Н. Д. Локшин[1], *А. В. Хвостиков*[1], *А. С. Крылов*[1]

АУГМЕНТАЦИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ В ЗАДАЧЕ КЛАССИФИКАЦИИ ГИСТОЛОГИЧЕСКИХ ИЗОБРАЖЕНИЙ СЛАБЫМИ АДВЕРСАТИВНЫМИ АТАКАМИ*

Введение

Некоторые модели машинного обучения, в частности нейронные сети, часто подвергаются *адверсативным атакам* - то есть неверно классифицируют входы, получающиеся в результате наложения на входные данные малого шума [1–3] (Рис. 1). Данному феномену подвержены нейросети, применяющиеся во многих популярных областях, включая компьютерное зрение [4, 5] и обработку естественного языка [6, 7]. Из-за этого надежность нейросетевых методов до сих пор является актуальной областью изучения. На данный момент имеется большое количество способов генерации таких атак, а также и методов повышения устойчивости к ним [1, 8–10]. Большинство способов повышения устойчивости к адверсативным атакам либо предусматривают изменение структуры исходной модели предсказания, например, *защитная дистилляция* нейросетей [8, 11], либо строят предположения о возможных атаках.

Одним из самых эффективных методов повышения устойчивости нейросетей к адверсативным атакам на данный момент является так называемая *адверсативная аугментация* - то есть аугментация обучающих выборок заранее сгенерированными адверсативными атаками. Данный метод был впервые предложен авторами [1], а затем дополнен авторами [4, 12]. На основе работ [1, 4, 12] было разработано огромное множество различных методов повышения устойчивости к адверсативным атакам [10, 13–16].

Тем не менее, метод адверсативной аугментации имеет и недостатки. Метод аугментации, предложенный в [4], зачастую значительно замедляет обучение нейронных сетей и не позволяет добиться роста точности на чистых данных. А в работе [12], несмотря на

[1]Лаборатория математических методов обработки изображений, факультет Вычислительной математики и кибернетики, МГУ имени М. В. Ломоносова, e-mail: phd0230028@gse.cs.msu.ru, khvostikov@cs.msu.ru, kryl@cs.msu.ru ,

*Исследование выполнено за счет гранта Российского научного фонда №22-21-00081.

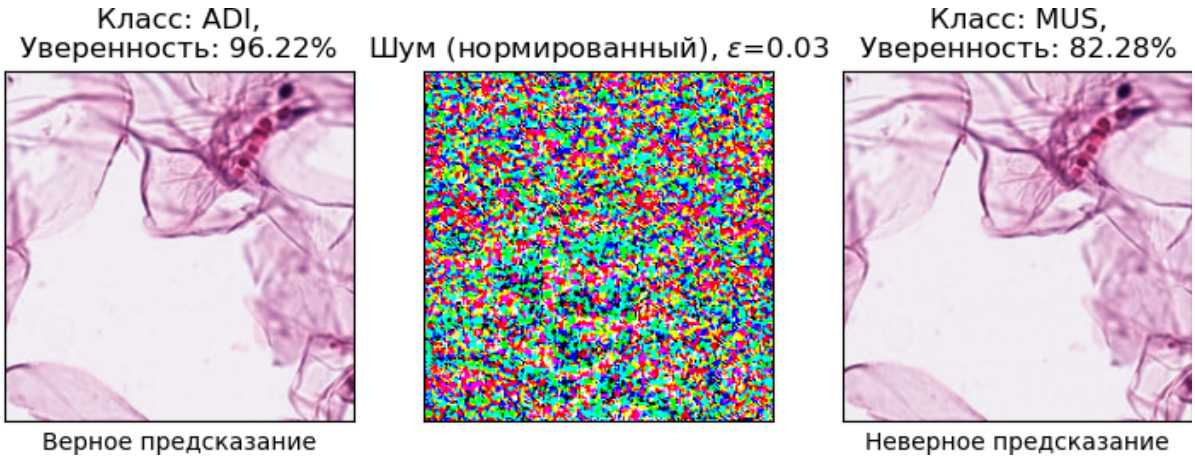

Рис. 1. Пример применения адверсативной атаки на гистологическое изображение, реализованной методом *Fast Gradient Sign Method* (FGSM), предложенного в [1]. Слева - исходное изображение, которое было верно определено классификатором, в середине - адверсативное возмущение, накладываемое на изображение с шагом $\varepsilon = 0.03$, справа - результат наложения, который был неверно определен классификатором.

