SELF-HEALING ROBUST NEURAL NETWORKS VIA CLOSED-LOOP CONTROL

Zhuotong Chen[1], Qianxiao Li[2] and Zheng Zhang[1]

[1]University of California, Santa Barbara
[2]National University of Singapore

ABSTRACT

Despite the wide applications of neural networks in engineering and scientific computing, there have been increasing concerns about their vulnerability issue. While numerous attack and defense techniques have been developed, we investigate the robustness issue from a new angle: can we design a self-healing neural network that can automatically detect and fix the vulnerability issue by itself? A typical self-healing mechanism is the immune system of a human body. This biology-inspired idea has been used in many engineering designs (e.g., semiconductor chip design), but is rarely investigated in deep learning. We consider the post-training self-healing of a neural network, and propose a closed-loop control formulation to automatically detect and fix the errors caused by various attacks or perturbations. Since the self-healing method does not need a-priori information about data perturbations/attacks, it can handle a broad class of unforeseen perturbations.