Обзор состязательных атак и методов защиты для детекторов объектов

Е. А. Чехонина, В. В. Костюмов

***Аннотация*—В настоящее время детекция объектов — одна из самых востребованных задач глубоких нейронных сетей, которая находит применение во многих критически важных областях: обработка естественного языка, обработка бóльших данных, анализ ДНК, управление автономными транспортными средствами. Однако недавние исследования показывают, что системы детекции объектов чувствительны к небольшим возмущениям во входных данных. Они незаметны для человеческого глаза, но могут полностью обмануть детектор. Уязвимость моделей для состязательных атак затрудняет их внедрение в реальные приложения. Существующие состязательные атаки можно разделить на цифровые и физические. Атаки в цифровом мире показывают хороший результат в лабораторных условиях, но теряют свою эффективность в реальном мире, в отличие от физических атак. Методы защиты можно разделить на эмпирические и сертифицированные. Последние имеют теоретически подтверждённую гарантию надёжности. Эмпирические же в будущем могут быть взломаны более сложными состязательными атаками. Несмотря на популярность темы состязательной робастности нейронных сетей, бóльшая часть работ посвящена задаче классификации, которая по структуре проще задачи детекции объектов. В этой статье мы представляем обзор известных методов атак и защиты для детекторов объектов и приводим их классификацию.**

***Ключевые слова*—детекция объектов, методы состязательной защиты, робастность детекторов, состязательные атаки, цифровые атаки, физические атаки**

I. Введение

Глубокие нейронные сети произвели революцию в области машинного обучения и были внедрены в широкий спектр областей: обработка естественного языка [1], обработка больших данных [2], анализ ДНК [3] и управление автономными транспортными средствами [4], [5], [6], [7]. Однако для внедрения в критически важные области модели глубокого обучения должны быть робастными.

Робастные модели обладают способностью хорошо работать даже при некотором уровне неопредёленности [8]. Однако ряд публикаций [9], [10] показали, что алгоритмы машинного обучения обладают низкой робастностью, и вызвали впечатляющую волну публикаций. Выяснилось, что глубокие нейронные сети очень чувствительны к небольшим умышленным возмущениям во входных данных. Они незаметны для человеческого глаза, но могут привести к сбоям в работе глубоких нейронных сетей. Атаки с применением таких небольших возмущений называют состязательными атаками.

Статья получена 20 мая 2023.

Екатерина Андреевна Чехонина, МГУ им. М.В. Ломоносова, (email: catherine.fish2.0@yandex.ru).

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

По мере того, как глубокие нейронные сети становятся всё более востребованными в коммерческом применении, всё более актуальными становятся и физические состязательные атаки — атаки, применение которых в реальном мире показывает высокую эффективность.

На данный момент бóльшая часть исследовательских работ в области робастности к состязательным атакам и методов защиты сосредоточена на задаче классификации. Она более проста по своей постановке, однако детекция объектов более востребована в приложениях реального мира.

Учитывая важность состязательной робастности глубоких нейронных сетей для задачи детекции, мы рассмотрим наиболее важные статьи на тему состязательных атак и защит и приведём их классификацию.

II. Терминология

Состязательные атаки являются небольшими возмущениями, которые применяются к изображениям. Обычно они незаметны для человеческого глаза, но могут вводить в заблуждение модели глубокого обучения и снижать их точность [11].

Модели детекции объектов, как правило, имеют одну и ту же структуру ввода-вывода. Все они берут входное изображение и выдают ограничивающие рамки, чтобы обеспечить локализацию каждого целевого объекта и его классификацию.

Для входного изображения x, детектор генерирует потенциальные ограничивающие рамки $B(x) = \{o_1, o_2, ..., o_S\}$, где $o_i = \{b_i^x, b_i^y, b_i^W, b_i^H, C_i, p_i\}$ — одна из рамок с центром в точке (b_i^x, b_i^y), размерностью (b_i^W, b_i^H), вероятностью $C_i \in [0; 1]$ наличия в данной рамке объекта, а также K-мерным вектором $p_i = (p_i^1, p_i^2, ..., p_i^K)$ вероятностей принадлежности к определенному классу. Выход детектора $O(x)$ получается путём отсеивания рамок с низкой вероятностью наличия в них объекта и исключением рамок с высокой степенью пересечения с другими.

Состязательный пример $x' = x + \delta$ генерируется через наложение возмущения на исходное изображение с целью обмануть детектор. Процесс генерации состязательного примера может быть сформулирован как:

$$\min_{\delta} \|x' - x\|_p : O(x') \neq O(x),$$

где p — метрика расстояния. Чаще всего в качестве нормы p рассматривают: L_0, которая равна количеству различающихся пикселей между двумя изображениями; L_2, которая равна евклидову расстоянию между пикселями двух изображений; L_∞, которая равна максимальной

разности между соотвествующими пикселями двух изображений.

Таким образом, перед детектором объектов стоит три подзадачи: обнаружить объект — понять, если ли он в конкретной части изображения; определить границы объекта — нарисовать ограничительные рамки; выставить метку, к какому классу относится обнаруженный объект.

III. Классификация состязательных атак

В этом разделе мы представим широко используемую в литературе [12] классификацию существующих состязательных атак в задачах компьютерного зрения.

A. Пространство атаки

Состязательные атаки могут быть реализованы в цифровом и физическом мире. В цифровом мире характеристики целевого объекта при обучении и применении атаки не изменяются, а злоумышленник стремится использовать нестандартные конфигурации, которые могут обмануть нейронную сеть.

