ПРОГРАММИРОВАНИЕ, 2023, № 3, с. 65–70

АНАЛИЗ ДАННЫХ

УДК 519.6+004.891.3+616-018

АУГМЕНТАЦИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ ГИСТОЛОГИЧЕСКИХ ИЗОБРАЖЕНИЙ АДВЕРСАТИВНЫМИ АТАКАМИ

© 2023 г. Н. Д. Локшин[a,*] (ORCID: 0000-0001-7777-7035),
А. В. Хвостиков[a,] (ORCID: 0000-0002-4217-7141),**
А. С. Крылов[a,*] (ORCID: 0000-0001-9910-4501)**

[a] Московский государственный университет имени М.В. Ломоносова, 119991 Москва, ГСП-1, Ленинские горы, д. 1, Россия

**E-mail: lockshin1999@mail.ru*

***E-mail: khvostikov@cs.msu.ru*

****E-mail: kryl@cs.msu.ru*

Поступила в редакцию 09.01.2023 г.
После доработки 15.01.2023 г.
Принята к публикации 20.01.2023 г.

В работе рассматривается задача аугментации выборки гистологических изображений адверсативными атаками для повышения устойчивости нейросетевых классификаторов, обученных на аугментированной выборке, к адверсативным атакам. В последние годы нейросетевые методы стремительно развивались, новые нейросетевые методы показывают впечатляющие результаты, однако они подвергаются так называемым адверсативным атакам – то есть совершают неверные предсказания на входах, получающихся в результате наложения на изображение малого шума. Из-за этого надежность нейросетевых методов до сих пор является актуальной областью изучения. В этой статье мы представляем и сравниваем между собой различные методы аугментации обучающей выборки, позволяющие повысить устойчивость нейросетевых классификаторов гистологических изображений к адверсативным атакам. Для этого мы предлагаем добавлять в обучающую выборку адверсативные атаки, полученные несколькими актуальными методами.

DOI: 10.31857/S0132347423030020, **EDN:** DEEBYB

1. ВВЕДЕНИЕ

Некоторые модели машинного обучения, в частности нейронные сети, часто подвергаются *адверсативным атакам* – то есть неверно классифицируют входы, получающиеся в результате наложения на изображение малого шума [1–4] (рис. 1). На данный момент имеется большое количество способов генерации таких атак, а также и методов защиты от них [1, 7, 8]. Большинство способов защиты от адверсативных атак либо предусматривают изменение структуры исходной модели предсказания, например, *защитная дистилляция* нейросетей [8], либо строят предположения о возможных атаках.

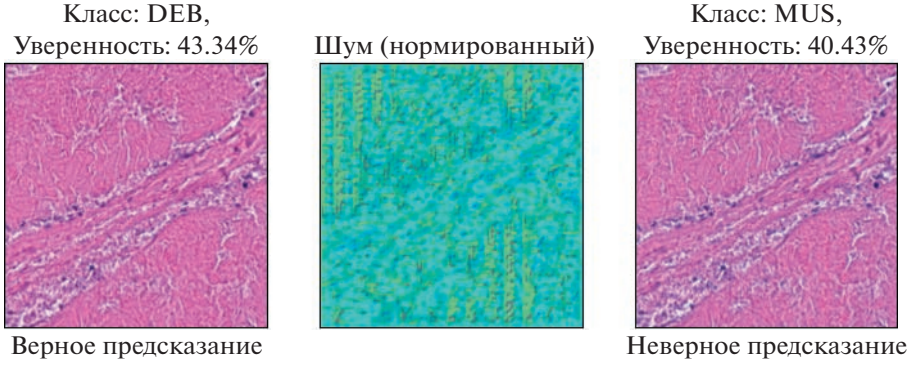

Рис. 1. Пример адверсативной атаки на модель ResNet50 [5] на изображение из набора данных NCT-CRC-HE-100K [6].

Самым эффективным методом генерации адверсативных возмущений на данный момент является метод *AdvGAN* [9]. Основные преимущества AdvGAN перед другими способами генерации адверсативных атак заключаются в более низкой степени искажения исходных изображений при совершении атаки и более высокой вероятности совершить атаку на изображения успешно. Данный метод заключается в *генеративно-состязательном* [10] обучении нейросети, которая после этапа обучения способна генерировать адверсативные возмущения по переданным на вход изображениям.

