УДК 004.896 DOI 10.18522/2311-3103-2023-2-227-239

Р.М. Ауси, Е.В. Заргарян, Ю.А. Заргарян

ГЛУБОКОЕ ОБУЧЕНИЕ МЕТОДАМ ЗАЩИТЫ ОТ АТАК

В последние годы алгоритмы машинного обучения, а точнее алгоритмы глубокого обучения, широко используются во многих областях, включая кибербезопасность. Однако системы машинного обучения уязвимы для атак со стороны злоумышленников, и это ограничивает применение машинного обучения, особенно в нестационарных средах со враждебными действиями, таких как область кибербезопасности, где существуют настоящие злоумышленники (например, разработчики вредоносных программ). С быстрым развитием методов искусственного интеллекта (ИИ) и глубокого обучения (ГО) важно обеспечить безопасность и надежность реализованных алгоритмов. В последнее время уязвимость алгоритмов глубокого обучения к конфликтующим паттернам получила широкое признание. Изготовленные сфабрикованным образом образцы для анализа могут привести к различному нарушению поведения моделей глубокого обучения, в то время как люди будут считать их безопасными для использования. Успешная реализация атак противника в реальных физических ситуациях и сценариях реального физического мира еще раз доказывает их практичность. В результате методы состязательной атаки и защиты привлекают все большее внимание со стороны сообществ безопасности и машинного обучения и стали горячей темой исследований в последние годы не только на территории России, но и других странах. Компании «Сбербанк», «Яндекс», «Группа Т1», «Медицинский центр Атлас» и многие другие ведут разработку конкурентоспособных решений, в том числе и на международном рынке. К сожалению, в списке 10 крупнейших ИТ-компаний направление Big Data, в частности и защита от атак представлено только компанией «Группа Т1», но потенциал роста рынка огромный. В данной работе представляются теоретические основы, алгоритмы и применение методов состязательных атак противника. Затем описывается ряд исследовательских работ по методам защиты, охватывающих широкий спектр исследований в этой области. В этой статье исследуется и обобщаются состязательные атаки и средства защиты, которые представляют собой самые современные исследования в этой области и отвечают последним требованиям, предъявляемым к информационной безопасности.

Состязательное машинное обучение; глубокая нейронная сеть; состязательная атака; информационная безопасность; кибербезопасность.

R.M. Ausi, E.V. Zargaryan, Yu.A. Zargaryan

DEEP TRAINING IN METHODS OF PROTECTION AGAINST ATTACKS

In recent years, machine learning algorithms, or rather deep learning algorithms, have been widely used in many fields, including cybersecurity. However, machine learning systems are vulnerable to attacks by attackers, and this limits the use of machine learning, especially in non-stationary environments with hostile actions, such as the cybersecurity field, where real attackers exist (for example, malware developers). With the rapid development of artificial intelligence (AI) and deep learning (GO) methods, it is important to ensure the safety and reliability of the implemented algorithms. Recently, the vulnerability of deep learning algorithms to conflicting patterns has been widely recognized. Fabricated samples for analysis can lead to various violations of the behavior of deep learning models, while people will consider them safe to use. The successful implementation of enemy attacks in real physical situations and scenarios of the real physical world once again proves their practicality. As a result, methods of adversarial attack and defense are attracting increasing attention from the security and machine learning communities and have become a hot topic of research in recent years not only in Russia, but also in other countries. Sberbank, Yandex, T1 Group, Atlas Medical Center and many others are developing competitive solutions, including on the international market. Unfortunately, in the list of the 10 largest IT companies, the direction of Big Data, in particular, and protection against attacks is represented only by the T1 Group company, but the market growth potential is huge. In this paper, the theoretical foundations, algorithms and application of methods of adversarial attacks of the enemy are

presented. Then a number of research papers on protection methods are described, covering a wide range of research in this area. This article explores and summarizes adversarial attacks and defenses, which represent the most up-to-date research in this field and meet the latest requirements for information security.

Adversarial machine learning; deep neural network; adversarial attack; information security; cybersecurity.

Введение. Многомиллиардное увеличение вычислительной мощности популяризировало использование глубокого обучения (ГО) для обработки многих задач машинного обучения (МО), таких как классификация изображений [1], обработка естественного языка и теория игр. Тем не менее, серьезная угроза безопасности современных алгоритмов ГО была обнаружена исследовательским сообществом: взломщики могут легко обмануть модели ГО, нарушив искажая доброкачественные образцы, не будучи обнаруженными человеком [2]. Возмущения, которые незаметны для человеческого зрения / слуха, достаточны, чтобы заставить модель делать ложные прогнозы с высокой степенью достоверности. Это явление, известное как состязательный паттерн (состязательная выборка), считается основным препятствием для массового развертывания моделей ГО в производстве. Для изучения этой открытой проблемы были предприняты значительные исследовательские усилия.

В соответствии с моделью угроз, существующие состязательные атаки можно классифицировать на атаки белого ящика, серого ящика и черного ящика. Разница между тремя моделями заключается в знании противников. В модели угроз атак «белого ящика» предполагается, что противники обладают полными знаниями о своей целевой модели, включая архитектуру и параметры модели. Следовательно, они могут напрямую создавать состязательные образцы на целевой модели любыми способами. В модели угроз «серого ящика» знания противников ограничены структурой целевой модели. В модели угроз черного ящика злоумышленники могут прибегать к доступу к запросам только для создания состязательных образцов. В рамках этих моделей угроз был предложен ряд алгоритмов атаки для генерации состязательной выборки, таких как алгоритм Бройдена-Флетчера Гольдфарба Шанно с ограниченной памятью (limited-memory Broyden–Fletcher Goldfarb Shanno – L-BFGS), метод быстрого градиентного знака (fast gradient sign method – FGSM) , базовый итерационный метод (basic iterative method – BIM) / проектируемый градиентный спуск (projected gradient descent – PGD), распределенно-состязательная атака [3, 4], атаки Карлини и Вагнера (C&W) [5], атака карты значимости на основе Якобиана (Jacobian based saliency map attack – JSMA). Эти алгоритмы атак разработаны в модели угроз белого ящика. Тем не менее, они также эффективны во многих настройках серого и черного ящиков из-за переносимости состязательных образцов между моделями [1].

