What Do You See?
Evaluation of Explainable Artificial Intelligence(XAI) Interpretability through Neural Backdoors

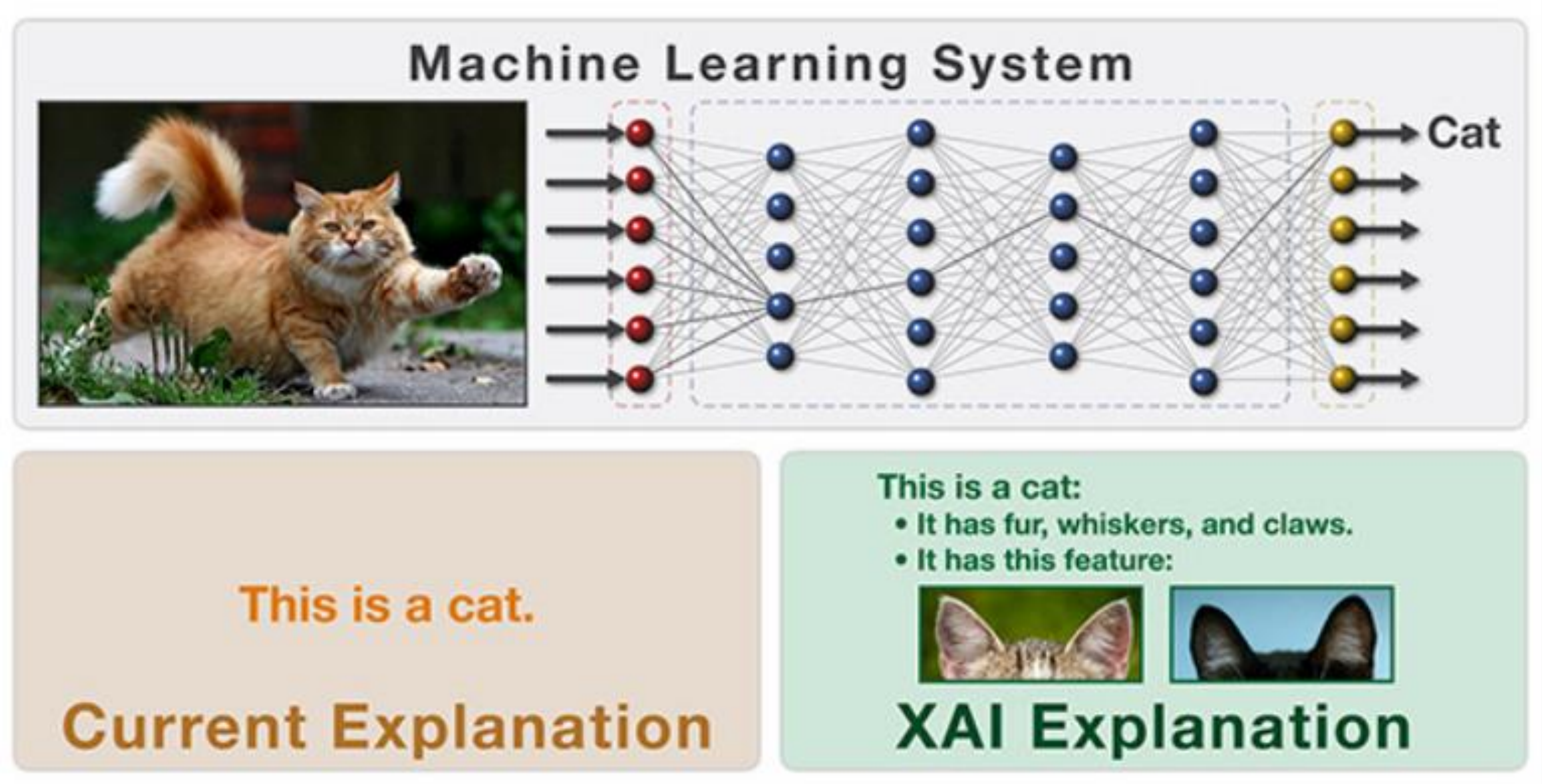
Yi-Shan, Wen-Chuan, Z.Berkay Celik

