УДК 004.932.2
DOI: 10.25559/SITITO.18.202204.861-871

Оригинальная статья

Исследование эффективности трансферного обучения при проведении состязательных атак

Д. И. Парфёнов*, И. П. Болодурина, Л. С. Гришина, А. Ю. Жигалов, С. В. Толмачев
ФГБОУ ВО «Оренбургский государственный университет», г. Оренбург, Российская Федерация
Адрес: 460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13
* parfenovdi@mail.ru

Аннотация

Глубокие нейронные сети становятся всё более эффективным инструментом решения широкого спектра сложных прикладных задач, т.к. способны устанавливать закономерности в неструктурированных данных, таких как изображения, видео- и аудиоинформация. Несмотря на то, что вероятность ошибки современных моделей нейронных сетей снижается, ML-модели остаются уязвимыми к состязательным атакам. В связи с этим, вопросы обеспечения безопасности моделей машинного обучения от состязательных атак являются одними из наиболее актуальных. В рамках данной работы проведено исследование эффективности применения технологии трансферного обучения для повышения устойчивости сверточных нейронных сетей от состязательных атак. Исследование проводилось на примере задачи распознавания дорожных знаков на снимках, которая является ключевой в области автомобильных беспилотных транспортных сетей. В ходе эксперимента построена базовая модель нейронной сети для решения задачи классификации дорожных знаков и реализованы атаки на основе алгоритмов PGD и FGSM, которые снизили точность классификации примерно в 2,8 раза. Для защиты от рассмотренных атак белого ящика на нейронную сеть проведено состязательное обучение, которое повысило точность модели на 12,11%. Кроме того, для рассмотрена технология трансферного обучения, которое позволило добиться повышения надежности после состязательного обучения на 42,38% по сравнению с исходным набором данных. Полученные результаты исследования могут быть использованы при разработке защищенной интеллектуальной системы поддержки принятия решений для обнаружения на кадре с видеорегистратора области, содержащей дорожный знак, и последующей классификации выделенного сегмента изображения.

Ключевые слова: свёрточная нейронная сеть, состязательные атаки, трансферное обучение, распознавание изображений, автомобильные беспилотные транспортные сети

Финансирование: исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 20-07-01065 «Разработка интеллектуальных методов адаптивного управления безопасностью и верификации работы виртуальной сетевой инфраструктуры мультиоблачной платформы для обработки больших данных», а также гранта Президента Российской Федерации для государственной поддержки молодых российских ученых – кандидатов наук № МК-258.2022.1.6 «Методы и алгоритмы автоматической интеллектуальной обработки больших массивов слабоструктурированных данных для защиты сетей 5G» и стипендии Президента Российской Федерации для молодых ученых и аспирантов № СП-3652.2021.5 «Методы и алгоритмы автоматизации машинного обучения для выявления инцидентов кибербезопасности в беспроводных транспортных сетях с динамической топологией».

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Исследование эффективности трансферного обучения при проведении состязательных атак / Д. И. Парфёнов [и др.] // Современные информационные технологии и ИТ-образование.2022.Т.18,№4.С.861-871.doi:https://doi.org/10.25559/SITITO.18.202204.861-871

Original article

Investigation of Transfer Learning Efficiency in Adversarial Attacks

D. I. Parfenov*, I. P. Bolodurina, L. S. Grishina, A. Yu. Zhigalov, S. V. Tolmachev
Orenburg State University, Orenburg, Russian Federation
Address: 13 Prospect Pobedy, Orenburg 460018, Russian Federation
* parfenovdi@mail.ru

Abstract
Deep neural networks are becoming an increasingly effective tool for solving a wide range of complex applied tasks, because they are able to establish patterns in unstructured data, such as images, video and audio information. Despite the fact that the probability of error of modern neural network models decreases, ML models remain vulnerable to adversarial attacks. In this regard, the issues of ensuring the security of machine learning models from adversarial attacks are among the most urgent. Within the framework of this work, a study was conducted on the effectiveness of the transfer learning technology to increase the stability of convolutional neural networks from adversarial attacks. The study was conducted on the example of the task of recognizing road signs in images, which is key in the field of automotive unmanned transport networks. During the experiment, a basic neural network model was built to solve the problem of classifying road signs and attacks based on PGD and FGSM algorithms were implemented, which reduced the classification accuracy by about 2.8 times. To protect against the considered white box attacks on the neural network, competitive training was conducted, which increased the accuracy of the model by 12.11%. In addition, the technology of transfer training was considered, which made it possible to achieve an increase in reliability after competitive training by 42.38% compared to the original data set. The obtained research results can be used in the development of a secure intelligent decision support system for detecting an area containing a road sign on a DVR frame and subsequent classification of the selected image segment.

Keywords: convolutional neural network, adversarial attacks, transfer learning, image recognition, automotive unmanned transport networks

Funding: The study was carried out with the financial support of the Russian Foundation for Basic Research within the framework of scientific project No. 20-07-01065 "Development of Intelligent Methods of Adaptive Security Management and Verification of the Virtual Network Infrastructure of the Multicloud Platform for Big Data Processing", as well as the grant of the President of the Russian Federation for state support of young Russian scientists – candidates of sciences No. MK-258.2022.1.6 "Methods and Algorithms for Automatic Intellectual Processing of Large Arrays of Semi-Structured Data for 5G Network Protection" and scholarships of the President of the Russian Federation for young scientists and graduate students No. SP-3652.2021.5 "Methods and algorithms of machine learning automation for detection of cybersecurity incidents in wireless trans

Conflict of interests: The authors declare no conflict of interest.

