УДК 004 DOI: 10.18698/2541-8009-2019-02-443

ИССЛЕДОВАНИЕ И СТАТИСТИЧЕСКИЙ АНАЛИЗ АТАК НА НЕЙРОННЫЕ СЕТИ В ЗАДАЧАХ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Л.И. Капитонова kapitonova@bmstu.ru
SPIN-код: 1664-6050

А.А. Ушакова ushkova.anna@icloud.com
SPIN-код: 7948-9609

Н.А. Шална shalnene@mail.ru
SPIN-код: 8609-9258

А.А. Сторожева nastya-stor28@yandex.ru
SPIN-код: 9869-8831

МГТУ им. Н.Э. Баумана, Москва, Российская Федерация

Аннотация

Рассмотрены различные виды атак на нейронные сети в задачах компьютерного зрения, проведены их сравнение и классификация. Представлен анализ таких видов атак из класса «враждебных изображений», как атаки, основанные на градиентном методе, и пиксельные атаки. Проанализирована статистика использования наборов данных для обучения нейросети, имеющихся в открытом доступе. На ее основе получена зависимость вероятности успешного проведения атаки для наборов данных, имеющихся в открытом доступе. Выявлены и проанализированы наиболее эффективные методы защиты от различных видов атак на нейронные сети.

Ключевые слова

Нейронная сеть, датасет, враждебная атака, защита от атак, уязвимости нейронных сетей, градиентный метод, пиксельные атаки, администратор безопасности

Поступила в редакцию 17.01.2019

Введение. Глубокое машинное обучение (deep learning) используется в различных областях знаний, таких как медицина, системы компьютерного зрения, системы автоматизации производства и др. Однако с ростом популярности какого-либо направления науки и техники активность злоумышленников увеличивается соразмерно этому росту [1]. Так, на сегодняшний день исследовательское сообщество продемонстрировало, что методы машинного обучения (в том числе нейронные сети с учителем) уязвимы для атак со стороны противника.

Одним из наиболее популярных классов атак на нейронные сети с учителем является класс атак «враждебных изображений» (adversarial image). Принцип атаки данного класса состоит в модификации исходного изображения таким образом, что изменения почти не заметны человеческому глазу, однако весьма ощутимы для нейронной сети, работающей с таким изображением. Мерой модификации обычно является норма, которая измеряет максимум абсолютного изменения в одном пикселе.

Класс атак «враждебных изображений» подразделяют на подклассы по следующим принципам:

1) по наличию у злоумышленника доступа к исходным параметрам модели:

– white box — злоумышленник имеет доступ;

– black box — злоумышленник не имеет доступа;

2) по способу создания «враждебного изображения»:

– нецелевые — злоумышленник относит «враждебное изображение» в любой класс, независимо от класса истинного изображения;

– целевые — злоумышленник относит «враждебное изображение» в определенный класс.

Анализ фальсификации существующих «датасетов». Среди существующих на сегодняшний день открытых наборов данных («датасетов») для обучения нейронных сетей можно выделить CIFAR, MNIST, ImageNet и др. Однако использование данных «датасетов» для обучения сети не обезопасит ее от возможных ошибок при работе. Эти ошибки могут быть вызваны тем, что данные, хранящиеся в сети, неустойчивы к различного вида атакам на нейронные сети [2–4]. В качестве подтверждения в работе был проведен анализ различных исследований в области уязвимостей нейронных сетей и предложена модель обеспечения безопасного использования методов машинного обучения. Исследования группы Moosavi в 2016 г. показали, что применение атаки «враждебного изображения», основанной на градиентном методе, позволяет увеличить уровень ошибок нейронной сети, обученной на «датасете» ImageNet, до 80 % [5]. Команда исследователей из Adrien Chan-Hon-Tong установила, что использование алгоритма добавления шума в исходное изображение позволило достичь уровня ошибки 77 % для таких открытых наборов данных, как CIFAR10 и VGG6 [6]. Анализ устойчивости современных классификаторов глубоких нейронных сетей к атаке «враждебного изображения», проводимый на открытом наборе ILSVRC 2012, показал, что для всех сетей достигается очень высокая вероятность ошибочной классификации. Значения вероятности реализации ошибочной классификации для различных нейронных сетей, которые обучались на открытых «датасетах», приведены ниже:

