Основания для работ по 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. Для так называемых критических применений, например, систем, принимающих решения в реальном времени, специальных систем и т.п. вопросы устойчивости являются определяющими с точки зрения практического использования систем машинного обучения. Применение систем машинного обучения (систем искусственного интеллекта, что сейчас является, фактически, синонимом) в такого рода областях возможно только при доказательстве устойчивости (определении гарантированных параметров работы). Проблемы с устойчивость возникают из-за разных характеристик данных на этапе обучения (тренировки) и тестирования (практического применения). При этом дополнительную сложность создает тот факт, что помимо естественных причин (несбалансированные выборки, ошибки измерений и т.п.) данные могут модифицироваться сознательно. Это так называемые атаки на системы машинного обучения. Соответственно, без защиты от подобного рода действий говорить о надежности систем машинного обучения нельзя. Атаки при этом могут быть направлены как на данные, так и на сами модели.**

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

I. Введение

Эта статья написана в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по подготовке и развитию магистерской программы "Artificial Intelligence in Cybersecurity" [1]. Работа является продолжением серии публикаций, начатых в [2].

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 [3]. 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. И, 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. И эти атаки, как таковые, могут быть направлены на все аспекты (элементы) системы машинного обучения. Атаке (специальным модификациям) могут быть подвергнуты тренировочные данные (отравление данных), сама модель (бэкдоры), а также тестовые данные (состязательные примеры). Соответственно, имея в виду возможные атаки (а для критических систем это особенно важно, поскольку вероятность таких атак повышается), система машинного обучения (глубинного обучение) не может быть автоматически объявлена надежной (устойчивой). Такого рода свойства должны подтверждаться и гарантироваться.

Остальная часть статьи структурирована следующим образом. В разделе II кратко приводится история проблемы. В разделе III приводится характеристика направлений работ в данной области. И, наконец, раздел IV представляет собой заключение

Статья получена 12 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).

II. О НАЧАЛЕ РАБОТ ПО УСТОЙЧИВОМУ МАШИННОМУ ОБУЧЕНИЮ

Собственно говоря, все началось именно с атак. Возможно, просто потому, что это были просто яркие (и простые) примеры воздействия на работу систем машинного обучения. Было продемонстрировано, что существующие модели уязвимы для тщательно созданных состязательных (состязательный здесь и далее – опровергающий) примеров [4,5,6]. Это проиллюстрировано на рисунках 1 и 2, где модификации в данных (изображения и аудио) вызывали неправильную работу классификатора.

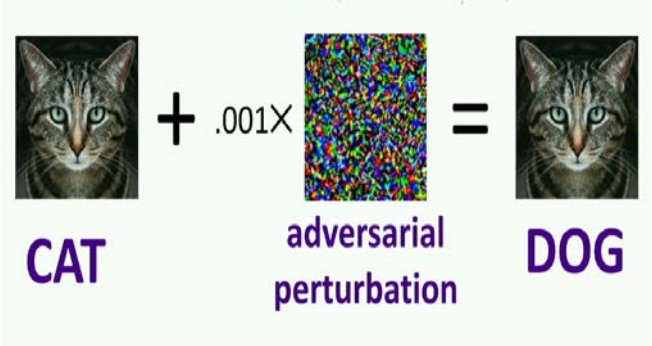

Рис 1. Состязательный пример для изображений

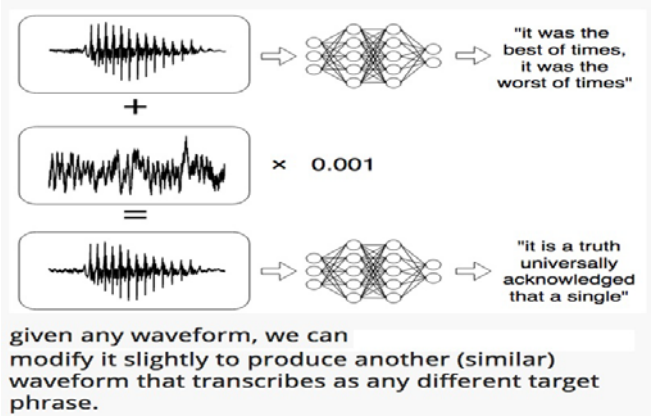

Рис. 2. Состязательный пример для аудио

С этого момента и началась история исследований в области устойчивого (надежного) машинного обучения. Форма появления состязательных примеров, фактически, определила и форму ответа. Исследования в данной области развивались в форме атака – защита (для известной атаки или атак предлагались методы снижения их влияния). А исследование атак именно на системы анализа (классификации) изображений остается предпочтительной областью исследования.

