УДК 004.891

Анализ алгоритмов, влияющих на эффективность методов машинного обучения

И.К. Радыгин [1], Л.И. Васильева[1], М.Р. Богданов [2]

[1]sensys96@gmail.com, [2] bogdanov_marat@mail.ru

[1] ФГБОУ ВО «Башкирский государственный педагогический университет» (БГПУ)
[2] ФГБОУ ВО «Уфимский университет науки и технологий» (УУНиТ)

Аннотация. Большинство актуальных исследований и работ в области машинного обучения ставят главной целью повышение точности распознавания, в то время как проблема состязательных атак на глубокие нейронные сети и их последствия пока не были в полной мере изучены. Работа посвящена анализу существующих атак, основанных на состязательных примерах на алгоритмы машинного обучения, применимые в цифровой образовательной среде. В данном материале анализируется исследовательский опыт по изучению эффективности атак, подготовленных с помощью алгоритма спроецированного градиентного спуска (*PGD*), алгоритма «глубокого обмана» (*DeepFool*), алгоритма Карлини – Вагнера (*CW*). Анализируются результаты атак обоих типов (по методам белого и черного ящика) на нейронные сети с архитектурами *InceptionV3*, *Densenet121*, *ResNet50*, *MobileNet* и *Xception*. Основной вывод работы заключается в том, что проблема состязательных атак актуальна для задач распознавания различных изображений, поскольку протестированные алгоритмы успешно атакуют обученные нейронные сети так, что их точность падает ниже 15 %.

Ключевые слова: машинное обучение, состязательные атаки, защита информации, цифровая образовательная среда.

ВВЕДЕНИЕ

Пандемия коронавируса внесла большие изменения в области образования. Ввиду невозможности проведения традиционных очных занятий и при наличии большого влияния цифровизации самого образовательного процесса, вопрос создания и развития цифровой образовательной среды стал востребованным. Анализ результатов применения дистанционного образования и попытки создания цифровой образовательной среды выявили ряд проблем, таких как: низкая информационная безопасность существующих решений, различные подходы к реализации сервисов, не позволяющих их логично встраивать в образовательный процесс, и малое количество верифицированного контента. В данной работе уделено внимание организации информационной безопасности.

В вопросах информационной безопасности при организации цифровой образовательной среды большую роль играет идентификация личности. Идентифицировать личность необходимо в различных случаях, в частности, при организации учебной деятельности, текущего контроля успеваемости, промежуточной аттестации, государственной итоговой аттестации, оказания учебной помощи и иных образовательных процедур.

Как правило, учетные данные обучающихся формируются автоматически случайным образом сгенерированным цифровым идентификатором, логин и пароль – это комбинации символов и цифр. В цифровой образовательной среде используется система идентификации личности обучающихся, получающих доступ к электронной информационно-образовательной

среде, позволяющая программными и (или) иными средствами осуществлять идентификацию личности обучающихся, а также обеспечивать контроль в сфере учета и хранения образовательных результатов. Идентификация личности обучающихся осуществляется путем использования электронной идентификации личности.

Также на практике применяется прием, когда в процессе сдачи экзамена с применением видеоконференцсвязи для идентификации личности экзаменующегося ему необходимо показать на камеру разворот паспорта или иной документ, удостоверяющий личность.

Велика вероятность того, что низкой информационной безопасностью существующих решений в организации работы цифровой образовательной среды могут воспользоваться злоумышленники, действующие в своих интересах, что безусловно может считаться уязвимостью и это является проблемой.

Для решения выявленной проблемы необходимо разработать безопасную и эффективную систему идентификации личности для цифровой образовательной среды. Мы предполагаем использование систем, основанных на распознавании человека по фотографии. Нейросети могут послужить качественным инструментом для решения поставленной задачи. Но в области алгоритмов машинного обучения в последнее время все больше обсуждается и исследуется вопрос состязательных примеров, поскольку это так же является прямой угрозой для цифровой безопасности, в частности, при создании цифровой образовательной среды. Ввиду этого требуется комплексный анализ существующих атак, основанных на состязательных примерах, на алгоритмы машинного обучения, предполагаемые к использованию в цифровой образовательной среде.

