МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М. В. ЛОМОНОСОВА
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ
КАФЕДРА ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Пак Юлия Руслановна

Обнаружение состязательных примеров в нейронных сетях

КУРСОВАЯ РАБОТА

Научный руководитель:
ассистент
Терёхина Ирина Юрьевна

Москва, 2023

Оглавление

Аннотация

Состязательными примерами называют данные, к которым путем некоторой процедуры добавлено возмущение, приводящее к неправильной классификации. Применение состязательных примеров является одним из самых распространенных типов атаки на системы машинного обучения. Такие атаки универсальны и практичны. Поэтому для улучшения безопасности систем машинного обучения необходимо повышать их устойчивость к подобным атакам.

В данной работе рассматривается задача обнаружения состязательных примеров в нейронных сетях. Она заключается в разработке методов детекции состязательных примеров в данных, подающихся на вход классификаторам на нейронных сетях. Описаны некоторые способы детекции состязательных примеров, продемонстрированы результаты их практического применения. На основе полученных результатов проведен сравнительный анализ методов, даны оценки их эффективности и универсальности.

1. Введение

Нейронные сети широко применяются для решения различных задач, таких как распознавание образов, прогнозирование и принятие решений. В частности, классификаторы на нейронных сетях демонстрируют точность, сравнимую с человеческой. Тем не менее, добавив к входным данным неощутимо малое возмущение, возможно заставить нейросеть выдать некорректный результат. Примеры, к которым путем некоторой процедуры добавлен малый вектор, приводящий к неправильной классификации, называются состязательными.

Как правило, выделяют три типа состязательных атак на системы машинного обучения [1]:

1) *Отравление данных.* Злоумышленник заражает обучающие данные, что впоследствии приводит к нарушению обучения модели и снижает качество ее работы на тестовом этапе и во время практического использования. Например, злоумышленник может внедрить вредоносные примеры в обучающую выборку классификатора.

2) *Атаки уклонения.* Злоумышленник манипулирует входными данными с целью обмануть раннее обученную модель на этапе ее практического использования.

3) *Кража модели.* Злоумышленник, отправляя запросы к модели, может воссоздать ее примерный аналог или выделить элементы выборки, на которой она была обучена.

Данная работа сосредоточена на атаках второго типа — атаках уклонения. Атаки данного типа наиболее практичны, так как осуществляются во время эксплуатации модели. В частности, рассматриваются атаки, предполагающие построение состязательных примеров.

В зависимости от уровня осведомленности злоумышленника об атакуемой модели можно выделить следующие категории состязательных атак [1]:

1) *Атаки белого ящика (полная осведомленность).* Атакующему известны: обучающие данные, множество признаков, алгоритм обучения, обучаемые параметры, гиперпараметры целевой модели.

2) *Атаки серого ящика (частичная осведомленность).* Чаще всего предполагается, что атакующему известно множество признаков, а также алгоритм обучения модели. Злоумышленник не имеет доступа к параметрам модели и к обучающим данным. Тем не менее существует возможность получить доступ к некоторому подмножеству обучающих данных, или использовать данные схожего типа. Атакующий путем прямого взаимодействия с моделью может получить метки для этих данных.

3) *Атаки черного ящика (нулевая осведомленность).* Атакующему не доступны никакие сведения о внутреннем строении модели. Однако в практическом сценарии злоумышленнику, вероятнее всего, известно, для какой задачи построена модель. Исходя из этого, атакующий может предположить, какие данные были использованы для обучения модели, и затем путем прямого взаимодействия с моделью получить метки для своей выборки.

Некоторые атаки уклонения, рассматриваемые в работе, соответствуют критериям сценария «белого ящика». Однако, исследования показали, что состязательные примеры, разработанные для атаки на определенную модель, обученную на некоторой выборке, обобщаются на модели с другими гиперпараметрами, а также на модели, обученные на непересекающихся выборках [2, 3]. Из этого можно сделать вывод, что уязвимость к состязательным атакам является общим для нейронных сетей феноменом. Наконец, эксперименты показали, что состязательные примеры, подвергнутые нетривиальным преобразованиям (например, обработке фотокамерой) сохраняют эффективность [4]. Все эти факторы указывают на практическую возможность осуществления состязательных атак в сценариях «черного ящика».