Появилось достаточно много работ, посвященных классификации атак. Делятся они по месту приложения (тренировочные данные, тестовые данные и собственно модель), а также осведомленности атакующего об архитектуре и работе атакуемой системы. На рисунке 3 приведен график по количеству публикаций, посвященных состязательным примерам. На нем явно виден резкий рост, начиная с 2018 года.

Большая часть работ касается именно анализа изображений. Другие рассматриваемые области – распознавание речи, анализ текстов и временных рядов.

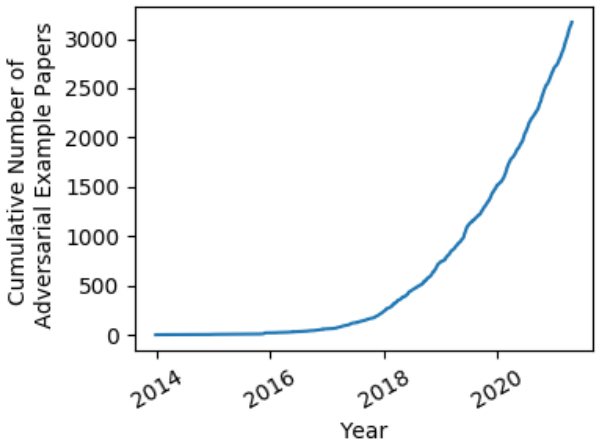

Рис.3. Публикации, посвященные состязательным примерам [7]

Но общим, на сегодняшний день, является доминирование атак над защитами. Большинство работ следует одной и той же модели. Сначала появляется статья об атаке того или иного рода, которая сравнивается с ранее известными атаками, а затем появляется работа о подходах для противодействия, которая сравнивается, в свою очередь, с существовавшими методами защиты. Но в любом случае – атаки предшествуют защите.

III. УСТОЙЧИВЫЕ СИСТЕМЫ МАШИННОГО ОБУЧЕНИЯ

Безопасность и надежность, о которых говорилось выше – это получение на реальных примерах данных (показателей), которые демонстрировались на этапе обучения. Причины, по которым приходится говорить о проблемах надежности, на самом деле, деле носят принципиальный характер для систем ML (DL) – данные, на которых проводилось обучение, отличаются от данных в реальных (рабочих) примерах. Отсюда и возникает тема устойчивости – система должна быть (теоретически, по крайней мере) устойчивой по отношению к такому изменению данных. Устойчивость в таком случае есть синоним надежности – работа системы сохраняется при изменении данных.

Классически, алгоритм, в котором погрешность, допущенная в начальных данных или допускаемая при вычислениях, с каждым шагом не увеличивается или увеличивается незначительно, называется устойчивым. В противном случае, если погрешность существенно увеличивается от шага к шагу, алгоритм называется неустойчивым. Устойчивость алгоритма – это мера его чувствительности к изменениям в исходных данных [8].

В компьютерных науках (информатике) устойчивость - это способность компьютерной системы справляться с ошибками во время выполнения [9, 10], а также способность справляться с ошибочным вводом [10]. Устойчивость может охватывать многие области информатики, такие как устойчивое (надежное)

программирование, устойчивое машинное обучение и устойчивая сеть безопасности и т.д. Формальные методы, такие как нечеткое тестирование (фаззинг), необходимы для демонстрации устойчивости, поскольку этот тип тестирования включает неверные или неожиданные входные данные. В качестве альтернативы для проверки устойчивости может использоваться искусственное внедрение неисправностей (в англоязычной литературе - fault injection).

Под устойчивым (надежным) машинным обучением обычно понимается устойчивость (надежность) алгоритмов машинного обучения. Чтобы алгоритм машинного обучения считался надежным, либо ошибка тестирования должна согласовываться с ошибкой обучения, либо производительность должна быть стабильной после добавления некоторого шума в набор данных.

