НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2025 Том 25 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
Jule–August 2025 Vol. 25 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-4-694-702
УДК 004.89

Исследование влияния состязательных атак на классификацию и кластеризацию изображений на примере модели ResNet50

Роман Ростиславович Болозовский[1], Алла Борисовна Левина[2]✉, Константин Сергеевич Красов[3]

[1,2,3] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация

[1] bolozovskii@gmail.com, https://orcid.org/0009-0004-5725-0526
[2] Alla_levina@mail.ru✉, https://orcid.org/0000-0003-4421-2411
[3] iflup@ya.ru, https://orcid.org/0009-0005-9232-5463

Аннотация

Введение. Прогресс в области компьютерного зрения привел к созданию мощных моделей, способных точно распознавать и интерпретировать визуальную информацию в различных областях знаний. На этом фоне растет уязвимость таких моделей к состязательным атакам — преднамеренному манипулированию входными данными с целью исказить модель машинного обучения и привести к неверным результатам распознавания. В работе приведены результаты исследования влияния различных типов состязательных атак на модель ResNet50 в задачах классификации и кластеризации изображений. **Метод.** Исследованы следующие типы состязательных атак: метод быстрого градиентного знака, базовый итерационный метод, метод проецируемого градиентного спуска, метод Карлини и Вагнера, состязательная атака с использованием Elastic-Net, Expectation Over Transformation Predicted Gradient Descent, атаки на основе джиттера. Для визуализации областей внимания модели применен метод Gradient-Weighted Class Activation Mapping (Grad-CAM). Для визуализации кластеров в пространстве признаков использован алгоритм t-SNE. Устойчивость к атакам оценивалась по показателям успешности атак с использованием алгоритмов *k*-ближайших соседей иерархического маленького мира с различными метриками сходства. **Основные результаты.** Выявлены существенные различия в воздействии атак на внутренние представления модели и области фокусировки внимания. Показано, что итеративные методы атак вызывают значительные изменения в пространстве признаков и заметно влияют на визуализации Grad-CAM, тогда как простые атаки оказывают меньшее воздействие. Установлена высокая чувствительность большинства алгоритмов кластеризации к возмущениям. Наибольшую устойчивость среди исследованных подходов показала метрика внутреннего произведения. **Обсуждение.** Полученные результаты указывают на зависимость устойчивости модели от параметров атак и выбора метрик сходства, что проявляется в особенностях формирования кластерных структур. Выявленные закономерности в перераспределении пространства признаков в условиях целенаправленных атак открывают перспективы для дальнейшей оптимизации алгоритмов кластеризации, способных обеспечить более высокую степень защиты систем компьютерного зрения.

Ключевые слова

состязательные атаки, компьютерное зрение, ResNet50, кластеризация изображений, KNN, HNSW

Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации № 075-00003-24-02 от 08.02.2024 (проект FSEE-2024-0003).

Ссылка для цитирования: Болозовский Р.Р., Левина А.Б., Красов К.С. Исследование влияния состязательных атак на классификацию и кластеризацию изображений на примере модели ResNet50 // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 4. С. 694–702. doi: 10.17586/2226-1494-2025-25-4-694-702

The impact of adversarial attacks on a computer vision models perception of images

Roman R. Bolozovskii[1], Alla B. Levina[2]✉, Konstantin S. Krasov[3]

[1,2,3] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation

[1] bolozovskii@gmail.com, https://orcid.org/0009-0004-5725-0526
[2] Alla_levina@mail.ru✉, https://orcid.org/0000-0003-4421-2411
[3] iflup@ya.ru, https://orcid.org/0009-0005-9232-5463

Abstract

Advances in computer vision have led to the development of powerful models capable of accurately recognizing and interpreting visual information in various fields of knowledge. However, these models are increasingly vulnerable to adversarial attacks – deliberate manipulations of input data designed to mislead the machine-learning model and produce incorrect recognition results. This article presents the results of an investigation into the impact of various types of adversarial attacks on the ResNet50 model in image classification and clustering tasks. Various types of adversarial attacks have been investigated: Fast Gradient Sign Method, Basic Iterative Method, Projected Gradient Descent, Carlini&Wagner, Elastic-Net Attacks to Deep Neural Networks, Expectation Over Transformation Projected Gradient Descent, and jitter-based attacks. The Gradient-Weighted Class Activation Mapping (Grad-CAM) method was used to visualize the attention areas of the model. The t-SNE algorithm was applied to visualize clusters in the feature space. Attack robustness was assessed by attack success rate using *k*-Nearest Neighbors algorithm and Hierarchical Navigable Small World algorithms with different similarity metrics. Significant differences in the effects of attacks on the internal representations of the model and areas of focus have been identified. It is shown that iterative attack methods cause significant changes in the feature space and significantly affect Grad-CAM visualizations, whereas simple attacks have less impact. The high sensitivity of most clustering algorithms to perturbations has been established. The metric of the inner product showed the greatest stability among the studied approaches. The results obtained indicate the dependence of the stability of the model on the attack parameters and the choice of similarity metrics, which is manifested in the peculiarities of the formation of cluster structures. The observed feature-space redistributions under targeted attacks suggest avenues for further optimizing clustering algorithms to enhance the resilience of computer-vision systems.

Keywords

adversarial attacks, computer vision, ResNet50, image clustering, KNN, HNSW

Acknowledgements

The work was performed within the framework of the State Assignment of the Ministry of Science and Higher Education of the Russian Federation No. 075-00003-24-02 dated 08.02.2024 (FSEE-2024-0003 project).

