Иванов Ю. С., Горькавый М. А., Грабарь Д. М.
Yu. S. Ivanov, M. A. Gorkavyy, D. M. Grabar

АНАЛИЗ УСТОЙЧИВОСТИ ПРЕДИКТИВНЫХ МОДЕЛЕЙ К СОСТЯЗАТЕЛЬНЫМ АТАКАМ В РОБОТОТЕХНИЧЕСКИХ КОМПЛЕКСАХ

ANALYSIS OF PREDICTIVE MODELS STABILITY TO ADVERSARIAL ATTACKS IN ROBOTICS COMPLEXES

Иванов Юрий Сергеевич – кандидат технических наук, доцент кафедры «Промышленная электроника» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре); 681013, Хабаровский край, г. Комсомольск-на-Амуре, ул. Ленина, д. 27. E-mail: ivanov_ys@icloud.com.
Yurii S. Ivanov – PhD in Engineering, Associate Professor, Industrial Electronics Department, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur); 681013, Khabarovsk territory, Komsomolsk-on-Amur, 27 Lenin str. E-mail: ivanov_ys@icloud.com.

Горькавый Михаил Александрович – кандидат технических наук, заведующий кафедрой «Управление инновационными процессами и проектами» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: mixkomsa@gmail.com.
Mikhail A. Gorkavyy – PhD in Engineering, Head of Management of Innovative Processes and Projects Department, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: mixkomsa@gmail.com.

Грабарь Даниил Михайлович – магистр кафедры «Управление инновационными процессами и проектами» Комсомольского-на-Амуре государственного университета (Россия, Комсомольск-на-Амуре). E-mail: grabardm@ml-dev.ru.
Daniil M. Grabar – Master's Degree Student, Management of Innovative Processes and Projects Department, Komsomolsk-na-Amure State University (Russia, Komsomolsk-on-Amur). E-mail: grabardm@ml-dev.ru.

Аннотация. В статье проведён анализ эффективности современных моделей для распознавания объектов в системах технического зрения роботизированных комплексов. Приведён обзор технологий состязательных атак на предиктивные модели. Проведены эксперименты по реализации существующих атак на различные модели. Подготовлен сравнительный анализ киберустойчивости различных наиболее часто используемых моделей в действующих системах к деструктивным информационным воздействиям.

Summary. The paper analyzes the effectiveness of modern models for object recognition in vision systems of robotic complexes. A review of technologies of adversarial attacks on predictive models is given. Experiments on the implementation of existing attacks on different models have been conducted. A comparative analysis of the cyber resistance of various most commonly used models in existing systems to destructive information influences has been prepared.

Ключевые слова: распознавание объектов, глубокие нейронные сети, компьютерное зрение, FGSM, безопасность, состязательные примеры.

Key words: object recognition, deep neural network, computer vision, FGSM, security, adversarial examples.

Работа выполнена при поддержке Российского научного фонда проект №22-71-100093 «Разработка и синтез перспективных мультимодальных адаптивныхалгоритмов и методов управления поведением коллаборативных робототехнических систем с учётом нештатных ситуаций и экстремальных условий в недетерминированной среде» https://rscf.ru/project/22-71-10093/.

УДК 004.8

Введение. Компьютерное зрение используется во многих приложениях робототехники для обнаружения людей, нештатных ситуаций, навигации и ориентирования в пространстве [1; 2].

Распознавание объектов для коллаборативной робототехники также является важной задачей, решаемой специалистами по машинному обучению.

Важно учитывать, что коботы предназначены для совместной работы с людьми, а значит у них нет «права на ошибку». Любая ошибка системы искусственного интеллекта (ИИ), может привести к травме человека-оператора, работающего в одном рабочем пространстве с коботом.

Одной из актуальных задач коллаборативной робототехники является задача обнаружения и распознавания инструментов и изделий. Коботу необходимо учитывать изменчивость освещения, неопределённость расположения и многообразие подвидов инструментов. Как правило, с данной задачей успешно справляются алгоритмы, основанные на применении глубокого обучения.

В лаборатории [3] исследователи изучали механику работы кобота с различными типами инструментов.

В работе [4] авторы разработали алгоритм распознавания движений человека и определения используемого инструмента в руке человека. Стоит отметить, что алгоритм показал высокую точность при тестировании в реальной производственной задаче.

