Invited Talk: Interpretable Deep Learning for Physics via Symbolic Regression

Miles Cranmer

Princeton University

Abstract

If we train a neural network on some dynamical system in some region of phase space, and it learns a way to execute the dynamics more efficiently than a handwritten code, how do we distill physical insight from the learned model? In this talk, I will argue that symbolic learning should play a major role in the process of interpreting a machine learning model for physical systems. I will discuss our method for converting an MLP-based deep neural network - which has been trained on a physical system - into a symbolic model. Our method makes use of our genetic algorithm-based symbolic regression tool, "PySR", applied to each latent space of the network. One of the problems with this process is working with the fact that neural networks have high-dimensional latent spaces, and genetic algorithms scale poorly with the number of features. To work around this issue, I'll then introduce our "Disentangled Sparsity Network", which encourages a neural network to learn an easy-to-interpret representation. I will then share several recent applications of our techniques to real physical systems, and the various insights we have discovered and rediscovered.