2021
[10] NeuIPS https://www.llnl.gov/news/neurips-papers-aim-improve-understanding-and-robustness-machine-learning-algorithms
[11] Pei, Kexin, et al. "Towards practical verification of machine learning: The case of computer vision systems." arXiv preprint arXiv:1712.01785 (2017).
[12] Katz, Guy, et al. "The marabou framework for verification and analysis of deep neural networks." International Conference on Computer Aided Verification. Springer, Cham, 2019.
[13] Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." IEEE Design & Test 37.2 (2020): 30-57.
[14] The Pentagon Is Bolstering Its AI Systems—by Hacking Itself https://www.wired.com/story/pentagon-bolstering-ai-systems-hacking-itself Retrieved: Aug, 2021
[15] Poison in the Well Securing the Shared Resources of Machine Learning https://cset.georgetown.edu/publication/poison-in-the-well/ Retrieved: Aug, 2021
[16] Hacking AI A PRIMER FOR POLICYMAKERS ON MACHINE LEARNING CYBERSECURITY https://cset.georgetown.edu/wp-content/uploads/CSET-Hacking-AI.pdf Retrieved: Aug, 2021
[17] Guaranteeing AI Robustness Against Deception https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception Retrieved: Aug, 2021
[18] DARPA is pouring millions into a new AI defense program. Here are the companies leading the charge https://www.protocol.com/intel-darpa-adversarial-ai-project Retrieved: Aug, 2021
[19] UT Austin Selected as Home of National AI Institute Focused on Machine Learning https://news.utexas.edu/2020/08/26/ut-austin-selected-as-home-of-national-ai-institute-focused-on-machine-learning/ Retrieved: Aug, 2021
[20] UT Austin Launches Institute to Harness the Data Revolution https://ml.utexas.edu/news/611 Retrieved: Aug, 2021
[21] National Security Education Center https://www.lanl.gov/projects/national-security-education-center/ Retrieved: Aug, 2021
[22] 2021 Project Descriptions Creates next-generation leaders in Machine Learning for Scientific Applications https://www.lanl.gov/projects/national-security-education-center/information-science-technology/summer-schools/applied-machine-learning/project-descriptions-2019.php Retrieved: Aug, 2021
[23] Assured Machine Learning: Robustness, Fairness, and Privacy https://computing.llnl.gov/casc/ml/robust Retrieved: Aug, 2021
[24] Explainable Artificial Intelligence https://computing.llnl.gov/casc/ml/ai Retrieved: Aug, 2021
[25] Advancing Machine Learning for Mission-Critical Applications https://computing.llnl.gov/sites/default/files/COMP_ROADSHOW_ML_CASC-final.pdf Retrieved: Aug, 2021
[26] Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information Fusion (2021).
[27] Intelligence Advanced Research Projects Activity (IARPA) https://www.iarpa.gov/ Retrieved: Aug, 2021
[28] Trojans in Artificial Intelligence https://www.iarpa.gov/index.php/research-programs/trojai Retrieved: Aug, 2021
[29] Trojans in Artificial Intelligence bibliography https://scholar.google.com/scholar?hl=en&as_sdt=0%2C47&q=W911NF20C0034+OR+W911NF20C0038+OR+W911NF20C0045+OR+W911NF20C0035+OR+IARPA-20001-D2020-2007180011 Retrieved: Aug, 2021
[30] ELLIS Programs launched https://ellis.eu/news/ellis-programs-launched Retrieved: Aug, 2021
[31] ELLIS programs https://ellis.eu/programs Retrieved: Aug, 2021
[32] Robust Machine Learning https://ellis.eu/programs/robust-machine-learning Retrieved: Aug, 2021
[33] Semantic, Symbolic and Interpretable Machine Learning https://ellis.eu/programs/semantic-symbolic-and-interpretable-machine-learning Retrieved: Aug, 2021
[34] Oomen, Thomas L. "Why the EU lacks behind China in AI development–Analysis and solutions to enhance EU’s AI strategy." rue 33.1: 7543.
[35] MIT Reliable and Robust Machine Learning https://www.csail.mit.edu/research/reliable-and-robust-machine-learning Retrieved: Aug, 2021
[36] Alexander Madry http://people.csail.mit.edu/madry/ Retrieved: Aug, 2021
[37] Center for Deployable Machine Learning (CDML) https://www.csail.mit.edu/research/center-deployable-machine-learning-cdml Retrieved: Aug, 2021
[38] Madry Lab http://madry-lab.ml/ Retrieved: Aug, 2021
[39] Robustness package https://github.com/MadryLab/robustness Retrieved: Aug, 2021
[40] Adversarial ML tutorial https://adversarial-ml-tutorial.org/ Retrieved: Aug, 2021
[41] RobustML https://www.robust-ml.org/ Retrieved: Aug, 2021
[42] Andriushchenko, Maksym, and Matthias Hein. "Provably robust boosted decision stumps and trees against adversarial attacks." arXiv preprint arXiv:1906.03526 (2019).
[43] Identifying and eliminating bugs in learned predictive models https://deepmind.com/blog/article/robust-and-verified-ai Retrieved: Aug, 2021
[44] Nandy, Abhishek, and Manisha Biswas. "Google’s DeepMind and the Future of Reinforcement Learning." Reinforcement Learning. Apress, Berkeley, CA, 2018. 155-163.
[45] Fairness & Robustness in Machine Learning https://perso.math.univ-toulouse.fr/loubes/fairness-robustness-in-machine-learning/ Retrieved: Aug, 2021
[46] Safe Artificial Intelligence http://safeai.ethz.ch/ Retrieved: Aug, 2021
[47] Latticeflow https://latticeflow.ai/ Retrieved: Aug, 2021
[48] Reliability Assessment of Traffic Sign Classifiers https://latticeflow.ai/wp-content/uploads/2021/01/Reliability_assessment_of_traffic_sign_classifiers_short.pdf Retrieved: Aug, 2021
[49] The Alan Turing Institute Robust machine learning https://www.turing.ac.uk/research/interest-groups/robust-machine-learning Retrieved: Aug, 2021
[50] AI roadmap https://www.gov.uk/government/publications/ai-roadmap Retrieved: Aug, 2021
[51] The Alan Turing Institute Adversarial machine learning https://www.turing.ac.uk/research/research-projects/adversarial-machine-learning Retrieved: Aug, 2021
[52] Malinin, Andrey, et al. "Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks." arXiv preprint arXiv:2107.07455 (2021).
[53] Oxford Applied and Theoretical Machine Learning Group https://oatml.cs.ox.ac.uk/ Retrieved: Aug, 2021