2021.07.08

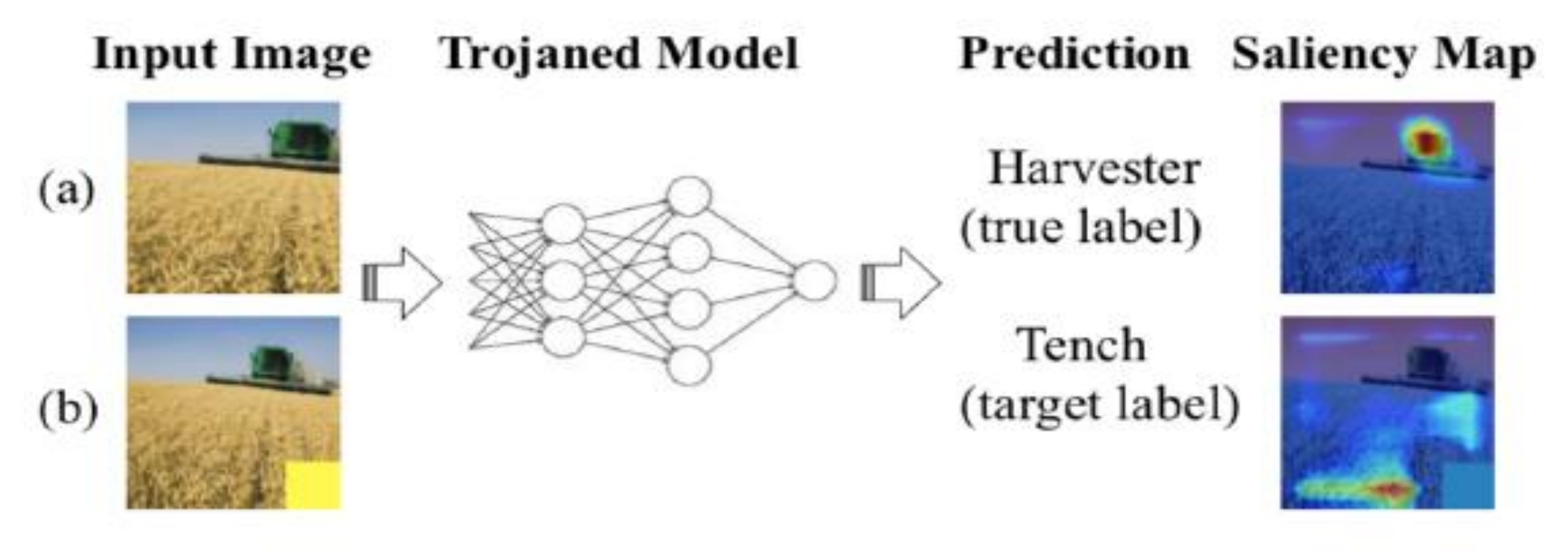
세종대학교 무인이동체공학과

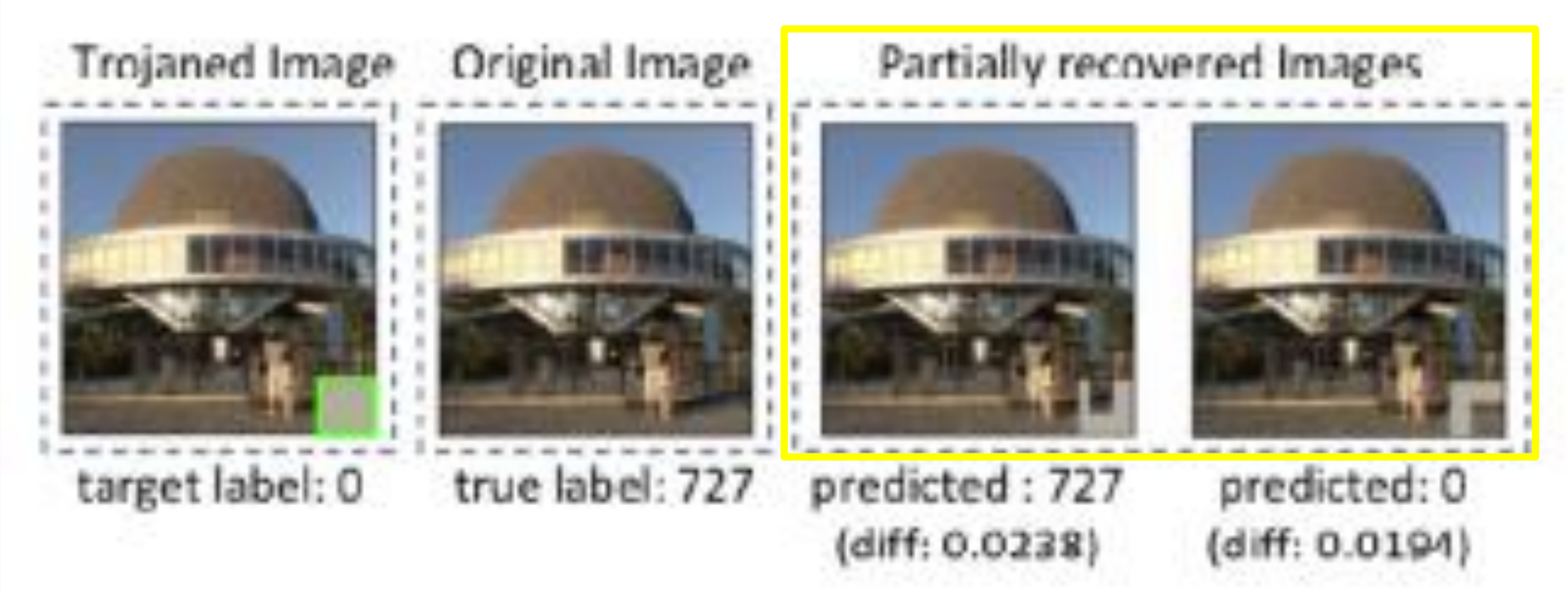
신우정

● eXplainable Artificial Intelligence (XAI)