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

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

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

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

При этом важно, что мы отмечаем именно проблемы (изменения) в данных. Ничего не говорится о природе этих изменений. Это может быть и ошибка тренировки – подобранный набор данных сильно отличается от известной генеральной совокупности, это могут быть неверные заключения (предположения) в алгоритмах, неверный выбор и работа со свойствами (features), равно как и сознательно внесенные измерения в наборы исходных данных, ставящие целью, например, требуемое изменение работы системы. Как отмечается в [11], эффективное машинное обучение сложно потому, что сложно найти закономерности, и часто недостаточно данных для обучения. В результате программы машинного обучения часто не работают. В [12], например, рассматриваются ошибки в “стандартном” наборе данных ImageNet, который используется во множестве задач, связанных с распознаванием изображений. Приводимые там цифры говорят о том, что около 6% изображений размечены (классифицированы) неверно. Чувствительность классификаторов на базе ImageNet к таким ошибкам обсуждается в работе [13].

Например, целью учебного курса по основам ML в университете Беркли [14], заявлена именно надежность машинного обучения: “Как мы можем построить модели, устойчивые к сдвигу распределения (данные на этапе тренировки и эксплуатации имеют разные статистические распределения – то есть, разнятся), к противникам (к опровергающим примерам), к неправильной спецификации модели и к приближениям, налагаемым вычислительными ограничениями? Как правильно оценивать такие модели?”

Типичный пример сдвига распределения представлен на рисунке 4. Тренировочные и тестовые данные аппроксимируются разными прямыми

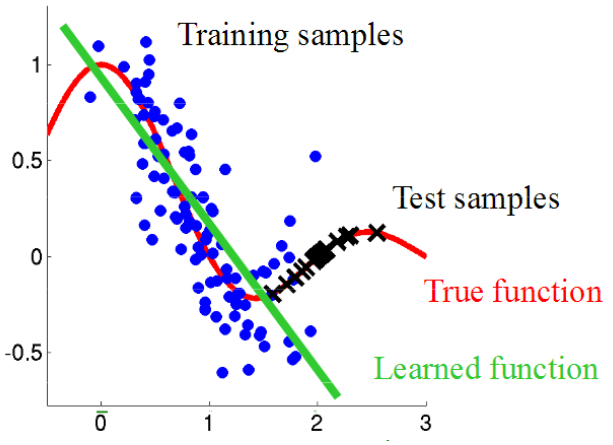

Рис. 4 Интерпретация тестовых и тренировочных данных [15]

А причиной этого различия являются разные распределения (рис. 5)

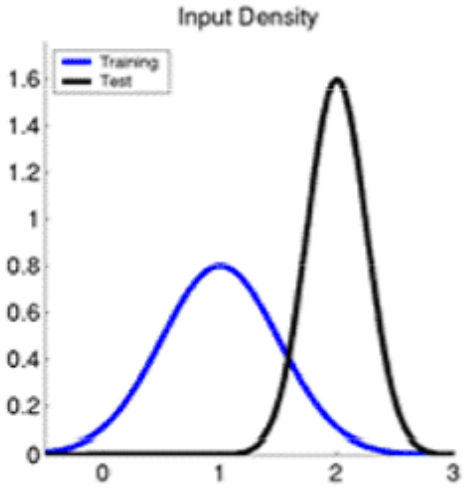

Рис. 5. Распределение данных [15]

Эти проблемы требуют переосмысления как теоретической, так и эмпирической парадигм машинного обучения. Например, теория принятия решений не учитывает случаи, когда функция вознаграждения является только приближенной. Между тем, измерения точности эмпирического теста на фиксированном распределении недостаточно для анализа таких явлений, как устойчивость к сдвигу распределения [16].

Google (Deepmind) в обзорной публикации своей исследовательской группы Robust and Verified Deep Learning group отмечает [17], что системы машинного

обучения по умолчанию не являются надежными. Даже системы, которые превосходят людей в определенной области, могут потерпеть неудачу в решении простых проблем, если будут внесены различия в исходные данные.

С точки зрения программиста, ошибка - это любое поведение, несовместимое со спецификацией, то есть предполагаемой функциональностью системы. В рамках работ DeepMind по решению задач интеллекта проводятся 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.

Основные задачи Robust and Verified Deep Learning group в Deepmind описываются так:

1. Эффективное тестирование соответствия спецификациям. Группа исследует эффективные способы проверки соответствия систем машинного обучения свойствам (таким как инвариантность или надежность), заданным 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. Обучение моделей машинного обучения в соответствии со спецификациями. Даже при большом количестве обучающих данных стандартные алгоритмы машинного обучения могут создавать прогностические модели, которые делают прогнозы несовместимыми с желательными спецификациями, такими как надежность или справедливость - это требует пересмотра алгоритмов обучения, которые создают модели, которые не только хорошо соответствуют данным обучения, но и соответствуют наборам спецификаций.