[54] Adversarial and Interpretable ML — Publications https://oatml.cs.ox.ac.uk/tags/adversarial_interpretability.html#title Retrieved: Aug, 2021
[55] Allen Institute for AI https://allenai.org/ Retrieved: Aug, 2021
[56] AI2 Machine Learning Seminars https://www.cs.washington.edu/research/ml/seminars Retrieved: Aug, 2021
[57] Verified AI https://berkeleylearnverify.github.io/VerifiedAIWebsite/ Retrieved: Aug, 2021
[58] Sanjit A. Seshia research group https://people.eecs.berkeley.edu/~sseshia/ Retrieved: Aug, 2021
[59] Bosch AI https://www.bosch-ai.com/ Retrieved: Aug, 2021
[60] Rich and Explainable Deep Learning https://www.bosch-ai.com/research/research-fields/rich_and_explainable_deep_learning_perception/ Retrieved: Aug, 2021
[61] Research Engineer – Robust and Explainable AI Methods https://www.mendeley.com/careers/job/research-engineer-robust-and-explainable-ai-methods-690764 Retrieved: Aug, 2021
[62] Yandex Shift Challenge https://research.yandex.com/shifts Retrieved: Aug, 2021
[63] Adversa https://adversa.ai/ Retrieved: Aug, 2021
[64] De Jimenez, Rina Elizabeth Lopez. "Pentesting on web applications using ethical-hacking." 2016 IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI). IEEE, 2016.
[65] The Road to Secure and Trusted AI https://adversa.ai/report-secure-and-trusted-ai/ Retrieved: Aug, 2021
[66] Robust AI https://www.robust.ai/ Retrieved: Aug, 2021
[67] Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton. "Deep learning for AI." Communications of the ACM 64.7 (2021): 58-65.
[68] Center for Long-Term Cybersecurity University of California, Berkeley. https://cltc.berkeley.edu/ Retrieved: Sep, 2021
[69] Center for Long-Term Cybersecurity University of California, Berkeley, Robust ML. https://cltc.berkeley.edu/?s=robust Retrieved: Sep, 2021
[70] Куприяновский, В. П., et al. "Оптимизация использования ресурсов в цифровой экономике." International Journal of Open Information Technologies 4.12 (2016).
[71] Куприяновский, В. П., et al. "Цифровая экономика и Интернет Вещей-преодоление силоса данных." International Journal of Open Information Technologies 4.8 (2016): 36-42.
[72] Incident Database https://incidentdatabase.ai/ Retrieved: Sep, 2021
[73] Robust and Secure AI https://resources.sei.cmu.edu/asset_files/WhitePaper/2021_019_001_735346.pdf Retrieved: Sep, 2021
[74] Artificial Intelligence Engineering https://sei.cmu.edu/our-work/artificial-intelligence-engineering/ Retrieved: Sep, 2021