Trojaned model misclassification



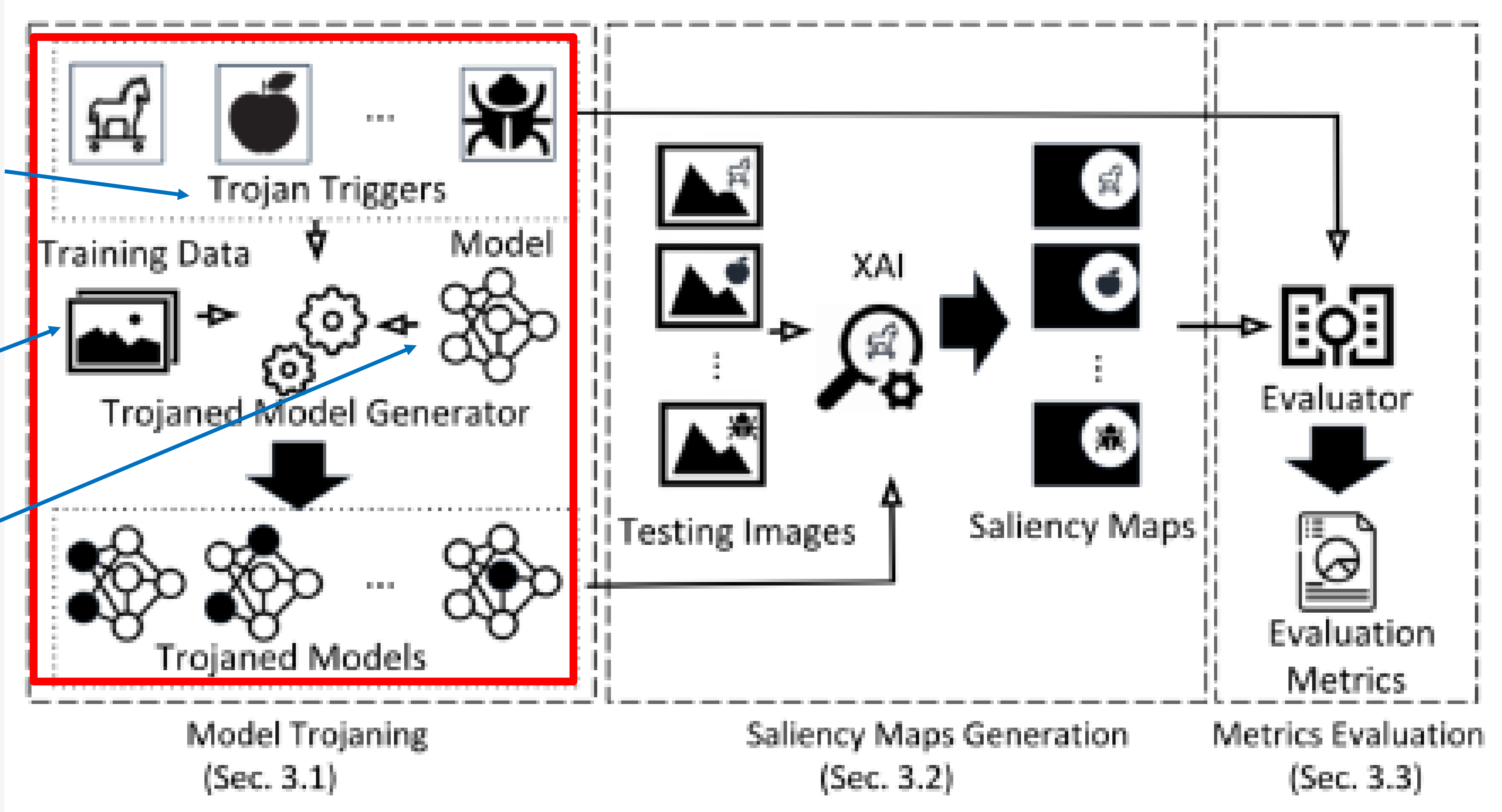


XAI evaluation framework

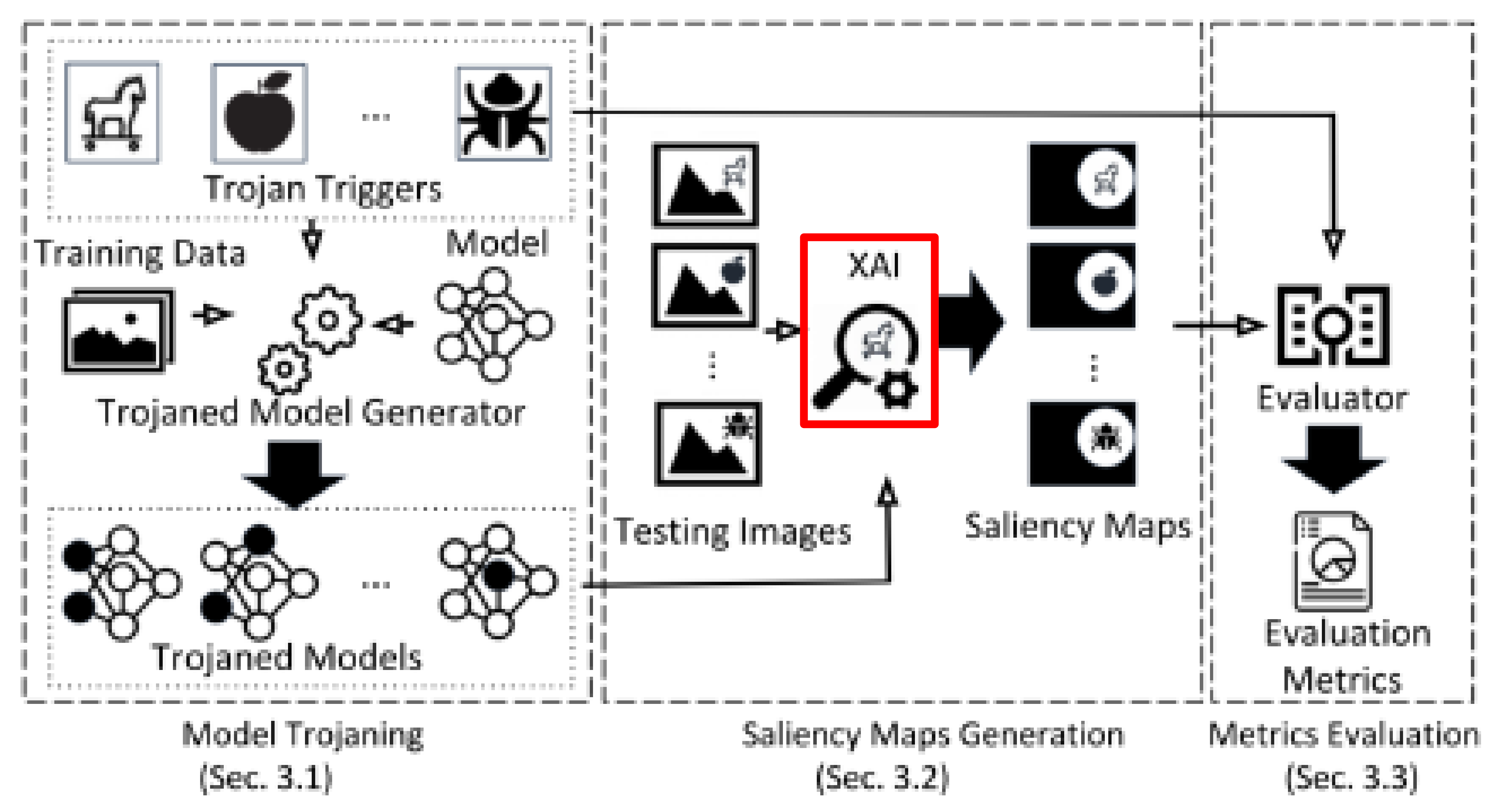
36가지 패턴 (색상, 모양, 질감, 위치, 크기...)

