基于全局和单调递减鲁棒性策略的

鲁棒神经网络训练方法

梁震[1]，吴陶然[2,3]，刘万伟[4,5]，薛白[2]，杨文婧[1]，王戟[1]，庞征斌[4]
[1]国防科技大学量子信息研究所兼高性能计算国家重点实验室，中国长沙市，410073
[2]中国科学院软件研究所计算机科学国家重点实验室，中国北京市，100190
[3]中国科学院大学计算机科学与技术学院，中国北京市，100190
[4]国防科技大学计算机学院，中国长沙市，410073
[5]国防科技大学复杂系统软件工程实验室，中国长沙市，410073

摘要：深度神经网络的鲁棒性引发了学术界和工业界的高度关注，特别是在安全攸关领域。相比于验证神经网络的鲁棒性是否成立，本文关注点在于给定扰动前提下的鲁棒神经网络训练。现有的代表性训练方法——区间边界传播（IBP）和CROWN-IBP——在较小扰动下表现良好，但在较大扰动下性能显著下降，本文称之为衰退风险。具体来说，衰退风险是指与较小扰动情况相比，IBP系列训练方法在较大扰动情况下不能提供预期的鲁棒神经网络的现象。为了缓解这种衰退风险，我们提出一种全局的、单调递减的鲁棒神经网络训练策略，该策略在每个训练轮次考虑多个扰动（全局鲁棒性训练策略），并将其相应的鲁棒性损失以单调递减的权重进行组织（单调递减鲁棒性训练策略）。实验证明，所提策略在较小扰动下能够保持原有算法的性能，在较大扰动下的衰退风险得到很大程度改善。值得注意的是，与原有训练方法相比，所提训练策略保留了更多的模型准确度，这意味着该训练策略更加平衡地考虑了模型的鲁棒性和准确性。

关键词：鲁棒神经网络；训练方法；衰退风险；全局鲁棒性训练；单调递减鲁棒性
https://doi.org/10.1631/FITEE.2300059