Камуфляж как состязательные атаки на модели машинного обучения

Д.Е. Пришлецов, С.Е. Пришлецов, Д.Е. Намиот

***Аннотация*—Статья посвящена состязательным атакам на модели машинного обучения. Под такими атаками понимается сознательное манипулирование данными на разных этапах конвейера машинного обучения, призванное либо воспрепятствовать правильной работе модели, либо добиться от нее желаемого результата. В данном случае рассматривались физические атаки уклонения, то есть модифицировались сами объекты, используемые в модели. В статье рассматривается использование камуфлированных изображений для обмана системы распознавания. Эксперименты проводились с моделью машинного обучения, которая распознает изображения автомобилей. Были использованы два типа камуфляжных рисунков – классический камуфляж и нанесение изображения другого автомобиля. Практически, такие манипуляции могут быть осуществлены с помощью аэрографии. В работе подтверждено успешное осуществление таких атак, получены метрики, характеризующие эффективность атак, а также возможности противодействия им с помощью состязательных тренировок. Все результаты открыто опубликованы, что дает возможность использовать их как программный стенд для отработки других атак и методов защиты.**

***Ключевые слова*—состязательные атаки, камуфляж, распознавание изображений, машинное обучение.**

I. Введение

Настоящая статья посвящена состязательным атакам на системы распознавания изображений. Атаки на системы машинного обучения – это сознательные модификации данных на любом из этапов конвейера машинного обучения, призванные либо воспрепятствовать работе модели, либо наоборот, добиться необходимого атакующему результата работы модели [1]. Есть еще атаки, связанные с модификацией готовых (предобученных) моделей машинного обучения [2], но в данной работе мы рассматриваем только модификации данных.

Атаки (модификации данных) на этапе исполнения модели называются атаками уклонения [3]. Атаки могут быть цифровые и физические [2]. Физические атаки для задач распознавания изображений – это какие-либо модификации самих объектов, изображения которых должны будут распознаваться. Интерес именно к физическим атакам связан с тем, что их, в принципе, нельзя запретить. Нельзя запретить внешние модификации объектов, изображения которых должны распознаваться. Соответственно, во всех системах необходимо считаться с их возможным наличием.

Основной момент во всех физических атаках – это способ модификации объектов или способ доведения модифицированных данных до модели. Ниже приведены некоторые примеры.

В работе [4] изменялся фон для знаков дорожного движения (рис. 1.1)

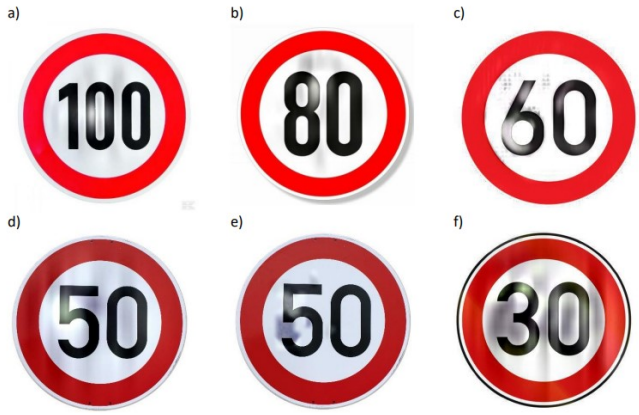

Рис 1.1. Модифицированные знаки дорожного движения [4].

В работе [5] проецировали с помощью дрона фальшивые знаки дорожного движения и фигуры людей на дороге (рис. 1.2)

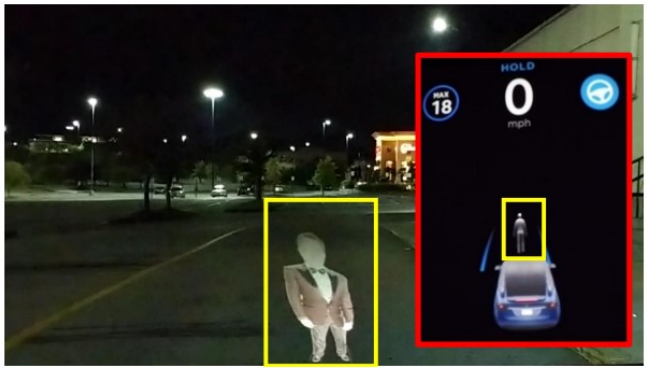

Рис. 1.2. Автопилот Тесла не начинает движение, “видя” человека на дороге [5].