Таким образом, состязательные примеры представляют реальную угрозу безопасности систем машинного обучения. Ведутся исследования в области защиты от атак данного типа. На сегодняшний день можно выделить следующие основные стратегии защиты:

- Повышение робастности модели путем улучшения ее архитектуры и процедуры обучения.
- Разработка методов, позволяющих отличать истинные примеры от состязательных.

Данная работа сфокусирована на втором подходе. Рассматриваются методы, позволяющие обнаружить состязательные примеры в данных, подающихся на вход классификаторам на нейронных сетях. Дается оценка их эффективности и универсальности - то есть, их применимости к детекции состязательных примеров, полученных различными способами.

1.1. Постановка задачи

1) Изучить предметную область. Проанализировать существующие на данный момент результаты исследований относительно способов построения состязательных примеров, причин их существования и методов защиты от атак, основанных на использовании состязательных примеров.
2) Построить, обучить и протестировать классификатор на нейронной сети.
3) Построить различными способами состязательные примеры. С их помощью атаковать построенный на предыдущем шаге классификатор. Оценить эффективность атак.
4) Реализовать методы обнаружения состязательных примеров. Применить их к примерам, полученным на предыдущем шаге. Провести сравнительный анализ методов, оценить их эффективность и универсальность.

1.2. Цель работы

На основе проведенного анализа выбрать наиболее эффективный и универсальный метод обнаружения состязательных примеров.

2. Обзор предметной области

Рост вычислительной мощности компьютеров позволил применить нейронные сети к широкому кругу задач. Нейронные сети продемонстрировали впечатляющие результаты в обработке естественного языка [5], беспилотных автомобилях [6], системах рекомендаций [7] и медицине [8].

Данная работа сфокусирована на применении нейронных сетей к задаче классификации изображений - одной из базовых задач компьютерного зрения. Одним из наиболее распространенных и эффективных решений для данной задачи считаются свёрточные нейронные сети (СНС) [9]. Было предложено множество архитектур СНС. Одними из наиболее популярных можно назвать LeNet [10], AlexNet [11], VGG [12], и ResNet [13]. Модели с данными архитектурами очень распространены на практике, и поэтому чаще всего являются целевыми в сценариях проведения состязательных атак. По этой причине именно такие модели используются при проведении исследований методов защиты.

В машинном обучении различают дискриминативный и генеративный типы моделирования. Дискриминативные модели учатся моделировать условное распределение метки класса y относительно признаков x как P(Y|X). Другими словами, такие модели моделируют границы принятия решений между классами. Генеративные модели обучаются приближать вероятностное распределение p(x), т. е. распределение вероятности обнаружения x в обучающей выборке. Нейронные сети являются дискриминативными моделями. Известно, что такие модели уязвимы не только к состязательным примерам, но и так называемым «fooling images» [14]. Примеры такого типа представляют собой изображения, которые человек не может распознать и отнести к какому-либо из рассматриваемых в задаче классов. Тем не менее модель с высокой степенью уверенности (>99.99%) относит их к одному из классов. Авторы статьи объясняют это дискриминативной природой глубоких нейронных сетей. Поскольку модели данного типа моделируют границы принятия решений, пространство примеров разделяется на области, каждая из которых соответствует одному из рассматриваемых классов. Следовательно, примеры, расположенные далеко от естественных изображений, но не выходящие за пределы области классификации, с высокой степенью уверенности распознаются моделью как объекты соответствующего области класса.

Рассматривая проблему состязательных примеров, следует отметить, что первое упоминание понятия "состязательный"в том смысле, в котором оно применяется в настоящих исследованиях, встречается в статье Дальви (и др.) [15]. Авторы описывают итеративную игру, в которой участвует противник (adversary) и ученик (learner). В качестве ученика выступает наивный байесовский классификатор, который делает предсказание. Противник прослеживает алгоритм, с помощью которого было сделано предсказание, и модифицирует входные данные с целью вызвать их некорректную классификацию. Зная стратегию противника, ученик улучшает алгоритм классификации.