более высокую скорость работы метода, на чистых данных точность обученного таким методом классификатора слегка ухудшается. Однако, исходя из того факта, что добавление случайного шума ко входным изображениям как способ аугментации при обучении позволяет повышать точность нейросетевых классификаторов на чистой выборке [17, 18], можно предположить, что аугментация адверсативными данными также на это способна. В данной работе на основе [1, 12] впервые предлагается такой метод повышения устойчивости нейросетевых классификаторов гистологических изображений к адверсативным атакам путем адверсативной аугментации, который позволяет также значительно повысить точность на чистых тестовых данных, то есть в некоторой степени нивелировать эффект переобучения на адверсативных атаках, используя для этого *слабые адверсативные атаки*, то есть адверсативные атаки с малыми коэффициентами. Такая же постановка задачи аугментации для задачи сегментации изображений адверсативными данными была поставлена в работе [19], однако в ней проиллюстрировано только ускорение обучения при добавлении в обучающую выборку атакованных изображений, повышения точности на чистых тестовых данных не показано.

Постановка задачи

Необходимо разработать такой метод повышения устойчивости нейросетевых классификаторов к адверсативным атакам, при котором точность обученных таким методом классификаторов на чистых данных

растет по сравнению с классификаторами той же архитектуры, обученными без применения этого метода.

Набор данных

В качестве входных данных имеется 107180 размеченных изображений - объединение наборов *NCT-CRC-HE-100K, CRC-VAL-HE-7K* [20, 21]. Изображения являются непересекающимися участками больших гистологических изображений, содержащих злокачественные новообразования толстого кишечника а также здоровые ткани, окрашенные гематоксилин-эозином. Изображения из CRC-VAL-HE-7K (7180 шт) образуют *скрытую выборку*, на которой проводится финальное тестирование нейросетевых классификаторов. Термин "скрытая"выборка означает, что изображения из этой выборки не принадлежат обучающей выборке. Изображения имеют размер $224 \times 224 \times 3$. Данные равномерно размечены на 9 классов тканей: *adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM)*. Примеры изображений для каждого класса приведены на Рис. 2.

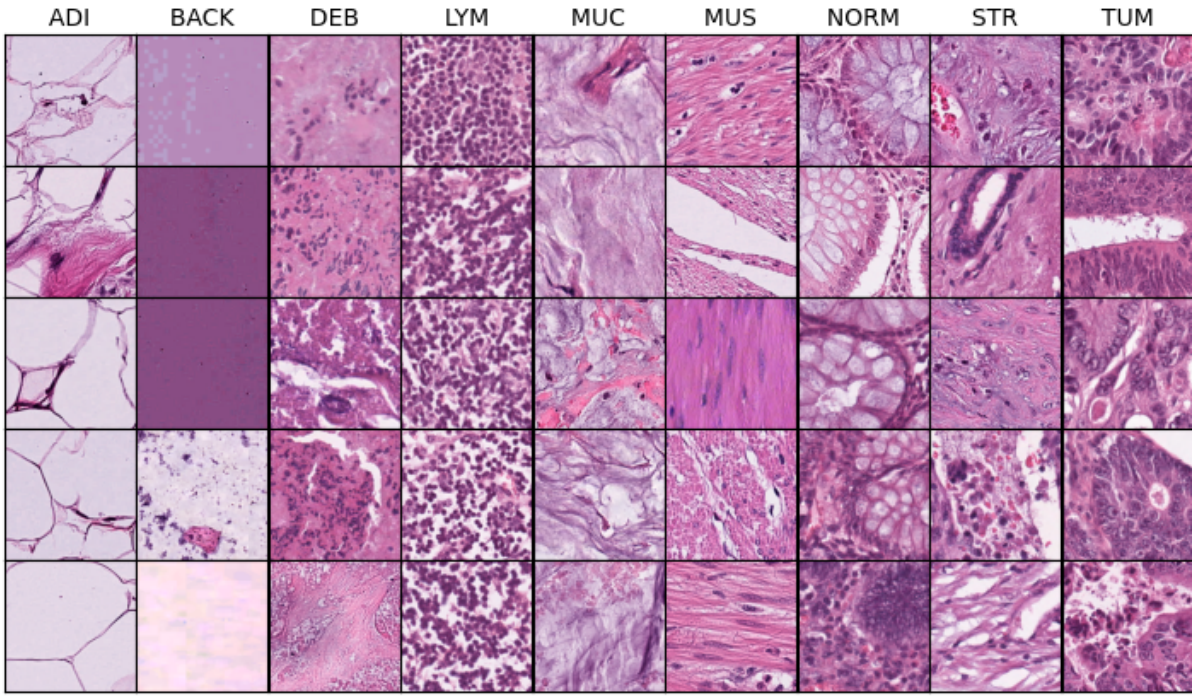

Рис. 2. Примеры изображений из набора NCT-CRC-HE-100K [20, 21]

.