Между тем, недавно были предложены различные защитные методы для обнаружения/классификации состязательных атак, включая эвристические и сертифицированные средства защиты. Эмпирическая защита относится к защитному механизму, который успешно защищает от определенных атак без теоретической гарантии точности. В настоящее время наиболее эффективной защитной эвристикой является состязательное обучение, которое пытается повысить надежность модели ГО путем введения нежелательных образцов на этапе обучения.

Другие эвристические средства защиты в значительной степени полагаются на переходы ввода/функции и шумоподавление для смягчения неблагоприятных последствий в областях данных/функций. С другой стороны, сертифицированные системы защиты всегда могут засвидетельствовать свою минимальную точность против четко определенного типа атаки противника. В последнее время популяр-

ным подходом к сертификации сети является построение состязательного многогранника и определение его верхней границы с помощью выпуклых расширений. Ослабленная верхняя граница – это сертификация для обученных моделей ГО, которая гарантирует, что ни одна атака с определенными ограничениями не может превышать приблизительный стандартизированный уровень успеха атаки, аппроксимированный верхней границей. Тем не менее, фактические результаты этих сертифицированных защит были все еще намного хуже, чем обучение [3, 4].

В этой статье исследуется и обобщаются состязательные атаки и средства защиты, которые представляют собой самые современные исследования в этой области.

Общие сведения. В этом разделе описываются определения и обозначения, используемые в этой статье. В частности, набор данных определяется как $\{X_i, Y_i\}_{i=1}^{N}$, где X_i- образец данных с меткой y_i, N - размер набора данных. Нейронная сеть обозначается как $f(.)$. с входом x и предсказанием $f(x)$. Соответствующие потери оптимизации (также называемые состязательными потерями) обозначаются $J(\theta, x, y)$, где θ обозначает веса модели. Для задачи классификации перекрестная энтропия между $f(x)$ и меткой y всегда применяется как потеря оптимизации, которая обозначается $J(f(x); y)$. Образец данных x' считается состязательным образцом x, когда x' близок к x при определенной метрике расстояния, в то время как $f(x') \neq y$. Формально состязательный образец x.

$$x' : D(x, x') < \eta, f(x') \neq y, \tag{1}$$

где D(.,.) метрика расстояния и η предопределенное ограничение расстояния, которое также известно как разрешенное возмущение. Эмпирически, а значение η принято для обеспечения сходства между x и x' такой, что x' неотличим от x.

По определению, состязательный образец x' должен находиться близко к состязательному образцу x под определенной метрикой расстояния. Наиболее часто используемой метрикой расстояния является L_P метрика расстояния [8]. L_P расстояние между x и x' обозначается как:

$$\|\upsilon\|_{\mathrm{p}} = (|\mathrm{v}_1|^{\mathrm{p}} + |\mathrm{v}_2|^{\mathrm{p}} + .. |\mathrm{v}_{\mathrm{d}}|^{\mathrm{p}})^{1/\mathrm{p}}, \tag{2}$$

где p – вещественное число; d – размерность вектора расстояния v.

Конкретно, L_0 расстояние соответствует количеству элементов в выборке x, измененной состязательной атакой. L_2 расстояние измеряется по формуле Евклидова расстояния между x и x'. Самая популярная метрика расстояния – L_∞ расстояние измеряет максимальную поэлементную разницу между доброкачественными и состязательными образцами. Существует также несколько состязательных атак на дискретные данные, которые применяются к другим метрикам расстояния, таким как количество выпавших точек и семантическое сходство [6].

Модели угроз. Существует три основные модели угроз для состязательных атак и защиты: модели «черного ящика», «серого ящика» и «белого ящика». Эти модели определяются в соответствии со знанием противников. В модели «черного ящика» злоумышленник не знает структуру целевой сети или параметры, но может взаимодействовать с алгоритмом глубокого обучения для запроса прогнозов для конкретных входных данных. Противники всегда создают состязательные образцы на суррогатном классификаторе, обученном приобретенными парами данных, прогнозировании и другими доброкачественными /состязательными выборками. Из-за переносимости состязательных образцов атаки «черного ящика» всегда могут поставить под угрозу естественно обученную незащищенную модель [6–8].

В модели «серого ящика» предполагается, что злоумышленник знает архитектуру целевой модели, но не имеет доступа к весам в сети. Противник также может взаимодействовать с алгоритмом ГО. В этой модели угроз ожидается, что противник создаст состязательные образцы на суррогатном классификаторе той же архитектуры. Из-за дополнительной информации о структуре противник «серого ящика» всегда показывает лучшую производительность атаки по сравнению с противником «черного ящика». Самый сильный противник, то есть противник «белого ящика», имеет полный доступ к целевой модели, включая все параметры, что означает, что противник может адаптировать атаки и напрямую создавать состязательные образцы на целевой модели. В настоящее время многие методы защиты, которые были продемонстрированы как эффективные против атак «черного ящика» / «серого ящика», уязвимы для адаптивной атаки «белого ящика» [5, 7].

Состязательные атаки. В этой части статьи представлено несколько типичных алгоритмов и методов состязательной атаки. Эти методы могут быть использованы для атаки на другие модели ГО, а также на модели классификации изображений ГО. Подробно описаны конкретные состязательные атаки на другие модели ГО.

1. АЛГОРИТМ L-BFGS. Первое сообщение об уязвимости глубоких нейронных сетей (ГНС) к состязательным выборкам появляется в 2013 году. В частности, изображение подвергается едва заметным состязательным возмущениям, чтобы повлиять на результат классификации ГНС. Для выявления состязательных возмущений с минимальной нормой L_p предлагается следующая формулировка метода:

$$\min_x \|x - x'\|_p, f(x') \neq y', \tag{3}$$

где $\|x - x'\|_p$ – l_p норма состязательных возмущений и y' – это состязательная метка цели ($y' \neq y$). Однако эта проблема оптимизации неразрешима.

