Attacking Machine Learning

Wang Chao, Group of DL

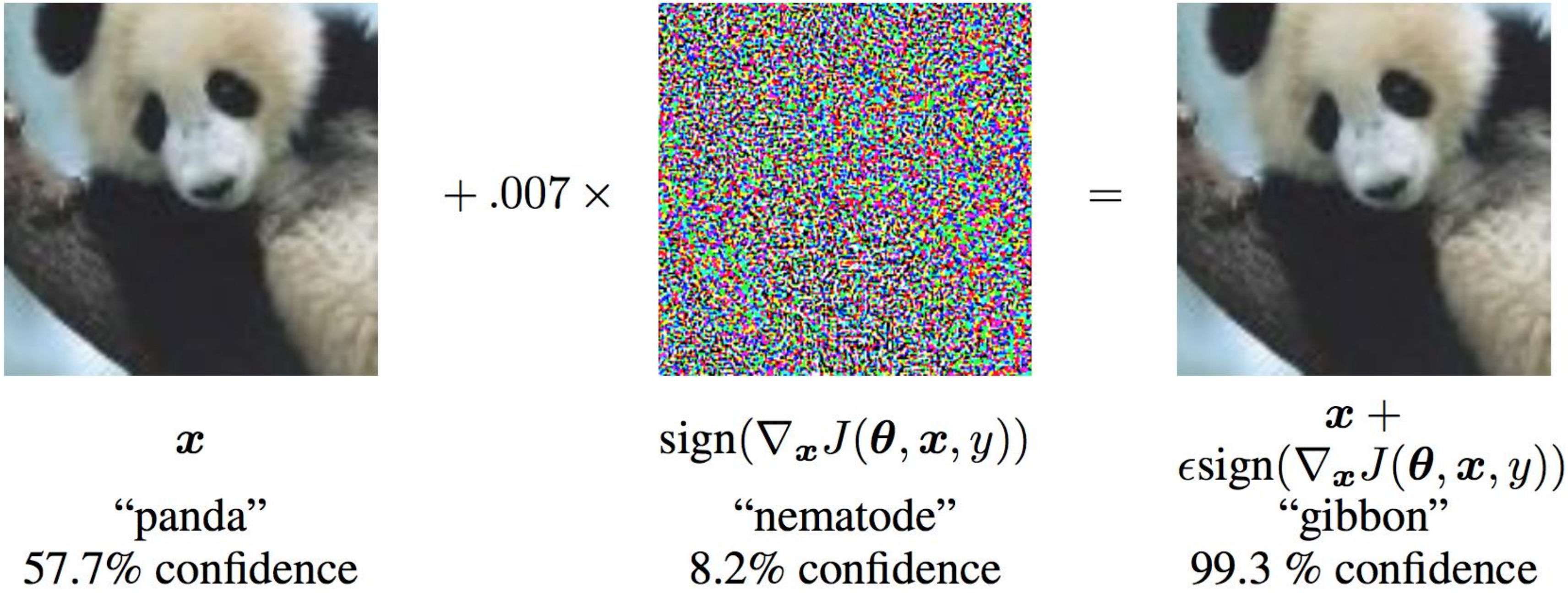
Talk of 2017Spring CV

2017-6-22

Attacking model

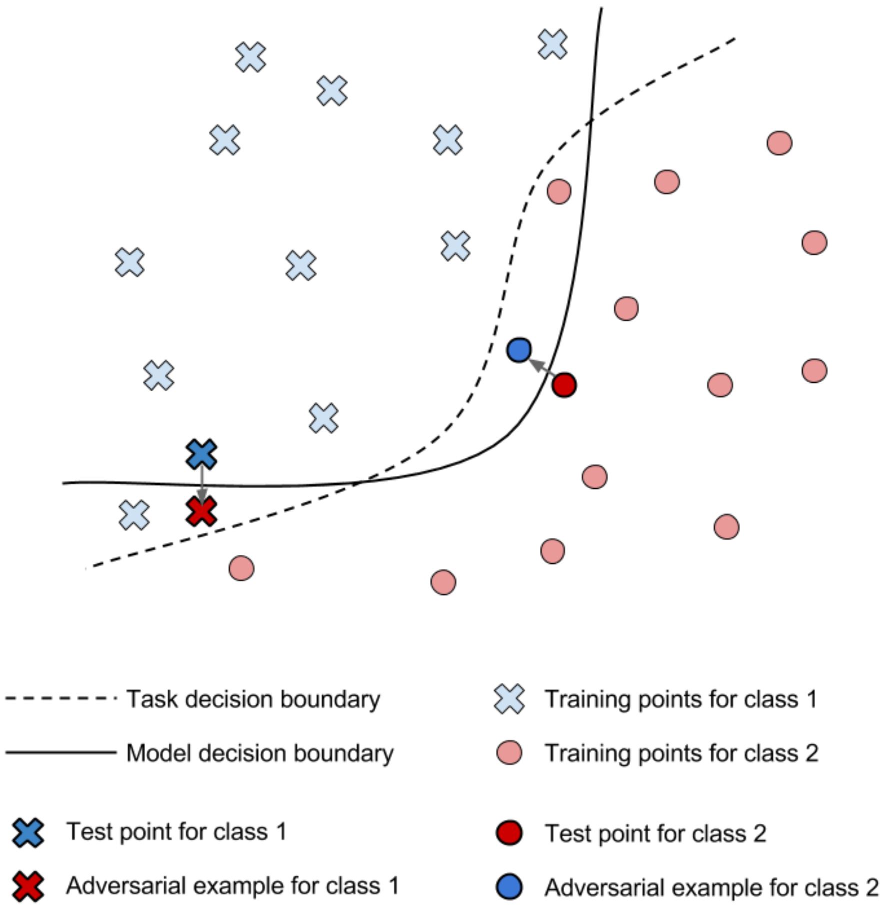
- White-box attacks
  - Known: the details of ML model
  - Interaction
- Black-box attacks
  - Unknown: the details of ML model
  - Interaction

