“相约星期四咖啡时间”之深度神经网络的后门攻击与防御

发布日期：2021-03-22 作者：刘洁 访问量： 261

2021年3月18日（周四）下午3：00，“相约星期四咖啡时间”如期举行。学院新入职教师陈艳姣研究员主持本周咖啡时间的交流，并为老师们带来了以“针对深度神经网络的后门攻击和防御”为主题的学术报告。

陈艳姣研究员首先介绍了当下人工智能在工业界应用越来越广泛的现状，并在此基础上引入机器学习中可能存在的后门攻击。攻击者寻求让机器学习模型学习结果中带有一个基于中毒样本的“后门”，可能是一个水印型模式，或者是不容易被察觉的事物，例如佩戴的眼镜。后门攻击使得分类器只在出现特定特征时才会进行不常规的分类，在其他情况下不会出现。另外，她介绍了目前较为常用的BadNets和Trojan等多种后门攻击方式，以及只使用部分网络信息即可完成攻击的Partial-Model-Based Attack。

陈艳姣研究员从以模型为基础的防御和以数据为基础的防御两个角度介绍了当前研究者针对后门攻击所尝试的防御方法。针对模型的防御方法目前可以从模型中恢复出可能存在的网络后门进而分析，也可以针对网络中神经元的激活情况分析当前是否被后门攻击。针对数据的防御方法目前可以通过数据消毒的方式对数据进行处理以达到良好的分类效果。最后，陈艳姣研究员介绍了自己提出的一种基于后门攻击的触发器生成方法，以及未来的进一步研究方向。

陈艳姣研究员的报告深入浅出，与工业界最新的应用相结合，给老师们带来了很多新的灵感启发。各位老师就机器学习技术在电气工程中的应用、深度学习未来的发展方向进行了探讨，并对各自的研究中所遇到的实际问题进行讨论，碰撞产生了很多新奇的学术想法，大家都表示相互交流中受到很大的启发。

“相约星期四咖啡时间”是学院为加强教师间合作交流打造的系列活动，每周四下午3点由学院一位老师主持，各位老师根据主题进行交流和讨论，3-5月主要由学院的新进校老师主持，在活动中增进交流，帮助新进校的老师更好的适应和融入新的工作环境。

下期相约星期四咖啡时间活动预告：


浙江大学 电气工程学院
COLLEGE OF ELECTRICAL ENGINEERING
ZHEJIANG UNIVERSITY
相约 | 星期四
MEET
THURSDAY
COFFEE
TIME
咖啡时间
罗皓泽 博士
交流内容
功率器件热状态感知原理与方法
时间 TIME
3月25日 下午3:00
地点 LOCATION
教二313 教工活动中心
罗皓泽博士，浙江大学百人计划研究员。研究工作主要集中于功率半导体器件封测与应用技术，包括功率半导体器件的应用特性分析、功率半导体器件的可靠性研究、功率半导体器件的封装与工艺开发等。

联系我们

电话：0571-87952707
传真：0571-87951625
地址：杭州市西湖区浙大路38号
邮编：310027

网站地图

院情总览
科学研究
管理登录

关于本站

浙江大学
浙江大学综合服务网
浙江大学图书馆
旧版回顾

官方微信