Можно минимизировать гибридные потери, то есть $C\|x - x'\|_p + J(\theta, x', y')$, где с – параметр, определяется, как приближение к решению задачи оптимизации, где оптимальное значение с можно найти с помощью линейного поиска по сетке [9].

2. МЕТОД СО ЗНАКАМИ БЫСТРОГО ГРАДИЕНТА. Предлагается создать эффективную нецелевую атаку, называемую FGSM. Для создания данной атаки L_∞ необходимо состязательные образцы добавить в доброкачественные образцы, как показано на рис. 1. FGSM – типичный алгоритм одношаговой атаки, который выполняет одношаговое обновление по направлению (т.е. знаку) градиента состязательного проигрыша $j(\theta, x, y)$, что позволяет увеличить потери в самом большом направлении. Формально сгенерированная FGSM состязательная выборка формулируется следующим образом:

$$x' = x + \epsilon.sign[\nabla_x J(\theta, x, y)], \tag{4}$$

где ϵ – является величиной возмущения. FGSM может быть легко расширен до алгоритма целенаправленной атаки путем погружения в градиент $J(\theta, x, y')$, в котором y' – это метка цели. Эта стратегия обновления может уменьшить перекрестную энтропию между ожидаемым вектором вероятности и объективным вектором вероятности, если перекрестная энтропия применяется в качестве антагонистической потери. Правило обновления для целевого FGSM можно сформулировать следующим образом:

$$x' = X - \epsilon.sign[\nabla_x J(\theta, x, y')]\,. \tag{5}$$

Кроме того, было обнаружено, что случайное возмущение перед выполнением FGSM на доброкачественных образцах может повысить производительность и разнообразие состязательных образцов FGSM.

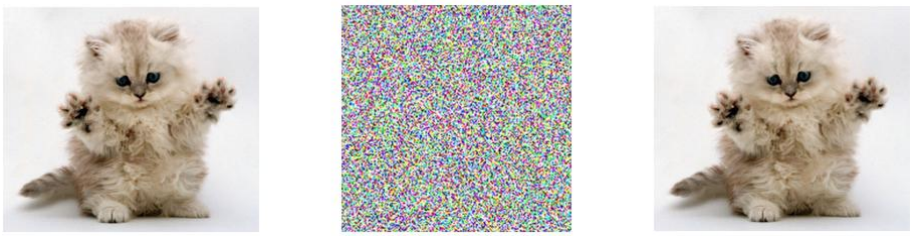

Рис. 1. Демонстрация состязательного образца, сгенерированного путем применения FGSM к GoogleNet [10]

3. BIM и PGD. BIM дополняет FGSM с меньшим размером шага и обрезает обновленный состязательный образец в допустимый диапазон для итераций T; то есть в t-й итерации правило обновления выглядит следующим образом:

$$x'_{t+1} = CLIP\{x'_t + \alpha \cdot sign[\nabla_x J(\theta, x'_t, y)]\}, \quad (6)$$

где $\alpha T = \epsilon$ и α – величина возмущения на каждой итерации. PGD можно рассматривать, как обобщенную версию BIM без ограничения в $\alpha T = \epsilon$. Чтобы ограничить конфликтные возмущения, PGD проецирует конфликтные выборки, извлеченные из каждого $\epsilon - L_\infty$ в соседний. Следовательно, размер враждебного возмущения меньше, чем ϵ. Формально процедура обновления выполняется следующим образом

$$x'_{t+1} = Proj\{x'_t + \alpha \cdot sign[\nabla_x J(\theta, x'_t, y)]\}, \quad (7)$$

где Proj проецирует обновленный состязательный образец в $\epsilon - L_\infty$ d допустимый диапазон [10].

4. ИМПУЛЬСНАЯ ИТЕРАТИВНАЯ АТАКА. Предлагается интеграция импульсной памяти в итеративный процесс BIM и разработывается новый итеративный алгоритм, называемый итеративным импульсным FGSM (MI-FGSM). В частности, MI-FGSM итеративно обновляет шаблон противника следующим образом:

$$x^1_{t+1} = clip\{x^1_t + \alpha . sign(g_{t+1})\}, \quad (8)$$

где градиент g обновляется по формуле: $g_{t+1} = \xi g_t + \frac{\nabla_x J(\theta, x'_t, y)}{\|\nabla_x J(\theta, x'_t, y)\|}$, ξ – является фактором распада данных.

Предлагается также план, который означает изготовление труппы моделей, чтобы исследовать модель в настройках «черного ящика» / «серого ящика». Основная идея состоит в том, чтобы рассмотреть градиенты нескольких моделей по отношению к входным данным и определить направление градиента, которое с большей вероятностью будет передано другим моделям. Комбинация MI-FGSM и ансамблевой схемы атаки заняла первые места в соревнованиях по нецелевой атаке противника и целевой атаке противника (настройка черного ящика) [11].

5. ДИСТРИБУТИВНО-АНТАГОНИСТИЧЕСКАЯ АТАКА. Предлагается новая состязательная атака, действующая в пространстве вероятностных мер, известная как распределенная состязательная атака (DAA). В отличие от PGD, где конфликтующие образцы генерируются независимо для каждого доброкачественного образца, DAA выполняет оптимизацию потенциально конфликтующих распределений. Кроме того, предлагаемая цель состоит в том, чтобы сначала включить расхождение Крафта-Макмиллана (KL) между диссонирующим и доброкаче-

ственным распределением данных в расчет потерь несоответствия, чтобы увеличить риск противоречивых обобщений в процессе. оптимизатор. Эта задача оптимизации распределения сформулирована следующим образом:

$$\max_{\mu} \int_{\mu} J(\theta, x', y) d\mu + kl[\mu(x') \parallel \pi(x)], \quad (9)$$

где обозначает состязательное распределение данных и доброкачественное $\pi(x)$распространение данных.

Поскольку прямая оптимизация над распределением неразрешима, авторы используют два метода оптимизации частиц для аппроксимации. По сравнению с PGD, DAA исследует новые состязательные паттерны, как показано на рис. 2 [12–14]. Данная модель считается одной из самых эффективных атак на несколько оборонительных моделей l_{∞}.

