Обзор и систематизация атак уклонением на модели компьютерного зрения

В.В. Костюмов

Аннотация—**Глубокое обучение привлекло огромное внимание научного сообщества в последние годы за счет отличных результатов в различных областях задач, в том числе компьютерном зрении. Например, в задаче классификации изображений некоторыми авторами даже было объявлено, что нейронные сети превзошли человека по качеству распознавания. Однако открытие состязательных примеров для моделей машинного обучения показало, что современные архитектуры компьютерного зрения очень уязвимы перед злоумышленниками и требуется дополнительное внимание при внедрении их в критические области инфраструктуры. С тех пор было предложено множество новых атак в разных моделях угроз и показана возможность осуществления подобных атак в реальном мире. При этом до текущего момента не предложено ни одного метода защиты, который был бы надежен относительно уже существующих атак, не говоря уже о гарантиях относительно всего возможного множества угроз. В данной работе проводится обзор и систематизация атак уклонением в области компьютерного зрения. В данном типе атак, который является наиболее распространенным, злоумышленник способен взаимодействовать с моделью только во время вывода, меняя при этом входные данные.**

Ключевые слова—**машинное обучение, компьютерное зрение, состязательные атаки**

I. Введение

Применение нейронных сетей в компьютерном зрении позволило добиться значительных результатов во многих задачах, таких как классификация изображений, детектирование объектов, семантическая сегментация, распознавание лиц, отслеживание объектов. Это привело к повсеместному внедрению этих моделей во многие системы, например, в беспилотные автомобили, системы видеонаблюдения, мобильные телефоны (FACE ID), системы поиска по картинкам.

В работе [1] была впервые показана уязвимость нейронных сетей в задаче классификации изображений. Оказалось, что несмотря на высокое качество этих моделей на валидационной и тестовой выборках, современные архитектуры подвержены состязательным атакам в форме незначительных и незаметных для человека изменений изображений, приводящих к неверной классификации возмущенного изображения. При этом модель присваивает изображению неверную метку с высокой вероятностью, а одно и то же возмущение способно обмануть сразу несколько моделей. Это привлекло внимание научного сообщества к вопросам безопасности нейронных сетей и устойчивости их предсказаний.

С тех пор множество атак было значительно расширено. Наибольшее количество из них было посвящено задаче классификации. Были открыты состязательные возмущения, которые являются универсальными не только для одного изображения, но и для целых выборок [2]. Были предложены атаки, в которых злоумышленник не знает архитектуру и веса модели, но может лишь наблюдать выходные распределения моделей ([3], [4], [5], [6]). Еще меньше требуют атаки, которые подбирают состязательный шум лишь на основе выдаваемой моделью метки ([7], [8]).

Отдельной областью атак уклонением стали атаки в реальном мире, когда злоумышленник способен изменять объекты окружающего мира и не ограничен условиями незаметности вносимого возмущения, но не способен напрямую влиять на то, в каком виде изображение этого объекта попадет в нейронную сеть ([9], [10], [11], [12]).

Статья получена 1 сентября 2022.

Василий Владимирович Костюмов, МГУ им. М.В. Ломоносова, (email: kostyumov@yandex.ru).

II. Классификация атак уклонением

В такой модели угроз злоумышленник не может влиять ни на обучающие данные (как это происходит в атаках отравлением), ни каким-либо образом изменять параметры модели. Злоумышленник может лишь иметь доступ к информации про архитектуру и веса модели или про ее поведение на различных входам. Можно декомпозировать модель угроз рассматриваемых атак на следующие аспекты: знания злоумышленника о модели, возможность напрямую контролировать вход модели, цель злоумышленника, итеративная мощность атаки.

1) *Классификация по знаниям злоумышленника*

- **Атаки в режиме белого ящика (white-box).** В этой модели угроз злоумышленник обладает значительной информацией 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 [13] авторы, не прикладывая значительных усилий, полностью извлекают модели машинного обучения на мобильных устройствах с операционной системой Android.

  Чаще всего информация об архитектуре и параметрах модели используется для вычисления градиента выхода модели по ее входу.
- **Атаки в режиме черного ящика (black-box).** 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. Во многих случаях суть данных атак сводится к численной оценке градиента ([4]), [5] так как прямое его вычисление невозможно при отсутствии доступа к параметрам модели. В других случаях используются оптимизационные алгоритмы, не требующие градиента, такие как эволюционные алгоритмы ([3], [14]) или случайный поиск ([6]).
– *Атаки на основе решения модели.* Предполагается наличие доступа к предсказанной моделью метке. Чаще всего такие атаки основы на случайной инициализации и постепенном приближении к оригинальному изображению ([7], [8]).
– *Атаки на основе переносимости.* В [1] было открыто, что состязательные примеры могут быть успешно перенесены с одной модели на другую. В этой модели угроз предполагается наличие доступа или к полному датасету, на котором обучалась целевая модель, или к его части для обучения модели-суррогата ([15], [16]), для которой и будут синтезироваться состязательные примеры, а затем переноситься на атакуемую модель (здесь также требуется доступ к выходному распределению модели).

2) *Классификация по контролю над входом модели*
- **Прямой контроль над входом модели.** Злоумышленник напрямую может контролировать, в каком виде изображение попадет в модель.
- **Непрямой контроль над входом модели.** Злоумышленник не может контролировать пиксели, которые попадут в модель. Чаще всего речь идет об атаках в реальном мире, когда злоумышленник может менять окружающий мир, например наклеивать стикеры, менять внешность и т.п. Целью таких атак является подобрать такие возмущения, которые сохранятся через весь пайплайн данных и будут робастны относительно различных изменений окружающей среды, которые злоумышленник не сможет контролировать, например, освещение, контрастность, угол, под которым будет сделана фотография и т.п.

3) *Классификация по цели злоумышленника*
- **Нецелевая атака.** Атака считается успешной, если модель присваивает состязательному примеру любой из классов, не являющийся правильным.
- **Целевая атака.** Атака считается успешной лишь в том случае, если модель присвоила состязательному примеру заранее выбранный определенный класс.

4) *Классификация по итеративной мощности*
- **Атака в один шаг.** Злоумышленник обращается к целевой модели лишь один раз.
- **Итеративная атака.** Злоумышленник несколько раз обращается к целевой модели для итеративного обновления состязательного примера.

