Об улучшении робастности моделей машинного обучения

Д.Е. Намиот, В.Ю. Романов

***Аннотация*—В данной статье рассматриваются вопросы улучшения робастности для моделей машинного обучения. Робастность является одной из важнейших характеристик моделей машинного обучения, которая определяет возможность практического использования моделей. Однако с определением этой характеристики для конкретных моделей не все оказывается просто. Во-первых, не все однозначно с самим определением робастности. Если мы рассматриваем робастность как сохранение поведения модели при малых возмущениях исходных данных, то возникают, по крайней мере, два вопроса – насколько малы должны быть эти изменения, и как такое определение соотносится с другими характеристиками модели? В первую очередь, среди других характеристик нужно рассмотреть обобщающую способность модели (генерализацию), которая определяется по работе модели с ранее неизвестными данными. Рассматривается также понятие надежности моделей, предложенное Google. Основное содержание статьи посвящено рассмотрению состязательных тренировок, которые, при всей своей ограниченности, остаются на сегодняшний день основным инструментом повышения робастности.**

***Ключевые слова*—машинное обучение, робастность, обобщение.**

I. Введение

Глоссарий IEEE определяет робастность как "The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions" [1]. Иными словами – сохранение корректного поведения (правильной работы) при возмущении исходных данных.

Другие определения (формулировки) следуют схожим идеям. Под робастностью в статистике понимают нечувствительность к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами [2].

Робастность для системы управления означает малое изменение выхода замкнутой системы управления при малом изменении параметров объекта управления. Это иногда формулируется также как устойчивость к помехам [3].

Статья получена 9 января 2024.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
В.Ю. Романов – МГУ имени М.В. Ломоносова (email: vladimir.romanov@gmail.com)

Алгоритм, в котором погрешность, допущенная в начальных данных или допускаемая при вычислениях, с каждым шагом не увеличивается или увеличивается незначительно, называется робастным [4]. В работе [5] алгоритмическая робастность определяется так: робастный алгоритм демонстрирует “схожую” производительность на данных, которые “близки”. “Схожесть” и “близость” должны определяться конкретными метриками в рамках решаемой задачи.

Очевиден интерес к робастности именно для моделей машинного обучения. Независимо от модели, они обучаются на конкретном наборе данных, а используются уже с другими данными. Будет ли наша модель показывать ту же самую производительность, которая была на этапе тренировки/тестирования? Данные на этапе использования (вывода) могут, в общем случае, отличаться от тренировочных [6].

Робастность очевидным образом связана с генерализацией - возможностью модели выработать обобщения на основе тренировочных данных. Обобщение — это способность обученной модели МО точно прогнозировать на примерах, которые не использовались для обучения [7]. Хорошая производительность обобщения является ключевой целью любого практического алгоритма обучения. При обучении, мы хотим подогнать модель к обучающему набору, в котором данные обучения и тестирования берутся из одного и того же распределения, которое, по нашему представлению, хорошо описывает генеральную совокупность. Если обучающие данные смещены в том смысле, что некоторые части совокупности недостаточно представлены или отсутствуют в обучающих данных, то подобранная модель на этих обучающих данных будет смещена в сторону от оптимальной функции [8].

Google в работе [9] использует термин робастная генерализация (рис. 1). Под этим понимается сохранение производительности модели на новых данных с тем же распределением, что и тренировочные (это и есть in-distribution performance), на данных с ковариантным сдвигом и данных со сдвигом субпопуляции.

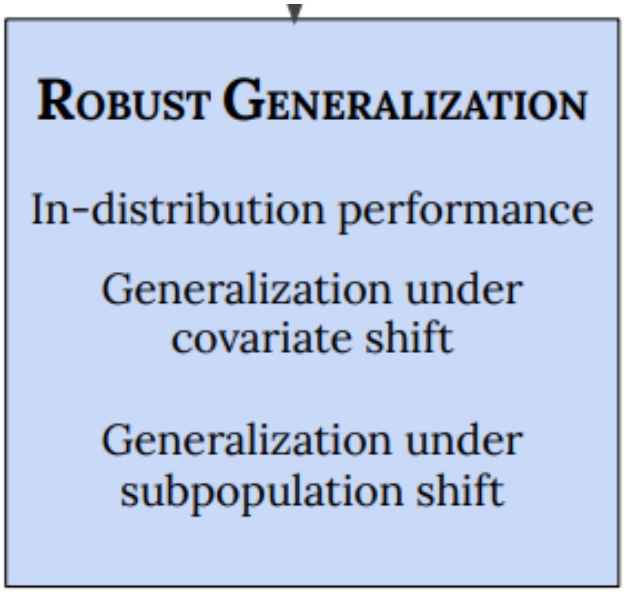

Рис. 1. Робастная генерализация [9]

Напомним, что под ковариантным сдвигом (в некоторых работах используется термин ковариационный сдвиг) понимается ситуация, когда изменяется распределение входных значений, а условное распределение выходных значений остается прежним. Сдвиг субпопуляции описывает сценарии, в которых интересующее распределение является лишь частью полного распределения. Тренировочные данные отбираются из распределения отдельных субпопуляций, которые сами по себе выбираются из распределения метапопуляций. Классический пример: мы обучаем модель классифицировать собак по их изображениям, а тренировочный набор включает изображения собак только одной породы.