В данной работе на основе [1, 3, 9, 11, 12] предлагается метод повышения устойчивости нейросетевых классификаторов гистологических изображений к адверсативным атакам.

2. АДВЕРСАТИВНЫЕ АТАКИ

В данной работе рассматривается 5 алгоритмов генерации адверсативных атак: *Fast Gradient Sign Method* (FGSM) [1], *Carlini and Wagner* (C&W) [3], *AdvGan* [9], *Projected Gradient Descent* (PGD) [11] и *Decoupled Direction and Norm* (DDN) [13].

Алгоритмы генерации адверсативных атак делятся на White-Box и Black-Box. White-Box-алгоритмы принимают на вход, помимо изображения, по которому необходимо построить адверсативную атаку, обученную нейросеть. Black-Box-алгоритмы принимают на вход только изображение. Все алгоритмы, рассматриваемые в данной статье, являются White-Box.

Fast Gradient Sign Method

Имея на входе изображение I с меткой класса y, функцию потерь J, обученный классификатор C, размер шага ϵ, адверсативное изображение вычисляется следующим образом:

$$\tilde{I} = I + \epsilon \cdot \operatorname{sign}\nabla_I J(C, I, y) \tag{2.1}$$

В данном алгоритме число ϵ является гиперпараметром. Как правило, ϵ выбирается достаточно малым, чтобы накладываемое на изображение возмущение было слабо заметным. В данной работе $\epsilon = 0.07$.

Projected Gradient Descent

Имея на входе изображение I с меткой класса y, функцию потерь J, обученный классификатор C, размер шага α и количество итераций K, для $k = 1 \ldots K$ совершается вычисление:

$$I_{k+1} = \prod_{I+S}(I_k + \alpha \cdot \operatorname{sign}(\nabla_{I_k} J(C, I_k, c))) \tag{2.2}$$

I_0 принимается равным I. На выходе алгоритма имеем адверсативное изображение I_K. В данной работе $\alpha = 3 \times 10^{-3}$, количество шагов равно 100, S – l_∞-сфера вокруг изображения I с радиусом 0.01.

Decoupled Direction and Norm

Имея на входе изображение I с меткой класса y, функцию потерь J, обученный классификатор C, размер шага α, шаг обновления радиуса γ и количество итераций K, для $k = 1 \ldots K$ совершаются следующие действия:

1. Вычисление градиента:

$$g \leftarrow \nabla J_{\tilde{I}_{k-1}}(C, \tilde{I}_{k-1}, y) \tag{2.3}$$

2. Нормирование и умножение на размер шага:

$$g \leftarrow \alpha \frac{g}{\|g\|_2} \tag{2.4}$$

3. Обновление адверсативного возмущения:

$$\sigma_k = \sigma_{k-1} + g \tag{2.5}$$

Возмущение σ_0 принимается равным 0.

4. Проекция адверсативного изображения $\tilde{I}_{k-1} + \sigma_k$ на ϵ_k-сферу вокруг исходного изображения I. ϵ_k берется равным $(1+\gamma)\epsilon_{k-1}$, если $\tilde{I}_{k-1}$ успешно нарушает работу C, либо $(1-\gamma)\epsilon_{k-1}$, если иначе:

$$\tilde{I}_k = \tilde{I}_{k-1} + \epsilon_k \frac{\sigma_k}{\|\sigma_k\|_2} \tag{2.6}$$

5. Клиппирование значений, выходящих за допустимые границы: $\tilde{I}_k$: $\tilde{I}_k \leftarrow clip(\tilde{I}_k, 0, 1)$.

В конце итераций на выход подается такое $\tilde{I}_k$, которое успешно нарушает работу C и имеет наименьшую l_2-норму.

