1. Определение понятий.

В этой секции перечислены основные термины, используемые в литературе, посвященной adversarial-атакам на системы, использующие глубокое обучение для решения задач компьютерного зрения. [слишком длинно, переписать своими словами]

* Adversarial-пример/изображение – намеренно искаженная (например, путём добавления шума) версия исходного изображения, используемая для обмана алгоритма машинного обучения.
* Adversarial-искажение – шум, добавляемый к исходному изображению для того, чтобы сделать из него adversarial-пример.
* Adversarial-обучение – добавление adversarial-примеров к исходным данным в процессе обучения модели.
* Black-box-атака – атака на выбранную модель, совершаемая без знания её структуры. Иногда допускается, что атакующий обладает ограниченными знаниями о структуре модели (например, о её процессе обучения или архитектуре), но не знает точные значения её параметров.
* Детектор – позволяет определить, является ли изображение adversarial-примером.
* Коэффициент обмана [fooling ratio? проблема в том, что в русском лучше звучит перевод из диапазона 0-1 в проценты, т.е. 0.5 в 50% и т.д., а в формулах используются коэффициенты с диапазоном 0-1] – процент изображений, на котором модель меняет результат предсказания после добавления к ним искажений.
* Одношаговые-методы – способы создания adversarial-искажений, предполагающие единственное вычисление, например, вычисление градиента функции потерь модели. С другой стороны – итеративные методы, совершающие множество последовательных однотипных операций и часто являющиеся вычислительно более сложными.
* Квазинезаметные искажения [в русском используется «квази»?] – практически не изменяющие изображения для человека.
* Очиститель [rectifier – по аналогии с «детектором», лучше перевести] – позволяет модифицировать adversarial-пример, восстанавливая исходное предсказание модели.
* Направленные атаки – используются для того, чтобы модель предсказала какой-то конкретный неверный результат. Их противоположность – ненаправленные атаки, цель которых заставить модель выбирать любой неверный ответ. […сделать любое неверное предсказание]
* Threat model [должен быть устоявшийся аналог] – описание потенциальных атак, в зависимости от подхода (например, black-box атаки).
* Переносимость – возможность использования adversarial-примера на моделях, которые не были использованы для его создания.
* Универсальное искажение - имеющее высокую вероятность обмануть модель на любом изображении. Следует отметить, что универсальность – свойство независимости искажения от класса изображения, в то время как переносимость – независимость искажения от модели.
* White-box-атака – предполагает, что у атакующего имеются полные сведения о модели, включая её параметры, архитектуру, метод обучения и, в некоторых случаях, обучающую выборку.

2. Классификация.

3. Adversarial-атаки.

В этом разделе рассмотрены наиболее известные методы атак на нейронные сети в «лабораторных условиях»; их эффективность оценивается на стандартных наборах данных (например, MNIST или CIFAR-10) в типичных задачах компьютерного. Следует заметить, что подходы, рассмотренные в данном разделе, составляют основу для атак в реальных условиях и каждый из рассмотренных подходов потенциально способен повлиять на развитие глубокого обучения на практике.  
 Обзор в данном разделе ведётся в хронологическом порядке, к основных подходам и главным направлениям атак даётся техническое описание, в то время как остальные подходы описаны в общих чертах, подробности читатель может найти в приложенных отсылках к исходным работам.

В части 3.1 рассмотрены атаки на наиболее распространённую задачу компьютерного зрения – классификацию/распознавание. Подходы, направленные на совершение атак на другие задачи (сегментация, атаки на обучение с подкреплением и др.) рассмотрены в части 3.2.

3.1 Атаки на классификацию.

3.1.1 Box-constrained L-BFGS (L-BFGS) [переводить на русский?]

Впервые существование adversarial-искажений продемонстрировал Szegedy и др. в [22].  
Пусть исходное изображение представлено в виде вектора , индекс «c» подчеркивает, что изображение «чистое» («clear»). Чтобы посчитать искажение , прибавление которого к исходному изображению изменит результат классификации, визуально не изменяя изображение было предложено найти решение следующей задачи:

Авторы предложили найти нетривиальное решение задачи (1), в котором предсказание классификатора отличается от исходного класса изображения. В этом случае нахождение становится нетривиальной проблемой, и для нахождения приблизительного решения авторы предлагают использовать алгоритм L-BFGS: находится минимальное для которого выполняется следующее условие:

В случаях, когда функция потерь классификатора представляет собой выпуклую функцию, решение (2) даёт единственное решение. Но, в случае глубоких нейронных сетей, это, чаще всего, не так.  
 Для получения adversarial-примера, вычисленное искажение просто прибавляется к исходному изображению. Как показано на рис. (2), полученные adversarial-изображения визуально неотличимы от исходных изображений. Также было замечено, что искажения, вычисленные для нейронной сети могут работать и для других сетей.