For citation: Parfenov D.I., Bolodurina I.P., Grishina L.S., Zhigalov A.Yu., Tolmachev S.V. Investigation of Transfer Learning Efficiency in Adversarial Attacks. *Modern Information Technologies and IT-Education.* 2022;18(4):861-871. doi: https://doi.org/10.25559/SITITO.18.202204.861-871

Введение

В настоящий момент активное распространение приложений и устройств, использующих в своей основе модели машинного обучения, предоставили новые возможности к автоматизации большинства процессов обработки больших данных, но вместе с тем подтолкнули развитие так называемых состязательных атак [1, 2]. Перехват информации при обучении ML-моделей, а также злонамеренная корректировка входного сигнала при использовании моделей зачастую недостаточно продумана на этапе проектирования приложений искусственного интеллекта и оставляет место для злоумышленника. В связи с этим, вопросы обеспечения безопасности от состязательных атак являются одними из наиболее актуальных и требуют разработки новых концепций защиты.

В прикладным областях, где от ошибки ML-модели могут произойти критически опасные ситуации, особенно важно уделять внимание защите от состязательных атак. Например, автономные транспортные средства (ТС) реализуют стратегии передвижения от системы управления на основе интеллектуальных алгоритмов. В большинстве случаев такие механизмы основаны на реализации методов компьютерного зрения [3, 4], для распознавания дорожной разметки, знаков, других ТС и т.д. На данный момент методы глубокого обучения – одни из самых популярных и эффективных инструментов решения подобных задач, однако и они подвержены состязательным атакам.

Цель исследования

Проведение злонамеренной корректировки входных изображений дорожных знаков может провести к авариям и тяжелым последствиям. В рамках данной статьи проведено исследование устойчивости модели сверточной нейронной сети (CNN), классифицирующей знаки, к состязательным атакам для разработки защищенного ML-сервиса [5].

Особенность данной работы заключается в реализации трансферного обучения и использования его основных преимуществ: построения алгоритмов с высокой обобщающей способностью, а также оптимального управления ресурсами, данными, временем и вычислительной мощностью [6]. В рамках данного исследования представлен подробный анализ эффективности трансферного обучения при построении модели CNN, устойчивой к состязательным атакам белого ящика.

Технологии обеспечения безопасности нейросетевых моделей

Существует довольно большое количество технологий проведения состязательных атак и предприняты различные попытки построить средства защиты, которые повышают надежность нейронной сети. Рассмотрим основные подходы, использующиеся на данный момент, для построения состязательных атак и моделей машинного обучения, устойчивых к ним.

Так, команда Facebook AI Research в исследовании [7] предложила первую универсальную состязательную атаку на основе градиента против моделей-трансформеров. Вместо поиска одного состязательного примера рассмотрено распределение состязательных примеров, параметризованное матрицей с непрерывными значениями, что обеспечивает оптимизацию на основе градиента.

Авторы исследования [8] представили подход к проведению состязательной атаки на систему распознавания лиц, добавляя невидимый человеческому глазу шум в исходные данные, в результате чего изображения не распознаются в большинстве случаев.

В работе [9] авторами предложены три новых алгоритма состязательной атаки, которые успешно обходят защитную дистилляцию при классификации изображений. Авторы предлагают использовать состязательные примеры с высокой степенью достоверности при переносимости атаки, что также может быть использовано для преодоления защитной дистилляции. В исследовании [10] предложен подход к генерации состязательных примеров с помощью генеративно-состязательных нейронных сетей. Предложенные подходы показали высокую эффективность среди атак серого и черного ящика.

Авторы исследования [11] предложили систематический метод для вычисления универсальных возмущений и показали, что современные глубокие нейронные сети к ним уязвимы. Данные возмущения были почти незаметными для человеческого глаза, при этом они хорошо обобщаются в нейронных сетях. Существование универсальных возмущений выявило геометрическую корреляцию между многомерными границами решений классификаторов, что может ухудшать модели распознавания. В работе [12] представлена структура для создания универсальных и зависящих от изображения возмущений для задач классификации и семантической сегментации изображения. Результаты работы алгоритмов генерации атак по скорости превосходят итерационные.

Для защиты от атак злоумышленников было предложено несколько методов укрепления нейронных сетей. Так, в работе [13] авторами проведен обзор эффективности состязательных атак таких как, атаки отравления, атаки черного ящика и белого ящика на различные типы архитектур нейронных сетей, кроме того рассмотрены стратегии защиты от представленных атак. В рамках статьи [14] исследованы вопросы использования технологии автоматического машинного обучения (AutoML) для создания комплексных и защищённых моделей, а также вопросы интерпретируемости выходных данных нейронной сети для повышения устойчивости к состязательным атакам.

Авторы статьи [15] предложили два подхода к состязательному обучению с целью улучшения производительности и устойчивости модели машинного обучения. В первом случае выполняется сужение исходного образца данных и его состязательного образца в пространстве представления, увеличивая их расстояние от различных размеченных образцов, во втором случае модель восстанавливает исходный образец данных из его состязательного представления.

В рамках исследования [16] представлена библиотека Foolbox Native на языке Python для тестирования устойчивости моделей машинного обучения по отношению к состязательным атакам. Авторы работы [17] исследовали состязательную устойчивость нейронных сетей с точки зрения робастной оптимизации, показывая высокую эффективность разрабатываемых решений на наборе данных MNIST. В работе [18] авторы

рассмотрели состязательные атаки в беспилотных транспортных системах с точки зрения статистической механики и предложили модель интерпретации состязательной устойчивости, основанную на статистической механике. В исследовании [19] представлена ансамблевая стратегия защиты от градиентных состязательных атак распознавания изображений на основе техники переобучения для набора данных CIFAR-10.

В работе [20] авторы представили структуру автоматической проверки полносвязных нейронных сетей на основе теории выполнимости по модулю (SMT), а в исследовании [21] другим научным коллективом представлен алгоритм защитной дистилляции для снижения эффективности враждебных выборок в глубоких нейронных сетях.

В исследовании [22] авторами представлен метод защиты от состязательных атак на основе ReLU. Предложенная идея состоит в том, что функция активации ReLU имеет ограничения сверху, благодаря чему сглаживается карта градиентов, и в каждой точке входного распределения уже нельзя будет получить всплеск.

