

Пояснительная записка

к дипломной работе на тему:

**"Состязательные атаки на нейронные сети"**

|  |  |
| --- | --- |
| Выполнил: | Карпов И. |
| Группа: | DSU-5 |

2023

**Оглавление**

Введение………………………………………………………………3

1. Описание проблемы………………………………………………4
2. Наборы данных…………………………………………………....6
3. Существующие типы и методы атак…………………………….7
   1. Классификация атаки по типу цели……………………….7
   2. Классификация по типу применения……………………...8
   3. Классификация по знаниям злоумышленника, методы генерации вредоносных искажений……………………….9
4. Существующие методы защиты от состязательных атак……..13
5. Эксперименты…………………………………………………....14
6. Заключение……………………………………………………….15

Список источников…………………………………………………..16

**Введение**

В современном мире нейронные сети используются повсеместно: при распознавании речи, жестов, поиске объектов на изображении, прогнозировании временных рядов, медицинской и технической диагностике, в биоинформатике, для высокочастотной торговли, обнаружения фрода, кредитного скоринга и во многих других сферах.

За последние 10 лет нейронные сети стали стандартным решением различных задач компьютерного зрения. Победа сверточной нейронной сети AlexNet на соревновании по распознаванию изображений ImageNet Large-Scale Visual Recognition Challenge в 2012 году с качеством более 63% и большим преимуществом над конкурентами вызвала в научном сообществе огромный интерес к исследованию сверточных нейронных сетей.

Дальнейший прогресс в этой области привел к тому, что по состоянию на конец 2022 года качество классификации на валидационной выборке набора данных ImageNet [повысилось до 91%](https://paperswithcode.com/sota/image-classification-on-imagenet). Также были достигнуты успехи в других задачах компьютерного зрения: детекции объектов, семантической сегментации, распознавании лиц, обработке видео и т.д. Это дало повод многим исследователям в данной области утверждать, что нейронные сети вышли на человеческий уровень решения задач компьютерного зрения, или даже превзошли человека.

К сожалению, многие модели машинного обучения чувствительны к некорректным входным данным. К примеру, метод опорных векторов очень чувствителен к шуму. При недостаточно большом размере набора данных для обучения деревья решений часто имеют области признакового пространства, в котором объекты могут неправильно классифицироваться. А глубокая сверточная нейронная сеть может быть обманута специальными изображениями. Такие изображения создаются атакующими сетями.

На некоторые модели машинного обучения намеренно производятся атаки. Они могут быть использованы для защиты от автоматического ввода капчи, сокрытия от распознавания лиц, обхода антиспам-систем или для некорректного распознавания дорожных знаков беспилотными автомобилями. Именно такие методы, направленные на введение в заблуждение моделей машинного обучения, и называются состязательными атаками. Для защиты от подобных атак необходимо использовать специальные модификации моделей машинного обучения.

1. **Описание проблемы**

Уже в 2013 году в статье ["Intriguing properties of neural networks"](https://arxiv.org/abs/1312.6199) было открыто существование состязательных примеров (adversarial examples). Авторы статьи показали, что добавление к обычным изображениям незначительного и незаметного для человека состязательного шума может привести к тому, что полученное состязательное изображение будет неверно классифицировано нейронной сетью с большой уверенностью в неправильном классе. Такая уязвимость нейронных сетей может стать критической для многих систем, которые основаны на распознавании изображений. С тех пор область состязательных атак стала активно развиваться. Множество различных атак было придумано для различных задач компьютерного зрения в разных моделях угроз (при разных знаний злоумышленника об атакуемой модели, при атаках в цифровом и реальном мире). Та же ситуация и с другими задачами, где используются нейронные сети: обработка естественного языка, распознавание речи, рекомендательные системы, задачи с временными рядами, задачи с графами и тд. Соответственно, для противодействия этим атакам стали предлагаться различные методы защиты моделей машинного обучения.

