中国矿业大学
CHINA UNIVERSITY OF MINING AND TECHNOLOGY

学术报告

受中国矿业大学信息与控制工程学院邀请，浙江大学陈艳姣教授在我校举行学术报告。欢迎广大师生踊跃参加！

报告题目：基于注意力自蒸馏的深度学习模型后门防御技术

时　　间：北京时间 3 月 2日下午 1：30

地　　点：中国矿业大学南湖校区 信控学院A101报告厅

主办单位：信息与控制工程学院无线通信与智能感知研究所

报告人简介：陈艳姣，浙江大学百人计划研究员、博士生导师。2010年本科毕业于清华大学电子工程系，2015年博士毕业于香港科技大学计算机科学与工程系，曾任加拿大多伦多大学博士后、武汉大学研究员。主要从事智能物联网安全研究，在计算机网络和信息安全等领域国际权威期刊和会议上发表论文100余篇，入选“中国科协青年人才托举工程”。获得浙江省科学技术进步奖一等奖等。担任ACM CCS、USENIX Security、NDSS、IEEE INFOCOM等国际会议程序委员会成员。担任IEEETIFS等国际期刊编委。

报告摘要：深度学习模型被广泛应用于人脸识别、自动驾驶等重要领域。然而，最近研究表明，深层神经网络容易受到隐蔽的后门攻击。深度学习模型的后门不影响干净输入的正确识别，但在特定激活器触发下可以实现定向识别出错。本报告，我们将介绍一种利用注意力自蒸馏的深度学习模型后门防御技术，利用不易受后门影响的模型浅层来进行层间注意力正。与现有方法相比，该防御技术在触发器未知的条件下，对多种后门模式均可实现有效的后门清除。