НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2024-24-6-1066-1070
УДК 004.89

Анализ уязвимости нейросетевых моделей YOLO к атаке Fast Sign Gradient Method

Николай Валерьевич Тетерев[1], Владислав Евгеньевич Трифонов[2], Алла Борисовна Левина[3]✉

[1,2,3] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация
[1] АО «Научно-инженерный центр Санкт-Петербургского электротехнического университета», Санкт-Петербург, 194021, Российская Федерация

[1] teterevkolya21@gmail.com, https://orcid.org/0009-0001-3394-9883
[2] vtr1f0nov@yandex.ru, https://orcid.org/0009-0002-5839-2812
[3] Alla_levina@mail.ru✉, https://orcid.org/0000-0003-4421-2411

Аннотация
Представлен анализ формализованных условий создания универсальных изображений, ложно классифицируемых алгоритмами компьютерного зрения, называемыми состязательными примерами, на нейросетевые модели YOLO. Выявлена и исследована закономерность успешного создания универсального деструктивного изображения в зависимости от сгенерированного набора данных, на котором происходило обучение нейронных сетей с помощью атаки Fast Sign Gradient Method. Указанная закономерность продемонстрирована для моделей классификатора YOLO8, YOLO9, YOLO10, YOLO11, обученных на стандартном наборе данных COCO.

Ключевые слова
генеративные атаки, состязательный пример, YOLO, COCO, набор данных, нейронная сеть

Благодарности
Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации № 075-00003-24-01 от 08.02.2024 (проект FSEE-2024-0003).

Ссылка для цитирования: Тетерев Н.В., Трифонов В.Е., Левина А.Б. Анализ уязвимости нейросетевых моделей YOLO к атаке Fast Sign Gradient Method // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 1066–1070. doi: 10.17586/2226-1494-2024-24-6-1066-1070

Analysis of the vulnerability of YOLO neural network models to the Fast Sign Gradient Method attack

Nikolai V. Teterev[1], Vladislav E. Trifonov[2], Alla B. Levina[3]✉

[1,2,3] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation
[1] Research & Engineering Center JSC "R&EC ETU", Saint Petersburg, 194021, Russian Federation

[1] teterevkolya21@gmail.com, https://orcid.org/0009-0001-3394-9883
[2] vtr1f0nov@yandex.ru, https://orcid.org/0009-0002-5839-2812
[3] Alla_levina@mail.ru✉, https://orcid.org/0000-0003-4421-2411

Abstract
The analysis of formalized conditions for creating universal images falsely classified by computer vision algorithms, called adversarial examples, on YOLO neural network models is presented. The pattern of successful creation of a universal destructive image depending on the generated dataset on which neural networks were trained using the Fast Sign Gradient Method attack is identified and studied. The specified pattern is demonstrated for YOLO8, YOLO9, YOLO10, YOLO11 classifier models trained on the standard COCO dataset.

Keywords
adversarial attacks, adversarial example, YOLO, COCO, dataset, neural network

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. 075-00003-24-01 dated 08.02.2024 (FSEE-2024-0003 project).

For citation: Teterev N.V., Trifonov V.E., Levina A.B. Analysis of the vulnerability of YOLO neural network models to the Fast Sign Gradient Method attack. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 1066–1070 (in Russian). doi: 10.17586/2226-1494-2024-24-6-1066-1070

Развитие искусственного интеллекта вызвало рост исследований, посвященных созданию угроз, влияющих на результат работы нейронных сетей (НС) [1]. В настоящее время НС, используемые в алгоритмах компьютерного зрения, тесно вошли в повседневную жизнь каждого человека. Атаки, рассмотренные в данной работе, используются злоумышленниками, чьими жертвами может стать любая организация, применяющая алгоритмы компьютерного зрения в своей работе [2].