Можно провести классификацию атак по какому-то из аспектов вносимого возмущения: по универсальности, по норме, по оптимальности.

1) *Классификация по универсальности возмущения.*
- **Индивидуальные возмущения.** Целью атаки является поиск возмущения для отдельного изображения.
- **Универсальные возмущения.** Поиск возмущений, которые будут применимы ко многим изображениям из конкретной выборки.

2) *Классификация по норме возмущения.*
- **Возмущения по норме l_0.** Норма l_0 возмущения равна количеству пикселей изображения, которые изменяются при атаке. При этом диапазон изменения каждого пикселя не ограничен. Существует варианты атаки, где у злоумышленника стоит цель обмануть модель, изменяя при этом минимальное число пикселей, расположенных произвольно относительно друг друга ([17], l_0-версия атаки Карлини и Вагнера [18], [3]). Есть и другой вариант - состязательные патчи [19], где изменяется какой-то связный регион изображения, чаще всего прямоугольный.
- **Возмущения по норме l_2.** Евклидова норма вносимого возмущения. $\|\delta\|_2 = \sqrt{\sum_{i=1}^{n} \delta_i^2}$
- **Возмущения по норме l_∞.** Максимальное значение вносимого возмущения для отдельного пикселя. $|\delta\|_\infty = \max\{|x_i|\}_{i=1}^{n}$

Многие атаки имеют сразу по несколько версий для разных метрик, например атака Карлини и Вагнера [18], имеющая версии для каждой из трех метрик. Иногда рассматривают атаки по норме l_1: $\|\delta\|_1 = \sum_{i=1}^{n} |\delta_i|$ ([20], [21]), но она довольно мало популярна.

3) *Классификация по оптимальности.*
- **Оптимальное возмущение.** Атака проводится с целью найти минимальное возмущение, которое сможет изменить предсказание модели.
- **Ограниченное возмущение.** Атака считается успешной, если найденное возмущение ограничено сверху по норме заранее определенным значением и изменяет предсказание модели.

Были предложены методы атак не только для классификаторов изображений, но и для других задач компьютерного зрения, в том числе для детекции объектов, семантической сегментации, распознавания лиц и т.д. В дальнейшем мы рассмторим наиболее интересные с нашей точки зрения, разделив их по знаниям злоумышленника.

III. Атаки в режиме белого ящика

Самая первая предложенная атака, L-BFGS [1], относится к этому классу атак. Пусть $f : [0,1]^m\{p_1, \ldots, p_k\}$ - классификатор, принимающий на вход нормализованные изображения из m пикселей и возвращающий на выходе вероятности принадлежности к каждому из k возможных классов. Пусть также определена функция потерь $J_f : [0,1]^m \times \{1, \ldots, k\} \to \mathbb{R}^+$. Это может быть, к примеру,

кросс-энтропия, штрафующая за несоответствие предсказаний модели требуемой метке. Возьмем некоторое изображение x, которое модель верно классифицирует с меткой l. Пусть $t \neq l$ - целевая метка, и мы хотим чтобы ее выдавала модель. Тогда сформулируем следующую задачу оптимизации с ограничениями:

$$\min_{\delta} \|\delta\|_2 \quad \text{т.ч.} \quad \arg\max_{c} f_c(x+\delta) = t, \quad x+\delta \in [0,1]^m \tag{1}$$

Эту задача сложна, поэтому ищется приблизительное решение задачи с помощью алгоритма L-BFGS при ограничениях.

$$\min_{\delta} c|\delta| + J_f(x,t) \quad \text{т.ч.} \quad x+\delta \in [0,1]^m, \tag{2}$$

где c - константа, отвечающая за относительную важность каждого из слагаемых в оптимизируемой функции.

Карлини и Вагнер улучшили этот метод, предложив заменить нелинейные граничные условия на аналогичные, но более простые для оптимизации [18]. Во-первых, они предложили подобрать функцию g, для которой условие $g(x+\delta,t) \leq 0$ будет эквивалетно условию $\arg\max_c f_c(x+\delta) = t$. Проанализировав несколько вариантов, авторы остановились на следующем:

$$g(\hat{x},t) = (\max_{i \neq t}[Z(\hat{x})_i] - Z(\hat{x})_t)^+, \tag{3}$$

где $Z(\hat{x}) \in \mathbb{R}^k$ - логиты модели, то есть вход на последний слой softmax, а $u^+ = \max\{0,u\}$

Данную задачу можно было бы решать методом градиентного спуска, если бы не стояло граничного условия $x+\delta \in [0,1]^m$. Для него предлагается три подхода:

1) Проецируемый градиентный спуск: при каждой итерации градиентного спуска мы ограничиваем получающееся значение каждого пикселя так, чтобы оно принадлежало отрезку $[0,1]$.
2) Ограниченный градиентный спуск: вместо оригинальной функции $g(x+\delta)$ используется $g(\min(\max(x+\delta,0),1))$
3) Замена переменных: вводится новая переменная w, такая что:

$$\delta_i = \frac{1}{2}(\tanh w_i + 1) - x_i$$

Таким образом, при использовании замены переменных, задача оптимизации по норме l_2 примет вид:

$$\min \|\frac{1}{2}(\tanh w + 1) - x\|_2^2 + c \cdot g(\frac{1}{2}(\tanh w + 1), t), \tag{4}$$

Данная атака может быть обобщена на метрики l_0 и l_∞.

В случае атак по метрике l_∞ существует ряд методов, основанных на знаке градиента функции потерь по входу модели. Начало было положено в работе [22], где в нецелевой версии атаки предлагалось делать один шаг, сонаправленный с градиентом функции потерь по входу. Данный метод получил название FGSM. Если размер состязательного возмущения ограничен значением $\|\delta\|_p <= \epsilon$, то состязательный пример вычисляется по следующей формуле

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x,l)) \tag{5}$$

В статье [23] была предложена целевая версия этой атаки. Если ранее мы двигались в сторону от верной метки l, то теперь предлагается двигаться в сторону к целевой метке t.

$$x^{adv} = x - \epsilon \cdot \text{sign}(\nabla_x J(x,t)) \tag{6}$$

В базовом итеративной варианте [24] этой атаки предлагается инициализировать состязательное изображение исходным изображением, а затем совершить несколько итераций в соответствии со знаком градиента на текущей итерации и размером шага α. При этой на каждой итерации t необходимо совершать усечение значений текущего состязательного изображения для того чтобы оставаться в границах шара $x_{t+1} \in \mathbb{B}_\epsilon(x)$ и допустимых значениях пикселей $x_{t+1} \in [0,1]^m$:

$$\begin{aligned} x_0^{adv} &= x \\ x_{t+1}^{adv} &= Clip_{[0,1]^m,\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x J(x_t^{adv}, l))\} \end{aligned} \tag{7}$$

В статье [25] было предложено инициализировать x_0^{adv} случайной точкой внутри шара $\mathbb{B}_\epsilon(x)$. Также там предлагалось несколько раз повторить весь алгоритм с различными случайными инициализациями для поиска состязательного возмущения, которое приведет к наибольшему значению функции потерь. Этот метод был назван проецируемым градиентным спуском (PGD).

