软件分析与测试
智能系统测试技术

严俊

中国科学院大学
University of Chinese Academy of Sciences
ISCAS 中国科学院软件研究所
Institute of Software, Chinese Academy of Sciences

Artificial Intelligence (Wikipedia)

- **Artificial intelligence** (**AI**) is **intelligence** demonstrated by **machines**, in contrast to the **natural intelligence** (**NI**) displayed by **humans and other animals**.
- In computer science AI research is defined as the study of "intelligent agents": **any device** that **perceives its environment** and **takes actions** that maximize its chance of successfully achieving its goals.

House Prices Prediction

面积	区域	学区房	商圈距离	地铁	价格
100	1	1	0.5	1	100
70	1	1	0.5	0	80
150	0	0	1	1	60
89	0	0	0	0	50
100	0	1	1	0	?

Feature/Representation
x

Target
y

- Hypotheses(假设/模型)
  - $y=w\cdot x+b$
- Loss Function

$$\sum_i^n(\hat{y}_i - y_i)^2 = \sum_i^n(\hat{y}_i - (w * x_i + b))^2$$

- •Optimization
  - Gradient descent
  - Analytical solution

Deep Learning

- Traditional Pattern Recognition: Fixed/Handcrafted Feature Extractor

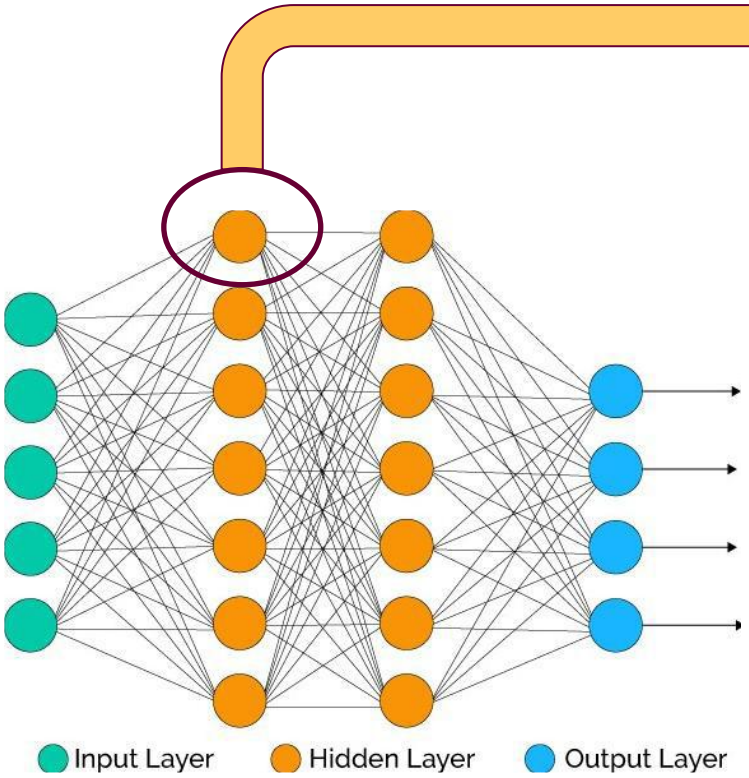
- Deep Learning: Representations are hierarchical and trained

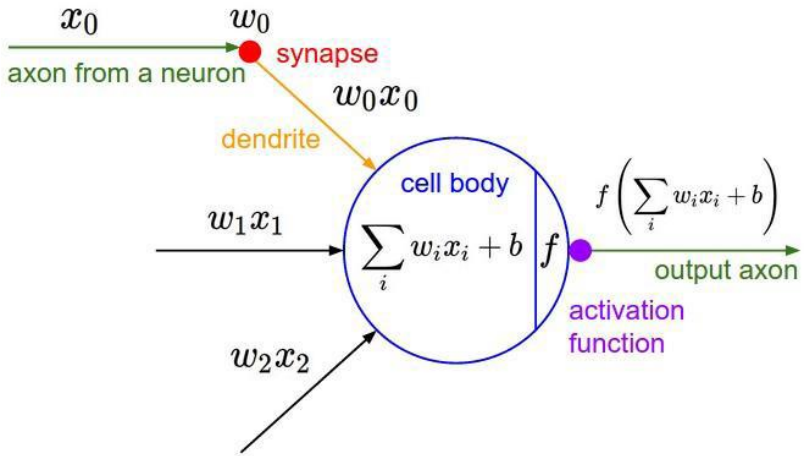
Deep Learning Approaches

- Deep neural network (DNN)
  - Fully-connected neural network
  - Convolutional Neural Network
  - Recurrent Neural Network

