УДК 378:004
DOI: 10.25686/978-5-8158-2474-4-2025-811-815

Адверсативная аугментация физически информированных нейронных операторов для задачи одномерного уравнения диффузии

Н. Д. Локшин, А. С. Крылов
Московский государственный университет имени М. В. Ломоносова, Москва, Россия

Аннотация. Известно, что нейронные сети уязвимы к adversarial-атакам – данным с тщательно подобранными возмущениями, которые незаметны человеческому глазу. В задачах медицинской визуализации это может представлять серьёзную угрозу при построении предсказаний на основе глубоких нейронных сетей. В данной статье мы изучаем влияние adversarial-атак на MLP и Fourier Neural Operator (FNO) при решении уравнения диффузии, которое иногда используется в задачах подавления шума в медицинских изображениях. Мы показываем, что при использовании adversarial augmentation ошибка L_2 на adversarial-решениях для возмущённых начальных условий немного улучшается, в то время как ошибка на решениях для чистых начальных условий уменьшается, что указывает на более высокую устойчивость по сравнению с обучением без adversarial augmentation. Дополнительно мы вводим простую, но новую adversarial-атаку, которая эффективна против Fourier Neural Operator – Low Frequency Attack.

***Ключевые слова*:** PINO, Fourier Neural Operator, Adversarial Attacks, Deep Learning, Diffusion Equation, Medical Imaging

Adversarial augmentation in the task of PINO for 1D diffusion equation

N. D. Lockshin, A. S. Krylov
Lomonosov Moscow State University, Moscow, Russia

Abstract. Neural networks are known to be vulnerable against adversarial attacks – data with carefully crafted adversarial perturbations that are imperceptible to the human eye. In medical imaging tasks this can be a major threat for making predictions based on deep neural network solutions. In this paper we study the effect of adversarial attacks on MLP and Fourier Neural Operator (FNO) for solving the diffusion equation, sometimes used in medical image denoising tasks. We show that with adversarial augmentation the L_2 error on the adversarial solutions for perturbed initial conditions marginally improves, whereas the error on the solutions for clean initial conditions decreases, indicating higher robustness than without adversarial augmentation. Additionally, we introduce a simple yet novel adversarial attack that is shown to be effective against Fourier Neural Operators – the Low Frequency Attack.

***Keywords*:** PINO, Fourier Neural Operator, Adversarial Attacks, Deep Learning, Diffusion Equation, Medical Imaging

Введение

Многие задачи науки и техники связаны с многократным решением сложных систем уравнений в частных производных (PDE) для различных значений некоторых параметров. Примеры возникают в молекулярной динамике, микромеханике и турбулентных течениях. Часто такие системы требуют тонкой дискретизации, чтобы уловить моделируемое явление. В результате традиционные численные решатели оказываются медленными и иногда неэффективными. Например, при проектировании материалов, таких как аэродинамические профили, необходимо решать обратную задачу, где требуется провести тысячи прогонов прямой модели. Быстрый метод делает такие задачи осуществимыми.

Традиционные решатели PDE, такие как методы конечных элементов (FEM) и конечных разностей (FDM), решают уравнение путём дискретизации пространства. Таким образом, возникает компромисс между скоростью и точностью: грубые сетки быстры, но менее точны; мелкие сетки точны, но медленны. Сложные PDE-системы обычно требуют очень мелкой дискретизации и поэтому крайне трудоёмки для традиционных методов. С другой стороны, методы, основанные на данных, могут напрямую обучаться по данным, что делает их на порядки быстрее традиционных решателей.

Physics-Informed Neural Networks (PINNs) [1] произвели революцию в решении PDE, включая физические законы в процесс обучения. PINNs могут интегрировать экспериментальные данные разной точности и модальности с различными формулировками уравнений Навье–Стокса для несжимаемых потоков [2, 3], а также для сжимаемых потоков [4] и биомедицинских течений [5]. Однако PINNs ограничены в своей способности устойчиво решать широкий класс PDE, так как любое

изменение начальных или граничных условий требует полного переобучения нейросети. Для решения этой проблемы были предложены PhysicsInformed Neural Operators (PINOs) [6], которые расширяют PINNs. PINO преобразовали решение PDE, обучая отображения операторов, которые обобщаются на разные входные функции, предлагая более высокую эффективность для задач вроде одномерного уравнения диффузии. PINO интегрируют физические законы в операторные нейросетевые структуры, такие как DeepONet [7] или Fourier Neural Operator (FNO) [8], что позволяет моделировать сложные системы с различными начальными и граничными условиями.