Сеть ……	CaffeNet	VGG-F	VGG-16	VGG-19	GoogLeNet	ResNet-152
Вероятность, % …….	93,3	93,7	78,7	77,8	78,9	84,0

Виды атак «враждебного изображения». Атаки, основанные на градиентном методе (FGSM). Наиболее успешными являются атаки, основанные на градиентном методе. Суть атак данного класса состоит в том, что злоумышленники модифицируют исходное изображение в направлении градиента функции потерь относительно входного изображения. Существует два основных подхода к реализации таких атак: однократный метод (FGSM и T-FGSM), при котором атакующий делает один шаг в направлении градиента, и итерационный метод (I-FGSM), при котором предпринимается несколько шагов.

Для класса атак white box однократные методы атак имеют более низкие показатели успеха по сравнению с итерационными методами. Для класса black box однократные методы оказываются более эффективными. Это вызвано тем, что итерационные методы, как правило, переобучают определенную модель.

Дополнительно выделяют такой тип атак, как выигрышные атаки с импульсом. Этот метод универсален — эффективность основанных на нем атак не зависит от того, является ли атака целевой или нецелевой.

Общие средства защиты от атак градиентного метода. Наиболее распространенная защита состоит во внедрении состязательных образов для обучения более надежной сети, которые генерируются с использованием целевой модели. Было показано, что этот подход имеет некоторые ограничения — в частности, этот вид защиты менее эффективен против атак black box, в которых состязательные образы генерируются с использованием другой модели [7]. Это связано с градиентной маскировкой, т. е. в такого рода защите вводится возмущение в градиентах, что делает атаки white box менее эффективными.

Пиксельные атаки. Генерирование враждебных изображений — это проблема оптимизации с ограничениями. Предположим, что входное изображение может быть представлено вектором, в котором каждый скалярный элемент представляет один пиксель [8].

Пусть f — функция классификатора, которая имеет размерность n, $x=(x_1,...,x_n)$ — исходное естественное изображение, правильно классифицированное как класс t. Следовательно, вероятность принадлежности изображения x к классу t равна $f_t(x)$. Вектор $e(x)=(e_1,...,e_n)$ является аддитивным враждебным искажением, целевым классом adv и ограничением максимальной модификации L. Заметим, что L всегда измеряется длиной вектора $e(x)$. Целью злоумышленника в случае *целевых пиксельных атак* является поиск оптимизированного решения $e(x)^*$. Данные утверждения можно представить в математическом виде следующим образом:

$$\frac{\text{maximize}}{e(x)^*} f_{adv}(x+e(x));$$
$$\text{subject to } \|e(x)\| \le L.$$

Таким образом, задача сводится к тому, чтобы найти два значения: размер возмущения (L) и соответствующую этому возмущению силу модификации для каждого измерения (d):

$$\frac{\text{maximize}}{e(x)^*} f_{adv}(x+e(x));$$
$$\text{subject to } \|e(x)\|_0 \le d.$$

В случае однопиксельной атаки сила модификации $d=1$. Фактически пиксельные атаки предполагают произвольную модификацию изображения по выбранному направлению из n возможных направлений.

Статистика реализуемых пиксельных атак. Задача исследования состояла в сравнении целевой и нецелевой однопиксельной атаки с тремя- и пятипиксельными атаками. Стоит отметить, что в данном исследовании эффективность нецелевой атаки оценивается на основе результатов целевой атаки.

Атакам подвергались следующие нейронные сети: All convolution network (AllConv), Network in Network (NiN). В качестве целевых классификаторов изображений использовался набор данных cifar-10. Эффективность проведения пиксельных атак представлена ниже:

Тип атаки ……..	однопиксельная	трехпиксельная	пятипиксельная
Успешность атаки, % ……...	79,40	79,17	77,09

Обобщая результаты, полученные в ходе исследования, выделим следующие:

– уровень успешности целевых однопиксельных атак на два типа сетей показывает эффективность использования;

– при увеличении порядка атаки (числа пикселей) количество целевых классов, которые могут быть достигнуты, значительно увеличивается;

– вероятности реализации успешных целевых атак составляют 79,40 % — однопиксельная, 79,17 % — трехпиксельная и 77,09 % — пятипиксельная;

– чем больше пикселей подвергается изменению, тем сложнее классифицировать изображения. Вероятность отнесения изображения к «своему» классу снижается и примерно выравнивается с вероятностями распределения в остальные классы;

– судя по вероятности ошибочной классификации, можно сделать вывод о том, что рассматриваемые сети AllConv и NiN уязвимы для этого типа атак.

