Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

人工智能系统安全综述

王琛 硕士研究生

2020年4月12日

内容提要

- 背景简介
- 基本概念
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解人工智能系统安全基本思想
  - 2. 了解人工智能系统攻击手段与防御方法
  - 3. 了解人工智能系统安全的发展方向与应用领域

背景简介

- 人工智能
  - 人工智能是利用人为制造来实现智能机器或者机器上的智能系统，模拟、延伸和扩展人类智能，感知环境，获取知识并使用知识获得最佳结果的理论、方法和技术

- 人工智能系统
  - 人工智能系统是指由人工智能软件系统和相关配套硬件组成的系统，人工智能软件系统是以基于底层运行库、通用库和专用库等组件实现的机器学习框架和基于框架开发的模型为核心，结合应用开发框架及业务逻辑代码形成的一套软件系统

背景简介

- 人工智能安全体系架构

背景简介

- 信息系统安全CIA三要素
  - 机密性（Confidentiality）
    - 只有授权用户可以获取信息
  - 完整性（Integrity）
    - 信息在输入和传输的过程中，不被非法授权修改和破坏，保证数据的一致性。
  - 可用性（Availability）
    - 保证合法用户对信息和资源的使用不会被不正当地拒绝。

背景简介

- 人工智能安全风险
  - 网络安全风险
  - 数据安全风险
  - 算法安全风险
  - 信息安全风险
  - 社会安全风险
  - 国家安全风险

基本概念

基本概念

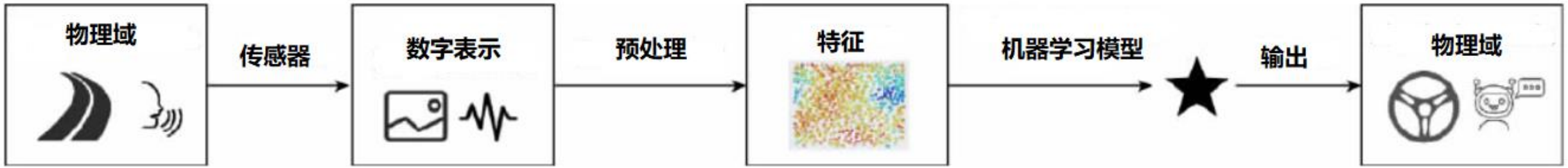
- 人工智能系统攻击手段分类
  - 攻击者掌握的情报
    - 白盒攻击
      - 指攻击者对于目标系统的详细信息十分了解，包括数据预处理方法、模型结构、模型参数等，甚至还能掌握部分的训练数据。
    - 黑盒攻击
      - 系统对于攻击者而言并不透明，关键细节都被隐藏，攻击者仅能够接触输入和输出环节
  - 攻击者的攻击阶段
    - 训练阶段攻击
    - 推断阶段攻击

基本概念

- 人工智能系统攻击目标分类
  - 根据攻击目标
    - 机密性攻击
    - 完整性攻击
    - 可用性攻击
  - 根据攻击效果
    - 目标攻击
    - 无目标攻击