Исследователи [5] предложили новый набор объектов для распознавания роботом, а также метод адаптации алгоритмов под условия съёмки.

Ранее авторами [6] предлагался подход к коррекции изображений с целью повышения точности распознавания целевых инструментов.

В работе [7] предлагается перспективный метод автоматического распознавания инструментов и их использование без предварительного обучения, тем самым реализуя one-shot или few-shot обучение.

Таким образом, методы глубокого обучения позволили достичь значительных успехов при построении коллаборативных робототехнических комплексов. Однако, как показывают многочисленные исследования, системы предиктивной аналитики всё чаще подвергаются атакам [8]. Незначительные и незаметные изменения входных изображений достаточны для того, чтобы обмануть большинство нейросетевых подходов.

В работе [9] проводится всесторонний обзор состязательных атак и защиты в области компьютерного зрения.

Применительно к задачам робототехники состязательные атаки рассматриваются в работе [10]. Авторами демонстрируются примеры атак на системы оценки позы.

Для обнаружения состязательных атак применяются различные методы обнаружения аномалий, в том числе показавшие свою эффективность в других областях информационной безопасности [11].

В данной работе решается задача распознавания инструментов в рабочей зоне кобота с использованием предобученных наиболее часто используемых глубоких моделей. Проведено исследование киберустойчивости моделей компьютерного зрения к наиболее популярным видам состязательных атак. Предложены методы повышения надёжности моделей.

Постановка задачи. Необходимо обучить наиболее современные модели для визуальной классификации изображений и протестировать их по метрикам точности, скорости и размеру модели. Математическая постановка задачи распознавания объектов в кадрах видеопотока приведена в работах [1; 12; 13].

Пусть имеются: множество образов $\omega \in \Omega$, заданных признаками $x_i, i = \overline{1,n}$, совокупность которых для образа ω представлена векторными описаниями $\Phi(\omega) = \big(x_1(\omega), x_2(\omega), \dots, x_n(\omega)\big) = \mathbf{x}$; множество классов $\mathbb{B} = \{\beta_1, \beta_2, \dots, \beta_c\}$, где c– количество классов. Априорная информация представлена обучающим множеством (датасетом) $\mathbb{D} = \{(\mathbf{x}^j, \beta^j)\}$, $j = \overline{1,L}$, заданным таблицей, каждая строка j которой содержит векторное описание образа $\Phi(\omega)$ и метку класса β_k, $k = \overline{1,c}$. Заметим, что обучающее множество характеризует неизвестное отображение $^*\mathbf{F}: \Omega \longrightarrow \mathbb{B}$.

Требуется по имеющимся кадрам $\mathbf{I}_t$ непрерывного видеопотока $\mathbf{V} = (\mathbf{I}_1 .., \mathbf{I}_t, \dots, \mathbf{I}_\tau)$ решить задачу распознавания образов: обнаружить образы ω в виде оценки признаков $\tilde{\mathbf{x}}$ с помощью отображения $\mathbf{F_1}: \mathbf{I}_t \longrightarrow \tilde{\mathbf{x}}$ и классифицировать их с использованием отображения $\mathbf{F_2}: \tilde{\mathbf{x}} \longrightarrow \beta_k$, $k = \overline{1,c}$ в соответствии с заданным критерием $\mathrm{P}(\tilde{\mathbf{x}})$, минимизирующим вероятность ошибки.

Таким образом, необходимо найти отображение $\mathbf{F_1}: \mathbf{I}_t \longrightarrow \beta_k$, при котором $\mathbf{F}$ является набором функций и алгоритмов $\mathbf{f}_i$, $i = \overline{1, N_f}$.

В качестве критерия эффективности будут использоваться общепринятые метрики для оценки качества классификатора: Accuracy, Precision, Recall, *F*-мера.

Целью состязательных атак на изображения является генерация нового изображения путём небольшого изменения исходного. Изменение происходит таким образом, чтобы максимизировать функционал ошибки модели машинного обучения.