На данный момент, вышло более 6 тысяч статей (["A Complete List of All (arXiv) Adversarial Example Papers"](https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html)), посвященных атакам, защитам и вопросам робастности нейронных сетей, решающих задачи для разных типов данных. Текущее состояние этой области машинного обучения таково, что нейронные сети по-прежнему уязвимы для множества атак и методов защиты от них не существует. Модели, обученные традиционными способами, могут показывать высокие метрики на валидационных или тестовых выборках известных наборов данных, однако при применении к этим же изображениям состязательных атак, то есть наложении небольшого шума, качество классификации этих моделей снижается до 0%. Различные методы позволяют повысить качество классификации при атаках. Однако получаемое после проведения атак качество классификации значительно меньше, чем качество классификации изображений из тестовых выборок.

Таким образом, задача создания моделей машинного обучения, которые были бы устойчивыми к различным атакам и возмущениям остается нерешенной. Несмотря на это, многие модели машинного обучения внедряются в промышленную эксплуатацию, в том числе в критически важные приложения, использующиеся в авионике, автономном вождении, биометрии (в том числе в системах безопасности, банковской сфере) и других важнейших сферах. Поэтому необходимо понимание угроз, которым подвержены подобные приложения, в том числе с точки зрения уязвимости моделей машинного обучения.

Последствиями пренебрежения к вопросам безопасности могут стать следующие явления: в цифровом мире злоумышленники могут попытаться с помощью атаки обойти фильтры нежелательного контента и загрузить на сайт или в социальную сеть запрещенную информацию. Также злоумышленник может попытаться обмануть различные поисковые системы, индексирующие изображения, связав, например, продукцию какой-либо компании с каким-то неприятным для людей понятием и тем самым нанеся компании репутационные издержки.

В физическом мире (часть атак созданы именно для этого) злоумышленник может попытаться дезориентировать систему автономного вождения, заставив его неверно классифицировать дорожный знак или не заметить существующий объект. Также злоумышленник может попытаться пройти незамеченным мимо системы автоматического видеонаблюдения или попытаться выдать себя за другую личность при биометрии и, тем самым, получить несанкционированный доступ к чему-либо.

1. **Наборы данных.**

Состязательные изображения генерируются при наличии доступа к наборам данных изображений, на которых обучали компьютерные модели.  Наиболее популярные наборы данных (и многие обученные модели) без труда можно скачать без каких-либо проблем.

Например, Голиаф среди наборов данных компьютерного зрения, ImageNet, доступен на торрентах во всех его многочисленных итерациях, без его обычных ограничений, позволяя создавать наборы тестов.

Имея данные, ничего не мешает выполнить «реверсинжиниринг» любого популярного набора данных вроде Cityscapes или CIFAR. Есть и другие датасеты, которые можно использовать. Наиболее популярными наборами данных изображений в компьютерном зрении являются:

1. CIFAR-10 и CIFAR-100;
2. CALTECH-101 и 256;
3. MNIST;
4. ImageNet;
5. Pascal VOC;
6. MS COCO;
7. Sports-1M;
8. YouTube-8M.

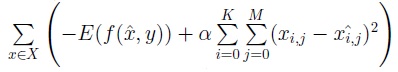
Атаки на состязательные изображения становятся возможными не только благодаря практике машинного обучения с использованием opensource подхода, но и благодаря корпоративной культуре разработки. Повторное использование хорошо зарекомендовавших себя наборов данных компьютерного зрения — вполне обычная практика, ведь реализовать их намного дешевле, чем создавать заново, они хорошо поддерживаются и обновляются.

Вот и получается, что нехватка новых наборов данных, высокая стоимость разработки наборов изображений, зависимость от старых «фаворитов» и тенденция просто адаптировать старые наборы данных — все это усугубляет проблему.

1. **Существующие типы и методы атак**
   1. **Классификация атаки по типу цели**
      1. **Атаки на изображения**

Целью состязательных атак на изображения является создание нового изображения путем небольшого изменения исходного. Изменение происходит таким образом, чтобы максимизировать функционал ошибки модели машинного обучения. Созданное изображение называется атакованным.