Ongoing academic and industrial projects dedicated to robust machine learning

Dmitry Namiot, Eugene Ilyushin, Ivan Chizhov

***Abstract*— With the growing use of systems based on machine learning, which, from a practical point of view, are considered as systems of artificial intelligence today, the attention to the issues of reliability (robustness) of such systems and solutions is also growing. Naturally, for critical applications, for example, systems that make decisions in real time, robustness issues are the most important from the point of view of the practical use of machine learning systems. In fact, it is the robustness assessment that determines the very possibility of using machine learning in such systems. This, in a natural way, is reflected in a large number of works devoted to the issues of assessing the robustness of machine learning systems, the architecture of such systems and the protection of machine learning systems from malicious actions that can affect their operation. At the same time, it is necessary to understand that robustness problems can arise both naturally, due to the different distribution of data at the stages of training and practical application (at the stage of training the model, we use only part of the data from the general population), and as a result of targeted actions (attacks on machine learning systems). In this case, attacks can be directed both at the data and at the models themselves.**

***Keywords*—robust machine learning, adversarial machine learning.**

REFERENCES

[1] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199 , 2013

[3] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in Advances in Neural Information Processing Systems, 2018, pp. 6103–6113.

[4] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," IEEE transactions on neural networks and learning systems, 2019.

[5] A Complete List of All (arXiv) Adversarial Example Papers https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html. Retrieved: Aug, 2021

[6] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: Aug, 2021

[7] Koh, Pang Wei, et al. "Wilds: A benchmark of in-the-wild distribution shifts." International Conference on Machine Learning. PMLR, 2021.

[8]Nair, Nimisha G., Pallavi Satpathy, and Jabez Christopher. "Covariate shift: A review and analysis on classifiers." 2019 Global Conference for Advancement in Technology (GCAT). IEEE, 2019.

[9] Major ML datasets have tens of thousands of errors https://www.csail.mit.edu/news/major-ml-datasets-have-tens-thousands-errors Retrieved: Aug, 2021

[10] NeuIPS https://www.llnl.gov/news/neurips-papers-aim-improve-understanding-and-robustness-machine-learning-algorithms

[11] Pei, Kexin, et al. "Towards practical verification of machine learning: The case of computer vision systems." arXiv preprint arXiv:1712.01785 (2017).

[12] Katz, Guy, et al. "The marabou framework for verification and analysis of deep neural networks." International Conference on Computer Aided Verification. Springer, Cham, 2019.

[13] Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." IEEE Design & Test 37.2 (2020): 30-57.

[14] The Pentagon Is Bolstering Its AI Systems—by Hacking Itself https://www.wired.com/story/pentagon-bolstering-ai-systems-hacking-itself Retrieved: Aug, 2021

[15] Poison in the Well Securing the Shared Resources of Machine Learning https://cset.georgetown.edu/publication/poison-in-the-well/ Retrieved: Aug, 2021

[16] Hacking AI A PRIMER FOR POLICYMAKERS ON MACHINE LEARNING CYBERSECURITY https://cset.georgetown.edu/wp-content/uploads/CSET-Hacking-AI.pdf Retrieved: Aug, 2021