基本概念

- 人工智能系统基本框架



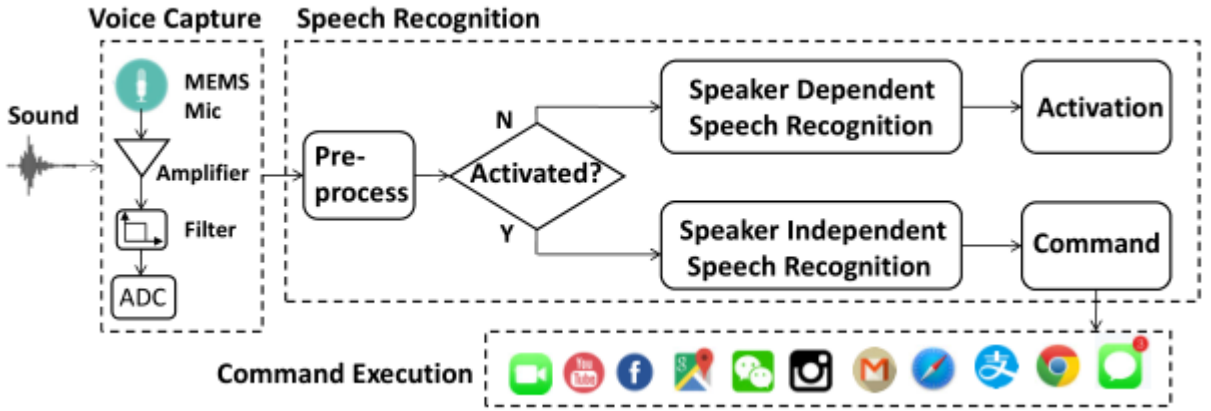
基本概念

- 输入环节攻击–传感器欺骗

T	干扰人工智能系统输入环节，从源头发起攻击
I	恶意攻击样本
P	针对传感器的工作特性，恶意构造相应的攻击样本并输送至传感器，造成人类和传感器对数据的感知差异，达成欺骗效果
O	错误的系统输出

基本概念

- 传感器欺骗-举例
  - Zhang Guoming, Chen Yan, Ji Xiaoyu, et al. DolphinAttack:Inaudible voice commands[C].CCS2017
  - 原理：人耳能听到的声音频率：20-20KHz，麦克风录制范围上限为24KHz，攻击者可以在麦克风超出人类听觉的接收频率范围内发送声音信号，从而使得设备能够感知而不被听众察觉
  - 结果：激活Siri启动iPhone上的FaceTime通话，激活谷歌Now将手机切换到飞行模式，甚至操纵奥迪汽车的导航系统



基本概念

- 传感器欺骗防御
  - 硬件防御
    - 传感器增强
  - 软件防御
    - 根据语音命令的独特特征，识别攻击信号

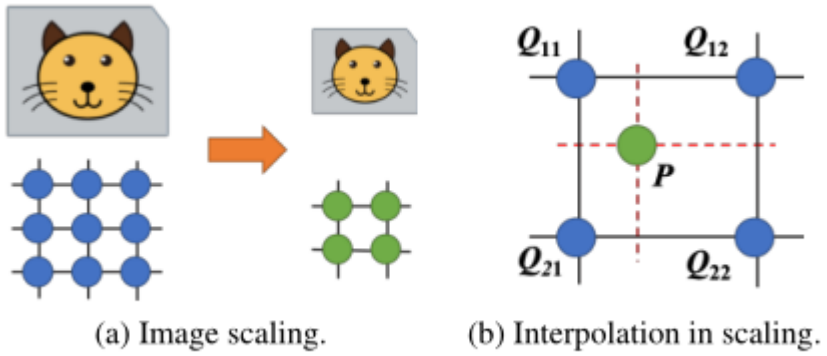
基本概念

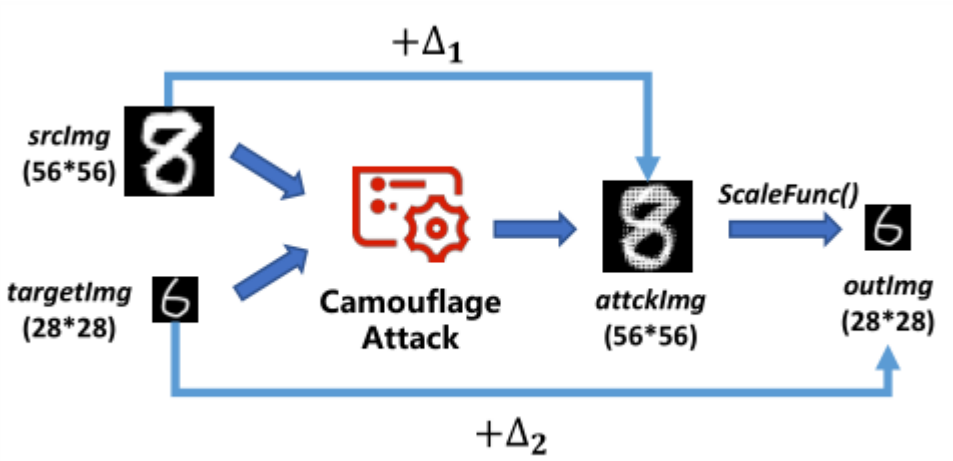
- 数据预处理环节攻击-重采样攻击
  - 重采样
    - 改变数据信息格式
    - 信息压缩