Например, атакующая модель может минимизировать такой функционал:



где *E( )* обозначает ошибку атакуемой модели, *x* – исходное изображение размера *K×M*, *ẋ*– соответствующее ему атакованное изображение, *X* – множество изображений, на которые производится атака, *f( )* – атакуемая модель, *α* – параметр, отвечающий за приоритетность малости *l2* нормы между атакованным и исходным изображениями.

Атаки на изображения являются одними из самых распространенных. Наиболее известным видом применения является защита от автоматического ввода капчи.

* + 1. **Атаки на временные ряды**

Целью атак на временные ряды является изменение наименьшего количества объектов в любом из промежутков времени заданной длины, максимизируя при этом ошибку модели. К этому типу атак также относятся те, которые увеличивают ошибку атакуемой модели, минимизируя суммарные изменения признаков объектов за определенный промежуток времени. Такие атаки используются в высокочастотной торговле и алгоритмическом трейдинге.

* + 1. **Атаки на агента в задачах обучения подкреплением**

Как правило, под атаками на агента подразумеваются атаки при которых в среду взаимодействия добавляется атакующий агент. Такой агент пытается уменьшить награду атакуемого агента.

* + 1. **Атаки на аудио**

Атаки на аудио в данный момент представлены, в основном, в виде атак на распознавание речи. Одним из простейших примеров является атака на VAD (Voice Activity Detector) – детектор речи в аудиопотоке. Обычно такой компонент присутствует в моделях распознавания речи. Атаки на распознавание речи распространены в голосовой капче.

* + 1. **Атаки на обработку естественного языка**

Среди атак на обработку естественного языка можно выделить следующие:

* атаки на машинный перевод;
* атаки на частеречную разметку;
* атаки на классификацию текстов;
* атаки на анализ тональности текста.
  1. **Классификация по типу применения**

Под атакой на систему машинного обучения в общем виде понимается некоторое преднамеренное вмешательство злоумышленника в процесс ее обучения или инференса. Целью такого вмешательства может являться воспрепятствование ее правильной работе, работа системы таким образом, который желает злоумышленник, кража информации про модель или приватных данных, на которых модель обучалась. Соответственно, обычно выделяют 4 типа атак:

* + 1. **Состязательные атаки (**evasion**)**

Воздействие лишь на вход уже обученной модели. Целью является неверная работа уже готовой продукта со встроенной в него моделью.

* + 1. **Атаки отравлением** (poisoning)

Изменение данных (загрязнение, зашумление данных, порча меток), на которых затем будет обучаться модель. Целью обычно является неверная работа модели.

* + 1. **Атаки извлечением**

Создание своей локальной копии некоторой модели по работе изучаемой закрытой модели.

* + 1. **Атаки инверсией**

Восстановление данных, на которых модель обучалась.

### С другой стороны можно провести ввести классификацию атак по различным свойствам результата атаки. Среди них можно выделить 4 аспекта: цель злоумышленника, универсальности атаки, норме вносимого возмущения, оптимальность/ограниченность вносимого возмущения.

### Классификация по цели злоумышленника:

* Нецелевая атака;
* Целевая атака.

### Классификация по универсальности атаки:

* Атака отдельного изображения;
* Универсальная атака для выборки.

### Классификация по норме возмущения:

Когда речь идет о наложении состязательных возмущениях, то чаще всего они предполагаются небольшими и задача атаки формулируется в рамках одной из норм. Пространство изображений одного и того же размера, представленные как вектора [0,1]*m*, можно рассматривать как линейное нормированное пространство с нормой *lp*.

Выбрав перед атакой одну из норм *lp*, злоумышленник затем или ставит задачу поиска состязательного возмущения , ограниченного по норме *lp* сверху каким-то заранее выбранным небольшим значением ϵ, или ставит задачу оптимизации с целью поиска минимального по норме *lp* состязательного возмущения .