Задача атаки с использованием состязательных примеров может быть сформулирована следующим образом. Для каждого $\mathbf{I}_t$, принадлежащего некоторому классу β_k, необходимо определить такую функцию $\boldsymbol{G}$, аддитивно изменяющую оригинальное изображение с заданным коэффициентом ε, таким образом реализуя отображение $\boldsymbol{G}: \mathbf{I}_t \longrightarrow {}^{G}\mathbf{I}_t$, где ${}^{G}\mathbf{I}_t$ – сгенерированное изображение ${}^{G}\mathbf{I}_t = \mathbf{I}_t + \varepsilon$.

Атака будет успешной при выполнении следующего условия: выполняется ошибочная классификация при $\mathbf{F}(\mathbf{I}_t) \neq \mathbf{F}({}^{G}\mathbf{I}_t)$ минимальной ε, гарантирующей визуальную неотличимость изображения $\mathbf{I}_t$ от ${}^{G}\mathbf{I}_t$.

Обучение классификаторов для распознавания. Для обучения и тестирования классификаторов использовался датасет KTH handtool (см. рис. 1), содержащий 5559 изображений 3 классов инструментов: молоток, плоскогубцы и отвёртка – в разном освещении и в разных местах. Каждый из классов разбивается на подклассы.

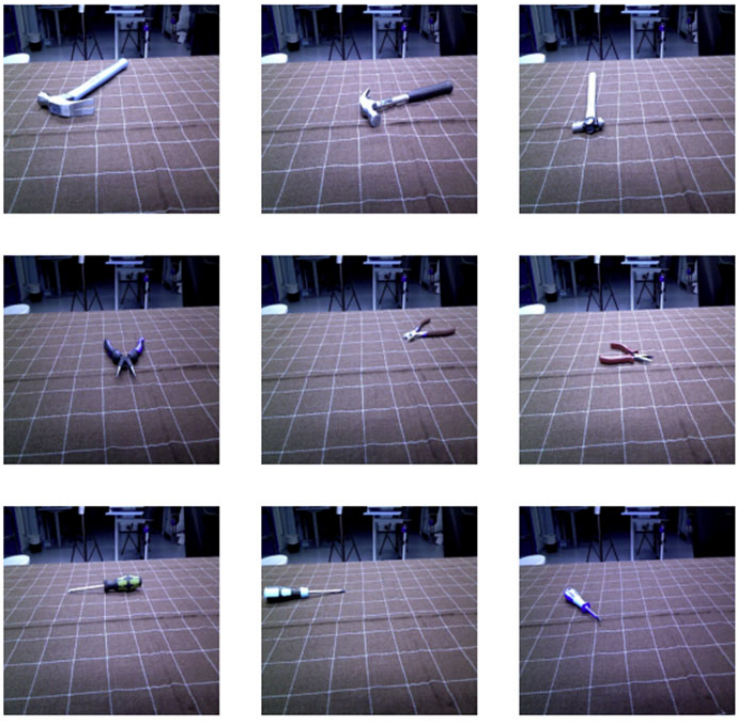

Рис. 1. Примеры изображений из датасета KTH handtool

В качестве базовых моделей используются следующие: MobileNetV3, EfficientNetB0, EfficientNetB3, EfficientNetV2. Архитектуры MobileNet и GhostNet подробно описаны в работах [12; 14].

Дальнейшее развитие ИИ послужило к разработке метода Neural architecture search (NAS). NAS использовался для проектирования сетей, которые соответствуют или превосходят по производительности созданные вручную архитектуры.

EfficientNet – класс новых моделей, который получился при изучении масштабирования (скейлинг, scaling) моделей и балансирования между собой глубины и ширины (количества кана-

лов) сети, а также разрешения изображений в сети. Авторы статьи [15] предлагают новый метод составного масштабирования (compound scaling method), который равномерно масштабирует глубину/ширину/разрешение с фиксированными пропорциями между ними (см. рис. 2).