В некоторых (во многих, на самом деле) работах используется такой термин как состязательная робастность [10]. Состязательные атаки (атаки уклонения) определяются, исторически, как минимальные модификации входных данных, которые изменяют решение системы. Классический пример, положивший начало состязательным атакам, представлен на рис. 2:

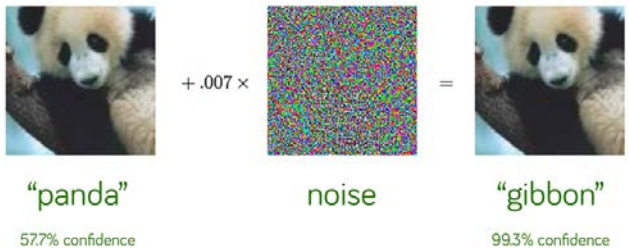

Рис. 2. Состязательные атаки [11]

Соответственно, состязательная робастность – это сохранение работы модели при состязательных возмущениях исходных данных (иными словами – это устойчивость к атакам уклонения [12]).

Формально, мерой состязательной робастности для модели $f(\cdot)$ может быть, например, согласно [10], наихудшее значение функции потерь L при заданном бюджете возмущений ε и метрике D

$$\mathop{\mathrm{E}}_{(x,y)\sim X}\left[\max_{x':D(x,x')<\varepsilon} L(f(x'),y)\right] \quad (1)$$

Другое общепринятое определение — это среднее (или медианное) минимальное расстояние до враждебного возмущения

$$\mathop{\mathrm{E}}_{(x,y)\sim X}\left[\min_{x'\in A_{x,y}} D(x,x')\right] \quad (2)$$

Где определение $A_{x,y}$ зависит от атаки. Для нецелевого нарушения классификации $A_{x,y}=\{x' \mid f(x')\neq y\}$ а для целевой атаки на класс t $A_{x,y}=\{x' \mid f(x')=t\}$

Формально, классическое определение робастности, например, классификатора определяется следующим образом. При заданных входных данных x и интересующей модели f мы хотим, чтобы предсказание модели оставалось одинаковым для всех входных данных x' в окрестности x, где окрестности определяются некоторой функцией расстояния δ и некоторым максимальным расстоянием Δ:

$$\forall x'.\ \delta(x,x')\leq\Delta \Rightarrow f(x)=f(x') \quad (3)$$

В качестве расстояния обычно используется мера ℓp. Из этой формулы, в частности, следует то, что отождествлять робастность и надежность не корректно. Надежность всегда связана с работоспособностью. По ГОСТу – это свойство объекта выполнять и сохранять во времени заданные функции в заданных режимах и условиях применения, технического обслуживания, ремонтов, хранения и транспортировки. Но в определении (3) ничего не говорится о работоспособности модели (правильности классификатора). Классификатор может работать неверно и сохранять это неверное значение при заданных возмущениях входных данных.

Надежность (reliability) для моделей машинного обучения согласно работе Google [3] – это сочетание трех характеристик: (1) модели должны точно сообщать о неопределенности своих прогнозов («знать то, чего они не знают»); (2) они должны быть надежно обобщены на новые сценарии (сдвиг распределения); и (3) они должны быть способны эффективно адаптироваться к новым данным (адаптация). Важно отметить, что надежная модель должна быть нацелена на успешную работу во всех этих областях одновременно, без необходимости какой-либо настройки для отдельных задач (рис.3).

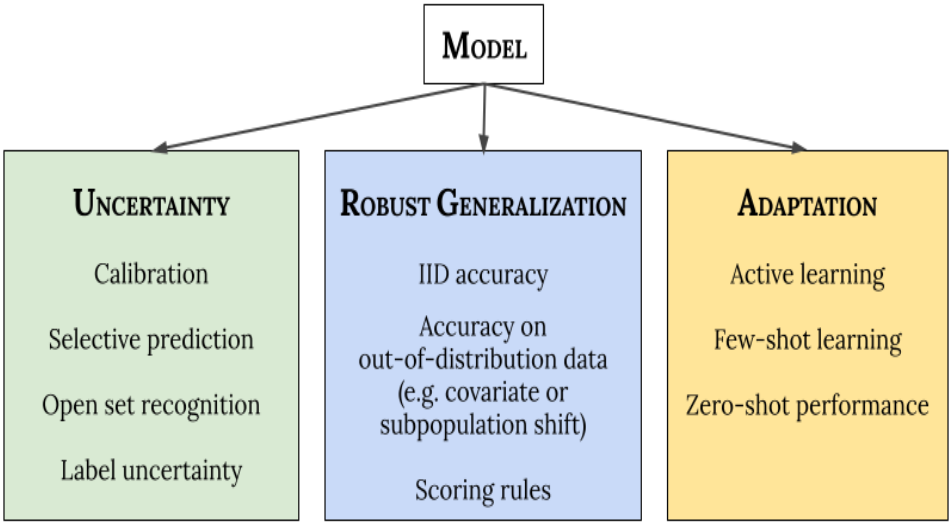

Рис. 3. Надежность моделей ML [9]

Понятие безопасности тесно связано с надежностью, поскольку дополнительно к работоспособности подразумевает еще и отсутствие ущерба от этой работы. Надежность и безопасность, с практической точки зрения, должны быть характеристиками не модели, а сущности более высокого уровня – системы, которая использует эту модель. Например, на этапе вывода модель ведь должна исполняться к некоторой вычислительной среды и рассмотрение надежности только модели – недостаточно.

В работе [14] авторы определили статистические характеристики робастности, которые можно использовать на практике. Согласно этим определениям:

Пусть есть классификатор $f{:}X \to L$, где $X \subseteq R^n$ есть входное пространство и $L = \{1, ..., L\}$ – метки. Мы предполагаем, что тренировочные и тестовые данные $x \in X$ имеют распределение D. Вводимые статистики зависят от параметра ε (бюджет модификаций), который соответствует идее о том, что рассматриваются состязательные примеры x, расположенные не дальше заданного расстояния (бюджета) в метрике $L\infty$ от рассматриваемого значения x_*.

