WUYANG CHEN, Simon Fraser University

Towards Data-Efficient and OOD Generalization of Scientific Machine Learning Models

In recent years, there has been growing promise in coupling machine learning methods with domain-specific physical insights to solve scientific problems based on partial differential equations (PDEs). However, there are two critical bottlenecks that must be addressed before scientific machine learning (SciML) can become practically useful. First, SciML requires extensive pretraining data to cover diverse physical systems and real-world scenarios. Second, SciML models often perform poorly when confronted with unseen data distributions that deviate from the training source, even when dealing with samples from the same physical systems that have only slight differences in physical parameters. In this line of work, we aim to address these challenges using data-centric approaches. To enhance data efficiency, we have developed the first unsupervised learning method for neural operators. Our approach involves mining unlabeled PDE data without relying on heavy numerical simulations. We demonstrate that unsupervised pretraining can consistently reduce the number of simulated samples required during fine-tuning across a wide range of PDEs and real-world problems. Furthermore, to evaluate and improve the out-of-distribution (OOD) generalization of neural operators, we have carefully designed a benchmark that includes diverse physical parameters to emulate real-world scenarios. By evaluating popular architectures across a broad spectrum of PDEs, we conclude that neural operators achieve more robust OOD generalization when pretrained on physical dynamics with high-frequency patterns rather than smooth ones. This suggests that data-driven SciML methods will benefit more from learning from challenging samples.