Таким образом, анализ работ в области состязательных атак на глубокие нейронные сети показывал, что рассматриваемая в данном исследовании тема является актуальной. Кроме того, в настоящее время существует большое множество алгоритмов генерации состязательных примеров и нейронные сети для задач компьютерного зрения подвержены «переносу состязательных атак». В ходе данного исследования будет рассмотрено применение стратегии трансферного обучения для повешения устойчивости от состязательных атак.

Формальная постановка задачи классификации дорожных знаков

Рассмотрим задачу построения модели распознавания дорожного знака по изображению как задачу классификации, для решения которой были рассмотрены открытые наборы данных изображений дорожных знаков, а именно: немецкий (GTSRB, [23]) и российский (RTSD, [24]). Количественные характеристики наборов дорожных знаков представлены в таблице 1.

Таблица 1. **Информация о наборах данных**
Table 1. **Information about data sets**

	GTSRB	*RTSD*
Количество типов знаков	43	156
Количество физических знаков	1728	15630
Количество изображений знаков	51839	104358

Набор данных RTSD имеет большое количество кадров, классов и физических знаков, поэтому будем строить модель распознавания российских дорожных знаков. За основу для построения защищенного от состязательных атак классификатора возьмем немецкий набор данных GTSRB, обучение с которого затем перенесем на российский набор автодорожных знаков.

Формальная постановка задачи многоклассовой классификации, которую будем решать с помощью сверточных нейронных сетей:

Пусть задано X – множество изображений дорожных знаков (30´30 пикселей), Y – множество классов дорожных знаков и существует некоторая неизвестная зависимость $y: X \to Y$ – классификатор. Необходимо по объектам $\{x_1,...,x_l\} \subset X$ обучающей выборки и известным ответам $y_i = y(x_i), i = 1,...,l$ построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$, представленный соответствующей матрицей значений цветов.

Для распознавания изображений дорожных знаков, воспользуемся моделью сверточной нейронной сети, архитектура которой имеет 4 сверточных слоя без кросс-соединений, со слоями MaxPool2D и BatchNormalization после 2 и 4 свёрточного слоя. Слои классификатора состоят из 2-х полносвязных слоев с операцией нормализации и dropout. Структура сверточной нейронной сети представлена на рис. 1.

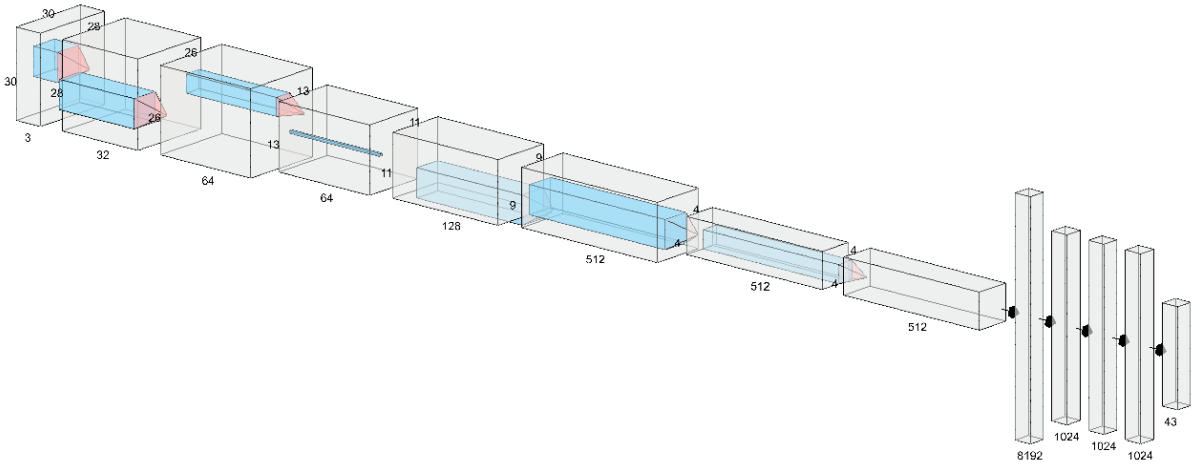

Рис. 1. Структура сверточной нейронной сети
Fig. 1. Structure of a Convolutional Neural Network

Обучим данную модель CNN и проанализируем ее устойчивость к состязательным атакам. Обучающий и тестовый набор данных разбиты в соотношении 5:1. Точность модели составила 98%, что соответствует высокому качеству идентификации дорожных знаков (Таблица 2).

Таблица 2. **Результаты обучения CNN на исходных данных**
Table 2. **CNN training results based on the source data**

	Precision	**Recall**	**F1-score**
accuracy	–	–	0,98
macro avg	0,97	0,97	0,97
weighted avg	0,98	0,98	0,98

Алгоритмы для генерации вредоносных данных

В зависимости от уровня доступа злоумышленника к модели разделяют состязательные атаки белого и черного ящика. В первом случае методы генерации вредоносных изображений основываются на полной информации о модели, во втором – информации нет и корректировка изображения осуществляется на основе результатов выводов ML-моделей. В размах данного исследования предполагается, что информация о структуре модели CNN не была должным образом защищена.

Наиболее распространенная состязательная атака белого ящика – *метод быстрой смены знака градиента (FGSM)*, который генерирует состязательные образцы в L_∞ окрестности с неатакованными примерами. Данный алгоритм выполняет одноэтапное обновление вдоль направления градиента состязательных потерь $J(\theta, x, y)$, чтобы увеличить потери в значимом направлении (например, верного класса изображения), т.е. состязательный пример FGSM имеет вид:

$$x' = x + \varepsilon \cdot sign\left[\nabla_x J(\theta, x, y)\right], \quad (1)$$

где ε – величина возмущения.