Хотя итеративный метод является довольно мощным с точки зрения успешности поиска состязательного возмущения, он обладает и некоторыми недостатками. В частности, было замечено, что подбираемые этим методом состязательные примеры плохо переносятся на другие модели ([26]) из-за того, что сильно ориентируются на границы принятия решений данной модели, как бы «переобучаясь» под нее.

Для лучшей переносимости в [26] было предложено использовать градиентный спуск с импульсом. Вектор импульса g_{t+1} учитывает последние обновления, накапливая среднее градиентов:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, l)}{\|\nabla_x J(x_t^{adv}, l)\|_1} \tag{8}$$

Соответственно, обновление состязательного примера происходит по формуле:

$$x_{t+1}^{adv} = Clip_{[0,1]^m,\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})\} \tag{9}$$

В другой серии работ было предложено использовать различные аугментации к изображению во время итеративного знакового градиентного метода. В работе [27] предлагается на каждой итерации с некоторой вероятностью применять к текущему состязательному примеру случайное изменение размера и случайное добавление нулей по краям изображения. Тогда обновление на текущей итерации примет вид:

$$x_{t+1}^{adv} = \text{Clip}_{[0,1]^m,\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(\nabla_x J(T(x_t^{adv}; p), y_{true}))\}, \tag{10}$$

где $T(x_t^{adv}; p)$ - случайная функция трансформации, которая определяется следующим образом:

$$T(x_t^{adv}; p) = \begin{cases} T(x_t^{adv}) & \text{с вероятностью } p \\ x_t^{adv} & \text{с вероятностью } 1-p \end{cases} \tag{11}$$

Этот метод получил название Diverse Inputs Iterative Fast Gradient Sign Method - M-DI^2-FGSM.

В [28] предложено в качестве аугментаций осуществлять трансляцию изображения, а в [29] - изменение масштаба

изображения. Также в [29] было предложено вместо импульса использовать момент Нестерова [30].

Другим слабым метом базовой итеративной версии алгоритма является наличие гиперпараметра - размера шага α. Кроме того, размер шага в базовой версии алгоритма является постоянным на протяжении всех итераций, что можно препятствовать нахождению состязательных примеров, приводящих к максимальному значению функции потерь. Для борьбы с этим в статье [31] был предложен адаптивный алгоритм Auto-PGD. Этот алгоритм разбивается на несколько эпизодов, и после каждого эпизода решается, нужно ли уменьшить размер шага вдвое. Размер шага уменьшается вдвое в том случае, если в течение прошедшего эпизода не был достигнут значительный прогресс.

В одной из версий Auto-PGD авторы предлагают вместо кросс-энтропии использовать новый вариант функции потерь - Difference of Logits Ratio (DLR). Его преимущество заключается в том, что он является инвариантным относительно масштабирования как выходного распределения модели, так и логитов модели. В результате этого алгоритм с такой функцией потерь может успешно найти состязательный пример даже для некоторых защищенных моделей. Если π - логиты в порядке убывания, то:

$$DLR(x, y) = -\frac{Z_y - \max\limits_{i \neq y} Z_i}{Z_{\pi_1} - Z_{\pi_3}} \tag{12}$$

Геометрический метод нахождения ближайшего по норме l_2 состязательного изображения под название DeepFool был предложен в [32]. Основа метода - поиск ближайшей к изображению границы разделения с каким-то другим классом и движении по направлению к этой границе. Если текущее изображение имеет класс l, то получить разделящую границу с каким-то классом c вблизи текущего изображения x_t можно с помощью линейной аппроксимации по Тейлору:

$$\pi(z) : f_c(x_t) - f_l(x_t) + \langle \nabla f_c(x_t) - \nabla f_l(x_t), z - x_t \rangle = 0 \tag{13}$$

Для нахождения ближайшего состязательного изображения необходимо определить класс s, такой что расстояние до разделяющей гиперплоскости между l и искомым классом было минимальным среди всех неправильных классов:

$$s = \arg\min_{l \neq c} \frac{|f_l(x_t) - f_c(x_t)|}{\|\nabla f_l(x_t) - \nabla f_c(x_t)\|_2} \tag{14}$$

Обновление текущего состязательного изображения происходит по нормали к этой ближайшей разделяющей гиперплоскости:

$$x_{t+1} = x_t + \frac{|f_l(x_t) - f_s(x_t)|}{\|\nabla f_l(x_t) - \nabla f_s(x_t)\|_2^2} (\nabla f_l(x_t) - \nabla f_s(x_t)) \tag{15}$$

Улучшенная версия DeepFool была предложена в [20].

В статье [2] авторы показали существования универсальных состязательных возмущений, которые могут быть наложены на множество изображений из выборки и привести к неверной классификации. Конкретно, авторы нашли $\|\delta\|_p \leq \epsilon$, что для него выполняется:

$$\mathbb{P}_{x \sim \mathcal{D}} \Big(\arg\max_c f_c(x + \delta) \neq \arg\max_c f_c(x) \Big) \geq 1 - \eta \tag{16}$$

для какого-то небольшого η.