Fully-Connected Neural Network

Multi-layer nonlinear transformations



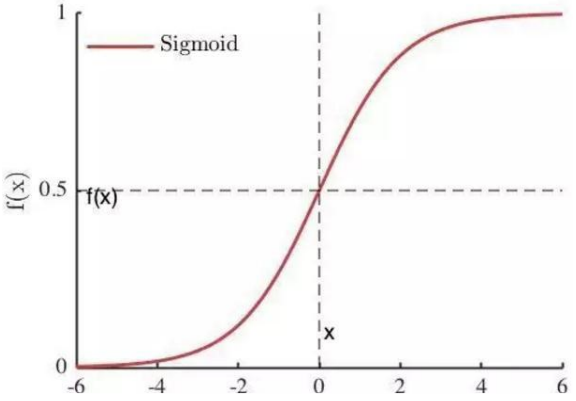


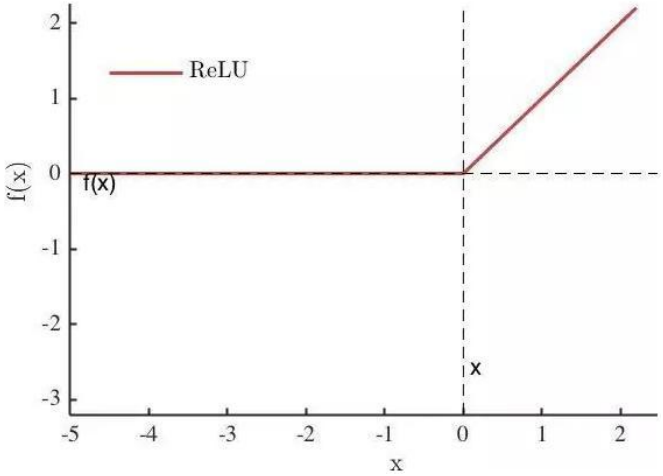
Activation Functions

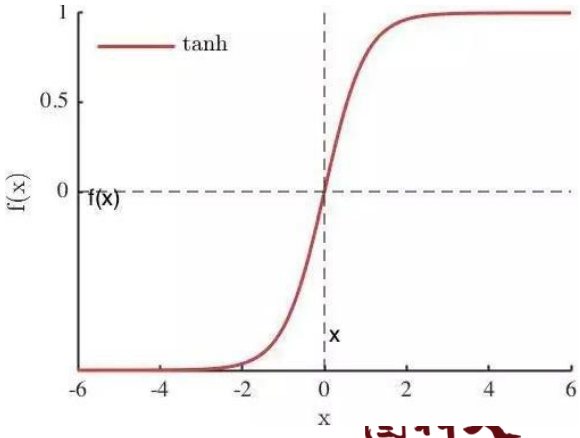
Sigmoid $f(x) = \frac{1}{1+e^{-x}}$

Tanh $\tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$

Relu $f(x) = \max(0,x)$







人工智能系统

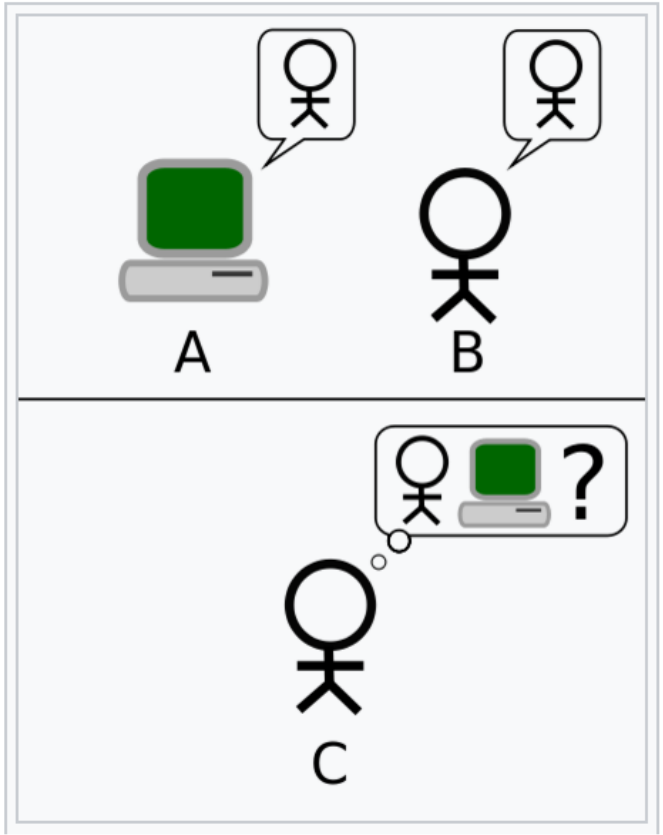
- 高度复杂的非线性系统
- 缺乏可解释性
- 效果和训练集相关

今天的内容

- 智能系统测试的一些探索
  - 测试规约
  - 测试覆盖率
  - 测试预言
- 自动驾驶系统的测试

图灵测试

- 假如你来提问，你会问什么问题
- 假如你来判定，你应该何时终止
- 假如你来回答，你应该怎么回答



软件测试的要素

- 如何设计测试用例
  - 白盒测试
  - 黑盒测试
  - 基于规范的测试
- 如何评估测试充分性
  - Test coverage
- 如何判断某个测试用例是否通过
  - Test Oracle

1.1 如何设计测试用例？

系统的规范

- 需要行业内人士的参与
- 完整地描述系统的行为很难
- 如果没有规范，可以采用
  - 组合测试
  - 随机测试
  - ……

Formal Specification for Deep Neural Networks

- Semantic classification
  - **System-level specification:**
    - over the entire system, e.g. automatic emergence braking system
  - **Input-output robustness:**
    - whether the model is robustness with adversarial perturbations
  - **Input-output relation:**
    - The guarantee of post-condition with a pre-input on the inputs
  - **Semantic invariance**
    - Equivalence class partition for the input space
  - **…**
- Trace-theoretic classification
  - **Trace properties: specified in LTL or MTL for stateful NNs**
  - **Hyperproperties: security policies**

Seshia, S. A., Desai, A., Dreossi, T., Fremont, D. J., Ghosh, S., Kim, E., ... & Yue, X. Formal specification for deep neural networks. In *ATVA 2018* (pp. 20-34).

Scenic: Language-Based Scene Generation

- Domain-specific probabilistic programming language is used for data synthesis
- Spectrum of scenarios can be defined from general to specific
- Corner cases could be specified and synthesized

Fremont, D., Yue, X., Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A.L., Seshia, S.A.: Scenic: Language-based scene generation. Technical report UCB/EECS-2018-8. EECS Department, UC, Berkeley, 2018

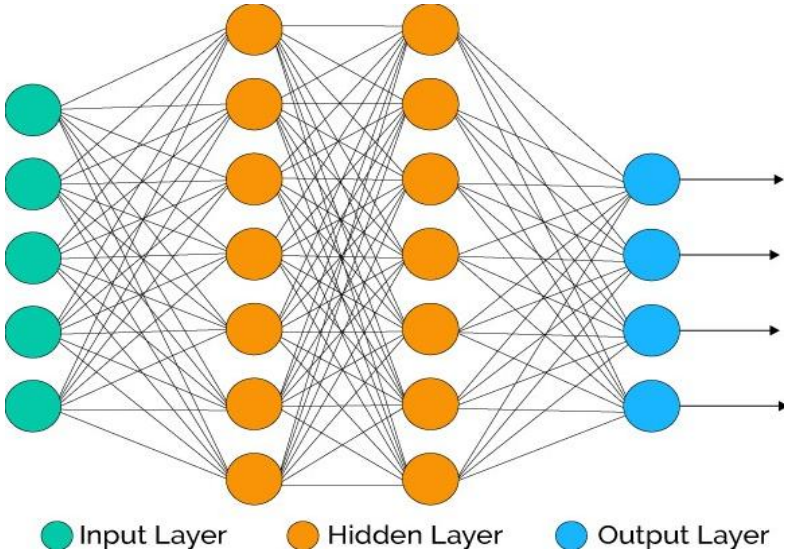
测试数据生成方法

- 采用现有的测试集合成
- 路径分析 + SMT
- Random / Fuzzing
- 各种黑盒白盒覆盖
  - 神经元（组合）覆盖

1.2 如何评估测试充分性

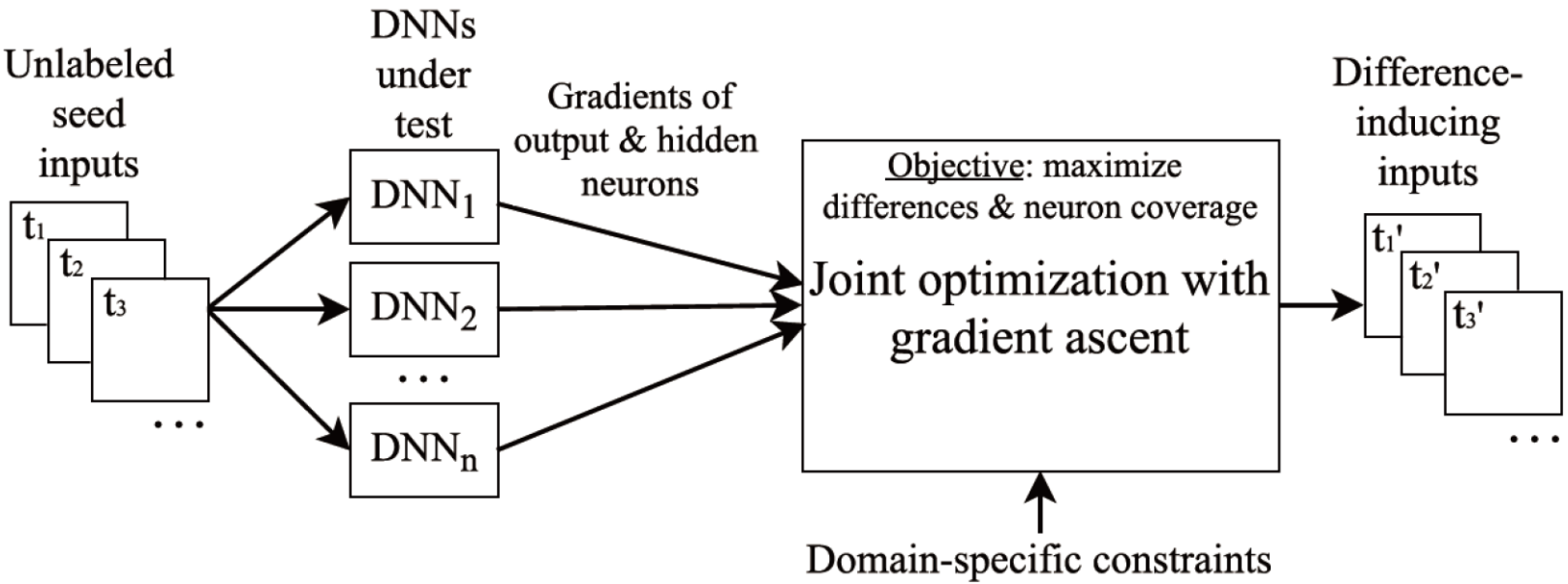
测试充分性

- 基于代码的覆盖率
- 基于条件的覆盖率[CH18]
- 基于结构的覆盖率
  - 神经元覆盖[PC17]



[CN17] Chih-Hong Chen,g Chung-Hao Huang, Hirotoshi Yasuoka. Quantitative Projection Coverage for Testing ML-enabled Autonomous Systems. 126-142 2018 ATVA
[PC17] Kexin Pei, Yinzhi Cao, Junfeng Yang, Suman Jana. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. SOSP 17

Test adequacy [PC17]



[PC17] Kexin Pei, Yinzhi Cao, Junfeng Yang, Suman Jana. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. SOSP 17

组合测试

- 将一层的神经元的输出离散化作为运行时的状态，用组合测试的方法对不同层的神经元相互作用进行测试。

- 同时定义了*t*-way的sparse、dense覆盖性等。

Lei Ma, Fuyuan Zhang, Minhui Xue, Bo Li, Yang Liu, Jianjun Zhao, Yadong Wang: Combinatorial Testing for Deep Learning Systems. CoRR abs/1806.07723 (2018).