При атаках в физическом мире после создания цифровой атаки, необходимо её реализовать, например распечатать патч. При реализации атаки её эффективность может уменьшиться по следующим причинам: различия цветопередачи из цифрового в реальный мир, несовершенства устройства воспроизведения, сложные условия окружающей среды — различное освещение, тень, перекрытие целевого объекта, изменение яркости, вращение, деформация и так далее.

B. Влияние на детектор объектов

Исходя из постановки задачи детекции, состязательные атаки могут влиять на все три подзадачи: на обнаружение объекта, на правильное определение границ объектов и на определение класса объекта.

C. Цель злоумышленника

В литературе состязательные атаки традиционно делят на целевые и нецелевые атаки. При нецелевых атаках злоумышленник хочет изменить метку объекта с корректной, на любую некорректную [13]. А при целевых атаках модель обманывается так, чтобы она назначала конкретную метку для конкретного объекта. Фактически, целевые атаки разрабатываются для определенного класса объектов [14].

D. Знания злоумышленника

Если злоумышленник обладает полными знаниями о модели, ее параметрах и может полностью воспроизвести атакуемую модель, говорят об атаке белого ящика. Если же злоумышленник ничего не знает о модели и может только подавать на вход изображения и получать результат работы модели, то такую атаку называют атакой чёрного ящика.

Будем описывать атаки согласно приведённой классификации. Итоговый результат обзора представлен в Таблице I. Мы рассмотрим сначала цифровые атаки, затем обратим внимание на атаки в реальном мире.

IV. Цифровые атаки на детекторы объектов

A. TOG: Targeted Adversarial Objectness Gradient Attacks

Представленная Chow и др. в [15] серия атак TOG состоит из 6 типов атак белого ящика, которые могут влиять на каждую из подзадач детектирования отдельно и на все сразу.

TOG-атака генерируется следующим образом:

$$x'_t = \prod_{x,\epsilon} x'_{t-1} - \alpha_{TOG}\,\mathrm{sign}\left(\frac{\partial \mathcal{L}^*(x'_{t-1}; \mathcal{O}^*, \mathcal{W})}{\partial x'_{t-1}}\right)$$

Здесь α_{TOG} обозначает размер шага, $L^*(\cdot)$ обозначает функцию потерь, которая может влиять на обнаружение объекта, формирование ограничительной рамки и выставление метки.

Также авторы метода предлагают универсальную TOG-атаку, которая обеспечивает переносимость сформированного возмущения, т.е. «обучившись» на известной нейронной сети атака может применяться и к другой сети, о которой ничего не известно. Примеры TOG-атак представлены на Рисунке 1.