[17] Guaranteeing AI Robustness Against Deception https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception Retrieved: Aug, 2021

[18] DARPA is pouring millions into a new AI defense program. Here are the companies leading the charge https://www.protocol.com/intel-darpa-adversarial-ai-project Retrieved: Aug, 2021

[19] UT Austin Selected as Home of National AI Institute Focused on Machine Learning https://news.utexas.edu/2020/08/26/ut-austin-selected-as-home-of-national-ai-institute-focused-on-machine-learning/ Retrieved: Aug, 2021

[20] UT Austin Launches Institute to Harness the Data Revolution https://ml.utexas.edu/news/611 Retrieved: Aug, 2021

[21] National Security Education Center https://www.lanl.gov/projects/national-security-education-center/ Retrieved: Aug, 2021

[22] 2021 Project Descriptions Creates next-generation leaders in Machine Learning for Scientific Applications https://www.lanl.gov/projects/national-security-education-center/information-science-technology/summer-schools/applied-machine-learning/project-descriptions-2019.php Retrieved: Aug, 2021

[23] Assured Machine Learning: Robustness, Fairness, and Privacy https://computing.llnl.gov/casc/ml/robust Retrieved: Aug, 2021

[24] Explainable Artificial Intelligence https://computing.llnl.gov/casc/ml/ai Retrieved: Aug, 2021

[25] Advancing Machine Learning for Mission-Critical Applications https://computing.llnl.gov/sites/default/files/COMP_ROADSHOW_ML_CASC-final.pdf Retrieved: Aug, 2021

[26] Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information Fusion (2021).

[27] Intelligence Advanced Research Projects Activity (IARPA) https://www.iarpa.gov/ Retrieved: Aug, 2021

[28] Trojans in Artificial Intelligence https://www.iarpa.gov/index.php/research-programs/trojai Retrieved: Aug, 2021

[29] Trojans in Artificial Intelligence bibliography https://scholar.google.com/scholar?hl=en&as_sdt=0%2C47&q=W911NF20C0034+OR+W911NF20C0038+OR+W911NF20C0045+OR+W911NF20C0035+OR+IARPA-20001-D2020-2007180011 Retrieved: Aug, 2021

[30] ELLIS Programs launched https://ellis.eu/news/ellis-programs-launched Retrieved: Aug, 2021

[31] ELLIS programs https://ellis.eu/programs Retrieved: Aug, 2021

[32] Robust Machine Learning https://ellis.eu/programs/robust-machine-learning Retrieved: Aug, 2021

[33] Semantic, Symbolic and Interpretable Machine Learning https://ellis.eu/programs/semantic-symbolic-and-interpretable-machine-learning Retrieved: Aug, 2021

[34] Oomen, Thomas L. "Why the EU lacks behind China in AI development–Analysis and solutions to enhance EU's AI strategy." rue 33.1: 7543.