Для поиска такого возмущения для каждого изображения $x_i \in X$ из обучающей выборки проверяется, меняется ли предсказываемая классификатором метка при наложении текущего универсального возмущения δ. Если же $\arg\max_c f_c(x + \delta) = \arg\max_c f_c(x)$, то вычисляется минимальная добавка $\Delta\delta_i$, которая способна привести на границу разделения классов (для его поиска используется DeepFool):

$$\begin{aligned} &\Delta\delta_i \leftarrow \arg\min_r \|r\|_2 \\ &\text{т.ч.} \quad \arg\max_c f_c(x_i + \delta + r) \neq \arg\max_c f_c(x_i) \end{aligned} \tag{17}$$

Далее универсальное возмущение обновляется с помощью оператора проецирования на шар $\mathbb{B}_\epsilon(0)$ по норме p:

$$\delta \leftarrow \Pi_\epsilon(\delta + \Delta\delta_i) \tag{18}$$

В работе [19] была предложена атака для поиска универсального состязательного патча. Такой патч при наложении на множество разных изображений из выборки в любое место изображения должен приводить к неправильной классификации в целевой класс. Авторы не стремятся к тому, чтобы патч был незаметен для человека. Единственное ограничение, которое они между собой ставят - чтобы патч занимал не более какой-то доли площади изображения (например, 10%), то есть это атака по норме l_0.

Пусть $x \in \mathbb{R}^{w \times h \times c}$ - изображение, p - подбираемый патч, l - место, куда накладывается патч, t - преобразования, применяемые к патчу (повороты, масштабирование). Можно определить оператор наложения патча $A(p, x, l, t)$, который применяет к патчу p преобразование t, а затем накладывает преобразованный патч на изображение x на место l.

Тогда задачу можно определить формально следующим образом. Если X - обучающая выборка, T - распределение преобразований, L - распределение мест, куда можно наложить патч, а y_t - целевая метка, то решается следующая задача оптимизации:

$$\hat{p} = \arg\max_p \mathbb{E}_{x \sim X, t \sim T, l \sim L} [\log f_{y_t}(A(p, x, l, t))] \tag{19}$$

IV. Атаки в режиме черного ящика

A. На основе выходного распределения

Рассмотренные в предыдущей главе методы требует доступа к весам модели для вычисления градиента. Если же у злоумышленника есть доступ лишь до выходного распределения модели, то необходимо или аппроксимировать реальное значение градиента с помощью каких-то алгоритмов, или применять другие оптимизационные алгоритмы.

Первыми предложили аппроксимировать градиент в работе [4], используя метод конечных разностей. Авторы проводят атаку, подобную атаке Карлини и Вагнера. При этом в качестве $g(x, t)$ авторы берут не логиты, а выходы всей модели, применяя к ним логарифм для уменьшения доминирования данного слагаемого:

$$g(\hat{x}, t) = \max\{\max_{c \neq t} \log[f_c(\hat{x})] - \log[f_t(\hat{x})], -\kappa\} \tag{20}$$

Реальный градиент $\nabla_x g(x, t)$ на каждой итерации градиентного спуска заменяется на его аппроксимацию $v(x, t)$.

Для каждого пикселя i из входного пространства это приблизительное значение считается с помощью конечной разности:

$$v_i(\hat{x}, t) = \frac{\partial f(\hat{x})}{\partial \hat{x}_i} \approx \frac{f(\hat{x} + h\mathbf{e}_i) - f(\hat{x} - h\mathbf{e}_i)}{2h}, \quad (21)$$

где h - маленькая константа (0.0001 у авторов), $\mathbf{e}$ - стандартный базисный вектор, у которого все элементы равны нулю, кроме i-го, равного 1.

Разумеется, в таком виде алгоритм требует крайне много вычислений лишь для одной итерации градиентного спуска. Поэтому авторы сразу предлагают несколько улучшений, позволяющих сократить вычислительную сложность, таких как применение координатного спуска, сокращение размерности атаки, выбор и оптимизация только лишь наиболее важных пикселей. В других работах, также основанных на идее аппроксимации градиента, предлагаются алгоритмы, которые еще больше помогают сократить количество запросов к модели. Среди них можно выделить [33], [5].

Среди алгоритмов, не требующих для оптимизации градиента, можно выделить *эволюционные алгоритмы*. В работе [3] предлагается атака по норме l_0, где с помощью дифференциальной эволюции выбирается всего лишь 1 пиксель, изменение которого способно обмануть нейронную сеть. То есть формально задача имеет следующий вид:

$$\underset{\delta}{\text{maximize}}\ f_t(x + \delta) \quad \text{т.ч.} \quad \|\delta\|_0 \leq 1 \quad (22)$$

При такой постановке решение задачи будет состоять из 5 значений: двух координат искомого пикселя и трех значений интенсивности в палитре RGB, на которые нужно изменить значения этого пикселя.

В работе [14] также с помощью эволюционного алгоритма синтезируются бессмысленные для человека изображения, которые классифицируются моделями с большой (99.99%) уверенностью. Алгоритм инициализируется сэмплированным из шума изображением, а затем случайным образом выбираются пиксели для мутации и значения, на которые изменяется интенсивность пикселей.

В работе [6] авторы использовали наблюдения предыдущх работ о том, что зачастую при успешной атаке по нормам l_∞, l_2 изменяются не случайные пиксели, а квадратные регионы изображения. Для поиска этих квадратов и значений, на которых в этих квадратах меняются интенсивности пикселей, авторы использовали *случайный поиск* [34]. Это позволило авторам добиться более низкого среднего числа запросов к модели, необходимого для успешной атаки.

B. На основе выходной метки

Первая атака в этой модели угроз была предложена в работе [7]. Ее идея состоит в старте с изображения, которое и так уже классифицируется с целевой меткой t и постепенном приближении к оригинальному изображению x. Каждая итерация алгоритма состоит из 3 шагов:

1) Выбрать *случайное направление* $\delta_k \sim \mathcal{N}(0, 1)$, затем провести масштабирование и усечение, так чтобы каждый пиксель лежал в допустимом диапазоне:
$$\hat{x}_{t-1} + \delta_t \in [0, 1]^m, \quad \|\delta_t\|_2 = \eta \cdot \|x, \hat{x}_{t-1}\|_2 \quad (23)$$
2) Спроектировать δ_t на сферу вокруг оригинального изображения, так чтобы:
$$\|x, \hat{x}_{t-1} + \delta_t\| = \|x, \hat{x}_{t-1}\|, \quad \hat{x}_{t-1} + \delta_t \in [0, 1] \quad (24)$$
3) Сделать маленький шаг в направлении оригинального изображения, так чтобы расстояние от текущего состязательного изображения до оригинального уменьшилось бы на μ:
$$\begin{aligned} &\|x, \hat{x}_{t-1}\| - \|x, \hat{x}_{t-1} + \delta_t\| = \mu \cdot \|x, \hat{x}_{t-1}\| \\ &\hat{x}_{t-1} + \delta_t \in [0, 1] \end{aligned} \quad (25)$$

Атака позволяет достичь близких к оригиналу состязательных изображений, однако для этого требуется много запросов к модели. Дальнейшие алгоритмы в этой модели угроз улучшают описанную выше процедуру, выбирая уже не случайное направление на шаге 1. Среди таких алгоритмов особо можно выделить [8], где для выбора направления осуществляется бинарный поиск границы разделения классов, а затем для более эффективного смещения к оригинальному изображению происходит аппроксимация направления градиента.