3. Формальное доказательство соответствия моделей машинного обучения спецификациям. Существует потребность в алгоритмах, которые могут проверить, что предсказания модели доказуемо согласуются с интересующей спецификацией для всех возможных входных данных. Хотя в области формальной проверки такие алгоритмы изучаются в течение нескольких десятилетий, эти подходы нелегко масштабировать до современных систем глубокого обучения с миллионами параметров, несмотря на впечатляющий прогресс.

Из других кратких характеристик области исследования, можно отметить презентацию Madry-lab (MIT) [18] и представленные в ней три заповеди Secure / Safe ML

I. Вы не должны тренироваться на данных, которым не полностью доверяете (из-за возможного отравления данных – изменения данных с целью обмана модели)

II. Вы не должны позволять никому использовать вашу модель (или наблюдать за ее работой), если вы полностью им не доверяете (из-за кражи модели и атак черного ящика). Это можно представить как аналогию декомпилирования или reverse engineering в программных системах – работа (поведение) модели изучается с целью построения состязательного примера.

III. Вы не должны полностью доверять предсказаниям вашей модели (из-за возможных состязательных примеров).

Первые два положение можно отнести к компьютерной безопасности. Последнее же не обязательно связано именно с атаками на систему безопасности. Как отмечается в [19], несмотря на их высокую точность вывода в практических приложениях, системы машинного обучения очень уязвимы для угроз безопасности и надежности, как в облаке, так и на периферии. Отравление обучающих данных (например, путем вставки случайного или искусственного шума в данные) с неверно помеченными входами, вставкой вредоносных компонентов в оборудование системы, загрязнением входов незаметным шумом во время вывода (т. е. во время работы системы), а также мониторинг побочных каналов системы для определения базовой модели - вот некоторые из способов, которыми злоумышленник может нарушить безопасность системы машинного обучения. Но даже при отсутствии явного злоумышленника, изменение процесса во время работы, ошибки памяти, условия окружающей среды вокруг системы во время обучения и вывода могут поставить под угрозу надежность системы машинного обучения.

Почему эти вопросы возникли сейчас (см. рисунок 3 с графиком роста числа публикаций)? Это, очевидно, связано с историей применения систем на базе машинного обучения. Вопросы устойчивости (надежности) просто не рассматривались в ранних применениях (как, впрочем, и в большинстве работ сейчас). В [20] описаны одни из ранних применений машинного обучения в Yahoo (Yahoo Research – один из пионеров в этой области). Какова цена ошибки при работе спам-фильтра или определении единичного профиля для рекламы? Очевидно, что об оценке одиночных ошибок речь вообще не идет. Во всех таких проектах есть некоторый приемлемый уровень точности, которая заведомо не достигает 100%. Аналогичная картина и с техническими проектами. Например, машинное обучение используется для оценки местоположения по результатам замеров параметров беспроводных сетей, заменяя более простые подходы [21, 22]. Используемые модели [23] будут заведомо неустойчивыми, поскольку измерения на практике зависят от клиентских устройств и текущего окружения, что практически всегда будет отличаться от тренировочных данных. Но опять-таки, цена ошибки здесь – это единицы (максимум – десятки) метров, что

несущественно для решаемых задач. Это может быть существенно, например, для каких-то роботехнических систем, но там и используются другие измерения.

Также необходимо отметить следующее. Использование программного обеспечения в критических областях (авионика, атомная промышленность и т.д.) уже давно сопровождается отработанным процессом сертификации (например, DO-178C и др.) [24]. Подобного рода решения, охватывающие не только тестирование программного обеспечения, но и процесс его разработки, необходимы и для систем машинного обучения [25].

В настоящее время исследования устойчивости (практически все) следуют из формулы (1): найти такие минимальные модификации данных, которые “обманывают” систему или, в обратную сторону, гарантировать работу системы при некоторых минимальных изменениях данных. Такая форма неявно предполагает присутствие человека в контуре принятия решения (минимальные изменения – незаметные человеку). Но для критических применений человек может отсутствовать и, следовательно, ограничений на изменения нет. Это означает, например, что если мы говорим о распознавании дорожных знаков (рис. 6),

Рис. 6. Дорожный знак

то необходимо говорить (доказывать) не только распознавание незначительно модифицированных изображений, но и, например, таких же изображений, но под другим углом (что уже проблема – рис.7), а также то, что нет других изображений, которые будут распознаны как такой знак.