Вышивка на свитере содержит карты значимости из модели детектирования лиц, и программа “видит” множество лиц на изображении [6] (рис. 1.3).

Статья получена 15 августа 2023
Д.Е. Пришлецов – МГУ имени М.В. Ломоносова (email: i@pukuluka.ru)
С.Е. Пришлецов – МГУ имени М.В. Ломоносова (email: prisha@inbox.ru)
Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com).

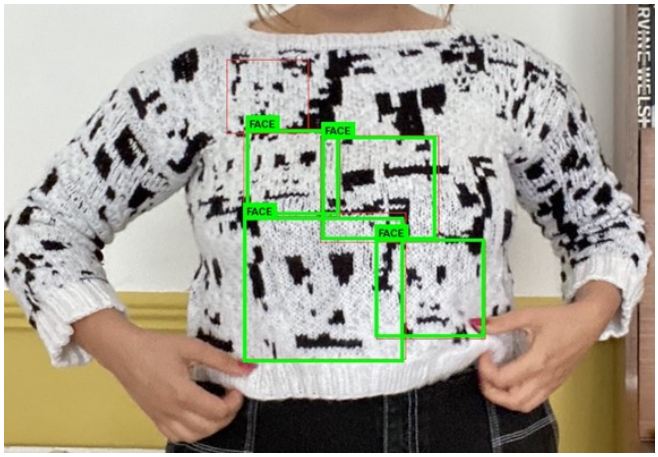

Рис.1.3 Состязательные рисунки на одежде [6]

Камуфляж можно признать исторически первой состязательной атакой. Камуфляжная (защитная) раскраска была призвана, как раз, обмануть наблюдателя. В принципе, ничего не меняется, если вместо человека-наблюдателя будет камера и модель машинного обучения, которая распознает изображения. Можно привести следующие примеры работ. В работе [7] рассматриваются физические состязательные атаки против моделей определения объектов (целей) на космических снимках. По сути - имитировалась известная атака патчами, которые накладывались на крышу объекта (рис. 1.4)

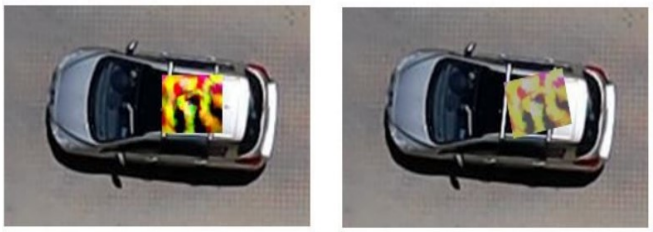

Рис. 1.4. Атака патчами [7].

Похожая схема была применена и в работе [8] (рис. 1.5)

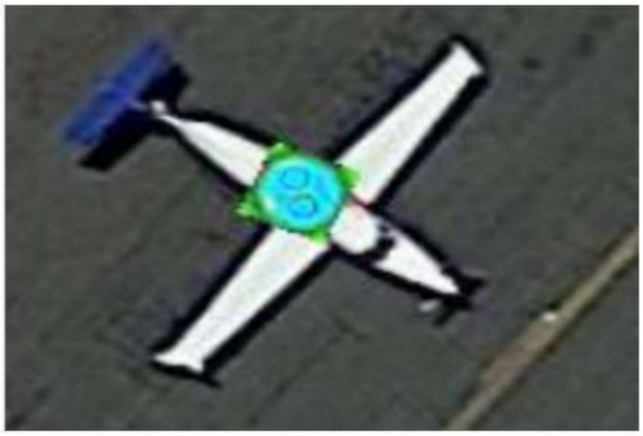

Рис. 1.5. Патч против распознавания на космических снимках [8].