Carlini and Wagner

Имея на входе изображение I с меткой класса y и обученный нейросетевой классификатор C, адверсативная атака выполняется путем решения задачи оптимизации:

$$\begin{gathered}\text{minimize } \|\sigma\|_2 - \epsilon \cdot f(I + \sigma) \\ \text{suchthat } I + \sigma \in [0,1]^n,\end{gathered} \tag{2.7}$$

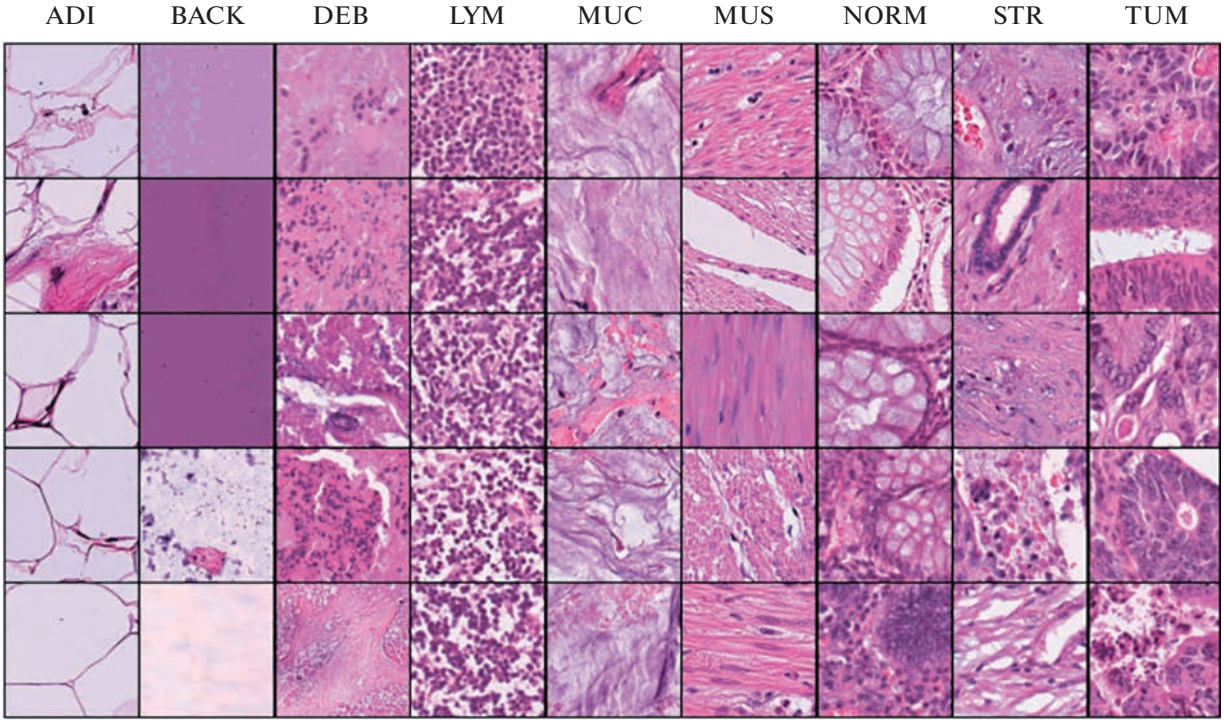

Рис. 2. Примеры изображений из набора NCT-CRC-HE-100K [6, 14].

где функция f подобрана таким образом, что $C(I+\sigma) = c$ тогда и только тогда, когда $f(x+\sigma) \le 0$. В данном эксперименте,

$$f(I+\sigma) = \max\left(\max_{i \neq y}(Z(I+\sigma)_i\right) - Z(I)_y, 0), \quad (2.8)$$

где $Z(I)$ – ненормированный выход классификатора C с принятием на вход I. Данная функция признана самой эффективной из возможных в оригинальной статье. Оптимальное значение константы ϵ таково что оно наименьшее, для которого $f(I+\sigma) \le 0$. Чтобы найти ϵ, совершается 20 шагов бинарного поиска, где на каждом шаге находится решение $I+\sigma$ решением задачи оптимизации 2.7 методом градиентного спуска. В данной работе количество шагов градиентного спуска равно 100 с шагом 0.01.