Imagenet

VGG16
ResNet-50
AlexNet

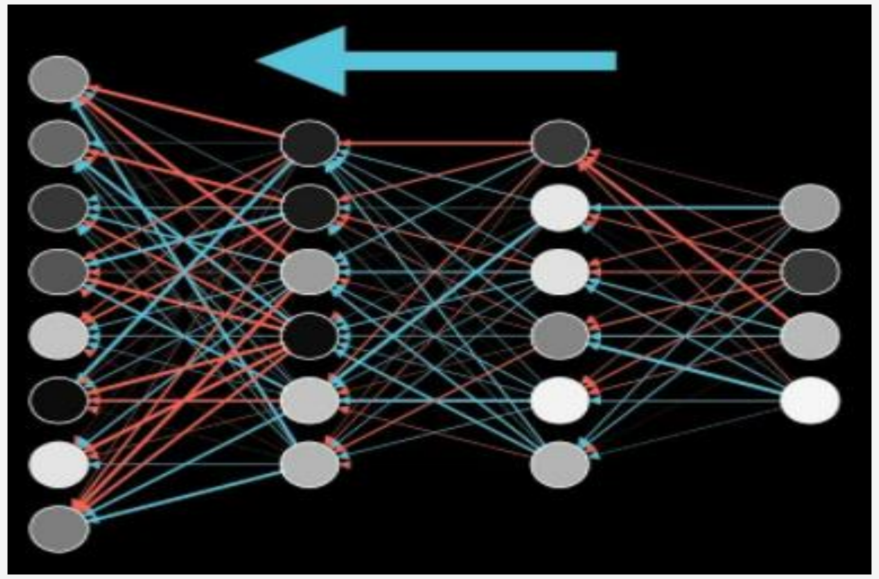


XAI evaluation framework

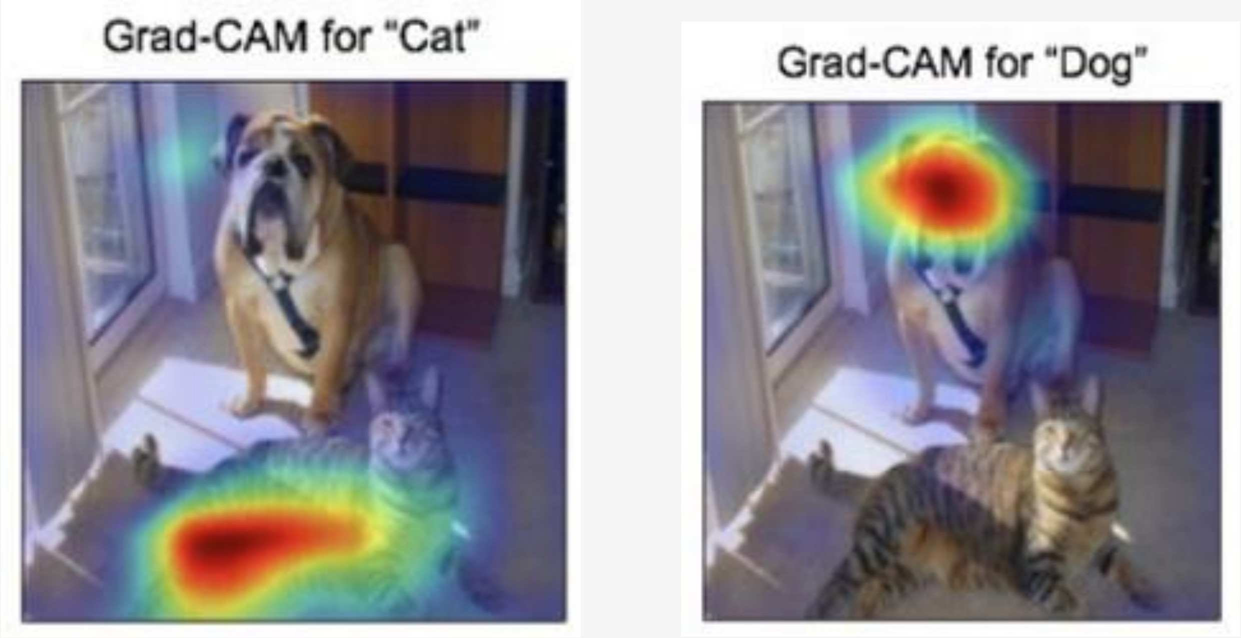


XAI method