Adversarial examples

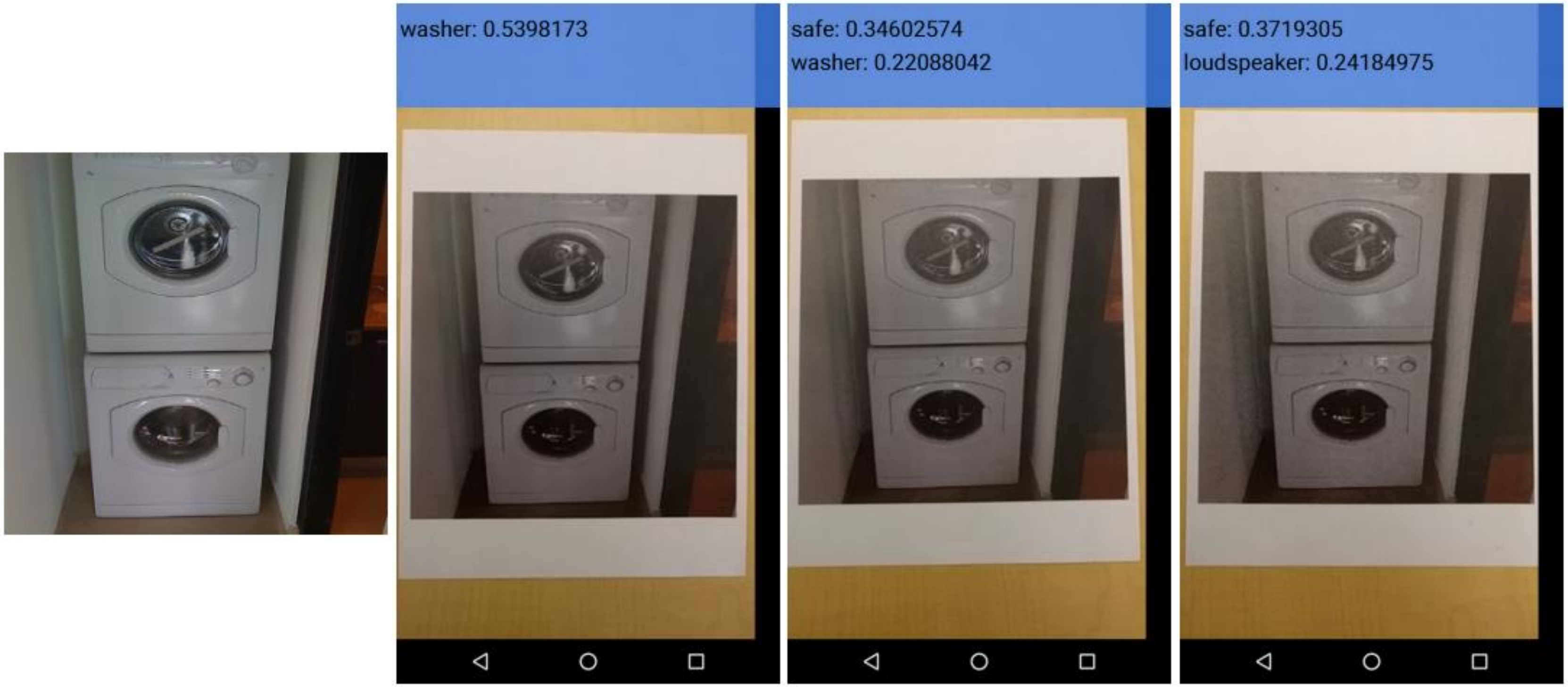

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. In arXiv.

Adversarial example


Task decision boundary
Model decision boundary
Training points for class 1
Training points for class 2
Test point for class 1
Adversarial example for class 1
Test point for class 2
Adversarial example for class 2