For citation: Bolozovskii R.R., Levina A.B., Krasov K.S. The impact of adversarial attacks on a computer vision models perception of images. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 694–702 (in Russian). doi: 10.17586/2226-1494-2025-25-4-694-702

Введение

За последние годы в области компьютерного зрения произошел значительный прогресс, приведший к созданию мощных моделей, способных точно распознавать и интерпретировать визуальную информацию. Эти модели нашли применение в различных областях, начиная от автономного вождения транспортными средствами и визуализации медицинских данных, заканчивая системами безопасности и наблюдения. Однако на этом фоне растет обеспокоенность по поводу уязвимости таких моделей к состязательным атакам.

Состязательные атаки — преднамеренное манипулирование входными данными с целью обмануть модели машинного обучения и заставить их выдавать неверные результаты. Эти атаки создают значительную угрозу для надежности и безопасности систем компьютерного зрения, потенциально приводя к ошибочным решениям с серьезными последствиями. Понимание влияния состязательных атак на восприятие изображений моделями компьютерного зрения является актуальной областью исследований [1].

История состязательных атак начинается с работы [2], где впервые было продемонстрировано существование незаметных изменений, способных заставить глубокие нейронные сети неверно классифицировать изображения. Исследования в работах [3, 4] существенно продвинули методы защиты, сосредоточив внимание на устойчивости моделей, решающих задачи классификации. В [5, 6] были усовершенствованы алгоритмы защиты, однако вопросы влияния состязательных атак на внутренние представления нейронных сетей остаются менее изученными. При этом кластеризация играет ключевую роль в задачах поиска и сопоставления изображений, где устойчивость к атакам критически важна. Изменение пространственного распределения признаков под воздействием атак может не только повлиять на точность классификации, но и изменить структуру кластеров, что влечет за собой искажение результатов анализа данных.

Анализ научных работ по данной тематике показывает, что большинство исследований сосредоточено на оценке того, как состязательные атаки влияют на точность классификаторов и детектирование объектов. Однако изменения метрик сходства между признаковыми представлениями изображений изучены менее подробно. В данной работе предлагается комплексный подход, объединяющий визуализацию и кластеризацию признакового пространства, что дает возможность исследовать, как состязательные атаки трансформируют

структуру представлений. Таким образом, данное исследование нацелено на выявление изменений, вносимых состязательными атаками в работу алгоритмов кластеризации с различными метриками сходства, и на анализ трансформаций, происходящих в нейронной сети при обработке изображений.

Несмотря на появление новых архитектур, ResNet50 [7] по-прежнему используется в исследованиях состязательных атак (например, [8]). Компактность ее архитектуры и наличие предобученных весов делают ResNet50 удобным базовым тестовым стендом для анализа поведения модели при атаках.

Предложенный комплексный подход включает несколько этапов анализа влияния состязательных атак на модель ResNet50. Применены методы различных состязательных атак для генерации состязательных примеров и оценки их влияния на классификацию. Внимание модели визуализировалось с помощью Gradient-Weighted Class Activation Mapping (Grad-CAM) для выявления областей изображения, которые становятся более или менее значимыми под воздействием возмущений. Несмотря на то, что влияние атак на интерпретационные механизмы ранее изучалось [9], использование Grad-CAM остается наиболее простым и наглядным способом демонстрации изменений в восприятии модели. Исходные изображения и их атакованные версии преобразовывались в признаковое пространство ResNet50, после чего применялся алгоритм t-SNE для визуализации кластеров в пониженной размерности. Кроме того, анализировались скалярные произведения векторов признаков изображений для оценки изменений в метрике сходства. И, наконец, рассчитывался коэффициент успешности атак (Attack Success Rate, ASR) при помощи алгоритмов *k*-ближайших соседей (K-Nearest Neighbors, KNN) и иерархического маленького мира (Hierarchical Navigable Small World, HNSW), что позволило количественно оценить влияние возмущений на точность кластеризации и стабильность кластеров. Такой подход позволяет проанализировать, каким образом возмущения изменяют внутренние представления модели и влияют на результаты кластеризации и распределение признаков.

Целью данного исследования является анализ влияния различных типов состязательных атак на восприятие изображений нейросетевыми моделями и оценка их устойчивости в задачах классификации и кластеризации.

Методика проведения исследования

Исследуемые атаки. Рассмотрим ряд наиболее актуальных состязательных атак.

Метод быстрого градиентного знака (Fast Gradient Sign Method, FGSM) [2] можно записать в виде:

$$\eta = \varepsilon \text{sign}(\nabla_x L(\theta, x, y)),$$

где η — состязательный пример; ε — порог искажения; $L(\theta, x, y)$ — функция потерь; θ — параметры модели; x — исходное изображение; y — распределение по всем возможным классам.

Метод проецируемого градиентного спуска (Projected Gradient Descent, PGD) [3] возможно описать уравнением:

$$x^{t+1} = \prod_{x+S}(x^t + \alpha \text{sign}(\nabla_x L(\theta, x, y))),$$

где t — номер итерации; x^t — состязательный пример на итерации t; S — область допустимых изменений; α — уровень «шума», добавляемый к исходному примеру.

Атака Карлини и Вагнера (Carlini&Wagner, CW) [4] работает путем оптимизации следующего выражения и поиска w:

$$\text{minimize} \| 0{,}5(\tanh(w) + 1 - x \|_2^2 + cf(0{,}5(\tanh(w) + 1)),$$

где c — константа, отвечающая за относительную важность каждого из слагаемых в оптимизируемой функции; w — оптимизирующая переменная; f — функция потерь.

