The Weizmann Institute of Science
Faculty of Mathematics and Computer Science

Machine Learning and Statistics Seminar

Room 1, Ziskind Building
on Wednesday, May 15, 2024
at 11:15

Dan Vilenchik
BGU

will speak on

Towards Reverse Algorithmic Engineering of Neural Networks

Abstract:
As machine learning models get more complex, they can outperform traditional algorithms and tackle a broader range of problems, including challenging combinatorial optimization tasks. However, this increased complexity can make understanding how the model makes its decisions difficult. Explainable models can increase trust in the model's decisions and may even lead to improvements in the algorithm itself. Algorithms like GradCAM or SHAP provide good explanations in terms of feature importance, typically for classification tasks. Still, they provide little insight when the ML pipeline is designed to work, for example, as an algorithm for solving optimization problems. In this talk, we present a concept-learning framework for explaining a neural machine-learning model's decision-making process from an algorithmic point of view. Using the NeuroSAT algorithm for SAT solving as a case study, we demonstrate how our framework finds the algorithmic concepts that drive the operation of NeuroSAT. Using the concepts that we discover, we can re-write the black box NeuroSAT net as a text-book algorithm that performs typical algorithmic moves like (a) compute confidence levels for every variable, (b) fix variables with the highest confidence and simplify the instance, (c) solve the residual formula using some simple technique. (Such a principle guides, for example, the well-known Belief-Propagation-Decimation algorithm).

Joint work with Elad Shoham (PhD student BGU), Kahalil Wattad (MSc student BGU), Hadar Cohen (MSc student BGU), and Havana Rika (Tel-Aviv-Yafo Academic College).

Short bio:

Dan Vilenchik holds a PhD in computer science from Tel Aviv University. He did a postdoc at UC Berkeley and UCLA. He is currently a tenured member of the Electrical Engineering School at Ben-Gurion University. His research includes various aspects of machine learning, such as the challenges of high-dimensional data, explainable AI, NLP, and multidisciplinary projects.