В данной работе анализ состязательных атак проводился на основе исследования задач классификации (распознавания) изображений, хотя подобные атаки могут быть проведены и в рамках области анализа и распознавания звука.

СОСТЯЗАТЕЛЬНЫЕ АТАКИ

Понятие состязательных атак. Под состязательной атакой понимается процесс, в результате которого атакуемый классификатор пытается предсказать класс изображения неправильно как с точки зрения человека, например, так и с точки зрения обученной и протестированной нейросетевой модели. Под ошибкой предсказания понимается такой случай, когда изображения, которые классификатор идентифицирует неверно, считаются, на первый взгляд, абсолютно допустимыми для соответственной предметной области, но при этом относятся к другому классу. При этом отличие таких ошибок от простых ошибок обобщения в том, что для данных изображений обычно есть практически не отличающееся от него парное изображение, причем верно распознаваемое сетью.

Рассмотрим создание атакующих изображений.

Генерация атакующих изображений. Развитие области состязательных атак начиналось с разработки и создания атакующих изображений (*adversarial examples*). Это такие изображения, изменения в которых не видны человеческому глазу, однако получившая на вход для обработки нейронная сеть вероятнее всего ошибется в предсказании. Работа состязательной атаки на основе атакующего изображения возможна благодаря использованию специального алгоритма, а также ряду других факторов.

Алгоритм генерации атакующего изображения состоит из трех основных пунктов, реализуемых последовательно друг за другом:

- выбор из предметной области атакуемой нейронной сети изображения;
- создание специального шумоподобного возмущения при использовании особого генерирующего алгоритма;
- наложение полученного шумоподобного возмущения на исходное изображение с применением попиксельного сложения.

В результате всех взаимодействий с изображением оно все равно остается узнаваемым для человека, однако нейронная сеть с большой вероятностью совершит ошибку при распознавании объекта, расположенного на изображении. Приведем формальное определение атакующего изображения.

Атакующие изображения.

Определение атакующего изображения. Пусть $x = \mathbb{R}^d$ – нормализованное входное изображение; $y : \mathbb{R}^d \to (0,1)^p$ – выход классификационной нейронной сети как функции от входного изображения с количеством классов p; $F : (0,1)^p \to \{1, \dots, p\}$ – решающая функция классификации.

Пусть $\varepsilon > 0$ – некоторое небольшое положительное число. Тогда ε-атакующим изображением называется такое изображение, для которого действует неравенство:

$$F\big(y(x)\big) \neq F\big(y(x^*)\big) \quad (1)$$

при выполнении ограничения:

$$\| x - x^* \| < \varepsilon. \quad (2)$$

В последнем уравнении в качестве нормы рассматривают, как правило, L_2 или L_∞; как обычно, допустима любая норма.

Первое уравнение означает, что результат работы нейронной сети в виде предсказания или классификации объекта у атакующего изображения отличен от подобного предсказания или непосредственно классификации объекта исходного изображения. Однако можно заметить, что такой результат возможен и в случае, если изображения относятся к разным классам. В целях демонстрации неестественности процесса вводится ограничение на малую величину ε для второго выражения. Величину этого параметра выбирает атакующая сторона, подбирая подходящее значение, чтобы атакующее изображение не слишком сильно отличалось от исходного. Обычно достаточно небольшого значения ε, чтобы классификатор совершил ошибку в работе. Этот параметр называется магнитудой модификации, поскольку условие (2) можно представить в виде:

$$\| \Delta x \| < \varepsilon, \quad x^* = x + \Delta x.$$

Таким образом, ε – магнитуда модификации изображения x. Следовательно, модификация изображения может быть выражена в различной степени, основываясь на различные значения ε. Практика работы с состязательными атаками на примере атакующих изображений показывает, что атакуемое изображение может быть изменено на столько, чтобы заставить классификатор ошибиться, но при этом изменения в изображении не заметны человеческому глазу.

Алгоритмы генерации атакующих изображений.