Тем не менее, глубокие нейронные сети уязвимы к adversarial-атакам — малым возмущениям на входе, которые приводят к значительным ошибкам предсказания, — что представляет собой вызов для надёжных научных вычислений. Adversarial-атаки и методы защиты широко изучены в компьютерном зрении [9, 10, 11, 12, 13]. Однако лишь немногие работы посвящены adversarial-атакам в задачах PINN и PINO. Недавние исследования, такие как [14], предлагают методы вроде AT-PINNs для повышения устойчивости, хотя их применение к уравнению теплопроводности пока мало изучено. Другие работы, включая исследования PINNs для задач теплопередачи [15], подчёркивают необходимость устойчивых моделей, но не рассматривают adversarial-атаки. Как и PINNs, PINOs уязвимы к adversarial-атакам, что ставит под угрозу их надёжность [16]. Кроме того, методы компьютерного зрения, основанные на решении PDE, например гибридные подходы [17], использующие метод Перона–Малика для устранения шума [18], также могут быть уязвимы.

В этой статье мы исследуем влияние таких атак на PINO для одномерного уравнения теплопроводности и эффективность adversarial augmentation в повышении устойчивости моделей, представляя метрики для оценки устойчивости и выделяя направления для будущих исследований, чтобы обеспечить надёжные предсказания в критических приложениях, таких как медицинская визуализация и теплотехника.

Постановка задачи

1. Одномерное уравнение теплопроводности

Мы рассматриваем одномерное уравнение теплопроводности

$$\frac{\partial t}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}, \quad x \in (0,1), t \in (0,1], \tag{1}$$

где $u(x, t)$ — функция распределения температуры, t — время, x — пространственная координата, а $\alpha > 0$ — коэффициент теплопроводности.

Для решения этого уравнения нам необходимо задать начальное и граничные условия. Пусть начальное распределение температуры задано:

$$u(x,0) = f(x), \forall x \in (0,1), \tag{2}$$

где $f(x)$ — произвольная функция начального распределения температуры вдоль стержня длиной 1. Рассмотрим однородные граничные условия Дирихле

$$u(0,t) = 0, \forall t > 0, \tag{3}$$

$$u(1,t) = 0, \forall t > 0. \tag{4}$$

С помощью синус-преобразования решение можно записать как

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin(n\pi x) e^{-\alpha(n\pi)^2 t}, \tag{5}$$

где коэффициенты b_n определяются начальными условиями

$$b_n = 2\int_0^1 f(x)\sin(n\pi x)dx. \tag{6}$$

Подставив эти коэффициенты, получаем полное решение $u(x,t)$, которое используется в наших экспериментах как ground truth.

2. Adversarial-атаки

Adversarial-атаки — это малые возмущения, добавляемые к исходному образцу, которые могут приводить к ошибкам классификации или регрессии нейросети. В контексте PDE атаки могут быть

направлены на начальные или граничные условия, а также на входную сетку. В данной работе мы сосредотачиваемся на атаках на начальные условия.

Формулировка задачи:

$$\max_{\xi} L(r(v+\xi;\theta), r(v+\xi)), \tag{7}$$

$$\text{s.t.} \quad \|\xi\|_\infty \le \epsilon, \tag{8}$$

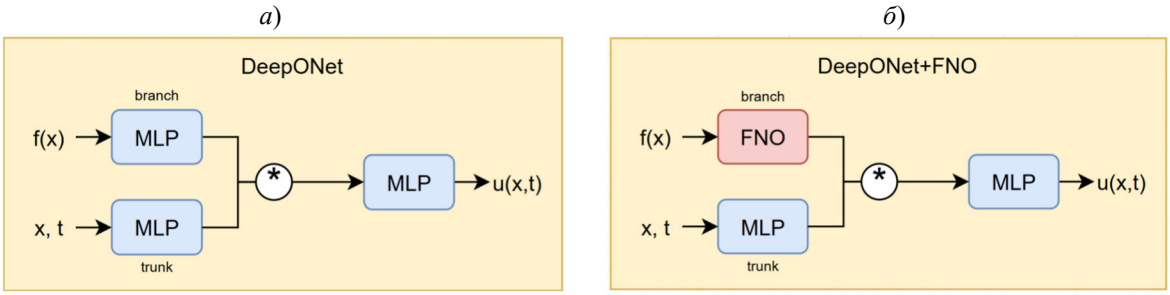
где $L(\cdot)$ — функция расстояния между $(r(v+\xi;\theta)$ и $r(v+\xi)$, а ϵ — максимально допустимая величина возмущения. В нашей работе используется L_2 – норма.