C. На основе переноса

В первой статье [1], посвященной атакам уклонением, было установлено, что состязальные примеры хорошо переносятся между моделями, обученными на одной и той же выборке данных. Однако у злоумышленника не всегда может быть доступ к данным, на которых обучалась атакуемая модель.

В работе [15] была предложена атака, где злоумышленник может наблюдать за выходным распределением целевой модели (не имея доступа к данным, на которых она обучалась) на произвольных сэмплах и использует эту информацию как метки для обучения модели-суррогата. Затем состязательные примеры синтезируются для суррогата и переносятся на целевую модель. Было выяснено, что случайные шумные изображения плохо подходят для обучения суррогата. Авторы предложили синтезировать изображения для обучения суррогата на основе определения направлений, в которых варьируются выходы целевой модели. Эти направления определяются с помощью знака якобиана $J_f = [\partial f_i(x)/\partial x_j]_{ij}$ выходов модели по входам.

Обучение суррогатной модели разделено на эпохи. Во время каждой эпохи текущая выборка из изображений отправляется на разметку целевой модели, а модель суррогат затем обучаются на полученных выходных распределениях. После этого происходит процедура пополнения выборки $\mathcal{D}$, получившая название Jacobian-based dataset augmentation:

$$\mathcal{D}_{t+1} \leftarrow \left\{x + \lambda \cdot \text{sign}\left([\partial f_i(x)/\partial x_j]_{ij}\right) : x \in \mathcal{D}_t\right\} \cup \mathcal{D}_t \quad (26)$$

В своей следующей работе [16] эти же авторы показали, что состязательные примеры могут переноситься даже между разными классами моделей, то есть например между нейронными сетями и простыми линейными моделями. Поэтому предложенная выше атака может быть осуществлена даже тогда, когда мы не знаем класс атакуемой модели.

Однако в работе [35] было показано, что хорошую переносимость имеют лишь нецелевые состязательные

примеры. В случае же целевой атаки состязательные примеры крайне редко переносятся с целевой меткой. Авторы выдвинули гипотезу, что у разных моделей существуют регионы входного пространства, которые классифицируются разными моделями одинаково, однако существуют атаки попадают туда довольно редко. Для более вероятного попадания в регионы, которые разные модели классифицируют одинако неправильно с целевой меткой, предлагается синтезировать состязательные примеры на ансамблях моделей.

V. Атаки в реальном мире

Первое исследование влияния условий реального мира на состязательные изображения было проведено в работе [24]. Авторы распечатывали состязательные примеры на бумаге и затем фотографировали их на мобильный телефон, проверяя, продолжат ли фотографии распечатанных состязательных примеров обманывать классификатор. Оказалось, что изображения, на которые наложены состязательные возмущения с помощью атаки FGSM, гораздо лучше сохраняют свою состязательность по сравнению с теми, которые подверглись атаке итеративной версией FGSM.

Аналогичные результаты были получены и при наложении на состязательные изображения различных трансформаций, таких как изменение контрастности и яркости, гауссовское размытие, гауссовское зашумление и кодирование JPEG. Таким образом, авторы сделали вывод, что атака FGSM более робастна в условиях реального мира по сравнению с ее итеративной версией.

В работе [9] решелась задача поиска таких состязательных примеров, которые бы были робастными в реальной жизни относительно разных физических условий, при разных углах обзора и разном расстоянии до камеры. Авторы предложили алгоритм Robust Physical Perturbations (RP2), в котором состязательные примеры сначала синтезируются в лабораторных условиях, а затем распечатываются и переносятся в реальную жизнь. Успешная реализация атаки была осуществлена для модели, классифицирущей дорожные знаки.

По сравнению с обычной постановкой задачи 2 авторы учитывают, что в реальном мире любой физический объект будет подвергаться физическим и цифровым преобразованиям, для моделирования изображений со всевозможными преобразованиями вводится распределение X^V. Также авторы вносят ограничения, согласно которым наносить изменения можно только на сам дорожный знак, но не на окружающие условия - то есть допустимая область нанесения возмущения для каждого изображения x ограничена маской M_x.

В этой атаке учтена также ограниченная способность принтера передавать цвета. Для этого используется оценка непечатности (non-printability score): для данного множества возможных к печати цветов P и множества цветов $R(\delta)$, необходимых для состязательного возмущения эта оценка вычисляется следующим образом

$$NPS = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'| \tag{27}$$

В результате, в целевой версии атаки с меткой y_t авторы решают следующую задачу оптимизации:

$$\arg\min_{\delta} \quad \lambda \|M_x \cdot \delta\|_p + \mathbb{E}_{x_i \sim X^V} J_f(x_i + T_i(M_x \cdot \delta), y_t) + NPS, \tag{28}$$

где функция T используется для выравнивания вносимых возмущений: при повороте дорожного знака должно поворачиваться и возмущение.

В следующей своей статье эта же группа авторов [11] модифицировала описанную выше атаку для обмана детектора дорожных знаков. Главным нововведением стала замена функции потерь: вместо кросс-энтропии, которая используется для классификации, авторы предложили использовать максимальную уверенность детектора YOLOv2 в классе объекта среди всех рамок, на которые YOLOv2 разбивает изображение.

В работе [10] была проведена атака на систему распознавания лиц (реализованную в виде классификатора) с помощью надевания на человека специальных состязательных очков. Оправа таких очков будет подобрана так, чтобы человек мог выдать себя за желаемую целевую личность с меткой y_t.