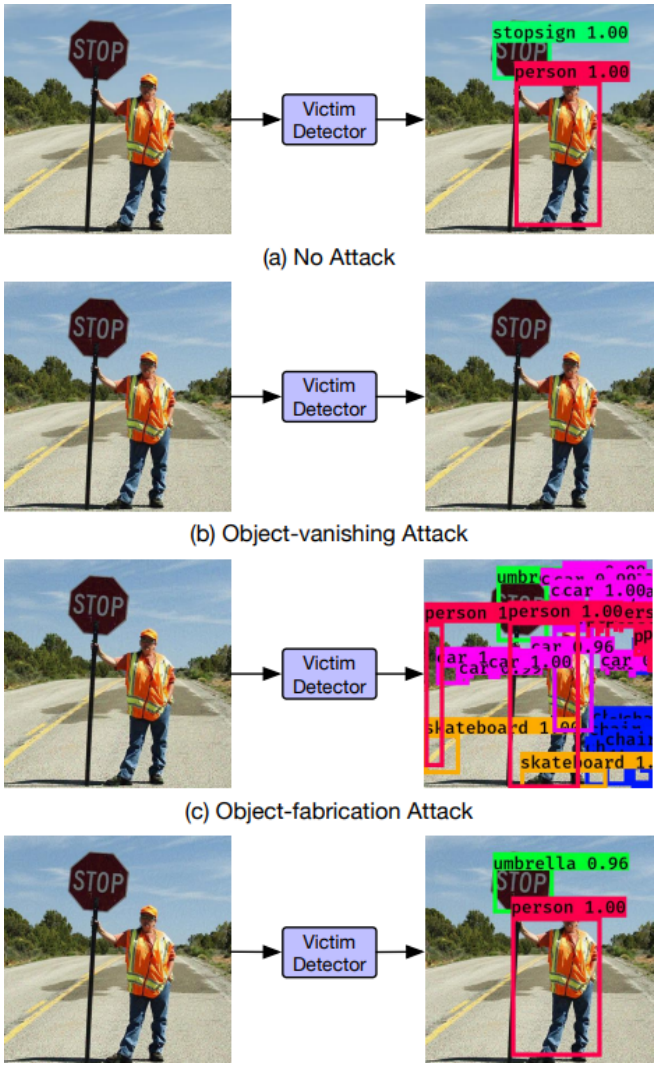

Рис. 1: Примеры трёх TOG-атак

Таблица I: Сравнение состязательных атак

Название	Тип атаки	Знания злоумышленника	Цель атаки	Влияние на детектор	Атакуемые модели
TOG	Цифровая	Атака белого ящика	Целевая и нецелевая	Все три подзадачи	YOLOv3, SSD, Faster R-CNN
DAG	Цифровая	Атака белого ящика	Целевая и нецелевая	Обнаружение объекта, выставление метки	FCN, Faster R-CNN
Evaporate attack	Цифровая	Атака чёрного ящика	Целевая	Обнаружение объекта	-
RAP	Цифровая	Атака белого ящика	Целевая	Обнаружение объекта, определение границ	Faster-RCNN, RFCN, FCIS, Mask-RCNN
Poster attack	Физическая, патч-атака	Атака белого/черного ящика	Нецелевая	Обнаружение объекта	Faster R-CNN, YOLOv2
Патч	Физическая, патч-атака	Атака белого ящика	Нецелевая	Обнаружение объекта	YOLOv2
Футболка	Физическая, патч-атака	Атака белого/черного ящика	Нецелевая	Обнаружение объекта	Faster R-CNN, YOLOv2
UPS	Физическая, камуфляжная атака	Атака белого ящика	Целевая и нецелевая	Обнаружение объекта, выставление метк	VGG-16, ResNet-101
SLAP	Физическая, оптическая атака	Атака белого ящика	Нецелевая	Обнаружение объекта, определение границ	Yolov3, Mask-RCNN, Lisa-CNN, Gtsrb-CNN

B. DAG: Dense Adversary Generation

В [16] авторы предлагают атаку белого ящика и также используют градиентный метод для формирования состязательных примеров. Последовательно искажаются метки объектов, которые нейронная сеть определяет правильно, в том числе после добавленных некоторых возмущений. Для этого минимизируется следующая лосс-функция:

$$L(X, \mathcal{T}, \mathcal{L}, \mathcal{L}') = \sum_{i=1}^{N} \left[f_{l_n}(X, t_n) - f_{l'_n}(X, t_n) \right],$$

где $\mathcal{T} = \{t_1, t_2, ..., t_N\}$ — целевые объекты, $\mathcal{L} = \{l_1, l_2, ..., l_N\}$ — истинные метки классов целевых объектов, $\mathcal{L}' = \{l'_1, l'_2, ..., l'_N\}$ — состязательные метки классов целевых объектов, которые можно определить как $l'_i \neq l_i$, $f_{l_n}(X, t_n)$ — степень уверенности модели в выставлении объекту t_n метки l_n.

Реализация DAG включает в себя только искажение метки для каждого объекта и выполнение итеративного обратного распространения градиента. Чаще и эффективнее эта атака применяется для семантической сегментации, но также применима к детектированию объектов.

C. Evaporate Attack

В работе [17] представлена атака чёрного ящика, основная задача которой — скрыть объект от детектора. Для этого сначала генерируется несколько начальных состязательных примеров и отправляются в нейронную сеть вместе с исходными изображениями. Затем генерация итогового состязательного примера x' происходит с помощью следующего оптимизационного уравнения:

$$\min_{x'} \mathcal{L}(x') = d(x', x) - \delta(D(x'))$$

Здесь вычисляется разница между состязательным критерием $\delta(D(x'))$ и расстоянием между состязательным примером и исходным изображением $d(x', x)$. Затем, решая задачу оптимизации, уменьшаем расстояние между состязательным и реальным примерами и увеличиваем состязательный критерий (он принимает значение 0, если требование атаки выполнено, и $-\infty$ в противном случае). Общая схема данной атаки представлена на Рисунке 2.

D. RAP: Robust Adversarial Perturbation

В этом методе [18] авторы фокусируется на атаке сети RPN (Region Proposal Network [19]), часто используемой для извлечения объектоподобных областей в детекторах объектов.

Для атаки оптимизируется функция потерь, которая пытается нарушить обнаружение объекта и определение границ объекта, так что даже если объект идентифицирован правильно, ограничивающая рамка не может быть точно установлена.

RAP может быть объединена с существующими методами состязательных атак, поскольку она специально фокусируется на атаке RPN, которая является промежуточной стадией сети.

V. Физические атаки на детекторы объектов

Физические состязательные атаки разрабатываются для изменения объекта или его области, чтобы ввести в заблуждение целевую модель, и называются атаками на уровне экземпляра. Именно их можно провести в реальном мире, так как изменить фоновую среду, например небо, невозможно.

В зависимости от способа, используемого для модификации целевого объекта, можно выделить атаки с постером, с патчом, камуфляжные атаки и оптические атаки.

Состязательные патчи это небольшие изображения, которые можно наклеить на поверхность целевого объекта. Эти атаки очень просты в реализации, но применимы в основном к плоским объектам.