Интуитивно, f is robust at $x_* \in X$, если “малые” изменения x_* не изменяют решение классификатора. “Малость” (незаметность) изменений описывается следующим условием: $\|x - x_*\|_{L\infty} \le \varepsilon$

Формально, мы будем говорить, что классификатор f (x_*, ε)-робастный, если для каждого x, такого что $\|x - x_*\|_{L\infty} \le \varepsilon$, $f(x) = f(x_*)$. А точечная робастность $\rho(f, x_*)$ классификатора f в точке x_* есть минимальное значение ε, при котором f перестает быть (x_*, ε)-робастным:

$$\rho(f, x_*) \overset{def}{=} \inf \{ \varepsilon \ge 0 \mid f \text{ не } (x_*, \varepsilon) \text{ робастен} \} \quad (4)$$

Тогда состязательная частота определяется для заданного параметра ε следующим образом.

$$\varphi(f, \varepsilon) \overset{def}{=} \Pr_{x^* \sim D}[\rho(f, x_*) \le \varepsilon] \quad (5)$$

и измеряет, как часто f не является (x_*, ε)-робастным. Другими словами, высокая состязательная частота свидетельствует о том, что классификатор не является (x_*, ε)-робастным для многих входных значений x_*.

Состязательная строгость для заданного параметра ε определяется следующим образом

$$\mu(f, \varepsilon) \overset{def}{=} E_{x^* \sim D}[\rho(f, x_*) \mid \rho(f, x_*) \le \varepsilon] \quad (6)$$

и оценивает возможность того, что f не является робастным в точке x_*, при условии, что f не является (x_*, ε)-робастным.

Меньшее значение $\mu(f, \varepsilon)$ соответствует худшей строгости состязательности, поскольку f более восприимчив к состязательным примерам, если расстояние до ближайшего состязательного примера невелико.

Частота и строгость отражают различное поведение устойчивости. Нейронная сеть может иметь высокую частоту состязательности, но низкую строгость состязательности, что указывает на то, что большинство состязательных примеров находятся на некотором расстоянии от исходной точки x_*. И наоборот, нейронная сеть может иметь низкую частоту состязательности, но высокую состязательную строгость, что указывает на то, что она обычно робастна, но иногда ее робастность резко снижается. Частота обычно является более важным показателем, поскольку нейронная сеть с низкой состязательной частотой большую часть времени робастна.

Строгость можно использовать для различения нейронных сетей с одинаковой состязательной частотой.

Отдельно необходимо отметить вопрос о размерах возмущения исходных данных во всех этих определениях. В большом количестве работ изучается модель угроз, в которой противник ограничен возмущениями, определяемыми через ℓp. Еще в работе Carlini [10] отмечается, что эта модель угроз весьма

ограничена и не полностью соответствует реальным угрозам. Но им же отмечается, что четко определенный характер этой модели угроз полезен для выполнения принципиальной работы по построению сильной защиты. И хотя ℓp-устойчивость не подразумевает устойчивость в более реалистичных моделях угроз, почти наверняка отсутствие устойчивости к возмущениям, ограниченным ℓp, будет означать отсутствие устойчивости в более реалистичных моделях угроз. Таким образом, работа над решением проблемы устойчивости этих четко определенных моделей угроз, ограниченных ℓp, является полезным упражнением. С этим можно согласиться, но ведь главное в этих утверждениях как раз то, что ℓp-устойчивость не гарантирует устойчивость в более реалистичных моделях угроз. То есть, вообще говоря, является достаточно бесполезной с практической точки зрения. ℓp-устойчивость позволяет использовать формальный аппарат для оценки робастности. Это важно, например, для оценочных подходов на основе константы Липшица [13]. Но практический эффект минимальных изменений весьма низок. Идея о том, что состязательные атаки должны быть незаметны для человека, не учитывает тот факт, что наибольшую проблему такие атаки представляют для критических приложений (авионика, автоматическое вождение и т.п.). Но именно в этих приложениях человек не участвует в принятии решения. Соответственно, размер модификации входных данных не имеет значения. Наоборот, для такого рода систем необходимо доказывать работоспособность для всех возможных модификаций входных данных.

В целом, очевидно, генерализация является более общим понятием, чем робастность и, что самое главное, более ориентированным на практику. Но, как отмечается во многих работах, эти два понятия часто связаны. Хорошая генерализация повышает восприимчивость к состязательным примерам и наоборот, можно добиться хорошей устойчивости к состязательным примерам за счет плохого обобщения [15]. В этом плане представляет интерес описанная выше робастная генерализация от Google.

А в остальном, остается только согласиться с утверждением работы [16], где говорится о многозначности понятия робастности.

Оставшаяся часть статьи структурирована следующим образом. В разделе II обсуждаются состязательные тренировки. В разделе III обсуждаются расширения и модификации стандартной процедуры состязательных тренировок. Раздел IV посвящен работе с исходными данными. В разделе V рассматриваются вопросы сертификации и аудита для моделей машинного обучения.

II Состязательные тренировки

Идея состязательных тренировок абсолютно прозрачна и полностью укладывается в базовую парадигму машинного обучения. Проще всего пояснить это на примере классификации изображений. Предположим, у нас есть классификатор, который был натренирован на некотором наборе изображений. При использовании построенной модели на практике мы сталкиваемся с изображениями, которые, кажется, совсем немного отличаются от тех, которые присутствовали в тренировочном наборе, но классификатор выдает неправильные метки. Или это те же самые объекты из тренировочного набора, но, например, снятые под другим ракурсом, при другом освещении, в другую погоду и т.п., что также классифицируется неверно.