В данной работе как раз и рассматривается использование камуфляжа для состязательных атак на модели машинного обучения. В основу статьи положены две выпускные квалификационные работы, выполненные на факультете ВМК МГУ имени М.В. Ломоносова (программа дополнительного образования) [12] и [13]. Идея работ состояла в следующем. Взять готовую модель распознавания, научить ее распознавать изображения автомобилей и получить метрики работы такой системы. Далее – протестировать работу модели на изображениях автомобилей с камуфляжем и посмотреть, как изменятся (ухудшатся) метрики. После чего провести так называемые состязательные тренировки, когда к тренировочному набору добавляются модифицированные (камуфлированные) изображения, но с правильной меткой, и модель переобучается. Далее – проверить распознавание камуфлированных изображений уже на переобученной модели.

В одном эксперименте использовался классический камуфляжный рисунок, в другом – идея, позаимствованная из методов маскировки в военно-морском флоте еще в Первую мировую войну. Тогда на борт корабля наносился силуэт другого (меньшего) корабля, что обманывало наблюдателя на большом расстоянии. В данном случае на автомобиль наносился силуэт другого автомобиля. Физическая (техническая) реализация обоих подходов обеспечивается услугой (сервисом), который называется аэрография.

Оставшаяся часть статьи структурирована следующим образом. В разделе II описана работа с камуфляжной окраской. В разделе III речь идет о нанесении дополнительных изображений. Раздел IV содержит заключение.

II. Камуфляжная раскраска

1) Подготовка и предварительная проверка

Выбран набор данных Stanford Car Dataset [9], содержащий более 190 моделей автомобилей.

В качестве алгоритма классификации подобран «Pytorch car classifier» в котором использовалась предварительно обученная нейронная сеть «Resnet34», а точность определения классов в 90% достигается за 10 эпох [10].

Для проверки работоспособности, был добавлен новый класс (модель автомобиля) - Toyota Camry 2006. Для решения задачи классификации с использованием нового класса систему необходимо переобучить.

После переобучения система классификации успешно справилась с поставленной задачей, отнеся автомобиль к верному классу (модели) с высокой долей уверенности (рис. 2.1).

Рис. 2.1. Результат работы классификатора на реальном примере

1) Подготовка и предварительная проверка

Из набора данных Stanford Car Dataset [9] выбран автомобиль Toyota Sequoia SUV 2012 для дальнейшего использования в качестве основной модели для нанесения камуфляжа. Изображения автомобиля были подобраны таким образом, чтобы они не входили в тренировочный набор данных классификатора.

Всего использовалось четыре изображения автомобиля. Три изображения - для тренировочного набора переобучения (рис. 2.2, рис. 2.3 и рис. 2.4):

Рис. 2.2. Toyota Sequoia SUV 2012_001.jpg

Рис. 2.3. Toyota Sequoia SUV 2012_002.jpg

Рис. 2.4. Toyota Sequoia SUV 2012_003.jpg

И одно - в качестве тестового (рис. 2.5):

Рис. 2.5. Toyota Sequoia SUV 2012_004.jpg

Для нанесения камуфляжа на изображение автомобиля было выбрано изображение, сгенерированное программным комплексом camogen [11] (рис. 3.5):

