Разработка метода защиты нейронных сетей от состязательных атак на основе adversarial training с использованием множества алгоритмов генерации атакованных изображений

Д. Д. Скалдин
Университет «Дубна»
danya.skaldin@mail.ru

И. С. Соколов
«Университет «Дубна»
ilya20834@gmail.com

А. В. Шевченко
Университет «Дубна»
leviathan0909@gmail.com

Ю. В. Трофимов
Университет «Дубна»
ura_trofimov@bk.ru

А. Н. Аверкин
Университет «Дубна»
averkin2003@inbox.ru

Аннотация. **В работе рассматривается проблема устойчивости нейронных сетей к состязательным атакам. Предлагается метод повышения устойчивости моделей машинного обучения на основе комбинированного adversarial training с использованием множества алгоритмов генерации атакованных данных. В рамках исследования реализован программный модуль формирования обучающего набора данных, включающий примеры, сформированные алгоритмами FGSM, BIM, JSMA и Carlini & Wagner. Проведено экспериментальное исследование влияния различных комбинаций атакованных данных и их долевого распределения в обучающем наборе на устойчивость нейронной сети к состязательным воздействиям. Полученные результаты позволяют определить наиболее эффективные комбинации методов генерации атакованных примеров для повышения устойчивости моделей машинного обучения при сохранении точности классификации на исходных данных.**

Ключевые слова: состязательные атаки; adversarial training; нейронные сети; FGSM; BIM; JSMA; Carlini & Wagner; робастность; глубокое обучение

I. Введение

Нейронные сети глубокого обучения демонстрируют высокую эффективность в задачах классификации изображений, однако обладают критической уязвимостью к состязательным (adversarial) атакам. Состязательные примеры представляют собой входные данные, модифицированные путём добавления малых, зачастую визуально неразличимых возмущений, которые приводят к ошибочной классификации [1]. Данная проблема особенно актуальна для систем автономного управления транспортом, медицинской диагностики и биометрической аутентификации.

Одним из наиболее эффективных подходов к повышению устойчивости является adversarial training – включение атакованных примеров в обучающую выборку [2]. Существующие работы, как правило, используют единственный алгоритм генерации возмущений, что ограничивает спектр атак, к которым модель приобретает устойчивость [3]. Различные алгоритмы оперируют в разных нормах возмущений и используют принципиально различные стратегии, что создаёт необходимость комбинированного подхода.

Целью работы является разработка и экспериментальное исследование метода adversarial training, основанного на формировании обучающего набора из примеров, сгенерированных множеством алгоритмов атак (FGSM, BIM, JSMA, C&W), с различными комбинациями и долевыми распределениями.

II. Математическое описание алгоритмов атак

A. FGSM

Метод FGSM [1] генерирует состязательный пример за один шаг в направлении знака градиента функции потерь (рис. 1):

$$x_{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x L(\theta, x, y)), \quad (1)$$

где x – исходное изображение, ε – величина возмущения, L – функция потерь модели с параметрами θ.

B. BIM

BIM [4] – итеративное расширение FGSM (рис. 2):

$$x_{t+1} = \text{Clip}_{x,\varepsilon} \{ x_t + \alpha \cdot \text{sign}(\nabla_x L(\theta, x_t, y)) \}, \quad (2)$$

где α – размер шага итерации, Clip – проекция на ε–окрестность в норме $L\infty$.

C. JSMA

JSMA [5] оперирует в норме L_0, модифицируя минимальное число пикселей (рис. 3). Карта значимости:

$$S(x, t)[i] = \partial F_t/\partial x_i \cdot |\sum_{j \neq t} \partial F_j/\partial x_i|, \quad (3)$$

где F_t – выход для целевого класса. Пиксель с максимальным S модифицируется до достижения ошибочной классификации или исчерпания бюджета γ.