В 2013 году было дано четкое определение состязательных примеров как входных данных, к которым добавлена неуловимая пертурбация, максимизирующая ошибку предсказания нейронной сети [3]. В этой же статье авторы приводят результаты исследований,

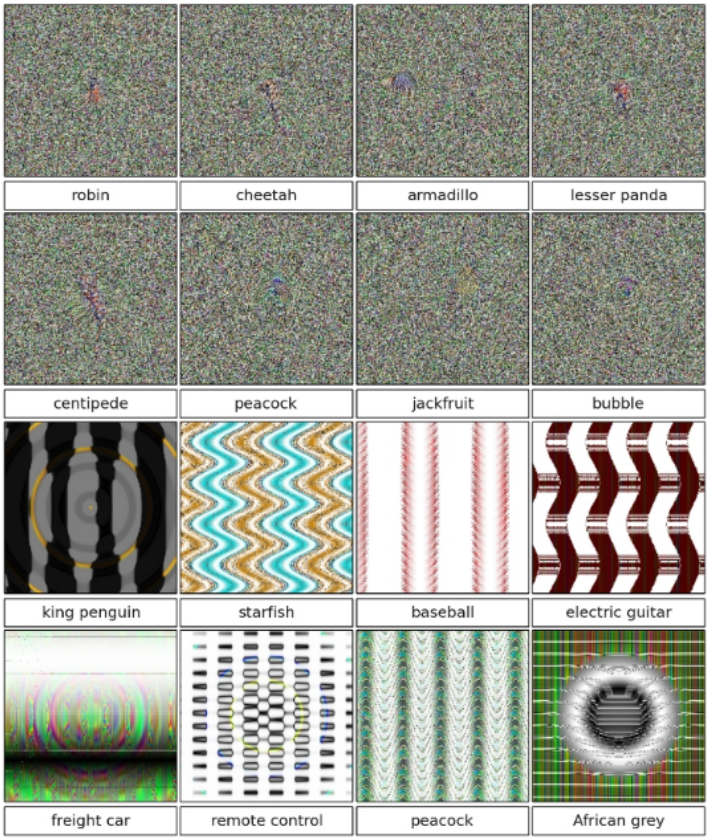

Рис. 1. "Fooling images"[14]

которые демонстрируют переносимость состязательных примеров между различными архитектурами нейронных сетей, а также между моделями, обученными на непересекающихся друг с другом наборах данных. Дальнейшие эксперименты [2] не только подтвердили предыдущие результаты, но и обнаружили, что состязательные примеры обобщаются и на другие алгоритмы машинного обучения (логистическая регрессия, k ближайших соседей, деревья решений, SVM).

2.1. Причины возникновения состязательных примеров

На сегодняшний день остается открытым вопрос относительно причин возникновения состязательных примеров. Были выдвинуты следующие гипотезы:

- Множество состязательных примеров представляет собой объединение небольших "карманов"в пространстве входных данных. Этим "карманам"соответствуют очень малые вероятности, но они плотно заполняют пространство, и поэтому их можно обнаружить вблизи любого "чистого"входного примера [3].

- Несмотря на то, что глубокие нейронные сети, как правило, моделируют нелинейные отображения, им присуще линейное поведение. Оно вызвано проектировочными решениями, нацеленными на упрощение процесса оптимизации. Кроме того, нейронные сети применяются для решения задач большой размерности. Пусть x - входной пример,

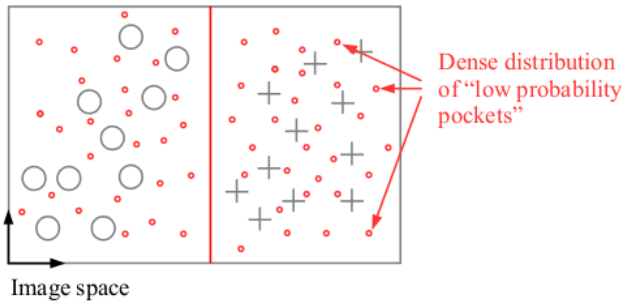

Рис. 2. Состязательные примеры в небольших "карманах"пространства данных [3, 16]

η - состязательная пертурбация, $\tilde{x} = x + \eta$ - состязательный пример, w - вектор весов некоторого слоя модели. Тогда на этом слое получим следующий результат:

$$w^\top \tilde{x} = w^\top x + w^\top \eta$$

Это означает, что значение активации для состязательного примера на $w^\top \eta$ больше, чем для соответствующего "чистого"примера. Таким образом, уязвимость модели к состязательным примерам возрастает с ростом размерности задачи [17].