На сегодняшний день существует несколько алгоритмов генерации атакующих изображений. Рассмотрим общую идею работы этих алгоритмов.

Классификация алгоритмов генерации атакующих изображений. По классификации, основанной на информации, необходимой для работы, алгоритмы делятся на атаки по методу белого ящика и атаки по методу черного ящика.

Для проведения атаки по методу белого ящика необходимо знать конфигурацию сети, включая ее архитектуру и все параметры, полученные в результате обучения. Кроме того, требуется наличие оригинального изображения соответствующей предметной области для генерации самого́ атакующего изображения.

Для проведения атаки по методу черного ящика достаточно иметь доступ ко входу сети, куда подаются изображения, и к результатам предсказания, в то время как конфигурация сети может оставаться неизвестной. Как и обычно, еще нужна информация о предметной области распознаваемых изображений.

Кроме того, атаки разделяются на направленные и ненаправленные по принципу наличия определенного атакуемого класса или же его отсутствию.

Алгоритм спроецированного градиентного спуска (*projected gradient descent, PGD*).

Данный метод – это применение классической техники градиентного спуска с учетом ограниченности возмущения [1–3]. В качестве целевой функции можно рассмотреть компоненту

вероятностей y'. В таком случае минимизация функции приводит к снижению вероятности принадлежности атакующего изображения этому классу, а максимизация – к повышению.

В результате генерация направленного на класс t атакующего изображения по этому методу зависит от коэффициента обучения $\alpha > 0$, количества итераций $n \in \mathbb{N}$, магнитуды возмущений ε и определяется следующим образом:

$$x_{k+1} = clip_{x,\varepsilon}\big(x_k + \alpha \nabla y_t(x_k)\big),$$

где $x_0 = x, k \in [0, n-1], x^* = x_n$, а функция $clip_{x,\varepsilon}$ отделяет те элементы ее аргумента, которые отличаются от тех же элементов x более чем на ε. Генерация ненаправленного атакующего изображения по этому методу зависит от исходного класса m и определяется следующим образом:

$$x_{k+1} = clip_{x,\varepsilon}(x_k + \alpha \nabla y_m(x_k))$$

Поскольку на каждой итерации применяется функция $clip_{x,\varepsilon}$, то в результате работы алгоритма получится изображение x^*, автоматически удовлетворяющее ограничению, выраженному (2) для L_∞-нормы.

Алгоритм «глубокого обмана» (*DeepFool*). Данный алгоритм основан на идее линеаризации функции выхода нейронной сети и итеративном вычислении атакующего изображения как точки проекции на некоторую псевдополуплоскость [4; 5]. Любая итерация этого алгоритма задается формулой

$$x_{k+1} = x_k \frac{y_l(x_k) - y_m(x_k)}{\| w_l - w_m \|^2} (w_l - w_m),$$

где $w = \nabla y(x_k), x_0 = x; m$ – исходный класс объекта x, а l – класс, выбираемый на каждой итерации так, чтобы возмущения объекта были минимальны. В качестве атакующего изображения используется значение $x^* = (1+\eta)x_p$ конкретно взятой итерации, когда атака была успешной.

Алгоритм Карлини – Вагнера (*CW*). Данный алгоритм генерации атакующих изображений основан на модификации применения алгоритма *L-BFGS* к специально поставленной задаче оптимизации [6]. Для генерации направленных на класс t атакующих изображений авторами работы [7] формулируется следующая задача минимизации:

$$\left\| \frac{1}{2}(\tanh(w)+1) - x \right\| + cf_t\left(\frac{1}{2}(\tanh(w)+1)\right) \to min, w \in \mathbb{R},$$

где c – конфигурируемый параметр, а функция $f_t(x)$ задается формулой

$$f_t(x) = \max(\max_{k \neq t}\big(y_k(x)\big) - y_t(x), -\text{к}).$$

Функция f_t также зависит от параметра k, показывающего желаемую степень уверенности в классификации атакующего изображения. При $k=0$ достаточно найти успешное атакующее изображение, вероятность принадлежности целевому классу которого больше, чем вероятности принадлежности другим классам, но при увеличении к она повышается. Далее эта функция минимизируется известным оптимизатором *Adam*, и в результате после ограниченного количества итераций получается атакующее изображение.