Анализ методов защиты от атак на нейронные сети. Самый простой и основополагающий способ обезопасить применение нейронных сетей — не применять открытые наборы данных для обучения. В связи с этим специалисты советуют накапливать и использовать для обучения сети собственные наборы данных. Однако зачастую это требование выполнить невозможно, поэтому ниже будут рассмотрены другие методы, которые позволяют противостоять атакам, направленным на глубокое машинное обучение, в частности, атакам «враждебного изображения».

Исследовательское сообщество в области нейронных сетей выделяет две стратегии противодействия атакам на нейронные сети. Согласно исследованиям команды Айана Гудфеллоу, существуют такие способы защиты нейронных сетей, как состязательная подготовка (adversarial training) и оборонительная «дистилляция» (defensive distillation), которые позволяют сделать модель машинного обучения более надежной в отношении атак, основанных на использовании возмущений [9, 10].

Стратегия состязательной подготовки впервые была продемонстрирована командой Szegedy, однако практических реализаций на данный момент не так много. Это обусловлено тем, что для ее реализации необходимы высокие вычис-

лительные методы. Суть состязательной подготовки состоит в том, чтобы в процессе обучения обучить нейронную сеть правильно классифицировать «враждебное изображение», которое подвергалось неоднократным модификациям.

Группа исследователей Ian Goodfellow продемонстрировала упрощенный способ генерирования «враждебных изображений» с помощью метода быстрых градиентных знаков. Данный метод позволяет сократить требуемые вычислительные мощности для реализации стратегии состязательной подготовки. Оборонительная «дистилляция» сглаживает поверхность решения модели в состязательных направлениях, используемых противником. Дистилляция — это обучающая процедура, при которой одна модель обучается предсказывать вероятности, полученные другой моделью, которая была обучена ранее. Дистилляция была впервые введена группой исследователей Hinton, использующей в качестве цели небольшую модель для имитации серьезной модели.

Оборонительная «дистилляция», которая нас интересует, имеет другую цель — просто сделать ответы окончательной модели менее строгими, поэтому она работает, даже если обе модели имеют одинаковый размер. Может показаться нелогичным обучать одну модель предсказывать выходные данные другой модели с той же архитектурой. Однако причина успешности стратегии состоит в том, что первая модель подвергается обучению с «жесткими» метками, а для обучения второй модели используются «нестрогие» метки. В результате вторая («дистиллированная») модель более устойчива к атакам градиентного метода и атакам на основе якобиан-преобразования.

Ведущие исследователи также утверждают, что наличие человеческого контроля позволяет повысить уровень защиты предприятия, использующего глубокое машинное обучение [11]. Другими словами, если «линия защиты» предприятия включает в себя сервис машинного обучения в роли средства защиты, то рекомендуется дополнить его постоянным мониторингом со стороны человека.

Говоря об атаках «враждебного изображения», выделяют два наиболее популярных метода защиты. Суть первого метода состоит во внедрении состязательных образов для обучения более надежной сети, которые генерируются с использованием целевой модели. Однако этот подход имеет некоторые ограничения — в частности, он менее эффективен против атак black box, в которых состязательные образы генерируются с использованием другой модели. Второй метод предполагает использование шумоподавляющей сети (DUNET), которая похожа на автокодер с шумоподавлением и использует сетевую структуру, имеющую прямые соединения между соответствующими уровнями в кодере и декодере. В результате применения вышеуказанных методов защиты от атаки «враждебного изображения» удалось снизить эффективность целевых атак до 20 %, а нецелевых — до 35 % [12].

Отметим, что рекомендуется проводить такое обучение нейронной сети, при котором возможна лишь атака класса black box. Это делается для того, чтобы поведение сети могло быть определено атакующим только по итогу наблюдения за результатом работы нейронной сети.