Очевидно, что при всех вопросах к генерализации, модель будет лучше работать (показывать более высокие метрики) на тех данных, на которых она тренировалась. Вот идея состязательных тренировок именно это и использует – добавляет модифицированные данные в тренировочный набор с правильными метками.

Типичный пример представлен в работе [17]. Рассматривалась модель классификации автомобилей, работающая с точностью >98%. Модель не распознавала камуфлированные изображения (на рис. 4 показано использование в качестве камуфляжа рисунка другого автомобиля).

Рис.4. Камуфляжный рисунок (аэрография) [17]

После добавления камуфлированных изображений в тренировочный набор и переобучения системы, точность вернулась к исходным значениям.

Вопросы, которые возникают к этому подходу, абсолютно очевидны. Например:

- Какие модификации необходимо добавить к тренировочному набору?
- Можем ли мы знать обо всех возможных модификациях?
- Добавление искаженных изображений, очевидно, понижает точность работы?

Естественно, что все возможные модификации нам неизвестны. И да, добавление искаженных данных влияет на точность. Более того, добавляя искаженные данные к тренировочному набору, мы можем даже отравить модель. Тем не менее, состязательные

тренировки являются самым простым и наиболее часто используемым подходом к увеличению робастности моделей машинного обучения.

Что может использоваться при дополнении данных?

Мы можем использовать некоторые предопределенные модификации (как в атаках черного ящика), какие-то стандартные преобразования (например, для изображений – повороты) и, наконец, какие-то естественные (семантически определенные) преобразования.

На рисунке 5 изображен фрагмент документации пакета Foolbox [18], который описывает модификации изображений для выполнения состязательных атак

L2AdditiveGaussianNoiseAttack	Samples Gaussian noise with a fixed L2 size.
L2AdditiveUniformNoiseAttack	Samples uniform noise with a fixed L2 size.
L2ClippingAwareAdditiveGaussianNoiseAttack	Samples Gaussian noise with a fixed L2 size after clipp
L2ClippingAwareAdditiveUniformNoiseAttack	Samples uniform noise with a fixed L2 size after clippi
LinfAdditiveUniformNoiseAttack	Samples uniform noise with a fixed L-infinity size
L2RepeatedAdditiveGaussianNoiseAttack	Repeatedly samples Gaussian noise with a fixed L2 siz
L2RepeatedAdditiveUniformNoiseAttack	Repeatedly samples uniform noise with a fixed L2 size
L2ClippingAwareRepeatedAdditiveGaussianNoiseAttack	Repeatedly samples Gaussian noise with a fixed L2 siz
L2ClippingAwareRepeatedAdditiveUniformNoiseAttack	Repeatedly samples uniform noise with a fixed L2 size
LinfRepeatedAdditiveUniformNoiseAttack	Repeatedly samples uniform noise with a fixed L-infini
InversionAttack	Creates "negative images" by inverting the pixel value:

Рис.5 Foolbox атаки [18]

Это различные методы добавления шума, черных и белых точек (Salt&Pepper атака) и т.п. Такого рода аугментацию тренировочных данных легко осуществить, но нужно помнить об одном факте. Наибольший интерес проблема с робастностью моделей вызывает именно для критических применений, где возможные состязательные атаки будут осуществляться в реальном мире (физические атаки по классификации [12]). Конечно, такого рода преобразования имеют очень малое отношение к возможным реальным изменениям для, например, изображений. С другой стороны, простота таких подходов и развитые методы формирования атак уклонений [19], обеспечивают их широкое применение.

Стандартные преобразования изображений, такие как повороты или изменение угла зрения уже могут быть реализованы в реальном мире. Естественно, что можно говорить и об одновременно используемых модификациях данных [20].

В работе [21] приводится следующая классификация методов добавления данных для состязательных тренировок – рис. 6.

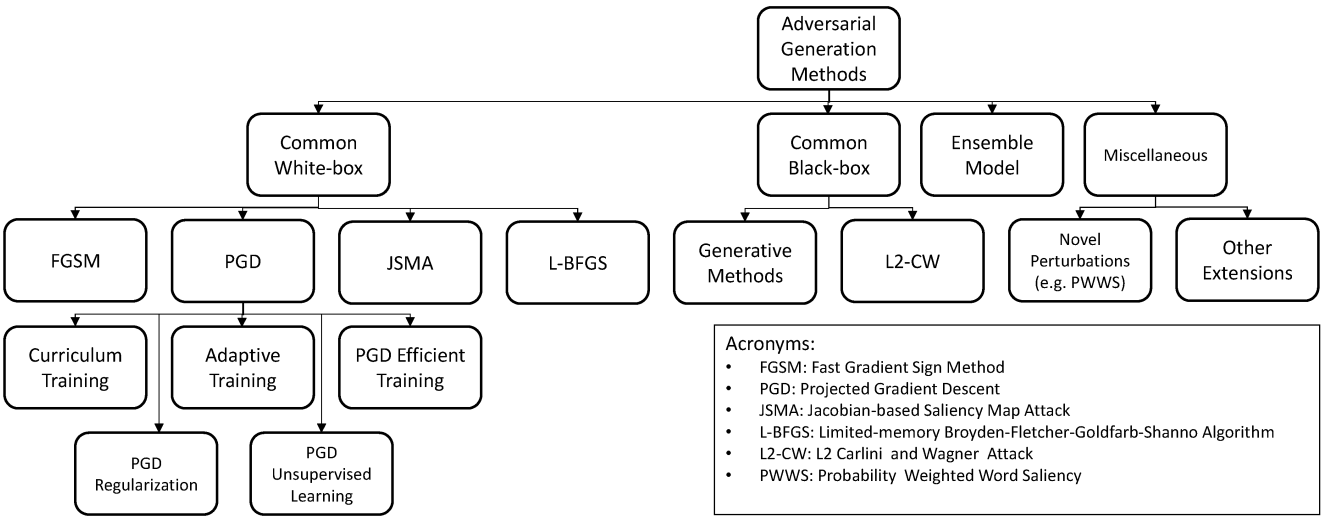

Рис.6. Порождение данных для состязательных тренировок [21]