При заданном x, целевом классе t, выходных данных модели $Z(x)$ и f, определяемым, как:

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_{t'} - \kappa,$$

где x' — целевой состязательный пример; κ — порог достоверности для успешной атаки.

Базовый итерационный метод (Basic Iterative Method, BIM) [10] описывается с помощью:

$$X_0^{adv} = X, X_{N+1}^{adv} = \text{Clip}_{X,\varepsilon}\{X_N^{adv} + \alpha \text{sign}(\nabla_X J(X_N^{adv}, y_{true}))\},$$

где $\text{Clip}_{X,\varepsilon}\{X'\}$ — функция, выполняющая попиксельную обрезку изображения X', так что результат будет находиться в L_∞ ε-окрестности исходного изображения X; $J(X, y)$ — функция потерь на основе перекрестной энтропии нейронной сети, получающей изображение X и класс y; y_{true} — истинный класс для изображения X.

Функция потерь $f(x)$ для атак с использованием Elastic-Net (EAD) [11] определяется как:

$$f(x, t) = \max\{\max_{j \neq t}[\text{Logit}(x)]_j - [\text{Logit}(x)]_t, - \kappa\},$$

где $\text{Logit}(x) = [[\text{Logit}(x)]_1, \ldots, [\text{Logit}(x)]_\kappa] \in \mathbb{R}^\kappa$ — logit-слой (слой, предшествующий слою softmax) представляющий x в рассматриваемой глубокой нейронной сети; κ — количество классов для классификации; $\kappa \geq 0$ — порог достоверности, гарантирующий постоянный разрыв между $\max_{j \neq t}[\text{Logit}(x)]_j$ и $[\text{Logit}(x)]_t$.

Атака Expectation Over Transformation (EOT) [12] в сочетании с PGD оценивает реальный градиент сети как среднее значение градиентов по нескольким случайным векторам $\boldsymbol{\varepsilon}$:

$$\hat{x}_{t+1} \leftarrow \prod_{x+S}\left[\hat{x}_t + \eta \underset{\boldsymbol{\varepsilon}}{\text{M}}(\nabla_x L(f(x; w, \boldsymbol{\varepsilon})\hat{y})|_{x=\hat{x}_t}\right],$$

где $\hat{x}$ — состязательный пример; $\prod_{x+S}[\ldots]$ — проекция на множество допустимых возмущений S; η — коэффициент скорости обучения; $\underset{\boldsymbol{\varepsilon}}{\text{M}}(\ldots)$ — математическое ожидание по распределению $\boldsymbol{\varepsilon}$.

Функция потерь для атак на основе джиттера [13] имеет следующий вид:

$$\mathcal{L}_{jitter} = \begin{cases} \dfrac{\|\hat{z} - \mathbf{Y} + \mathcal{N}(0, \sigma)\|_2}{\|\gamma\|_p}, & \text{если неправильно классифицированы} \\ \|\hat{z} - \mathbf{Y} + \mathcal{N}(0, \sigma)\|_2, & \text{если еще не были неправильно классифицированы} \end{cases},$$

где $\hat{z} = \text{softmax}\left(\alpha \dfrac{z}{\|z\|_\infty}\right)$; $\mathbf{Y}$ — вектор истинных меток в унитарном коде; $\mathcal{N}(0, \sigma)$ — функция нормального шума; σ — уровень шума; γ — состязательное возмущение; p — норма.

Визуализация и используемые методы. Для интерпретации решений модели классификации был применен метод Grad-CAM [14], визуализирующий области изображения, наиболее влияющие на прогноз. Для визуализации кластеров изображений был задействован алгоритм t-SNE [15], заключающийся в минимизации расхождения Кульбака–Лейблера (Kullback–Leibler, KL) между распределениями вероятностей, построенными в исходном многомерном пространстве $p_{j|i}$ и в пространстве пониженной размерности $q_{j|i}$. Исходное сходство $p_{j|i}$ оценивается на основе гауссова ядра следующим образом:

$$p_{j|i} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right) \Big/ \sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right),$$

где x_j, x_i — точки данных; σ_i — дисперсия гауссовой функции, центрированной в x_i.

Аналогично в пространстве пониженной размерности:

$$q_{j|i} = \exp\left(-\|y_i - y_j\|^2\right) \Big/ \sum_{k \neq i} \exp\left(-\|y_i - x_k\|^2\right),$$

где y_j, y_i — аналоги точек данных в пространстве пониженной размерности.

Целевая функция для минимизации:

$$C = \sum_i \text{KL}(P_i|Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

где P_i и Q_i — условные распределения вероятностей по всем другим точкам данных и карты, заданным x_i и y_i соответственно.

Далее рассматриваются алгоритмы KNN, используемые в задачах кластеризации.

Алгоритм KNN [16] используется для поиска k-ближайших соседей в пространстве признаков. При поиске вычисляются расстояния (например, евклидово, манхэттенское или Минковского) между точкой запроса и всеми точками обучающего набора.

Алгоритм KNN с помощью HNSW [17] — усовершенствованный метод поиска ближайших соседей. HNSW может использоваться с различными показателями сходства: косинусным ($d = 1{,}0 - \sum(A_i B_i) \big/ \sqrt{\sum A_i^2 \sum B_i^2}$; L_2 в квадрате ($d = \sum((A_i - B_i)^2)$); внутреннего произведения ($d = 1{,}0 - \sum((A_i B_i))$. A_i, B_i — значения i-го признака векторов признаков **A** и **B** соответственно.