Работа выполнена при поддержке государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2)

D. *Carlini & Wagner*

C&W [6] минимизирует норму возмущения при условии ошибочной классификации (Рис. 4):

$$\text{minimize } \|\delta\|^2 + c \cdot f(x + \delta), \tag{4}$$

где $f(x') = \max(Z(x')_y - \max_{i \neq y} Z(x')_i, -\kappa)$, Z – логиты модели, κ – параметр уверенности, c – балансирующая константа. Замена переменных через tanh обеспечивает ограничение $x + \delta \in [0, 1]$.

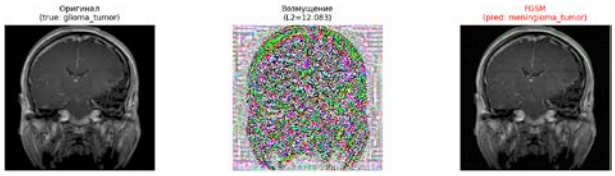

Рис. 1. Зашумление FGSM

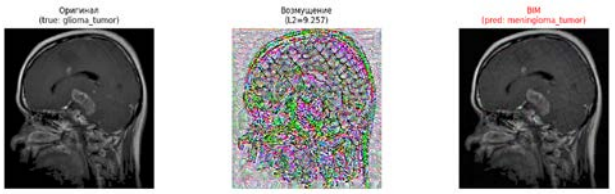

Рис. 2. Зашумление BIM

Рис. 3. Зашумление JSMA

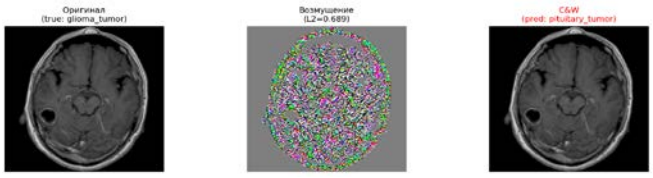

Рис. 4. Зашумление C&W

III. Предлагаемый метод

Метод основан на формировании обучающего набора из чистых и атакованных примеров. Пусть D – исходный набор, $A = \{A_1, \ldots, A_k\}$ – множество алгоритмов, $w = (w_1, \ldots, w_k)$ – вектор весов ($\sum w_j = 1$). Для каждого батча B формируется смешанный батч:

$$B' = (1 - r) \cdot B \cup r \cdot (\biguplus_j w_j \cdot A_j(B)), \tag{5}$$

где $r \in [0, 1]$ – доля adversarial–примеров. Атакованные примеры генерируются на лету на основе текущего состояния модели. Исследованы 10 комбинаций: baseline, каждая из четырёх атак, попарные комбинации, полная комбинация с равными и неравными весами. Архитектура – ResNet–18 (предобученная на ImageNet).

IV. Результаты экспериментов

Обучение выполнялось на протяжении 30 эпох с применением оптимизатора Adam при скорости обучения 0,001 и косинусной схемы её изменения; параметр r был установлен равным 0,5. Для атаки FGSM использовалось значение $\varepsilon = 0{,}03$. В случае BIM сохранялось то же ε, шаг выбирался равным 0,007 и проводилось 10 итераций. Для JSMA брались параметры $\theta = 1{,}0$ и $\gamma = 0{,}1$. В методе C&W использовалось $\kappa = 0$ и выполнялось 200 итераций. Результаты приведены в табл. 1.

ТАБЛИЦА I. Точность классификации (%) при различных атаках

Комбинация	Clean	FGSM 0,03	FGSM 0,05	BIM 0,03	BIM 0,05	JSMA	C&W
Baseline	94,2	31,5	18,7	22,4	11,3	42,8	15,6
FGSM	91,8	68,3	52,1	55,7	38,2	48,5	29,3
BIM	90,5	62,4	48,6	64,8	46,1	47,2	31,8
FGSM+BIM	89,3	71,5	56,8	67,2	48,5	50,1	34,6
FGSM+BIM+JSMA	88,6	69,8	54,2	65,1	45,7	58,3	36,2
Все атаки (=)	87,4	70,2	55,6	66,8	47,3	59,7	48,5
Акцент FGSM	88,1	72,4	57,3	63,5	44,1	55,8	38,7
Акцент C&W	87,9	64,8	50,1	60,2	42,6	56,4	54,1