Состязательные тренировки могут быть представлены как оптимизационная задача, в которой мы стремимся минимизировать максимальные потери модели

$$\min \sum_i \max L(f(x_i + \delta), y_i) \quad (7)$$

Внутренняя часть формулы предназначена для максимизации потерь L модели f относительно выходной метки y путем добавления возмущения δ во

входные данные *x*. В общем, внутренние функции аппроксимации строятся с использованием различных методов атаки. Это приближение важно для определения верхнего предела состязательной оптимизации. Как правило, метод состязательного обучения с аналогичной функцией внутренней аппроксимации будет иметь схожие преимущества и ограничения, которые и описываются для атак [19]. В целом, состязательные атаки можно разделить на две более обширные категории: атаки «белого ящика» и «черного ящика» [12], что и показано на рисунке 6. Методы состязательной генерации «белого ящика» или «черного ящика» обычно имеют некоторые общие свойства [22].

Гораздо больший интерес, на наш взгляд, представляют собой семантически обусловленные (часто их еще называют естественные) методы аугментации данных для состязательных тренировок. Если мы говорим об изображениях (а это наиболее используемая и описываемая область для обсуждения состязательных атак на модели машинного обучения), то такие дополнения включают, например, имитацию различных погодных условий (рис.7) или проблем с камерой (рис. 8, 9).

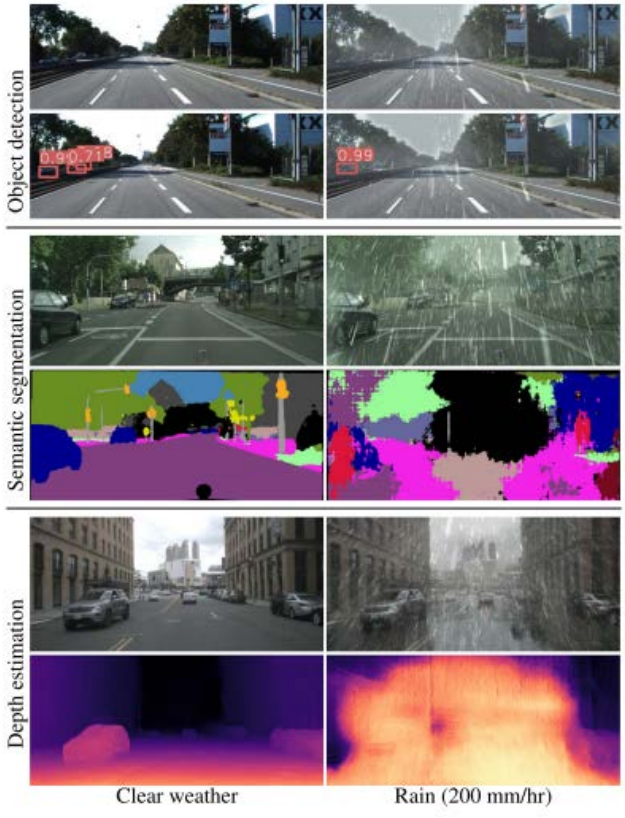

Рис.7 Дождь и распознавание объектов [23]

Подобного рода преобразования могут использоваться при сертификации промышленных систем видеонаблюдения.

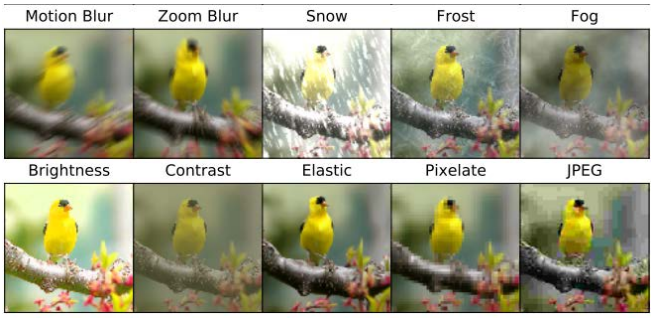

Рис.8 Проблемы с камерой и погодные условия [24]

Рис. 9 Проблемы с камерой [25]

III Модификации состязательных тренировок

Обучение признакам или обучение представлениям (feature learning или representation learning в англоязычной литературе) - это набор техник, которые позволяют системе автоматически обнаружить представления, необходимые для выявления признаков или классификации исходных данных.

Обучение представлениям, как описано, например, в работах [27, 28] позволило точно контролировать атрибуты (признаки) изображения путем использования и манипулирования низкоразмерными кодировками представления (скрытыми кодировками), что облегчает непрерывное и интерпретируемое редактирование атрибутов (признаков) в изображениях. Каждое измерение скрытых кодировок соответствовало определенному типу вариаций признаков [29]. Манипулирование одним измерением при сохранении фиксированных других размеров позволило целенаправленно изменять признаки, сводя к минимуму влияние на другие атрибуты (признаки) изображения. Это позволяет обеспечить плавные изменения конкретных атрибутов изображения. Похожим образом строятся, например, некоторые атаки отравления данных. Проводятся модификации в пространстве признаков, которые внешне могут быть незаметны. В этом плане интересна работа [30], где такие преобразования описываются как состязательные примеры с легитимной семантикой. В работе как раз описывается новый подход к созданию многообразных состязательных примеров с легитимной семантикой (MAELS).