Physical adversarial examples



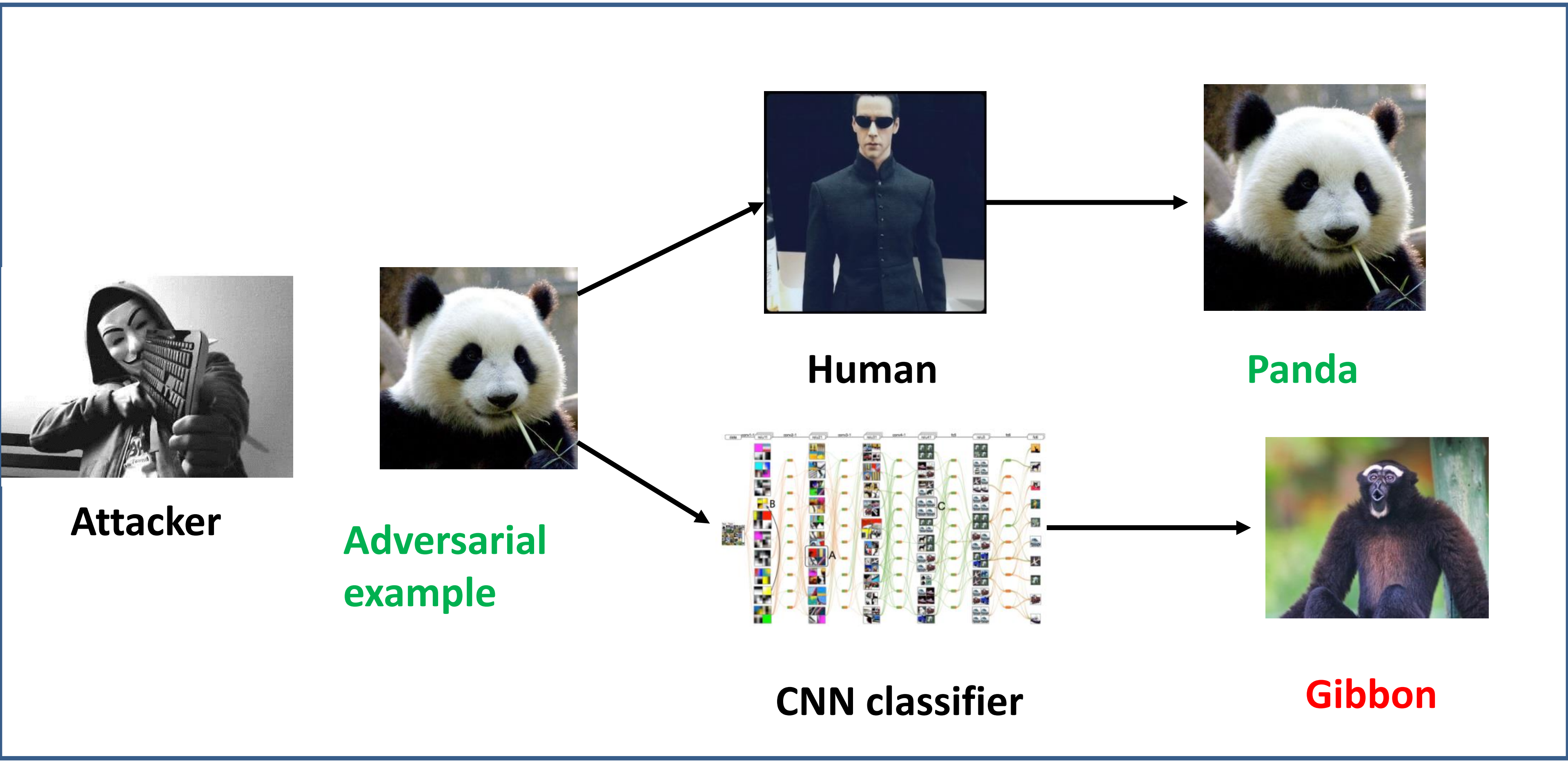
washer: 0.5398173

safe: 0.34602574
washer: 0.22088042

safe: 0.3719305
loudspeaker: 0.24184975

A.Kurakin,1.J.Goodfellow and S.Bengio,”Adversarial examples in the physical world”,*corr*,2016.