### Классификация по оптимальности:

* Минимальное возмущение. Атака проводится с целью найти минимальное по какой-то норме *lp* состязательное возмущение, то есть задача формулируется как  при различных граничных условиях.
* Ограниченное возмущение. В такой постановке задачи достаточно найти любое состязательное возмущение, такое что .
  1. **Классификация по знаниям злоумышленника, методы генерации вредоносных искажений.**
     1. **Атаки в режиме белого ящика (white-box attacks).**

В этой модели угроз известна значительная информация о внутреннем устройстве модели. Чаще всего предполагается знание архитектуры модели и значения всех ее параметров. Это самая сильная модель угроз и она вполне может встречаться в реальной жизни, например в работе ["Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps "](https://www.usenix.org/conference/usenixsecurity21/presentation/sun-zhichuang) авторы, не прикладывая значительных усилий, полностью извлекают модели машинного обучения из приложений на мобильных устройствах с операционной системой Android.

К методам атаки типа «белый ящик» относят:

* **FGSM** – наиболее известный тип атаки. Его суть заключается в том, чтобы сделать небольшой шаг в сторону знака градиента ошибки атакуемой модели, так, чтобы норма разности исходного и атакуемого изображений была ниже определенного порога ϵ.

Атака вычисляет градиент:

*,*

а далее ищет минимальное значение (шаг) ϵ так, что

- состязательный пример.

FGSM заключается в добавлении шума, направление которого совпадает с градиентом функции стоимости по отношению к данным. Шум масштабируется по параметру ϵ. В этой формуле имеет значения не величина градиента, а его направление.

Это похоже на градиентный спуск, который предназначен для обновления весов модели, чтобы минимизировать функцию стоимости путем получения градиента по отношению к весу.

* **L-BFGS** – первая проведенная атака на сверточные нейронные сети. Эта атака минимизирует квадрат расстояния между исходным изображением и атакованным одновременно с минимизацией ошибки атакуемой модели до класса *t*. Оптимизируемый функционал выглядит так:

*,*

где коэффициент *с* подбирается таким образом, чтобы сеть продолжала классифицировать объект *x′* как класс *t*. Коэффициент *с* подбирается оптимизатором второго порядка L-BFGS, откуда и произошло название метода.

* **PGD** – это итеративный метод поиска атакованного изображения в ϵ -шаре от исходного изображения *x*. Метод осуществляет проекцию найденного на очередной итерации атакованного изображения в сторону максимизации ошибки, откуда он и получил свое название (projected gradient descent).
* **C&W** – Атака Карлини и Вагнера появилась как решение против защиты от состязательных атак типа FGSM и L-BFGS. Вместо минимизации функционала, предложенного в методе L-BFGS, предлагается использовать минимизацию зазора вероятностей всех классов с целевым. Проведение подобной оптимизации позволяет найти устойчивые к защите от состязательных атак изображения.
* **DeepFool Attack** На каждой итерации DeepFool вычисляет для каждого класса *ℓ≠ℓ0* минимальное расстояние *d(ℓ,ℓ0),* необходимое для достижения границы класса, путем аппроксимации классификатора модели линейным классификатором. Затем он делает соответствующий шаг в направлении класса с наименьшим расстоянием.
* **Universal атаку** – нахождение общего для большинства изображе-

ний набора данных возмущения *δ* с *p*-нормой ограниченной *ϵ*.

* **Атаку искривления пространства** – небольшая трансформация ло-

кальных областей изображения.

И другие типы атак.

* + 1. **Атаки в режиме черного ящика (black-box attacks).**

В таких атаках не предполагается детального знания об архитектуре и параметрах модели. Однако предполагается взаимодействие с моделью и возможность наблюдать ее выход в зависимости от посылаемого ей входа. Такие атаки можно дополнительно разделить на следующие подкатегории:

* Атаки на основе выходного распределения модели - имеются в виду предсказанные моделью вероятности или логиты (входы на softmax-слой). Во многих случаях суть данных атак сводится к численной оценке градиента (его прямое вычисление невозможно при отсутствии доступа к параметрам модели), в других - используются неградиентные методы оптимизации, например генетические алгоритмы или случайный поиск.
* Атаки на основе решения модели - предполагается наличие доступа только к предсказанному моделью классу объекта.
* Атаки на основе переносимости - в самой первой статье по состязательным атакам “[Intriguing properties of neural networks](https://arxiv.org/abs/1312.6199)” было открыто, что состязательные примеры могут быть успешно перенесены с одной модели на другую, обученную на том же датасете. В этой модели угроз предполагается наличие доступа или к полному датасету, на котором обучалась целевая модель, или к его части для создания модели-суррогата, для которой и будут синтезироваться состязательные примеры (например, одним из методов белого ящика), а затем эти изображения будут использоваться для атаки на исхожную модель. В некоторых подходах помимо датасета требуется также доступ к выходному распределению модели.

К методам атаки типа «черный ящик» относят:

* **Аддитивная равномерная шумовая атака** – Эта атака проверяет надежность модели на добавление равномерного шума. Внутренний поиск выполняется для нахождения минимальных посторонних возмущений.
* **Аддитивная шумовая атака по Гауссу** – Эта атака проверяет надежность модели на добавление нормального шума. Внутренний поиск выполняется для нахождения минимальных посторонних возмущений.
* **Шумовая атака Salt And Pepper** – Эта атака проверяет надежность модели на добавление специального шума (соли – белые пиксели и перца – черные пиксели). Внутренний поиск выполняется для нахождения минимальных посторонних возмущений.
* **Атака уменьшения контрастности** – Эта атака проверяет устойчивость модели к снижению контрастности. Внутренний поиск выполняется для нахождения минимальных посторонних возмущений.
* **Внутренний поиск** может выполняться следующим образом: 1) строится дискретная сетка параметров, значения считаются по сетке и выбирается наихудшее 2) случайно выбираются *k* значений параметра. Из результатов выбирается худший.
* **Размытие по Гауссу** – проверка устойчивости к размытию по Гауссу.
* **Однопиксельная атака** – устойчивость к изменению отдельных пикселей, устанавливая для одного пикселя черный или белый цвет.
* **Атака на локальный поиск** – измеряет чувствительность модели к отдельным пикселям, применяя экстремальные возмущения и наблюдая влияние на вероятность правильного класса, так ищутся наиболее чувствительные пиксели, и делается состязательное изображение.

1. **Существующие методы защиты от состязательных атак.**

Наличие состязательных примеров и возможность их синтезировать в режиме черного ящика на основании выходного распределения модели или только лишь используя предсказанный моделью класс, их переносимость с одной модели на другую, а также методы генерации робастных состязательных примеров для физического мира вызвали серьезные вопросы, связанные с безопасностью внедрения моделей компьютерного зрения в реальные системы.

Вместе с развитием техник атак на модели машинного обучения параллельно широко развивалась область защит от состязательных примеров. Однако задача защиты модели от состязательных примеров довольно плохо формализуется.

На данный момент, все методы можно разделить на эмпирические и сертификационные. Эмпирические защиты от состязательных атак можно разделить на 3 категории:

* **Робастная оптимизация**. Методы из этой категории нацелены на повышение эффективности процедуры обучения моделей. Эту категории можно дополнительно разделить на 3 подкатегории:
  + **Состязательное обучение**. Так как существующие наборы данных не покрывают полностью распределения, на которых классификаторы работают в реальной жизни, а также не включают в себя состязательные примеры, то некоторые техники предлагают синтезировать состязательные примеры прямо во время обучения и подмешивать их к обучающей выборке. Данные методы являются чисто эмпирическими и не предоставляют никаких теоретических гарантий.
  + **Регуляризация**. Эти методы применяют различные техники регуляризации для уменьшения влияния небольших входных возмущений на выход модели.
  + **Байесовские подходы**.
* **Усложнение синтеза состязательных примеров (сокрытие/запутывание градиента)**. Методы из данной категории нацелены на создание моделей, градиенты которых не получится использовать для оптимизации, что помешает выбрать правильное направление для оптимизации. Однако данное методы имеют лишь ограниченную применимость, так как многие атаки (особенно в режиме черного ящика) не используют градиенты.
* **Детектирование состязательных примеров**. Обучение дополнительных моделей или применение методов из статистики для отделения состязательных примеров от обычных сэмплов.