T	干扰人工智能系统数据预处理环节，从源头发起攻击
I	恶意攻击样本
P	在样本中隐藏恶意信息，预处理后进行攻击
O	错误的系统输出

基本概念

- 重采样攻击-举例
  - Xiao Qixue, Chen Yufei, Shen Chao, et al. Seeing is not believing: Camouflage attacks on image scaling algorithms [C] USENIX 2019
  - 原理：针对图像缩放算法进行攻击，缩放后产生语义发生巨大变化的视觉图像





基本概念

- 重采样攻击–防御
  - 攻击预防
    - 忽略大小与深度学习模型使用的输入大小不同的输入
    - 在缩放图像之前随机地从图像中删除一些像素(按行或按列)
  - 攻击检测
    - 检测在缩放过程中输入特征的明显变化，如颜色直方图和颜色散射分布

基本概念

- 机器学习模型攻击
  - 安全风险
    - 数据投毒
    - 对抗样本
  - 隐私风险
    - 模型提取
    - 模型逆向

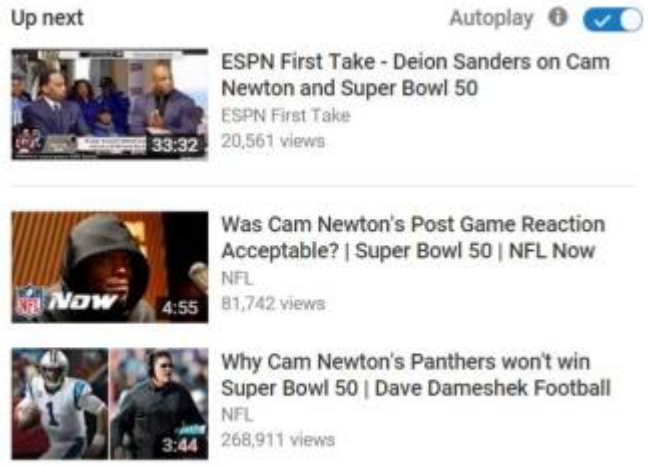
基本概念

- 数据投毒
  - 原理：攻击者通过修改训练数据内容和分布，来影响模型的训练结果
  - 举例：推荐系统攻击
  - Yang Guolei, Gong N Zhengiang, Cai Ying. Fake covisitation injection attacks to recommender systems [C] NDSS 2017

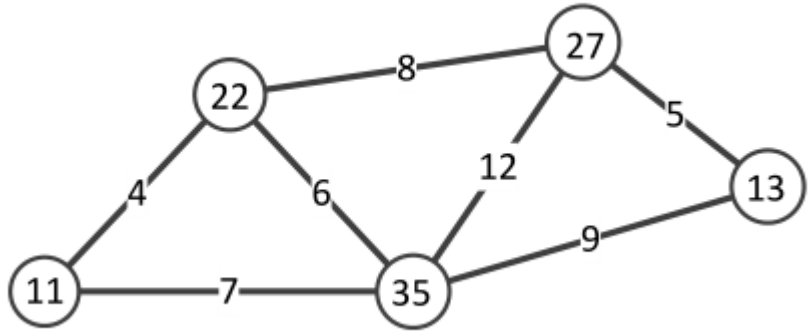
T	欺骗推荐系统，按照攻击者意图进行推荐
I	虚假联合访问
P	在系统中注入虚假的联合访问
O	攻击者设定的推荐信息