Заключение. В результате представленного в данной работе анализа можно выделить следующие общие положения:

1. Существующие на сегодняшний день атаки на нейронные сети имеют показатели успешности ниже 60 %.

2. Атаки градиентного метода в большей степени зависят от наличия у злоумышленника доступа к исходной модели. Так, однократные методы градиентных атак оказываются более эффективными для класса black box, а итерационный метод — white box.

3. Также для атак, основанных на градиентном методе, важно наличие цели: целевые атаки имеют вероятность успеха 40 %, тогда как нецелевые атаки оказываются эффективны.

4. Использование методов защиты от атак градиентного метода позволяет снизить их эффективность до 35 % в случае нецелевых атак и до 20 % в случае целевых атак.

5. В случае нецелевой атаки наибольший процент реализации имеют атаки градиентного метода — 80…94 %. Эффективность пиксельных атак в данном случае составляет 72 %.

Отметим, что существующие типы атак постоянно подвергаются улучшениям со стороны отдельных исследователей как из России, так и из зарубежных стран. Они предлагают новые способы построения алгоритма атаки с целью повышения ее успешности. Для предотвращения атак необходимо обезопасить нейронную сеть предприятия от внутренних и внешних угроз и инцидентов. Администратор безопасности предприятия, взаимодействуя с базой данных, должен управлять протоколом доступа в подсистеме идентификации и аутентификации, а также следить за работоспособностью подсистемы контроля целостности сети. Администратору безопасности необходимо отслеживать информацию об уязвимостях нейронной сети (т. е. периодически тестировать ее) и своевременно принимать меры по их устранению. Также при наличии нарушения он обязан локализовать это нарушение, установить причину возникновения, принять меры по ликвидации последствий и оценить ущерб.

Литература

[1] Комашинский В.И., Смирнов Д.А. Нейронные сети и их применение в системах управления и связи. М., Горячая линия-Телеком, 2002.

[2] Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. М., Горячая линия-Телеком, 2001.

[3] Рутковская Д., Пилиньский М., Рутковский Л. Нейронный сети, генетические алгоритмы и нечеткие системы. М., Горячая линия-Телеком, 2006.

[4] Галушкина А.И. Теория нейронных сетей. Москва, ИПРЖР, 2000.

[5] Gomes J. Adversarial attacks and defences for convolutional neural networks. *medium.com*: веб-сайт. URL: https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7 (дата обращения: 03.06.2018).

[6] Nguyen A., Yosinski J., Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *Proc. IEEE Conf. CVPR*, 2015. DOI: 10.1109/CVPR.2015.7298640 URL: https://ieeexplore.ieee.org/document/7298640

[7] Chan-Hon-Tong A. On the simplicity to produce falsified deep learning results. *hal.archives-ouvertes.fr*: веб-сайт. URL: https://hal.archives-ouvertes.fr/hal-01676691v1 (дата обращения: 29.06.2018).

[8] Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks. *arxiv.org*: веб-сайт. URL: https://arxiv.org/pdf/1710.08864.pdf (дата обращения: 07.07.2018).

[9] Moosavi-Dezfooli S.M., Fawzi O., Fawzi A., et al. Universal adversarial perturbations. *arxiv.org*: веб-сайт. URL: https://arxiv.org/pdf/1610.08401.pdf (дата обращения: 28.07.2018).

[10] Huang S., Papernot N., Goodfellow I., et al. Adversarial attacks on neural network policies. *arxiv.org*: веб-сайт. URL: https://arxiv.org/abs/1702.02284 (дата обращения: 28.07.2018).

[11] Papernot N., McDaniel P., Goodfellow I., et al. Practical black-box attacks against machine learning. *arxiv.org*: веб-сайт. URL: https://arxiv.org/abs/1602.02697 (дата обращения: 28.07.2018).

[12] Papernot N., McDaniel P. Extending defensive distillation. *arxiv.org*: веб-сайт. URL: https://arxiv.org/abs/1705.05264 (дата обращения: 28.07.2018).

Капитонова Людмила Ивановна — студентка кафедры «Защита информации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Ушакова Анна Андреевна — студентка кафедры «Защита информации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Шална Никита Андреевич — студент кафедры «Защита информации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

Сторожева Анастасия Андреевна — студентка кафедры «Защита информации», МГТУ им. Н.Э. Баумана, Москва, Российская Федерация.