FGSM может быть легко расширен до алгоритма целевой атаки, путем спуска по градиенту $J(\theta, x, y')$, в котором y' – целевая метка. Данная процедура обновления может уменьшить перекрестную энтропию между прогнозируемым вектором вероятности и целевым вектором вероятности, если перекрестная энтропия применяется в качестве состязательной потери. Кроме того, было обнаружено, что случайное возмущение перед выполнением FGSM может повысить производительность и разнообразие состязательных образцов.

Повышение производительности путем итеративного оптимизатора реализовано в методе BIM, который выполняет FGSM с меньшим размером шага и ограничивает состязательный образец в пределах допустимого диапазона для T итераций. Таким образом, в t -й итерации правило выглядит следующим образом:

$$x'_{t+1} = Clip\left\{x'_t + \alpha \cdot sign\left[\nabla_x J(\theta, x'_t, y)\right]\right\}, \quad (2)$$

где α – величина возмущения в каждой итерации, $\alpha T = \varepsilon$.

Projected Gradient Descent (PGD) можно рассматривать как обобщенную версию BIM без ограничения $\alpha T = \varepsilon$: для ограничения состязательного возмущения, алгоритм проецирует состязательные образцы, полученные на каждой итерации, в ε - L_∞ окрестность доброкачественных образцов:

$$x'_{t+1} = \mathrm{Proj}\left\{x'_t + \alpha \cdot sign\left[\nabla_x J(\theta, x'_t, y)\right]\right\}, \quad (3)$$

где функция Proj проецирует обновленный состязательный образец в ε - L_∞ окрестность и допустимый диапазон.

В таблице 3 представлены результаты оценки точности исходной модели CNN при проведении состязательных атак FGSM и PGD. Отметим, что атака PGD значительнее повлияла на качество классификации дорожных знаков в 2,8 раза. В связи с этим, рассмотрим возможные способы защиты от состязательных атак подобного типа.

Таблица 3. **Результаты состязательных атак на исходную модель**
Table 3. **Results of competing attacks on the original model**

	FGSM	*PGD*
Accuracy	50,08%	34,71%

Стоит отметить, что обеспечить полную защиту от данного рода угроз невозможно по причине отсутствия генеральной совокупности данных, но минимизация влияния вредоносных примеров на этапе обучения или эксплуатации модели вполне решаемая задача. Большинство современных стратегий защиты не адаптируются ко всем типам атак, поскольку один метод может блокировать один тип атаки, но оставляет другую уязвимость открытой для злоумышленника, который знает лежащий в основе защитный механизм. Кроме того, реализация таких стратегий защиты может привести к снижению производительности, а также может снизить точность прогнозирования реальной модели.

Стандартным методом защиты грубой силы является состязательное обучение [25], при котором надежность модели повышается за счет включения в обучающую выборку состязательных примеров, получаемых преимущественно с использованием алгоритмов белого ящика, описанных ранее. Состязательное обучение повышает устойчивость модели к атакам белого ящика.

Другим методом упрочнения моделей является сжатие элементов входных данных [26]. Основная идея защиты заключается в том, что оно снижает сложность представления данных, тем самым устраняя вредоносные возмущения. Для задачи классификации изображений может уменьшаться глубина цвета на уровне пикселей или использоваться сглаживающий фильтр. Данный метод хорошо предотвращает состязательные атаки, но имеет побочный эффект в виде ухудшении точности модели на реальных примерах.

В рамках данного исследования будет рассмотрен механизм защиты модели CNN от атак FGSM и PGD на основе состязательного обучения: сгенерируем состязательные образцы на основе атаки PGD, так как она оказала наибольшее влияние на исходную модель, добавим их к обучающему набору данных и проведем заново обучение модели CNN. В таблице 4 представлены результаты оценки точности модели CNN после состязательного обучения при проведении атак FGSM и PGD.

Таблица 4. **Результаты состязательных атак на исходную модель**
Table 4. **Results of competing attacks on the original model**

	Bening	*FGSM*	*PGD*
Accuracy	96,84% (-0,62%)	56,29% (+6,21%)	46,82% (+12,11%)

Отметим, что точность модели распознавания дорожных знаков на нормальных данных незначительно снизилась (-0.62%), а на состязательных примерах рассмотренных типов атак повысилась. Однако, полученные результаты устойчивости модели CNN к атакам белого ящика свидетельствуют о необходимости применения дополнительных мер защиты. В связи с этим, проведем исследование эффективности применения технологии трансферного обучения для повышения устойчивости сверточных нейронных сетей от состязательных атак.

Трансферное обучение

В машинном обучении под трансферным обучением понимается техника, когда в качестве отправной точки используется предобученная модель на данных похожей задачи. Используя трансферное обучение можно повысить точность для задачи на малом наборе данных дообучая нейронную сеть.

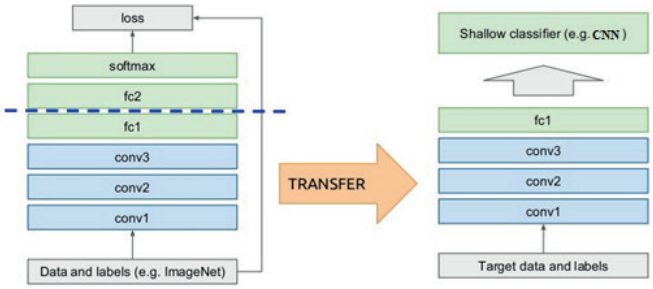

Рис. 2. Схема переноса обучения
Fig. 2. Training transfer scheme

К достоинствам трансферного обучения можно отнести то, что нет необходимости в наличии большого набора данных для задачи, так как для корректировки весов можно использовать небольшой набор данных, в то время как нейронная сеть будет знать зависимости и веса на обучившись на схожей задаче. В отличие от этого, трансферное обучение использует знания, полученные из предварительно обученной модели, для продолжения выполнения задачи.

Обучим построенный защищенный классификатор немецких дорожных знаков распознавать российские дорожные знаки, используя трансферное обучение.