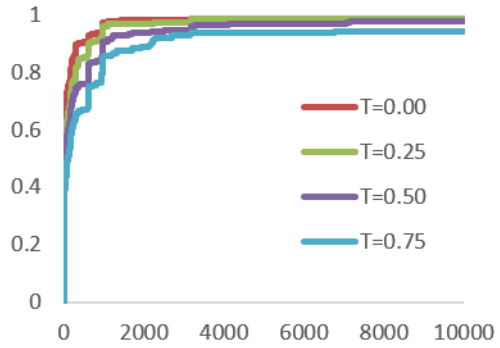
其他的覆盖标准

- 一组scalable、能以各种粒度级别监视和评估神经元活动和连接的测试标准。
- Neuron-Level
  - *k*-multisection Neuron Coverage
  - Neuron Boundary Coverage
  - Strong Neuron Activation Coverage
- Layer-Level
  - Top-*k* Neuron Coverage
  - Top-*k* Neuron Patterns

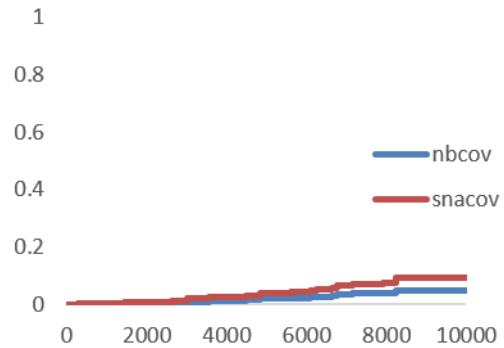
Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, Jianjun Zhao, Yadong Wang: DeepGauge: multi-granularity testing criteria for deep learning systems. ASE 2018.

神经网络覆盖率实验评估 LeNet-5

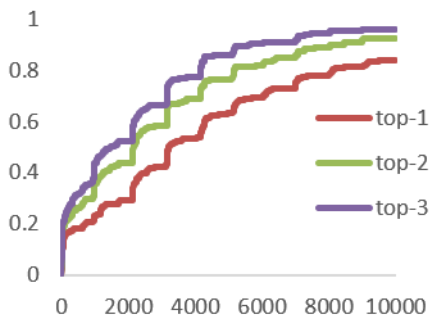
层次	网络结构	NCov	2-W Spar.	2-W Den.	(0.5, 2)-C	(0.75,2)-C	KMNC		NBcov	SNAcov	TKNC		
							k=100	k=1000			k=1	k=2	k=3
1	ReLU(conv(5,5,6))	97.49					68.41	47.56	7.34	14.60	13.16	15.52	17.58
2	MaxPool(2,2)	98.98					74.36	55.09	7.61	14.97	18.28	21.94	24.49
3	ReLU(conv(5,5,16))	100.0					79.53	53.85	7.97	15.94	31.13	39.81	46.00
4	MaxPool(2,2)	100.0					85.84	66.68	7.75	15.50	50.00	58.50	64.00
5	ReLU(fc(120))	99.17	96.68	98.75	100.0	98.33	88.66	69.07	8.33	16.67	84.17	92.50	93.33
6	ReLU(fc(84))	100.0	99.54	99.88	100.0	100.0	94.35	80.83	10.11	20.24	79.76	92.86	92.86



NCov全局神经元累积覆盖率



NBcov & SNACov神经元覆盖率累积增量图



Top-k全局神经元累积覆盖率

覆盖率指导的测试集约减

NCov神经元覆盖	最小测试集	覆盖率
NCov(T=0.00)	7	98.79%
NCov(T=0.25)	8	98.79%
NCov(T=0.50)	13	97.58%
NCov(T=0.75)	13	95.97%

Top-k 神经元覆盖	最小测试集	覆盖率
k=1	122	83.37%
k=2	80	92.33%
k=3	60	95.96%

组合覆盖率	最小测试集		覆盖率
2-Way Den	Layer5	88	97.47%
	Layer6	31	99.97%

K-multisection 神经元覆盖	最小测试集	覆盖率
KNMCov(k=100)	1532	86.81%

神经元边界覆盖	最小测试集	覆盖率
NBCov	17	4.83%
SNACov	16	9.27%

1.3 如何判断某个测试用例是否通过

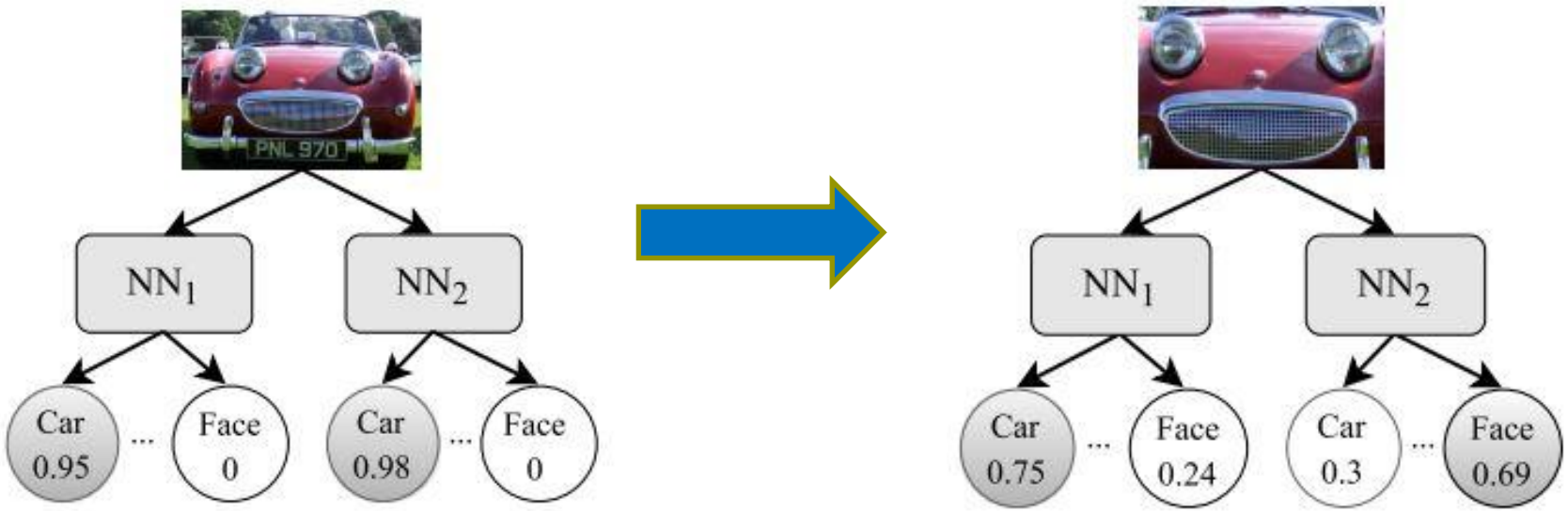
蜕变测试 (Metamorphic Testing)

- 以图像识别为例
  - 一幅图片，加入少量扰动，结果不变
  - 图像经过特定的处理，识别结果不变
    - 晴天、雨天、雪天……
- 目前使用较多的人工智能测试技术
- 主要研究集中在手工构造 Metamorphic Relation

[XH09] Xiaoyuan Xie, et al., Application of Metamorphic Testing to Supervised Classifiers. QSIC 2009: 135-144