Основные результаты

Оценка влияния состязательных атак на Grad-CAM в задаче классификации. Для изучения влияния состязательных атак на Grad-CAM в задаче классификации проводились эксперименты с использованием эталонного набора данных ImageNet-1k [18] при различных сценариях атак. Применялись такие методы, как: FGSM (BIM), PGD, CW (с метрикой L_2), EAD с нормой L_∞, Expectation Over Transformation Projected Gradient Descent (EOTPGD) и атаки на основе джиттера. Для каждого метода генерировались возмущения и оценивалось их влияние на результат Grad-CAM. Результаты экспериментов представлены на рис. 1 и 2.

FGSM и атака на основе джиттера влияют на Grad-CAM меньше, чем другие. FGSM генерирует состязательные примеры, делая один шаг в направлении градиента, используя фиксированный размер шага. Из-за относительно неглубокой природы атаки внимание сети в основном остается на важных областях, что приводит к минимальному влиянию на визуализацию. Атака на основе джиттера предполагает внесение небольших случайных изменений (например, незначительных сдвигов или поворотов) во входное изображение. Эти изменения вызывают некоторую вариативность изображения, но не изменяют существенных характеристик, на которые опирается модель глубокого обучения, такая как ResNet50, при классификации. Поскольку данный подход не нацелен непосредственно на границы принятия решений или градиенты модели в очень агрессивной манере, он оказывает менее разрушительное воздействие по сравнению с градиентными методами.

Более продвинутые атаки используют многочисленные итерации и методы оптимизации, оказывая более агрессивное влияние на процесс принятия решений в модели, что может привести к значительным изменениям в областях фокусировки нейронной сети.

На основе полученных визуализаций можно сделать вывод, что различные атаки оказывают разное влияние на разные слои сети. Атаки на основе FGSM и джиттера приводят к размытию фокуса модели, сохраняя наиболее значимую часть исходного изображения. В то же время изучение визуального представления шаблонов внимания начальных слоев нейронной сети показывает, что атаки на основе FGSM и джиттера вызывают меньшую степень изменения восприятия по сравнению с другими типами атак.

Оценка влияния состязательных атак на кластеризацию изображений. Для изучения влияния состязательных атак на визуализацию кластеризации изображений проводились эксперименты с целью изменения класса изображения при различных сценариях атак. Использовались такие методы, как: FGSM (BIM), PGD, CW, EAD с нормой , EOTPGD и атаки на основе джиттера. Для каждого метода генерировались возмущения и оценивалось их влияние на визуализацию.

Изображения с классом «королевский пингвин» подвергались атаке с целью изменить класс исход-

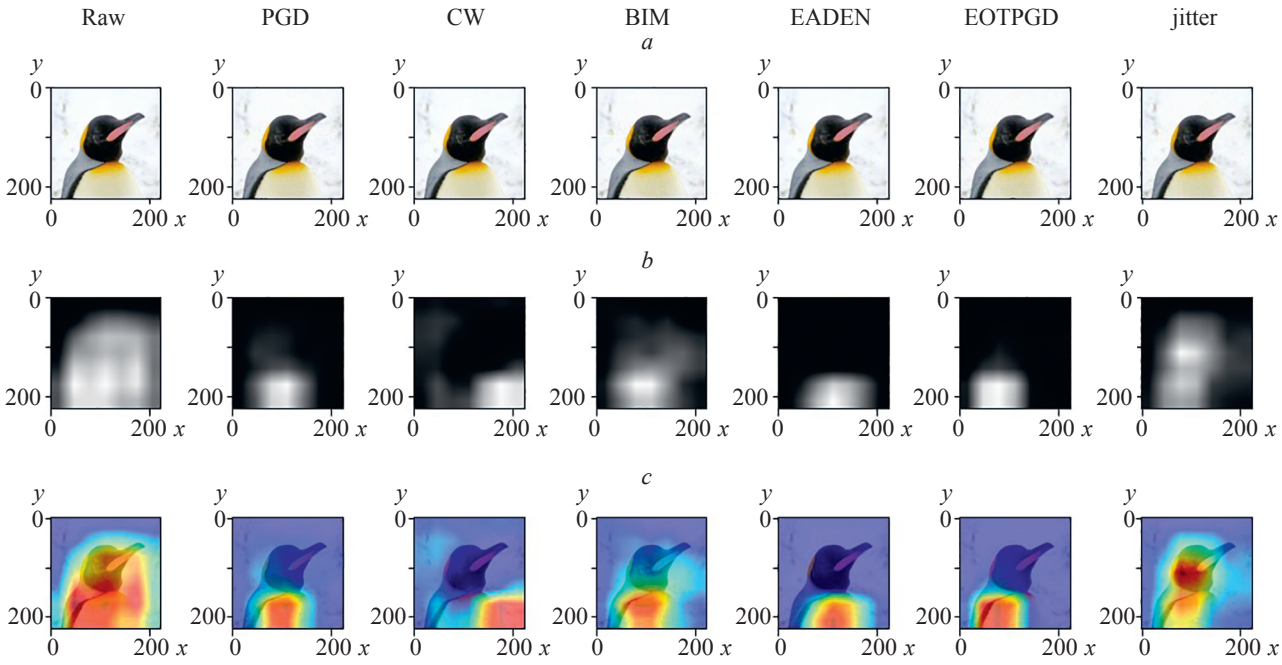

Рис. 1. Сравнение исходного изображения (*a*) с атакуемыми при помощи Grad-CAM (*b*, *c*) на четвертом слое модели ResNet50. Значения по осям *x* и *y* — в пикселах. Jitter — атаки на основе джиттера

Fig. 1. Comparison of the original image (*a*) with the attacked images using Grad-CAM (*b*, *c*) on the fourth layer of the ResNet50 model. The values on the *x* and *y* axes are in pixels. Jitter distortions are attacks based on jitter