AdvGan

Данный алгоритм строит адверсативную атаку путем обучения генеративно-состязательной сети (GAN). Во время обучения генератор $G(z)$ минимизирует следующий функционал:

$$l = \gamma l_{nce} + \theta l_{GAN} + \zeta l_{threshold}, \quad (2.9)$$

где l_{nce} – это обратный функционал кросс-энтропии:

$$l_{nce} = \frac{1}{l_{ce}}, \quad (2.10)$$

$$l_{ce} = -\frac{1}{N}\sum_{n=1}^{N} \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{N_c} \exp(x_{n,c})}, \quad (2.11)$$

l_{GAN} – это функционал, оптимизируемый в рамках генеративно-состязательной сети:

$$l_{GAN} = \mathbb{E}_{I \sim p_{data}(I)} \log D(I) + \\ + \mathbb{E}_{I \sim p_{adversarial}(I)} \log(1 - D(I + aG(z))), \quad (2.12)$$

где $G(z)$ имеет на выходе изображение, $D(I)$ имеет на выходе вероятность того, что изображение I – настоящее. $l_{threshold}$ задан следующим образом:

$$l_{threshold} = \sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{C} \max\left(\left|I_{x,y,z}\right| - thr, 0\right)^2, \quad (2.13)$$

где $I_{x,y,z}$ – значения пикселя изображения I в координатах (x, y, z). В данном эксперименте значение thr равно 0.25. Коэффициенты γ, θ и ζ были эмпирически подобраны и равны 10, 1, 1 соответственно.

Таблица 1. Результаты тестирования классификаторов, имеющих структуру ResNet50. В левой колонке приведены обучающие выборки: 18K – исходная обучающая выборка, 36K – исходная обучающая выборка, аугментированная 18000 синтетическими изображениями, сгенерированными StyleGan2, 72K – набор 36K, аугментированный $AdvGan_{a=1}$, 216K – набор 36K, аугментированный $AdvGan_{a=1}$, FGSM, DDN, PGD, C&W. Названия алгоритмов генерации адверсативных атак в столбцах обозначают адверсативную атаку, примененную к каждому изображению в тестовой выборке

Обуч. выборка	Точность на тестовой выборке с применением различных видов атак							
	No attack	FGSM	DDN	PGD	C&W	$AdvGan_{a=0.5}$	$AdvGan_{a=1}$	$AdvGan_{a=2}$
18K	99.49	19.43	0	0.01	0	96.22	69.57	44.60
36K	99.37	36.69	97.76	1.25	95.60	95.99	72.73	46.64
72K	99.23	53.24	98.30	8.85	97.73	99.36	99.32	95.36
216K	98.89	84.16	98.61	72.33	98.46	98.78	98.47	94.82

3. ПОСТАНОВКА ЗАДАЧИ

В качестве входных данных имеется 22500 размеченных изображений – подмножество набора *NCT-CRC-HE-100K* [6, 4]. Изображения являются непересекающимися участками больших гистологических изображений, содержащих злокачественные новообразования толстого кишечника а также здоровые ткани, окрашенные гематоксилин-эозином. Подмножество набора данных выбрано для уменьшения временных затрат на эксперименты и обучение нейросетевых методов. Изображения имеют размерности $224 \times 224 \times 3$. 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.

Приведенный набор разбит на тренировочную выборку, содержащую 18000 изображений, и тестовую, содержащую 4500 изображений. Требуется исследовать и реализовать методы аугментации данного набора данных различными адверсативными атаками и сравнить качество различных нейросетевых классификаторов на скрытых выборках, также аугментированных адверсативными атаками. Термин “скрытая” выборка означает, что изображения из этой выборки не принадлежат обучающей выборке.

4. ПРЕДЛАГАЕМЫЙ МЕТОД

В данной работе предлагается рассматривать несколько наборов данных для обучения, аугментированных различными наборами адверсативных атак, а именно:

• Исходная обучающая выборка, состоящая из 18000 гистологических изображений;

• Обучающая выборка, состоящая из 18000 исходных гистологических изображений, и 18000 синтетических изображений, сгенерированных хорошо известным алгоритмом StyleGAN2 – всего 36000 изображений;

• 36000 изображений из предыдущего пункта, аугментированных 36000 адверсативными атаками AdvGan (т.е применение AdvGan к каждому из 36000 изображений в обучающей выборке с последующим добавлением результата в обучающую выборку) – всего 72000 изображений;

• 216000 изображений в результате аугментации алгоритмами DDN, C&W, FGSM, PGD и AdvGan. Каждый из перечисленных алгоритмов был применен к каждому из 36000 изображений с последующим добавлением результата в обучающую выборку.