Построим базовую модель на основе нейронной сети, прошедшей состязательное обучение на наборе данных GTSRB, изменив размерность последнего полносвязного слоя для классов набора данных RTSD. Проанализируем эффективность применения трансферного обучения при построении модели CNN, устойчивой к состязательным атакам белого ящика.

Результаты экспериментального исследования

В рамках данного исследования с помощью метода PGD созданы вредоносные изображения на основе валидационного набора данных и дообучена модель CNN на полученных изображениях. Результат эксперимента показывает, что нейронная сеть стала чуть более устойчива к состязательным атакам, а наибольшее влияние на точность по-прежнему оказывает метод PGD.

Проанализируем эффективность применения трансферного обучения и перенесем модель на набор данных RTSD: Заморозим для обучения все слои базовой модели, кроме последнего, и обучим модель в течение 30 эпох. После этого разморозим дополнительно 2 верхних слоя и продолжим обучение в течении ещё 20 эпох.

Точность нейронной сети после трансферного обучения и настройки составила 95,24%, сбалансированная точность равняется 95,00%. Изменение точности (accuracy) и ошибки (loss) по эпохам зафиксировано на рис. 3 и 4 соответственно.

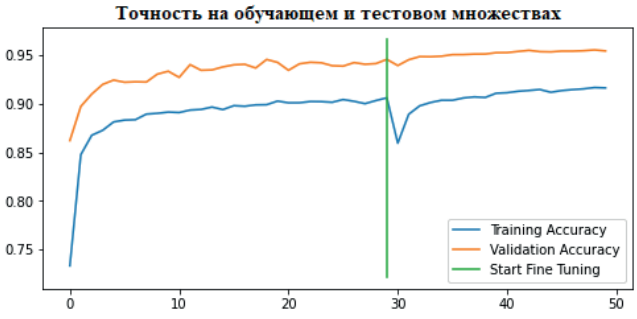

Рис. 3. Зависимость точности (accuracy) от количества эпох обучения
Fig. 3. Dependence of accuracy on the number of training epochs

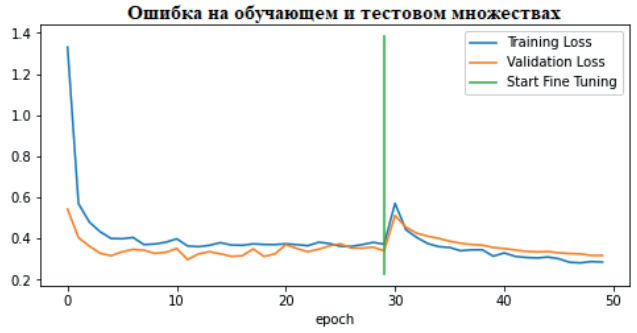

Рис. 4. Зависимость ошибки (loss) от количества эпох обучения
Fig. 4. Dependence of the error (loss) on the number of training epochs

Точность модели CNN по распознаванию дорожных знаков после трансферного обучения при проведении атак FGSM и PGD представлена в таблице 5.

Таблица 5. **Результаты атак на модель после трансферного обучения**
Table 5. **Results of attacks on the model after transfer training**

Набор данных	*Bening*	*FGSM*	*PGD*
GTSRB	96,84%	56,29%	46,82%
RTSD	95,24%	92,65%	92,61%

Можно заметить, что после переноса обучения защитные свойства сохранились, а большее количество классов в целевой модели значительно повлияло на улучшение значений метрики при состязательных атаках. Трансферное обучение позволило добиться повышения надежности после состязательного обучения на 42,38% по сравнению с исходным набором данных. Таким образом, построен защищенный от состязательных атак классификатор российских дорожных знаков.

Заключение

В рамках данной работы проведено исследование эффективности применения технологии трансферного обучения для повышения устойчивости сверточных нейронных сетей от состязательных атак. Исследование проводилось на примере задачи распознавания дорожных знаков на снимках, которая является ключевой в области автомобильных беспилотных транспортных сетей.

В ходе эксперимента построена базовая модель нейронной сети для решения задачи классификации дорожных знаков и реализованы атаки на основе алгоритмов PGD и FGSM, которые снизили точность классификации примерно в 2,8 раза. Для защиты от рассмотренных моделей атак на нейронную сеть проведено состязательное обучение на сгенерированных изображениях с помощью алгоритма PGD, который оказывал наибольшее влияние на точность классификации дорожных знаков. Точность модели после состязательного обучения повысилась на 12,11%. Кроме того, для повышения устойчивости нейронной сети также рассмотрена технология трансферного обучения. Модель CNN, прошедшая состязательное обучение на наборе данных GTSRB, использована при решении схожей задачи классификации дорожных знаков на российском наборе данных RTSD. Трансферное обучение позволило добиться повышения надежности после состязательного обучения на 42,38% по сравнению с исходным набором данных. Таким образом, построен защищенный от состязательных атак классификатор российских дорожных знаков.

Список использованных источников

[1] Ковалев И. А., Косов Н. А. Состязательные атаки в нейронных сетях // Студенческий вестник. 2021. № 20-11(165). С. 36-43. URL: https://elibrary.ru/item.asp?id=46213350 (дата обращения: 16.10.2022).

[2] Колесниченко М. Д. Об уязвимости искусственных нейронных сетей к состязательным атакам // Актуальные проблемы авиации и космонавтики / под ред. Ю. Ю. Логинова. Т. 2. Красноярск : СибГУ имени М. Ф. Решетнёва, 2020. С. 228-229. URL: https://elibrary.ru/item.asp?id=45617596 (дата обращения: 16.10.2022).

[3] Бодунков Н. Е., Арефин В. В., Кобринец С. К. Исследование методов атак на сверточные нейронные сети // Беспилотные транспортные средства с элементами искусственного интеллекта (БТС-ИИ-2021). М. : РАИИ, 2021. С. 122-125. URL: https://elibrary.ru/item.asp?id=48235793 (дата обращения: 16.10.2022).