基本概念

- 数据投毒–举例
  - 关键思想：
    - 将攻击建模为受约束的线性优化问题
    - 使用脚本进行联合访问注入



YouTube



联合访问图

基本概念

- 关键难点
  - 如何选择合适的注入项目?
  - 如何选择注入次数?

$$\text{maximize } IUI = \sum_{j \in V_k} a_j \cdot p_j \tag{8}$$

$$\text{subject to} \sum_{j \in V_k} a_j \cdot m_{jk} \leq m \tag{9}$$

$$s'_{ji_t} > s'_{jk_j}, \quad \forall j \in V_k \tag{10}$$

$$w_{i_t} + m_{jk} \geq \tau, \quad \forall j \in V_k \tag{11}$$

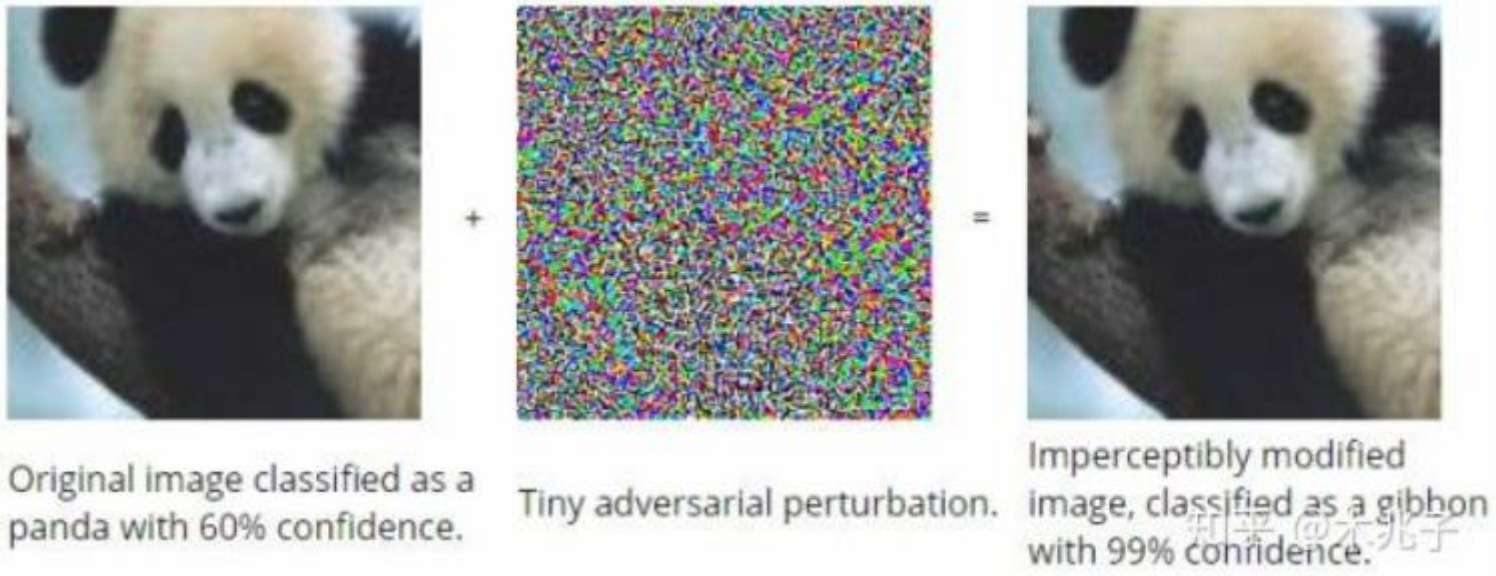
$$a_j \in \{0, 1\}, \quad \forall j \in V_k \tag{12}$$

基本概念

- 数据投毒-防御
  - 核心思想：考虑污染数据与正常数据的差异
  - 防御措施
    - 数据清洗
      - 增加访问限制
    - 提高模型的鲁棒性
      - 检测虚假访问