В научных работах рассматривается множество различных атак на нейросетевые алгоритмы компьютерного зрения [3], основная часть которых была обнаружена с 2016 по 2020 года. Данные атаки разделяют на два основных класса: black box и white box [4]. Различия между атаками заключаются в знании параметров НС. При атаках white box злоумышленник знает все параметры модели для проведения успешной атаки. В сценарии атаки с black box злоумышленник рассматривает модель как «черный ящик» и пытается найти уязвимости, не имея подробных знаний о внутреннем устройстве модели. Эти атаки особенно сложны, потому что злоумышленник должен полагаться на внешние наблюдения и запросы для создания эффективных примеров состязательности.

Основная идея состязательных атак на алгоритмы компьютерного зрения заключается в создании состязательного примера, за счет добавления к исходному изображению сгенерированных данных («возмущений») [5], как показано на рис. 1.

После добавления «возмущения» к исходному изображению, классифицирующая НС распознает объект на изображении иначе, чем его видит человеческий глаз.

НС отличаются друг от друга не только архитектурой, но и способом их обучения. Наибольшее распространение получил метод «обратного распространения ошибок», основанный на вычислении градиента функции ошибок [6]. В рамках обучения НС градиент используется для вычисления минимума функции ошибки. Данный градиент был использован для обучения НС в настоящей работе.

Одним из наиболее известных классов атак на компьютерное зрение является атака на градиент спуска. Графически поиск градиента при обучении НС показан на рис. 2.

На каждой итерации поиска минимума функции ошибки высчитывается градиент и сравнивается с его предыдущим значением. В случае если предыдущее значение больше, то направление спуска остается прежним, а если меньше, то направление меняется.

Основная идея состязательных атак на градиент заключается в уязвимости метода обратного распространения ошибок (алгоритм, используемый для обучения НС). Одной из наиболее распространенных атак на основе градиента является атака Fast Sign Gradient Method (FGSM).

FGSM [7] — атака быстрого градиентного спуска, нацеленная на уязвимость НС к линейным состязательным «возмущениям». Это означает, что незначительные изменения в исходном изображении способны повлиять на результат работы НС. FGSM описывается в виде:

$$\eta = \epsilon \text{sign}(\nabla_x J(\theta, x, y)),$$

где J — функция затрат; θ — параметр модели; x — входные данные для модели; y — цель, ϵ — коэффициент, который определяет размер «возмущения»; sign — знак каждого элемента вектора, определяется знаком входного градиента в элементе. В ходе атаки берется минимальный градиент функции потерь, и, отталкива-

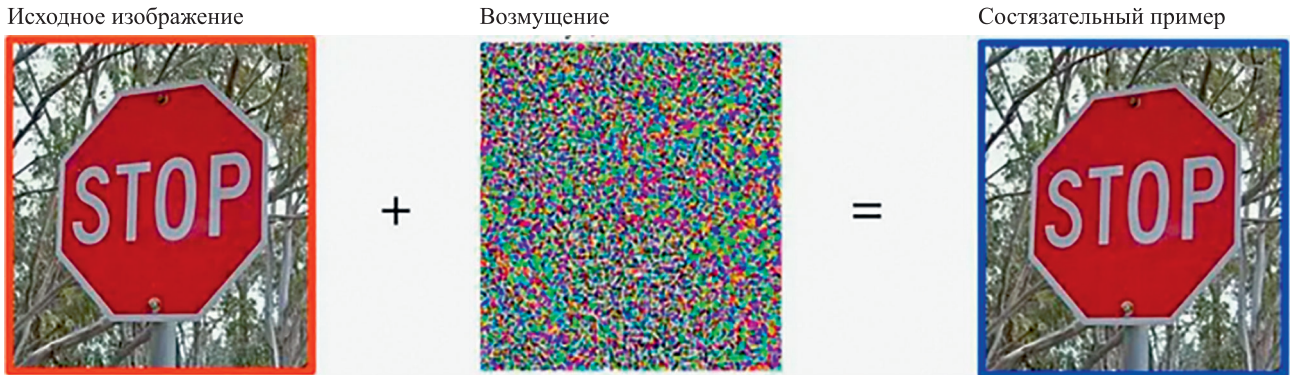

Рис. 1. Добавление «возмущений» для создания состязательного примера
Fig. 1. Adding “perturbations” to create an adversarial example