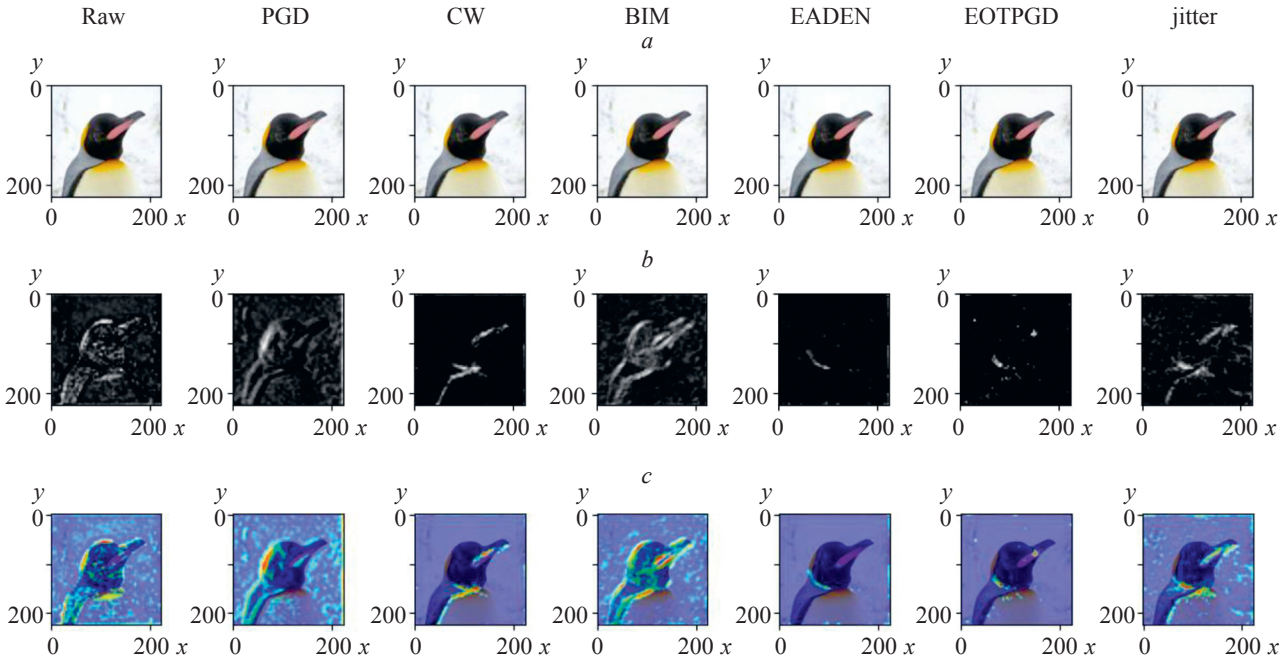

Рис. 2. Сравнение исходного изображения (*a*) с атакуемыми при помощи Grad-CAM (*b*, *c*) на первом слое модели ResNet50. Значения по осям *x* и *y* — в пикселах. Jitter — атаки на основе джиттера

Fig. 2. Comparison of the original image (*a*) with the attacked ones using Grad-CAM (*b*, *c*) on the first layer of the ResNet50 model. The values on the *x* and *y* axes are in pixels. Jitter distortions are attacks based on jitter

ного изображения на «футбольный мяч». На рис. 3 представлена визуализация с использованием t-SNE, а на рис. 4 — интересующая область рис. 3 в увеличении.

На рис. 3 и 4 оси отражают двумерную проекцию признаков, полученных с использованием t-SNE. Несмотря на то, что оси не обладают конкретной физической интерпретацией, взаимное расположение точек показывает степень схожести внутренних представлений изображений в модели [15]. Смещение точек под действием атак свидетельствует о существенном изменении восприятия модели.

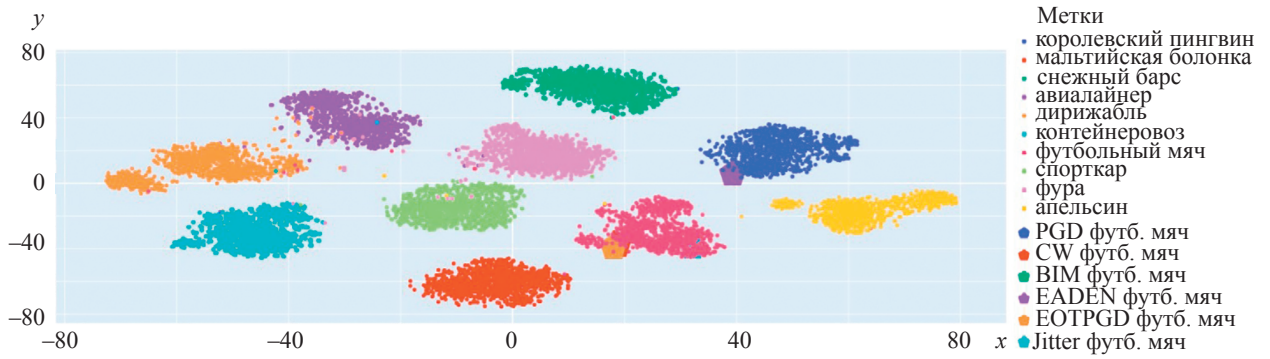

Рис. 3. Кластеризация исходных изображений и изображений, подвергшихся атаке различными методами. Название было изменено с «королевский пингвин» на «футбольный мяч». Значения по осям *x* и *y* – безразмерная компонента проекции признакового вектора

Fig. 3. Clusterization of the original images and images attacked by various methods. The name was changed from “king penguin” to “soccer ball”. The values on the *x* and *y* axes are the dimensionless component of the feature vector projection