[4] Войнов Д. М. Исследование взаимосвязей между параметрами состязательных атак и ошибок классификации изображений нейронными сетями // 76-я научная конференция студентов и аспирантов Белорусского государственного университета. Минск : БГУ, 2019. С. 47-51. URL: https://elibrary.ru/item.asp?id=41203721 (дата обращения: 16.10.2022).

[5] Voynov D. M., Kovalev V. A. A comparative study of white-box and black-box adversarial attacks to the deep neural networks with different architectures // Proceedings of the II International Scientific Conference on Computer Technologies and Data Analysis (CTDA'2020). Minsk : BSU, 2020. P. 185-189. URL: https://elib.bsu.by/handle/123456789/248666 (дата обращения: 16.10.2022).

[6] Чертилин К. Э., Ивченко В. Д. Разработка и трансферное обучение искусственной нейронной сети для классификации изображений // Приборы. 2022. № 7(265). С. 37-47. URL: https://elibrary.ru/item.asp?id=49450885 (дата обращения: 16.10.2022).

[7] Bose A. J., Aarabi P. Adversarial Attacks on Face Detectors Using Neural Net Based Constrained Optimization // 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). Vancouver, BC, Canada : IEEE Computer Society, 2018. P. 1-6. doi: https://doi.org/10.1109/MMSP.2018.8547128

[8] Gradient-based Adversarial Attacks against Text Transformers / Ch. Guo [и др.] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic : Association for Computational Linguistics, 2021. P. 5747-5757. doi: https://doi.org/10.18653/v1/2021.emnlp-main.464

[9] Carlini N., Wagner D. Towards evaluating the robustness of neural networks // 2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA : IEEE Computer Society, 2017. P. 39-57. doi: https://doi.org/10.1109/SP.2017.49

[10] Generating adversarial examples with adversarial networks / Ch. Xiao [и др.] // Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018) ; ed. by J. Lang. Stockholm, Sweden : IJCAI, 2018. P. 3905-3911. doi: https://doi.org/10.24963/ijcai.2018/543

[11] Universal Adversarial Perturbations / S.-M. Moosavi-Dezfooli [и др.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA : IEEE Computer Society, 2017. P. 86-94. doi: https://doi.org/10.1109/CVPR.2017.17

[12] Generative Adversarial Perturbations / O. Poursaeed [и др.] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA : IEEE Computer Society, 2018. P. 4422-4431. doi: https://doi.org/10.1109/CVPR.2018.00465

[13] Towards Deep Learning: A Review On Adversarial Attacks / M. M. Irfan [и др.] // 2021 International Conference on Artificial Intelligence (ICAI). Islamabad, Pakistan : IEEE Computer Society; 2021. P. 91-96. doi: https://doi.org/10.1109/ICAI52203.2021.9445247

[14] Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX / J. Rauber [и др.] // Journal of Open Source Software. 2020. Vol. 5, issue 53. Article number: 2607. doi: https://doi.org/10.21105/joss.02607

[15] Towards Deep Learning Models Resistant to Adversarial Attacks / A. Madry [и др.] // Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, BC, Canada : Vancouver Convention Center, 2018. P. 1-23. URL: https://openreview.net/pdf?id=rJzIBfZAb (дата обращения: 16.10.2022).

[16] Interpreting Adversarial Examples and Robustness for Deep Learning-Based Auto-Driving Systems / K. Wang [и др.] // IEEE Transactions on Intelligent Transportation Systems. 2022. Vol. 23, no. 7. P. 9755-9764. doi: https://doi.org/10.1109/TITS.2021.3108520

[17] Mani N., Moh M., Moh T. S. Defending Deep Learning Models Against Adversarial Attacks // International Journal of Software Science and Computational Intelligence (IJSSCI). 2021. Vol. 13, issue 1. P. 72-89. doi: http://doi.org/10.4018/IJSSCI.2021010105

[18] Two to Trust: AutoML for Safe Modelling and Interpretable Deep Learning for Robustness / M. Amirian [и др.] // Trustworthy AI – Integrating Learning, Optimization and Reasoning. TAILOR 2020. Lecture Notes in Computer Science ; ed. by F. Heintz, M. Milano, B. O'Sullivan. Vol. 12641. Cham : Springer, 2021. P. 268-275. doi: https://doi.org/10.1007/978-3-030-73959-1_23

[19] Qiu Y., Zhang J., Zhou J. Improving Gradient-based Adversarial Training for Text Classification by Contrastive Learning and Auto-Encoder // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021. P. 1698-1707. doi: https://doi.org/10.18653/v1/2021.findings-acl.148

[20] Safety Verification of Deep Neural Networks / X. Huang [и др.] // Computer Aided Verification. CAV 2017. Lecture Notes in Computer Science ; ed. by R. Majumdar, V. Kunčak. Vol. 10426. Cham : Springer, 2017. P. 3-29. doi: https://doi.org/10.1007/978-3-319-63387-9_1

[21] Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks / N. Papernot [и др.] // 2016 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA : IEEE Computer Society, 2016. P. 582-597. doi: https://doi.org/10.1109/SP.2016.41

[22] Wong E., Kolter Z. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope // Proceedings of the 35 th International Conference on Machine Learning (PMLR). Vol. 80. Stockholm, Sweden, 2018. P. 5286-5295. URL: https://proceedings.mlr.press/v80/wong18a.html (дата обращения: 16.10.2022).