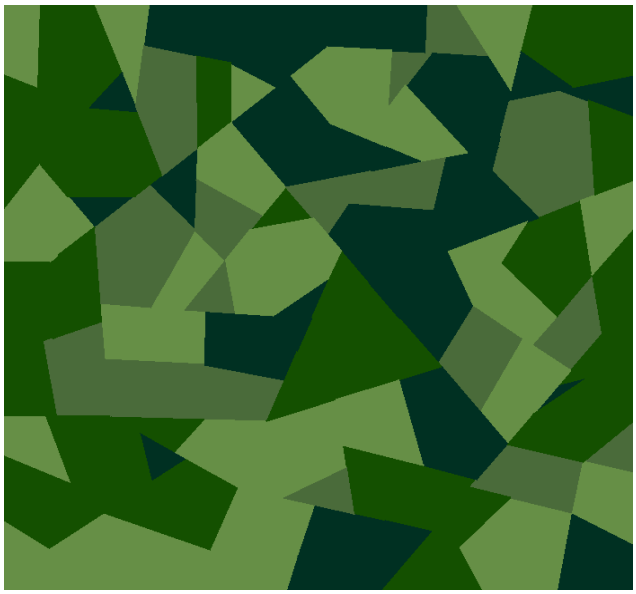

Рис. 2.6. Изображение камуфляжа

На изображения автомобиля были наложены полупрозрачные изображения камуфляжа для создания «камуфляжного» набора (рис. 2.7.1, рис. 2.7.2, рис. 2.7.3 и рис. 2.7.4):

Рис. 2.7.1. Toyota Sequoia SUV 2012_001_changed.jpg

Рис. 2.7.2. Toyota Sequoia SUV 2012_002_changed.jpg

Рис. 2.7.3. Toyota Sequoia SUV 2012_003_changed.jpg

Рис. 2.7.4. Toyota Sequoia SUV 2012_004_changed.jpg

Результаты классификации камуфляжного набора и составляющих его прототипов без добавления изображений в обучающий набор и переобучения модели:

Таблица 2.1

Изображение	**Оценка, %**
Toyota Sequoia SUV 2012_001 (Рис. 2.2)	78.54
Toyota Sequoia SUV 2012_001_changed (Рис. 2.7.1)	-
Toyota Sequoia SUV 2012_002 (Рис. 2.3)	78.34
Toyota Sequoia SUV 2012_002_changed (Рис. 2.7.2)	-
Toyota Sequoia SUV 2012_003 (Рис. 2.4)	89.12
Toyota Sequoia SUV 2012_003_changed (Рис. 2.7.3)	-
Toyota Sequoia SUV 2012_004 (Рис. 2.5)	*64.78*
Toyota Sequoia SUV 2012_004_changed (Рис. 2.7.4)	-

Из таблицы результатов видно, что классификатор не «угадал» ни одного автомобиля с нанесённым рисунком.

Некамуфлированные изображения система классификации узнала и отнесла их к соответствующим классам.

Атака успешна!

2) Включение прототипов в тренировочный набор

Добавим в тренировочный набор не камуфлированные изображения автомобиля Toyota Sequoia SUV 2012 (рис. 3.2, рис. 3.3 и рис. 3.4) и переобучаем классификатор:

Таблица 2.2

Изображение	**Оценка, %**
Toyota Sequoia SUV 2012_001 (Рис. 2.2)	99.98
Toyota Sequoia SUV 2012_001_changed (Рис. 2.7.1)	-
Toyota Sequoia SUV 2012_002 (Рис. 2.3)	99.89
Toyota Sequoia SUV 2012_002_changed (Рис. 2.7.2)	-

Изображение	Оценка, %
Toyota Sequoia SUV 2012_003 (Рис. 2.4)	99.96
Toyota Sequoia SUV 2012_003_changed (Рис. 2.7.3)	86.90
Toyota Sequoia SUV 2012_004 (Рис. 2.5)	*99.66*
Toyota Sequoia SUV 2012_004_changed (Рис. 2.7.4)	-

Как видно из таблицы, классификатор «узнал» все изображения автомобиля (рис. 2.2, рис. 2.3, рис. 2.4 и рис. 2.5) – что неудивительно, так как рис. 2.2, рис. 2.3 и рис. 2.4 были добавлены в тренировочный набор перед переобучением, а также изображение автомобиля с нанесённым рисунком Toyota Sequoia SUV 2012_003_changed (рис. 2.7.3).

Значительно увеличилась точность классификации изображения, не вошедшего в тренировочный набор - Toyota Sequoia SUV 2012_004 (рис. 2.5).