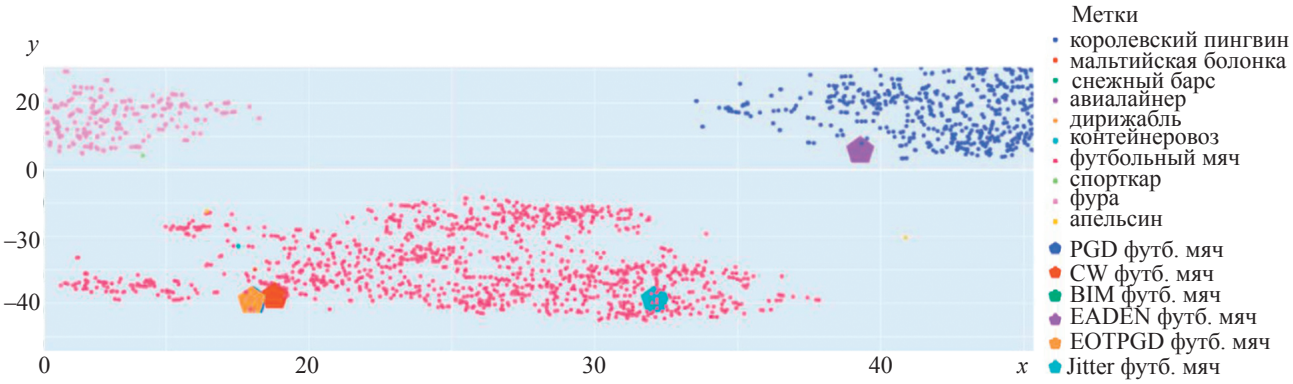

Рис. 4. Кластеризация исходных изображений и изображений, подвергшихся атаке различными методами. Увеличенный фрагмент рис. 3. Название было изменено с «королевский пингвин» на «футбольный мяч». Значения по осям *x* и *y* — безразмерная компонента проекции признакового вектора

Fig. 4. Clusterization of the original images and images attacked by various methods. A closer look at the original and attacked inserts and images. The name was changed from “king penguin” to “soccer ball”. The values on the *x* and *y* axes are the dimensionless component of the feature vector projection

На рис. 4 изображение, сгенерированное алгоритмом EADEN, попадает в тот же кластер, что и исходное изображение. Также из общей группы атакованных изображений выделилось внедрение, сгенерированное в ходе атаки на основе джиттера. Остальные атаки образовали плотный кластер, от которого немного отделился алгоритм CW.

Сравнительные результаты проведенных атак. Для сравнения сходства исходных и атакованных изображений был использован метод скалярного произведения, который позволяет оценить сходство между векторами в заданном пространстве. Полученные сходства представлены в табл. 1.

Таблица 1. Сходства для атакованных и различных классов исходных изображений

Table 1. Similarities for attacked and different classes of pure images

Тип атаки	Скалярное произведение с изображением случайного класса	Скалярное произведение с изображением оригинального класса	Различие между случайным и оригинальным классами	Скалярное произведение с изображением целевого класса	Различие между случайным и целевым классами
PGD	427,38	639,34	211,95	521,20	91,38
CW	376,59	569,24	192,65	430,78	54,18
BIM	415,03	582,20	167,18	450,36	35,33
EADEN	439,24	779,36	340,12	553,97	114,73
EOTPGD	381,65	565,35	183,69	471,10	89,45
Атаки на основе джиттера	413,08	581,48	168,40	588,21	175,14

Атаки, использующие итеративные методы, такие как PGD и BIM, приводят к значительному смещению пространства признаков, что отражается в увеличенных значениях скалярного произведения с изображениями исходного класса. При этом EADEN, основанный на комбинированной $L_1 - L_2$ регуляризации, демонстрирует особо высокие показатели сходства с исходным классом, что говорит о том, что возмущения сохраняют высокий уровень корреляции с оригинальными характеристиками изображения. В то же время атака CW, стремящаяся минимизировать величину возмущений, приводит к относительно невысоким изменениям в представлении, что подтверждается меньшей разницей между случайным и целевым классами. Наиболее эффективной оказалась атака на основе джиттера. Полученные результаты демонстрируют, что в большинстве случаев наблюдается заметная разница в скалярных произведениях атакованных и исходных изображений.

В данной работе атакованные изображения были сгруппированы с помощью алгоритма KNN (с использованием евклидова расстояния), алгоритма поиска с HNSW с косинусным сходством, со сходством L_2 и со сходством внутреннего произведения. В табл. 2 представлен показатель успешности атаки (Attack Success Rate, ASR) различных алгоритмов.

ASR для KNN у большинства методов достаточно высок, что говорит о чувствительности данного метода к возмущениям. Исключением является EADEN с ASR равным 0,0, что указывает на специфику данного метода, при котором возмущения не позволяют существенно сместить пространство признаков.

Для HNSW с косинусным сходством ASR в целом также остается высоким, указывая на чувствительность данного метода в связи с тем, что угловое расстояние между векторами легко изменяется под воздействием атак. Наихудшие результаты были продемонстрированы метрикой HNSW со сходством L_2. ASR равный 1,0 для всех типов атак говорит о практически полном отсутствии устойчивости данной метрики к ним. Подобный результат связан с тем, что для генерации возмущений использовались L_2-оптимизированные методы. Наилучший результат продемонстрировала последняя метрика, где ASR не поднимался выше значения 0,5. Это означает, что внутреннее представление оказывается менее подверженным влиянию атакующих возмущений.