- Входные данные принадлежат многообразию в пространстве изображений. Границы принятия решений между классами пересекают это многообразие, тем самым формируя подмногообразия различных классов. В то же время, границы расширяются за пределы многообразия входных примеров, и в некоторых случаях могут располагаться очень близко к нему, поэтому пертурбации, направленные к границам, могут их пересекать [16].

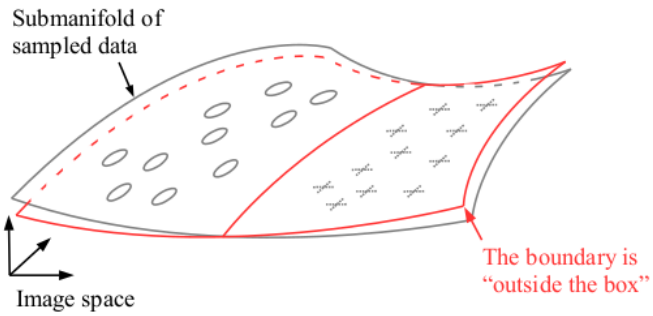

Рис. 3. Границы принятия решений выходят за пределы многообразия данных [16]

3. Обзор рассматриваемых методов атаки

Введем следующие условные обозначения:

x: входной пример.

$f()$: отображение, моделируемое целевой моделью.

l: корректная метка класса.

$\tilde{l}$: некорректная метка класса.

η: состязательная пертурбация.

θ: параметры модели.

J: функция потерь.

Любую атаку уклонения, основанную на построении состязательного примера, можно свести к поиску η такого, что

$$f(x) = l$$

$$f(x + \eta) = \tilde{l}$$

На сегодняшний день существует множество методов построения состязательных примеров. Опишем некоторые из них.

3.1. Fast gradient sign method (FGSM)

Данный метод основан на предположении о том, что уязвимость нейронных сетей к состязательным примерам объясняется их линейностью [17]. Возмущение вычисляется по формуле

$$\eta = \varepsilon \text{sign} \nabla_x J_\theta(x, l)$$

Здесь ε - величина возмущения, ∇J_θ - градиент функции потерь. Приведенная выше формула линеаризует функцию потерь в окрестности θ. Градиент вычисляется при помощи обратного распространения (back propagation). FGSM - один из самых распространенных методов состязательной атаки. Он является простым и быстрым с вычислительной точки зрения, так как требует лишь одного обновления значения градиента.

3.2. Basic iterative method (BIM)

Метод BIM представляет собой расширение метода FGSM [4]. Небольшие возмущения добавляются к входному примеру в несколько итераций.

$$\tilde{x}_0 = x, \ \tilde{x}_{N+1} = \text{Clip}_{x,\varepsilon}\{\tilde{x}_N + \alpha \text{sign}(\nabla_x J(\tilde{x}_N, l))\}$$

Функция $\text{Clip}_{x,\varepsilon}$ выполняет попиксельное отсечение изображения на каждой итерации метода. Таким образом, состязательный пример не выходит за пределы ε-окрестности исходного

примера. На практике встречаются две разновидности этого метода. Первая разновидность предполагает фиксированное число итераций, вторая предполагает остановку в тот момент, когда метка класса модифицируемого примера сменяется на некорректную.

3.3. Jacobian-based saliency map attack (JSMA)

Данный метод использует карту состязательной значимости [18]. Она показывает, какие признаки входного изображения следует подвергнуть модификации для того, чтобы построить состязательный пример. Состязательный пример строится в несколько итераций.

