ABBA NEURAL NETWORKS

JEAN-CHRISTOPHE PESQUET

We introduce ABBA networks, a novel class of (almost) non-negative neural networks, which are shown to possess a series of appealing properties. In particular, we demonstrate that these networks are universal approximators while enjoying the advantages of non-negative weighted networks. We derive tight Lipschitz bounds both in the fully connected and convolutional cases.

We propose a strategy for designing ABBA nets that are robust against adversarial attacks, by finely controlling the Lipschitz constant of the network during the training phase. We show that our method outperforms other state-of-the-art defenses against adversarial white-box attackers. Experiments are performed on image classification tasks on four benchmark datasets.