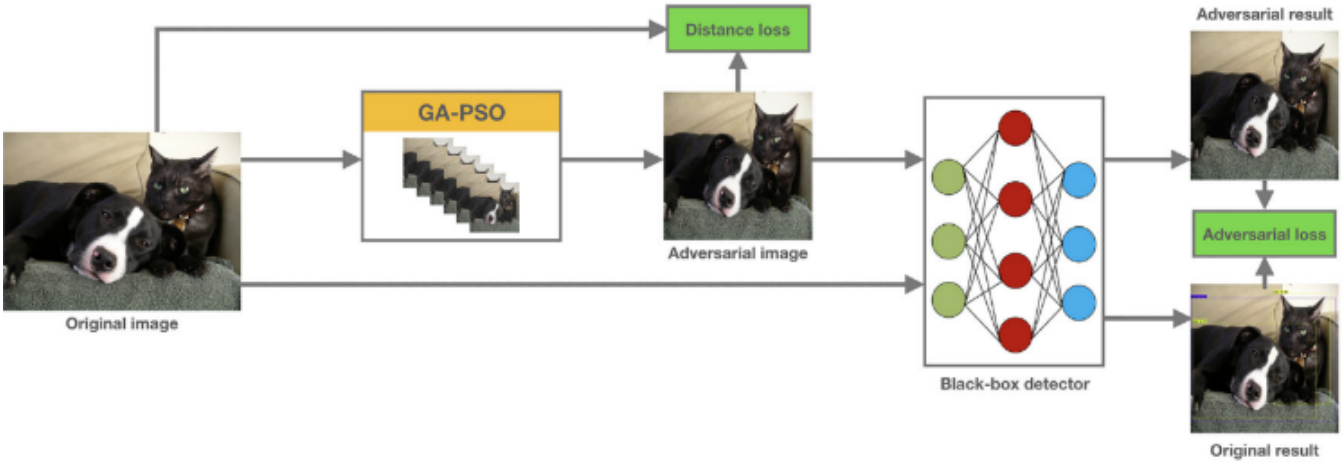

Рис. 2: Схема Evaporate Attack

Камуфляжные атаки предназначены для изменения внешнего вида 3D-объекта с помощью специальной текстуры, которая может быть нанесена на поверхность целевого объекта, например, в качестве рисунка на одежде или машине.

При **оптических атаках** злоумышленник испускает свет или лазер с помощью осветительного устройства, например проектора или лазерного излучателя, в направлении целевого объекта, чтобы изменить его внешний вид. Такие атаки поддаются контролю и хорошо скрываются. Однако они чувствительны к освещению окружающей среды, что ограничивает их использование на практике.

A. Poster attack

В работе [20] была предложена атака с наложением на дорожный знак состязательного постера, который можно рассматривать как патч полного размера. В результате атаки нейронные сети становятся не способны задетектировать дорожные знаки. В этой атаке авторы минимизируют максимальную уверенность детектора в классе объекта (который необходимо скрыть) среди всех рамок, генерируемых моделью после наложения постера.

Также авторы делают получившийся постер устойчивым относительно различным преобразованиям, которым он может подвергнуться в реальных условиях. Для этого используется фреймворк Expectation over Transformation [21]. Внешний вид получившихся в результате дорожных знаков представлен на Рисунке 3.

Рис. 3: Результат физической атаки с постером

B. Adversarial patch

Впервые атаки с применением состязательного патча были предложены в статье [22]. Авторы провели атаку на классификатор, накладывая на некоторый регион изображения специальный патч, который они подбирали исходя из требований о том, что он должен подходить к большинству изображений из некоторого датасета, а также приводить к обману модели даже при применении к патчу различных преобразований.

Такие свойства позволили сделали атаку с помощью патча применимой и для детекторов. В статье [23] была предложена атака на детектор YOLOv2 с использованием специального распечатанного патча, который позволил бы человеку скрыться от обнаружения. Минимизируемая функция потерь выглядит в этой атаке следующим образом:

$$L = \alpha L_{nps} + \beta L_{tv} + L_{det},$$

где

- L_{nps} - коэффициент непечатности (non-printability score), позволяющий подбирать те цвета, которые возможно распечатать на принтере;
- L_{tv} - общая дисперсия, минимизируемая с целью сделать изображение более естественным;
- L_{det} - функция, штрафующая за высокую уверенность детектора о наличии на изображении человека или просто за наличие какого-то объекта.

C. Wearable attack

Однако в статье [24] было показано, что если распечатать подобранный в предыдущей атаке патч на футболке, то из-за различных деформаций он перестанет обманывать детектор. Таким образом, данный патч плохо применим в случае реальной атаки. Необходимо учитывать движения, которые приводит к таким положениям футболки, описать которые обычными трансформациями вроде поворотов, изменений яркости и т.п. нельзя.

Для подбора такого патча, который продолжал бы обманывать модель, будучи распечатанным на футболке, авторы предложили помимо обычных преобразований патча рассмотреть множество преобразований *thin plate*

spline(TPS), которое широко используется в качестве нежесткой модели преобразований при выравнивании изображения и сопоставления форм. TPS выучивает зависящее от параметров отображение с оригинального изображения x на целевое изображение z с помощью набора контрольных точек с данными позициями.

В результате, авторы используют фреймворк Expectation over Transformation [21] для минимизации математического ожидания функции потерь атаки, взятому по обычным преобразованиям t, нежестким преобразованиям t_{TPS}, а также случайному шуму ν:

$$\min_{\delta} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{t,t_{TPS},v}[f(x_i')] + \beta L_{tv},$$

где x_i' - кадр видео после наложения на него преобразований из упомянутых множеств, а функция потерь $f(\cdot)$ для YOLOv2 — максимальная вероятность класса «человек» среди всех предсказанных рамок.

D. UPC: Universal Physical Camouflage Attack

Huang и др. в [25] представляют Universal Physical Camouflage Attack — атаку, которая создает универсальный камуфляж, позволяющий скрыть объектов от обнаружения или обмануть детектор при определении класса объекта. Она атакует все экземпляры, принадлежащие к одной и той же категории объектов, например человек или автомобили, поэтому называется универсальной.

Чтобы эффективно обрабатывать деформации сложных объектов в физическом мире, вводится набор преобразований, например обрезка, изменение размера, аффинная гомография. Также обеспечивается визуальное сходство между сгенерированным камуфляжем и естественными изображениями. Камуфляжные узоры визуально похожи на естественные изображения и могут рассматриваться как текстурные узоры на поверхностях объектов, такие как аксессуары для людей и рисунки для автомобилей. Примеры полученных камуфляжей для автомобилей можно видеть на Рисунке 3.

