第9讲

第9讲：智能汽车的安全挑战

牛温佳 教授

北京交通大学 · 网络空间安全学院

本讲内容

学习目标

理解对抗攻击的基本原理和分类

了解数字世界和物理世界中的对抗攻击

认识智能汽车面临的特有安全威胁

9.1 对抗攻击——AI的"视觉欺骗"

Section 9.1 对抗攻击——AI的"视觉欺骗"

9.1 对抗攻击——AI的"视觉欺骗"

9.1.1 什么是对抗攻击?

对抗攻击是指向正常输入中故意添加人眼几乎无法察觉的微小扰动，使得机器学习模型做出错误的预测。例如，在一张"停止"标志的图片上叠加特定的噪声，人类看起来仍然是"停止"标志，但AI模型却可能将其识别为"限速"标志。

对抗攻击的几个关键特点：

- **微小扰动**：对图像像素的改动极小，人眼难以察觉
- **模型错误**：扰动后模型输出严重偏离正确结果
- **攻击意图**：扰动是精心设计的，不是随机噪声

9.1.2 对抗攻击的分类

按攻击者的信息获取程度：

- **白盒攻击**：攻击者知道目标模型的全部信息（结构、参数、训练数据），可以精确计算梯度。经典方法包括FGSM、PGD、CW攻击等。

- **黑盒攻击**：攻击者只能通过"提问"（输入样本，获得输出）来猜测模型的弱点。分为查询攻击（反复向模型查询）和迁移攻击（利用另一个替代模型生成对抗样本）。

按攻击目标：

9.1 对抗攻击——AI的"视觉欺骗"（续）

- **非目标攻击**：只要让模型预测错误即可
- **目标攻击**：让模型预测为攻击者指定的特定类别

9.2 数字世界的对抗攻击

Section 9.2 数字世界的对抗攻击

9.2 数字世界的对抗攻击

快速梯度符号法（FGSM）

2014年由Goodfellow提出，是最简单、最经典的对抗攻击方法。它利用损失函数对输入样本的梯度方向，沿梯度上升方向一步修改样本，使损失最大化。虽然简单，但为后续更强大的攻击方法奠定了基础。

投影梯度下降法（PGD）

PGD被称为"最强的一阶攻击方法"。它在FGSM的基础上引入迭代和随机起点——从样本附近的随机点开始，多步小步长地沿梯度上升方向移动，每步后把对抗样本"投影"回允许的扰动范围内。

CW攻击

2017年由Carlini和Wagner提出，将噪声大小直接放入优化目标中，是目前最强大的白盒攻击之一，曾攻破许多被认为有效的防御策略。

9.2 数字世界的对抗攻击（续）

AutoAttack

集成了多种攻击方法，可以自动调参，是目前公认的最可靠的模型鲁棒性评估工具。

9.3 物理世界的对抗攻击

Section 9.3 物理世界的对抗攻击

9.3 物理世界的对抗攻击

数字攻击直接修改像素值，在真实世界中通过摄像头拍摄后，由于光照、视角、打印质量等因素的影响，攻击效果会大打折扣。物理对抗攻击就是要解决这个问题。

9.3.1 对抗补丁

Brown等人提出的"对抗补丁"是一种具有高度物理鲁棒性的攻击手段。它不追求全图微小扰动，而是在图像特定区域施加高强度的对抗噪声。这些补丁经过打印后，放置在真实场景中，依然可以有效欺骗AI系统。

例如，一张专门设计的黑白图案贴纸，贴在路牌上，可以让自动驾驶汽车完全忽略该路牌的存在。

9.3.2 对抗伪装

Duan等人提出的"对抗伪装"更加隐蔽——利用风格迁移技术，将攻击代码"画"成路牌上的铁锈、积雪或褪色痕迹。人类看起来是年久失修的旧路牌，但在AI眼中却是致命的陷阱。

9.3.3 3D对抗攻击

最新研究指出，对抗攻击已经扩展到三维空间。Cao等人提出了针对多传感器融合感知系统的物理攻击方法——生成可以3D打印的物体，放在路边既能欺骗摄像头，又能欺骗激光雷达，是第一个能让融合感知系统"失效"的物理攻击。

9.4 智能汽车安全面临的四大挑战

Section 9.4 智能汽车安全面临的四大挑战

9.4 智能汽车安全面临的四大挑战

1. **感知安全**：攻击者通过物理手段（贴纸、激光、伪装）干扰传感器，使系统"看不见"或"看错"障碍物
2. **决策安全**：攻击者诱使规划模块做出危险的路径决策
3. **执行安全**：攻击者通过CAN总线注入虚假控制指令
4. **通信安全**：攻击者拦截或篡改V2X（车-车/车-路）通信

延伸阅读

📖Goodfellow, I. J. et al. Explaining and Harnessing Adversarial Examples. ICLR, 2015.

📖Carlini, N. & Wagner, D. Towards Evaluating the Robustness of Neural Networks. S&P, 2017.

📖Brown, T. B. et al. Adversarial Patch. NIPS, 2017.

思考题

1. 1. 为什么物理世界的对抗攻击比数字攻击更有挑战性？为什么也更危险？

2. 2. 对抗补丁为什么能同时在数字和物理世界保持攻击效果？

3. 3. 如果你是自动驾驶的安全工程师，你会如何设计防御体系来抵御这些攻击？

📌本讲小结

- ✓ 9.1 对抗攻击——AI的"视觉欺骗"
- ✓ 9.2 数字世界的对抗攻击
- ✓ 9.3 物理世界的对抗攻击
- ✓ 9.4 智能汽车安全面临的四大挑战
- ✓ 思考与讨论

感谢聆听

第9讲 · 智能汽车的安全挑战