[23] Detection of traffic signs in real-world images: The German traffic sign detection benchmark / S. Houben [и др.] // The 2013 International Joint Conference on Neural Networks (IJCNN). Dallas, TX, USA : IEEE Computer Society, 2013. P. 1-8. doi: https://doi.org/10.1109/IJCNN.2013.6706807

[24] Шахуро В. И., Конушин А. С. Российская база изображений автодорожных знаков // Компьютерная оптика. 2016. Т. 40, № 2. С. 294-300. doi: https://doi.org/10.18287/2412-6179-2016-40-2-294-300

[25] Witchcraft: Efficient PGD Attacks with Random Step Size / P. Y. Chiang [и др.] // ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain : IEEE Computer Society, 2020. P. 3747-3751. doi: https://doi.org/10.1109/ICASSP40776.2020.9052930

[26] Xu J., Cai Z., Shen W. Using FGSM Targeted Attack to Improve the Transferability of Adversarial Example // 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE). Xi'an, China : IEEE Computer Society; 2019. p. 20-25. doi: https://doi.org/10.1109/ICECE48499.2019.9058535

Поступила 16.10.2022; одобрена после рецензирования 24.11.2022; принята к публикации 09.12.2022.

Об авторах:

Парфёнов Денис Игоревич, начальник отдела цифровых образовательных платформ, ФГБОУ ВО «Оренбургский государственный университет» (460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13), кандидат технических наук, **ORCID: https://orcid.org/0000-0002-1146-1270**, parfenovdi@mail.ru

Болодурина Ирина Павловна, заведующий кафедрой прикладной математики, ФГБОУ ВО «Оренбургский государственный университет» (460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13), доктор технических наук, профессор, **ORCID: https://orcid.org/0000-0003-0096-2587**, prmat@mail.osu.ru

Гришина Любовь Сергеевна, старший преподаватель кафедры прикладной математики, ФГБОУ ВО «Оренбургский государственный университет» (460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13), **ORCID: https://orcid.org/0000-0003-2752-7198**, zabrodina97@inbox.ru

Жигалов Артур Юрьевич, ведущий программист сектора автоматизированной поддержки организации учебного процесса, ФГБОУ ВО «Оренбургский государственный университет» (460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13), **ORCID: https://orcid.org/0000-0003-3208-1629**, leroy137.artur@gmail.com

Толмачев Сергей Владимирович, магистрант факультета математики и информационных технологий, ФГБОУ ВО «Оренбургский государственный университет» (460018, Российская Федерация, г. Оренбург, пр. Победы, д. 13), **ORCID: https://orcid.org/0000-0003-2556-3267**, tolmachev411@yandex.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Kovalev I.A., Kosov N.A. Adverse Attacks in Neural Networks. *Studencheskij vestnik* = Student Herald. 2021;(20-11):36-43. Available at: https://elibrary.ru/item.asp?id=46213350 (accessed 16.10.2022). (In Russ., abstract in Eng.)

[2] Kolesnichenko M.D. Vulnerability of artificial neural networks to adversarial attacks. In: Loginov Yu.Yu. ed. Proceedings of the International Scientific Conference on Actual problems of aviation and astronautics. Vol. 2. Krasnoyarsk: Reshetnev University; 2020. p. 228-229. Available at: https://elibrary.ru/item.asp?id=45617596 (accessed 16.10.2022). (In Russ., abstract in Eng.)

[3] Bodunkov N.E., Arefin V.V., Kobrinets S.K. [Investigation of attack methods on convolutional neural networks]. In: Proceedings of the Sixth All-Russian scientific and practical seminar "Unmanned vehicles with artificial intelligence phenomena" (BTS-II-2021). Moscow: RAAI; 2021. p. 122-125. Available at: https://elibrary.ru/item.asp?id=48235793 (accessed 16.10.2022). (In Russ.)

[4] Voinov D.M. [Investigation of the relationship between the parameters of adversarial attacks and image classification errors by neural networks]. In: Proceedings of the 76th Scientific Conference of Students and Postgraduates of the Belarusian State University. Minsk: BSU; 2019. p. 47-51. Available at: https://elibrary.ru/item.asp?id=41203721 (accessed 16.10.2022). (In Russ.)

[5] Voynov D.M., Kovalev V.A. [A comparative study of white-box and black-box adversarial attacks to the deep neural networks with different architectures]. In: *Kompyuternye tehnologii i analiz dannyh (CTDA'2020) : materialy II Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, 23-24 aprelya 2020* = Proceedings of the II International Scientific Conference on Computer Technologies and Data Analysis (CTDA'2020), Minsk, 23-24 April 2020. Minsk: BSU; 2020. p. 185-189. Available at: https://elib.bsu.by/handle/123456789/248666 (accessed 16.10.2022). (In Russ.)

[6] Chertilin K.E., Ivchenko V.D. Development and transfer training of an artificial neural network for image classification. *Instruments*. 2022;(7):37-47. Available at: https://elibrary.ru/item.asp?id=49450885 (accessed 16.10.2022). (In Russ., abstract in Eng.)

[7] Bose A.J., Aarabi P. Adversarial Attacks on Face Detectors Using Neural Net Based Constrained Optimization. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP). Vancouver, BC, Canada: IEEE Computer Society; 2018. p. 1-6. doi: https://doi.org/10.1109/MMSP.2018.8547128

[8] Guo Ch., Sablayrolles A., Jégou H., Kiela D. Gradient-based Adversarial Attacks against Text Transformers. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 5747-5757. doi: https://doi.org/10.18653/v1/2021.emnlp-main.464

[9] Carlini N., Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE Computer Society; 2017. p. 39-57. doi: https://doi.org/10.1109/SP.2017.49

[10] Xiao C., Li B., Zhu J.-Y., He W., Liu M., Song D. Generating adversarial examples with adversarial networks. In: Lang J. ed. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018). Stockholm, Sweden: IJCAI; 2018. p. 3905-3911. doi: https://doi.org/10.24963/ijcai.2018/543

[11] Moosavi-Dezfooli S.-M., Fawzi A., Fawzi O., Frossard P. Universal Adversarial Perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE Computer Society; 2017. p. 86-94. doi: https://doi.org/10.1109/CVPR.2017.17

[12] Poursaeed O., Katsman I., Gao B., Belongie S. Generative Adversarial Perturbations. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE Computer Society; 2018. p. 4422-4431. doi: https://doi.org/10.1109/CVPR.2018.00465

[13] Irfan M.M., Ali S., Yaqoob I., Zafar N. Towards Deep Learning: A Review On Adversarial Attacks. In: 2021 International Conference on Artificial Intelligence (ICAI). Islamabad, Pakistan: IEEE Computer Society; 2021. p. 91-96. doi: https://doi.org/10.1109/ICAI52203.2021.9445247

[14] Rauber J., Zimmermann R., Bethge M., Brendel W. Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. *Journal of Open Source Software*. 2020;5(53):2607. doi: https://doi.org/10.21105/joss.02607

[15] Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, BC, Canada: Vancouver Convention Center; 2018. p. 1-23. Available at: https://openreview.net/pdf?id=rJzIBfZAb (accessed 16.10.2022).

