Abstracts

Sebastijan Dumančić

Machine learning models that provably satisfy constraints

In this talk, I will cover several of our works on training machine learning models such that they probably satisfy constraints their users impose on them. For instance, we might want models that are guaranteed to be fair, avoid catastrophic scenarios, and so on. This kind of problem is a perfect avenue for a fruitful integration of logic and statistical learning. I will also briefly outline our key insights, as well as the challenges we faced and future opportunities.

Atticus Geiger

Causal Abstraction as a Theoretical Foundation for Mechanistic Interpretability

Causal abstraction provides a theoretical foundation for mechanistic interpretability, the field concerned with providing intelligible algorithms that are faithful simplifications of the known, but opaque low-level details of black box AI models. There are good reasons to be optimistic about productive interplay between theoretical work on causal abstraction and applied work on mechanistic interpretability, as we can measure and manipulate the micro variables of deep learning models with perfect precision and accuracy, and thus empirical claims about their structure can be held to the highest standard of rigorous falsification through experimentation.

Martin Grohe

The Logic of Graph Neural Networks

Graph neural networks (GNNs) are deep learning models for graph data that play a key role in machine learning on graphs. A GNN describes a distributed algorithm carrying out local computations at the vertices of the input graph. Typically, the parameters governing this algorithm are acquired through data-driven learning processes.

After introducing the basic model, in this talk I will focus on the expressiveness of GNNs: which functions on graphs or their vertices can be computed by GNNs? Understanding the expressiveness will help us understand the suitability of GNNs for various application tasks and guide our search for possible extensions.

Surprisingly, the expressiveness of GNNs has a clean and precise characterisation in terms of logic and Boolean circuits, that is, computation models of classical (descriptive) complexity theory.

Levin Hornischer

Semantics for Non-symbolic Computation: Including Neural Networks and Analog Computers

Despite the great technological progress, we are lacking a foundational theory of modern artificial intelligence. We want to interpret, explain, and verify the 'sub-symbolic' computation performed by neural networks that drive this success. For classical 'symbolic' computation, this problem was solved by semantics: the mathematical description of the meaning of program code. In this talk, we develop one approach to analogous semantics for non-symbolic computation performed by neural networks and other analog computers. To do so, we first summarize the three semantics for symbolic computation (operational, denotational, logical), and then we describe our counterpart for non-symbolic computation (dynamical systems, domains, and modal logic). The key idea is to represent the dynamics of the non-symbolic computation as a limit of symbolic approximations, which are given by interpretable observations. We implement these techniques to illustrate the training dynamics of a neural network in a standard machine learning task.

Thomas Icard

TBA

Herbert Jaeger

What a mathematical foundation for unconventional computing should deliver and how it might look like

For digital computing we possess a formal theory foundation that deeply roots in Western philosophical history, is mathematically transparent, has been worked out and stabilized and codified into a standard textbook format, and obviously has changed the world through digital computers. For information processing in neuromorphic microchips, or in other new and to-be-found hardware substrates based on unconventional physical effects, or in biological brains or other natural systems, we do not have anything like a unifying formal theory foundation. But we need it, and it should not only be academically acceptable but really practically useful. In my talk I will draw a quick overall picture of this situation, and then outline my own approach toward formulating such a general formal theory for information processing in non-digital, non-symbolic dynamical systems.

Giuseppe Marra

From Statistical Relational to Neuro-Symbolic AI

The integration of learning and reasoning is one of the key challenges in artificial intelligence and machine learning today. The area of Neuro-Symbolic computation tackles this challenge by integrating symbolic reasoning with neural networks. This talk will provide a description of Neuro-Symbolic Artificial Intelligence (NeSy) by drawing several parallels to another field that has a rich tradition in integrating learning and reasoning, namely Statistical Relational Artificial Intelligence (StarAI). The talk will discuss seven dimensions for introducing and categorizing several StarAI and NeSy approaches. A wide variety of systems will be positioned along some of these dimensions, naturally leading to an underlying “recipe” of Nesy. Finally, recent developments in generative and explainable NeSy approaches will be discussed.

Lena Strobl

Expressivity of Transformers: What Formal Languages Can They Represent?

A major advancement in language modeling is the use of the transformer architecture. But what problems can transformers solve, what problems can they not solve, and how can we prove it? This talk examines the expressivity of transformers through the lens of computability and complexity theory. We will situate transformers within the landscape of automata, boolean circuits, and formal logics. We will discuss what is currently known about transformers’ capabilities and limitations, address the practical implications of these results for natural language processing, and identify some directions for future work. The goal is for attendees to gain a comprehensive understanding of transformers’ expressive power, specifically in terms of the problems they fundamentally can and cannot solve.