Таким образом, состязательные атаки оказывают неоднородное воздействие на внутреннее представление изображений в модели, что проявляется, как в изменении скалярных произведений, так и в успешности атак на алгоритмы кластеризации. Итеративные методы, направленные на минимизацию возмущений, по-разному смещают признаки, что указывает на сложность их влияния на пространство признаков. Кроме того, результаты демонстрируют важность выбора метрики для кластеризации. Применение HNSW со сходством внутреннего произведения позволяет добиться более высокой устойчивости к атакам, что является критически важным аспектом при разработке систем компьютерного зрения, способных работать в условиях потенциальных угроз.

Обсуждение

Полученные результаты подчеркивают необходимость дальнейших исследований в области повышения стойкости моделей компьютерного зрения к состязательным атакам и разработке методов, позволяющих минимизировать их влияние на качество классификации и кластеризации изображений. Одним из направлений будущих исследований является детализация характеристик, определяющих их восприимчивость к атакам, и определение возможности предсказания по ним успешности атаки без ее проведения. Развитие подобных методов предсказания может позволить автоматизировано адаптировать защитные механизмы модели, в зависимости от потенциальных угроз. Другим перспективным направлением является исследование способов адаптации метрик сходства для повышения устойчивости к атакам. Полученные результаты показали, что использование внутреннего произведения в алгоритме HNSW снижает успешность атак. В этом контексте целесообразно исследовать возможности гибридных метрик, комбинирующих преимущества и учитывающих особенности пространства признаков в процессе кластеризации. Также важным практическим направлением является разработка механизмов защиты, учитывающих специфику выявленных закономерностей. В частности, использование методов регуляризации, направленных на увеличение расстояний между представлениями различных классов или на стабилизацию их относительных позиций в пространстве, может снизить уязвимость модели к атакам.

Таблица 2. Показатели успешности атак при использовании различных алгоритмов кластеризации

Table 2. Attack success rates for different clustering algorithms

Тип атаки	KNN поверх пространства признаков с использованием L_2	Выходные данные модели	HNSW с косинусным сходством	HNSW со сходством L_2	HNSW со сходством внутреннего произведения
PGD	1,00	1,0	1,00	1,0	0,25
CW	0,63	1,0	0,95	1,0	0,11
BIM	1,00	1,0	1,00	1,0	0,37
EADEN	0,00	1,0	0,57	1,0	0,14
EOTPGD	0,89	1,0	0,95	1,0	0,21
Атаки на основе джиттера	1,00	1,0	1,00	1,0	0,50

Заключение

Проведенный анализ показал, что целенаправленные возмущения оказывают неоднозначное влияние на распределение признаковых представлений в модели ResNet50. Сопоставление результатов визуализации посредством метода Gradient-Weighted Class Activation Mapping, изменения показателей сходства и динамики кластеризации продемонстрировали, что итеративные методы атак вызывают выраженное смещение внутренних представлений, в то время как методы с элементами случайных преобразований позволяют сохранить ключевые характеристики исходных изображений.

Полученные результаты указывают на зависимость устойчивости модели от параметров атак и выбора метрик сходства, что проявляется в особенностях формирования кластерных структур. Выявленные закономерности в перераспределении пространства признаков в условиях целенаправленных атак открывают перспективы для дальнейшей оптимизации алгоритмов кластеризации, способных обеспечить более высокую степень защиты систем компьютерного зрения.

Таким образом, выполненное исследование не только подтверждает известные уязвимости современных моделей, но и выявляет дополнительные нюансы взаимодействия атакующих воздействий с внутренней структурой сети, особенно в аспекте кластеризации, что может служить основой для дальнейших научных разработок в данной области.

Литература

1. Liu A. Guo J., Wang J., Liang S., Tao R., Zhou W., Liu C., Liu X., Tao D. X-adv: Physical adversarial object attacks against x-ray prohibited item detection // arXiv. 2023. arXiv:2302.09491. https://doi.org/10.48550/arXiv.2302.09491
2. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // arXiv. 2015. arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
3. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks // arXiv. 2019. arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083
4. Carlini N., Wagner D. Towards evaluating the robustness of neural networks // Proc. of the IEEE Symposium on Security and Privacy (SP). 2017. P. 39–57. https://doi.org/10.1109/SP.2017.49
5. Qian Y., He S., Zhao C., Sha J. Wang W., Wang B. Lea2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 4487–4498. https://doi.org/10.1109/iccv51070.2023.00416
6. Schlarmann C., Singh N.D., Croce F., Hein M. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models // Proc. of the 41st International Conference on Machine Learning. 2024. N 1779. P. 43684–43704.
7. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 770–778. https://doi.org/10.1109/CVPR.2016.90
8. Liu X., Hu J., Yang Q., Jiang M., He J., Fang H. A divide-and-conquer reconstruction method for defending against adversarial example attacks // Visual Intelligence. 2024. V. 2. P. 30. https://doi.org/10.1007/s44267-024-00061-y
9. Zhang J., Wu W., Huang J., Huang Y., Wang W., Su Y., Lyu M. Improving adversarial transferability via neuron attribution-based attacks // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 14973–14982. https://doi.org/10.1109/CVPR52688.2022.01457
10. Kurakin A., Goodfellow I., Bengio S. Adversarial examples in the physical world // Artificial Intelligence Safety and Security. 2018. P. 14. https://doi.org/10.1201/9781351251389-8
11. Chen P.-Y., Sharma Y., Zhang H., Yi J. & Hsieh C.-J. Ead: Elastic-net attacks to deep neural networks via adversarial examples // Proc. of the 32nd AAAI Conference on Artificial Intelligence. 2018. V. 32. N 1. P. 10–17. https://doi.org/10.1609/aaai.v32i1.11302
12. Zimmermann R.S. Comment on “adv-bnn: Improved adversarial defense through robust bayesian neural network” // arXiv. 2019. arXiv:1907.00895. https://doi.org/10.48550/arXiv.1907.00895
13. Schwinn L., Raab R., Nguyen A., Zanca D., Eskofier B. Exploring misclassifications of robust neural networks to enhance adversarial attacks // Applied Intelligence. 2023. V. 53. N 17. P. 19843–19859. https://doi.org/10.1007/s10489-023-04532-5
14. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization // International Journal of Computer

