Understanding Deep Representation Learning via Neural Collapse

Qing Qu
University of Michigan, United States
qingqu@umich.edu

Recently, an intriguing phenomenon in the final stages of network training has been discovered and caught great interest, in which the last-layer features and classifiers collapse to simple but elegant mathematical structures: all training inputs are mapped to class-specific points in feature space, and the last-layer classifier converges to the dual of the features' class means while attaining the maximum possible margin. This phenomenon, dubbed Neural Collapse, persists across a variety of different network architectures, datasets, and even data domains. Moreover, a progressive neural collapse occurs from shallow to deep layers. This talk leverages the symmetry and geometry of Neural Collapse, and develops a rigorous mathematical theory to explain when and why it happens under the so-called unconstrained feature model. Based upon this, we show how it can be used to provide guidelines to understand and improve transferability with more efficient fine-tuning.