6. АТАКИ НА ОСНОВЕ ОПТИМИЗАЦИИ. Предлагается набор состязательных атак на основе оптимизации (C&W атак), которые могут генерировать l_0, l_1, и l_{∞} нормы измеренных состязательных образцов, а именно: CW_0, CW_1, CW_{∞} Похожий на L-BFGS, формулируется цель оптимизации следующим образом:

$$\min_{\delta} D(x, x+\delta) + c.F)(x+\delta), \text{ при условии } x+\delta \in [0,1]. \quad (10)$$

где δ обозначает состязательное возмущение, D (.,.), l_0, l_1, l_{∞} – матричное расстояние, и $f(x+\delta)$ ообозначает настройку состязательно потери, которая удовлетворяет $f(x+\delta) \leq 0$, если методика DNNs является прогнозом, то – это цель атаки для обеспечения $(x+\delta)$, что выдает допустимое изображение, вводит новый заменитель переменной К δ следующим образом:

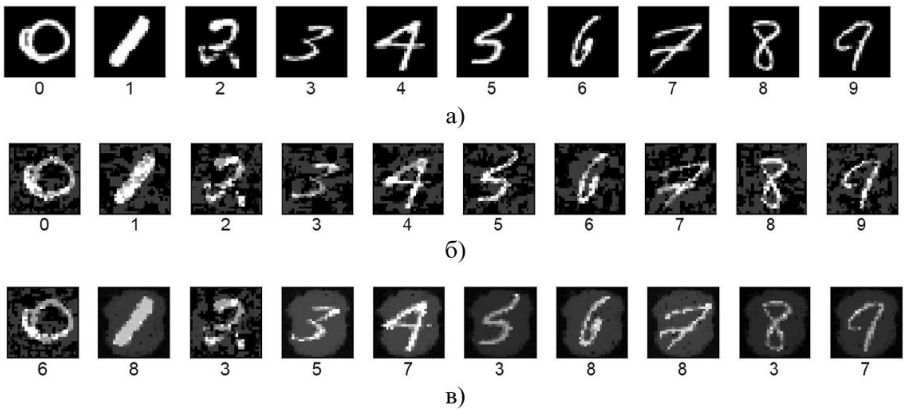

Рис. 2. Сравнение между PGD и DAA. DAA имеет тенденцию генерировать более структурированные возмущения [7] а – изначальные значение, б – после применения PGD, в – после применения DAA

$$\delta = \frac{1}{2}[\tanh(k) + 1], \quad (11)$$

так что $x+\delta = \frac{1}{2}[\tanh(k)+1]$, который всегда находится в диапазоне [0,1] в процессе оптимизации.

Атаки C&W обеспечивают 100-процентный прогресс атаки на нормально подготовленных DNN для MNIST, CIFAR-10 и ImageNet. Они также компрометируют охраняемые утонченные модели, на которых L-BFGS и Deep Fool пренебрегают поиском недоброжелательных результатов.

7. ОСНОВАННЫЙ НА ПОДХОДЕ К КАРТЕ ЗНАЧИМОСТИ. Предложено эффективная целевая атака под названием JSMA, которая может обмануть DNN с маленькими l_0 возмущений. Метод сначала вычисляет матрицу:

$$\nabla l(x) = \frac{\partial(x)}{\partial x} = \left[\frac{ax_j(x)}{ax_y}\right], y \in 1, \dots M_{in}.j \in 1, M_{OUT}\ , \quad (12)$$

где M_{in} – число нейронов во входном слое; M_{out} – количество нейронов на выходном слое; c – индекс для входного компонента x; j – индекс для выходного компонента l.

Матрица определяет, как элементы ввода влияют на выходы различных классов. Согласно якобианской матрице, карта состязательности $xS(X, Y^1)$ определяется для выбора объектов/пикселей, которые должны быть возмущены для получения желаемых изменений в выходных данных [11, 12]. В частности, предлагаемый алгоритм возмущает элемент $X[Y]$ с наибольшим значением $S(X, Y^1)[\gamma]$ и значительно увеличивает/уменьшает выходы целевого/другого состояния. Следовательно, возмущения на небольшой части элементов уже могут влиять на $l(x)$ и обмануть нейронную сеть.

8. УНИВЕРСАЛЬНАЯ СОСТЯЗАТЕЛЬНАЯ АТАКА. Во всех упомянутых выше атаках искусственно созданные враждебные возмущения специфичны для доброкачественных образцов. Другими словами, враждебные возмущения не передаются по доброкачественным выборкам.

Здесь возникает прямой вопрос: существует ли универсальное возмущение, которое может обмануть сеть на большинстве доброкачественных выборок?

Такой вектор возмущения путем итеративного обновления возмущения с использованием всех целевых доброкачественных выборок существует. На каждой итерации для безвредных выборок, которые текущее возмущение не может обмануть, решается задача оптимизации, которая аналогична L-BFGS [15] и которая направлена на обнаружение минимального дополнительного возмущения, необходимого для компрометации выборок. Затем дополнительное возмущение добавляется к текущему возмущению. В конечном счете, возмущение позволяет большинству доброкачественных выборок обмануть сеть. Эксперименты показывают, что этот простой итеративный алгоритм эффективен для атаки на глубокие сети. Удивительно, но эта переносимость между выборками также сохраняется в разных моделях; например, универсальные возмущения, созданные на VGG, также могут достигать коэффициента обманчивости выше 53% , чем в других моделях.

Враждебные атаки на широко распространенные приложения, представляющие промышленный интерес.