Из других подходов к состязательным тренировкам

можно отметить так называемый куррикулум. В этом подходе [31] используются последовательно слабые и сильные атаки. Метод оценивает, может ли модель достичь достаточной точности при более слабых настройках силы атаки, а затем постепенно увеличивает силу после достижения высокой точности.

Порождающие модели – еще один подход к построению состязательных примеров [32]. Хорошие обзоры различных методов построения примеров для состязательных тренировок есть в работах [21, 33].

Вместе с тем во многих работах отмечается, что, несмотря на свою значимость, проблема обобщения состязательных тренировок к неизвестным атакам в настоящее время изучается лишь изредка. Одна из возможных причин заключается в том, что наше понимание состязательных примеров ограничено и неполно.

IV Работа с исходными данными

В работе [26], авторы оценивали влияние на робастность моделей очистки тренировочных данных. Использованные методы для автоматической очистки данных представлены ниже на рисунке

Error Type	Detection Method	Repair Method
Missing Values	Empty Entries	Deletion
		Mean_Mode, Mean_Dummy Median_Mode, Median_Dummy Mode_Mode, Mode_Dummy
		HoloClean
Outliers	SD	Mean, Median, Mode
	IQR	
	IF	HoloClean
Duplicates	Key Collision	Deletion
	ZeroER	
Inconsistencies	OpenRefine	Merge
Mislabels	cleanlab	cleanlab

Рис. 10. Очистка данных [26]

Наибольший эффект в плане повышения робастности моделей был от восстановления пропущенных значений.

V Сертификация и аудит

Сертификация – это процедура, которая призвана гарантировать достижение заданных показателей (метрик). Соответственно, если мы говорим о моделях, то это гарантии для метрик модели. Аудит – это процедура, которая оценивает выполнение некоторых предписанных действий или соблюдение заданных условий.

Говоря о сертификации моделей машинного обучения, необходимо учитывать следующее. Модель машинного обучения на этапе вывода (inference) – эта программа. Программное обеспечение для критических применений (например, в авионике) сертифицируется. Процедура сертификации для программного обеспечения, безусловно, включает в себя функциональное тестирование, которое должно подтвердить работоспособность системы для всех допустимых входных данных [34]. Вот, по крайней мере, такой момент для моделей машинного обучения и не выполняется. И наличие состязательных атак препятствует полному функциональному тестированию, и сами модели тестируются на конечных датасетах, а вовсе не на всех допустимых входных данных. Поэтому одно и то же слово “сертификация” имеет разный смысл для моделей машинного обучения и для программ, которые эти модели реализуют [35].

Настоящей сертификацией для моделей машинного обучения могла бы быть формальная верификация моделей машинного обучения, которая позволяла бы, например, точно устанавливать границы значений выходных параметров нейронной сети. Но реальные достижения в этой области еще далеки от повсеместного практического применения [36].

Сертифицированная же робастность - это установление теоретически подтвержденной нижней границы устойчивости при определенных ограничениях на возмущения. Именно последнее условие отличает сертификацию моделей от классической сертификации. Робастность определяется и сертифицируется при определенных ограничениях на возмущения. На рисунке 11 представлены типичные результаты сертифицированной робастности для классификатора изображений. Показана стандартная точность на выбранных датасетах и гарантированная точность, при заданных ограничениях на изменения.

Dataset	Train. Alg.	Test set Accuracy %		Certified Robust %	
		Standard	Generative	$\delta = 0.05$	$\delta = 0.1$
MNIST	ERM	97.9	71.6	73.2	62.4
	IRM	97.8	78.7	91.4	37.0
	PGD	97.0	79.5	91.0	73.8
	MDA	97.2	96.5	97.2	86.6

Рис.11. Сертифицированная робастность [37].

Аудит моделей машинного обучения является практической моделью оценки качества моделей машинного обучения. Идея аудита достаточно прозрачна. Аудит – это некоторый чеклист, который состоит из вопросов о характеристиках разработанной модели. Примеры вопросов [35]:

- Ведутся ли логи в процессе работы модели ?
- Есть возможность определить сдвиг распределения в реальных данных ?
- и т.д.

Естественно, что поддержка логов никак не влияет ни на одну метрику модели. Но - отсутствие логов в процессе эксплуатации не позволит расследовать (определить) состязательные атаки. То есть их отсутствие повышает риски использования такой модели.

Представьте, что каждый ответ “нет” на подобный вопрос оценивается в какое-то количество баллов (в разных предметных областях вопросы могут “стоить” по-разному). В итоге, у нас получится цифровая оценка рисков модели. Для чего это нужно? Во-первых, это позволит сравнивать между собой разные реализации. Во-вторых, это показывает пути улучшения качества модели (за что именно были получены штрафные баллы). Эти вопросы, по сути, описывают лучшие

практики при разработке моделей машинного обучения. В-третьих, эти вопросы определяют необходимые инструменты (компоненты) для доверенных платформ машинного обучения [38].