Приведенный набор разбит на тренировочную выборку, содержащую 100000 изображений из набора NCT-CRC-HE-100K, и скрытую, содержащую 7180 изображений из набора CRC-VAL-HE-7K.

Методы

Алгоритм аугментации, использующийся при обучении, состоит из следующих этапов:

1. Стандартная аугментация входного гистологического изображения, состоящая из этапов:
   - Горизонтальное отражение изображения с вероятностью 0.5.
   - Случайные небольшие изменения (95% - 105%) яркости, контраста, насыщенности и оттенка изображения.
   - Гауссово размытие с размером ядра 3 и среднеквадратическим отклонением 0.5.
   - Нормализация.
2. Применение алгоритма FGSM с вероятностью 0.5 с фиксированным коэффициентом ε.

Гауссово размытие и нормализация изображений выполняются всегда. На этапе тестирования некоторого обученного классификатора аугментация входных изображений имеет следующий вид:

1. Нормализация.
2. Применение алгоритма FGSM с фиксированным коэффициентом ε'.

Метод генерации адверсативных атак FGSM является *white-box* методом, то есть, помимо изображения, на которое необходимо наложить адверсативное возмущение, ему на вход передается и предобученная нейросеть, с помощью которой это возмущения вычисляется. При обучении и тестировании одного нейросетевого классификатора алгоритм FGSM применялся с использованием нейросети такой же архитектуры, обученной без адверсативной аугментации.

Архитектура нейронной сети, функция потерь и Fast Gradient Sign Method

Для классификации гистологических изображений в данной работе используются архитектуры ResNet50 [22] и EfficientNetB2 [23].

На вход нейронной сети подается тензор размерности $batchsize \times 3 \times 224 \times 224$, где batchsize – количество входных изображений (размер пакета). Выход нейронной сети представляет собой тензор размера $batchsize \times n_{out}$, где n_{out} – число каналов, и каждому каналу соответствует распределение вероятностей соответствующего класса для каждого исходного изображения.

В качестве функции потерь для обучения нейронной сети используется кросс-энтропия:

$$E = \sum_{\mathbf{x} \in \Omega} \log \left(p_{\ell(\mathbf{x})}(\mathbf{x}) \right),$$

где $\mathbf{x}$ – изображение обучающей выборки $\mathbf{x} \in \Omega$, $\Omega \subset \mathbb{Z}^{224\times224\times3}$, $\ell : \Omega \to \{1,\ldots,K\}$ – истинная метка каждого пикселя, K – количество классов (в данной работе K равно 9), $p_{\ell(\mathbf{x})}$ соответсвует вероятности правильного класса для данного пикселя и вычисляется как soft-max по всем каналам выхода нейронной сети:

$$p_k(\mathbf{x}) = \exp\left(a_k(\mathbf{x})\right) / \left(\sum_{k'=1}^{K} \exp\left(a_{k'}(\mathbf{x})\right)\right),$$

где $a_k(\mathbf{x})$ – выход нейронной сети для канала k в позиции $\mathbf{x}$.

Алгоритм FGSM работает по следущей формуле:

$$\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \cdot \mathrm{sign}\nabla_{\mathbf{x}} E(C, \mathbf{x}, \ell),$$

где ε - гиперпараметр, имеющий смысл размера адверсативного возмущения, C - предобученная нейросеть, с помощью которой вычисляется адверсативное возмущение, $\tilde{\mathbf{x}}$ - адверсативное изображение.