У базовой модели точность заметно проседает под атаками – с 94,2% до 31,5% при FGSM и до 15,6% при C&W. Обучение на одной атаке ожидаемо «подтягивает» устойчивость именно к ней, но на другие сценарии почти не переносится, например, 68,3% на FGSM сочетаются лишь с 29,3% на C&W. Когда атаки комбинируются, картина становится более ровной, и общая робастность растёт. При использовании полной комбинации среднее значение достигает 58,0%, хотя это сопровождается потерей точности на 6,8 п.п. Наиболее сбалансированным оказался вариант с FGSM и BIM, где прирост робастности составляет 31,1 п.п. при снижении точности на 4,9 п.п. Если же делать упор на C&W, можно получить 54,1% на этой атаке, однако по совокупной робастности такой подход всё же уступает более равномерной комбинации, что видно по данным табл. 2.

ТАБЛИЦА II. Сводные показатели робастности

Комбинация	Clean (%)	Средн. роб. (%)	Δ роб. от baseline (п.п.)	Потеря clean (п.п.)
Baseline	94.2	23.7	–	–
FGSM	91.8	48.7	25	−2.4
BIM	90.5	50.2	26.5	−3.7
FGSM+BIM	89.3	54.8	31.1	−4.9
Все атаки (=)	87.4	58	34.3	−6.8
Акцент C&W	87.9	54.7	31	−6.3

A. *Анализ процесса обучения*

Динамика обучения моделей с разными вариантами adversarial training показана на рис. 5. Слева приведены значения функции потерь, где сплошной линией отмечен test, а пунктиром – train; справа отображена точность на

тестовой выборке по эпохам. На старте у всех моделей с adversarial training значение loss выше, чем у baseline, но примерно к 15–20-й эпохе различия практически исчезают. Наиболее ровно ведёт себя модель cw_heavy, тогда как fgsm_bim и all_attacks в начале обучения колеблются сильнее, что, по–видимому, связано с более разнообразными возмущениями в обучающих данных. Тестовая точность у всех таких моделей выходит на плато в пределах 88–92% примерно к 20-й эпохе, и признаков переобучения не наблюдается.

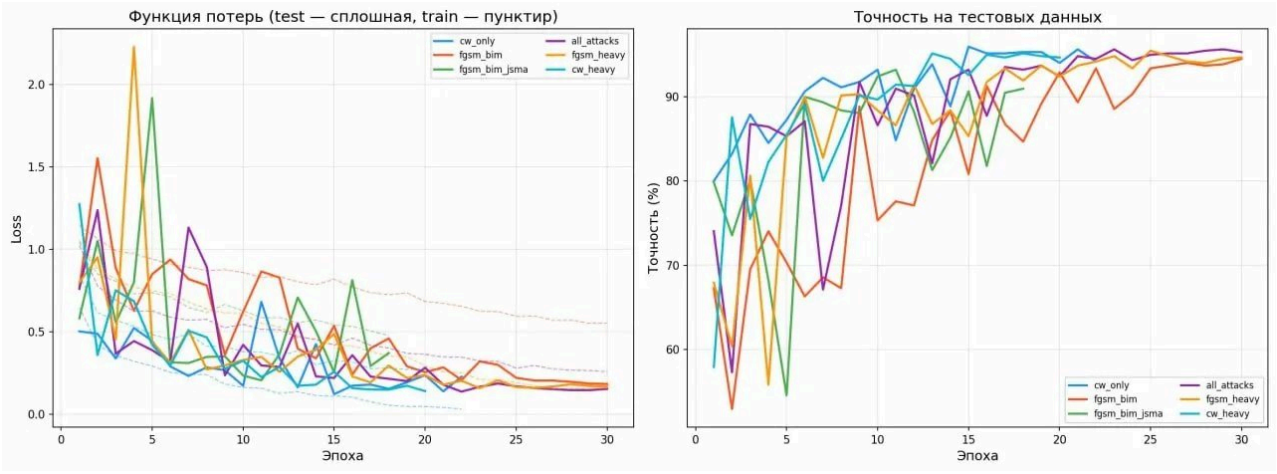

Рис. 5. Динамика обучения моделей с различными комбинациями adversarial training

B. Чувствительность к силе атаки

На рис. 6 и 7 показано, как меняется точность классификации при увеличении силы атак FGSM и BIM, задаваемой параметром ε. По этим зависимостям удобно проследить, насколько быстро модели теряют робастность по мере усиления возмущений.

