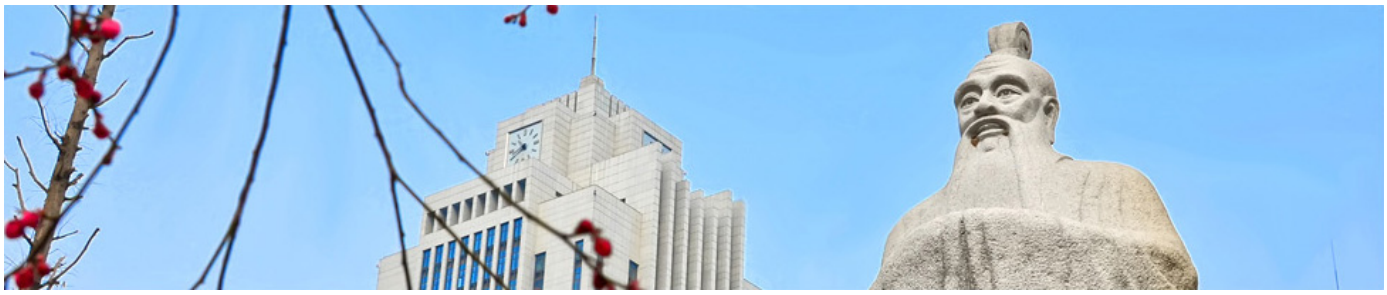
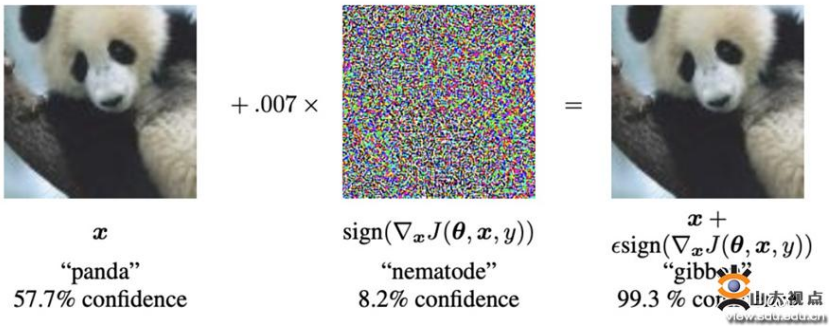


山大学生徐曦烈在国际机器学习顶级会议发表论文

发布日期：2020年08月05日 10:53 点击次数：2399

［本站讯］近日，山东大学泰山学堂计算机取向2017级本科生徐曦烈与新加坡国立大学Mohan Kankanhalli教授课题组合作，以共同第一作者在全球著名的机器学习会议ICML 2020（2020 International Conference on Machine Learning）发表了题为“Attacks Which Do Not Kill Training Make Adversarial Learning Stronger”的研究论文。

ICML是由国际机器学习学会（IMLS）主办的人工智能和机器学习领域的国际顶级学术会议，是中国计算机协会推荐A类会议（CCF-A）。近年来随着人工智能成为研究热点，ICML等顶级人工智能大会广泛受到学界、业界重视，文章评选门槛逐年抬高，今年录用率仅为20%左右，录选文章含金量大幅提升。



如今，深度神经网络应用已经非常广泛。近年来研究发现，深度学习模型容易受到对抗数据的攻击。对抗数据是由深度神经网络的输入自然数据和人工噪声合成而得，它可以轻易迷惑深度神经网络而不会被人类视觉系统识别错误。如上图所示，一张大熊猫图片在被加入噪声后，会被神经网络分类错误，并被认为是一张长臂猿图片。由此可见对抗数据的危害很大，尤其是对于无人驾驶、医疗诊断、金融分析这些安全性至关重要的领域。因此，提升神经网络的对抗鲁棒性变得十分重要。对抗学习是提升网络鲁棒性的重要方式之一。传统的对抗学习是根据极小极大公式，在内层通过最大化损失函数来寻找对抗数据，然后在外层学习对抗数据来最小化损失函数。但当前传统对抗学习存在很多问题，比如通过对抗训练而得的神经网络尽管有很大鲁棒性，但是以牺牲网络准确率为代价。

最新发布

- 寒假报道（9）：多措并举织密寒...
- 机械工程学院开展“家校协同育...
- 山东大学持续深入推进家校协同...
- 筑造梦港湾，山大二院青年文明...
- 春天里吹响山大创新创业教育的...
- 青岛校区开展走访慰问春节留校...
- 机械工程学院本科生党支部开展...
- 山东大学第二医院慰问健康守护者
- 材料科学与工程学院留校学生迎...
- 青岛市教育局局长刘鹏照一行来...

新闻排行

山东大学召开2020年度中层领导 ...
陈玉国教授团队在Nature Commun...
山东大学召开2021年校领导班子 ...
山大4基地入选基础学科拔尖学生...
山东大学召开2020年度学校领导 ...
Nature Communications发表高宁...
山大第13例，王子铭同学捐献造 ...
国家重点研发计划“工业窑炉协 ...
山东大学领导班子召开2020年度 ...
樊丽明一行检查在建基建工程并 ...