Рис. 7 Изменение точки зрения радикально меняет распознавание [28]

Также необходимо иметь в виду, что за исключением методов формальной верификации [26], методы проверки устойчивости сегодня представляют собой, фактически, еще один вычислительный эксперимент. Например, так называемое состязательное обучение – это просто использование известных состязательных примеров в тренировочных данных [27]. Это позволит распознавать такие состязательные примеры, но вряд ли может считаться доказательством устойчивости.

V Заключение

На сегодняшний день использование систем машинного обучения в критических приложениях невозможно без решения вопросов об устойчивости используемых моделей. Также на сегодняшний день нет универсальных решений в данной области. Более того, не вполне ясно, удастся ли получить такие решения. Наиболее полно по понятие доказательства устойчивости подходят системы формальной верификации для систем машинного обучения, но их текущее применение ограничивается проблемами с масштабируемостью. Как результат можно предположить появление моделей машинного обучения включающих непосредственную поддержку обработки сдвига распределения.

Благодарности

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

Библиография

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

[2] Namiot D., Ilyushin E., Chizhov I. Ongoing academic and industrial projects dedicated to robust machine learning //International Journal of Open Information Technologies. – 2021. – Т. 9. – №. 10. – С. 35-46.

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

[4] 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

[5] 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.

[6] 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.

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

[8] Xu, H., Mannor, S.: Robustness and generalization. In: COLT, pp. 503–515 (2010)

[9] "A Model-Based Approach for Robustness Testing" (PDF). Dl.ifip.org. Retrieved 2016-11-13.

[10] IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990

[11] Tyler Folkman Machine learning: introduction, monumental failure, and hope https://towardsdatascience.com/machine-learning-introduction-monumental-failure-and-hope-65a8c6098a92

[12] Foundations for AI errors https://www.wired.com/story/foundations-ai-riddled-errors/ Retrieved: Sep, 2021

[13] Tsipras, Dimitris, et al. "From imagenet to image classification: Contextualizing progress on benchmarks." International Conference on Machine Learning. PMLR, 2020.
[14] Stat 260 https://www.stat.berkeley.edu/~jsteinhardt/stat260/index.html Retrieved: Sep, 2021
[15] Francisco Herrera Dataset Shift in Classification: Approaches and Problems http://iwann.ugr.es/2011/pdf/InvitedTalk-FHerrera-IWANN11.pdf Retrieved: Sep, 2021
[16] Jacob Steinhardt https://www.stat.berkeley.edu/~jsteinhardt/index.html Retrieved: Sep, 2021
[17] DeepMind Robust and Verified Deep Learning group https://deepmind.com/blog/article/robust-and-verified-ai Retrieved: Sep, 2021
[18] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: Sep, 2021
[19] Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." IEEE Design & Test 37.2 (2020): 30-57.
[20] Yahoo Research AI https://www.financialexpress.com/archive/yahoo-research-uses-artificial-intelligence-everywhere/191870/ Retrieved: Sep, 2021
[21] Namiot, Dmitry. "Context-Aware Browsing--A Practical Approach." 2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies. IEEE, 2012.
[22] Namiot, Dmitry, and Manfred Sneps-Sneppe. "Proximity as a service." 2012 2nd Baltic Congress on Future Internet Communications. IEEE, 2012.
[23] Rojo, Jordi, et al. "Machine learning applied to wi-fi fingerprinting: The experiences of the ubiqum challenge." 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE, 2019.
[24] Souyris, Jean, et al. "Formal verification of avionics software products." International symposium on formal methods. Springer, Berlin, Heidelberg, 2009.
[25] DO-178C cert-kit for airborne machine learning to be researched by Intelligent Artifacts https://militaryembedded.com/avionics/software/do-178c-cert-kit-for-airborne-machine-learning-to-be-researched-by-intelligent-artifacts Retrieved: Sep, 2021
[26] Seshia, Sanjit A., et al. "Formal specification for deep neural networks." International Symposium on Automated Technology for Verification and Analysis. Springer, Cham, 2018.
[27] Li, Guofu, et al. "Security matters: A survey on adversarial machine learning." arXiv preprint arXiv:1810.07339 (2018).
[28] Ego4D: Around the World in 3,000 Hours of Egocentric Video https://ai.facebook.com/research/publications/ego4d-unscripted-first-person-video-from-around-the-world-and-a-benchmark-suite-for-egocentric-perception Retrieved: Sep, 2021