В первую очередь вычисляется якобиан отображения, моделируемого целевым классификатором:

$$\nabla F(x) = \frac{\partial F(x)}{\partial x} = \left[\frac{\partial F_j(x)}{\partial x_i}\right]_{i\in 1\ldots M, j\in 1\ldots N}$$

Здесь F - вектор выходных значений последнего скрытого слоя (слой логитов), M - размерность входных данных, N - размерность выхода модели. Якобиан вычисляется путем прямого распространения (forward propagation): производная каждого скрытого слоя вычисляется в терминах предыдущего скрытого слоя:

$$\frac{\partial H_k(x)}{\partial x_i} = \left[\frac{\partial f_{k,p}(W_{k,p}\cdot H_{k-1} + b_{k,p})}{\partial x_i}\right]_{p\in 1\ldots m_k}$$

Здесь H_k - вектор выходных значений k-го скрытого слоя, $f_{k,p}$ - функция активации p-го нейрона в k-ом слое, m_k - число нейронов в k-ом слое, $W_{k,p}$ и $b_{k,p}$ - соответственно веса и сдвиг p-го нейрона в k-ом слое. Затем, применяя рекурсивно цепное правило, получим:

$$\frac{\partial F_j(x)}{\partial x_i} = \left(W_{n+1,j}\cdot\frac{\partial H_n}{\partial x_i}\right)\times\frac{\partial f_{n+1,j}}{\partial x_i}(W_{n+1,j}\cdot H_n + b_{n+1,j})$$

Поскольку

$$l = \arg\max_j F_j(x),$$

для получения состязательного примера следует увеличить вероятность принадлежности изображения некорректному классу $\tilde{l}$ $(F_t(x))$ и уменьшить вероятности принадлежности всем остальным классам $(F_j(x), j \neq t)$. Для этого на основе вычисленного ранее якобиана строится карта состязательной значимости. Ее можно построить двумя способами:

1) Обнаружить признаки, которые необходимо увеличить для получения некорректной классификации.

$$S(x,t)\,[i] = \begin{cases} 0, & \text{если} \frac{\partial F_t(x)}{\partial x_i} < 0 \text{ или } \Sigma_{j\neq t}\frac{\partial F_j(x)}{\partial x_i} > 0 \\ \left(\frac{\partial F_t(x)}{\partial x_i}\right)\left|\Sigma_{j\neq t}\frac{\partial F_j(x)}{\partial x_i}\right| & \text{иначе} \end{cases}$$

2) Обнаружить признаки, которые необходимо уменьшить для получения некорректной классификации.

$$S(x,t)\,[i] = \begin{cases} 0, & \text{если} \frac{\partial F_t(x)}{\partial x_i} < 0 \text{ или } \Sigma_{j\neq t}\frac{\partial F_j(x)}{\partial x_i} > 0 \\ \left|\frac{\partial F_t(x)}{\partial x_i}\right|\left(\Sigma_{j\neq t}\frac{\partial F_j(x)}{\partial x_i}\right) & \text{иначе} \end{cases}$$

Наконец, когда обнаружен признак, который необходимо изменить, к нему добавляется возмущение величины θ. На этом шаге завершается итерация алгоритма.

Итерации продолжаются, пока норма разности между исходным и состязательным примером $\|x - \tilde{x}\|$ не превышает заданное γ.

3.4. DeepFool

Метод DeepFool [19] ищет минимальное возмущение, приводящее к некорректной классификации примера.

Рассмотрим случай, когда целевой классификатор моделирует аффинную функцию $f(x) = W^\intercal x + b$. Тогда множество $P = \{\hat{x} : f(\hat{x}) = l\} = \bigcap_{k=1}^{c} \{\hat{x} : f_l(\hat{x}) \geqslant f_k(\hat{x})\}$ - выпуклый многогранник, и $x \in P$. Следовательно, искомое возмущение - расстояние между x и дополнением к множеству P.

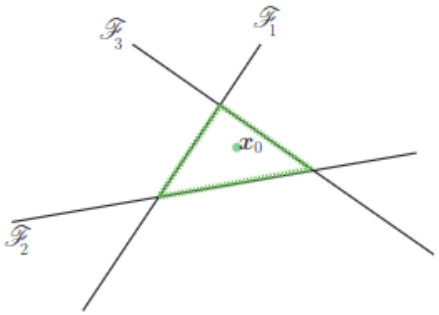

Рис. 4. [19]

Находим ближайшую к x гиперплоскость, ограничивающую множество P:

$$\tilde{l}(x) = \arg\min_{k \neq l} \frac{|f_k(x) - f_l(x)|}{\|w_k - w_l\|_2}$$

Здесь w_k - k-ый столбец матрицы W.