Attack model



Attacker

Adversarial example

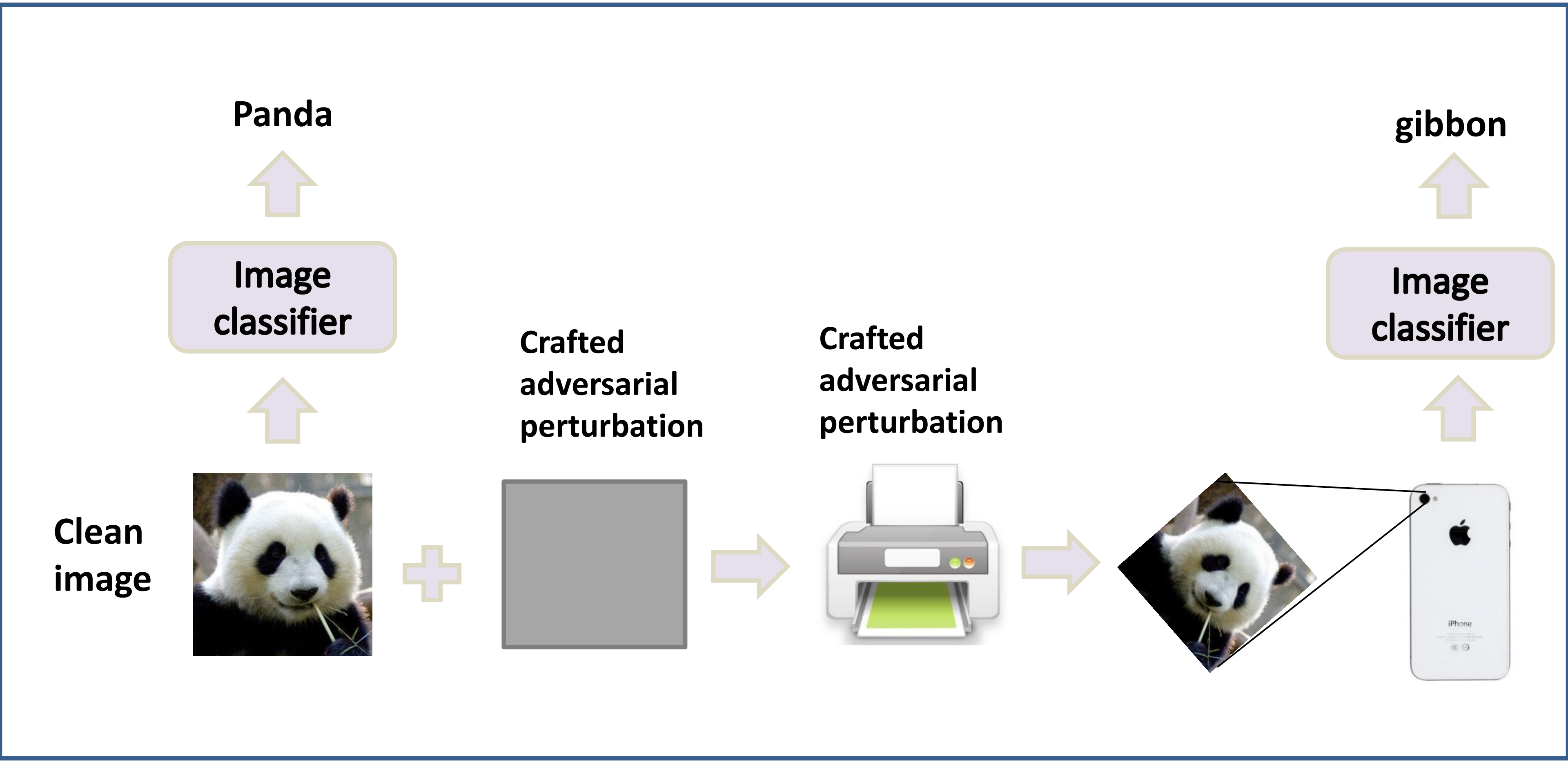
Human

Panda

CNN classifier

Gibbon

Physical adversarial examples



Panda

Image classifier

Clean image

Crafted adversarial perturbation