The rationale for working on 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, attention to the issues of reliability (stability) of such systems and solutions is also growing. For so-called critical applications such as real-time decision-making systems, special systems, etc. sustainability issues are crucial from the point of view of the practical use of machine learning systems. The use of machine learning systems (artificial intelligence systems, which is now, in fact, a synonym) in such areas is possible only with the proof of stability (determination of guaranteed performance parameters). Resiliency problems arise from different characteristics of the data during training (training) and testing (practical application). At the same time, additional complexity is created by the fact that, in addition to natural reasons (unbalanced samples, measurement errors, etc.), the data can be deliberately modified. These are the so-called attacks on machine learning systems. Accordingly, it is impossible to talk about the reliability of machine learning systems without protection against such actions. 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] Artificial Intelligence in Cybersecurity. http://master.cmc.msu.ru/?q=ru/node/3496 (in Russian) Retrieved: Sep, 2021

[2] Namiot D., Ilyushin E., Chizhov I. Ongoing academic and industrial projects dedicated to robust machine learning //International Journal of Open Information Technologies. – 2021. – Т. 9. – №. 10. – С. 35-46.

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

[4] 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

[5] 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.

[6] 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.

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

[8] Xu, H., Mannor, S.: Robustness and generalization. In: COLT, pp. 503–515 (2010)

[9]"A Model-Based Approach for Robustness Testing" (PDF). Dl.ifip.org. Retrieved 2016-11-13.

[10] IEEE Standard Glossary of Software Engineering Terminology, IEEE Std 610.12-1990

[11] Tyler Folkman Machine learning: introduction, monumental failure, and hope https://towardsdatascience.com/machine-learning-introduction-monumental-failure-and-hope-65a8c6098a92

[12] Foundations for AI errors https://www.wired.com/story/foundations-ai-riddled-errors/ Retrieved: Sep, 2021

[13] Tsipras, Dimitris, et al. "From imagenet to image classification: Contextualizing progress on benchmarks." International Conference on Machine Learning. PMLR, 2020.

[14] Stat 260 https://www.stat.berkeley.edu/~jsteinhardt/stat260/index.html Retrieved: Sep, 2021

[15] Francisco Herrera Dataset Shift in Classification: Approaches and Problems http://iwann.ugr.es/2011/pdf/InvitedTalk-FHerrera-IWANN11.pdf Retrieved: Sep, 2021

[16] Jacob Steinhardt https://www.stat.berkeley.edu/~jsteinhardt/index.html Retrieved: Sep, 2021

[17] DeepMind Robust and Verified Deep Learning group https://deepmind.com/blog/article/robust-and-verified-ai Retrieved: Sep, 2021

[18] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: Sep, 2021

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

[20] Yahoo Research AI https://www.financialexpress.com/archive/yahoo-research-uses-artificial-intelligence-everywhere/191870/ Retrieved: Sep, 2021

[21] Namiot, Dmitry. "Context-Aware Browsing--A Practical Approach." 2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies. IEEE, 2012.

[22] Namiot, Dmitry, and Manfred Sneps-Sneppe. "Proximity as a service." 2012 2nd Baltic Congress on Future Internet Communications. IEEE, 2012.

[23] Rojo, Jordi, et al. "Machine learning applied to wi-fi fingerprinting: The experiences of the ubiqum challenge." 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE, 2019.

[24] Souyris, Jean, et al. "Formal verification of avionics software products." International symposium on formal methods. Springer, Berlin, Heidelberg, 2009.

[25] DO-178C cert-kit for airborne machine learning to be researched by Intelligent Artifacts https://militaryembedded.com/avionics/software/do-178c-cert-kit-for-airborne-machine-learning-to-be-researched-by-intelligent-artifacts Retrieved: Sep, 2021

[26] Seshia, Sanjit A., et al. "Formal specification for deep neural networks." International Symposium on Automated Technology for Verification and Analysis. Springer, Cham, 2018.

[27] Li, Guofu, et al. "Security matters: A survey on adversarial machine learning." arXiv preprint arXiv:1810.07339 (2018).

[28] Ego4D: Around the World in 3,000 Hours of Egocentric Video https://ai.facebook.com/research/publications/ego4d-unscripted-first-person-video-from-around-the-world-and-a-benchmark-suite-for-egocentric-perception Retrieved: Sep, 2021