На каждой обучающей выборке был обучен нейросетевой классификатор имеющий архитектуру *ResNet50* [5], в итоговом сравнении участвует 4 нейросетевых классификатора, имеющих оди-

наковую структуру, но обученных на разных наборах данных. Для тестирования используются следующие скрытые наборы данных:

• Исходная тестовая выборка, состоящая из 4500 гистологических изображений;

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением FGSM [1] с параметром $\epsilon = 0.1$;

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением C&W [3];

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением PGD [11];

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением DDN [13];

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением AdvGan с параметром $a = 0.5$;

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением AdvGan с параметром $a = 1$;

• Тестовая выборка, состоящая из 4500 исходных гистологических изображений с применением AdvGan с параметром $a = 2$.

5. АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Результаты экспериментов приведены в табл. 1. Предложенный метод аугментации данных существенно повысил устойчивость нейросетевого классификатора ResNet50 к адверсативным атакам. Тем не менее, даже при увеличении объема обучающей выборки в 12 раз вышеприведенным методом точность классификатора на скрытой выборке без применения к ней адверсативных атак изменяется несущественно.

6. ЗАКЛЮЧЕНИЕ

В данной работе был предложен и реализован метод аугментации обучающей выборки адверсативными атаками для обучения нейросетевых классификаторов. В случае, когда архитектуры классификатора, использовавшегося для генерации адверсативных атак, и тестового классификатора совпадали, было продемонстрировано повышение устойчивости против ряда адверсативных атак. Для обучения моделей использовался вычислительный кластер с 4x Nvidia RTX A6000 факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

БЛАГОДАРНОСТИ

Исследования выполнены при финансовой поддержке Российского научного фонда в рамках научного проекта 22-21-00081.

СПИСОК ЛИТЕРАТУРЫ

1. *Goodfellow I.J., Shlens J., Szegedy Ch.* 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. № 5. С. 828–841.
3. *Carlini N., Wagner D.* Towards evaluating the robustness of neural networks // IEEE Symposium on Security and Privacy (SP). 2017. P. 39–57.
4. *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. P. 2574–2582.
5. *He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian.* Deep residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 770–778.
6. *Kather J.N., Halama N., Marx A.* 100000 histological images of human colorectal cancer and healthy tissue // Zenodo10, 2018.
7. *Liang Bin, Li Hongcheng, Su Miaoqiang, Li Xirong, Shi Wenchang, Wang Xiaofeng.* Detecting adversarial image examples in deep neural networks with adaptive noise reduction // IEEE Transactions on Dependable and Secure Computing. 2018. № 1. P. 72–85.
8. *Papernot N., McDaniel P., Wu Xi, Jha Somesh, Swami Ananthram.* Distillation as a defense to adversarial perturbations against deep neural networks // IEEE Symposium on Security and Privacy (SP), 2016. P. 582–597.
9. *Xiao Chaowei, Li Bo, Zhu Jun-Yan, He Warren, Liu Mingyan, Song Dawn.* Generating adversarial examples with adversarial networks // arXiv preprint arXiv:1801.02610, 2018.
10. *Goodfellow I., Pouget-Abadie J., Mirza Mehdi, Xu Bing, Warde-Farley D., Ozair Sherjil, Courville A., Bengio Yoshua.* Generative adversarial nets // Advances in Neural Information Processing Systems, 2014.
11. *Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards deep learning models resistant to adversarial attacks // arXiv preprint arXiv:1706.06083, 2017.

12. *Karras Tero, Laine Samuli, Aittala Miika, Hellsten Janne, Lehtinen Jaakko, Aila Timo.* Analyzing and improving the image quality of stylegan // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. P. 8110–8119.

13. *Rony J., Hafemann L.G., Oliveira L.S., Ayed Ismail Ben, Sabourin R., Granger E.* Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. P. 4322–4330.

14. *Khvostikov A., Krylov A., Mikhailov I., Malkov P., Danilova N.* Tissue Type Recognition in Whole Slide Histological Images. 2021.