The Information in Optimal Representations

Stefano Soatto

Stefano Soatto, University of California, Los Angeles - Samueli School of Engineering, Computer Science, Los Angeles, USA

Optimal representations are functions computed by solving an optimization problem using past data (training set). Their optimality is defined in terms of data that is not given to us (test set), whereas the inference criterion use a disjoint finite sample. I will start with defining optimality in a learned representation using classical notions of sufficiency and invariance, which can be formalized in classical information-theoretic language but that are not computable from the training set. I will then derive an inference criterion and variational principle based on generalization bounds, that is at face value unrelated to the optimality criteria. I will then describe the Emergence Bound, that shows that the optimality of a learned representation is connected with the Information Lagrangian, which is computed from the training data and minimized during the training process. In the process, I will show that the Information in the Weights of a deep neural network plays a key role in the optimality of the representation. Even defining this notion is non-trivial as classical tools for information theory yield degenerate results, as a trained network is a deterministic system. I will show how information in a trained network can be defined, computed, and related to notions of sufficiency, invariance, minimality, stability/sensitivity, and tied to the geometry of the loss landscape and the notion of “accessible information.” Once understood how to compute an optimal representation for a given data and a given task, I will turn to determining the relation among tasks, starting from defining a topology in the space of learning tasks. Again I will show that Information in the weights provides a metric for the space of learning tasks (Task2Vec). However, I will show that there are non-local and non-linear phenomena at play that make the transient dynamic critical in determining whether a task is “reachable” from another, no matter how close they are to one another in the information metric (Critical Learning Periods). This shows that regularization when training deep network does not act by modifying the geometry and topology of the loss landscape at convergence, but instead by influencing the early transient dynamics (Time Matters in Regularizing Deep Networks). I will conclude with open problems and promising directions based on recent empirical observations.