3. Расширение фотографий: форматы записи цифровых фотографий. – URL: www.fotoprizer.ru/articles/teoria-fotografii/rasshirenie-fotografiy-formati-zapisi-cifrovih-fotografiy/153/?n=153&q=1335 (дата обращения 09.04.2022).

4. Как разоблачить фотоманипуляции // Рамблер Групп Ferra.ru. – URL: www.ferra.ru/review/multimedia/79974.htm (дата обращения 09.04.2022).

5. Jeronymo, D.C. Image forgery detection by semi-automatic wavelet soft-Thresholding with error level analysis / D.C. Jeronymo, Y.C.C. Borges, L.S. Coelho // Expert Systems with Applications. – 2017. – С. 348-356.

6. Sudiatmika, I.B.K. Image forgery detection using error level analysis and deep learning [Текст] / I.B.K. Sudiatmika, F. Rahman // Telecommunication Computing Electronics and Control. – 2019. – С. 653-659.

С.В. Толмачев, И.П. Болодурина, Д.И. Парфёнов, Л.С. Гришина, А.Ю. Жигалов

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ СВЁРТОЧНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ ПРИ ПРОВЕДЕНИИ СОСТЯЗАТЕЛЬНЫХ АТАК

(Оренбургский государственный университет)

Машинное обучение всё глубже проникает во все сферы нашей жизни, включая виртуальные ассистенты с голосовым управлением, компьютерное зрение и т.п. Глубокие нейронные сети (ГНС) являются один из наиболее распространенных инструментов решения подобных задач, т.к. способны устанавливать закономерности в неструктурированных данных, таких как изображения, видео- и аудиоинформация. Несмотря на то, что современные модели глубокого обучения достаточно надежны и вероятность их ошибки с каждым годом стремится к нулю, они по-прежнему подвержены так называемым состязательным атакам. Состязательный пример это вектор входных данных, для которых модель стабильно выдает предсказания, неверные с точки зрения человека. В связи с тем, что на системы искусственного интеллекта зачастую возлагается функция принятия решения, необходимо обеспечить их устойчивость к уязвимостям данного рода.

С момента первого упоминания на Международной конференции по репрезентационному обучению в 2014 году о наличии состязательной угрозы [1] для алгоритмов глубокого обучения было разработано большое количество методов генерации вредоносных входных данных и способов защиты от них [2]. В рамках настоящей работы попытаемся изучить некоторые типы состязательных атак с различными моделями угроз для построения в дальнейшем устойчивой модели глубокой нейронной сети для решения задач компьютерного зрения.
Рассмотрим следующую постановку задачи классификации изображений.

Пусть дана база данных изображений дорожных знаков для проведения многоклассовой классификации. База данных содержит тренировочный набор из 39209 размеченных изображений и тестовый набор объемом в 12630 преце-

дентов. Распределение обучающего набора изображений по классам представлено на рисунке 1.

Пусть X – множество изображений дорожных знаков, Y – множество классов дорожных знаков. Тогда обучающая выборка X^l, представляет собой множество пар объект-ответ $X^l = (x_i, y_i)_{i=1}^l$, где $x_i \in X$ – 8-битное RGB изображение 30×30 пикселей, заданное матрицей значений цветов, $y_i \in Y$ – известный класс дорожного знака на объекте.

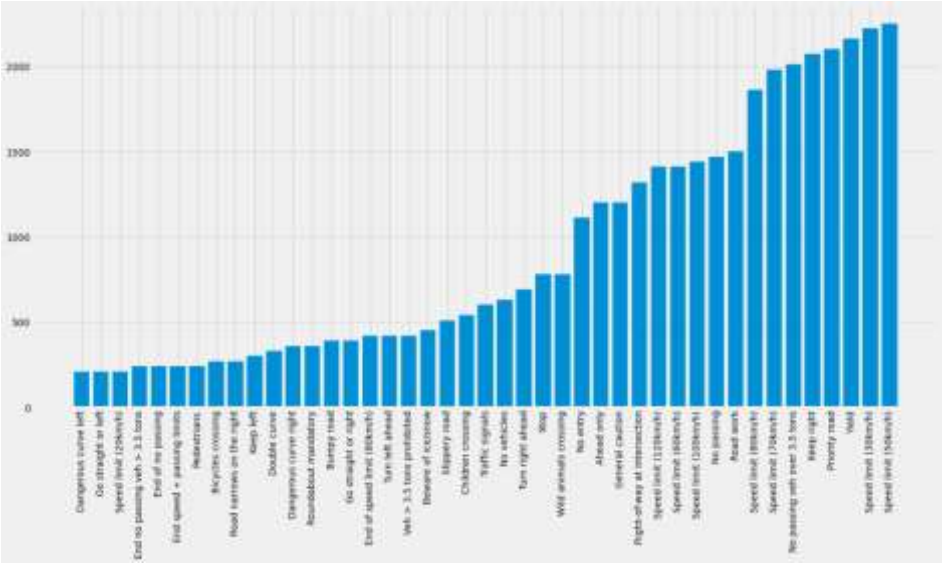

Рис. 1. Распределение изображений дорожных знаков

Пусть существует $y : X \to Y$ – некоторая зависимость, значения которой известны только на объектах обучающей выборки. Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

Таким образом, получена формальная постановка задачи многоклассовой классификации, которую будем решать с помощью ГНС. Архитектура свёрточной нейронной сети для распознавания изображений дорожных знаков представлена в таблице 1. Проведем исследование устойчивости данной модели к состязательной атаке на основе алгоритма быстрого градиента FGSM.