[35] MIT Reliable and Robust Machine Learning https://www.csail.mit.edu/research/reliable-and-robust-machine-learning Retrieved: Aug, 2021
[36] Alexander Madry http://people.csail.mit.edu/madry/ Retrieved: Aug, 2021
[37] Center for Deployable Machine Learning (CDML) https://www.csail.mit.edu/research/center-deployable-machine-learning-cdml Retrieved: Aug, 2021
[38] Madry Lab http://madry-lab.ml/ Retrieved: Aug, 2021
[39] Robustness package https://github.com/MadryLab/robustness Retrieved: Aug, 2021
[40] Adversarial ML tutorial https://adversarial-ml-tutorial.org/ Retrieved: Aug, 2021
[41] RobustML https://www.robust-ml.org/ Retrieved: Aug, 2021
[42] Andriushchenko, Maksym, and Matthias Hein. "Provably robust boosted decision stumps and trees against adversarial attacks." arXiv preprint arXiv:1906.03526 (2019).
[43] Identifying and eliminating bugs in learned predictive models https://deepmind.com/blog/article/robust-and-verified-ai Retrieved: Aug, 2021
[44] Nandy, Abhishek, and Manisha Biswas. "Google's DeepMind and the Future of Reinforcement Learning." Reinforcement Learning. Apress, Berkeley, CA, 2018. 155-163.
[45] Fairness & Robustness in Machine Learning https://perso.math.univ-toulouse.fr/loubes/fairness-robustness-in-machine-learning/ Retrieved: Aug, 2021
[46] Safe Artificial Intelligence http://safeai.ethz.ch/ Retrieved: Aug, 2021
[47] Latticeflow https://latticeflow.ai/ Retrieved: Aug, 2021
[48] Reliability Assessment of Traffic Sign Classifiers https://latticeflow.ai/wp-content/uploads/2021/01/Reliability_assessment_of_traffic_sign_classifiers_short.pdf Retrieved: Aug, 2021
[49] The Alan Turing Institute Robust machine learning https://www.turing.ac.uk/research/interest-groups/robust-machine-learning Retrieved: Aug, 2021
[50] AI roadmap https://www.gov.uk/government/publications/ai-roadmap Retrieved: Aug, 2021
[51] The Alan Turing Institute Adversarial machine learning https://www.turing.ac.uk/research/research-projects/adversarial-machine-learning Retrieved: Aug, 2021
[52] Malinin, Andrey, et al. "Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks." arXiv preprint arXiv:2107.07455 (2021).
[53] Oxford Applied and Theoretical Machine Learning Group https://oatml.cs.ox.ac.uk/ Retrieved: Aug, 2021
[54] Adversarial and Interpretable ML — Publications https://oatml.cs.ox.ac.uk/tags/adversarial_interpretability.html#title Retrieved: Aug, 2021
[55] Allen Institute for AI https://allenai.org/ Retrieved: Aug, 2021
[56] AI2 Machine Learning Seminars https://www.cs.washington.edu/research/ml/seminars Retrieved: Aug, 2021
[57] Verified AI https://berkeleylearnverify.github.io/VerifiedAIWebsite/ Retrieved: Aug, 2021
[58] Sanjit A. Seshia research group https://people.eecs.berkeley.edu/~sseshia/ Retrieved: Aug, 2021
[59] Bosch AI https://www.bosch-ai.com/ Retrieved: Aug, 2021
[60] Rich and Explainable Deep Learning https://www.bosch-ai.com/research/research-fields/rich_and_explainable_deep_learning_perception/ Retrieved: Aug, 2021
[61] Research Engineer – Robust and Explainable AI Methods https://www.mendeley.com/careers/job/research-engineer-robust-and-explainable-ai-methods-690764 Retrieved: Aug, 2021
[62] Yandex Shift Challenge https://research.yandex.com/shifts Retrieved: Aug, 2021
[63] Adversa https://adversa.ai/ Retrieved: Aug, 2021
[64] De Jimenez, Rina Elizabeth Lopez. "Pentesting on web applications using ethical-hacking." 2016 IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI). IEEE, 2016.
[65] The Road to Secure and Trusted AI https://adversa.ai/report-secure-and-trusted-ai/ Retrieved: Aug, 2021
[66] Robust AI https://www.robust.ai/ Retrieved: Aug, 2021
[67] Bengio, Yoshua, Yann Lecun, and Geoffrey Hinton. "Deep learning for AI." Communications of the ACM 64.7 (2021): 58-65.
[68] Center for Long-Term Cybersecurity University of California, Berkeley. https://cltc.berkeley.edu/ Retrieved: Sep, 2021
[69] Center for Long-Term Cybersecurity University of California, Berkeley, Robust ML. https://cltc.berkeley.edu/?s=robust Retrieved: Sep, 2021
[70] Kuprijanovskij, V. P., et al. "Optimizacija ispol'zovanija resursov v cifrovoj jekonomike." International Journal of Open Information Technologies 4.12 (2016).
[71] Kuprijanovskij, V. P., et al. "Cifrovaja jekonomika i Internet Veshhej-preodolenie silosa dannyh." International Journal of Open Information Technologies 4.8 (2016): 36-42.
[72] Incident Database https://incidentdatabase.ai/ Retrieved: Sep, 2021
[73] Robust and Secure AI https://resources.sei.cmu.edu/asset_files/WhitePaper/2021_019_001_735346.pdf Retrieved: Sep, 2021
[74] Artificial Intelligence Engineering https://sei.cmu.edu/our-work/artificial-intelligence-engineering/ Retrieved: Sep, 2021