Во время публикации работы считалось, что нейронная сеть выделяет признаки, определяющие пространство, где воспринимаемые между изображениями различия приблизительно равняются евклидовому расстоянию между ними. [At the time of this discovery the Computer Vision community was fast adapting to the impression that deep learning features define the space where perceptual distances are well approximated by the Euclidean distances.] Эти противоречивые результаты пробудили интерес исследователей к adversarial-атакам.

3.1.2 Fast Gradient Sign Method (FGSM).

Szegedy и др. в [22] было замечено, что устойчивость нейронных сетей к adversarial-примерам может быть повышена за счёт adversarial-обучения. Для того, чтобы повысить эффективность adversarial-обучения, Goodfellow и др. в [23] бы разработан метод быстрого расчёта adversarial-искажений:

градиент функции стоимости в точке, определяемой параметрами модели ,

по отношению к исходному изображению ,скаляр, ограничивающих размер искажения.

Adversarial-примеры, полученные с помощью FGSM, используют «линейность» нейронных сетей в многомерном пространстве, хотя в то время такие модели считались сильно нелинейными. Goodfellow и др. в [23] выдвинули гипотезу, что дизайн современных нейронных сетей, намеренно поощряющий линейное поведение для вычислительной простоты (ReLU), делает их уязвимыми к простому поиску подобных искажений. Эта идея получила название «гипотезы линейности» [linearity hypothesis] и поддерживается существованием FGSM похода.  
 Kurakin и др. [80] отмечают, что adversarial-примеры, полученные с помощью FGSM для изображений из ImageNet, достигают успешного обмана сетей (top-1 error rate) в 63-69% случаев при Также авторы предложили «одношаговый направленный FGSM метод», в котором вместо использования в (3) исходного класса , используется вычисляется градиент по отношению к классу , имеющему наименьшую вероятность среди предсказанных сетью для изображения [не совсем направленная?]

Полученное таким образом искажение вычитается из исходного изображения, чтобы сделать из него adversarial-изображение. Для нейронной сети, использующей кросс-энтропию как функцию потерь, такая операция максимизирует вероятность выбора в качестве класса adversarial-примера. Предполагается, что использование случайного класса в качестве также может привести к успешному обману сети, хотя это может привести к менее интересным результатам, например к классификации одной породы собак в качестве другой. Авторы также показали, что использование adversarial-обучения повышает устойчивость сети к атакам с помощью FGSM и его вариантов.

FGSM искажает изображение, увеличивая значение функции потерь для правильного класса. Функция позволяет максимизировать величину искажения, в то время как ограничивает его равномерную норму (-норму) [проверить мат. термин]. Miyato и др. [103] предложил похожий метод вычисления adversarial-искажений:

В приведенном выражении, градиент нормализуется делением на -норму. Kurakin и др. [80] назвали этот способ «FGL2» методом и предложили использовать для нормализации равномерную норму, назвав получившуюся технику «FGL∞». Все эти методы рассматриваются как «одношаговые» методы вычисления adversarial-искажений.

3.1.3 Basic & Least-Likely-Class Iterative Methods (BIM & ILCM).

Одношаговые методы искажают изображения, делая один большой шаг в направлении, увеличивающим значение функции потерь классификатора. Интуитивное продолжение этой идеи – итеративное совершение множества маленьких шагов, сопровождающееся корректировкой направления движения на каждом шаге.  
 Простой итеративный метод (BIM) [35] предполагает именно это:

где – искаженное изображение на -ой итерации, - ограничивает значения пикселей изображения на , – определяет размер шага (обычно, ).

Алгоритм начинается с и совершает количество итераций, определяемое выражением . Было замечено [55], что BIM является эквивалентом PGM (Projected Gradient Descent) – стандартному методу оптимизации выпуклых функций.  
[уточнить мат. правильность]

По аналогии с модификацией FGSM до его «направленной» версии, Kurakin и др. [35] предложили расширить BIM до Iterative Least-likely Class Method (ILCM). В этом случае, в качестве класса в (5) предлагается использовать класс (с наименьшей вероятностью, данной классификатором). Было показано, что adversarial-примеры, полученные с помощью ILCM, сильно влияют на точность классификации сетей с архитектурой Inception v3, даже для малых значений .