По сравнению с обычной постановкой задачи, где максимизируется вероятность целевого класса, авторы предлагают учитывать свойство, согласно которому цвета на естественных изображениях меняются постепенно в пределах какой-то окрестности. Также, резкие переходы цветов между соседними пикселями могут быть просто плохо захвачены камерой. Поэтому предлагается минимизировать общую дисперсию, которая для возмущения δ определяется как:

$$TV(\delta) = \sum_{i,j} \left((\delta_{i,j} - \delta_{i+1,j})^2 + (\delta_{i,j} - \delta_{i,j+1})^2 \right)^{\frac{1}{2}} \tag{29}$$

Таким образом, если для данного человека рассматривается выборка X его фотографий в обычных очках, то необходимо решить следующую задачу поиска универсальной состязательной оправы очков:

$$\arg\min_{\delta} \sum_{x \in X} J_f(x + \delta, y_t) + \lambda_1 TV(\delta) + \lambda_2 NPS(\delta) \tag{30}$$

В работе [36] авторы продемонстрировали метод создания состязательных примеров, которые будут устойчивы к шуму, искажениям и аффинным преобразованиям. Более того, они создали реальный физический 3d-объект, который при разных углах фотографирования будет обманывать модель в большой доле случаев.

Предложенный фреймворк получил название Expectation Over Transformation и предназначен для поиска состязательных примеров, робастных к распределению преобразований T. Авторы формулируют задачу следующим образом:

$$\begin{aligned} &\arg\max_{x^{adv}} \quad \mathbb{E}_{t \sim T}[\log f_{y_t}(t(x^{adv}))] \\ &\text{т.ч.} \quad \mathbb{E}_{t \sim T} \|t(x^{adv}) - t(x)\| < \epsilon, \quad x^{adv} \in [0,1]^n, \end{aligned} \tag{31}$$

где d - функция расстояния.

При оптимизации целевого функционала на каждом шаге градиентного спуска градиент математического ожидания аппроксимируется сэмплированием независимых

преобразований. Для решения поставленной оптимизационной задачи авторы переформулируют ее подобно задаче из атаки Карлини и Вагнера:

$$\arg\max_{x^{adv}} \quad \Big(\mathbb{E}_{t\sim T}[\log f_{y_t}(t(x^{adv}))] - \lambda \mathbb{E}_{t\sim T}\|t(x^{adv}) - t(x)\|\Big) \tag{32}$$

Однако в работе [12] было показано, что множество преобразований, рассмотренных в предыдущей работе, может быть недостаточно для некоторых атак в реальном мире. Авторы пытались создать футболку с состязательным принтом, которая сделана бы человека невидимым для детекторов. Футболка при движении будет деформироваться так, что обычных поворотов, изменений яркости и прочих подобных трансформаций будет недостаточно для описания ее новых положений относительно камеры. Поэтому авторы для описания деформаций футболки предлагают использовать множество преобразований *thin plate spline(TPS)* [37], которое широко используется в качестве нежесткой модели преобразований при выравнивании изображения и сопоставления форм. Это множество включает в себя как афинные, так и неаффинные компоненты деформации. Лишь после добавления неаффинных преобразований авторы добиваются успеха в этой атаке.

VI. Заключение

Обзор предложенных за последние годы состязательных атак уклонением на системы компьютерного зрения показывает, что современные нейронные сети являются уязвимыми в разных моделях угроз. Более того, во многих статьях были проведены атаки в реальном мире. Значительное число работ, посвященное атакам в режиме черного ящика, демонстрирует возможность атак даже при минимальных знаниях злоумышленника про модель. Поэтому крайне важно для каждой модели, внедряемой в промышленную эксплуатацию, оценивать ее робастность относительно атак и оценивать риски, связанные с ее потенциальным обманом в разных моделях угроз.

Благодарности

Мы благодарны сотрудникам кафедры Информационной безопасности факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова за ценные обсуждения данной работы.

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект»

Статья является продолжением серии публикаций, посвященных устойчивым моделям машинного обучения ([38], [39], [40]). Она подготовлена в рамках проекта кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова по созданию и развитию магистерской программы ”Искусственный интеллект в кибербезопасности”[41].

Список литературы

[1] Intriguing properties of neural networks / Christian Szegedy, Wojciech Zaremba, Ilya Sutskever et al. // arXiv preprint arXiv:1312.6199. — 2013.

[2] Universal adversarial perturbations / Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 1765–1773.

[3] Su Jiawei, Vargas Danilo Vasconcellos, Sakurai Kouichi. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. — 2019. — Vol. 23, no. 5. — P. 828–841.

[4] Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models / Pin-Yu Chen, Huan Zhang, Yash Sharma et al. // Proceedings of the 10th ACM workshop on artificial intelligence and security. — 2017. — P. 15–26.

[5] Ilyas Andrew, Engstrom Logan, Madry Aleksander. Prior convictions: Black-box adversarial attacks with bandits and priors // arXiv preprint arXiv:1807.07978. — 2018.

[6] Square attack: a query-efficient black-box adversarial attack via random search / Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, Matthias Hein // European Conference on Computer Vision / Springer. — 2020. — P. 484–501.

[7] Brendel Wieland, Rauber Jonas, Bethge Matthias. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models // arXiv preprint arXiv:1712.04248. — 2017.

[8] Chen Jianbo, Jordan Michael I, Wainwright Martin J. Hopskipjumpattack: A query-efficient decision-based attack // 2020 ieee symposium on security and privacy (sp) / IEEE. — 2020. — P. 1277–1294.

[9] Robust physical-world attacks on deep learning visual classification / Kevin Eykholt, Ivan Evtimov, Earlence Fernandes et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 1625–1634.

[10] Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition / Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K Reiter // Proceedings of the 2016 acm sigsac conference on computer and communications security. — 2016. — P. 1528–1540.

[11] Physical adversarial examples for object detectors / Dawn Song, Kevin Eykholt, Ivan Evtimov et al. // 12th USENIX workshop on offensive technologies (WOOT 18). — 2018.

[12] Adversarial t-shirt! evading person detectors in a physical world / Kaidi Xu, Gaoyuan Zhang, Sijia Liu et al. // European conference on computer vision / Springer. — 2020. — P. 665–681.

[13] Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps / Zhichuang Sun, Ruimin Sun, Long Lu, Alan Mislove // 30th USENIX Security Symposium (USENIX Security 21). — 2021. — P. 1955–1972.

[14] Nguyen Anh, Yosinski Jason, Clune Jeff. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — P. 427–436.