АНАЛИЗ РЕЗУЛЬТАТОВ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Построенные задачи классификации. Анализ эффективности атак, подготовленных с помощью алгоритма спроецированного градиентного спуска (*PGD*), алгоритма «глубокого бмана» (*DeepFool*), алгоритма Карлини – Вагнера (*CW*), приведен на результатах эксперимента, проведенного на пяти наборах биомедицинских изображений [8]. Конфигурирование данных позволило более детально изучить эффект состязательных атак. Опишем проведенные эксперименты.

Исследуя атаки по методу **белого ящика,** выполнялись следующие действия: проводилась атака на нейронную сеть, предназначенную для задачи классификации, в результате чего генерировалось атакующее изображение для оценки работы алгоритмов. Далее атакующее изображение подавалось на вход той же сети и полученные оценки принадлежности сохранялись для анализа.

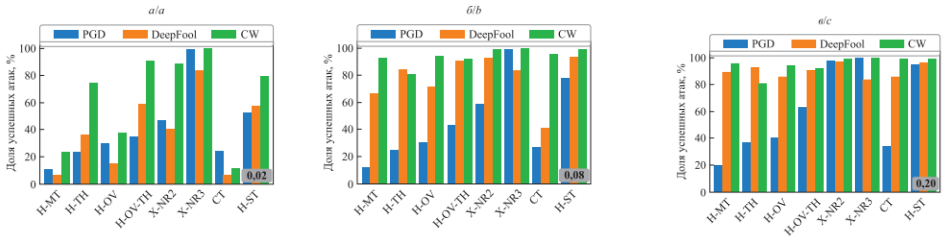

Рис. 1. Доля успешных атак для использованных наборов данных при ограничении L_∞-нормы возмущения в 0,02; 0,08 и 0,20

При использовании L_∞-нормы алгоритм *CW* показал себя лучше почти во всех случаях.

Из представленных данных видно, что в целом алгоритмы *DeepFool* и *CW* близки по эффективности: для ε, равного 1,0 и 2,0, в четырех из восьми случаев доли успешных атак этих алгоритмов почти равны, в оставшихся случаях однозначного фаворита не наблюдается.

Анализируя атаки по методу **черного ящика,** необходимо обратить внимание, что атаки совершаются без знания архитектуры нейронной сети. Поэтому требовался совершенно иной принцип проведения атак. Из выборки изображений в предметной области обучалась имитирующая сеть, проводилась атака по методу белого ящика, в результате чего получалось атакующее изображение, а потом проводилась сама атака и фиксировались результаты. Для проведения атак по данной методике в настоящей работе рассматриваются пять архитектур глубоких нейронных сетей – *InceptionV3, DenseNet121, ResNet50, MobileNet* и *Xception.*

После выполнения описанной ранее последовательности действий для каждой выбранной задачи классификации было получено 25 наборов результатов предсказаний, включая 20 наборов для каждой пары «целевая сеть – инструментальная сеть» и 5 наборов для пар, сети в которых совпадают. Следует отметить, что в последнем случае получались атаки по методу белого ящика, поскольку атаковалась та же самая сеть, для которой генерировались атакующие изображения. По полученным таким образом данным вычислялась доля успешных атак.

ЗАКЛЮЧЕНИЕ

В рамках проводимой работы был проведен анализ существующих атак, основанных на состязательных примерах, на алгоритмы машинного обучения, предполагаемые к использованию в цифровой образовательной среде. По результатам анализа сделан ряд выводов.

1. В новой цифровой образовательной среде особую актуальность приобретает проблема состязательных атак, поскольку после тестов алгоритмы успешно атакуют обученные нейронные сети, в результате чего их точность становится ниже 15%.

2. Алгоритм спроецированного градиентного спуска (*PGD*) при тех же величинах злонамеренных пертурбаций изображения показал себя как самый неэффективный в сравнении с алгоритмом «глубокого обмана» (*DeepFool*) и алгоритмом Карлини – Вагнера (*CW*).