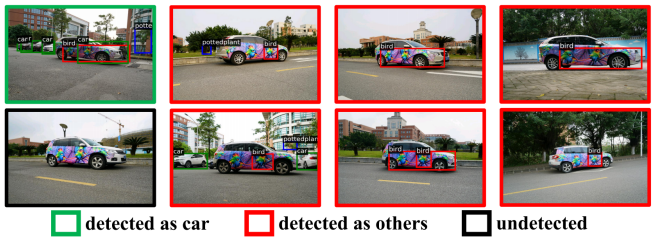

Рис. 4: Применение атаки UPC в реальном мире

E. SLAP: Short-Lived Adversarial Perturbations

В [26] авторы предлагают оптическую атаку SLAP, которая может быть реализована с помощью проектора. Злоумышленник может проецировать специально созданные состязательные возмущения на объекты реального мира, что даёт больший контроль над атакой, поскольку проекции можно включать и выключать по мере необходимости и не оставлять очевидных следов атаки.

SLAP моделирует трехстороннюю аддитивную взаимосвязь между поверхностью, проекцией и изображением, воспринимаемым камерой. Надежность атаки повышается путем систематического выявления и учета изменяющихся условий окружающей среды в процессе обучения. Пример проекции, которую создаёт SLAP можно видеть на Рисунке 4.

Рис. 5: Иллюстрация действия атаки SLAP

VI. Классификация методов защиты

В этом разделе мы опишем классификацию существующих методов защиты от состязательных атак.

A. Гарантии надёжности

Существует две категории методов защиты по уровню гарантии надежности.

Эмпирические методы не имеют теоретически подтверждённой гарантии надёжности: в будущем они могут быть взломаны более сложными состязательными атаками.

Если существует процедура сертификации надёжности метода защиты для определённых входных данных в отношении определённой модели угроз, то это метод сертифицированной защиты. При сертификации важно доказать, что результаты будут действительны для любой атаки в рамках модели угроз, включая те, которые обладают полным знанием алгоритма защиты, настроек и т.д.

B. Цель защиты

Существует также две различные цели защиты.

Улучшение робастности предсказания нацелено на то, чтобы принимать правильные решения, например правильно определять ограничивающую рамку или класс объекта, даже при наличии состязательной атаки. Это также часто называют защитой, основанной на восстановлении.

Обнаружение атак, с другой стороны, направлено только на обнаружение. Если защитная система обнаруживает атаку, она отсеивает состязательный пример, и модель не выполняется предсказание.

Мы рассмотрим сначала методы сертифицированной защиты, затем перейдём к эмпирическим методам. Итоговый результат обзора представлен в Таблице II.

VII. Сертифицированные методы защиты

A. Медианное сглаживание

В работе [27] был предложен метод случайного сглаживания, который позволяет провести вероятностную сертификацию классификатора относительно искажений входа по норме l_2 с помощью наложения на исходное

Таблица II: Сравнение методов защиты

Название метода	Гарантии надёжности	Цель защиты
Медианное сглаживание	Сертифицированный	Сертификация по норме l_2
ObjectSeeker	Сертифицированный	Сертификация от атак патчами
Состязательное обучение	Эмпирический	Повышение робастности
Использование пространственного контекста	Эмпирический	Повышение робастности
APM	Эмпирический	Обнаружение атаки и повышение робастности
Gabor convolutional layers	Эмпирический	Повышение робастности

изображение гауссовского шума и оценки математического ожидания классификации. В статье [28] была предложена адаптация этого метода для задачи детекции, которая была рассмотрена как задача регрессии. Было показано, что оценка математического ожидания дает плохие результаты, поэтому авторы перешли к оценке медианы, которая меньше подвержена влиянию выбросов распределения.

Таким образом, предложенный авторами пайплайн сертификации включает в себя добавление к исходному изображению различных сэмплов шума, а затем сопоставление предсказаний модели на получившихся зашумленных изображений. При значительном совпадении предсказаний на разных зашмуленных изображений можно с очень высокой вероятностью сертифицировать предсказания модели на исходном изображении.

B. ObjectSeeker

В работе [29] была начата, а в работе [30] продолжена разработка метода для сертифицированной защиты детекторов от атак патчами. Предложенный метод ObjectSeeker основан на маскировании входного изображения горизонтальными и вертикальными полосами. При каждой маскировке проводится детекция на незакрытой части изображения.

После прохождения такими полосами по всему изображению, проводится детектирование объектов на всем изначальном изображении.

Затем отфильтровываются те предсказанные рамки с первого этапа, которые значительно пересекаются с рамками на втором этапе. Если в результате этого процесса отфильтровались все рамки, то изображение считается чистым, не содержащим патчей. Если же после первого этапа остались рамки, аналоги которых не были предсказаны на втором этапе, считается, что на изображении присутствует патч. В работе была доказана теорема, согласно которой описанная процедура дает формальные гарантии безопасности.

Большим преимуществом данного метода является его эффективность против патчей любых размеров и форм.

VIII. Эмпирические методы защиты

A. Состязательное обучение

Состязательное обучение — это общий подход к повышению робастности модели, заключающийся в попеременной генерации состязательных примеров и обучении модели на этих же состязательных примерах.

В работе [31] авторы предлагают декомпозировать задачу детекции объектов на задачу классификации и задачу локализации и обратить внимание на взаимное влияние лосс-функций каждой из подзадач.