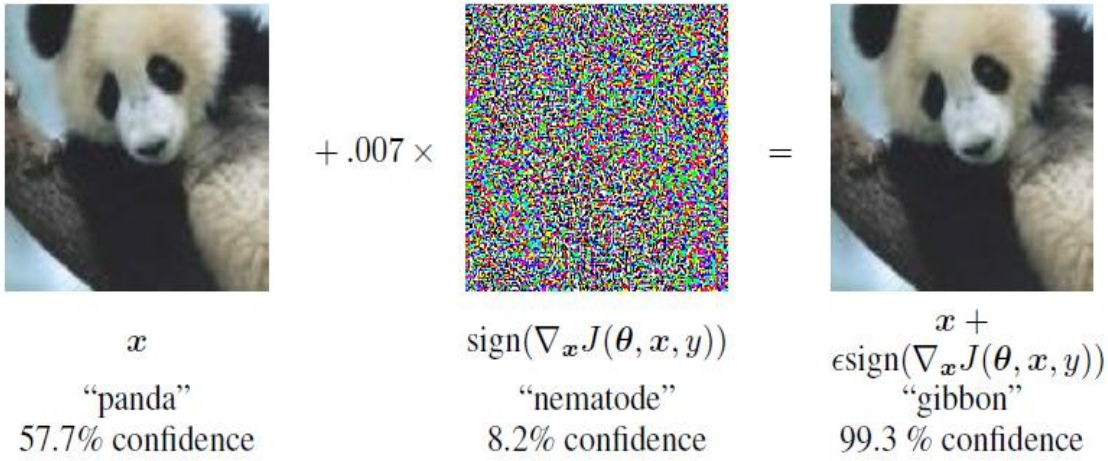
差分测试 Differential Testing

- If two DL systems with the same functionality generate different results, there might be mistakes



对抗样本

- AI系统并不能完全准确地判断样本，存在安全隐患
- 生成使模型预测错误的样本，找到模型的缺陷
- 多样本融合和样本变异方法是一种对抗样本生成方法



Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. CoRR abs/1412.6572 (2014)

案例：Cracking Google's phishing pages filter (GPPF)

- crack the classification model of GPPF and extract sufficient knowledge from it, including the classification algorithm, scoring rules and features, etc
- all the phishing pages (100%) can be easily manipulated to bypass the detection of GPPF

Bin Liang, Miaoqiang Su, Wei You, Wenchang Shi, Gang Yang: Cracking Classifiers for Evasion: A Case Study on the Google's Phishing Pages Filter. WWW 2016: 345-356

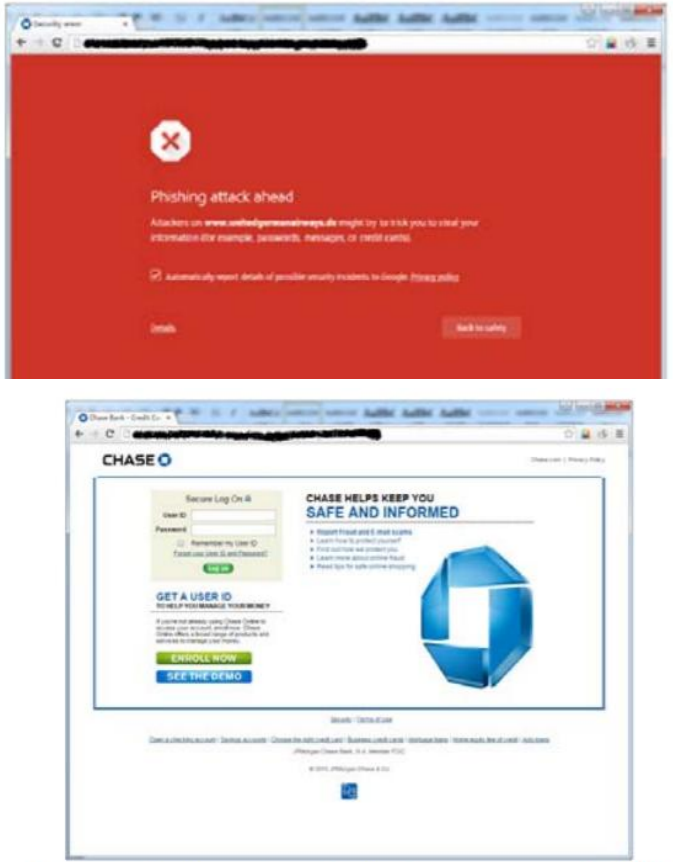


Figure 4. The dressed-up phishing page can evade GPPF.

案例：Attacking automatic speech recognition (ASR) systems

- attacks can even be automatically construct
- the voice commands can be stealthily embedded into songs

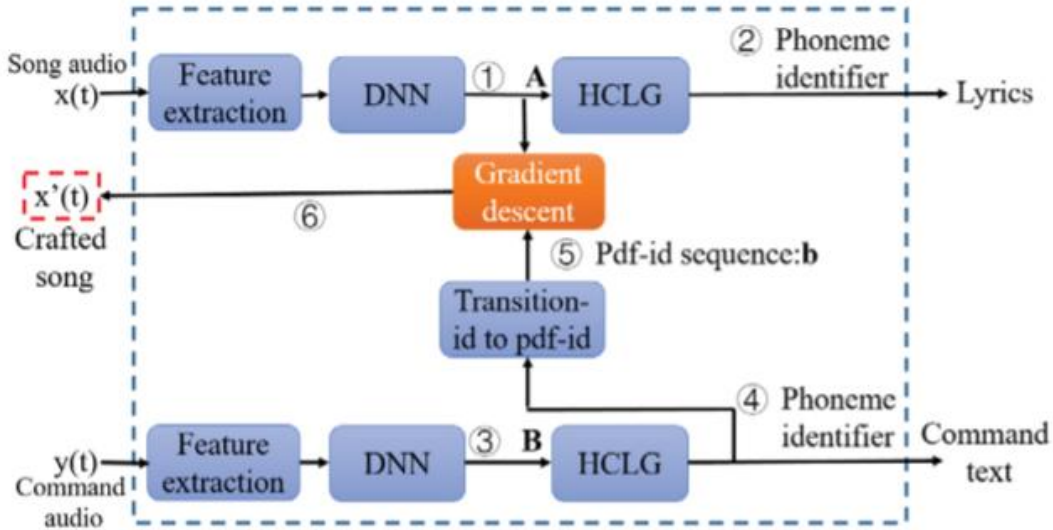


Figure 3: Steps of attack.

Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, Carl A Gunter, "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition", USENIX Security, 2018