3. DeepONet

DeepONet — архитектура нейронной сети для обучения операторов [7]. Она вдохновлена теоремой об универсальной аппроксимации операторов и состоит из двух подсетей: branch network и trunk network. Branch network кодирует входные функции, а trunk network кодирует точки, в которых вычисляется оператор. В наших экспериментах мы используем три варианта DeepONet:

- оригинальный DeepONet с MLP в качестве branch и trunk;
- DeepONet с FNO в качестве branch;
- Stacked DeepONet с FNO в branch и разными Nmodes.

Также мы добавляем дополнительный слой MLP после произведения branch и trunk сетей, как показано на рисунке.



Варианты архитектуры DeepONet, используемые в данной статье: *а*) DeepONet с MLP в качестве branch и trunk сетей, с дополнительным слоем MLP в конце; *б*) DeepONet с нейронным оператором Фурье в качестве branch сети и MLP в качестве trunk, с дополнительным слоем MLP в конце

Теория

Для обучения нейросети на каждой итерации мы сэмплируем $x \sim \text{Uniform}(0, 1)$, $t \sim \text{Exponential}(\lambda = 10.0)$. Экспоненциальное распределение времени выбрано, чтобы избежать переобучения на состояниях с температурами, близкими к нулю. Формула плотности этого распределения имеет вид

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \ge 0 \\ 0, & if\ x < 0 \end{cases} \tag{9}$$

Начальные условия формируются как сумма синусоид с случайными коэффициентами

$$f(x) = \sum_{i=1}^{Nc} a_i \sin(i\pi x), \tag{10}$$

где $Nc = 100$.

Для генерации adversarial-атак на начальные условия мы используем:

1. Low Frequency Attack (LFA) — добавление низкочастотной компоненты

$$f_{adv}(x) = f(x) + \epsilon \sin(k\pi x), \tag{11}$$

где $k = 1$, $\epsilon = 0.1$.

2. Fast Gradient Sign Method (FGSM) [9]:

$$f_{adv}(x) = f(x) + \epsilon \text{sign} \nabla_x L(u(x,t;\theta), u(x,t)). \tag{12}$$

При adversarial training на каждой итерации с вероятностью p = 0.5 мы используем adversarial начальное условие, сгенерированное с помощью предварительно обученной модели. Решение для такого условия берётся как ground truth для обучения.

Результаты экспериментов

Мы оцениваем модели, вычисляя ошибку L_2 между предсказаниями и двумя решениями: чистым (без возмущений) и adversarial (с возмущением в начальных условиях). Формула

$$L_2 \approx \frac{1}{N_x N_t} \sum_{i=0}^{N_x-1} \sum_{j=0}^{N_t-1} (u(\frac{i}{N_x}, \frac{j}{N_t}; \theta) - u(\frac{i}{N_x}, \frac{j}{N_t}))^2, \quad (13)$$

при $N_x = N_t = 1024$. Результаты приведены в таблице. Первый столбец – названия моделей. Надстрочный индекс adv в названии модели указывает, что модель обучалась с использованием adversarial-усиления. Далее названия столбцов обозначают тип adversarial-атаки на начальное условие; надстрочный индекс adv обозначает ошибку в норме L_2 относительно adversarial-решения, а отсутствие индекса – на ошибку относительно чистого решения.