[16] Wang K., Li F., Chen C.-M., Hassan M.M., Long J., Kumar N. Interpreting Adversarial Examples and Robustness for Deep Learning-Based Auto-Driving Systems. *IEEE Transactions on Intelligent Transportation Systems*. 2022;23(7):9755-9764. doi: https://doi.org/10.1109/TITS.2021.3108520

[17] Mani N., Moh M., Moh T.S. Defending Deep Learning Models Against Adversarial Attacks. *International Journal of Software Science and Computational Intelligence (IJSSCI)*. 2021;13(1):72-89. doi: http://doi.org/10.4018/IJSSCI.2021010105

[18] Amirian M., Tuggener L., Chavarriaga R., Satyawan Y., Schilling F., Schwenker F., Stadelmann T. Two to Trust: AutoML for Safe Modelling and Interpretable Deep Learning for Robustness. In: Heintz F., Milano M., O'Sullivan B. (Eds.) Trustworthy AI – Integrating Learning, Optimization and Reasoning. TAILOR 2020. Lecture Notes in Computer Science. Vol. 12641. Cham: Springer; 2021. p. 268-275. doi: https://doi.org/10.1007/978-3-030-73959-1_23

[19] Qiu Y., Zhang J., Zhou J. Improving Gradient-based Adversarial Training for Text Classification by Contrastive Learning and Auto-Encoder. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics; 2021. p. 1698-1707. doi: https://doi.org/10.18653/v1/2021.findings-acl.148

[20] Huang X., Kwiatkowska M., Wang S., Wu M. Safety Verification of Deep Neural Networks. In: Majumdar R., Kunčak V. (Eds.) Computer Aided Verification. CAV 2017. Lecture Notes in Computer Science. Vol. 10426. Cham: Springer; 2017. p. 3-29. doi: https://doi.org/10.1007/978-3-319-63387-9_1

[21] Papernot N., McDaniel P., Wu X., Jha S., Swami A. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In: 2016 IEEE Symposium on Security and Privacy (SP). San Jose, CA, USA: IEEE Computer Society; 2016. p. 582-597. doi: https://doi.org/10.1109/SP.2016.41

[22] Wong E., Kolter Z. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In: Proceedings of the 35 th International Conference on Machine Learning (PMLR). Vol. 80. Stockholm, Sweden; 2018. p. 5286-5295. Available at: https://proceedings.mlr.press/v80/wong18a.html (accessed 16.10.2022).

[23] Houben S., Stallkamp J., Salmen J., Schlipsing M., Igel C. Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In: The 2013 International Joint Conference on Neural Networks (IJCNN). Dallas, TX, USA: IEEE Computer Society; 2013. p. 1-8. doi: https://doi.org/10.1109/IJCNN.2013.6706807

[24] Shakhuro V.I., Konushin A.S. Russian traffic sign images dataset. *Computer Optics*. 2016;40(2):294-300. (In Russ., abstract in Eng.) doi: https://doi.org/10.18287/2412-6179-2016-40-2-294-300

[25] Chiang P.Y., et. al. Witchcraft: Efficient PGD Attacks with Random Step Size. In: ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE Computer Society; 2020. p. 3747-3751. doi: https://doi.org/10.1109/ICASSP40776.2020.9052930

[26] Xu J., Cai Z., Shen W. Using FGSM Targeted Attack to Improve the Transferability of Adversarial Example. In: 2019 IEEE 2nd International Conference on Electronics and Communication Engineering (ICECE). Xi'an, China: IEEE Computer Society; 2019. p. 20-25. doi: https://doi.org/10.1109/ICECE48499.2019.9058535

Submitted 16.10.2022; approved after reviewing 24.11.2022; accepted for publication 09.12.2022.

About the authors:

Denis I. Parfenov, Head of Digital Educational Platforms Department, Orenburg State University (13 Prospect Pobedy, Orenburg 460018, Russian Federation), Cand. Sci. (Eng.), **ORCID: https://orcid.org/0000-0002-1146-1270**, parfenovdi@mail.ru

Irina P. Bolodurina, Head of the Department of Applied Mathematics, Orenburg State University (13 Prospect Pobedy, Orenburg 460018, Russian Federation), Dr. Sci. (Eng.), Professor, **ORCID: https://orcid.org/0000-0003-0096-2587**, prmat@mail.osu.ru

Lyubov S. Grishina, Lecturer of Department Applied Mathematics, Orenburg State University (13 Prospect Pobedy, Orenburg 460018, Russian Federation), **ORCID: https://orcid.org/0000-0003-2752-7198**, zabrodina97@inbox.ru

Artur Yu. Zhigalov, Software Developer of the Sector of Automated Support for the Organization of the Educational Process, Orenburg State University (13 Prospect Pobedy, Orenburg 460018, Russian Federation), **ORCID: https://orcid.org/0000-0003-3208-1629**, leroy137.artur@gmail.com

Sergey V. Tolmachev, Master degree student, Orenburg State University (13 Prospect Pobedy, Orenburg 460018, Russian Federation), **ORCID: https://orcid.org/0000-0003-2556-3267**, tolmachev411@yandex.ru

All authors have read and approved the final manuscript.