Обучение нейронных сетей

Используемые нейронные сети были программно реализованы на языке Python 3.9 с использованием библиотеки PyTorch. Обучение моделей велось на графическом ускорителе NVIDIA RTX A6000 (48Gb). В качестве оптимизатора был выбран Adam с начальным значением скорости обучения 0.001. На вход нейронной сети подавался пакет из 256 гистологических изображений размером $224 \times 224 \times 3$. Сеть обучалась на протяжении около 150 эпох, каждая эпоха включала 504 итерации. Точное количество эпох зависело от скорости обучения. В конце каждой эпохи, если за предыдущие 10 эпох среднее по пакетам значение функции потерь не понизилось, скорость обучения умножалась на коэффициент 0.1. При снижении скорости обучения до значения 0.000001 обучение прекращалось.

Результаты

Было проведено 10 итераций обучения нейросетей ResNet50 и EfficientNetB2 на пяти различных значениях коэффициента ε: 0.0, 0.0125, 0.025, 0.05 и 0.075. На Рис. 3 показан эффект переобучения нейросетевого классификатора, при котором точность на адверсативных примерах с коэффициентом $\varepsilon' = 0.025$ выше, чем на чистых изображениях.

Выбор значения ε обусловлен тем, что при меньших значениях адверсативное возмущение, накладываемое на гистологическое изображение, становится более заметным невооруженным глазом. На Рис. 4 показаны примеры адверсативных атак, совершенных методом FGSM, с различными значениями параметра ε.

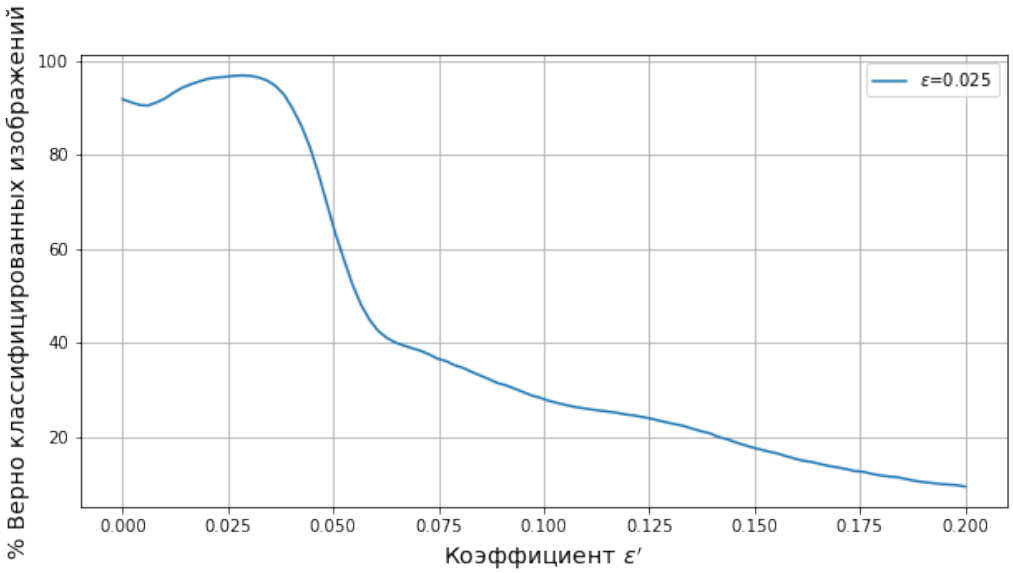

Рис. 3. Результаты тестирования классификатора ResNet50, обученного с коэффициентом $\varepsilon = 0.025$ в адверсативной аугментации, на скрытой выборке CRC-VAL-HE-7K. На графике показана зависимость точности предсказаний классификатора от коэффициента ε', с которым был применен алгоритм FGSM к каждому изображению скрытой выборки.