1. Модели семантической сегментации являются объектом состязательных атак. Предложен систематический алгоритм - для генерации состязательных образцов для задач обнаружения объектов и сегментации. Основная идея данного алгоритма заключается в рассмотрении всех целей в задаче обнаружения/сегментации одновременно и оптимизации общих потерь. Более того, чтобы справиться с большим количеством предложений в задаче обнаружения объектов на уровне пикселей, алгоритм сохраняет увеличенное, но разумное количество предложений, изменяя скорость пересечения над объединением в процессе оптимизации. Отмечается, что для задачи сегментации связь между широко используемыми состязательными потерями и точностью не так хорошо установлена, как в задаче классификации. Поэтому предлагается новая альтернативную потеря, чтобы оценить истинные потери противника, которые являются продуктом случайных ошибок и потерь миссии. Случайная ошибка характеризуется разницей между предсказанной вероятностью, лежащей в основе истины и вероятностью предсказания цели. Независимые от модели потери соответствуют цели максимизации.

Кроме того, выводится новое приближение градиента потерь замещения для входных данных, чтобы обеспечить оптимизацию на основе градиента на входе. Эксперименты показывают, что предложенный алгоритм достигает наивысшей производительности атаки при сегментировании семантики.

Методы состязательной защиты

1. Состязательной обучение. Состязательное обучение, которое направлено на повышение надежности нейронной сети путем обучения ее с помощью состязательных образцов, является интуитивно понятной защитой. Официально это игра min-max, которую можно спланировать следующим образом:

$$\min_{\theta} \max_{D(x1,x^1)} J(\theta, x^1, y), \quad (13)$$

где $J(\theta, x^1, y)$ – состязательная потеря, с сетевым весом h, состязательный ввод x', и метка истинности y. $D(x, x')$ представляет определенную метрику расстояния между x и x'. Задача внутренней максимизации заключается в поиске наиболее эффективных состязательных образцов, что решается хорошо продуманной состязательной атакой, такой как FGSM [5] и PGD [6]. Внешняя минимизация является стандартной процедурой обучения для минимизации потерь. Предполагается, что полученная сеть должна быть устойчивой к состязательной атаке, используемой для генерации состязательной выборки на этапе обучения. Противоборствующая подготовка является одной из наиболее эффективных защит от состязательных атак. В частности, он достигает самой высокой точности по нескольким критериям.

2. Случайный шум. Предлагается использовать случайный механизм шумоподавления, известный как случайный (RSE), для защиты от состязательных возмущений. Чтобы стабилизировать выходы DNN, RSE исследуются результаты прогнозирования над случайными шумами и добавляется шумовой слой перед каждым сверточным слоем во время фаз обучения и тестирования. Предлагается защита на основе методов глубокого обучения. Чтобы обеспечить границы ГО по изменению распределения по сравнению с его прогнозами входных данных разработанный метод можно использовать для защиты L_1/L_2 атаки с использованием механизмов ГО. Далее предлагается напрямую добавлять случайный шум к пикселям состязательных примеров перед классификацией, чтобы устранить последствия состязательных возмущений. Этот простой метод может превышать размер состязательного возмущения, к которому он устойчив, что зависит от первой и второй по величине вероятностей распределения выходной вероятности (вектора).

3. Шумоподавление. Шумоподавление является очень простым методом для уменьшения шумовых / контрастных эффектов. Есть два направления проектирования таких защит: входное шумоподавление и шумоподавление функций. Первая составляющая удаляет некоторые или все конфликтующие возмущения из входных данных, а вторая пытается свести к минимуму влияние конфликтующих возмущений на высокоуровневые функции.

Чтобы свести к минимуму конфликтующие эффекты сначала используются два метода сжатия (шумоподавления) уменьшения битов и размытия изображения – для снижения степеней свободы и устранения шумов столкновения, как показано на рис. 3. Обнаружение импульсных паттернов выполняется путем сравнения прогнозов модели на исходном и сжатом изображении. Если исходные и сжатые входные данные производят отличать выходные данные от модели, исходные входные данные могут быть конфликтующим образцом. Предполагается, что методы сжатия признаков, могут смягчить C&W-атаку. После каждого шага процесса оптимизации доступно промежуточное изображение. Уменьшенная глубина цвета версии этого промежуточного изображения проверяется системой обнару-

жения [16]. Такой процесс оптимизации выполняется несколько раз, и все промежуточные образцы конкурентов, которые могут пройти систему, агрегируются. Все эти адаптивные атаки могут нарушить систему сжатия входных данных с гораздо меньшими шумами.

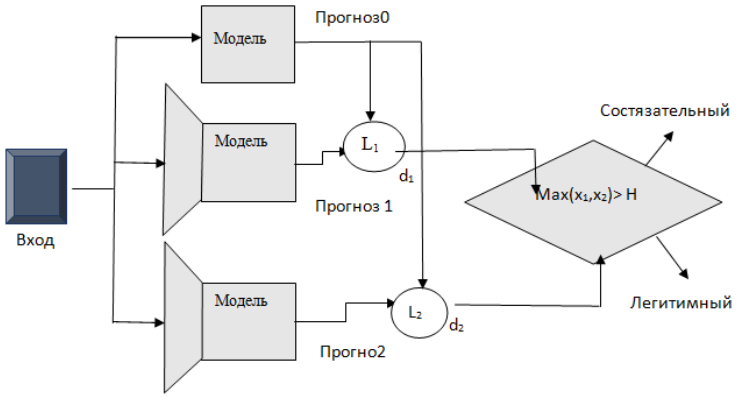

Рис. 3. Структура для сжатия признаков. d_1 и d_2 - разница между прогнозом модели на исходном входе и сжатом входе; H - порог, который используется для поиска примеров состязательного поведения

4. Очистка входов на основе методологии GAN. GAN является мощным инструментом для изучения обобщенной модели распределения данных. Поэтому многие задачи целесообразно решать с использованием GAN для изучения доброкачественного распределения данных для создания доброкачественной проекции для состязательного ввода. Защита GAN и устранение помех со стороны противника GAN (APE-GAN) являются двумя типичными алгоритмами среди всех аналогичных. Защита-GAN образует генератор для моделирования распределения доброкачественных изображений, как показано на рис. 4 [17]. На этапе тестирования защита-GAN очищает состязательный ввод, ища изображение, близкое к состязательному входу в его изученном распределении, и передает это доброкачественное изображение в классификатор. Эта стратегия может быть использована для защиты от различных вражеских атак. В настоящее время наиболее эффективная схема атаки против защиты-GAN основана на обратном дифференциальном приближении [18], что позволяет снизить его точность до 55% при возмущении противника 0,005 L_2. APE-GAN [80] непосредственно учит генератор очищать противоречивый образец, используя его в качестве входных данных, и генерирует доброкачественный аналог.