[15] Practical black-box attacks against machine learning / Nicolas Papernot, Patrick McDaniel, Ian Goodfellow et al. // Proceedings of the 2017 ACM on Asia conference on computer and communications security. — 2017. — P. 506–519.

[16] Papernot Nicolas, McDaniel Patrick, Goodfellow Ian. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples // arXiv preprint arXiv:1605.07277. — 2016.

[17] The limitations of deep learning in adversarial settings / Nicolas Papernot, Patrick McDaniel, Somesh Jha et al. // 2016 IEEE European symposium on security and privacy (EuroS&P) / IEEE. — 2016. — P. 372–387.

[18] Carlini Nicholas, Wagner David. Towards evaluating the robustness of neural networks // 2017 ieee symposium on security and privacy (sp) / IEEE. — 2017. — P. 39–57.

[19] Adversarial patch / Tom B Brown, Dandelion Mané, Aurko Roy et al. // arXiv preprint arXiv:1712.09665. — 2017.

[20] Croce Francesco, Hein Matthias. Minimally distorted adversarial examples with a fast adaptive boundary attack // International Conference on Machine Learning / PMLR. — 2020. — P. 2196–2205.

[21] Sharma Yash, Chen Pin-Yu. Breaking the madry defense model with l 1-based adversarial examples // arXiv preprint arXiv:1710.10733. — 2017.

[22] Goodfellow Ian J, Shlens Jonathon, Szegedy Christian. Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. — 2014.

[23] Kurakin Alexey, Goodfellow Ian, Bengio Samy. Adversarial machine learning at scale // arXiv preprint arXiv:1611.01236. — 2016.

[24] Kurakin Alexey, Goodfellow Ian J, Bengio Samy. Adversarial examples in the physical world // Artificial intelligence safety and security. — Chapman and Hall/CRC, 2018. — P. 99–112.

[25] Towards deep learning models resistant to adversarial attacks / Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt et al. // arXiv preprint arXiv:1706.06083. — 2017.

[26] Boosting adversarial attacks with momentum / Yinpeng Dong, Fangzhou Liao, Tianyu Pang et al. // Proceedings of the IEEE

conference on computer vision and pattern recognition. — 2018. — P. 9185–9193.
[27] Improving transferability of adversarial examples with input diversity / Cihang Xie, Zhishuai Zhang, Yuyin Zhou et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — P. 2730–2739.
[28] Evading defenses to transferable adversarial examples by translation-invariant attacks / Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — P. 4312–4321.
[29] Nesterov accelerated gradient and scale invariance for adversarial attacks / Jiadong Lin, Chuanbiao Song, Kun He et al. // arXiv preprint arXiv:1908.06281. — 2019.
[30] Nesterov Yurii E. A method for solving the convex programming problem with convergence rate o (1/k^ 2) // Dokl. akad. nauk Sssr. — Vol. 269. — 1983. — P. 543–547.
[31] Croce Francesco, Hein Matthias. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks // International conference on machine learning / PMLR. — 2020. — P. 2206–2216.
[32] Moosavi-Dezfooli Seyed-Mohsen, Fawzi Alhussein, Frossard Pascal. Deepfool: a simple and accurate method to fool deep neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 2574–2582.
[33] Black-box adversarial attacks with limited queries and information / Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin // International Conference on Machine Learning / PMLR. — 2018. — P. 2137–2146.
[34] Rastrigin LA. The convergence of the random search method in the extremal control of a many parameter system // Automaton & Remote Control. — 1963. — Vol. 24. — P. 1337–1342.
[35] Delving into transferable adversarial examples and black-box attacks / Yanpei Liu, Xinyun Chen, Chang Liu, Dawn Song // arXiv preprint arXiv:1611.02770. — 2016.
[36] Synthesizing robust adversarial examples / Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok // International conference on machine learning / PMLR. — 2018. — P. 284–293.
[37] Bookstein Fred L. Principal warps: Thin-plate splines and the decomposition of deformations // IEEE Transactions on pattern analysis and machine intelligence. — 1989. — Vol. 11, no. 6. — P. 567–585.
[38] Ilyushin Eugene, Namiot Dmitry, Chizhov Ivan. Attacks on machine learning systems-common problems and methods // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 3. — P. 17–22.
[39] Dmitry Namiot, Eugene Ilyushin, Ivan Chizhov. On a formal verification of machine learning systems // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 5. — P. 30–34.
[40] Huayu Li, Dmitry Namiot. A survey of adversarial attacks and defenses for image data on deep learning // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 5. — P. 9–16.
[41] Artificial intelligence in cybersecurity. — https://cs.msu.ru/node/3732. — Retrieved: May, 2022. (in Russian).

A survey and systematization of evasion attacks in computer vision

Vasily Kostyumov

***Abstract*—Deep learning has received a lot of attention from the scientific community in recent years due to excellent results in various areas of tasks, including computer vision. For example, in the problem of image classification, some authors even announced that neural networks have surpassed humans in the accuracy of recognition. However, the discovery of adversarial examples for machine learning models has shown that modern computer vision architectures are very vulnerable to adversaries and additional attention is required when implementing them in critical infrastructure areas. Since then, many new attacks in different threat models have been proposed and the possibility of such attacks in the real world has been shown. At the same time, no protection method has been proposed so far that would be reliable against existing attacks, not to mention guarantees against the entire possible set of threats. This article discusses and systematizes evasion attacks in the field of computer vision. In this type of attack which is most popular, an adversary can only interact with the model during inference and change its input.**

***Keywords*—machine learning, computer vision, adversarial attacks**

References

[1] Intriguing properties of neural networks / Christian Szegedy, Wojciech Zaremba, Ilya Sutskever et al. // arXiv preprint arXiv:1312.6199. — 2013.

[2] Universal adversarial perturbations / Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2017. — P. 1765–1773.

[3] Su Jiawei, Vargas Danilo Vasconcellos, Sakurai Kouichi. One pixel attack for fooling deep neural networks // IEEE Transactions on Evolutionary Computation. — 2019. — Vol. 23, no. 5. — P. 828–841.

[4] Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models / Pin-Yu Chen, Huan Zhang, Yash Sharma et al. // Proceedings of the 10th ACM workshop on artificial intelligence and security. — 2017. — P. 15–26.

[5] Ilyas Andrew, Engstrom Logan, Madry Aleksander. Prior convictions: Black-box adversarial attacks with bandits and priors // arXiv preprint arXiv:1807.07978. — 2018.