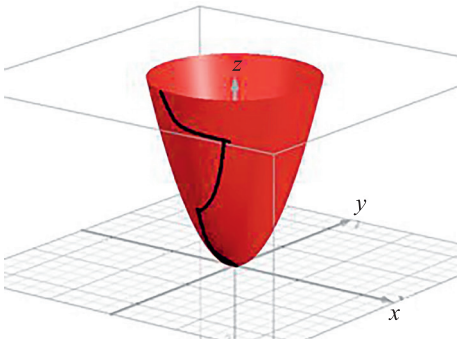

Рис. 2. 3D-модель метода обучения, основанного на градиентном спуске, где черным цветом проиллюстрировано направление скорейшего убывания

Fig. 2. Graphical illustration of the gradient descent learning method

ясь от него, происходит вычисление состязательного примера.

FGSM не требует итеративной процедуры (многократного повторения) для вычисления состязательных примеров, что обеспечивает ее выигрыш в скорости. Она относится к атакам white box, когда у злоумышленника есть доступ к НС модели и ее коэффициентам. Идея атаки заключается в использовании градиента в обратном направлении при градиентном спуске.

К атакам на градиент также относятся:

— CW (Carlini and Wagner Attacks — атаки Карлини и Вагнера) [8] — рассматривает задачу создания состязательных примеров как задачу оптимизации, стремясь найти наименьшее возмущение входных данных, которое вызывает неправильную классификацию целевой моделью.

Целевая функция:

$$J(x') = \alpha \text{dist}(\mathbf{x}, \mathbf{x}') + \beta \text{loss}(f(x'), y_t),$$

где α, β — положительные константы; **x** — вектор входных данных; **x′** — вектор возмущенных данных; dist ∈ $\{L_2, L_\infty\}$ — измерение возмущений, где L_2, L_∞ — множество норм; loss($f(x')$, y_t) — потери при неправильной классификации целевой модели f в следствии добавления возмущенных входных данных, относительно целевого класса y_t.

— ZOO (Zero-Order Optimization — оптимизация нулевого порядка) [9] — атака, которая используется в основном в ситуациях, когда информация о градиентах недоступна, например, в сценариях оптимизации «черного ящика», где доступны только оценки функций (выходные данные) без прямого доступа к градиентам.

— SimBA (Simple Black-Box Attack — простая атака «черного ящика») [10] — атака, фокусирующаяся на генерации состязательных примеров путем итеративного изменения входных запросов и наблюдения за реакциями модели для определения эффекта возмущений. Этот итеративный процесс позволяет злоумышленнику исследовать пространство ввода и создавать эффективные примеры состязательности, не требуя знания архитектуры модели, или градиентов. Формула измерения значения целевой функции для каждого входного слоя имеет вид [10]

$$(P_X, P_Y)_i = \bigcup_{\{(a,\,b)\in(P^*_X,\,P^*_Y)_{i-1}\}} \bigcup_{\begin{cases} x\in[a-d,a+d], \\ y\in[b-d,b+d] \end{cases}} (x, y),$$

где P_X — значение входного слоя; P_Y — значение целевой функции для выходного слоя; d — параметр, обозначающий половину длины стороны квадрата, в котором производятся изменения; a, b — числовые значения пикселов.

В рамках работы проводилось тестирование нейронных моделей YOLO, классифицирующих объекты на изображении, на устойчивость к созданию универсальных состязательных примеров для атаки FGSM. Из-за различия архитектур НС YOLO идея о создании одного изображения для всех моделей кажется невозможной. Однако при одинаковом наборе данных возможна ситуация, при которой минимальные значения градиента функции потерь будут находиться на небольшом расстоянии друг от друга, как показано на рис. 3.