VI Заключение

В настоящей статье рассмотрены вопросы повышения робастности моделей машинного обучения. Естественно, что первый вопрос в такого рода исследовании – это вопрос о том, что именно понимается под робастностью и, самое главное, как это измеряется. Здесь, на самом деле, нет единого мнения. Это связано, как нам кажется, с тем, что реальным желаемым измерением является надежность работы системы машинного обучения. А это не совпадает с классическим определением робастности. Отмечается также, что хотя точность модели на IID (Independent and identically distributed) данных является сильным предиктором точности для OOD (out of distribution) данных, она не является решающей. И вопрос об улучшении (или гарантированном сохранении) метрик модели на OOD данных остается открытым. А это, собственно, и есть вопрос, который определяет надежность работы модели. Основным достоинством состязательных тренировок является то, что они предлагают некоторый алгоритмический путь, который, в принципе, может повысить робастность моделей путем увеличения по количеству и разнообразию тренировочного набора данных.

Благодарности

Авторы благодарны сотрудникам лаборатории Открытых информационных технологий кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова. Статья продолжает серию публикаций, написанных для поддержки магистерской программы Искусственный интеллект в кибербезопасности [39].

Библиография

[1] IEEE Std. 610.12 (IEEE Standard Glossary of Software Engineering Terminology).

[2] Робастность https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C Retrieved: Jan, 2024

[3] Робастное управление https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D0%BE%D0%B5_%D1%83%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5 Retrieved: Jan, 2024

[4] "A Model-Based Approach for Robustness Testing" (PDF). Dl.ifip.org. Retrieved: Jan, 2024

[5] Xu, Huan, and Shie Mannor. "Robustness and generalization." Machine learning 86 (2012): 391-423.

[6] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93. (in Russian)

[7] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag, 2006

[8] Chung, Yeounoh, et al. "Unknown examples & machine learning model generalization." arXiv preprint arXiv:1808.08294 (2018).

[9] Tran, Dustin, et al. "Plex: Towards reliability using pretrained large model extensions." arXiv preprint arXiv:2207.07411 (2022).

[10] Carlini, Nicholas, et al. "On evaluating adversarial robustness." arXiv preprint arXiv:1902.06705 (2019).

[11] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[12] Намиот, Д. Е. "Схемы атак на модели машинного обучения." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[13] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russian)

[14] Bastani, Osbert, et al. "Measuring neural net robustness with constraints." Advances in neural information processing systems 29 (2016).

[15] Pedraza, Anibal, Oscar Deniz, and Gloria Bueno. "On the relationship between generalization and robustness to adversarial examples." Symmetry 13.5 (2021): 817.

[16] Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[17] Prishletsov, Dmitry, Sergey Prishletsov, and Dmitry Namiot. "Camouflage as adversarial attacks on machine learning models." International Journal of Open Information Technologies 11.9 (2023): 41-49. (in Russian)

[18] Foolbox https://github.com/bethgelab/foolbox Retrieved: Jan, 2024

[19] Костюмов, В. В. (2022). Обзор и систематизация атак уклонением на модели компьютерного зрения. International Journal of Open Information Technologies, 10(10), 11-20.

[20] Tramer, Florian, and Dan Boneh. "Adversarial training and robustness for multiple perturbations." Advances in neural information processing systems 32 (2019).

[21] Zhao, Weimin, Sanaa Alwidian, and Qusay H. Mahmoud. "Adversarial Training Methods for Deep Learning: A Systematic Review." Algorithms 15.8 (2022): 283.

[22] E. A. Ilyushin, D. E. Namiot, and I. V. Chizhov, "Attacks on Machine Learning Systems – Common Problems and Methods," International Journal of Open Information Technologies, vol. 10, no. 3, pp. 17-22, 2022. (in Russian)

[23] Tremblay, Maxime, et al. "Rain rendering for evaluating and improving robustness to bad weather." International Journal of Computer Vision 129 (2021): 341-360.

[24] Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).

[25] Kar, Oğuzhan Fatih, et al. "3d common corruptions and data augmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[26] Li, Peng, et al. "CleanML: A study for evaluating the impact of data cleaning on ml classification tasks." 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 2021.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent.Representation learning: A review and new perspectives.IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[28] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel.Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems, 29, 2016.

[29] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou.Interfacegan: Interpreting the disentangled face representation learned by gans.IEEE transactions on pattern analysis and machine intelligence, 44(4):2004–2018, 2020.

[30] Li, Shuai, Xiaoyu Jiang, and Xiaoguang Ma. "Transcending Adversarial Perturbations: Manifold-Aided Adversarial Examples with Legitimate Semantics." arXiv preprint arXiv:2402.03095 (2024).

[31] Cai, Q.-Z.; Du, M.; Liu, C.; Song, D. Curriculum Adversarial Training. arXiv 2018, arXiv:1805.04807

[32] Намиот Д. Е., Ильюшин Е. А. Порождающие модели в машинном обучении //International Journal of Open Information Technologies. – 2022. – Т. 10. – №. 7. – С. 101-118.

[33] Bai, Tao, et al. "Recent advances in adversarial training for adversarial robustness." arXiv preprint arXiv:2102.01356 (2021).

[34] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.

[35] Namiot, Dmitry, and Eugene Ilyushin. "Trusted Artificial Intelligence Platforms: Certification and Audit." International Journal of Open Information Technologies 12.1 (2024): 43-60. (in Russian)
[36] Brix, Christopher, et al. "The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results." arXiv preprint arXiv:2312.16760 (2023)
[37] Wu, Haoze, et al. "Toward certified robustness against real-world distribution shifts." 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, 2023.
[38] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian)
[39] Магистерская программа «Искусственный интеллект в кибербезопасности» (ФГОС) https://cs.msu.ru/node/3732 Retrieved: Jan, 2024.