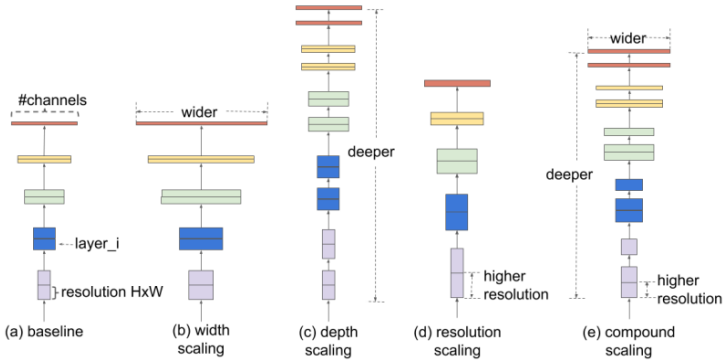

Рис. 2. Масштабирование EfficientNet

Чтобы улучшить работу нейросети, исследователи [16] выбирали начальную архитектуру (см. рис. 3) автоматически с помощью методов AutoML. Так были построены EfficientNet-B1 – EfficientNet-B7 с целым числом в конце имени, указывающим значение составного коэффициента.

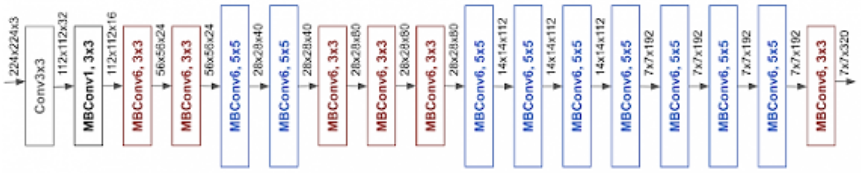

Рис. 3. Архитектура EfficientNet

EfficientNetV2 [17] (см. рис. 4) имеет более высокую скорость обучения и лучшую эффективность параметров, чем предыдущие модели данного класса EfficientNet. Данная архитектура имеет ряд важных отличий по сравнению с предыдущим поколением:

– используется как MBConv слой, так и недавно добавленный слой fused-MBConv на ранних уровнях;

– используется меньший коэффициент расширения для MBConv, что позволяет снизить накладные расходы на доступ к памяти.

С использованием тестирующей выборки, полученной из датасета, был проведён эксперимент на оборудовании со следующими параметрами: ЦПУ Intel Core i7-5820K, ГПУ 1080 Ti. Проведено сравнение эффективности описанных нейронных сетей с точки зрения стандартных метрик качества классификации.

Stage	Operator	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv1x1 & Pooling & FC	-	1280	1

Рис. 4. Архитектура EfficientNetV2-S

Атаки на обученные модели. Как правило, выделяют следующие виды состязательных атак [18]: FGSM, BIM, DeepFool, JSMA, CW, PGD. При этом FGSM (fast gradient sign method) является простым методом, который делает один шаг в направлении градиента:

$$\mathbf{I}_t = \mathbf{I}_t + \varepsilon * \text{sign}\left(\nabla_k J(\mathbf{I}_t, \beta_k)\right),$$

где J – функция потерь; ε – множитель для обеспечения малых возмущений.

Таблица 1

Сравнительный анализ работы алгоритмов

Модель	Метрики			
	Prec.	Recall	f1-score	Acc.
MobileNetV3	0.974	0.972	0.973	0.973
EfficientNetB0	0.9973	0.997	0.997	0.997
EfficientNetB3	0.960	0.956	0.957	0.958
EfficientNetV2	0.996	0.996	0.996	0.996

Фактически для генерации состязательного примера мы прибавляем шумовую карту к исходному изображению с некоторым ε (см. рис. 5).

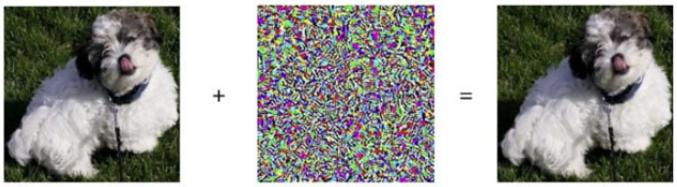

Рис. 5. Генерация изображения с использованием карты шумов

Для эксперимента были получены состязательные изображения из KTH handtool путём добавления шумов с порогами [0, .05, .1, .15, .2, .25]. На рис. 6 приведены примеры сгенерированных изображений.

Каждая из обученных моделей была подвергнута FGSM-атаке. Результаты эксперимента приведены в табл. 2.

Вывод по результатам моделирования. Как мы видим, даже небольшие целенаправленные шумы способны сделать SOTA-модель бесполезной. Стоит учитывать, что FGSM-атака требует наличия доступа к самой модели, однако существуют методы, способные сгенерировать состязательные примеры более эффективно. Большинство из используемых моделей для задач распознавания образов основано на сверточных операциях. Однако в статье [19] представлено использование трансформеров для классификации изображений. Одним из перспективных направлений дальнейшей работы является исследование устойчивости трансформеров к различным видам состязательных атак.