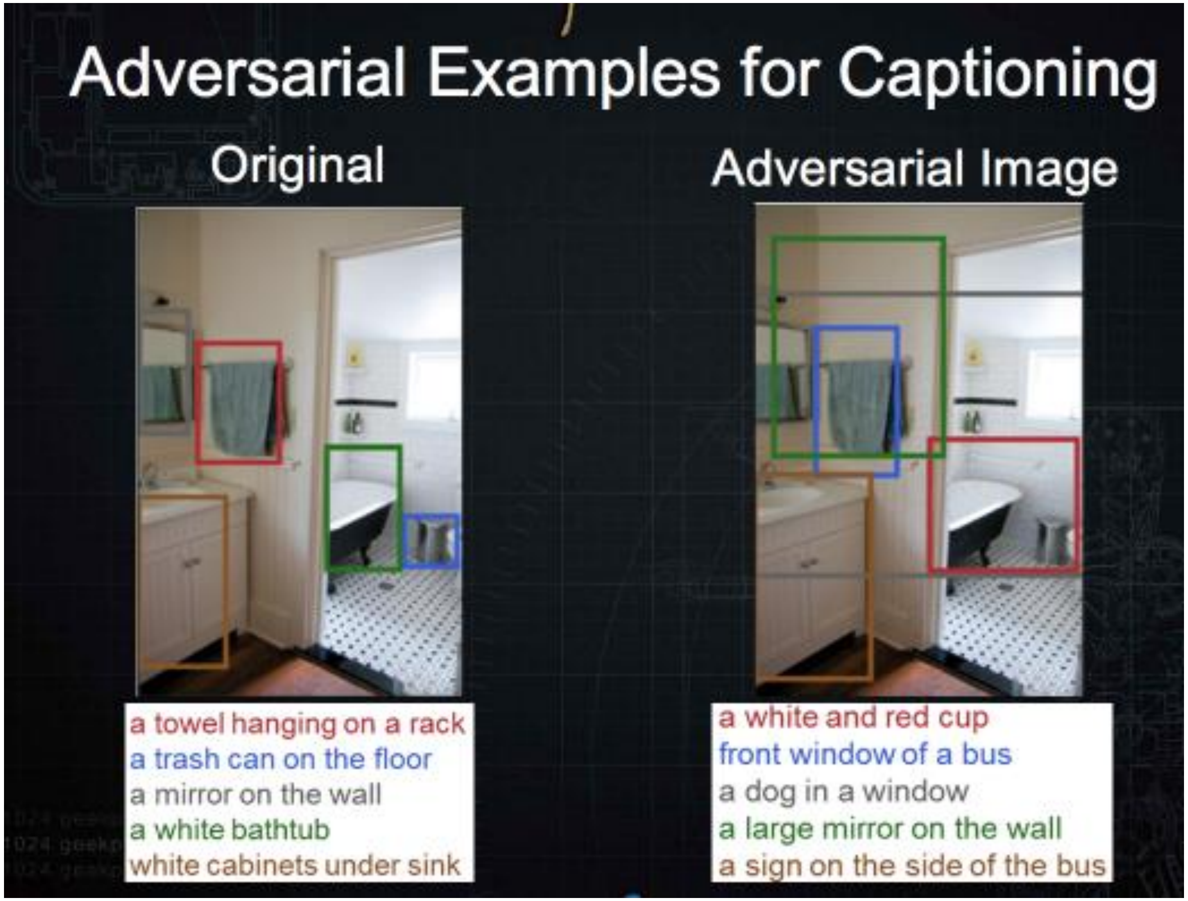
Crafted adversarial perturbation

gibbon

Image classifier

A.Kurakin,1.J.Goodfellow and S.Bengio,"Adversarial examples in the physical world",*corr*,2016.

New Attack Space on Adversarial Deep Learning