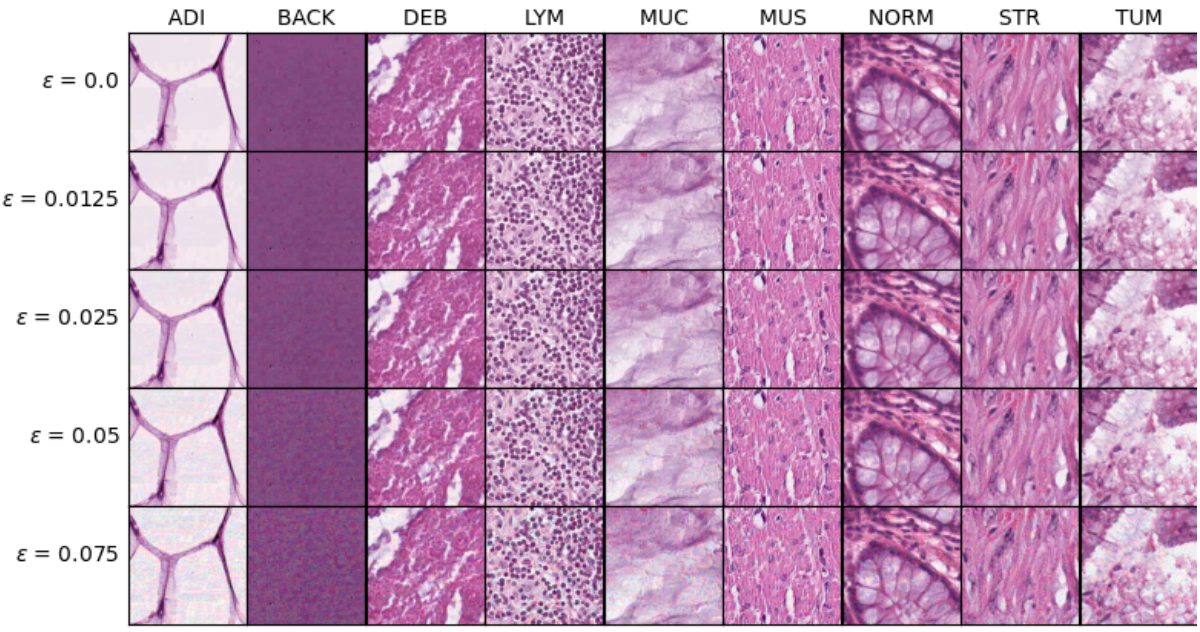

Рис. 4. Примеры адверсативных атак FGSM с различными коэффициентами ε на наборе NCT-CRC-HE-100K.

Эффект переобучения на адверсативных изображениях в той или иной степени присутствует при любом выборе коэффициента ε при обучении с адверсативной аугментацией. Выбор значения гиперпараметра

ε около 0.05 помогает уменьшить проявление такого эффекта на скрытой выборке, при этом повышает точность классификации на чистых данных. Результаты тестирования со всеми коэффициентами ε, с которыми проводилось обучение классификатора ResNet50, показаны на Рис. 5.