На приведенных 3D-моделях схематично изображено пересечение градиентов. На рис. 3 видно, что три глобальных минимума находятся в непосредственной близости. Это может быть использовано для создания деструктивного изображения.

Представленная последовательность действий реализована в проводимых исследованиях для моделей НС YOLO, которые имеют схожую архитектуру (YOLO8–YOLO11). Модели были обучены на одном и том же наборе данных Common Objects in Context (COCO) — набор данных из 330 тыс. изображений, из которых 200 тыс. имеют аннотации для задач обнаружения объектов, сегментации и создания надписей. В ходе эксперимента удалось найти точку пересечения для трех из четырех моделей, относительно которой получилось создать универсальный состязательный пример, работаю-

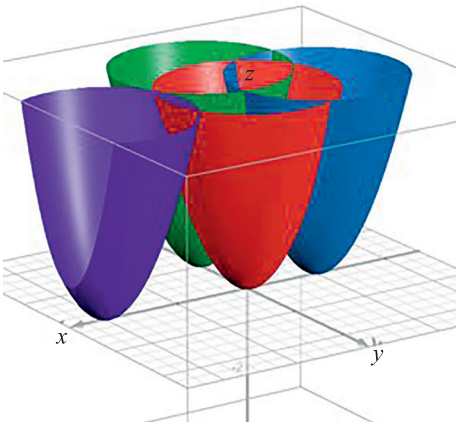

Рис. 3. Расположение градиентных минимумов для различных нейронных сетей: YOLO11 (фиолетовая фигура); YOLO10 (красная фигура); YOLO9 (синяя фигура); YOLO8 (зеленая фигура)

Fig. 3. Location of gradient minima for different neural network: YOLO11 (purple); YOLO10 (red); YOLO9 (blue); YOLO8 (green)

Таблица. Доля ложного распознавания состязательного примера, %

Table. False recognition rate of an adversarial example, %

Название модели	Размер возмущения є		
	0,07	0,10	0,15
YOLO8	38	46	59
YOLO9	63	72	78
YOLO10	0	24	45
YOLO11	0	0	2

щий для нескольких алгоритмов компьютерного зрения. Результаты эксперимента представлены в таблице, где для каждой модели YOLO написан процент ложного распознавания состязательного примера, в зависимости от коэффициента є, при котором внесенные возмущения остаются малозаметными для человеческого зрения.

Из результатов, приведенных в таблице, видно, что найденная точка общего градиентного минимума находится в непосредственной близости относительно градиента модели YOLO9. Также заметно, что значение не подходит для проведения состязательной атаки на YOLO11. Это объясняется тем, что разработчики ultralytics изменили архитектуру НС, что сместило точку глобального минимума для градиента функции потерь.

Заключение

В ходе проведения исследований была показана возможность создания состязательного примера с помощью атаки Fast Sign Gradient Method на модели нейросетей YOLO. Показано, что неуязвимой к данному классу атак является модель YOLO11.

Выявлена закономерность создания универсального деструктивного изображения с помощью атаки Fast Sign Gradient Method в зависимости от сгенерированного набора данных, на котором происходило обучение нейронных сетей.

Литература

1. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. Adversarial attacks and defences: A survey // arXiv. 2018. arXiv:1810.00069v1. https://doi.org/10.48550/arXiv.1810.00069
2. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey // IEEE Access. 2018. V. 6. P. 14410–14430. https://doi.org/10.1109/access.2018.2807385
3. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples // Proc. of the 3rd International Conference on Learning Representations, ICLR 2015. 2015.
4. Zhang C., Zhang H., Hsieh C.-J. An efficient adversarial attack for tree ensembles // Advances in Neural Information Processing Systems. 2020. V. 33.
5. Xiong P., Tegegn M., Sarin J.S., Pal S., Rubin J. It is all about data: A survey on the effects of data on adversarial robustness // ACM Computing Surveys. 2024. V. 56. N 7. P. 1–41. https://doi.org/10.1145/3627817
6. Zuo C. Regularization effect of fast gradient sign method and its generalization // arXiv. 2018. arXiv:1810.11711. https://doi.org/10.48550/arXiv.1810.11711
7. Yosinski J., Clune J., Nguyen A., Fuchs T., Lipson H. Understanding neural networks through deep visualization // arXiv. 2015. arXiv:1506.06579v1. https://doi.org/10.48550/arXiv.1506.06579
8. 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
9. Li Z., Chen P.-Y., Liu S., Lu S., Xu Y. Zeroth-order optimization for composite problems with functional constraints // Proceedings of the AAAI Conference on Artificial Intelligence. 2022. V. 36. N 7. P. 7453–7461. https://doi.org/10.1609/aaai.v36i7.20709
10. Guo C., Gardner J., You Y., Wilson A., Weinberger K. Simple black-box adversarial attacks // Proceedings of Machine Learning Research. 2019. V. 97. P. 2484–2493.