1. **Эксперименты.**

Для проведения экспериментов были выбраны открытые наборы данных MNIST и CIFAR10. Тестирование производилось на ResNet-18, сверточная нейронная сеть, [SqueezeNet](https://arxiv.org/abs/1602.07360). Использовались библиотеки, позволяющие работать с Adversarial примерами Foolbox и ART-IBM, для них использован датасет ImageNet. Атаки, исследованные во время эксперимента:

• FSGM;

• One pixel attack;

• PGD;

• Deep Fool;

• C&W;

• Basic Iterative Attack;

• Additive Uniform Noise Attack.

Таблица 1. Полученная точность предсказания верного класса для разных атак в зависимости от изменения , при *l* ∞, в процентах.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attack \ | 0.0 | 0.0005 | 0.001 | 0.002 | 0.003 | 0.005 | 0.01 | 0.1 | 0.5 | 1 |
| FGSM | 94 | 81 | 44 | 25 | 6 | 0 | 0 | 0 | 0 | 0 |
| PGD | 94 | 81 | 38 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Basic Iterative | 94 | 69 | 38 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Additive Uniform Noise | 94 | 94 | 94 | 94 | 94 | 94 | 94 | 81 | 12 | 0 |
| DeepFool | 94 | 75 | 31 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |

Был воспроизведен один из методов защиты от атаки – повторное обучение модели после добавления атакованных изображений в обучающую выборку. Точность такого подхода составила порядка 93%.

Минусом такого подхода, очевидно, является изменение исходной модели, что приводит к снижению ее точности. Также это заставляет модель подстраиваться к атакам именно того типа, который использовался при обучении, но может не защитить от атак другим методом. Кроме того, в некоторых реальных задачах, например, в детекции злокачественных опухолей на медицинских снимках, обучение модели – это очень долгий и дорогой процесс из-за объемов данных.

1. **Заключение.**

Хотя за последние несколько лет мы наблюдаем впечатляющий прогресс в машинном обучении, мы должны осознавать ограничения, которыми все еще обладают наши инструменты. Надежность является одной из ключевых проблем. Человеческое восприятие и когнитивные способности устойчивы к широкому спектру неприятных возмущений в реальном мире. Однако состязательные примеры показывают, что глубокие нейронные сети в настоящее время далеки от достижения того же уровня надежности.

С одной стороны, современные модели теперь достигают человеческого уровня точности в сложных задачах, таких как классификация Imagenet. С другой стороны, состязательные атаки показывают, что хорошая точность при выполнении конкретных задач не означает, что мы создали классификаторы изображений, которые так же надежны, как люди. Эта проблема также встречается и в других областях, таких как распознавание речи, где акценты или шумная среда все еще являются значительными препятствиями для глубоких сетей.

Подводя итог, можно сделать вывод, что состязательные примеры актуальны не только для проверки безопасности машинного обучения, а также представляют собой диагностическую основу для оценки обученных моделей. В отличие от стандартных процедур оценки, состязательный подход выходит за рамки статического набора тестов и позволяет нам выявить потенциально неочевидные недостатки. До тех пор, пока наши классификаторы подвержены небольшим изменениям, достижение гарантии надежности будет невозможно. В конце концов, цель состоит в том, чтобы создать модели, которые не только безопасны, но и согласуются с нашей интуицией о том, что значит “изучить” задачу, чтобы они были надежными, безопасными и простыми в развертывании в различных средах.

**Список источников**

# Aleksander Mądry, Ludwig Schmidt. A Brief Introduction to Adversarial Examples //gradientscience.org/intro\_adversarial/

# Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Mądry. Adversarial Examples Are Not Bugs, They Are Features. // arxiv.org/abs/1905.02175.

