Факультет вычислительной математики и кибернетики

МГУ имени М.В. Ломоносова

Программа профессиональной переподготовки «Разработчик компьютерных технологий»

«Физические атаки на систему классификации изображений посредством нанесения камуфляжа»

Пришлецов С.Е.
Научный руководитель: Намиот Д.Е.

Постановка задачи

Цель работы – произвести физическую атаку на систему классификации изображений автомобилей путём нанесения на автомобиль камуфляжа.
Провести анализ результатов и переобучить сеть с использованием маскирующих изображений.
Оценить результаты классификации переобученной системы.

Актуальность задачи

Результаты данной работы позволят повысить точность распознавания объектов системой классификации, а также увеличат устойчивость сети к физическим атакам.

Возможно применение в бортовых системах автомобилей и строительной техники для обработки окружающей обстановки и предотвращения инцидентов.

Несомненна актуальность данного направления для правоохранительных органов: возможность классифицировать объекты, не смотря на возможные отражения на их поверхностях, злонамеренное введение в заблуждение систем видеонаблюдения и т.д.

Подход к реализации задачи

1. Взять готовую систему классификации изображений и протестировать её
2. Получить оценки точности классификации на тренировочном наборе
3. Нанести камуфляж на автомобили из тренировочного набора
4. Оценить точность классификации на камуфлированных изображениях
5. Сравнить работу классификатора на разных камуфлированных изображениях
6. Добавить камуфлированные изображения к тренировочному набору и перетренировать сеть
7. Запустить классификатор на наборе без камуфляжа и с камуфляжем (камуфляжами). Сравнить точность работы с предыдущими результатами

Реализация задачи

Выбран набор данных Stanford Car Dataset, 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 (автор алгоритма - DEEPBEAR: https://www.kaggle.com/code/deepbear/pytorch-car-classifier-90-accuracy).

Для проверки работоспособности, был добавлен новый класс (модель автомобиля). Система успешно справилась с поставленной задачей, отнеся автомобиль к верному классу (модели) с высокой долей вероятности.

Реализация задачи

В качестве автомобиля для нанесения камуфляжа был выбран Toyota Sequoia SUV 2012. Были найдены дополнительные изображения этого автомобиля, которые предварительно не участвовали ни в обучении, ни в тестировании модели.

Реализация задачи

На изображения автомобилей были наложены полупрозрачные изображения «камуфляжного» рисунка:






Реализация задачи

Результаты трёх этапов классификаций:

	Оценка, %		
Имя машины/ResNet34, номер поколения	**0**	**1**	**2**
Toyota Sequoia SUV 2012_001	78,54	99,98	99,98
Toyota Sequoia SUV 2012_001_changed	Fail	Fail	99,99
Toyota Sequoia SUV 2012_002	78,34	99,89	99,99
Toyota Sequoia SUV 2012_002_changed	Fail	Fail	99,98
Toyota Sequoia SUV 2012_003	89,12	99,96	99,82
Toyota Sequoia SUV 2012_003_changed	Fail	86,9	99,99
Toyota Sequoia SUV 2012_004	**64,78**	**99,66**	**99,91**
Toyota Sequoia SUV 2012_004_changed	**Fail**	**Fail**	**99,95**

_changed – замаскированное изображение машины.
Fail – машина идентифицировалась ошибочно.
Поколения сети:
0 – изображения использовались только в качестве тестовых
1 – незамаскированные изображения были добавлены в train этап сети (кроме 004)
2 – замаскированные изображения были добавлены в train (кроме 004)

Выводы

- Исходный код работы выложен в открытом доступе: https://github.com/sergiussrussia/resnet34_attacks
- Состязательная тренировка работает и позволяет успешно бороться с физическими атаками на систему классификации.
- Неизвестно, какие ещё могут нанести искажения на исходные данные для классификатора.
- Физические атаки на нейросети возможны и реализуемы
- Общедоступные системы классификации изображений очень хорошо работают даже на «домашних» аппаратных комплексах
- Методы противодействия – увеличение объёма и разнообразия тренировочных наборов данных (аугментация) для обучения классификаторов