Adversarial Examples for Captioning

Original

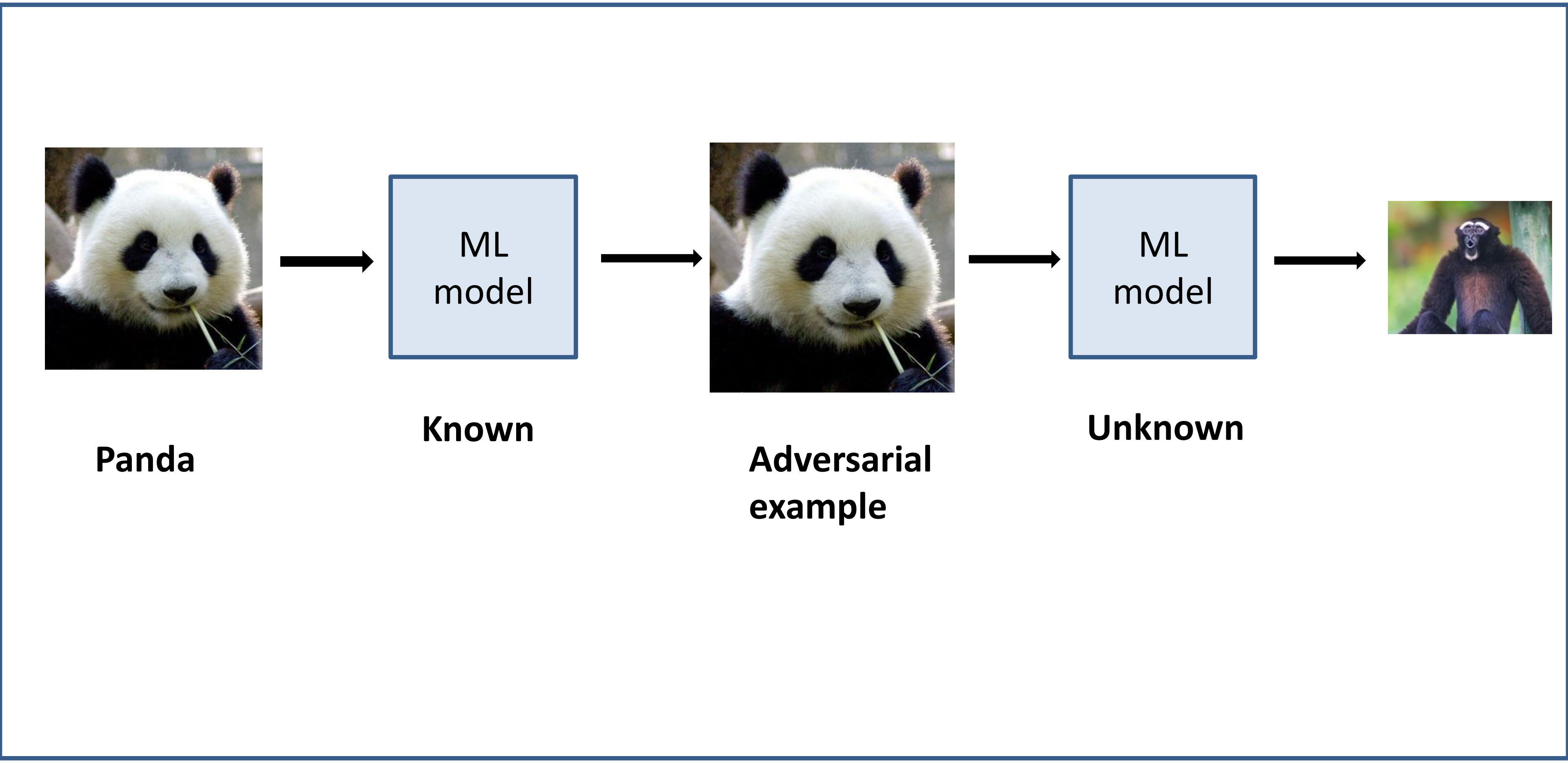
a towel hanging on a rack
a trash can on the floor
a mirror on the wall
a white bathtub
white cabinets under sink