Таблица 1 – Структура свёрточной нейронной сети

Layer (type)	Output Shape	Param
conv2d (Conv2D)	(None, 28, 28, 32)	896
conv2d_1 (Conv2D)	(None, 26, 26, 64)	18496
max_pooling2d (MaxPooling2D)	(None, 13, 13, 64)	0
batch_normalization (BatchNormalization)	(None, 13, 13, 64)	256
conv2d_2 (Conv2D)	(None, 11, 11, 128)	73856
conv2d_3 (Conv2D)	(None, 9, 9, 512)	590336
max_pooling2d_1 (MaxPooling2D)	(None, 4, 4, 512)	0
batch_normalization_1 (BatchNormalization)	(None, 4, 4, 512)	2048

flatten (Flatten)	(None, 8192)	0
dense (Dense)	(None, 1024)	8389632
batch_normalization_2 (BatchNormalization)	(None, 1024)	4096
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 43)	44075

Отметим, что методы генерации вредоносных данных классифицируются по уровню доступа злоумышленника к модели на три категории:
- методы белого ящика – злоумышленник имеет полную информацию о модели, об алгоритме, используемом при обучении и может получить доступ к распределению обучающих данных. Он также знает параметры полностью обученной архитектуры модели;
- методы черного ящика предполагают отсутствие знаний о модели и корректировку вредоносных входных данных на основе результата, выдаваемого моделью;
- методы серого ящика с оценкой занимают промежуточное положение между методами белого и черного ящика, при которых злоумышленник имеет доступ к предварительным предсказаниям (оценкам) ГНС (например, к четырем наибольшим).

Предположим, что злоумышленник изучил архитектуру нейросетевого решения (атака белого ящика), и решает провести атаку методом быстрого градиента FGSM, который вносит в каждое входное значение небольшие искажения в соответствии со знаком градиента функции потерь, которая используется при обучении сети, не учитывая относительную важность отдельных изменений [3].

Пример исходного изображения и соответствующего состязательного примера представлен на рисунке 2.

Рис. 2. Исходное изображение, состязательный пример атаки FGSM и разница изображений

В результате проведения ряда экспериментов на устойчивость глубокой нейронной сети было сгенерировано 50 состязательных примеров (из них 20 примеров знаков ограничения скорости в 30 км/ч, 10 знаков STOP и 20 знаков «главная дорога») на основе атаки FGSM. Данные изображения были классифицированы как знаки ограничения скорости, дорожных работ, скользкой дороги и знака «уступи дорогу». Кроме того, 5 изображений знаков STOP не были

корректно обработаны алгоритмом. Таким образом, эффективность проведения атаки FGSM соответствует 90% и модель глубокой нейронной сети для распознавания изображений дорожных знаков неустойчива к атаке белого ящика FGSM.

Исследование выполнено при финансовой поддержке РФФИ (проект № 20-07-01065) и гранта Президента Российской Федерации для государственной поддержки молодых российских ученых - кандидатов наук (№ МК-258.2022.1.6), а также стипендии Президента Российской Федерации молодым ученым и аспирантам (СП-3652.2021.5 и № СП-919.2022.5).

Литература

1. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. CoRR abs/1312.6199 (2013). http://arxiv.org/abs/1312.6199

2. Uri Shaham, Yutaro Yamada, and Sahand Negahban. 2015. Understanding Adversarial Training: Increasing Local Stability of Neural Nets through Robust Optimization. CoRR abs/1511.05432 (2015). http://arxiv.org/abs/1511.05432

3. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. CoRR abs/1412.6572 (2014). http://arxiv.org/abs/1412.6572

Ю.А. Ургалкина, И.В. Семенова

РЕАЛИЗАЦИЯ ПОИСКА КУЛИНАРНЫХ РЕЦЕПТОВ ПО РЕЗУЛЬТАМ РАСПОЗНАВАНИЯ ИНГРЕДИЕНТОВ НА ИЗОБРАЖЕНИИ

(Самарский университет)

Поиск информации составляет неотъемлемую часть современного мира. В настоящее время поиск, основанный на технологии компьютерного зрения, используется во многих отраслях, так как помогает пользователю упростить нахождение нужной информации. При этом систем, которые позволили бы осуществлять поиск кулинарных рецептов по результатам распознавания ингредиентов на изображении, на сегодняшний день не существуют.

Так как на одном изображении может находиться одновременно несколько продуктов, алгоритм определения ингредиентов, находящихся на изображении, можно разделить на два основных этапа: обнаружение (сегментация) и распознавание (классификация) образов.

Сегментация изображения является важным предварительным шагом большинства задач автоматического распознавания образов. Сегментация – это разбиение изображения на области, однородные по некоторому признаку и покрывающие всё изображение.