Они предлагают использовать следующую лосс-функцию для состязательного обучения:

$$\min_{\theta}\left[\max_{\overline{x}\in S_{cls}\cup S_{loc}}\mathcal{L}(f_\theta(\overline{x}),y_k,\mathbf{b_k})\right]$$

Где $S_{cls}\cup S_{loc}$ — ограничение области значений состязательного примера, которое определяется как набор изображений, максимизирующих потери при выполнении задачи классификации или потери при локализации.

B. Использование пространственного контекста

Цифровые состязательные атаки патчами могут быть реализованы даже без перекрытия целевого объекта, так как детекторы объектов активно используют пространство вокруг объекта. Поэтому в работе [32] авторы предлагают ограничить использование пространственного контекста моделью во время обучения.

Для достижения этой цели могут быть использованы инструменты интерпретации нейронных сетей, такие как Grad-CAM [33], которые выделяют области изображения, влияющие на конкретное решение нейронной сети. Grad-CAM работает путем визуализации производных выхода модели относительно промежуточного свёрточного слоя. Таким образом, необходимо ограничить производные для конкретного выхода так, чтобы они не выходили за пределы ограничивающей рамки целевого объекта.

Другим способом достижения цели является замена пространственного контекста у изображений. Чтобы создать такой набор данных, необходимо взять два случайных изображения из исходного датасета и целевой объект одного из них вставить в то же место на втором изображении.

C. APM: Adversarial Pixel Masking

В статье [34] авторы предлагают защиту от состязательных патчей, называемую Adversarial Pixel Masking, которую можно применять к предобученным детекторам объектов. APM дополняет состязательное обучение, добавляя сеть предварительной обработки данных MaskNet.

Во время состязательного обучения фиксируются веса детектора, и MaskNet обучается распознавать состязательные патчи на изображении. Затем они удаляются с изображения. Пример работы APM можно видеть на Рисунке 5.

D. Gabor convolutional layers

Для повышения робастности детектора объектов Amirkhani и Karimi [35] предлагают предобрабатывать

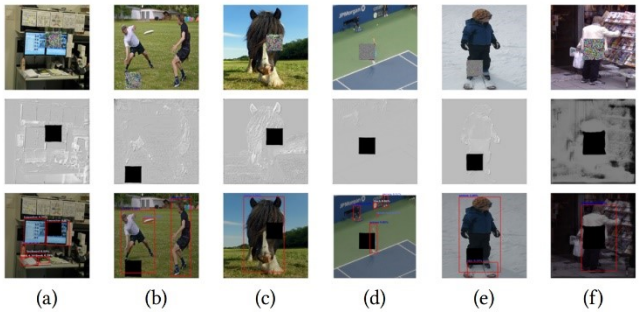

Рис. 6: Пример работы APM

изображения с помощью фильтров Габора. Для этого нужно разделить изображение на составляющие его каналы RGB, а затем подать эти каналы в банк фильтров Габора в виде тензора.

Каждый фильтр в банке сконструирован с определенным углом, и он может извлекать края и другие низкоуровневые объекты изображений, соответствующие этому углу.

Благодаря высокой способности извлекать низкоуровневые элементы изображения фильтры Габора могут повысить надежность нейронной сети.

IX. Заключение

В заключение хочется отметить, что состязательные атаки являются интересным феноменом нейронных сетей, однако их существование ставит под сомнение безопасность и надёжность применения глубоких нейронных сетей в реальных приложениях. Детекторы объектов имеют огромный потенциал внедрения в нашу жизнь, поэтому необходимо активно исследовать методы повышения их робастности к состязательным атакам.

Список литературы

[1] Sharma Akanksha Rai, Kaushik Pranav. Literature survey of statistical, deep and reinforcement learning in natural language processing // International Conference on Computing, Communication and Automation. — 2017. — P. 350–354.

[2] Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks / Hexuan Hu, Bo Tang, Xuejiao Gong et al. // IEEE Transactions on Industrial Informatics. — 2017. — Vol. 13, no. 4. — P. 2106–2116.

[3] Deng Lei, Wu Hui, Liu Hui. D2vcb: a hybrid deep neural network for the prediction of in-vivo protein-dna binding from combined dna sequence // IEEE International Conference on Bioinformatics and Biomedicine. — 2019. — P. 74–77.

[4] Ackerman Evan. How drive.ai is mastering autonomous driving with deep learning // IEEE Spectrum Magazine. — 2017. — URL: https://spectrum.ieee.org/how-driveai-is-mastering-autonomous-driving-with-deep-learning.

[5] Novel arithmetics in deep neural networks signal processing for autonomous driving: challenges and opportunities / Marco Cococcioni, Federico Rossi, Emanuele Ruffaldi et al. // IEEE Signal Processing Magazine. — 2020. — Vol. 38, no. 1. — P. 97–110.

[6] Cococcioni Marco, Ruffaldi Emanuele, Saponara Sergio. Exploiting posit arithmetic for deep neural networks in autonomous driving applications // International Conference of Electrical and Electronic Technologies for Automotive. — 2018. — P. 1–6.

[7] Okuyama Takafumi, Gonsalves Tad, Upadhay Jaychand. Autonomous driving system based on deep q learning // International Conference on Intelligent Autonomous Systems. — 2018. — P. 201–205.

[8] Ben-Tal Aharon, El Ghaoui Laurent, Nemirovski Arkadi. Robust optimization. — Princeton University Press, 2009.