3) Состязательное обучение

Добавляем камуфлированные изображения автомобиля (кроме тестовых) в тренировочный набор и переобучаем классификатор (состязательная тренировка):

Таблица 2.3

Изображение	Оценка, %
Toyota Sequoia SUV 2012_001 (Рис. 2.2)	99.98
Toyota Sequoia SUV 2012_001_changed (Рис. 2.7.1)	99.99
Toyota Sequoia SUV 2012_002 (Рис. 2.3)	99.99
Toyota Sequoia SUV 2012_002_changed (Рис. 2.7.2)	99.98
Toyota Sequoia SUV 2012_003 (Рис. 2.4)	99.82
Toyota Sequoia SUV 2012_003_changed (Рис. 2.7.3)	99.99
Toyota Sequoia SUV 2012_004 (Рис. 2.5)	*99.91*
Toyota Sequoia SUV 2012_004_changed (Рис. 2.7.4)	*99.95*

Прекрасные результаты!

Все изображения с высокой степенью уверенности верно классифицированы.

Код и другие материалы по данному эксперименту доступны на Github [12]

III. Нанесение дополнительного изображения

1) Подготовка и предварительная проверка

Из набора данных Stanford Car Dataset [9] выбран автомобиль Mazda Tribute SUV 2011 для дальнейшего использования в качестве основной модели для нанесения камуфляжа. Изображения автомобиля были подобраны таким образом, чтобы они не входили в тренировочный набор данных классификатора.

Всего использовалось три изображения автомобиля. Два изображения - для тренировочного набора переобучения (рис. 3.1 и рис. 3.2):

Рис. 3.1. Mazda Tribute SUV 2011 1.jpg

Рис. 3.2. Mazda Tribute SUV 2011 2.jpg

И одно - в качестве тестового (рис. 3.3):

Рис. 3.3. Mazda Tribute SUV 2011 3.jpg

Для нанесения рисунков на основную модель автомобиля выбраны по одному изображению автомобилей Lincoln Town Car Sedan 2011 (рис. 3.4):

Рис. 3.4. Lincoln Town Car Sedan 2011.jpg

И Tesla Model S Sedan 2012 (рис. 3.5):

Рис. 3.5. Tesla Model S Sedan 2012.jpg

На изображения основного автомобиля были наложены полупрозрачные изображения вспомогательных автомобилей для создания «камуфляжного» набора (рис. 3.6.1, рис. 3.6.2, рис. 3.7.1, рис. 3.7.2 и рис. 3.8.1, рис. 3.8.2):

Рис. 3.6.1. Mazda Tribute SUV 2011 101.jpg

Рис. 3.6.2. Mazda Tribute SUV 2011 102.jpg

Рис. 3.7.1. Mazda Tribute SUV 2011 201.jpg

Рис. 3.7.2. Mazda Tribute SUV 2011 202.jpg

Рис. 3.8.1. Mazda Tribute SUV 2011 301.jpg

Рис. 3.8.2. Mazda Tribute SUV 2011 302.jpg

Результаты классификации камуфляжного набора и составляющих его прототипов без добавления изображений в обучающий набор и переобучения модели:

Таблица 3.1

Изображение	**Оценка, %**
Lincoln Town Car Sedan 2011 (Рис. 3.4)	97.7810
Tesla Model S Sedan 2012 (Рис. 3.5)	98.2424
Mazda Tribute SUV 2011 1 (Рис. 3.1)	94.4616
Mazda Tribute SUV 2011 2 (Рис. 3.2)	99.2643
Mazda Tribute SUV 2011 3 (Рис. 3.3)	99.4660
Mazda Tribute SUV 2011 101 (Рис. 3.6.1)	-
Mazda Tribute SUV 2011 102 (Рис. 3.6.2)	-
Mazda Tribute SUV 2011 201 (Рис. 3.7.1)	-
Mazda Tribute SUV 2011 202 (Рис. 3.7.2)	59.4215
Mazda Tribute SUV 2011 301 (Рис. 3.8.1)	-
Mazda Tribute SUV 2011 302 (Рис. 3.8.2)	-

Из таблицы результатов видно, что классификатор «угадал» всего один автомобиль с нанесённым рисунком, но степень уверенности системы классификации близка к 50%, что сравнимо с подбрасыванием монетки.