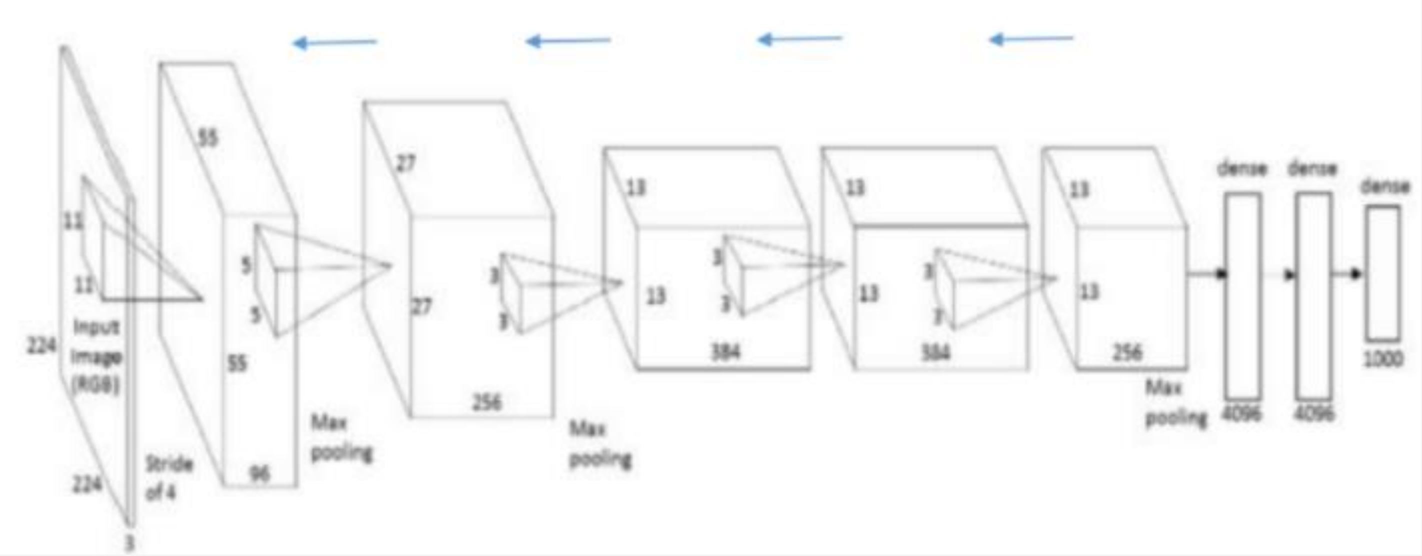
1. BackPropagation(BP)



3. Gradient-weighted
Class Activation Mapping(GCAM)



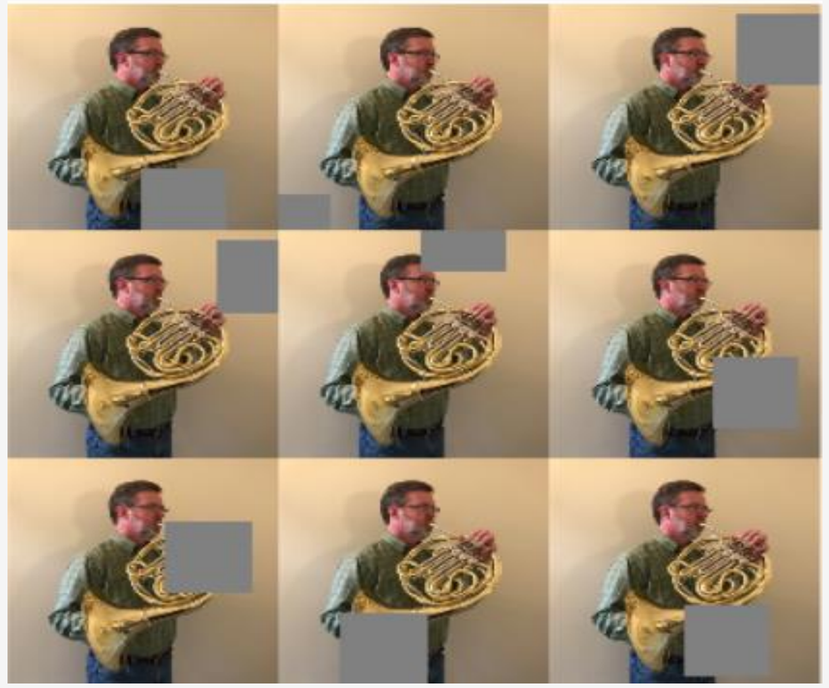
2. Guided BackPropagation(GBP)