В случае FGSM, что видно на рисунке 6, при ε = 0,03 лучше всего себя ведут модели fgsm_only и fgsm_heavy – их точность остаётся выше 90%. Модель fgsm_only удерживает лидерство примерно до ε = 0,05, где сохраняется около 80%, что выглядит логичным с учётом её узкой специализации на данном типе атаки. В то же время baseline и jsma_only теряют точность уже при ε = 0,01. Интереснее ведёт себя fgsm_bim. Она не показывает максимума на слабых возмущениях, зато остаётся заметно более устойчивой при их усилении и держит около 55% точности при ε = 0,1. Для сравнения, fgsm_only к этому моменту падает до 28%, что указывает на лучшую обобщающую способность комбинированного подхода при сильных атаках.

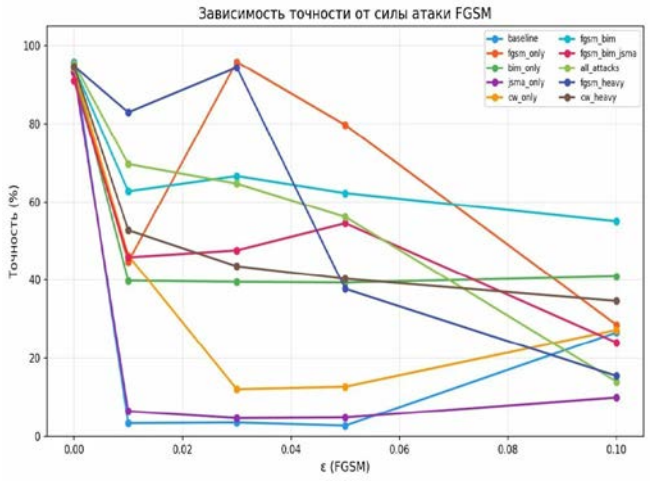

Рис. 6. Зависимость точности от силы атаки FGSM

Для BIM, как видно на рис. 7, точность убывает заметно быстрее по мере роста ε, что ожидаемо для итеративной атаки. Уже при ε = 0,01 наиболее устойчивой оказывается модель fgsm_heavy с результатом около 76%, тогда как jsma_only и cw_only почти сразу теряют работоспособность. При дальнейшем увеличении ε вперёд выходит bim_only и удерживает около 33% при ε = 0,05, что ещё раз подчёркивает пользу обучения, ориентированного именно на такие атаки. При этом baseline и cw_only сильнее других проседают уже на самом первом шаге, при переходе от ε = 0 к ε = 0,01, что указывает на их высокую чувствительность даже к слабым итеративным возмущениям.

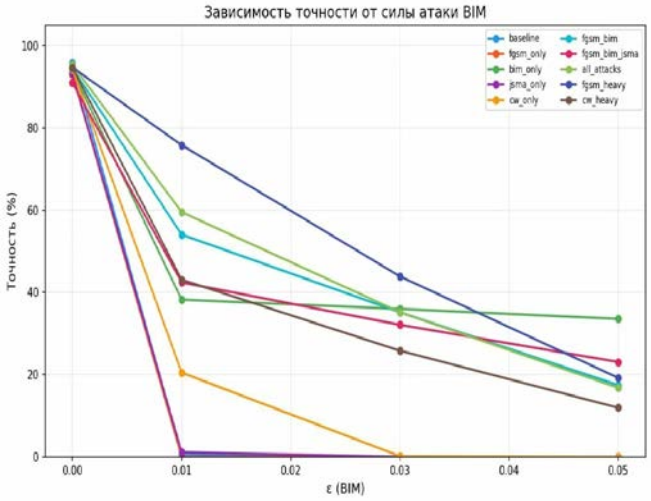

Рис. 7. Зависимость точности от силы атаки BIM

C. Обсуждение результатов

Проведённый анализ позволяет выделить несколько устойчивых тенденций. Прежде всего, довольно чётко проявляется компромисс между точностью и робастностью. По мере роста средней устойчивости примерно на 10 п.п. точность на «чистых» данных обычно снижается примерно на 2 п.п.