[6] Square attack: a query-efficient black-box adversarial attack via random search / Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, Matthias Hein // European Conference on Computer Vision / Springer. — 2020. — P. 484–501.

[7] Brendel Wieland, Rauber Jonas, Bethge Matthias. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models // arXiv preprint arXiv:1712.04248. — 2017.

[8] Chen Jianbo, Jordan Michael I, Wainwright Martin J. Hopskipjumpattack: A query-efficient decision-based attack // 2020 ieee symposium on security and privacy (sp) / IEEE. — 2020. — P. 1277–1294.

[9] Robust physical-world attacks on deep learning visual classification / Kevin Eykholt, Ivan Evtimov, Earlence Fernandes et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 1625–1634.

[10] Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition / Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K Reiter // Proceedings of the 2016 acm sigsac conference on computer and communications security. — 2016. — P. 1528–1540.

[11] Physical adversarial examples for object detectors / Dawn Song, Kevin Eykholt, Ivan Evtimov et al. // 12th USENIX workshop on offensive technologies (WOOT 18). — 2018.

[12] Adversarial t-shirt! evading person detectors in a physical world / Kaidi Xu, Gaoyuan Zhang, Sijia Liu et al. // European conference on computer vision / Springer. — 2020. — P. 665–681.

[13] Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps / Zhichuang Sun, Ruimin Sun, Long Lu, Alan Mislove // 30th USENIX Security Symposium (USENIX Security 21). — 2021. — P. 1955–1972.

[14] Nguyen Anh, Yosinski Jason, Clune Jeff. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — P. 427–436.

[15] Practical black-box attacks against machine learning / Nicolas Papernot, Patrick McDaniel, Ian Goodfellow et al. // Proceedings of the 2017 ACM on Asia conference on computer and communications security. — 2017. — P. 506–519.

[16] Papernot Nicolas, McDaniel Patrick, Goodfellow Ian. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples // arXiv preprint arXiv:1605.07277. — 2016.

[17] The limitations of deep learning in adversarial settings / Nicolas Papernot, Patrick McDaniel, Somesh Jha et al. // 2016 IEEE European symposium on security and privacy (EuroS&P) / IEEE. — 2016. — P. 372–387.

[18] Carlini Nicholas, Wagner David. Towards evaluating the robustness of neural networks // 2017 ieee symposium on security and privacy (sp) / IEEE. — 2017. — P. 39–57.

[19] Adversarial patch / Tom B Brown, Dandelion Mané, Aurko Roy et al. // arXiv preprint arXiv:1712.09665. — 2017.

[20] Croce Francesco, Hein Matthias. Minimally distorted adversarial examples with a fast adaptive boundary attack // International Conference on Machine Learning / PMLR. — 2020. — P. 2196–2205.

[21] Sharma Yash, Chen Pin-Yu. Breaking the madry defense model with l 1-based adversarial examples // arXiv preprint arXiv:1710.10733. — 2017.

[22] Goodfellow Ian J, Shlens Jonathon, Szegedy Christian. Explaining and harnessing adversarial examples // arXiv preprint arXiv:1412.6572. — 2014.

[23] Kurakin Alexey, Goodfellow Ian, Bengio Samy. Adversarial machine learning at scale // arXiv preprint arXiv:1611.01236. — 2016.

[24] Kurakin Alexey, Goodfellow Ian J, Bengio Samy. Adversarial examples in the physical world // Artificial intelligence safety and security. — Chapman and Hall/CRC, 2018. — P. 99–112.

[25] Towards deep learning models resistant to adversarial attacks / Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt et al. // arXiv preprint arXiv:1706.06083. — 2017.

[26] Boosting adversarial attacks with momentum / Yinpeng Dong, Fangzhou Liao, Tianyu Pang et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 9185–9193.

[27] Improving transferability of adversarial examples with input diversity / Cihang Xie, Zhishuai Zhang, Yuyin Zhou et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — P. 2730–2739.

[28] Evading defenses to transferable adversarial examples by translation-invariant attacks / Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2019. — P. 4312–4321.

[29] Nesterov accelerated gradient and scale invariance for adversarial attacks / Jiadong Lin, Chuanbiao Song, Kun He et al. // arXiv preprint arXiv:1908.06281. — 2019.

[30] Nesterov Yurii E. A method for solving the convex programming problem with convergence rate o (1/k^ 2) // Dokl. akad. nauk Sssr. — Vol. 269. — 1983. — P. 543–547.

[31] Croce Francesco, Hein Matthias. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks // International conference on machine learning / PMLR. — 2020. — P. 2206–2216.

[32] Moosavi-Dezfooli Seyed-Mohsen, Fawzi Alhussein, Frossard Pascal. Deepfool: a simple and accurate method to fool deep neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — P. 2574–2582.
[33] Black-box adversarial attacks with limited queries and information / Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin // International Conference on Machine Learning / PMLR. — 2018. — P. 2137–2146.
[34] Rastrigin LA. The convergence of the random search method in the extremal control of a many parameter system // Automaton & Remote Control. — 1963. — Vol. 24. — P. 1337–1342.
[35] Delving into transferable adversarial examples and black-box attacks / Yanpei Liu, Xinyun Chen, Chang Liu, Dawn Song // arXiv preprint arXiv:1611.02770. — 2016.
[36] Synthesizing robust adversarial examples / Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok // International conference on machine learning / PMLR. — 2018. — P. 284–293.
[37] Bookstein Fred L. Principal warps: Thin-plate splines and the decomposition of deformations // IEEE Transactions on pattern analysis and machine intelligence. — 1989. — Vol. 11, no. 6. — P. 567–585.
[38] Ilyushin Eugene, Namiot Dmitry, Chizhov Ivan. Attacks on machine learning systems-common problems and methods // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 3. — P. 17–22.
[39] Dmitry Namiot, Eugene Ilyushin, Ivan Chizhov. On a formal verification of machine learning systems // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 5. — P. 30–34.
[40] Huayu Li, Dmitry Namiot. A survey of adversarial attacks and defenses for image data on deep learning // International Journal of Open Information Technologies. — 2022. — Vol. 10, no. 5. — P. 9–16.
[41] Artificial intelligence in cybersecurity. — https://cs.msu.ru/node/3732. — Retrieved: May, 2022. (in Russian).