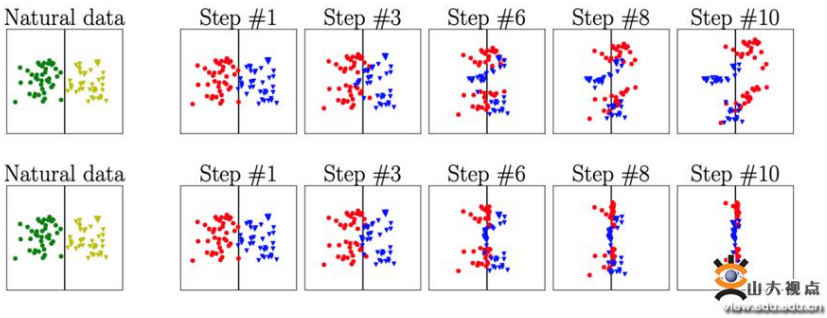
山大日记

山大人物

视点微信

互动话题

视点图志



在本论文中，作者指出传统对抗学习的问题出在极小极大公式，并提出了最小-最小公式，改变了内层的最大化损失函数，在内层寻找友好的对抗数据来解决这个问题。如上图所示，黄色和绿色的点是自然数据，红色和蓝色的点是对抗数据。第一行展示了传统对抗学习过程中在寻找极其恶劣的对抗数据，致使红色的点跑到了黄色区域，蓝色的点跑到了绿色区域，这样会导致自然数据和对抗数据严重的交叉混合问题。如第二行所示，作者通过寻找友好的对抗数据（在搜索对抗数据时，如果发现当前对抗数据被错误分类，则提前停止搜索。这样产生的对抗数据，称为友好的对抗数据），很好地缓解了上述交叉混合问题。

Table 1. Evaluations (test accuracy) of deep models (WRN-32-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$	PGD-100
Madry	87.30	56.10	45.80	46.80	-
CAT	77.43	57.17	46.06	42.28	-
DAT	85.03	63.53	48.70	47.27	-
FAT ($\epsilon_{train} = 8/255$)	**89.34** $\pm$ 0.221	65.52 $\pm$ 0.355	46.13 $\pm$ 0.409	46.82 $\pm$ 0.517	45.31 $\pm$ 0.531
FAT ($\epsilon_{train} = 16/255$)	87.00 $\pm$ 0.203	**65.94** $\pm$ 0.244	**49.86** $\pm$ 0.328	**48.65** $\pm$ 0.176	**49.56** $\pm$ 0.255

Results of Madry, CAT and DAT are reported in (Wang et al., 2019). FAT has the same evaluations.

Table 2. Evaluations (test accuracy) of deep models (WRN-34-10) on CIFAR-10 dataset

Defense	Natural	FGSM	PGD-20	C&W$_\infty$	PGD-100
TRADES ($\beta = 1.0$)	88.64	56.38	49.14	-	-
FAT for TRADES ($\epsilon_{train} = 8/255$)	**89.94** $\pm$ 0.303	**61.00** $\pm$ 0.418	**49.70** $\pm$ 0.653	49.35 $\pm$ 0.363	48.35 $\pm$ 0.240
TRADES ($\beta = 6.0$)	84.92	61.06	56.61	**54.47**	55.47
FAT for TRADES ($\epsilon_{train} = 8/255$)	**86.60** $\pm$ 0.548	**61.97** $\pm$ 0.570	55.98 $\pm$ 0.209	54.29 $\pm$ 0.173	[illegible] $\pm$ 0.291
FAT for TRADES ($\epsilon_{train} = 16/255$)	84.39 $\pm$ 0.030	61.73 $\pm$ 0.131	**57.12** $\pm$ 0.233	54.36 $\pm$ 0.177	[illegible] $\pm$ 0.155

Results of TRADES ($\beta = 1.0$ and 6.0) are reported in (Zhang et al., 2019b). FAT for TRADES has the same evaluations.

友好的对抗学习便是基于友好的对抗数据来进行学习，从而得到鲁棒的神经网络。友好的对抗学习具有以下好处：一是缓解了交叉混合问题。二是很省时，因为作者在训练中会提早停止搜索对抗数据，所以不需要更多的反向传递。三是能够实现更大的保护半径。四是如上表所示，极大提升了模型的准确率，并保持了模型鲁棒性，甚至提升了模型的鲁棒性。

泰山学堂十年来积极探索基础学科拔尖人才培养模式，通过“一制三化”即导师制、小班化、个性化、国际化培养，建立全面发展与个性发展相结合的因材施教的培养机制。以泰山学堂计算机取向为例，目前与剑桥大学、新加坡国立大学、新加坡南洋理工大学、加州大学洛杉矶分校、香港大学、华盛顿大学圣路易斯分校等国际一流院校建立了联合培养合作关系。该论文作者就是在学校支持下参加了去年新加坡国立大学暑期学校项目，通过与Mohan Kankanhalli教授课题组深入交流产生了该论文的想法。通过“请进来”，泰山学堂每年邀请海外著名学者为学生开设讲座或课程讲授前沿知识，同时选派优秀学生“走出去”参加海外交流学习，激发学生的专业兴趣，取得了良好成效。

文章链接：https://proceedings.icml.cc/static/paper_files/icml/2020/520-Paper.pdf

【供稿单位：本科生院 泰山学堂　作者：王蓓 徐曦烈　编辑：新闻网工作室　责任编辑：刘婷婷　】

相关阅读

- 网络空间安全学院6篇论文被国际对称密码...
- 山东大学承办2018年全省大学生思想政治...
- 【2015】山大学生在国际大学生iCAN中国...
- 【2017】Nature发表国际实验成果证实山...
- 青塔：近一个月，山大发表4篇顶级医学期...
- 【2015】山大与青岛高新区共建大学生创...

精彩视频

- Nature发表国际实验成果证实山大学者理...
- 山大学生会、研究生会疫情防控在行动
- 山东大学承办2018年全省大学生思想政治...
- 【2016】全国大学生地球物理知识竞赛在...
- 【2016】山大学生社团创新发展助推校园...
- 【2016】山大学生就业创业体验中心盈创...

验证码 7924 看不清楚,换张图片 提交

共0条评论 共1页 当前第1页 拖动光标可翻页查看更多评论