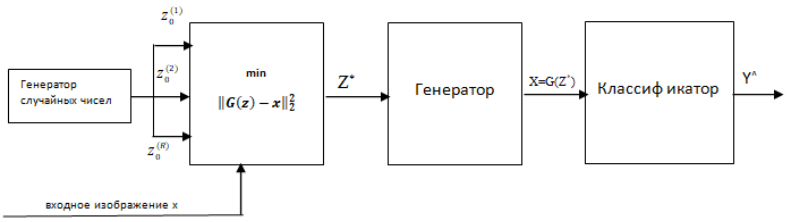

Рис. 4. Защита-ГАН. G – общая модель может отбирать высокоразмерные входные данные из низкоразмерного z-вектора; R – случайное векторное число, сгенерированное генератором случайных чисел

Выводы:

1. Атаки «белого ящика» и «черного ящика». С точки зрения оппонента, основное различие между настройками «белого» и «черного ящика» заключается в их уровне подхода к целевой модели. В настройках «белого ящика» злоумышленники могут получить доступ к структуре и весам модели, чтобы они могли рассчитать наклон реальной модели или оценить уклон с помощью методов [19]. Кроме того, противники могут корректировать свои методы атаки с помощью методов и настроек защиты. В этом контексте большинство ранее введенных эвристических защит неэффективны против таких мощных адаптивных противников. Однако в настройках «черного ящика» структура модели и веса являются секретными для противника. В этом контексте, чтобы применить вышеуказанные алгоритмы атаки на основе градиента, противник должен вывести значения модели из ограниченной информации. Без какой-либо информации, специфичной для модели, непредвзятая оценка параметров модели является ожиданием набора предварительно обученных моделей с различными случайными частицами. Таким образом, противник может вывести параметры из выхода целевой модели с хорошо спроектированными входами. В этом контексте предлагаемая конструкция может применять метод нулевого порядка, чтобы дать гораздо лучшую оценку параметров модели. Однако недостатком этого метода является то, что он требует большого количества целевых попаданий.

2. Различия между потоками состязательной атаки и защиты. Поток исследований состязательной атаки охватывает два основных направления. Первое направление заключается в разработке более эффективных и сильных атак для оценки различных систем защиты. Важность этого направления интуитивно понятна, поскольку ожидается понять все угрозы для потенциальных противников. Второе направление – это трансляция атак противника в физический мир. До сих пор эта тема исследования была сосредоточена на том, представляют ли эти вражеские атаки реальную угрозу в физическом мире. Некоторые исследователи предположили, что конфликтующие атаки, первоначально разработанные в цифровых пространствах, не будут эффективны в физическом мире из-за влияния определенных факторов окружающей среды. Недавно Cao et al. [20–25] удалось создать противоположные цели, чтобы обмануть систему обнаружения на основе LiDAR, еще раз подтвердив существование противоположных физических образцов. Когда дело доходит до обороны, сообщество начинает фокусироваться на сертифицированной безопасности, поскольку большинство эвристических мер безопасности не обеспечивают защиту от адаптивных атак белого ящика, а сертифицированная защита должна гарантировать, что защита эффективна в некоторых ситуациях, независимо от ситуации.

Однако до сих пор масштабируемость была распространенной проблемой для большинства сертифицированных систем безопасности. Например, доменно-коррелированный анализ является популярным новым направлением для сертификации DNN, но он не масштабируется до очень глубоких нейронных сетей и больших данных. Конечно, развитие защиты сталкивается с большими проблемами по сравнению с нападением. Это происходит главным образом потому, что атака может быть нацелена только на одну категорию защиты, но защита должна быть сертифицирована, т.е. должна быть эффективной против всех возможных методов нападения в определенных ситуациях.