[9] 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.

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

[11] Lu Jiajun, Issaranon Theerasit, Forsyth David. Safety net: detecting and rejecting adversarial examples robustly // IEEE International Conference on Computer Vision. — 2017.

[12] A survey on physical adversarial attack in computer vision / Donghua Wang, Wen Yao, Tingsong Jiang et al. // arXiv preprint arXiv:2209.14262. — 2022.

[13] 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.

[14] Adversarial examples: attacks and defenses for deep learning / Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li // IEEE Transactions on Neural Networks and Learning Systems. — 2019. — Vol. 30, no. 9. — P. 2805–2824.

[15] Adversarial objectness gradient attacks in real-time object detection systems / Ka-Ho Chow, Ling Liu, Margaret Loper et al. // Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications. — 2020.

[16] Adversarial examples for semantic segmentation and object detection / Cihang Xie, Jianyu Wang, Zhishuai Zhang et al. // IEEE International Conference on Computer Vision. — 2017. — P. 1369–1378.

[17] An adversarial attack on dnn-based black-box object detectors / Yajie Wang, Yu-an Tan, Wenjiao Zhang et al. // Journal of Network and Computer Applications. — 2020. — Vol. 161.

[18] Robust adversarial perturbation on deep proposal-based models / Yuezun Li, Daniel Tian, Ming-Ching Chang et al. // arXiv preprint arXiv:1809.05962 s. — 2018.

[19] Faster r-cnn: Towards real-time object detection with region proposal networks / Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun // Advances in neural information processing systems. — 2015. — Vol. 28.

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

[21] Synthesizing robust adversarial examples / Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok // International conference on machine learning / PMLR. — 2018. — P. 284–293.

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

[23] Thys Simen, Van Ranst Wiebe, Goedemé Toon. Fooling automated surveillance cameras: adversarial patches to attack person detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. — 2019. — P. 0–0.

[24] Adversarial t-shirt! evading person detectors in a physical world / Kaidi Xu, Gaoyuan Zhang, Sijia Liu et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 / Springer. — 2020. — P. 665–681.

[25] Universal physical camouflage attacks on object detectors / Lifeng Huang, Chengying Gao, Yuyin Zhou et al. // IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — P. 720–729.

[26] Slap: Improving physical adversarial examples with shortlived adversarial perturbations / Giulio Lovisotto, Henry Turner, Ivo Sluganovic et al. // 30th USENIX Security Symposium. — 2021. — P. 1865–1882.

[27] Cohen Jeremy, Rosenfeld Elan, Kolter Zico. Certified adversarial robustness via randomized smoothing // international conference on machine learning / PMLR. — 2019. — P. 1310–1320.

[28] Detection as regression: Certified object detection with median smoothing / Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader et al. // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 1275–1286.

[29] Xiang Chong, Mittal Prateek. Detectorguard: Provably securing object detectors against localized patch hiding attacks // Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. — 2021. — P. 3177–3196.

[30] Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking / Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, Prateek Mittal // arXiv preprint arXiv:2202.01811. — 2022.

[31] Zhang Haichao, Wang Jianyu. Towards adversarially robust object detection // IEEE/CVF International Conference on Computer Vision. — 2019. — P. 421–430.

[32] Role of spatial context in adversarial robustness for object detection / Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, Pirsiavash Hamed // IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. — 2020.

[33] Grad-cam: Visual explanations from deep networks via gradient-based localization / Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das et al. // IEEE International Conference on Computer Vision. — 2017.

[34] Chiang Ping-Han, Chan Chi-Shen, Wu Shan-Hung. Adversarial pixel masking: A defense against physical attacks for pre-trained object detectors // 29th ACM International Conference on Multimedia. — 2021.
[35] Amirkhani Abdollah, Karimi Mohammad Parsa. Adversarial defenses for object detectors based on gabor convolutional layers // The Visual Computer. — 2022. — Vol. 38, no. 6. — P. 1929–1944.

Overview of adversarial attacks and defenses for object detectors

Ekaterina Chekhonina, Vasily Kostyumov

***Abstract*—Nowadays object detection is considered as one of the most popular fields of deep neural networks with numerous applications in critical areas: natural language processing, big data processing, DNA analysis, autonomous vehicles. However, detection object systems are sensitive to small perturbations in input data. They are imperceptible to the human eye, but they can completely mislead the DNNs. Object detectors are vulnerable against adversarial attacks and hardly could be embedded in real-life applications. Existing adversarial attacks can be divided into digital and physical adversarial attacks. Attacks in the digital world have strong attack performance in lab environments but are not so effective in the real world, unlike physical attacks. Defenses can be divided into empirical and certified. Certified methods guarantee reliability. Empirical defenses can be vulnerable against complex adversarial attacks. While the field of adversarial robustness is very popular, the majority of the work has been focused on the task of image classification due to it being simpler in structure than object detection. We review the prominent attack and defense mechanism related to object detection and propose its classification.**

***Keywords*—object detection, adversarial defences, detectors robustness, adversarial attacks, digital attacks, physical attacks**

References

[1] Sharma Akanksha Rai, Kaushik Pranav. Literature survey of statistical, deep and reinforcement learning in natural language processing // International Conference on Computing, Communication and Automation. — 2017. — P. 350–354.

[2] Intelligent fault diagnosis of the high-speed train with big data based on deep neural networks / Hexuan Hu, Bo Tang, Xuejiao Gong et al. // IEEE Transactions on Industrial Informatics. — 2017. — Vol. 13, no. 4. — P. 2106–2116.