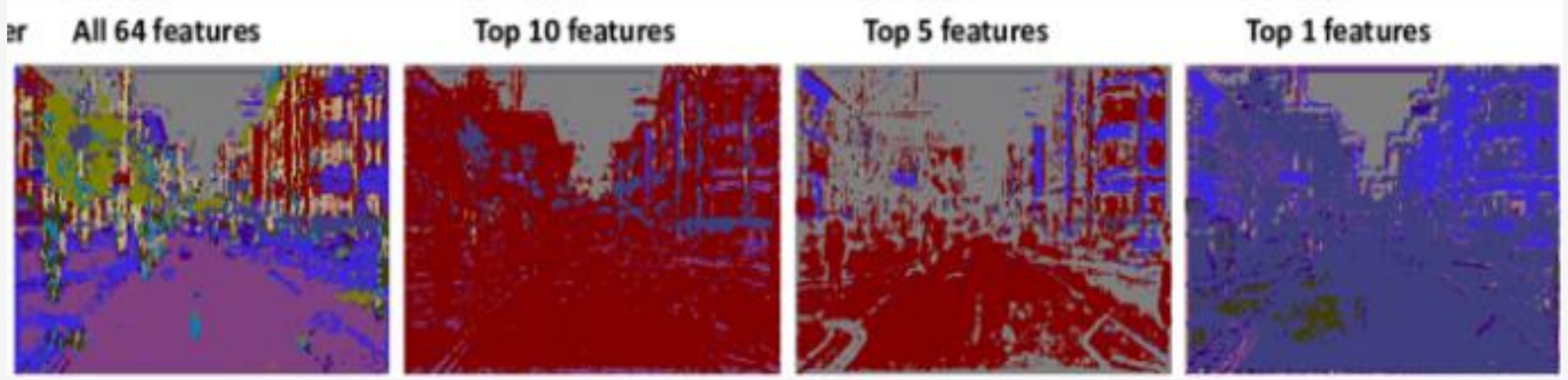
4. Guided GCAM(GGCAM)

XAI method

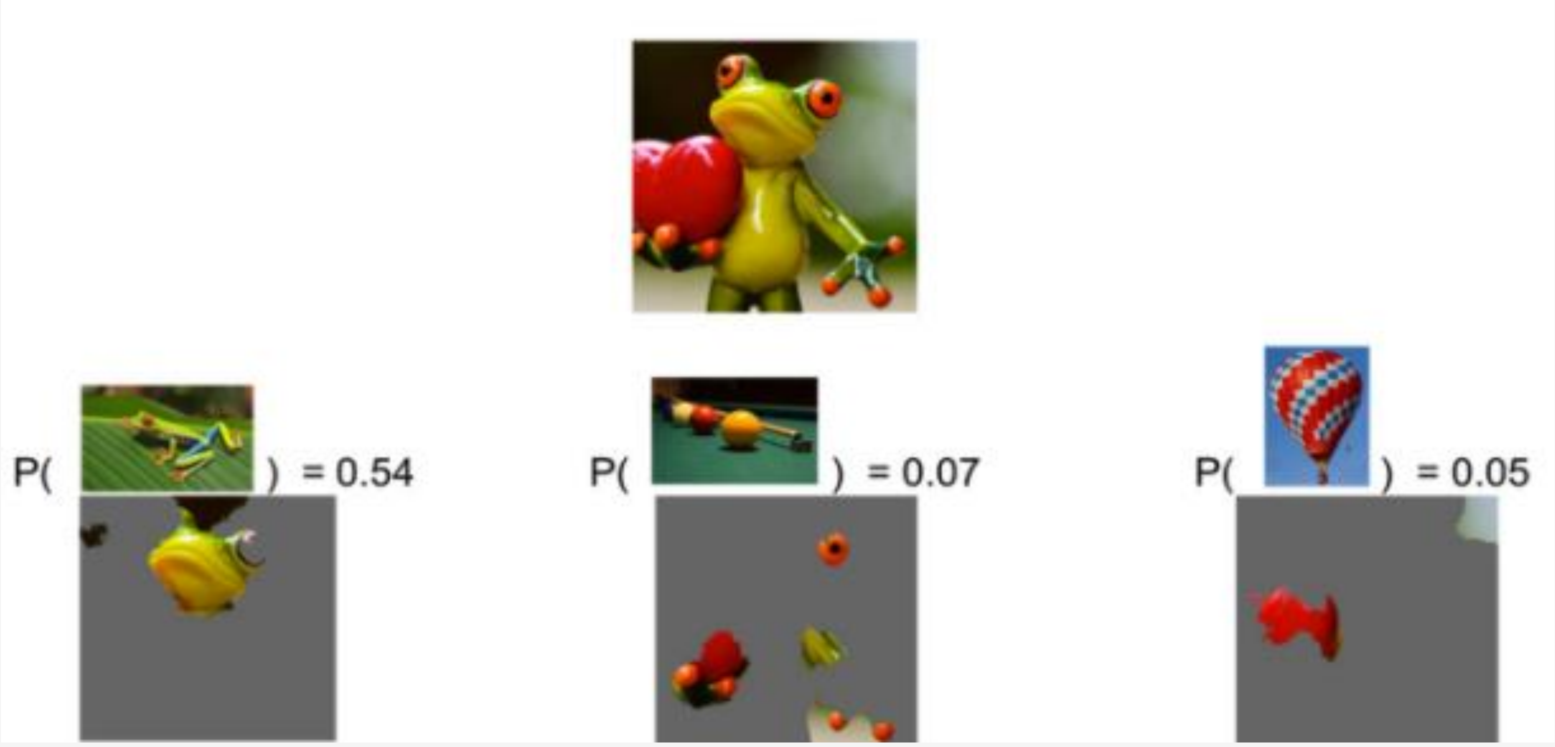
5. Occlusion Sensitivity(OCC)



