NJIT

Department of Computer Science

Restricting information flow in deep learning for interpretability and insight

Kieran Murphy
University of Pennsylvania

Hosted by: Ioannis Koutis

Date: Thursday, February 13, 2025
Coffee: 2:15 PM – 2:30 PM
Time: 2:30 AM – 3:30 PM
Location: GITC 4402 (4th floor Seminar Lecture Hall)

Zoom Link: https://njit-edu.zoom.us/j/94567889080?pwd=FVNxiAuNWSPEKwLCDwoAHriPJ1ISqi.1

Abstract:
Neural networks excel at information processing, distilling immense variation in data into reduced representations that are useful for diverse tasks. While learning objectives often involve quantities related to information content (in the formal sense of information theory), the actual processing of information throughout a network is generally opaque and difficult to access because of how point transforms preserve information. My research addresses this by integrating probabilistic encoders early in network architectures to restrict and monitor information flow from multiple sources, such as expression levels in genomics datasets, structural descriptors of materials, timesteps of dynamic processes, or latent dimensions of trained representation spaces.

By tracking what information a model uses about data, we gain a powerful new avenue for interpretability that preserves the full expressivity of downstream processing. Simultaneously, this framework provides actionable insights into key relationships within the data and supports recently proposed assessments of multivariate structure in complex systems. Finally, it facilitates principled analysis and engineering of representation learning methods, as we demonstrate in the case of unsupervised disentanglement. The generality of my approach bridges disciplines, empowering advancements in machine learning, physics, and data science and addressing real-world challenges by combining the rigor of information theory with the scalability of deep learning.

Bio:
Kieran is a postdoc at the University of Pennsylvania, working with Dani Bassett on research at the intersection of machine learning, information theory, and physics. Before this, he studied the physics of amorphous materials during a Ph.D. at the University of Chicago, which included 3D-printing around 100,000 bespoke grains of sand. He also spent 1.5 years in New York City as part of Google's AI Residency program, where he developed methods in computer vision and representation learning.