Steps toward an AI scientist: neuro-symbolic models for concept generalization and theory learning

Tailin Wu, Stanford University

Abstract: In this talk, I will introduce my works that expand machine learning models' capability for learning novel concepts and theories. Such capability is essential for building an AI Scientist, an agent that has human's concept generalization capability and can propose scientific theories on the relevant concepts.

Firstly, I introduce a paradigm for learning theories for describing a dynamical system [1]. Instead of using one model to learn everything, we propose a novel paradigm centered around the learning and manipulation of *theories*, which parsimoniously predict both aspects of the future (from past observations) and the domain in which these predictions are accurate. We incorporate four common strategies with a long history in physics: divide-and-conquer, Occam's razor, unification and lifelong learning, for the discovery of theories. Specifically, we propose a novel generalized-mean-loss to encourage each theory to specialize in its comparatively advantageous domain, and a differentiable description length objective to downweight bad data and "snap" learned theories into simple symbolic formulas. Theories are stored in a "theory hub", which continuously unifies learned theories and can propose theories when encountering new environments. We test our implementation, the toy "AI Physicist" learning agent, on a suite of increasingly complex physics environments. From unsupervised observation of trajectories through worlds involving random combinations of gravity, electromagnetism, harmonic motion and elastic bounces, our agent typically learns faster and produces mean-squared prediction errors about a billion times smaller than a standard feedforward neural net of comparable complexity, typically recovering integer and rational theory parameters exactly. Our agent successfully identifies domains with different laws of motion also for a nonlinear chaotic double pendulum in a piecewise constant force field.

Secondly, I will introduce Zero-shot Concept Recognition and Acquisition (ZeroC) [2], a neuro-symbolic architecture that can recognize and acquire more complex, hierarchical concepts at inference time in a zero-shot way. ZeroC represents concepts (e.g., rectangles, tables) as graphs of constituent concept models (as nodes) and their relations (as edges). It establishes a one-to-one mapping between a symbolic graph structure of a concept and its corresponding probability model (in terms of energy-based models) for recognition and detection. Pretrained with recognizing simpler concepts and relations, ZeroC can acquire new concepts, communicate its graph structure, and apply it to classification and detection tasks (even across domains) at inference time.

Finally, I will discuss future directions that combine concept and theory learning.

References:

[1] **T. Wu**, Max Tegmark, "Toward an Artificial Intelligence Physicist for Unsupervised Learning." Physical Review E 100 (3), 033311. Featured in PRE Spotlight on Machine Learning in Physics. Featured in MIT Technology Review and MotherBoard. URL: https://journals.aps.org/pre/abstract/10.1103/PhysRevE.100.033311 .

[2] **T. Wu**, M. Tjandrasuwita, Z. Wu, X. Yang, K. Liu, R. Sosič, J. Leskovec, "ZeroC: A Neuro-Symbolic Model for Zero-shot Concept Recognition and Acquisition at Inference Time." In 36th Conference on Neural Information Processing Systems (NeurIPS 2022). URL: https://arxiv.org/abs/2206.15049 .