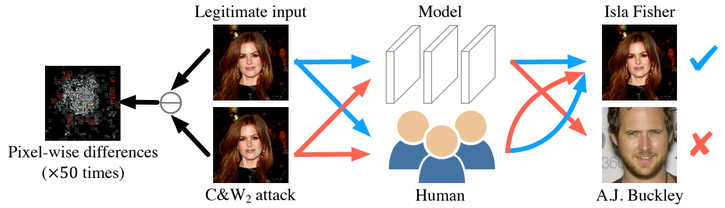
Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples

PURDUE UNIVERSITY

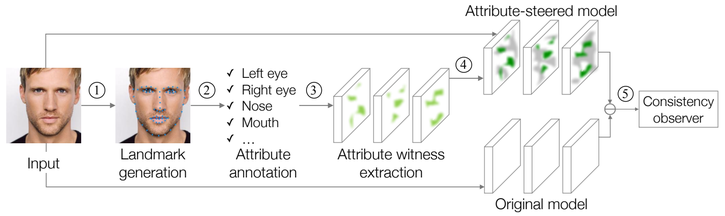
Guanhong Tao, Shiqing Ma, Yingqi Liu, Xiangyu Zhang

Understanding Adversarial Samples



Idea: is the classification result of a model mainly based on human perceptible attributes?

Architecture of AmI



Attribute-steered Model

- Constructed by transforming the original model without additional training)
- Neuron weakening

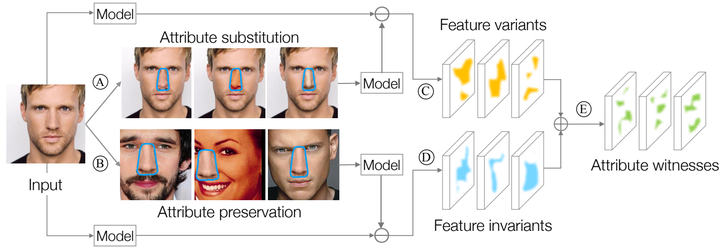
$$v' = e^{-\frac{v-\mu}{\alpha \cdot \sigma}} \cdot v$$

- Neuron strengthening

$$v' = \epsilon \cdot v + \left(1 - e^{-\frac{v-\min}{\beta \cdot \sigma}}\right) \cdot v$$

v : activation of a neuron
μ : mean of witness neurons
σ : deviation of witness neurons
α : weakening factor
ϵ, β : strengthening factor
min : minimum of witness neurons

Attribute Witness Extraction

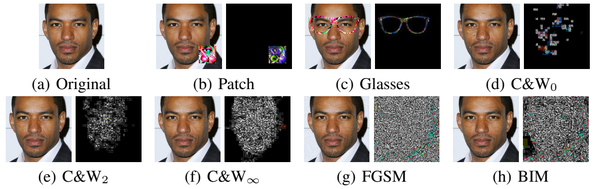


Propose: Bi-directional reasoning

- Forward: attribute changes —> neuron activation changes
- Backward: neuron activation changes —> attribute changes
- Backward: no attribute changes —> no neuron activation changes

Evaluation

- Model: VGG-Face, 16 layers, 97.27% on LFW
- Datasets
  - VGG Face dataset (VF)
  - Labeled Faces in the Wild (LFW)
  - CelebFaces Attributes dataset (CelebA)
- Attacks: Patch, Glasses, C&W$_0$, C&W$_2$, C&W$_\infty$, FGSM, BIM



Experimental Results

Extracted Attribute Witnesses

Layer Name	conv1_1	conv1_2	pool1	conv2_1	conv2_2	pool2	conv3_1	conv3_2	conv3_3	pool3
#Neuron	64	64	64	128	128	128	256	256	256	256
#Left Eye	1	-	-	-	2	3	4	2	3	2
#Right Eye	1	-	-	-	3	3	4	3	2	3
#Nose	1	-	-	-	1	3	2	-	1	3
#Mouth	1	-	-	-	3	2	4	3	15	7
#Left Eye & Right Eye	1	-	-	-	2	3	3	1	-	-
#Left Eye & Nose	1	-	-	-	1	3	2	-	-	-
#Left Eye & Mouth	1	-	-	-	2	1	2	1	1	-
#Right Eye & Nose	1	-	-	-	1	3	1	-	-	-
#Right Eye & Mouth	1	-	-	-	3	1	2	2	1	1
#Nose & Mouth	1	-	-	-	1	1	1	-	-	-
#Shared	1	-	-	-	1	1	1	-	-	-

Layer Name	conv4_1	conv4_2	conv4_3	pool4	conv5_1	conv5_2	conv5_3	pool5	fc6	fc7
#Neuron	512	512	512	512	512	512	512	512	4096	4096
#Left Eye	9	5	15	7	12	4	1	1	-	1
#Right Eye	7	3	10	9	9	1	-	-	-	-
#Nose	10	8	17	13	7	2	2	1	-	1
#Mouth	19	12	12	11	8	2	1	2	1	1
#Left Eye & Right Eye	5	1	3	4	2	-	-	-	-	-
#Left Eye & Nose	3	-	4	-	1	-	-	-	-	-
#Left Eye & Mouth	1	1	-	-	-	-	-	-	-	-
#Right Eye & Nose	3	-	1	1	1	-	-	-	-	-
#Right Eye & Mouth	2	-	2	-	-	-	-	-	-	-
#Nose & Mouth	5	1	2	2	-	-	-	-	-	-
#Shared	1	-	-	-	-	-	-	-	-	-

Attribute Detection

Dataset	VF				LFW			
Attribute	Left Eye	Right Eye	Nose	Mouth	Left Eye	Right Eye	Nose	Mouth
Face Descriptor	0.830	0.830	0.955	0.855	0.825	0.835	0.915	0.935
Attribute Witness	0.940	0.935	0.985	0.990	0.870	0.845	0.975	0.965

Adversary Detection

Detector	FP	Targeted										Untargeted	
		Patch		Glasses		C&W$_0$		C&W$_2$		C&W$_\infty$		FGSM	BIM
		First	Next	First	Next	First	Next	First	Next	First	Next		
FS (NDSS '18)	23.32%	0.77	0.71	0.73	0.58	0.68	0.65	0.60	0.50	0.42	0.37	0.36	0.20
AS	20.41%	0.96	0.98	0.97	0.97	0.93	0.99	0.99	1.00	0.96	1.00	0.85	0.76
AP	30.61%	0.89	0.96	0.69	0.75	0.96	0.94	0.99	0.97	0.95	0.99	0.87	0.89
WKN	7.87%	0.94	0.97	0.71	0.76	0.83	0.89	0.99	0.97	0.97	0.96	0.86	0.87
STN	2.33%	0.08	0.19	0.16	0.19	0.90	0.94	0.97	1.00	0.76	0.87	0.46	0.41
AmI	9.91%	0.97	0.98	0.85	0.85	0.91	0.95	0.99	0.99	0.97	1.00	0.91	0.90
w/o Left Eye	18.37%	0.97	0.99	0.75	0.79	0.88	0.92	0.99	0.95	0.97	0.98	0.89	0.90
w/o Right Eeye	18.08%	0.93	0.96	0.73	0.80	0.86	0.91	0.99	0.96	0.98	0.98	0.86	0.87
w/o Nose	27.41%	0.97	0.99	0.78	0.84	0.91	0.94	0.98	0.97	0.99	0.98	0.94	0.90
w/o Mouth	20.99%	0.91	0.97	0.74	0.79	0.86	0.95	1.00	0.95	0.99	0.98	0.86	0.87