В этой статье предоставлен обзор новейших репрезентативных методов защиты и атаки, которые более подробно будут рассматриваться при исследовании атак на предприятиях, используемых нейронные сети для работы. Рассмотрены идеи и методы предложенных методов и алгоритмов. К сожалению, в настоящее время нет никакого защитного механизма, который был бы эффективным и действенным против состязательных атак.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Krizhevsky A., Sutskever I., Hinton G.E.* ImageNet classification with deep convolutional neural networks // Proceedings of the 26th Conference on Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA; 2012. – P. 1097-105.
2. *Чо К., ван Мерриенбур Б., Гюльсехре С., Бахданау Д., Бугарес Ф., Швенк Х. и др.* Изучение фразовых представлений с использованием кодировщика-декодера RNN для статистического машинного перевода. 2014. arXiv:1406.1078.
3. *Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., et al*. Intriguing properties of neural networks. 2013. arXiv:1312.6199.
4. *Goodfellow I.J., Shlens J., Szegedy C.* Explaining and harnessing adversarial examples. 2014. arXiv:1412.6572.
5. *Заргарян Ю.А.* Задача управляемости в адаптивной автоматной обучаемой системе управления // Матер. X Международной научно-технической конференции "Технологии разработки информационных систем". – 2020.
6. *Zargaryan E.V., Zargaryan Y.A., Kapc I.V., Sakharova O.N., Kalyakina I.M and Dmitrieva I.A.* Method of estimating the Pareto-optimal solutions based on the usefulness. International Conference on Advances in Material Science and Technology - CAMSTech-2020 // IOP Conf. Series: Materials Science and Engineering. – 2020. – Vol. 919 (2). – P. 022027 (1-8). – DOI: 10.1088/1757-899X/919/2/022027.
7. *Zheng T., Chen C., Ren K.* Distribution ally adversarial attack. 2018. arXiv:1808.05537.
8. *Карлини Н., Вагнер Д.* К оценке надежности нейронных сетей // Матер. симпозиума IEEE 2017 года по безопасности и конфиденциальности; 22–26 мая 2017 г. Сан-Хосе, Калифорния, США, 2017. – С. 39-57.
9. *Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A.* The limitations of deep learning in adversarial settings // Proceedings of the 2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24;Saarbrucken, Germany, 2016. – P. 372-87.
10. *Moosavi-Dezfooli S.M., Fawzi A., Frossard P.* DeepFool: a simple and accurate method to fool deep neural networks // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30. Las Vegas, NV, USA, 2016. – P. 2574-82.
11. *Papernot N., McDaniel P., Goodfellow I.* Transferability in machine learning:from phenomena to black-box attacks using adversarial samples. 2016. arXiv:1605.07277.
12. *Liu Y., Chen X., Liu C., Song D.* Delving into transferable adversarial examples and black-box attacks. 2016. arXiv:1611.02770.
13. *Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards deep learning models resistant to adversarial attacks. 2017. arXiv: 1706.06083.
14. *Аламир Х.С., Заргарян Е.В., Заргарян Ю.А.* Модель прогнозирования транспортного потока на основе нейронных сетей для предсказания трафика на дорогах // Известия ЮФУ. Технические науки. – 2021. – № 6 (223). – С. 124-132.
15. *Zheng T., Chen C., Yuan J., Li B., Ren K.* Point Cloud saliency maps. 2018. arXiv:1812.01687.
16. *Beloglazov D., Shapovalov I., Soloviev V., Zargaryan E.* The hybrid method of path planning in non-determined environments based on potential fields // ARPN Journal of Engineering and Applied Sciences. – 2017. – Vol. 12, No. 23. – P. 6762-6772.
17. *Атали А., Карлини Н., Вагнер Д.* Запутанные градиенты дают ложное чувство безопасности: обход защиты к состязательным примерам. 2018.arXiv:1802.00420.
18. *Zargarjan E.V., Zargarjan Ju.A., Finaev V.I.* Information support for the training of fuzzy production account balance in the conditions of incomplete data // Innovative technologies and didactics in teaching (ITDT-2016): Collected papers. – 2016. – P. 128-138.
19. *Чен Ю., Шарма Ю., Чжан Х., И Дж., Се С.Дж.* EAD: атаки эластичной сети на глубокие нейронные сети на состязательных примерах // Матер. тридцать второй конференции AAAI по искусственному интеллекту; 2–7 февраля 2018 г. Новый Орлеан, Лос-Анджелес, США, 2019.
20. *Пушнина И.В.* Система управления подвижным объектом в условиях неопределенности // Наука и образование на рубеже тысячелетий: Сб. научно-исследовательских работ. – Кисловодск, 2018. – С. 65-74.
21. *Xiao C., Li B., Zhu J.Y., He W., Liu M., Song D.* Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.

22. *Роннебергер О., Фишер., Брокс Т.* U-net: сверточные сети для сегментации биомедицинских изображений // Матер. Международной конференции по вычислительной технике медицинских изображений и компьютерному вмешательству; 5–9 октября 2015 г. Мюнхен, Германия, 2015. – С. 234-41.
23. *Qi C.R., Su H., Mo K., Guibas L.J.* PointNet: deep learning on point sets for 3D classification and segmentation // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI,USA, 2017. – P. 652-60.
24. *Финаев В.И., Заргарян Ю.А., Заргарян Е.В., Соловьев В.В.* Формализация групп подвижных объектов в условиях неопределённости для выбора управляющих решений // Информатизация и связь. – 2016. – № 3. – С. 56-62.
25. *Бехзадан В., Мунир А.* Уязвимость глубокого обучения с подкреплением к атакам с целью индукции политики // Матер. Международной конференции по машинному обучению и интеллектуальному анализу данных в распознавании образов; 15–20 июля 2017 г. Нью-Йорк, Нью-Йорк, США, 2017. – С. 262-75.

REFERENCES

1. *Krizhevsky A., Sutskever I., Hinton G.E.* ImageNet classification with deep convolutional neural networks, *Proceedings of the 26th Conference on Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA; 2012,* pp. 1097-105.
2. *Cho K., van Merrienbur B., Gyul'sekhre S., Bakhdanau D., Bugares F., Shvenk Kh. i dr.* Izuchenie frazovykh predstavleniy s ispol'zovaniem kodirovshchika-dekodera RNN dlya statisticheskogo mashinnogo perevoda [The study of phrasal representations using RNN encoder-decoder for statistical machine translation], 2014. arXiv:1406.1078.
3. *Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., et al*. Intriguing properties of neural networks, 2013, arXiv:1312.6199.
4. *Goodfellow I.J., Shlens J., Szegedy C.* Explaining and harnessing adversarial examples. 2014. arXiv:1412.6572.
5. *Zargaryan Yu.A.* Zadacha upravlyaemosti v adaptivnoy avtomatnoy obuchaemoy sisteme upravleniya [The problem of controllability in an adaptive automatic trainable control system], *Mater. X Mezhdunarodnoy nauchno-tekhnicheskoy konferentsii "Tekhnologii razrabotki informatsionnykh sistem"* [Materials of the X International Scientific and Technical Conference. "Information Systems Development Technologies"], 2020.
6. *Zargaryan E.V., Zargaryan Y.A., Kapc I.V., Sakharova O.N., Kalyakina I.M and Dmitrieva I.A.* Method of estimating the Pareto-optimal solutions based on the usefulness. International Conference on Advances in Material Science and Technology - CAMSTech-2020, *IOP Conf. Series: Materials Science and Engineering,* 2020, Vol. 919 (2), pp. 022027 (1-8). DOI: 10.1088/1757-899X/919/2/022027.
7. *Zheng T., Chen C., Ren K.* Distribution ally adversarial attack. 2018. arXiv:1808.05537.
8. *Karlini N., Vagner D.* K otsenke nadezhnosti neyronnykh setey [To assess the reliability of neural networks], *Mater. simpoziuma IEEE 2017 goda po bezopasnosti i konfidentsial'nosti; 22–26 maya 2017 g. San-Khose, Kaliforniya, SShA, 2017* [Proceedings of the 2017 IEEE Symposium on Security and Privacy; May 22-26, 2017; San Jose, California, USA; 2017], pp. 39-57.
9. *Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A.* The limitations of deep learning in adversarial settings, *Proceedings of the 2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24;Saarbrucken, Germany, 2016,* pp. 372-87.
10. *Moosavi-Dezfooli S.M., Fawzi A., Frossard P.* DeepFool: a simple and accurate method to fool deep neural networks, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30. Las Vegas, NV, USA, 2016,* pp. 2574-82.
11. *Papernot N., McDaniel P., Goodfellow I.* Transferability in machine learning:from phenomena to black-box attacks using adversarial samples. 2016. arXiv:1605.07277.
12. *Liu Y., Chen X., Liu C., Song D.* Delving into transferable adversarial examples and black-box attacks. 2016. arXiv:1611.02770.
13. *Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards deep learning models resistant to adversarial attacks. 2017. arXiv: 1706.06083.
14. *Alamir KH.S., Zargaryan E.V., Zargaryan Yu.A.* Model' prognozirovaniya transportnogo potoka na osnove neyronnykh setey dlya predskazaniya trafika na dorogakh [A traffic flow prediction model based on neural networks for predicting traffic on the roads], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2021, No. 6 (223), pp. 124-132.