2. 自动驾驶系统测试

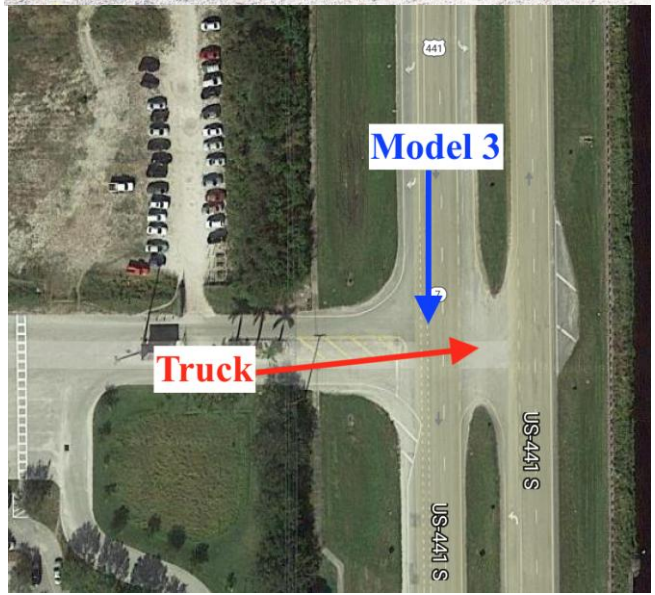
自动驾驶系统测试

- 自动驾驶的安全性

- 一些测试案例
  - 采用对抗样本的决策系统测试
  - 目标感知系统的运行时监控
  - 场景仿真测试

特斯拉事故

- 事故原因
  - 白色集装箱
  - 摄像头过曝
  - 卡车底盘高于毫米波雷达
- 同类事故发生两起
  - 2016.5
  - 2019.5

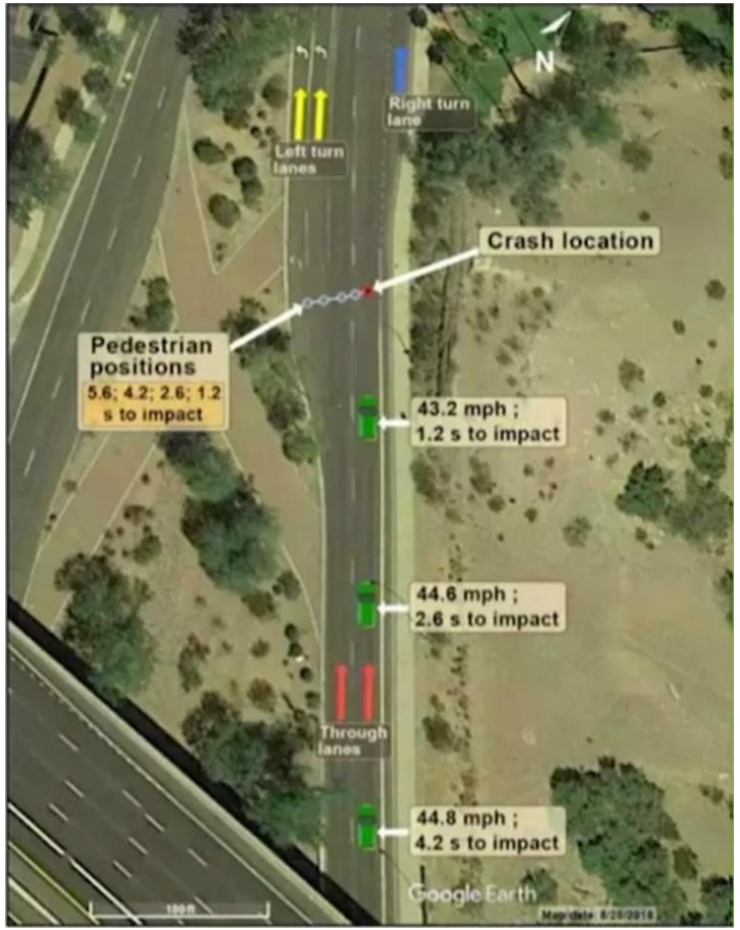


Autopilot was active when a Tesla crashed into a truck, killing driver
https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

Uber事故 2018.3

- 事故发生时
  - 驾驶员干预情况下，或可避免
  - 车速65km/h, 无减速行为，撞向行人
  - 车辆装载了LiDAR
- 发生了什么事情?

距离碰撞前时间（s）	系统判定结果
5.2	其他
4.2	车辆
2.7-3.8	车辆和其他之间摇摆
2.6	自行车
1.5	未知
1.2	自行车



国内自动驾驶

2021年7月8日，在世界人工智能大会上，当时还是华为自动驾驶项目负责人的苏箐就十分直白的表述了当前“自动驾驶”高事故率的观点。

2021年8月18日，林某某驾驶电动汽车，在沈海高速公路追尾碰撞前方同车道的轻型普通货车，造成电动汽车驾驶人当场死亡。

“当时交警请蔚来说明车辆信息时，蔚来上海总部的工程师已向蔚来福建闽南区负责人电话确认，林文钦驾驶的车辆在发生事故时正在高速公路上处于自动驾驶状态。”

2022年8月10日，浙江宁波一辆小鹏P7突发高架追尾交通事故，事故导致一人身亡。据报道，浙江宁波一辆小鹏P7疑似因在高架桥上使用智能驾驶辅助功能，在行驶过程中与前方检查车辆故障的人员和故障车辆相撞，事故导致一名前车乘客死亡。

2.1 决策系统测试

对抗样本测试

- DeepXplore：通过基于梯度下降的办法去提高神经覆盖率，找到触发不同行为的输入的过程。
- DeepTest：通过给原始图片加一些黑点、雾、雨等干扰来自动生成图片来模仿驾驶场景，来检测系统是否会做出正确的判断。
- 问题：合成的图片都是现实中不可能出现的场景。

(a) Patch

(b) Holes

(c) Translation

(d) Fog

(e) Rain

Figure 1: Driving scenes synthesized by DeepXplore (a)(b) and DeepTest (c)(d)(e)

生成对抗网络 (GAN)

- 通过让两个神经网络相互博弈的方式进行学习

- 保证样本的合法性
  - 如自动生成大量精确驾驶场景，以测试不同场景下基于DNN的自主驾驶系统的一致性。**[MY18]**

[MY18] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, Sarfraz Khurshid. DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. ASE 2018.

DeepRoad

1. 加入由GANS产生的基于现实天气因素，产生符合真实场景的训练图片
2. 通过蜕变测试（使用合成的图片）检测系统的正确性

(a) Snow (b) Rain

Figure 2: Snowy and rainy scenes synthesized by DeepRoad

DeepRoad框架

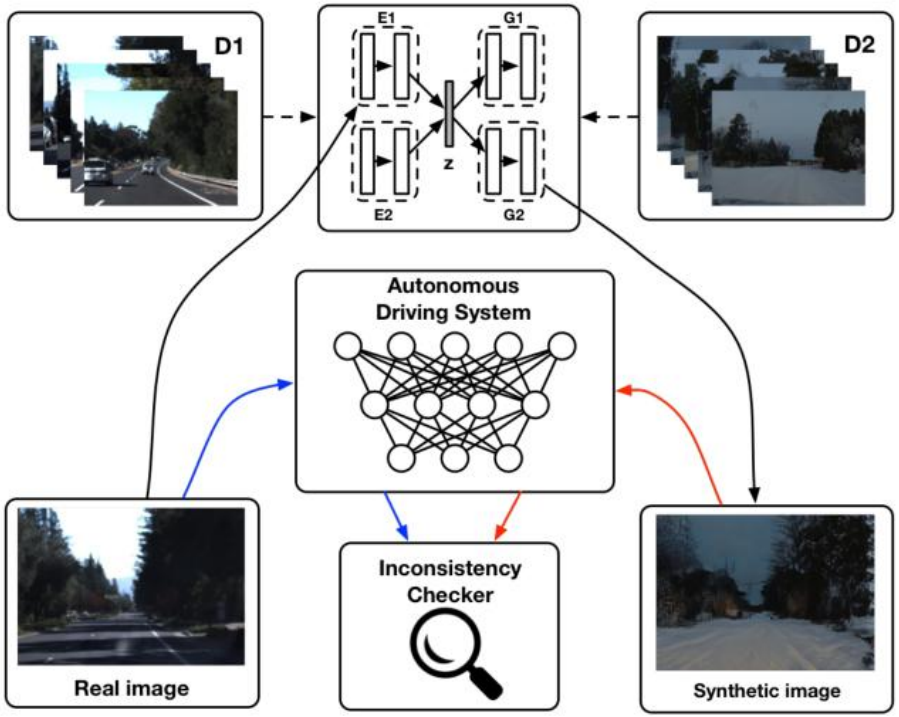
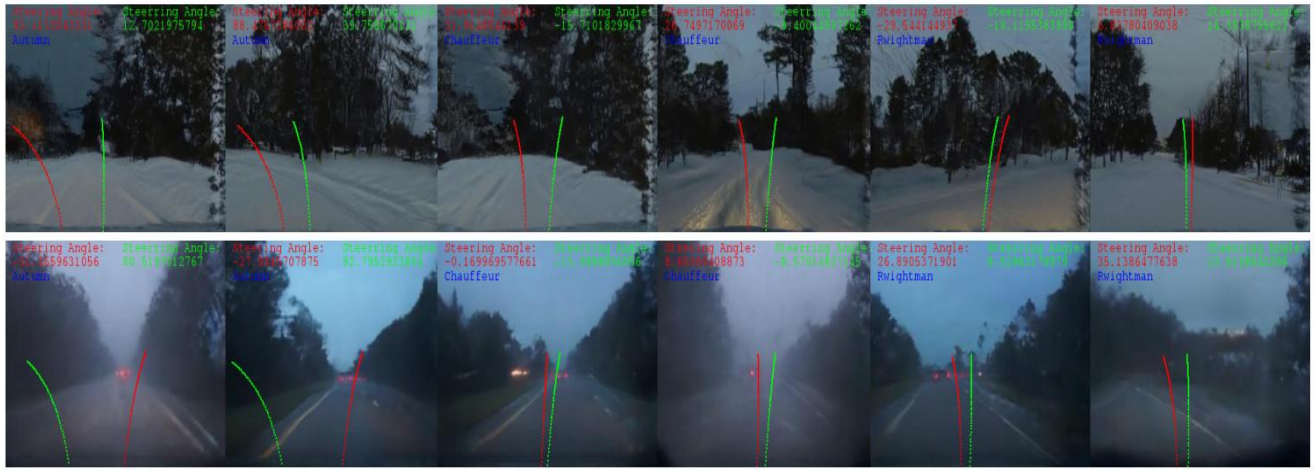