3. В трех из четырех задач распознавания изображений атаки по методу черного ящика с использованием алгоритма *PGD* показали низкую эффективность.

СПИСОК ЛИТЕРАТУРЫ

1. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv: 1706.06083v4 [Preprint]. 2017 [cited 2020 August 27]: [28 p.]. Available from: https://arxiv.org/abs/1706.06083. Журнал Белорусского государственного университета. Математика. Информатика. 2020; 3:60–72 Journal of the Belarusian State University. Mathematics and Informatics. 2020; 3:60–72.

2. Ozdag M. Adversarial attacks and defenses against deep neural networks: a survey. Procedia Computer Science.2018;140: 152–161. DOI: 10.1016/j.procs.2018.10.315.

3. Wang H, Yu C-N. A direct approach to robust deep learning using adversarial networks. arXiv:1905.09591v1 [Preprint]. 2019 [cited 2020 August 27]: [15 p.]. Available from: https://arxiv.org/abs/1905.09591.

4. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. arXiv:1704.01155v2 [Preprint]. 2017 [cited 2020 August 27]: [15 p.]. Available from: https://arxiv.org/abs/1704.01155.

5. Moosavi-Dezfooli S-M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. arXiv: 1511.04599v3 [Preprint]. 2015 [cited 2020 August 27]: [9 p.]. Available from: https://arxiv.org/abs/1511.04599.

6. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: 2 nd International conference on learning representations; 2014 April 14–16; Banff, Canada.Banff: Springer; 2014. p. 1–10.

7. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy; 2017 June 26; San Jose, CA, USA.[S. l.]: IEEE; 2017. p. 39–57. DOI: 10.1109/SP.2017.49.

8. Войнов ДМ, Ковалев ВА. Устойчивость нейронных сетей к состязательным атакам при распознавании биомедицинских изображений. Журнал Белорусского государственного университета. Математика. Информатика.2020;3:60–72. https://doi.org/10.33581/2520-6508-2020-3-60-72.

ОБ АВТОРАХ

РАДЫГИН Илья Константинович, аспирант БГПУ им. М.Акмуллы.
ВАСИЛЬЕВА Лидия Ильясовна, к.т.н., зав.кафедрой ИТ, БГПУ им. М.Акмуллы.
БОГДАНОВ Марат Робертович, к.б.н., доцент, УГАТУ.

METADATA

Title: Analysis of algorithms affecting the performance of machine learning algorithms.

Affiliation: Bashkir State Pedagogical University named after M. Akmulla (BSPU named after M. Akmulla), Russia.

Email: [1]sensys96@gmail.com, [2] bogdanov_marat@mail.ru.

Language: Russian.

Source: Molodezhnyj Vestnik UGATU (scientific journal of Ufa University of Science and Technology), no. 1(27), pp. 94-99, 2023. ISSN 2225-9309 (Print).

Abstract: Most of the current research and work in the field of machine learning is aimed at improving the accuracy of recognition, while the problem of adversarial attacks on deep neural networks and their consequences have not yet been fully studied. The work is devoted to the analysis of existing attacks based on adversarial examples on machine learning algorithms applicable in the digital educational environment. This article analyzes the research experience in studying the effectiveness of attacks prepared using the algorithm of projected gradient descent (PGD), the DeepFool algorithm, the Carlini–Wagner algorithm (CW). The results of attacks of both types (using white and black box methods) on neural networks with the architectures InceptionV3, Densenet121, ResNet50, MobileNet and Xception are analyzed. The main conclusion of the work is that the problem of adversarial attacks is relevant for the tasks of recognizing various images, since the tested algorithms successfully attack trained neural networks so that their accuracy drops below 15%.

Key words: machine learning, adversarial attacks, information security, digital educational environment.

About authors:

RADYGIN, Ilya Konstantinovich, postgraduate student of the BSPU named after M.Akmulla.

VASILYEVA, Lidiya Ilyasovna, Ph.D. in Technical sciences of the BSPU named after M.Akmulla

BOGDANOV, Marat Robertovich, Ph.D. in Biology, Ufa state aviation technical University