Некамуфлированные изображения система классификации уверенно узнала и отнесла их к соответствующим классам. Анализ степени уверенности классификатора на неверно классифицированных изображениях представляется нецелесообразным
Атака успешна!

2) Включение прототипов в тренировочный набор

Добавим в тренировочный набор не камуфлированные изображения базовой модели автомобиля – Mazda Tribute SUV 2011 (рис. 3.1 и рис. 3.2) и переобучаем классификатор:

Таблица 3.2

Изображение	**Оценка, %**
Mazda Tribute SUV 2011 1 (Рис. 3.1)	99.8312
Mazda Tribute SUV 2011 2 (Рис. 3.2)	99.8933
Mazda Tribute SUV 2011 3 (Рис. 3.3)	99.2513
Mazda Tribute SUV 2011 101 (Рис. 3.6.1)	82.4236
Mazda Tribute SUV 2011 102 (Рис. 3.6.2)	-
Mazda Tribute SUV 2011 201 (Рис. 3.7.1)	43.5934
Mazda Tribute SUV 2011 202 (Рис. 3.7.2)	-
Mazda Tribute SUV 2011 301 (Рис. 3.8.1)	*68.8769*
Mazda Tribute SUV 2011 302 (Рис. 3.8.2)	-

Как видно из таблицы, классификатор «узнал» все изображения базового автомобиля (рис. 3.1, рис. 3.2 и рис. 3.3) – что неудивительно, так как рис. 3.1 и рис. 3.2 были добавлены в тренировочный набор перед переобучением, а также изображения автомобиля с нанесённым рисунком Lincoln Town Car Sedan 2011 (рис. 3.4) – рис. 3.6.1, рис. 3.7.1 и рис. 3.8.1.

Вероятно, что такая «разборчивость» классификатора была обусловлена тем фактом, что при нанесении изображения Lincoln Town Car Sedan 2011 (рис. 3.4) на борт автомобиля была выбрана большая степень прозрачности, чем при нанесении изображения Tesla Model S Sedan 2012 (рис. 3.5).

Также незначительно изменилась точность классификации изображения, не вошедшего в тренировочный набор (Mazda Tribute SUV 2011 3), но осталась очень высокой.

3) Состязательное обучение

Добавляем камуфлированные изображения основной модели автомобиля (кроме тестовых – выделены курсивом) в тренировочный набор и переобучаем классификатор (состязательная тренировка):

Таблица 3.3

Изображение	**Оценка, %**
Mazda Tribute SUV 2011 1 (Рис. 3.1)	99.8106
Mazda Tribute SUV 2011 2 (Рис. 3.2)	99.8774
Mazda Tribute SUV 2011 3 (Рис. 3.3)	99.5169
Mazda Tribute SUV 2011 101 (Рис. 3.6.1)	98.1727
Mazda Tribute SUV 2011 102 (Рис. 3.6.2)	98.8750
Mazda Tribute SUV 2011 201 (Рис. 3.7.1)	99.1301
Mazda Tribute SUV 2011 202 (Рис. 3.7.2)	99.8778
Mazda Tribute SUV 2011 301 (Рис. 3.8.1)	*96.1642*
Mazda Tribute SUV 2011 302 (Рис. 3.8.2)	*94.9404*

Прекрасные результаты! Все изображения с высокой степенью уверенности верно классифицированы.

Код и другие материалы по данному эксперименту доступны на Gitlab [13]

IV. Заключение

Основной заслугой данной работы является практическая демонстрация использования камуфляжа для организации состязательных атак на системы распознавания изображений. Также продемонстрирована работа (эффект) состязательных тренировок. Все результаты опубликованы на Github/Gitlab. Соответственно, получился практический стенд, который можно использовать “как есть”, а также проводить эксперименты с различными типами маскировок.