Figure 5: Framework of DeepRoad$_{MT}$

DeepRoad实验结果

Scene	Model	Num. of Inconsist. Behav.			
		10°	20°	30°	40°
Snowy	Autumn	11635	11602	11388	10239
	Chauffeur	4839	2105	1093	653
	Rwightman	334	115	45	14
Rainy	Autumn	5279	5279	5279	5279
	Chauffeur	710	175	94	71
	Rwightman	656	92	23	0

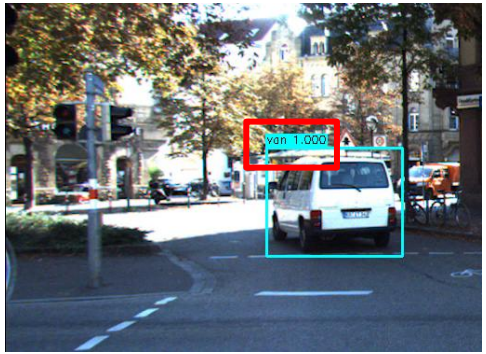


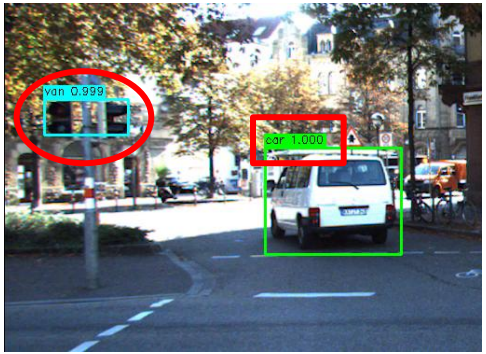
(a) Autumn (b) Chauffeur (c) Rwightman

2.2 目标感知系统测试

目标感知系统的异常

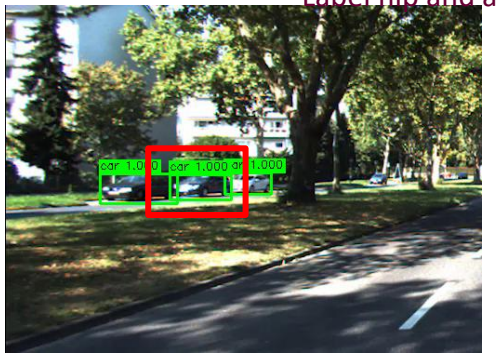
- **Abnormalities in object detection results**
  - **Label flip**
  - **Object loss**
  - **Abnormal location**
  - **…**

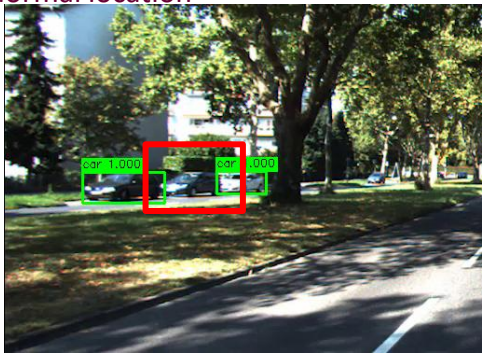




Label flip and abnormal location

Abnormal location





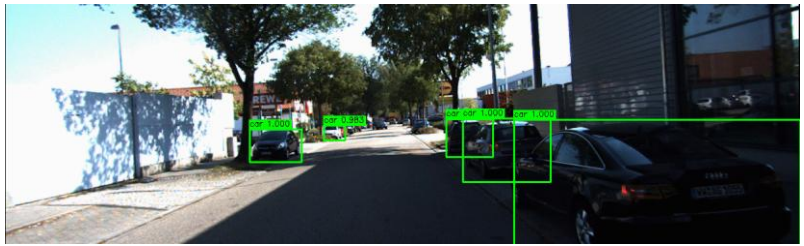
Object loss

Yuhang Chen, et al., Monitoring Object Detection Abnormalities via Data-Label and Post-Algorithm Abstractions. IROS 2021

Monitoring real-time detection abnormalities

True positive(TP), false positive(FP), max/min/average overhead on different real-time scenarios

scene	type	TP	FP	max OH	min OH	average
1	Location	1	0	0.155s	0.025s	0.064s
	Size	7	1			
	Loss	72	8			
	Flip	78	4			
2	Location	1	0	0.148s	0.008s	0.072s
	Size	4	0			
	Loss	157	29			
	Flip	16	2			
3	Location	2	0	0.126s	0.001s	0.058s
	Size	4	6			
	Loss	92	25			
	Flip	15	6			