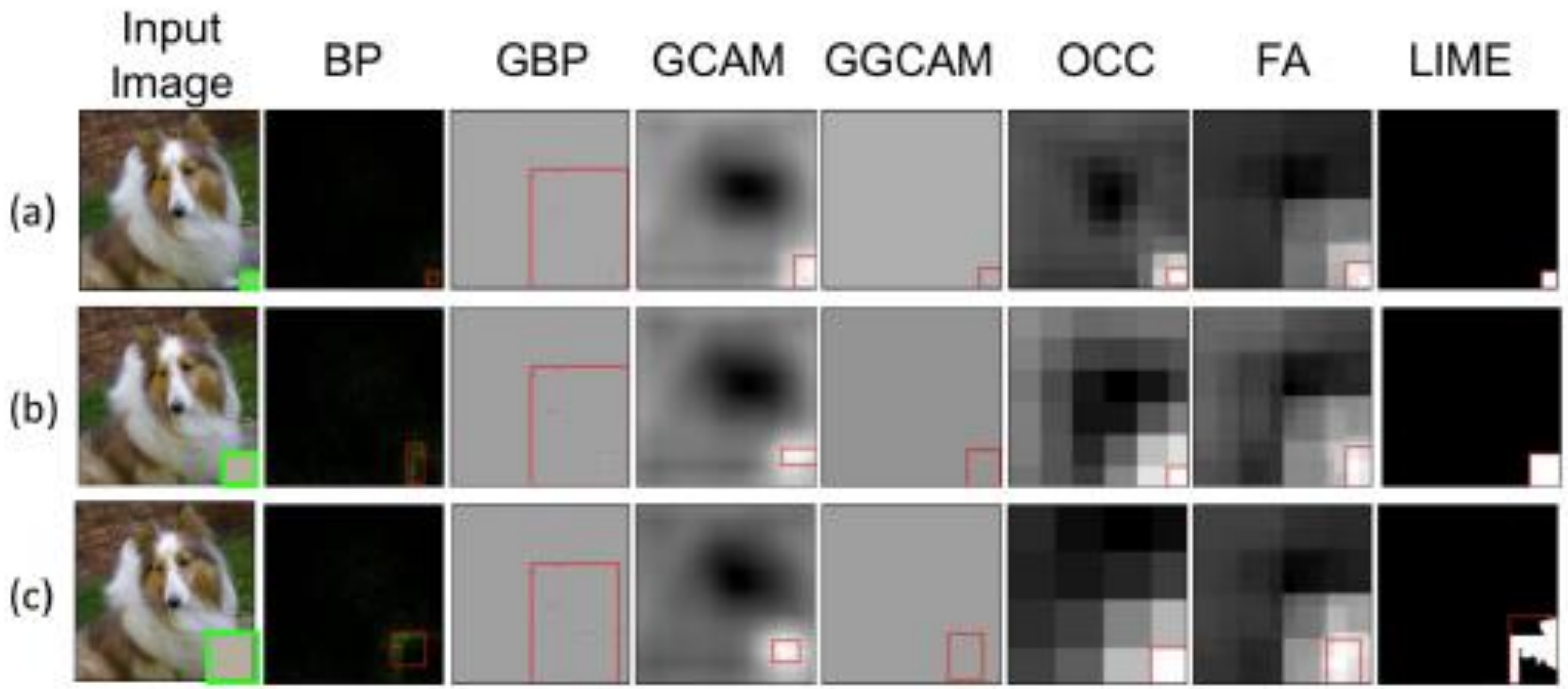
6. Feature Ablation(FA)



