Designed logic lucidity: Enhance neural network explainability via the concept of statistical molding

Ker-Chau Li[1,2]

[1]*Statistics Department, UCLA, U.S.A.*
[2]*Institute of Statistical Science, Academia Sinica, Taiwan*

Abstract

NN modeling consists of two components, network architecture and parameter training. Hinging on the loss function, a training data set is fed into the network. After many training cycles, there output the optimal values for the numerous parameters, the node connection weights and biases, which often outnumber the training data, thereby leading to dilemmas such as over/under f it and double-descent. Meanwhile, inside the network, numerous nodes are interlocked and their presumed roles are at best loosely designated by the network structure. How to explain or interpret the behavior of a well-trained network, such as the impulse responses or node-to-node, layer -to-layer correlations, has received more and more attention. To increase logic lucidity of the data flow underlying the network configuration, I will present a new concept called statistical molding (SM). SM requires (i) a modification of network configuration by a set of logic rules, (ii) molding distributions for the data passing through the modified nodes, and (iii) a molding bonus function. Applying to supervised learning, the tuning of SM is conducted with the aim of maximizing the molding bonus without compromising the misclassification error. If successfully trained, SM can reveal clusters hidden within and across classes and facilitate post-classification diagnostics in crystal clarity. Examples from several popular image datasets and networks will be presented in collaboration with Drs. Hao Ho and Yu-cheng Li, Institute of Statistical Science, Academia Sinica. This talk is dedicated to the memory of my friend, colleague and informal career mentor, Don Ylvisaker.