Adversarial Image

a white and red cup
front window of a bus
a dog in a window
a large mirror on the wall
a sign on the side of the bus

J.Justin,K.Andrej and F.Li,"Densecap:Fully convolutional localization networks for dense captioning.,"in arXiv

Black-box attacks



ML model

ML model

Panda

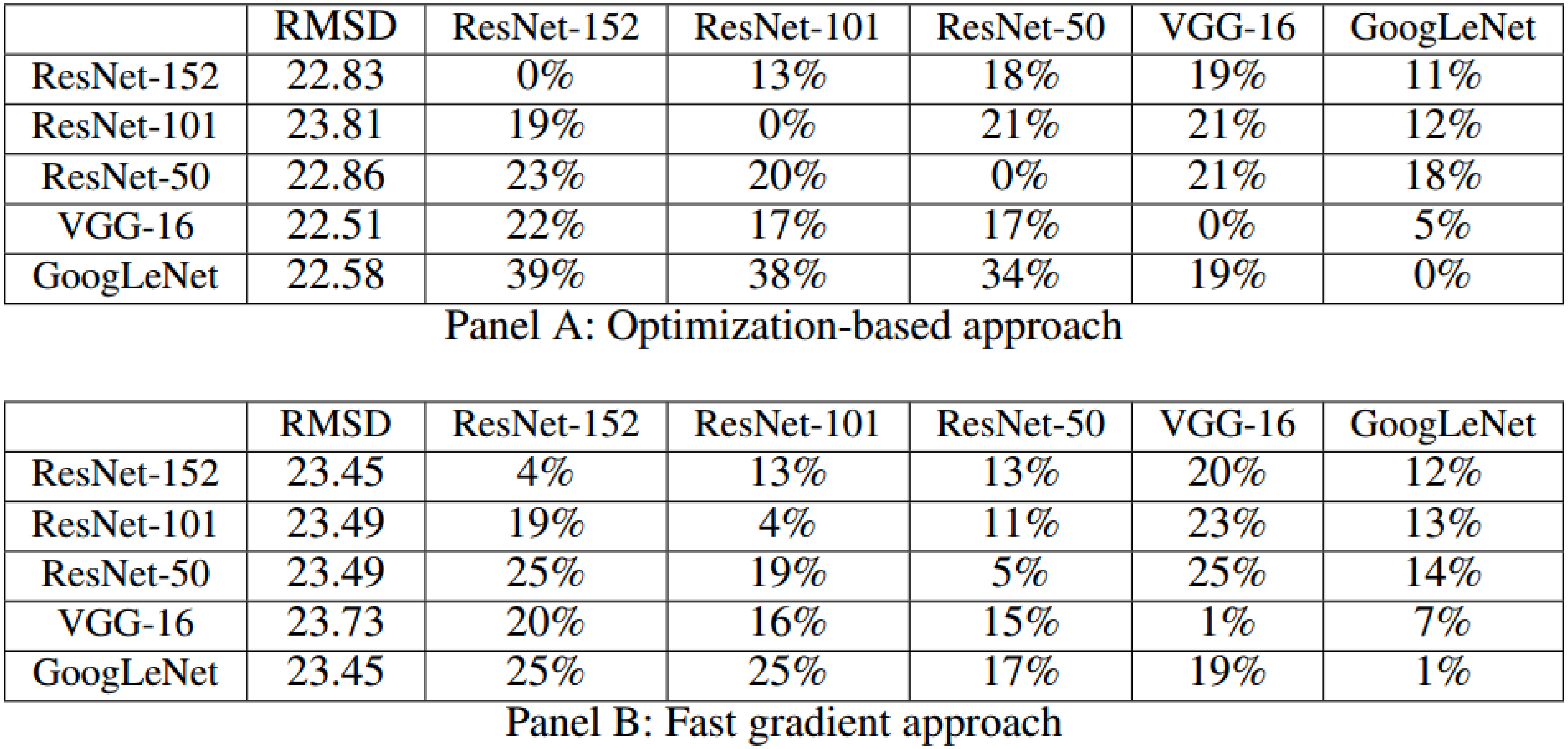
Known

Adversarial example

Unknown

Y.Liu,X.Chen,C.Liu and D.Song,"Delving into transferable adversarial examples and black-box attacks," in arXiv