基本概念

- 对抗样本
  - 对抗样本由Christian Szegedy等人提出，是指在数据集中通过故意添加细微的干扰所形成的输入样本，导致模型以高置信度给出一个错误的输出。
  - 人类完全无法识别对抗样本，可是深度学习模型会以高置信度将它们进行分类。



基本概念

- $y = w^T * x$
- $x' = x + t$
- $w^T * x' = w^T * x + w^T * t$
- 当W和t维数很大时，即使很小扰动，累加起来也很可观

Lets fool a binary linear classifier:

x	2	-1	3	-2	2	2	1	-4	5	1
w	-1	-1	1	-1	1	-1	1	1	-1	1
adversarial x	1.5	-1.5	3.5	-2.5	2.5	1.5	1.5	-3.5	4.5	1.5

← input example

← weights

class 1 score before:

-2 + 1 + 3 + 2 + 2 - 2 + 1 - 4 - 5 + 1 = -3

=> probability of class 1 is 1/(1+e^(-(-3))) = 0.0474

-1.5+1.5+3.5+2.5+2.5-1.5+1.5-3.5-4.5+1.5 = 2

=> probability of class 1 is now 1/(1+e^(-(2))) = 0.88

i.e. we improved the class 1 probability from 5% to 88%

$$P(y=1 \mid x; w, b) = \frac{1}{1+e^{-(w^T x+b)}} = \sigma(w^T x + b)$$

基本概念

- 对抗样本攻击举例-FGSM（Fast Gradient Sign Method）
  - Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

T	生成对抗样本，使模型分类错误
I	原始样本x
P	在原始样本中添加微小的扰动，最大化误差函数，对分类结果产生最大化变化
O	对抗样本$\tilde{x}$

基本概念

- FGSM（Fast Gradient Sign Method）
  - 原始样本：x
  - 对抗样本：$\tilde{x}$=x+η
  - 约束条件：$\|\eta\|_\infty < \varepsilon$
  - 将修改后的图像输入分类模型中，x与参数矩阵相乘。

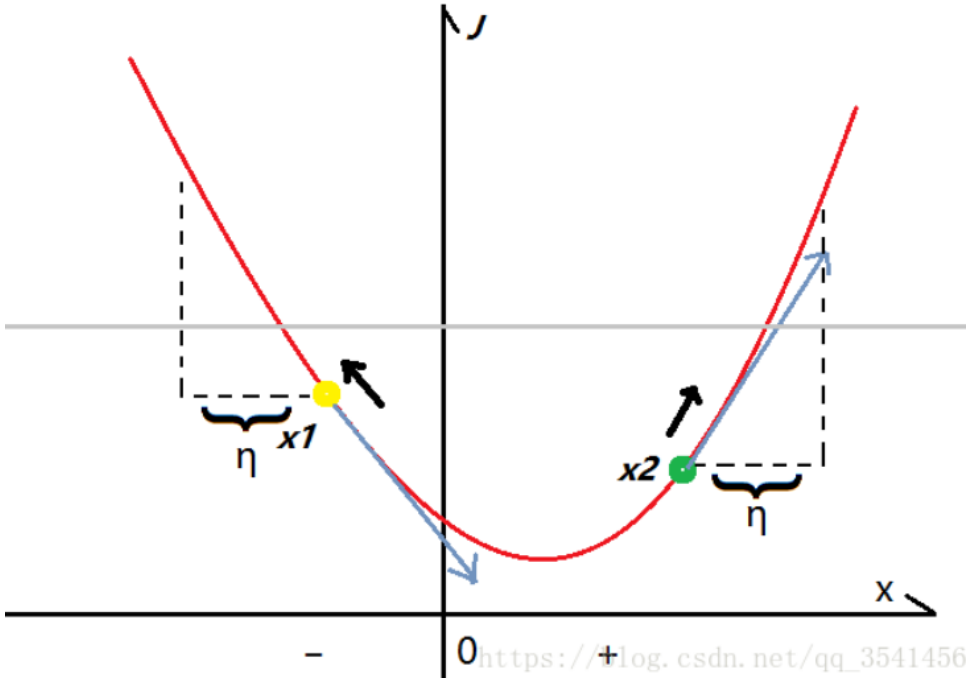
$$\boldsymbol{w}^{\boldsymbol{T}}\tilde{x} = \boldsymbol{w}^{\boldsymbol{T}}\mathrm{x} + \boldsymbol{w}^{\boldsymbol{T}}\eta$$

