知の物理学研究センター

ipi seminar

Institute for Physics of Intelligence

Hidenori Tanaka / 田中 秀宣 氏

(Stanford University/スタンフォード大学)

「From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction」

２０19年10月11日(金)　１０時００分〜

東京大学本郷キャンパス理学部１号館４階 447 号室

Recently, deep feedforward neural networks have achieved considerable success in modeling biological sensory processing, in terms of reproducing the input-output map of sensory neurons. However, such models raise profound questions about the very nature of explanation in neuroscience. Are we simply replacing one complex system (a biological circuit) with another (a deep network), without understanding either? Moreover, beyond neural representations, are the deep network's computational mechanisms for generating neural responses the same as those in the brain? Without an algorithmic approach to extracting and understanding computational mechanisms from deep neural network models, it can be difficult both to assess the degree of utility of deep learning approaches in neuroscience, and to extract experimentally testable hypotheses from deep networks. We develop such an algorithmic approach by combining dimensionality reduction and modern attribution methods for determining the relative importance of interneurons for specific visual computations. We apply this approach to deep network models of the retina, revealing a conceptual understanding of how the retina acts as a predictive feature extractor that signals deviations from expectations for diverse spatiotemporal stimuli. For each stimulus, our extracted computational mechanisms are consistent with prior scientific literature, and in one case yields a new mechanistic hypothesis. Thus overall, this work not only yields insights into the computational mechanisms underlying the striking predictive capabilities of the retina, but also places the framework of deep networks as neuroscientific models on firmer theoretical foundations, by providing an algorithmic path to go beyond comparing neural representations to extracting and understand computational mechanisms.