References

1. Liu A. Guo J., Wang J., Liang S., Tao R., Zhou W., Liu C., Liu X., Tao D. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. *arXiv*, 2023, arXiv:2302.09491. https://doi.org/10.48550/arXiv.2302.09491
2. Goodfellow I.J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *arXiv*, 2015, arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572
3. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards deep learning models resistant to adversarial attacks. *arXiv*, 2019, arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083
4. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. *Proc. of the IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. https://doi.org/10.1109/SP.2017.49
5. Qian Y., He S., Zhao C., Sha J. Wang W., Wang B. Lea2: A lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4487–4498. https://doi.org/10.1109/iccv51070.2023.00416
6. Schlarmann C., Singh N.D., Croce F., Hein M. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *Proc. of the 41st International Conference on Machine Learning*, 2024, no. 1779. pp. 43684–43704.
7. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
8. Liu X., Hu J., Yang Q., Jiang M., He J., Fang H. A divide-and-conquer reconstruction method for defending against adversarial example attacks. *Visual Intelligence*, 2024, vol. 2, pp. 30. https://doi.org/10.1007/s44267-024-00061-y
9. Zhang J., Wu W., Huang J., Huang Y., Wang W., Su Y., Lyu M. Improving adversarial transferability via neuron attribution-based attacks. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14973–14982. https://doi.org/10.1109/CVPR52688.2022.01457
10. Kurakin A., Goodfellow I., Bengio S. Adversarial examples in the physical world. *Artificial Intelligence Safety and Security*, 2018, pp. 14. https://doi.org/10.1201/9781351251389-8
11. Chen P.-Y., Sharma Y., Zhang H., Yi J. & Hsieh C.-J. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *Proc. of the 32nd AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1, pp. 10–17. https://doi.org/10.1609/aaai.v32i1.11302
12. Zimmermann R.S. Comment on “adv-bnn: Improved adversarial defense through robust bayesian neural network”. *arXiv*, 2019, arXiv:1907.00895. https://doi.org/10.48550/arXiv.1907.00895
13. Schwinn L., Raab R., Nguyen A., Zanca D., Eskofier B. Exploring misclassifications of robust neural networks to enhance adversarial attacks. *Applied Intelligence*, 2023, vol. 53, no. 17, pp. 19843–19859. https://doi.org/10.1007/s10489-023-04532-5
14. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer*

Vision. 2020. V. 128. N 2. P. 336–359. https://doi.org/10.1007/s11263-019-01228-7
15. van der Maaten L., Hinton G. Visualizing data using t-SNE // Journal of Machine Learning Research. 2008. V. 9. P. 2579–2605.
16. Fix E., Hodges J. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, 1951. 44 p.
17. Malkov Y.A., Yashunin D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020. V. 42. N 4. P. 824–836. https://doi.org/10.1109/TPAMI.2018.2889473
18. ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Stanford Vision Lab, Stanford University, Princeton University. ImageNet Data [Электронный ресурс]. URL: https://www.image-net.org/download.php. (дата обращения: 03.03.2025).

Vision, 2020, vol. 128, no. 2, pp. 336–359. https://doi.org/10.1007/s11263-019-01228-7
15. van der Maaten L., Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, vol. 9, pp. 2579–2605.
16. Fix E., Hodges J. *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951, 44 p.
17. Malkov Y.A., Yashunin D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 4, pp. 824–836. https://doi.org/10.1109/TPAMI.2018.2889473
18. *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*. Stanford Vision Lab, Stanford University, Princeton University. ImageNet Data [Электронный ресурс]. Available at: https://www.image-net.org/download.php. (accessed: 03.03.2025).

Авторы

Болозовский Роман Ростиславович — аспирант, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация, https://orcid.org/0009-0004-5725-0526, bolozovskii@gmail.com

Левина Алла Борисовна — кандидат физико-математических наук, доцент, доцент, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация, sc 56427692900, https://orcid.org/0000-0003-4421-2411, Alla_levina@mail.ru

Красов Константин Сергеевич — младший научный сотрудник, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация, https://orcid.org/0009-0005-9232-5463, iflup@ya.ru

Authors

Roman R. Bolozovskii — PhD Student, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, 197022, Russian Federation, https://orcid.org/0009-0004-5725-0526, bolozovskii@gmail.com

Alla B. Levina — PhD (Physics & Mathematics), Associate professor, Associate Professor, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, 197022, Russian Federation, sc 56427692900, https://orcid.org/0000-0003-4421-2411, Alla_levina@mail.ru

Konstantin S. Krasov — Junior Researcher, Saint Petersburg Electrotechnical University “LETI”, Saint Petersburg, 197022, Russian Federation, https://orcid.org/0009-0005-9232-5463, iflup@ya.ru

Статья поступила в редакцию 05.03.2025
Одобрена после рецензирования 30.05.2025
Принята к печати 20.07.2025

Received 05.03.2025
Approved after reviewing 30.05.2025
Accepted 20.07.2025