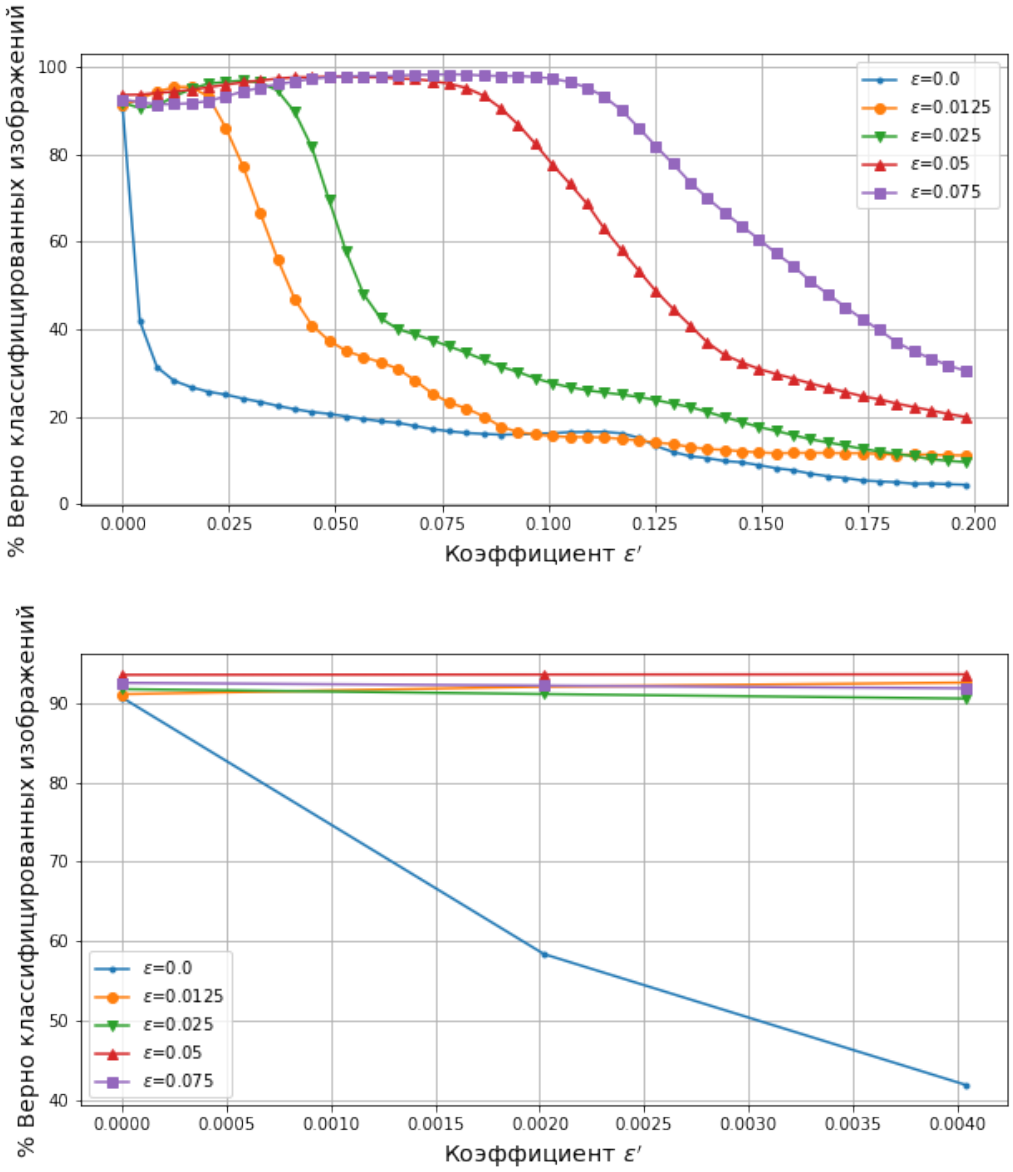

Рис. 5. Результаты тестирования классификатора ResNet50, обученного с указанным коэффициентом ε в адверсативной аугментации, на скрытой выборке CRC-VAL-HE-7K. На верхнем рисунке указаны результаты для $\varepsilon \in [0, 0.2]$, на нижнем - для $\varepsilon \in [0, 0.004]$.

Численные результаты тестирования на чистой скрытой выборке приведены в таблице 1.

	Коэффициент ε в адверсативной аугментации при обучении				
	$\varepsilon = 0$	$\varepsilon = 0.0125$	$\varepsilon = 0.025$	$\varepsilon = 0.05$	$\varepsilon = 0.07$
ResNet50	90.65 ±0.88	91.13 ±0.86	91.76 ±0.83	**93.56** ±0.74	92.55 ±0.79
EfficientNetB2	84.74 ±1.09	86.8 ±1.02	89.76 ±0.92	91.01 ±0.86	**93.15** ±0.76

Табл. 1. Результаты тестирования нейросетевых классификаторов, обученных с указанным ε, на чистой скрытой выборке CRC-VAL-HE-7K. Значения соответствуют точке $\varepsilon' = 0$ на Рис. 5.

Таким образом, основываясь на результатах Табл. 1, показана возможность выбора такого значения параметра адверсативного возмущения при осуществлении FGSM атаки, что для меньших значений параметра при включении атакованных изображений в обучающую выборку наравне с исходными, точность работы на тестовой выборке повышается по сравнению с исходным классификатором.

Заключение

В данной работе был предложен и реализован метод аугментации обучающей выборки адверсативными атаками для обучения нейросетевых классификаторов. Была показана возможность выбора такого значения параметра ε в адверсативной аугментации, использовавшегося во время обучения, что эффект переобучения нейросетевого классификатора на адверсативные возмущения становится менее выраженным, а также растет его точность на чистой скрытой выборке. Технология аугментации наборов изображений за счет использования результатов адверсативных атак, при которых точность работы на тестовой невозмущенной выборке повышается по сравнению с исходным классификатором, впервые разработанная в данной работе, может быть использована для широкого класса практических задач глубокого обучения.

Благодарности

Работа выполнена за счет гранта Российского научного фонда №22-21-00081.

Литература

1. *Goodfellow Ian J, Shlens Jonathon, Szegedy Christian*. Explaining and harnessing adversarial examples // *arXiv preprint arXiv:1412.6572*. — 2014.
2. *Su Jiawei, Vargas Danilo Vasconcellos, Sakurai Kouichi*. One pixel attack for fooling deep neural networks // *IEEE Transactions on Evolutionary Computation*. — 2019. — Vol. 23, no. 5. — Pp. 828–841.
3. *Moosavi-Dezfooli Seyed-Mohsen, Fawzi Alhussein, Frossard Pascal*. Deepfool: a simple and accurate method to fool deep neural networks //

Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 2574–2582.