3.1.4 Jacobian-based Saliency Map Attack (JSMA).

Наиболее частый подход к генерации квазинезаметных adversarial-изменений – это ограниение их или -норм. Papernot и др. [60] предложили способ проведения атак, путём ограничения -норм adversarial-искажения. На практике, это означает что такие искажения модифицируют только несколько пикселей исходного изображения. Основой их алгоритма является поочередная модификация пикселей исходного изображения и отслеживание изменений в результирующей классификации. Изменения отслеживаются путём построения карты градиентов выходных слоёв сети – в ней для каждого пикселя исходного изображения задаётся значение – большее значение соответствуют большей вероятности изменения предсказанного класса при изменении данного пикселя. После этого, алгоритм выбирает пиксель с наибольшей вероятностью обмануть сеть и изменяет его. Этот процесс повторяется до тех пор, пока не достигается максимальное количество измененных пикселей или пока не удастся обмануть классификатор.

3.1.5 One Pixel Attack

Крайним случаем adversarial-атаки можно считать изменение единственного пикселя исходного изображения. Su и др. [68] провели атаку на три разные сети с успехом на 70.97% тестовых изображений, изменяя значения всего лишь одного пикселя. Было также показано, что средняя уверенность классификатора на ложных примерах равнялась 97.47%. Примеры adversarial-изображений из [68] см. на (рис. 3).  
 Adversarial-примеры были получены используя метод дифференциальной эволюции [148]: были выбраны 400 векторов, содержащих xy-координаты и RGB значения случайных пикселей исходного изображения. Затем, значения векторов были случайным образом изменены, создавая следующее поколение, при этом потомок соревнуется со своим родителем, используя в качестве целевого критерия вероятность, полученную от классификатора. После нескольких итераций, для изменения изображения выбирается последний оставшийся потомок.

Даже с такой простой эволюционной стратегией Su и др. [68] удалось провести успешые атаки на нейронные сети. Следует заметить, что дифференциальная эволюция позволяет генерировать adversarial-примеры без доступа к параметрам и градиентам сети. Единственная необходимая информация – это вероятности, полученные от классификатора.

3.1.6 Carlini and Wagner Attacks (C&W).

При появлении методов защиты простив adversarial-изменений, Carlini и Wagner предложили набор из 3 атак, делающих изменений квазинезаметными путём ограничения их и -норм. Ими было показано, что данные атаки способны преодолеть механизм защиты, называемый defensive distillation [38][166]. Более того, было показано, что adversarial-примеры, полученные для незащищенных этим механизмом сетей, могут быть перенесены на защищенные, что допускает совершение black-box-атак.

В то время, когда наиболее частым методом совершения black-box-атак является использование переносимости adversarial-примеров, Chen и др. [41] предложили «Zeroth Order Optimization» атаку, заключающуюся в приблизительном вычислении градиентов атакуемой модели.

3.1.7 DeepFool.

Moosavi-Dezfooli и др. [72] предложили вычислять квазинезаметные искажения итеративным способом. Их алгоритм заключается в линеаризации функции потерь и вычисления минимального искажения, необходимого для изменения классификации при данной линеаризации. На каждом шаге алгоритма функция потерь вычисляется при добавлении полученного на прошлом шаге искажения. Так происходит до тех пор, пока классификатор не станет выдавать ошибочный результат (таким образом DeepFool можно отнести к ненаправленным атакам). Авторы [72] показывают что полученные adversarial- искажения имеют меньшие нормы, чем полученные с помощью FGSM и имеющие тот же процент успешной мисклассифиакции. [удобное слово, можно использовать?]

3.1.8 Universal Adversarial Perturbations.

В то время, как FGSM, ILCM, DeepFool и др. методы нацелены на вычисление adversarial-изменений, обманывающих классификатор на единственном изображении, «универсальные» adversarial-искажения, вычисленные Moosavi-Dezfooli и др. [16], с высокой вероятностью подходят для любого изображения. Как показано на (рис. 1) эти искажения могут быть квазинезаметными для человека. Для их формального определения предположим, что исходные «чистые» изображения получены из распределения Ω. Тогда изменение можно назвать «универсальным», если оно удовлетворяет следующему ограничению:

где – вероятность, – процент успешных обманов, -норма,

– константа, определяющая величину искажения.

Строго говоря, искажения, удовлетворяющие (6) должны называться   
«-универсальным», подчёркивая сильную зависимость от названных параметров. Как бы то ни было, в литературе их обычно называют просто «универсальными».

Авторы [16] вычислили универсальные adversarial-искажения, ограничивая их и   
-норму, и показали, что искажения в 4% от нормы исходного изображения достигают значительного процента успешных обманов (80%) на лучших классификаторах. Их метод похож на метод, используемый в DeepFool – последовательного сдвига изображения к границе класса, определенной классификатором. Но в этом случае «все» изображения из тренировочной выборки последовательно сдвигаются к своим границам, накапливая суммарное искажение, ограничиваемое по -норме величиной .