Искомое минимальное возмущение - вектор, проецирующий x на гиперплоскость с меткой $\tilde{l}$:

$$\eta = \frac{\left|f_{\tilde{l}}(x) - f_l(x)\right|}{\|w_{\tilde{l}} - w_l\|_2^2}(w_{\tilde{l}} - w_l)$$

3.5. Carlini & Wagner (CW)

4. Обзор рассматриваемых методов детекции

4.1. Kernel density estimation (KDE)

4.2. Bayesian uncertainty (BU)

4.3. Local intrinsic dimensionality (LID)

4.4. Mahalanobis distance (MD)

5. Реализация атак и защиты

Для демонстрации был выбран датасет CINIC-10 [20]. Этот датасет состоит из 270 тыс. изображений, равномерно распределенных в 10 классов. Изображения взяты из датасета CIFAR-10 и дополнены изображениями из ImageNet. CINIC-10 ранее не использовался для проведения исследований методов проведения состязательных атак и защиты от них.

В качестве целевой модели выбрана Resnet18 [13]. Точность модели после обучения в 300 эпох составила ....

6. Заключение

Список литературы

1. *Biggio B.*, *Roli F.* Wild patterns: Ten years after the rise of adversarial machine learning // Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. — 2018. — С. 2154—2156.
2. *Papernot N.*, *McDaniel P.*, *Goodfellow I.* Transferability in machine learning: from phenomena to black-box attacks using adversarial samples // arXiv preprint arXiv:1605.07277. — 2016.
3. Intriguing properties of neural networks / C. Szegedy [и др.] // arXiv preprint arXiv:1312.6199. — 2013.
4. *Kurakin A.*, *Goodfellow I. J.*, *Bengio S.* Adversarial examples in the physical world // Artificial intelligence safety and security. — Chapman, Hall/CRC, 2018. — С. 99—112.
5. *Otter D. W.*, *Medina J. R.*, *Kalita J. K.* A survey of the usages of deep learning for natural language processing // IEEE transactions on neural networks and learning systems. — 2020. — Т. 32, № 2. — С. 604—624.
6. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues / A. Gupta [и др.] // Array. — 2021. — Т. 10. — С. 100057.
7. Deep learning based recommender system: A survey and new perspectives / S. Zhang [и др.] // ACM computing surveys (CSUR). — 2019. — Т. 52, № 1. — С. 1—38.
8. Opportunities and obstacles for deep learning in biology and medicine / T. Ching [и др.] // Journal of The Royal Society Interface. — 2018. — Т. 15, № 141. — С. 20170387.
9. *Albawi S.*, *Mohammed T. A.*, *Al-Zawi S.* Understanding of a convolutional neural network // 2017 international conference on engineering and technology (ICET). — Ieee. 2017. — С. 1—6.
10. Gradient-based learning applied to document recognition / Y. LeCun [и др.] // Proceedings of the IEEE. — 1998. — Т. 86, № 11. — С. 2278—2324.
11. *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* Imagenet classification with deep convolutional neural networks // Communications of the ACM. — 2017. — Т. 60, № 6. — С. 84—90.
12. *Simonyan K.*, *Zisserman A.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. — 2014.
13. Deep residual learning for image recognition / K. He [и др.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 770—778.

14. *Nguyen A.*, *Yosinski J.*, *Clune J.* Deep neural networks are easily fooled: High confidence predictions for unrecognizable images // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — С. 427—436.

15. Adversarial classification / N. Dalvi [и др.] // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — 2004. — С. 99—108.

16. *Tanay T.*, *Griffin L.* A boundary tilting persepective on the phenomenon of adversarial examples // arXiv preprint arXiv:1608.07690. — 2016.

17. *Goodfellow I. J.*, *Shlens J.*, *Szegedy C.* Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. — 2014.

18. The limitations of deep learning in adversarial settings / N. Papernot [и др.] // 2016 IEEE European symposium on security and privacy (EuroS&P). — IEEE. 2016. — С. 372—387.

19. *Moosavi-Dezfooli S.-M.*, *Fawzi A.*, *Frossard P.* Deepfool: a simple and accurate method to fool deep neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 2574—2582.

20. Cinic-10 is not imagenet or cifar-10 / L. N. Darlow [и др.] // arXiv preprint arXiv:1810.03505. — 2018.