RESEARCH AND STATISTICAL ANALYSIS OF ATTACKS ON NEURAL NETWORKS IN TECHNICAL VISION TASKS

L.I. Kapitonova kapitonova@bmstu.ru
SPIN-code: 1664-6050

A.A. Ushakova ushkova.anna@icloud.com
SPIN-code: 7948-9609

N.A. Shalna shalnene@mail.ru
SPIN-code: 8609-9258

A.A. Storozheva nastya-stor28@yandex.ru
SPIN-code: 9869-8831

Bauman Moscow State Technical University, Moscow, Russian Federation

Abstract

Different types of attacks on neural networks in technical vision tasks are considered, compared and classified. The analysis of such types of attacks from the class of "hostile images", such as attacks based on the gradient method and pixel attacks, is presented. The statistics of using data sets for training the neural network, available in the public domain, is analyzed. Based on it, the dependence of the probability of a successful attack for data sets, that are publicly available, is obtained. The most effective methods of protection against various types of attacks on neural networks are identified and analyzed.

Keywords

Neural network, dataset, hostile attack, protection against attacks, neural network vulnerabilities, gradient method, pixel attacks, security administrator

Received 17.01.2019

References

[1] Komashinskiy V.I., Smirnov D.A. Neyronnye seti i ikh primenenie v sistemakh upravleniya i svyazi [Neural networks and using them in communication control systems]. Moscow, Goryachaya liniya-Telekom Publ., 2002 (in Russ.).

[2] Kruglov V.V., Borisov V.V. Iskusstvennye neyronnye seti. Teoriya i praktika [Artificial neural networks. Theory and practice]. Moscow, Goryachaya liniya-Telekom Publ., 2001 (in Russ.).

[3] Rutkovskaya D., Pilin'skiy Moscow, Rutkovskiy L. Neyronnyy seti, geneticheskie algoritmy i nechetkie sistemy [Neural networks, genetic algorithms and fuzzy systems]. Moscow, Goryachaya liniya-Telekom Publ., 2006 (in Russ.).

[4] Galushkina A.I. Teoriya neyronnykh setey [Neural networks theory]. Moscow, IPRZhR Publ., 2000 (in Russ.).

[5] Gomes J. Adversarial attacks and defences for convolutional neural networks. *medium.com*: website. URL: https://medium.com/onfido-tech/adversarial-attacks-and-defences-for-convolutional-neural-networks-66915ece52e7 (accessed 03.06.2018).

[6] Nguyen A., Yosinski J., Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *Proc. IEEE Conf. CVPR*, 2015. DOI: 10.1109/CVPR.2015.7298640 URL: https://ieeexplore.ieee.org/document/7298640

[7] Chan-Hon-Tong A. On the simplicity to produce falsified deep learning results. *hal.archives-ouvertes.fr*: website. URL: https://hal.archives-ouvertes.fr/hal-01676691v1 (accessed 29.06.2018).
[8] Su J., Vargas D.V., Sakurai K. One pixel attack for fooling deep neural networks. *arxiv.org*: website. URL: https://arxiv.org/pdf/1710.08864.pdf (accessed 07.07.2018).
[9] Moosavi-Dezfooli S.M., Fawzi O., Fawzi A., et al. Universal adversarial perturbations. *arxiv.org:* website. URL: https://arxiv.org/pdf/1610.08401.pdf (accessed 28.07.2018).
[10] Huang S., Papernot N., Goodfellow I., et al. Adversarial attacks on neural network policies. *arxiv.org*: website. URL: https://arxiv.org/abs/1702.02284 (accessed 28.07.2018).
[11] Papernot N., McDaniel P., Goodfellow I., et al. Practical black-box attacks against machine learning. *arxiv.org*: website. URL: https://arxiv.org/abs/1602.02697 (accessed 28.07.2018).
[12] Papernot N., McDaniel P. Extending defensive distillation. *arxiv.org*: website. URL: https://arxiv.org/abs/1705.05264 (accessed 28.07.2018).

Kapitonova L.I. — Student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Ushakova A.A. — Student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Shalna N.A. — Student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.

Storozheva A.A. — Student, Department of Information Security, Bauman Moscow State Technical University, Moscow, Russian Federation.