4. Towards deep learning models resistant to adversarial attacks / Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt et al. // *arXiv preprint arXiv:1706.06083*. — 2017.
5. *Wang Xiaosen, Zhang Zeliang, Zhang Jianping*. Structure invariant transformation for better adversarial transferability // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2023. — Pp. 4607–4619.
6. Universal and transferable adversarial attacks on aligned language models / Andy Zou, Zifan Wang, J Zico Kolter, Matt Fredrikson // *arXiv preprint arXiv:2307.15043*. — 2023.
7. CT-GAT: Cross-Task Generative Adversarial Attack based on Transferability / Minxuan Lv, Chengwei Dai, Kun Li et al. // *arXiv preprint arXiv:2310.14265*. — 2023.
8. *Özdenizci Ozan, Legenstein Robert*. Adversarially Robust Spiking Neural Networks Through Conversion // *arXiv preprint arXiv:2311.09266*. — 2023.
9. *Jeong Jongheon, Shin Jinwoo*. Multi-scale Diffusion Denoised Smoothing // *arXiv preprint arXiv:2310.16779*. — 2023.
10. *Puttagunta Murali Krishna, Ravi S, Nelson Kennedy Babu C*. Adversarial examples: attacks and defences on medical deep learning systems // *Multimedia Tools and Applications*. — 2023. — Pp. 1–37.
11. Distillation as a defense to adversarial perturbations against deep neural networks / Nicolas Papernot, Patrick McDaniel, Xi Wu et al. // 2016 IEEE Symposium on Security and Privacy (SP) / IEEE. — 2016. — Pp. 582–597.
12. *Kurakin Alexey, Goodfellow Ian J., Bengio Samy*. Adversarial Machine Learning at Scale // *CoRR*. — 2016. — Vol. abs/1611.01236. — URL: `http://arxiv.org/abs/1611.01236`.
13. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization / Dan Hendrycks, Steven Basart, Norman Mu et al. // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). — 2021. — October. — Pp. 8340–8349.
14. Unlabeled data improves adversarial robustness / Yair Carmon, Aditi Raghunathan, Ludwig Schmidt et al. // *Advances in neural information processing systems*. — 2019. — Vol. 32.
15. Augmix: A simple data processing method to improve robustness and uncertainty / Dan Hendrycks, Norman Mu, Ekin D Cubuk et al. // *arXiv preprint arXiv:1912.02781*. — 2019.

16. *Rice Leslie, Wong Eric, Kolter Zico*. Overfitting in adversarially robust deep learning // Proceedings of the 37th International Conference on Machine Learning / Ed. by Hal Daumé III, Aarti Singh. — Vol. 119 of *Proceedings of Machine Learning Research*. — PMLR, 2020. — 13–18 Jul. — Pp. 8093–8104. — URL: https://proceedings.mlr.press/v119/rice20a.html.

17. Land-use and land-cover classification using a human group-based particle swarm optimization algorithm with an LSTM Classifier on hybrid pre-processing remote-sensing images / Ganesh B Rajendran, Uma M Kumarasamy, Chiara Zarro et al. // *Remote Sensing*. — 2020. — Vol. 12, no. 24. — P. 4135.

18. *Nasonov Andrey, Chesnakov Konstantin, Krylov Andrey*. CNN based retinal image upscaling using zero component analysis // *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. — 2017. — Vol. 42. — Pp. 27–31.

19. Adversarial attack driven data augmentation for accurate and robust medical image segmentation / Mst Pervin, Linmi Tao, Aminul Huq et al. // *arXiv preprint arXiv:2105.12106*. — 2021.

20. *Kather Jakob Nikolas, Halama Niels, Marx Alexander*. 100,000 histological images of human colorectal cancer and healthy tissue // *Zenodo10*. — 2018. — Vol. 5281.

21. Tissue type recognition in whole slide histological images / Alexander Khvostikov, Andrey Krylov, Ilya Mikhailov et al. // CEUR Workshop Proceedings. — Vol. 3027. — 2021. — Pp. 496–507.

22. Deep residual learning for image recognition / Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — Pp. 770–778.

23. *Tan Mingxing, Le Quoc V.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // *CoRR*. — 2019. — Vol. abs/1905.11946. — URL: http://arxiv.org/abs/1905.11946.