15. *Zheng T., Chen C., Yuan J., Li B., Ren K*. Point Cloud saliency maps. 2018. arXiv:1812.01687.
16. *Beloglazov D., Shapovalov I., Soloviev V., Zargaryan E.* The hybrid method of path planning in non-determined environments based on potential fields, *ARPN Journal of Engineering and Applied Sciences*, 2017, Vol. 12, No. 23, pp. 6762-6772.
17. *Атали А., Карлини Н., Вагнер Д.* Запутанные градиенты дают ложное чувство безопасности: обход защиты к состязательным примерам. 2018.arXiv:1802.00420.
18. *Zargarjan E.V., Zargarjan Ju.A., Finaev V.I.* Information support for the training of fuzzy production account balance in the conditions of incomplete data, *Innovative technologies and didactics in teaching (ITDT-2016): Collected papers*, 2016, pp. 128-138.
19. *Chen Yu., Sharma Yu., Chzhan Kh., I Dzh.*, Se S.Dzh. EAD: ataki elastichnoy seti na glubokie neyronnye seti na sostyazatel'nykh primerakh [EAD: Elastic network attacks on deep neural networks on adversarial examples], *Mater. tridtsat' vtoroy konferentsii AAAI po iskusstvennomu intellektu; 2–7 fevralya 2018 g. Novyy Orlean, Los-Andzheles, SShA, 2019* [Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence; February 2-7, 2018; New Orleans, Los Angeles, USA; 2019].
20. *Pushnina I.V.* Sistema upravleniya podvizhnym ob"ektom v usloviyakh neopredelennosti [The control system of a moving object in conditions of uncertainty], *Nauka i obrazovanie na rubezhe tysyacheletiy: Sb. nauchno-issledovatel'skikh rabot* [Science and Education at the turn of the Millennium. Collection of research papers]. Kislovodsk, 2018, pp. 65-74.
21. *Xiao C., Li B., Zhu J.Y., He W., Liu M., Song D.* Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.
22. *Ronneberger O., Fisher., Broks T.* U-net: svertochnye seti dlya segmentatsii biomeditsinskikh izobrazheniy [U-net: convolutional networks for segmentation of biomedical images], *Mater. Mezhdunarodnoy konferentsii po vychislitel'noy tekhnike meditsinskikh izobrazheniy i komp'yuternomu vmeshatel'stvu; 5–9 oktyabrya 2015 g. Myunkhen, Germaniya, 2015* [Proceedings of the International Conference on Medical Imaging Computing and Computer Intervention; October 5-9, 2015; Munich, Germany; 2015], pp. 234-41.
23. *Qi C.R., Su H., Mo K., Guibas L.J.* PointNet: deep learning on point sets for 3D classification and segmentation, *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI,USA, 2017*, pp. 652-60.
24. *Finaev V.I., Zargaryan Yu.A., Zargaryan E.V., Solov'ev V.V.* Formalizatsiya grupp podvizhnykh ob"ektov v usloviyakh neopredelennosti dlya vybora upravlyayushchikh resheniy [Formalization of groups of mobile objects under uncertainty for the choice of control solutions], *Informatizatsiya i svyaz'* [Informatization and Communication], 2016, No. 3, pp. 56-62.
25. *Bekhzadan V., Munir A.* Uyazvimost' glubokogo obucheniya s podkrepleniem k atakam s tsel'yu induktsii politiki [Vulnerability of deep learning with reinforcement to attacks for the purpose of policy induction], *Mater. Mezhdunarodnoy konferentsii po mashinnomu obucheniyu i intellektual'nomu analizu dannykh v raspoznavanii obrazov; 15–20 iyulya 2017 g. N'yu-York, N'yu-York, SShA, 2017* [Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition; July 15-20, 2017. New York, New York, USA, 2017], pp. 262-75.

Статью рекомендовал к опубликованию д.т.н. Ю.А. Кравченко.

Ауси Рим Мохаммед Худхейр – Южный федеральный университет; e-mail: ausi@sfedu.ru; г. Таганрог, Россия; кафедра систем автоматического управления; аспирант.

Заргарян Елена Валерьевна – e-mail: ezargaryan@sfedu.ru; кафедра систем автоматического управления; к.т.н.; доцент.

Заргарян Юрий Артурович – e-mail: yazargaryan@sfedu.ru; кафедра систем автоматического управления; к.т.н.; доцент.

Aussi Rim Mohammed Hedhair – Southern Federal University; e-mail: ausi@sfedu.ru; Taganrog, Russia; the department of automatic control systems; postgraduate student.

Zargaryan Elena Valerevna – e-mail: ezargaryan@sfedu.ru; the department of automatic control systems; cand. of eng. sc.; associate professor.

Zargaryan Yuri Arturovich – e-mail: yazargaryan@sfedu.ru; the department of automatic control systems; cand. of eng. sc.; associate professor.