Scaling Up AI-driven Scientific Discovery via Embedding Physics Modeling into End-to-end Learning, Harnessing Randomized Algorithms and Control Variable Experimentation

Yexiang Xue, Department of Computer Science, Purdue University. yexiang@purdue.edu

Learning first-principle physics models directly from experiment data has been a grand goal of Artificial Intelligence, and will greatly accelerate the pace of scientific discoveries if succeeded. Nevertheless, despite a lot of exciting effort [6-20], AI-driven scientific discovery in a closed loop has not been fully realized, because of major computational bottlenecks, particularly, the lack of end-to-end frameworks embedding large-scale simulation of physics models into learning, and the lack of efficient algorithms to accelerate not only the forward simulation, but the even more computational demanding backward gradient propagation.

We address these key computational bottlenecks in AI-driven scientific discovery by proposing:

(1) an end-to-end framework[3] to learn physics models in the form of Partial Differential Equations (PDEs) directly from the experiment data. In our framework, PDE solvers are formulated as fully differentiable neural network layers, allowing for seamless embedding of large-scale physics simulation into learning and efficient gradient propagation. We integrate two neural networks, one neural differential equation net simulating the temporal dynamics of PDEs, and the second recognition net extracting the values of PDE variables from data. Back-propagation is carried out through both networks to match the predicted values of PDE variables with experiment observations. Recently, a series of related research aim at learning partial differential equations from data. However, they did not achieve fully automatic physics model identification from experiment data because the input of these models are the trajectories of differential equations, which may be unavailable from experiment data and need to be extracted as a separate step.

(2) We also propose to scale up learning first-principle models harnessing randomized algorithms[1,2]. We notice the temporal evolutions of many physical systems often consist of gradually changing updates across wide areas in addition to a few rapid updates concentrated in a small set of "interfacial" regions. In other words, the updates can be decomposed into the sum of a sparse signal in the value domain (for rapid updates in the interfacial regions) and another sparse one in the frequency domain (for gradual updates). This sparse decomposition allows us to accelerate both the "inner-loop" simulation and the "outer-loop" learning via randomized algorithms with provable guarantees on accuracy. Specifically, random projections can be used to handle sparse updates in compressed spaces for faster computation. Locality Sensitive Hashing (LSH) accelerates computation by handling similar updates at once.

(3) Finally, careful experiment designs lead to faster identification of ground-truth physics models. In this direction, we build increasingly more complex symbolic expressions with the aid of control variable experimentation. Initially, we focus on searching for the best reduced model under controlled environments where the values of most variables are held constant. The correct reduced model can be quickly identified because of the reduced modeling space. Then we extend the reduced model to more complicated models by experimenting on new variables in an incremental way. We show this approach leads to faster learning.

The development of AI-driven scientific discovery approaches was motivated by the real-world application of learning the physics model of nano-scale crystalline defects in materials[4-5]. Transmission Electron Microscope (TEM) cameras capture high resolution videos of these detects in the size of terabytes during in-situ radiation experiments, which justifies the need for fully automated analysis tools. In summary, our work provides a series of effective computational tools to accelerate scientific discovery.

References

[1] M. Nasim, X. Zhang, A. El-Azab, Y. Xue. *Efficient Learning of Sparse and Decomposable PDEs using Random Projection*. In UAI, 2022. https://proceedings.mlr.press/v180/nasim22a.html

[2] C. Sima, Y. Xue. *LSH-SMILE: Locality Sensitive Hashing Accelerated Simulation and Learning*. In NeurIPS, 2021. https://www.cs.purdue.edu/homes/yexiang/publications/2021Neurips_LSH_Physics.pdf

[3] Y. Xue, M. Nasim, M. Zhang, C. Fan, X. Zhang, A. El-Azab. *Physics Knowledge Discovery via Neural Differential Equation Embedding*. In ECML-PKDD, 2021.
https://www.cs.purdue.edu/homes/yexiang/publications/ecml_pkdd21.pdf
[4] Md Nasim, A.R.G. Sreekar, Tongjun Niu, Cuncai Fan, Zhongxia Shang, Jin Li, Haiyan Wang, Anter El-Azab, Yexiang Xue*, and Xinghang Zhang*. Unraveling the size fluctuation and shrinkage of nanovoids during in situ radiation of Cu by automatic pattern recognition and phase field simulation. In Journal of Nuclear Materials (JNM), 2022.
[5] Tongjun Niu, Md Nasim, Annadanam Rayaprolu Goutham Sreekar, Cuncai Fan, Jin Li, Zhongxia Shang, Yexiang Xue, Anter El-Azab, Haiyan Wang, Xinghang Zhang. Recent studies on void shrinkage in metallic materials subjected to in situ heavy ion irradiations. In The Journal of The Minerals, Metals & Materials Society (JOM), 2020.
[6] Abolafia, D.A., Norouzi, M., Le, Q.V.: Neural program synthesis with priority queue training. CoRR abs/1801.03526 (2018)
[7] Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., Loreggia, A., Murugesan, K., Mattei, N., Rossi, F., Srivastava, B.: Thinking fast and slow in AI. In: AAAI. pp. 15042–15046. AAAI Press (2021)
[8] Bradley, E., Easley, M., Stolle, R.: Reasoning about nonlinear system identification. Artificial Intelligence 133(1), 139–188 (2001)
[9] Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences 113(15), 3932–3937 (2016)
[10] Chen, C., Luo, C., Jiang, Z.: Elite bases regression: A real-time algorithm for symbolic regression. In: ICNC-FSKD. pp. 529–535. IEEE (2017)
[11] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems 31 (2018)
[12] Cranmer, M.D., Sanchez-Gonzalez, A., Battaglia, P.W., Xu, R., Cranmer, K., Spergel, D.N., Ho, S.: Discovering symbolic models from deep learning with inductive biases. In: NeurIPS (2020)
[13] Dzeroski, S., Todorovski, L.: Discovering dynamics: From inductive logic programming to machine discovery. J. Intell. Inf. Syst. 4(1), 89–108 (1995)
[14] Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F.A., Miranda, M., Pallarès, J., Sales-Pardo, M.: A bayesian machine scientist to aid in the solution of challenging scientific problems. Science advances 6(5), eaav6971 (2020)
[15] Iten, R., Metger, T., Wilming, H., Del Rio, L., Renner, R.: Discovering physical concepts with neural networks. Physical review letters 124(1), 010508 (2020)
[16] King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E., Clare, A.: The automation of science. Science 324(5923), 85–89 (2009)
[17] Langley, P.W., Simon, H.A., Bradshaw, G., Zytkow, J.M.: Scientific Discovery: Computational Explorations of the Creative Process. The MIT Press (02 1987)
[18] Liu, Z., Tegmark, M.: Machine learning conservation laws from trajectories. Phys. Rev. Lett. 126, 180604 (May 2021)
[19] Raissi, M., Perdikaris, P., Karniadakis, G.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686–707 (2019)
[20] Schmidt, M., Lipson, H.: Distilling free-form natural laws from experimental data. Science 324(5923), 81–85 (2009)
[21] Valdés-Pérez, R.: Human/computer interactive elucidation of reaction mechanisms:bapplication to catalyzed hydrogenolysis of ethane. Catalysis Letters 28, 79–87 (1994)