基于连续扰动生成方法的可持续对抗训练

林巍[1,3]，廖丽娟[2]
[1]福建理工大学计算机科学与数学学院，中国福州市，350118
[2]西安理工大学经济与管理学院，中国西安市，710048
[3]福建理工大学福建省大数据挖掘与应用技术重点实验室，中国福州市，350118

摘要：基于在线生成对抗性样本的对抗性训练在防御对抗性攻击和提高卷积神经网络（CNN）模型鲁棒性方面取得良好效果。然而，大多数现有对抗训练方法都致力于寻找强对抗例子迫使模型学习对抗数据分布，这不可避免地增加了大量计算开销并导致干净数据丢失。本文展示了在不同训练世代中渐进式地增强对抗样本本身的对抗强度能有效提高模型鲁棒性，适当的模型转换可以保持模型泛化性能，且这一转换过程的计算成本可忽略不计。因此，本文提出一种针对对抗训练的连续扰动生成方法(SPGAT)，该方法通过在前一训练世代转移的对抗样本上添加扰动逐步增强对抗样本，并跨世代转换模型以提高对抗训练效率。实验表明，本文所提SPGAT方法既高效又有效；例如，所提方法计算时间为900分钟，标准对抗训练持续时间为4100分钟，对抗精度和干净样本精度性能提升分别超过7%和3%。在不同数据集上对SPGAT进行广泛评估，包括小规模MNIST、中等规模CIFAR-10和大规模CIFAR-100。实验结果表明，相比于目前最优方法，所提方法更有效。

关键词：对抗训练；对抗攻击；随机权重平均；机器学习；模型的泛化
https://doi.org/10.1631/FITEE.2300474