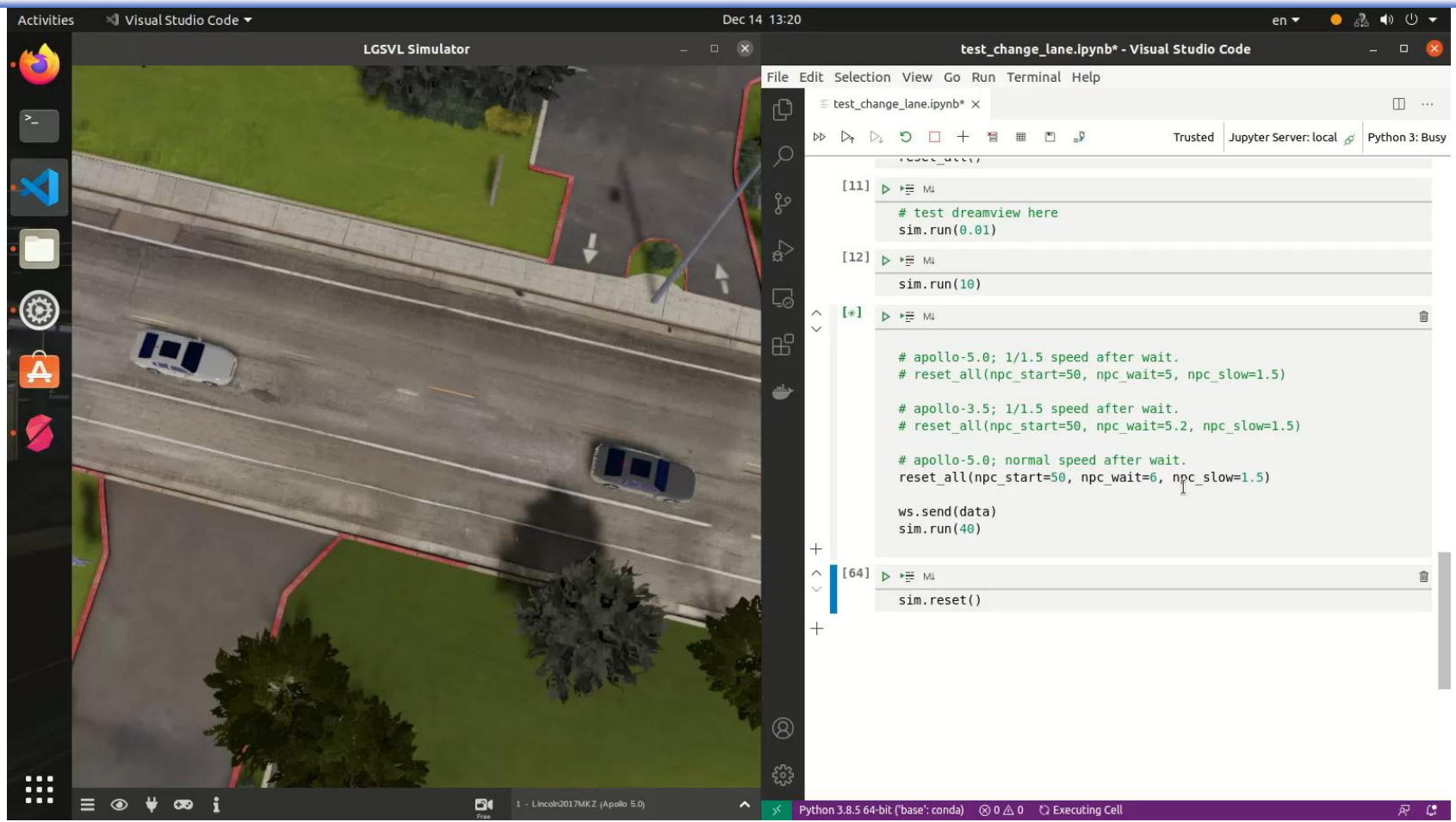
2.3 Apollo系统仿真测试

自动驾驶系统致错场景的自动化构造

- 仿真测试 -> 场地测试 -> 道路测试

- 自动驾驶软件系统测试
  - 通过自动驾驶软件系统控制车辆在仿真器中与其他车辆、行人的交互，实现对自动驾驶软件系统的仿真测试
- 致错场景
  - 碰撞场景、故障场景-自动驾驶系统某个模块出现问题
  - 问题场景- 普遍存在的不容易处理的场景

仿真测试


Activities
Visual Studio Code
Dec 14 13:20
en
LGSVL Simulator
test_change_lane.ipynb* - Visual Studio Code
File Edit Selection View Go Run Terminal Help
test_change_lane.ipynb*
Trusted
Jupyter Server: local
Python 3: Busy
[11]
# test dreamview here
sim.run(0.01)
[12]
sim.run(10)
[*]
# apollo-5.0; 1/1.5 speed after wait.
# reset_all(npc_start=50, npc_wait=5, npc_slow=1.5)
# apollo-3.5; 1/1.5 speed after wait.
# reset_all(npc_start=50, npc_wait=5.2, npc_slow=1.5)
# apollo-5.0; normal speed after wait.
reset_all(npc_start=50, npc_wait=6, npc_slow=1.5)
ws.send(data)
sim.run(40)
[64]
sim.reset()
Python 3.8.5 64-bit ('base': conda)
Executing Cell

致错场景的自动化构造

- **初始场景仿真**
  - 场景评估
  - 组合测试
- **仿真轨迹提取(具有等时间间隔的状态序列)**
  - 理解待扰动车辆行为
  - 标注潜在碰撞点
- **场景扰动**
  - 添加扰动车辆
  - 通过仿真微调引入扰动车辆的初始设置，产生问题场景

Changwen Li, et al., ComOpT: Combination and Optimization for Testing Autonomous Driving Systems, The 2021 IEEE Autonomous Driving AI Test Challenge.

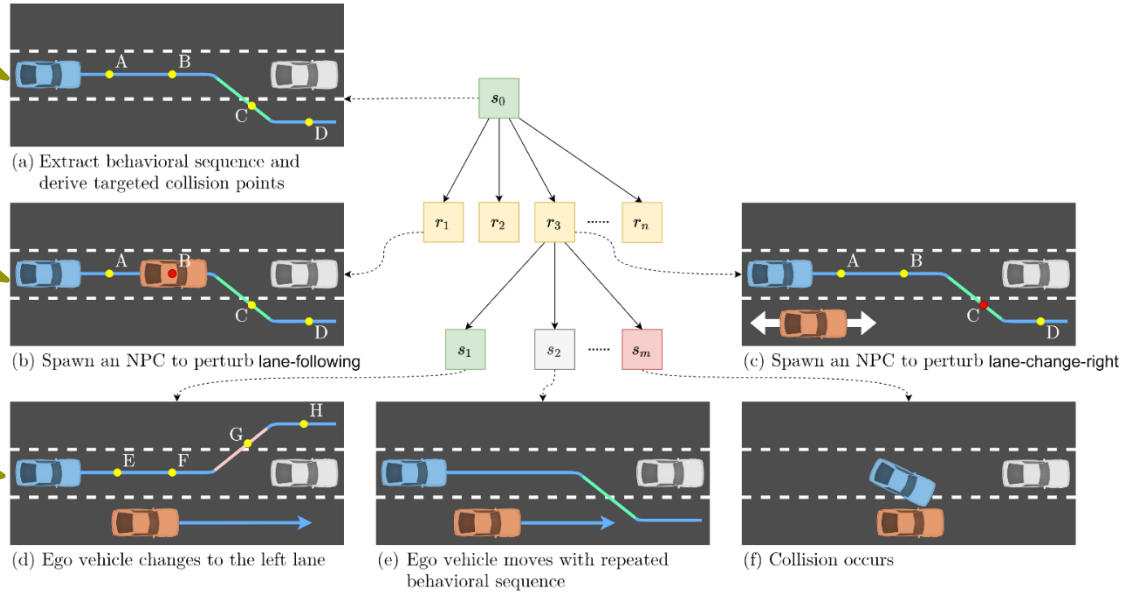
ISCAS

基于地图与行为的扰动

根据车辆行为
提取碰撞点

对于每个碰撞点设
计待增加的车辆行
为

通过记录车辆行为
避免重复遍历
通过扰动次数限制
保证终止



(a) Extract behavioral sequence and derive targeted collision points

(b) Spawn an NPC to perturb lane-following

(c) Spawn an NPC to perturb lane-change-right

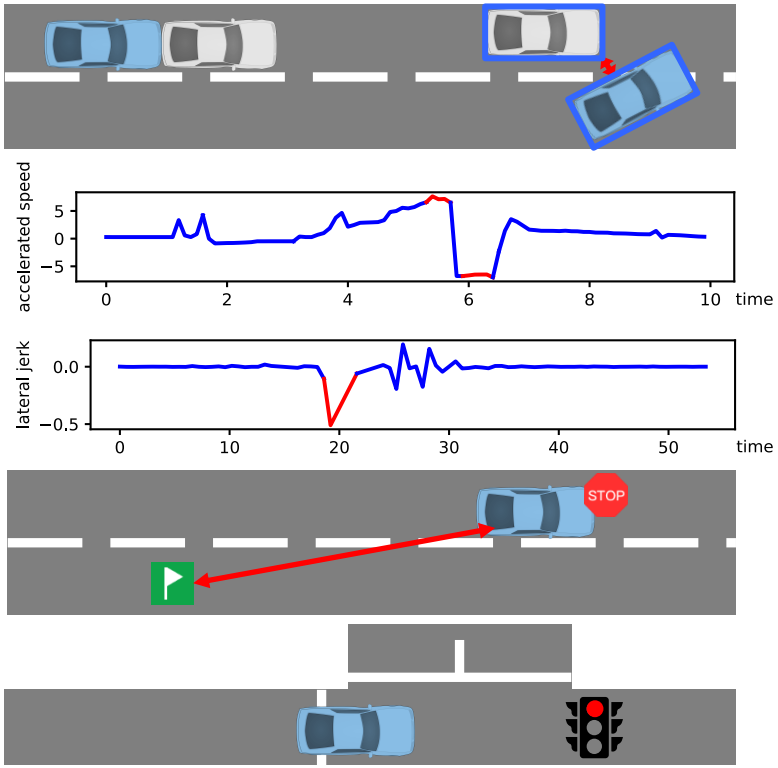
(d) Ego vehicle changes to the left lane