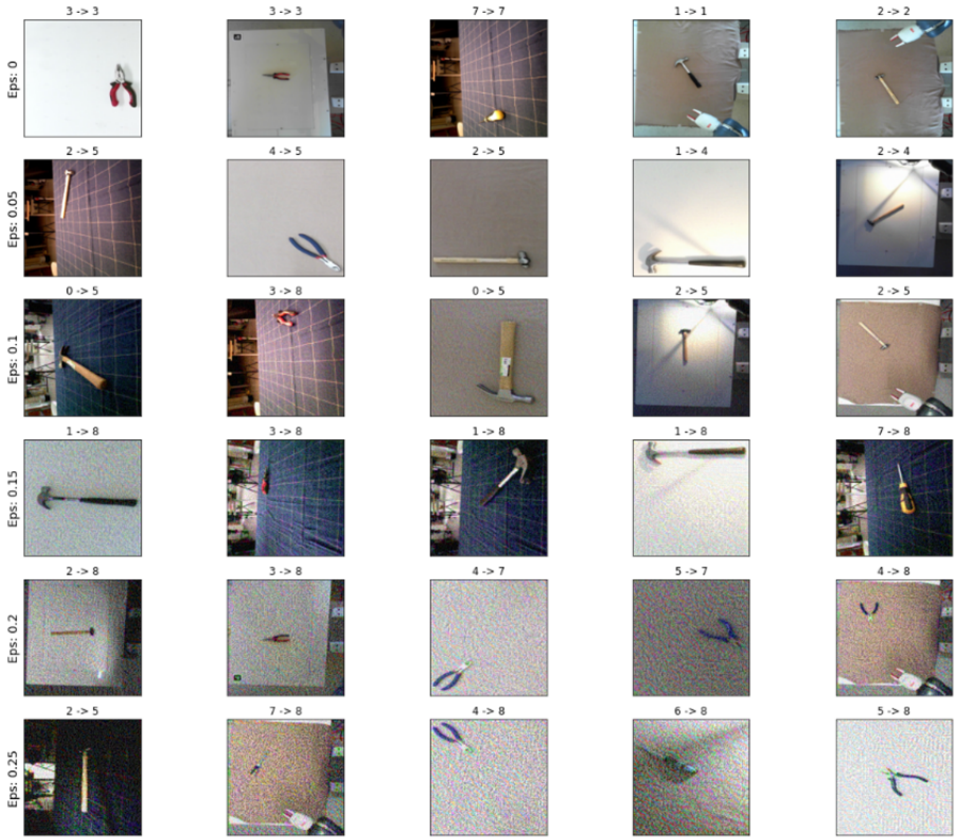

Рис. 6. Сгенерированные изображения KRH handtool путём добавления шумов

Таблица 2

Результаты эксперимента FGSM-атаки

Порог	Метрики	Модели			
		MobileNetV3	EfficientNetB0	EfficientNetB3	EfficientNetV2
0	Acc.	0.973	0.997	0.958	0.996
	f1	0.973	0.997	0.957	0.996
0.01	Acc.	0.272	0.532	0.423	0.343
	f1	0.265	0.543	0.417	0.354
0.05	Acc.	0.186	0.228	0.183	0.193
	f1	0.124	0.193	0.137	0.197
0.1	Acc.	0.127	0.192	0.151	0.163
	f1	0.041	0.136	0.089	0.174
0.15	Acc.	0.125	0.175	0.131	0.167
	f1	0.042	0.111	0.091	0.169
0.2	Acc.	0.128	0.157	0.102	0.149
	f1	0.042	0.091	0.054	0.129

Заключение. Приведён сравнительный анализ эффективности популярных нейросетевых архитектур распознавания изображений для задачи классификации инструмента. Представлены и проанализированы результаты исследований устойчивости разноплановых моделей к состязательным атакам. Полученные в ходе данного исследования результаты и выводы позволяют уточнить разделы информационной безопасности, связанные с использованием моделей для задач промышленной робототехники.