Алгоритм, предложенный в [16] вычисляет adversarial-искажения, нацеленные на единственную модель. Тем не менее, было показано, что эти искажения также распространяются и на другие сети, особенно имеющие сходную архитектуру. В этом отношении, авторы [16] называют искажения «вдвойне универсальными». Более того, ими было показано, что высокого процента успешных обманов (более 50%) можно достичь на тренировочной выборке, состоящей всего лишь из 2000 изображений.

Другие методы получения универсальных искажений были предложены   
Хрульковым и др. [190] (позволяет получать хорошие результаты, используя маленькое количество изображений), Mopuri и др. [135] (получили универсальные искажения, независимые от данных).

3.1.9 UPSET и ANGRI.

Эти две black-box-атаки были предложены Sarkar и др. [146].  
UPSET (Universal Perturbations for Steering to Exact Targets): для n классов алгоритм вычисляет n независимых от изображения искажений, соответствующих направленному изменению классификации для любого изображения. UPSET основан на генерирующей сети , принимающей на вход выбранный класс и создающей направленное adversarial-искажение . Полностью, UPSET-сеть решает следующую оптимизационную проблему:

где значения пикселей изображения нормализованы в промежутке , а s – скаляр.

Значения пикселей искаженного изображения также ограничиваются интервалом .

ANGRI похожим образом вычисляет индивидуальные для каждого изображения искажения. Было показано, что оба алгоритма достигают высоких показателей на выборках из MNIST [10] и CIFAR-10 [152].

3.1.10 Houdini.

Cisse и др. [131] предложили алгоритм для обмана основанных на градиентном обучении подходов, подходящий для различных функций потерь. Типичные adversarial-атаки предполагают наличие дифференцируемой функции потерь, что иногда может не соблюдаться. Например, функция потерь в задаче распознавании речи может быть основана на доле правильно разобранных слов, что не предполагает прямого использования градиента. Houdini используется для создания adversarial-примеров для таких задач. Кроме удачного создания adversarial-примеров для классификации, Houdini был использован для совершения атаки на систему распознавания речи [151]. Также авторами была продемонстрирована переносимость атак на распознавание речи, путём совершения black-box-атаки на Google Voice.

3.1.11 Adversarial Transformation Networks (ATNs).

Baluja и Fisher [42] обучили нейронные сеть создавать adversarial-примеры для набора других нейронных сетей. Adversarial-примеры, созданные этими сетями, вычисляются путём минимизации составной функции потерь: первая часть обеспечивает квазинезаметность adversarial-искажений, а вторая направлена на изменение классификации результирующего изображения.

Hayex и Danezis [47] также использовали нейронную сеть для создания adversarial-примеров для black-box-атак. В представленных результатах, adversarial-примеры, созданные сетью остаются визуально неотличимыми от исходных изображений, но уменьшают точность классификации с 99.4% до 0.77% на MNIST [10] и с 91.4% до 6.8% на CIFAR-10 [152].

3.1.12 Другие атаки.

Описанные выше adversarial-атаки являются либо самыми популярными в недавней литературе, либо главными представителями основных направлений исследований. Ниже представлено краткое описание других методов совершения adversarial-атак.

Sabour и др. [26] описали вероятность создания adversarial-примеров путём изменения начальных слоёв нейронной сети. Авторы показали, что возможно изменить внутреннее представление adversarial-изображения, повторяющее представление изображений различных классов. Papernot и др. в [109] изучили переносимость adversarial-атак. Narodytska и Kasiviswanathan в [54] представили способ совершения black-box-атаки, путём изменения нескольких пикселей изображения. Liu и др. в [31] представили «эпсилон-окрестность»-white-box-атаку, успешность которой по отношению к защищенным через defensive distillation сети достигла 100%. Oh и др. в [133] применили теорию игр для разработки стратегии противодействия существующим способам защиты от adversarial-атак. Mpouri и др. в [135] разработали метод вычисления универсальных искажений. Hosseini и др. в [98] ввели понятие «семантические-adversarial-примеры» - изображения, которые обозначают одинаковые объекты для человека, но разные для классификатора. Kanbak и др. [73] предложили ManiFool – алгоритм, измеряющий устойчивость нейронной сети к геометрически искаженным изображениям. Dong и др. [170] предложил итеративный метод совершения adversarial-black-box-атак. Недавно, Carlini и Wagner [59] показали, что десять методов защиты от adversarial-атак могут быть преодолены, используя новые атаки, использующие измененные функции потерь. Rozsa и др. [94] предложили метод создания нескольких различных adversarial-примеров для исходного изображения. Следует упомянуть, что adversarial-искажения можно использовать для улучшения качества классификации. Yoo и др. в [195] предложили метод улучшения классификации, путём добавления небольших искажений к исходному изображению.

3.2 Атаки на другие задачи глубокого обучения.