Благодарности

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

Библиография

[1] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22 (in Russian)
[2] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)
[3] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)
[4] Morgulis, Nir, et al. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019).
[5] Nassi, Ben, et al. "Phantom of the adas: Phantom attacks on driver-assistance systems." Cryptology ePrint Archive (2020).
[6] Knitting an anti-surveillance jumper https://kddandco.com/2022/11/02/knitting-an-anti-surveillance-jumper/ Retrieved: Aug, 2023
[7] Du, Andrew, et al. "Physical adversarial attacks on an aerial imagery object detector." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.
[8] Adhikari, Ajaya, et al. "Adversarial patch camouflage against aerial detection." arXiv preprint arXiv:2008.13671 (2020).
[9] Stanford Car Dataset https://www.kaggle.com/datasets/jutrera/stanford-car-dataset-by-classes-folder Retrieved: Aug, 2023
[10] DEEPBEAR “Pytorch car classifier” https://www.kaggle.com/code/deepbear/pytorch-car-classifier-90-accuracy Retrieved: Aug, 2023
[11] Gael Lederrey “camogen camouflage generator” https://github.com/glederrey/camogen Retrieved: Aug, 2023
[12] Пришлецов С.Е. “Физические атаки на систему классификации изображений посредством нанесения камуфляжа” https://github.com/sergiussrussia/resnet34_attacks Retrieved: Aug, 2023
[13] Пришлецов Д.Е. “Атака на систему классификации изображений: мимикрия” https://gitlab.com/Pukuluka/adversarial-training Retrieved: Aug, 2023

Camouflage as adversarial attacks on machine learning models

Dmitry Prishletsov, Sergey Prishletsov, Dmitry Namiot

Abstract **- The article is devoted to adversarial attacks on machine learning models. Such attacks are understood as the deliberate manipulation of data at different stages of the machine learning pipeline, designed to either prevent the correct operation of the model, or to achieve the desired result from it. In this case, physical evasion attacks were considered, that is, the objects themselves used in the model were modified. The article discusses the use of camouflaged images to deceive the recognition system. The experiments were carried out with a machine learning model that recognizes images of cars. Two types of camouflage patterns were used - classic camouflage and drawing an image of another car. In practice, such manipulations can be carried out using airbrushing. The work confirmed the successful implementation of such attacks, and obtained metrics characterizing the effectiveness of attacks, as well as the possibility of countering them using competitive training. All results are openly published, which makes it possible to use them as a software stand for testing other attacks and protection methods.**

Keywords **- adversarial attacks, camouflage, image recognition, machine learning.**

REFERENCES

[1] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22 (in Russian)

[2] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86. (in Russian)

[3] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)

[4] Morgulis, Nir, et al. "Fooling a real car with adversarial traffic signs." arXiv preprint arXiv:1907.00374 (2019).

[5] Nassi, Ben, et al. "Phantom of the adas: Phantom attacks on driver-assistance systems." Cryptology ePrint Archive (2020).

[6] Knitting an anti-surveillance jumper https://kddandco.com/2022/11/02/knitting-an-anti-surveillance-jumper/ Retrieved: Aug, 2023

[7] Du, Andrew, et al. "Physical adversarial attacks on an aerial imagery object detector." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[8] Adhikari, Ajaya, et al. "Adversarial patch camouflage against aerial detection." arXiv preprint arXiv:2008.13671 (2020).

[9] Stanford Car Dataset https://www.kaggle.com/datasets/jutrera/stanford-car-dataset-by-classes-folder Retrieved: Aug, 2023

[10] DEEPBEAR "Pytorch car classifier" https://www.kaggle.com/code/deepbear/pytorch-car-classifier-90-accuracy Retrieved: Aug, 2023

[11] Gael Lederrey "camogen camouflage generator" https://github.com/glederrey/camogen Retrieved: Aug, 2023

[12] Prishlecov S.E. "Fizicheskie ataki na sistemu klassifikacii izobrazhenij posredstvom nanesenija kamufljazha" https://github.com/sergiussrussia/resnet34_attacks Retrieved: Aug, 2023

[13] Prishlecov D.E. "Ataka na sistemu klassifikacii izobrazhenij: mimikrija" https://gitlab.com/Pukuluka/adversarial-training Retrieved: Aug, 2023