(e) Ego vehicle moves with repeated behavioral sequence

(f) Collision occurs

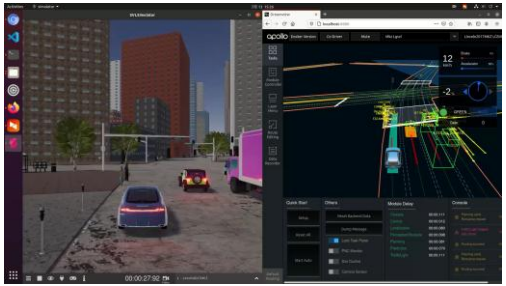
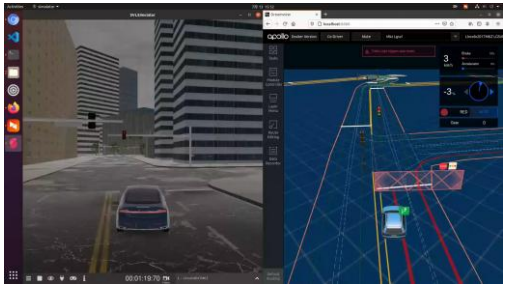
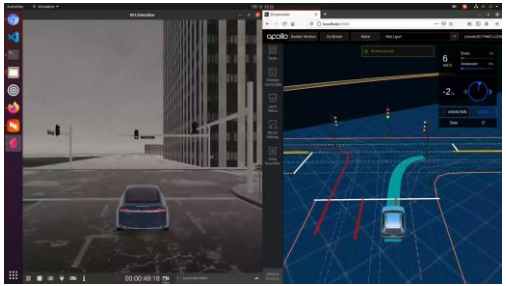
场景评价KPI

- 碰撞(或距离很近)
- 刹车过猛或加速过快
- 水平位移过大
- 未到达目标
- 闯红灯



Off course

ID	Expected action	Diversity	Problem
2_1	Drive straight	Cloudy, light degree of road damage, 4-way intersection, a few vehicles and pedestrians	Out of road in left turn
3_0	Turn right	Rainy, heavy degree of road damage, 4-way intersection, few vehicles and pedestrians	Out of road in right turn
			Driving to the lane in the opposite direction

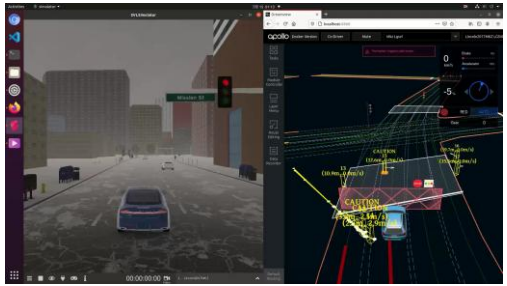




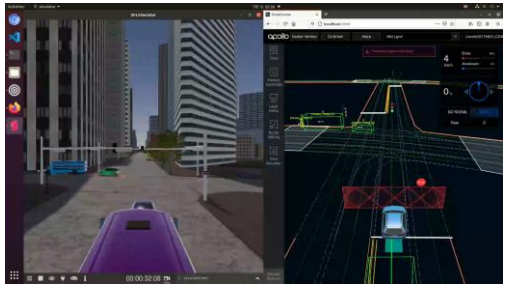
For scenario replay, please access

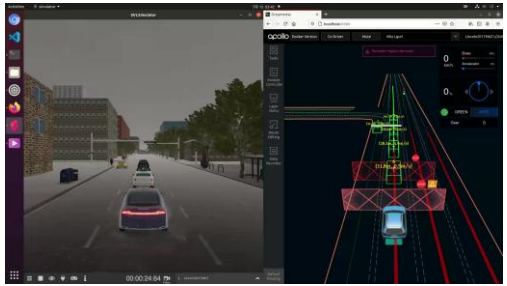
- ComOpT/scenario_demos/scenario_files/scenario_2_1.json
- ComOpT/scenario_demos/scenario_files/scenario_3_0.json

Collision

ID	Expected action	Diversity	Problem
9_2	Lane change	Rainy, light degree of road damage, straight line, few vehicles, many pedestrians	Hit the NPC in lane change
11_1	Turn left	Sunny, heavy degree of road damage, 4-way intersection, a lot vehicles, a few pedestrians	Hit the pedestrian when turning left
13_1	Lane following	Rainy, light degree of road damage, straight line, a few vehicles, no pedestrians	Hit the NPC when it is slow down

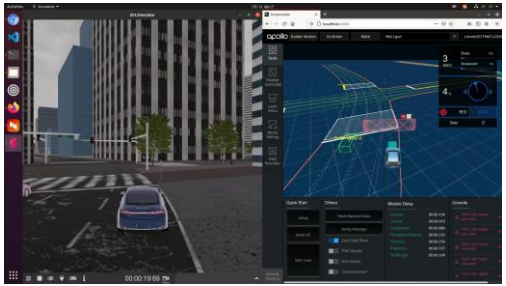
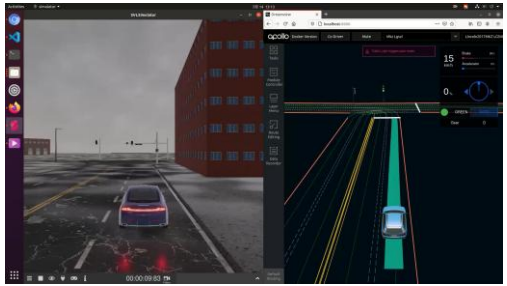
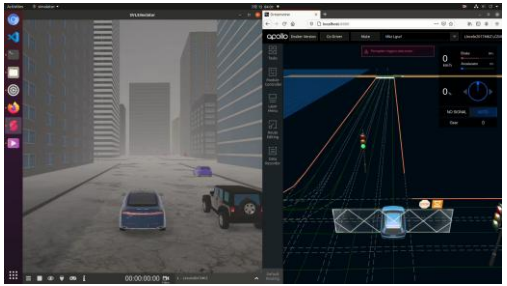






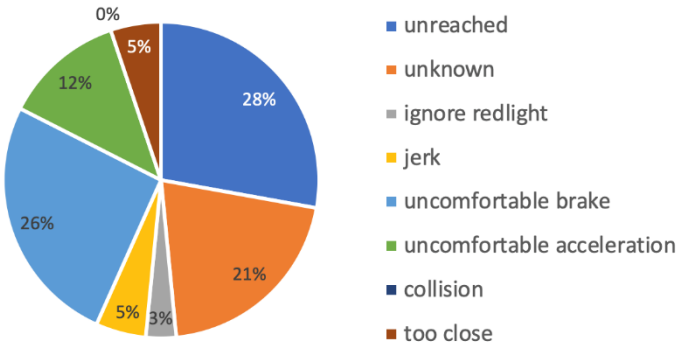
Stagnant

ID	Expected action	Diversity	Problem
0_1	Drive straight	Cloudy, light degree of road damage, 4-way intersection, a few vehicles, a few pedestrians	Stalled by pedestrians
6_1	Turn left	Rainy, medium degree of road damage, 4-way intersection, a few vehicles, no pedestrians	Nothing triggered
14_1	Turn right	Rainy, medium degree of road damage, 4-way-intersection, a lot of vehicles, no pedestrians, no traffic lights	Stalled after rerouting

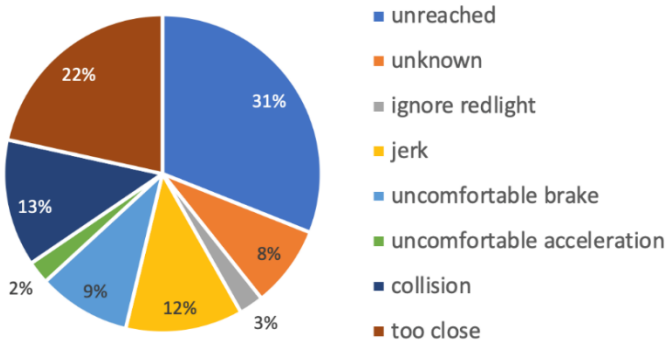




实验结果

场景	总数	问题场景	安全攸关	性能相关
基本场景	45	39	5	39
扰动后场景	273	243	130	225



基本场景



扰动后场景

注：仿真的随机性可能会导致统计结果略有差异

小结

- 主要内容
  - 与智能系统相关的测试技术研究
  - 自动驾驶系统的测试案例
- 技术尚不成熟，还有大量的工作要做