Также видно, что атаки, работающие в одной норме, частично «перекрывают» друг друга. В частности, FGSM и BIM, обе относящиеся к $L\infty$, дают заметный эффект переноса устойчивости. Обучение на одной из них повышает устойчивость и к другой. Если же рассматривать атаки разных типов, например $L\infty$, L_2 и L_0, такого эффекта почти нет, и прирост оказывается ограниченным. Это как раз объясняет, почему комбинированные схемы обучения выглядят более предпочтительными.

Отдельно проявляет себя стратегия с усиленным акцентом на конкретной атаке. Она имеет смысл, когда заранее понятно, с каким типом угрозы предстоит работать. Например, fgsm_heavy лучше подходит против быстрых градиентных атак, а cw_heavy – против оптимизационных. Если же такой информации нет, более ровное распределение, как в all_attacks, даёт наиболее сбалансированный результат без явных перекосов.

В контексте задачи классификации опухолей головного мозга по МРТ такие различия уже выходят за рамки теории. Потеря около 6,8 п.п. точности на чистых данных в случае all_attacks компенсируется кратным ростом устойчивости. Для медицинских систем поддержки принятия решений это принципиальный момент, поскольку ошибки, вызванные adversarial–возмущениями, могут напрямую повлиять на диагноз.

V. Заключение

В работе предложен и экспериментально исследован метод комбинированного adversarial training, основанный на формировании обучающего набора из примеров, сгенерированных множеством алгоритмов атак с управляемым долевым распределением. Эксперименты проведены на задаче классификации опухолей головного мозга по МРТ–снимкам (4 класса, 3095 изображений) с архитектурой ResNet–18. Основные результаты:

- комбинация всех атак даёт средн. робастность 58,0% vs 23,7% у baseline при 87,4% clean accuracy;
- одиночная атака не обеспечивает кросс–робастность;
- FGSM+BIM оптимальна при ограниченных ресурсах;
- равномерное распределение весов предпочтительно при неопределённости относительно типа потенциальных воздействий.

Анализ зависимости точности от силы атаки (ε) показал, что комбинированные модели демонстрируют более плавную деградацию робастности по сравнению со специализированными, что свидетельствует о формировании более устойчивых внутренних представлений. Предложенный подход может быть использован при разработке систем поддержки принятия решений в медицинской диагностике, автономном управлении и других областях, где модели искусственного интеллекта функционируют в условиях потенциальных состязательных воздействий. Направления дальнейших исследований включают интеграцию методов certified robustness, адаптивные стратегии управления весами атак в процессе обучения, а также масштабирование подхода на архитектуры трансформеров и более крупные наборы медицинских данных.

Список литературы

[1] Goodfellow I.J., Shlens J., Szegedy C. Explaining and Harnessing Adversarial Examples // Proceedings of ICLR. 2015.

[2] Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks // Proceedings of ICLR. 2018.

[3] Tramèr F., Boneh D. Adversarial Training and Robustness for Multiple Perturbations // NeurIPS. 2019.

[4] Kurakin A., Goodfellow I., Bengio S. Adversarial Examples in the Physical World // ICLR Workshop. 2017.

[5] Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z.B., Swami A. The Limitations of Deep Learning in Adversarial Settings // IEEE EuroS&P. 2016.

[6] Carlini N., Wagner D. Towards Evaluating the Robustness of Neural Networks // IEEE S&P. 2017.

[7] Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing Properties of Neural Networks // ICLR. 2014.

[8] He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // CVPR. 2016.

[9] Dong Y., Liao F., Pang T., Su H., Zhu J., Hu X., Li J. Boosting Adversarial Attacks with Momentum // CVPR. 2018.

[10] Croce F., Hein M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Attacks // ICML. 2020.