# [Nicholas Carlini](https://arxiv.org/search/cs?searchtype=author&query=Carlini%2C+N), [David Wagner](https://arxiv.org/search/cs?searchtype=author&query=Wagner%2C+D). Towards Evaluating the Robustness of Neural Networks. // arxiv.org/abs/1608.04644.

* 1. [Aleksander Madry](https://arxiv.org/search/stat?searchtype=author&query=Madry%2C+A), [Aleksandar Makelov](https://arxiv.org/search/stat?searchtype=author&query=Makelov%2C+A), [Ludwig Schmidt](https://arxiv.org/search/stat?searchtype=author&query=Schmidt%2C+L), [Dimitris Tsipras](https://arxiv.org/search/stat?searchtype=author&query=Tsipras%2C+D), [Adrian Vladu](https://arxiv.org/search/stat?searchtype=author&query=Vladu%2C+A). Towards Deep Learning Models Resistant to Adversarial Attacks. // [arxiv.org/abs/1706.06083](https://arxiv.org/abs/1706.06083).

# [Battista Biggio](https://arxiv.org/search/cs?searchtype=author&query=Biggio%2C+B), [Fabio Roli](https://arxiv.org/search/cs?searchtype=author&query=Roli%2C+F). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. // [arxiv.org/abs/1712.03141](https://arxiv.org/abs/1712.03141).

# [Logan Engstrom](https://arxiv.org/search/cs?searchtype=author&query=Engstrom%2C+L), [Brandon Tran](https://arxiv.org/search/cs?searchtype=author&query=Tran%2C+B), [Dimitris Tsipras](https://arxiv.org/search/cs?searchtype=author&query=Tsipras%2C+D), [Ludwig Schmidt](https://arxiv.org/search/cs?searchtype=author&query=Schmidt%2C+L), [Aleksander Madry](https://arxiv.org/search/cs?searchtype=author&query=Madry%2C+A). Exploring the Landscape of Spatial Robustness. // [arxiv.org/abs/1712.02779](https://arxiv.org/abs/1712.02779).

# [Martin Anderson](https://www.unite.ai/author/martinanderson/). Optical Adversarial Attack Can Change the Meaning of Road Signs. // unite.ai/optical-adversarial-attack-can-change-the-meaning-of-road-signs/

# [Jiawei Su](https://arxiv.org/search/cs?searchtype=author&query=Su%2C+J), [Danilo Vasconcellos Vargas](https://arxiv.org/search/cs?searchtype=author&query=Vargas%2C+D+V), [Sakurai Kouichi](https://arxiv.org/search/cs?searchtype=author&query=Kouichi%2C+S). One pixel attack for fooling deep neural networks. // arxiv.org/abs/1710.08864.

# [Nicolas Papernot](https://arxiv.org/search/cs?searchtype=author&query=Papernot%2C+N), [Patrick McDaniel](https://arxiv.org/search/cs?searchtype=author&query=McDaniel%2C+P), [Ian Goodfellow](https://arxiv.org/search/cs?searchtype=author&query=Goodfellow%2C+I), [Somesh Jha](https://arxiv.org/search/cs?searchtype=author&query=Jha%2C+S), [Z. Berkay Celik](https://arxiv.org/search/cs?searchtype=author&query=Celik%2C+Z+B), [Ananthram Swami](https://arxiv.org/search/cs?searchtype=author&query=Swami%2C+A). Practical Black-Box Attacks against Machine Learning. // arxiv.org/abs/1602.02697.

# [Naveed Akhtar](https://arxiv.org/search/cs?searchtype=author&query=Akhtar%2C+N), [Ajmal Mian](https://arxiv.org/search/cs?searchtype=author&query=Mian%2C+A). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. // arxiv.org/abs/1801.00553.

## Jonas Rauber, Roland S. Zimmermann. Foolbox: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX // [foolbox.readthedocs.io](https://foolbox.readthedocs.io/en/stable/)

# Adversarial Robustness Toolbox (ART) // adversarial-robustness-toolbox.readthedocs.io