第10讲

第10讲：深度感知系统的鲁棒防御策略

牛温佳 教授

北京交通大学 · 网络空间安全学院

本讲内容

1. 10.1 传统防御方法有哪些?
2. 10.2 为什么传统防御对物理3D攻击"力不从心"?
3. 10.3 新的防御思路
4. 10.4 防御效果的可视化
5. 思考题与小结

学习目标

了解传统对抗防御方法及其局限性

理解为什么传统防御难以应对物理3D攻击

认识到需要全新防御范式的紧迫性

10.1 传统防御方法有哪些?

Section 10.1 传统防御方法有哪些?

10.1 传统防御方法有哪些?

10.1.1 数据增强

在训练阶段向数据集中加入各种变换（旋转、缩放、添加噪声、模拟雨雾等），希望模型见过足够多的"花样"，从而在真实场景中更具鲁棒性。

局限性：难以覆盖物理世界中无限变化的对抗纹理与光照组合，模型始终存在未被覆盖的盲区。

10.1.2 对抗训练

将对抗样本纳入训练集中一起训练，让模型"见过"对抗攻击的模式，期望它能学会抵抗。对抗训练在数字攻击场景中效果显著。

局限性：对抗训练计算成本极高——每个训练样本都需要先生成对抗样本。更重要的是，物理3D攻击的参数空间极高，几乎不可能覆盖所有攻击变体。

10.1 传统防御方法有哪些？（续）

10.1.3 输入预处理与去噪

在图像输入模型之前先进行滤波、去噪等预处理，期望能够滤除对抗扰动。

局限性：精心设计的对抗纹理可能是低频的、自然的图案（如迷彩涂装），与真实物体纹理高度相似。预处理难以在不破坏关键视觉信息的前提下"剥离"对抗特征。

10.2 为什么传统防御对物理3D攻击"力不从心"?

Section 10.2 为什么传统防御对物理3D攻击"力不从心"?

10.2 为什么传统防御对物理3D攻击"力不从心"？

以针对单目深度估计的3D全纹理攻击为例，它能成功的原因在于利用了传统深度神经网络的两大脆弱性：

1. **完全连续可微性**：传统CNN从输入到输出都基于连续的浮点运算和可导的激活函数，为攻击者提供了一条"高速公路"般的梯度回传路径。

2. **局部纹理依赖**：CNN的感受野更容易被高频纹理特征主导，而非物体的宏观三维几何结构。当对抗伪装覆盖物体表面时，CNN会将对抗性纹理误认为是真实的深度线索。

10.3 新的防御思路

Section 10.3 新的防御思路

10.3 新的防御思路

10.3.1 从架构层面寻求突破

既然现有的CNN架构"天生"容易受到对抗攻击的欺骗，一个根本性的思路是：**改变计算范式本身**。这正是量子类脑计算在防御中的潜力所在。

10.3.2 量子-脉冲混合防御的三个维度

1. 阻断梯度路径

在模型中引入脉冲神经元或参数化量子线路，破坏计算图的完全可微性。当攻击者无法获得精确梯度时，就无法有效地优化对抗纹理。

2. 高维特征隔离

10.3 新的防御思路（续）

利用量子特征映射，将图像特征映射到指数级庞大的量子希尔伯特空间中。在这个高维空间中，"真实物体特征"和"对抗噪声特征"可以被非线性地拉开距离，形成更鲁棒的决策边界。

3. 离散脉冲编码

用脉冲编码替代连续激活值，使攻击者的微小扰动难以在稀疏离散的脉冲流中传播放大。膜电位的泄漏机制还会让扰动信号随时间自然衰减。

10.4 防御效果的可视化

Section 10.4 防御效果的可视化

10.4 防御效果的可视化

研究表明，传统的CNN在面对全车对抗纹理覆盖时，深度估计图中目标车辆的轮廓几乎完全"消失"。而在量子-脉冲混合架构下，即使受到同样强度的攻击，目标车辆的深度轮廓依然部分可辨——攻击扰动从车辆区域被"驱散"到了背景区域。

这种**"攻击能量空间重分配"**——将扰动从安全关键目标区域引导至非关键背景区域——是新型防御范式的核心价值所在。

延伸阅读

📖Madry, A. et al. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR, 2018.

📖Athalye, A. et al. Obfuscated Gradients Give a False Sense of Security. ICML, 2018.

思考题

1. 1. 为什么说"完全可微"既赋予深度学习强大的学习能力，也使其容易受到对抗攻击?

2. 2. "对抗训练"和"改变模型架构"两种防御思路的根本区别是什么?

3. 3. 为什么把对抗扰动从目标区域"驱散"到背景也是一种有效的防御?

📌本讲小结

- ✓ 10.1 传统防御方法有哪些?
- ✓ 10.2 为什么传统防御对物理3D攻击"力不从心"?
- ✓ 10.3 新的防御思路
- ✓ 10.4 防御效果的可视化
- ✓ 思考与讨论

感谢聆听

第10讲 · 深度感知系统的鲁棒防御策略