Численные результаты, показывающие эффективность ADVERSARIAL-усиления

Модель \ Атака:	L_2	L_2 LFA(k=1, ϵ=0.1)	L_2^{adv} LFA(k=1, ϵ=0.1)	L_2 FGSM(ϵ=0.1)	L_2^{adv} FGSM(ϵ=0.1)
MLP	0.0009	0.0009	0.006	0.001	0.001
MLP^{adv}	0.0005	0.002	0.001	0.001	0.0008
FNO_{16}	0.0009	0.001	0.005	0.016	0.015
FNO_{16}^{adv}	0.0003	0.003	0.0006	0.005	0.004
FNO_{32}	0.0006	0.001	0.003	0.009	0.005
FNO_{32}^{adv}	0.0006	0.002	0.002	0.004	0.001
FNO_{64}	0.04	0.04	0.05	0.05	0.05
FNO_{64}^{adv}	0.04	0.04	0.05	0.05	0.05
$StackFNO_{8,16,32,64}$	0.001	0.001	0.006	0.01	0.004
$StackFNO_{8,16,32,64}^{adv}$	0.0003	0.003	0.0009	0.005	0.001

Можно видеть, что практически в любом случае, при любом выборе $Nmodes$ для FNO, модели с состязательной аугментацией превосходят свои аналоги без состязательной аугментации на состязательных решениях, что свидетельствует о повышенной устойчивости. Кроме того, состязательная аугментация может повысить производительность на чистых данных, как показано во втором столбце.

Заключение

В этой статье предложен метод adversarial augmentation для обучения PINO. Показано, что он повышает устойчивость моделей на чистых и adversarial-данных. Необходимы дальнейшие исследования влияния adversarial augmentation и выбора атак для других PDE (поток Дарси, уравнение Бюргерса, Навье–Стокса), а также методов медицинской визуализации, основанных на 2D-фильтре Перона–Малика.

Список литературы

1. M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial dierential equations, Journal of Computational physics 378 (2019) 686–707.

2. X. Jin, S. Cai, H. Li, G. E. Karniadakis, Nsfnets (navier-stokes ow nets): Physics-informed neural networks for the incompressible navier-stokes equations, Journal of Computational Physics 426 (2021) 109951.

3. M. Raissi, A. Yazdani, G. E. Karniadakis, Hidden uid mechanics: Learning velocity and pressure fields from flow visualizations, Science 367 (2020) 1026–1030.

4. Z. Mao, A. D. Jagtap, G. E. Karniadakis, Physics-informed neural networks for high-speed flows, Computer Methods in Applied Mechanics and Engineering 360 (2020) 112789.

5. M. Yin, X. Zheng, J. D. Humphrey, G. E. Karniadakis, Non-invasive inference of thrombus material properties with physics-informed neural networks, Computer Methods in Applied Mechanics and Engineering 375 (2021) 113603.

6. N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar, Neural operator: Learning maps between function spaces with applications to PDEs, Journal of Machine Learning Research 24 (2023) 1–97.

7. L. Lu, P. Jin, G. Pang, Z. Zhang, G. E. Karniadakis, Learning nonlinear operators via deeponet based on the universal approximation theorem of operators, Nature machine intelligence 3 (2021) 218–229.

8. Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial dierential equations, arXiv preprint arXiv:2010.08895 (2020).

9. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

10. X. Chen, X. Gao, J. Zhao, K. Ye, C.-Z. Xu, Advdiuser: Natural adversarial example synthesis with diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4562–4572.

11. C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, D. Song, Generating adversarial examples with adversarial networks, arXiv preprint arXiv:1801.02610 (2018).

12. P. Zhu, G. Osada, H. Kataoka, T. Takahashi, Frequency-aware gan for adversarial manipulation generation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4315–4324.

13. W. Ma, Y. Li, X. Jia, W. Xu, Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4630–4639.

14. Y. Li, S. Shi, Z. Guo, B. Wu, Adversarial training for physics-informed neural networks, 2023. URL: https://arxiv.org/abs/2310.11789. arXiv:2310.11789.

15. S. Cai, Z. Wang, S. Wang, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks for heat transfer problems, Journal of Heat Transfer 143 (2021) 060801.

16. A. D. Adesoji, P.-Y. Chen, Evaluating the adversarial robustness for fourier neural operators, arXiv preprint arXiv:2204.04259 (2022).

17. N. Wang, Y. Shang, Y. Chen, M. Yang, Q. Zhang, Y. Liu, Z. Gui, A hybrid model for image denoising.

18. P. Pietro, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. PAMI 12 (1990).