On improving the robustness of machine learning models

Dmitry Namiot, Vladimir Romanov

***Abstract*— This article discusses how to improve robustness for machine learning models. Robustness is one of the most important characteristics of machine learning models, which determines the possibility of practical use of the models. However, not everything is simple when determining this characteristic for specific models. Firstly, not everything is clear with the very definition of robustness. If we consider robustness as the preservation of the behavior of the model under small perturbations of the initial data, then at least two questions arise - how small should these changes be, and how does such a definition relate to other characteristics of the model? First of all, among other characteristics, it is necessary to consider the generalizing ability of the model (generalization), which is determined by the model's work with previously unknown data. Google's concept of model reliability is also discussed. The main content of the article is devoted to the consideration of competitive training, which, despite all its limitations, remains today the main tool for increasing robustness.**

***Keywords*— machine learning, robustness, generalization.**

REFERENCES

[1] IEEE Std. 610.12 (IEEE Standard Glossary of Software Engineering Terminology).

[2] Robastnost' https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C Retrieved: Jan, 2024

[3] Robastnoe upravlenie https://ru.wikipedia.org/wiki/%D0%A0%D0%BE%D0%B1%D0%B0%D1%81%D1%82%D0%BD%D0%BE%D0%B5_%D1%83%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5 Retrieved: Jan, 2024

[4] "A Model-Based Approach for Robustness Testing" (PDF). Dl.ifip.org. Retrieved: Jan, 2024

[5] Xu, Huan, and Shie Mannor. "Robustness and generalization." Machine learning 86 (2012): 391-423.

[6] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93. (in Russian)

[7] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag, 2006

[8] Chung, Yeounoh, et al. "Unknown examples & machine learning model generalization." arXiv preprint arXiv:1808.08294 (2018).

[9] Tran, Dustin, et al. "Plex: Towards reliability using pretrained large model extensions." arXiv preprint arXiv:2207.07411 (2022).

[10] Carlini, Nicholas, et al. "On evaluating adversarial robustness." arXiv preprint arXiv:1902.06705 (2019).

[11] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[12] Namiot, D. E. "Shemy atak na modeli mashinnogo obuchenija." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[13] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134. (in Russian)

[14] Bastani, Osbert, et al. "Measuring neural net robustness with constraints." Advances in neural information processing systems 29 (2016).

[15] Pedraza, Anibal, Oscar Deniz, and Gloria Bueno. "On the relationship between generalization and robustness to adversarial examples." Symmetry 13.5 (2021): 817.

[16] Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[17] Prishletsov, Dmitry, Sergey Prishletsov, and Dmitry Namiot. "Camouflage as adversarial attacks on machine learning models." International Journal of Open Information Technologies 11.9 (2023): 41-49. (in Russian)

[18] Foolbox https://github.com/bethgelab/foolbox Retrieved: Jan, 2024

[19] Kostjumov, V. V. (2022). Obzor i sistematizacija atak ukloneniem na modeli komp'juternogo zrenija. International Journal of Open Information Technologies, 10(10), 11-20.

[20] Tramer, Florian, and Dan Boneh. "Adversarial training and robustness for multiple perturbations." Advances in neural information processing systems 32 (2019).

[21] Zhao, Weimin, Sanaa Alwidian, and Qusay H. Mahmoud. "Adversarial Training Methods for Deep Learning: A Systematic Review." Algorithms 15.8 (2022): 283.

[22] E. A. Ilyushin, D. E. Namiot, and I. V. Chizhov, "Attacks on Machine Learning Systems – Common Problems and Methods," International Journal of Open Information Technologies, vol. 10, no. 3, pp. 17-22, 2022. (in Russian)

[23] Tremblay, Maxime, et al. "Rain rendering for evaluating and improving robustness to bad weather." International Journal of Computer Vision 129 (2021): 341-360.

[24] Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).

[25] Kar, Oğuzhan Fatih, et al. "3d common corruptions and data augmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[26] Li, Peng, et al. "CleanML: A study for evaluating the impact of data cleaning on ml classification tasks." 2021 IEEE 37th International Conference on Data Engineering (ICDE). IEEE, 2021.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent.Representation learning: A review and new perspectives.IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[28] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel.Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems, 29, 2016.

[29] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou.Interfacegan: Interpreting the disentangled face representation learned by gans.IEEE transactions on pattern analysis and machine intelligence, 44(4):2004–2018, 2020.

[30] Li, Shuai, Xiaoyu Jiang, and Xiaoguang Ma. "Transcending Adversarial Perturbations: Manifold-Aided Adversarial Examples with Legitimate Semantics." arXiv preprint arXiv:2402.03095 (2024).

[31] Cai, Q.-Z.; Du, M.; Liu, C.; Song, D. Curriculum Adversarial Training. arXiv 2018, arXiv:1805.04807

[32] Namiot D. E., Il'jushin E. A. Porozhdajushhie modeli v mashinnom obuchenii //International Journal of Open Information Technologies. – 2022. – T. 10. – #. 7. – S. 101-118.

[33] Bai, Tao, et al. "Recent advances in adversarial training for adversarial robustness." arXiv preprint arXiv:2102.01356 (2021).

[34] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.

[35] Namiot, Dmitry, and Eugene Ilyushin. "Trusted Artificial Intelligence Platforms: Certification and Audit." International Journal of Open Information Technologies 12.1 (2024): 43-60. (in Russian)

[36] Brix, Christopher, et al. "The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results." arXiv preprint arXiv:2312.16760 (2023)

[37] Wu, Haoze, et al. "Toward certified robustness against real-world distribution shifts." 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). IEEE, 2023.

[38] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127. (in Russian)

[39] Magisterskaja programma «Iskusstvennyj intellekt v kiberbezopasnosti» (FGOS) https://cs.msu.ru/node/3732 Retrieved: Jan, 2024.