基本概念

- FGSM（Fast Gradient Sign Method）

$$\eta = \varepsilon sign\left(\nabla_x J\left(\theta, x, y\right)\right)$$

- 模型参数：θ
- 模型输入：x
- 结果标签：y
- 损失函数：J(θ,x,y)
- 符号函数：sign（）
  - sign函数可以保证变化方向和梯度函数方向一致



基本概念

- 对抗样本防御方法
  - 对抗训练
  - 梯度掩模
  - 对抗样本检测

基本概念

- 模型提取

T	还原原始模型或模拟等价模型
I	大量的预测数据
P	利用人工智能系统提供的API，输入预测数据，接收类标签和置信度系数，计算模型参数
O	模型参数

基本概念

- 模型提取举例
  - 输出： f(x)
  - 参数： w, b
  - 对f(x)求反函数，得到一个包含n+1个参数的线性函数，随机访问n+1次得到方程组求解

特征向量 n = 92 * 112 = 10,304

$$f(\mathbf{x}) = 1/(1+e^{-(w*x+b)})$$

$$\ln\left(\frac{f(x)}{1-f(x)}\right) = w*x+b$$

基本概念

- 模型提取防御
  - 对模型参数和输出结果进行近似处理
  - 模型水印

基本概念

- 模型逆向
  - 根据训练数据与非训练数据的拟合差异来窥探训练数据隐私

T	还原模型训练数据
I	测试样本和模型对应的置信度输出
P	利用人工智能系统提供的API，输入测试样本，根据置信度系数，逼近原始数据
O	模型训练数据

基本概念

- 模型逆向举例
  - FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C] CCS 2015

Algorithm 1 Inversion attack for facial recognition models.

function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
  $c(\mathbf{x}) \stackrel{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x}) + \text{AUXTERM}(\mathbf{x})$
  $\mathbf{x}_0 \leftarrow \mathbf{0}$
  for $i \leftarrow 1 \ldots \alpha$ **do**
    $\mathbf{x}_i \leftarrow \text{PROCESS}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
    if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \ldots, c(\mathbf{x}_{i-\beta}))$ **then**
      break
    if $c(\mathbf{x}_i) \leq \gamma$ **then**
      break
  return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$

基本概念

- 模型逆向防御
  - 同态加密
  - 差分隐私
  - 安全多方计算

应用总结

应用总结

- 人工智能安全应用
  - 网络防护
  - 数据管理
  - 信息审查
  - 智能安防
  - 金融风险
  - 舆情监测

应用总结

- 未来发展
  - 物理对抗样本
  - 更有效的对抗训练
  - 模型鲁棒性的形式化验证
  - 人工智能系统自动化测试方法
  - 隐私保护

参考文献

[1] Zhang Guoming, Chen Yan, Ji Xiaoyu, et al. DolphinAttack:Inaudible voice commands[C].CCS2017

[2] Xiao Qixue, Chen Yufei, Shen Chao, et al. Seeing is not believing: Camouflage attacks on image scaling algorithms [C] USENIX 2019

[3] Yang Guolei, Gong N Zhengiang, Cai Ying. Fake covisitation injection attacks to recommender systems [C] NDSS 2017

[4] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[5]中国信息通信研究院《人工智能安全白皮书(2018)》

道德经

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！