References

1. Chakraborty A., Alam M., Dey V., Chattopadhyay A., Mukhopadhyay D. Adversarial attacks and defences: A survey. *arXiv*, 2018, arXiv:1810.00069v1. https://doi.org/10.48550/arXiv.1810.00069
2. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 2018, vol. 6, pp. 14410–14430. https://doi.org/10.1109/access.2018.2807385
3. Goodfellow I., Shlens J., Szegedy C. Explaining and harnessing adversarial examples. *Proc. of the 3rd International Conference on Learning Representations, ICLR 2015*, 2015.
4. Zhang C., Zhang H., Hsieh C.-J. An efficient adversarial attack for tree ensembles. *Advances in Neural Information Processing Systems*, 2020, vol. 33.
5. Xiong P., Tegegn M., Sarin J.S., Pal S., Rubin J. It is all about data: A survey on the effects of data on adversarial robustness. *ACM Computing Surveys*, 2024, vol. 56, no. 7, pp. 1–41. https://doi.org/10.1145/3627817
6. Zuo C. Regularization effect of fast gradient sign method and its generalization. *arXiv*, 2018, arXiv:1810.11711. https://doi.org/10.48550/arXiv.1810.11711
7. Yosinski J., Clune J., Nguyen A., Fuchs T., Lipson H. Understanding neural networks through deep visualization. *arXiv*, 2015, arXiv:1506.06579v1. https://doi.org/10.48550/arXiv.1506.06579
8. 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
9. Li Z., Chen P.-Y., Liu S., Lu S., Xu Y. Zeroth-order optimization for composite problems with functional constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 7, pp. 7453–7461. https://doi.org/10.1609/aaai.v36i7.20709
10. Guo C., Gardner J., You Y., Wilson A., Weinberger K. Simple black-box adversarial attacks. *Proceedings of Machine Learning Research*, 2019, vol. 97, pp. 2484–2493.

Авторы

Тетерев Николай Валерьевич — младший научный сотрудник, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация; инженер, АО «Научно-инженерный центр Санкт-Петербургского электротехнического университета», Санкт-Петербург, 194021, Российская Федерация, https://orcid.org/0009-0001-3394-9883, teterevkolya21@gmail.com

Authors

Nikolai V. Teterev — Junior Researcher, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation; Engineer, Research & Engineering Center JSC "R&EC ETU", Saint Petersburg, 194021, Russian Federation, https://orcid.org/0009-0001-3394-9883, teterevkolya21@gmail.com

Трифонов Владислав Евгеньевич — младший научный сотрудник, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация, https://orcid.org/0009-0002-5839-2812, vtr1f0nov@yandex.ru

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

Vladislav E. Trifonov — Junior Researcher, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation, https://orcid.org/0009-0002-5839-2812, vtr1f0nov@yandex.ru

Alla B. Levina — PhD (Physica & 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

Статья поступила в редакцию 01.10.2024
Одобрена после рецензирования 06.11.2024
Принята к печати 24.11.2024

Received 01.10.2024
Approved after reviewing 06.11.2024
Accepted 24.11.2024