基于广义全变分低秩矩阵恢复的对抗样本防御

李文[1,2]，王恒友[1,5]，霍连志[3]，何强[1,5]，陈琳琳[1,5]，何志权[4]，吴永贤[2]
[1]北京建筑大学理学院，中国北京市，100044
[2]华南理工大学计算机科学与工程学院，中国广州市，510006
[3]中国科学院空天信息研究所，中国北京市，100094
[4]广东省智能信息处理重点实验室,，中国深圳市，518060
[5]北京建筑大学大数据建模与技术研究所，中国北京市，100044

摘要：一阶全变分（TV）正则化的低秩矩阵分解在恢复图像结构上表现出优异性能。利用全变分在图像去噪方面的优异性能，提高深度神经网络鲁棒性。然而，尽管一阶全变分正则化可以提高模型鲁棒性，但其过度平滑降低了干净样本的准确率。本文提出一种新的低秩矩阵恢复模型，称为LRTGV，该模型将广义全变分（TGV）正则化引入到重加权低秩矩阵恢复模型。在所构建的模型中，TGV可以在不过度平滑的情况下更好地重建图像纹理信息。重加权核范数和L_1范数可以增强全局结构信息。因此，本文所提出的LRTGV模型在破坏对抗噪声结构的同时能增强图像全局结构和局部纹理信息。为解决具有挑战性的最优模型问题，本文提出一种基于交替方向乘子法的算法。实验结果表明，该算法对黑盒攻击具有一定防御能力，并且在图像恢复方面优于现有低秩矩阵恢复方法。

关键词：广义全变分；低秩矩阵；交替方向乘子法；对抗样本
https://doi.org/10.1631/FITEE.2300017