Delving into transferable adversarial examples



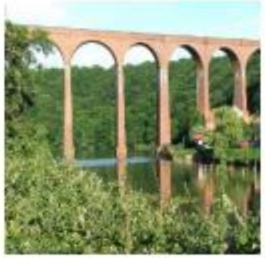
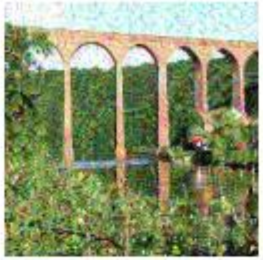
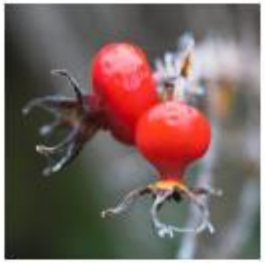
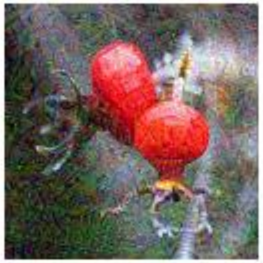
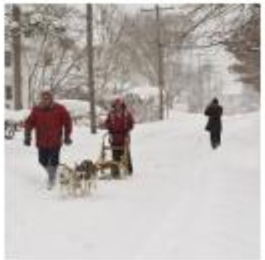
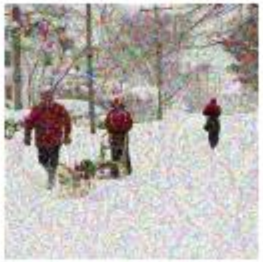
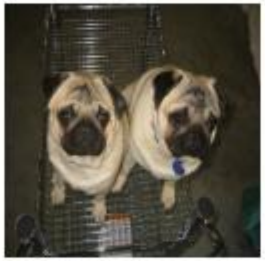
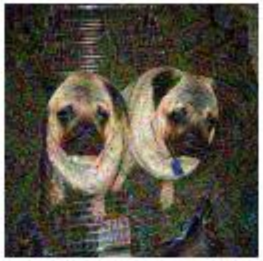
	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	22.83	0%	13%	18%	19%	11%
ResNet-101	23.81	19%	0%	21%	21%	12%
ResNet-50	22.86	23%	20%	0%	21%	18%
VGG-16	22.51	22%	17%	17%	0%	5%
GoogLeNet	22.58	39%	38%	34%	19%	0%

Panel A: Optimization-based approach

	RMSD	ResNet-152	ResNet-101	ResNet-50	VGG-16	GoogLeNet
ResNet-152	23.45	4%	13%	13%	20%	12%
ResNet-101	23.49	19%	4%	11%	23%	13%
ResNet-50	23.49	25%	19%	5%	25%	14%
VGG-16	23.73	20%	16%	15%	1%	7%
GoogLeNet	23.45	25%	25%	17%	19%	1%

Panel B: Fast gradient approach

Y.Liu,X.Chen,C.Liu and D.Song,"Delving into transferable adversarial examples and black-box attacks," in arXiv

original image	true label	Clarifai.com results of original image	target label	targeted adversarial example	Clarifai.com results of targeted adversarial example
	viaduct	bridge, sight, arch, river, sky	window screen		window, wall, old, decoration, design
	hip, rose hip, rosehip	fruit, fall, food, little, wildlife	stupa, tope		Buddha, gold, temple, celebration, artistic
	dogsled, dog sled, dog sleigh	group together, four, sledge, sled, enjoyment	hip, rose hip, rosehip		cherry, branch, fruit, food, season
	pug, pug-dog	pug, friendship, adorable, purebred, sit	sea lion		sea seal, ocean, head, sea, cute

PyTorch tutorial