ЛИТЕРАТУРА

1. Амосов, О. С. Вычислительный метод распознавания ситуаций и объектов в кадрах непрерывного видеопотока с использованием глубоких нейронных сетей для систем контроля и управления доступом / О. С. Амосов, С. Г. Амосов, С. В. Жиганов // Известия Российской академии наук. Теория и системы управления. – 2020. – № 5. – С. 73-88.
2. Амосов, О. С. Локализация человека в кадре видеопотока с использованием алгоритма на основе растущего нейронного газа и нечёткого вывода / О. С. Амосов, Ю. С. Иванов, С. В. Жиганов // Компьютерная оптика. – 2017. – № 1. – С. 46-58.
3. MCube Lab Manipulation and Mechanisms Laboratory at MIT // The MCube Lab – Tool Use Dataset: [сайт]. – URL: https://mcube.mit.edu/tool-use/ (дата обращения: 25.12.2022). – Текст: электронный.
4. Штехин, С. Е. Разработка алгоритма распознавания движений человека методами компьютерного зрения в задаче нормирования рабочего времени / С. Е. Штехин, Ю. К. Иванова // Труды ИСП РАН. – 2020. – № 32. – С. 121-136.
5. Mancini M. et al. Kitting in the wild through online domain adaptation // 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – IEEE, 2018. – P. 1103-1109.
6. Cheng, L., Target-tools recognition method based on an image feature library for space station cabin service robots / L. Cheng, Z. Jiang, H. Li, B. Wei, Q. Huang // Robotica. – 2016. – № 34. – P. 925-941.
7. Tee K. P. Towards Emergence of Tool Use in Robots: Automatic Tool Recognition and Use Without Prior Tool Learning / K. P. Tee, J. Li, L. T. Pang Chen, K. W. Wan and G. Ganesh // 2018 IEEE International Conference on Robotics and Automation (ICRA). – 2018. – № 1. – P. 6439-6446.
8. Chukhnov A. P. Algorithms for Detecting and Preventing Attacks on Machine Learning Models in Cyber-Security Problems / A. P. Chukhnov, Y. S Ivanov // Journal of Physics: Conference Series. – 2021. – № 1. – P. 2096.
9. Akhtar N. et al. Advances in adversarial attacks and defenses in computer vision: A survey // IEEE Access. – 2021. – Т. 9. – P. 155161-155196.
10. Chawla H. et al. Adversarial attacks on monocular pose estimation // 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). – IEEE, 2022. – P. 12500-12505.
11. Амосов, О. С. Сетевая классификация атак в задачах информационной безопасности на основе интеллектуальных технологий, фрактального и вейвлет-анализа / О. С. Амосов, Д. С. Магола, С. Г. Баена // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2017. – № IV-1 (32). – С. 19-29.
12. Ivanov Y. S. Intelligent Deep Neuro-Fuzzy System of Abnormal Situation Recognition for Transport Systems / Y. S. Ivanov, S. V. Zhiganov, T. I. Ivanova // Current Problems and Ways of Industry Development: Equipment and Technologies. – 2021. – № 1. – P. 224-233.
13. Amosov O. S. Recognition of Abnormal Traffic Using Deep Neural Networks and Fuzzy Logic / O. S. Amosov, Y. S. Ivanov and S. G. Amosova // Vladivostok: 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), 2019. – P. 1-5.
14. Ivanov Y. S. Comparative Analysis of Deep Neural Networks Architectures for Visual Recognition in the Autonomous Transport Systems / Y. S. Ivanov, S. V. Zhiganov, N. N. Liubushkina // Journal of Physics: Conference Series. – 2021. – № 1. – P. 2096.
15. Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks // International conference on machine learning. – PMLR, 2019. – P. 6105-6114.
16. Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks // International conference on machine learning. – PMLR, 2019. – P. 6105-6114.
17. Tan M., Le Q. Efficientnetv2: Smaller models and faster training // International conference on machine learning. – PMLR, 2021. – P. 10096-10106.
18. Behnia F. et al. Code-bridged classifier (cbc): A low or negative overhead defense for making a cnn classifier robust against adversarial attacks // 2020 21st International Symposium on Quality Electronic Design (ISQED). – IEEE, 2020. – P. 27-32.
19. Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale // arXiv preprint arXiv: 2010. 11929. – 2020.