[3] Deng Lei, Wu Hui, Liu Hui. D2vcb: a hybrid deep neural network for the prediction of in-vivo protein-dna binding from combined dna sequence // IEEE International Conference on Bioinformatics and Biomedicine. — 2019. — P. 74–77.

[4] Ackerman Evan. How drive.ai is mastering autonomous driving with deep learning // IEEE Spectrum Magazine. — 2017. — URL: https://spectrum.ieee.org/how-driveai-is-mastering-autonomous-driving-with-deep-learning.

[5] Novel arithmetics in deep neural networks signal processing for autonomous driving: challenges and opportunities / Marco Cococcioni, Federico Rossi, Emanuele Ruffaldi et al. // IEEE Signal Processing Magazine. — 2020. — Vol. 38, no. 1. — P. 97–110.

[6] Cococcioni Marco, Ruffaldi Emanuele, Saponara Sergio. Exploiting posit arithmetic for deep neural networks in autonomous driving applications // International Conference of Electrical and Electronic Technologies for Automotive. — 2018. — P. 1–6.

[7] Okuyama Takafumi, Gonsalves Tad, Upadhay Jaychand. Autonomous driving system based on deep q learning // International Conference on Intelligent Autonomous Systems. — 2018. — P. 201–205.

[8] Ben-Tal Aharon, El Ghaoui Laurent, Nemirovski Arkadi. Robust optimization. — Princeton University Press, 2009.

[9] 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.

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

[11] Lu Jiajun, Issaranon Theerasit, Forsyth David. Safety net: detecting and rejecting adversarial examples robustly // IEEE International Conference on Computer Vision. — 2017.

[12] A survey on physical adversarial attack in computer vision / Donghua Wang, Wen Yao, Tingsong Jiang et al. // arXiv preprint arXiv:2209.14262. — 2022.

[13] 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.

[14] Adversarial examples: attacks and defenses for deep learning / Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li // IEEE Transactions on Neural Networks and Learning Systems. — 2019. — Vol. 30, no. 9. — P. 2805–2824.

[15] Adversarial objectness gradient attacks in real-time object detection systems / Ka-Ho Chow, Ling Liu, Margaret Loper et al. // Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications. — 2020.

[16] Adversarial examples for semantic segmentation and object detection / Cihang Xie, Jianyu Wang, Zhishuai Zhang et al. // IEEE International Conference on Computer Vision. — 2017. — P. 1369–1378.

[17] An adversarial attack on dnn-based black-box object detectors / Yajie Wang, Yu-an Tan, Wenjiao Zhang et al. // Journal of Network and Computer Applications. — 2020. — Vol. 161.

[18] Robust adversarial perturbation on deep proposal-based models / Yuezun Li, Daniel Tian, Ming-Ching Chang et al. // arXiv preprint arXiv:1809.05962 s. — 2018.

[19] Faster r-cnn: Towards real-time object detection with region proposal networks / Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun // Advances in neural information processing systems. — 2015. — Vol. 28.

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

[21] Synthesizing robust adversarial examples / Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok // International conference on machine learning / PMLR. — 2018. — P. 284–293.

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

[23] Thys Simen, Van Ranst Wiebe, Goedemé Toon. Fooling automated surveillance cameras: adversarial patches to attack person detection // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. — 2019. — P. 0–0.

[24] Adversarial t-shirt! evading person detectors in a physical world / Kaidi Xu, Gaoyuan Zhang, Sijia Liu et al. // Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 / Springer. — 2020. — P. 665–681.

[25] Universal physical camouflage attacks on object detectors / Lifeng Huang, Chengying Gao, Yuyin Zhou et al. // IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — P. 720–729.

[26] Slap: Improving physical adversarial examples with shortlived adversarial perturbations / Giulio Lovisotto, Henry Turner, Ivo Sluganovic et al. // 30th USENIX Security Symposium. — 2021. — P. 1865–1882.

[27] Cohen Jeremy, Rosenfeld Elan, Kolter Zico. Certified adversarial robustness via randomized smoothing // international conference on machine learning / PMLR. — 2019. — P. 1310–1320.

[28] Detection as regression: Certified object detection with median smoothing / Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader et al. // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 1275–1286.

[29] Xiang Chong, Mittal Prateek. Detectorguard: Provably securing object detectors against localized patch hiding attacks // Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. — 2021. — P. 3177–3196.

[30] Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking / Chong Xiang, Alexan-

der Valtchanov, Saeed Mahloujifar, Prateek Mittal // arXiv preprint arXiv:2202.01811. — 2022.
[31] Zhang Haichao, Wang Jianyu. Towards adversarially robust object detection // IEEE/CVF International Conference on Computer Vision. — 2019. — P. 421–430.
[32] Role of spatial context in adversarial robustness for object detection / Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, Pirsiavash Hamed // IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. — 2020.
[33] Grad-cam: Visual explanations from deep networks via gradient-based localization / Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das et al. // IEEE International Conference on Computer Vision. — 2017.
[34] Chiang Ping-Han, Chan Chi-Shen, Wu Shan-Hung. Adversarial pixel masking: A defense against physical attacks for pre-trained object detectors // 29th ACM International Conference on Multimedia. — 2021.
[35] Amirkhani Abdollah, Karimi Mohammad Parsa. Adversarial defenses for object detectors based on gabor convolutional layers // The Visual Computer. — 2022. — Vol. 38, no. 6. — P. 1929–1944.