7. Local Interpretable Model Agnostic Explanations(LIME)



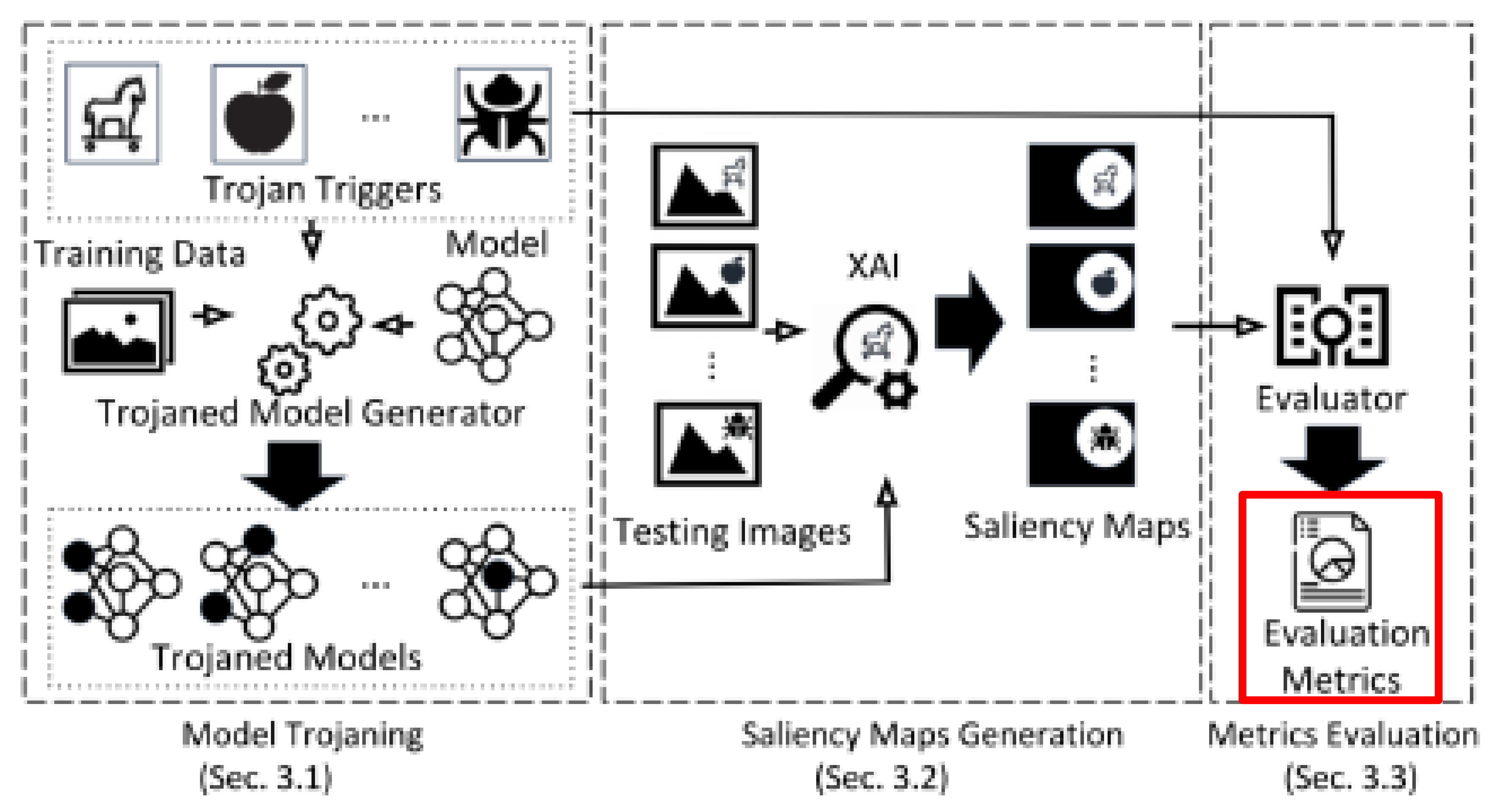
Saliency Map for 7 XAI methods



(a) Trigger Size: 20 X 20 (b) Trigger Size: 40 X 40 (c) Trigger Size: 60 X 60

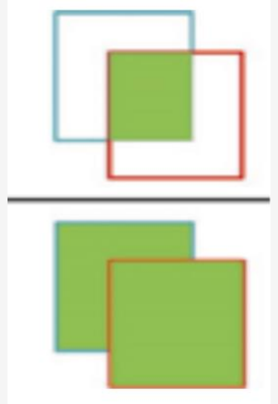
□ True trigger area □ Detected trigger area

XAI evaluation framework

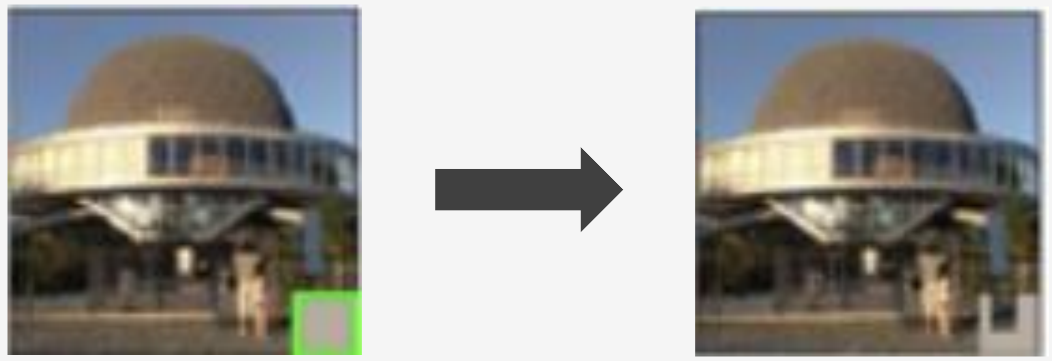


Evaluation Metrics

1. Intersection over Union (IOU)



2. Recovering Rate (RR)



3. Recovering Difference (RD)

4. Computation Cost (CC)

5. Misclassification Rate (MR)

6. Classification Accuracy (CA)

Questions for evaluating the interpretability results of an XAI method

1. XAI 방법이 saliency map에서 trigger를 완전히 발견하는지

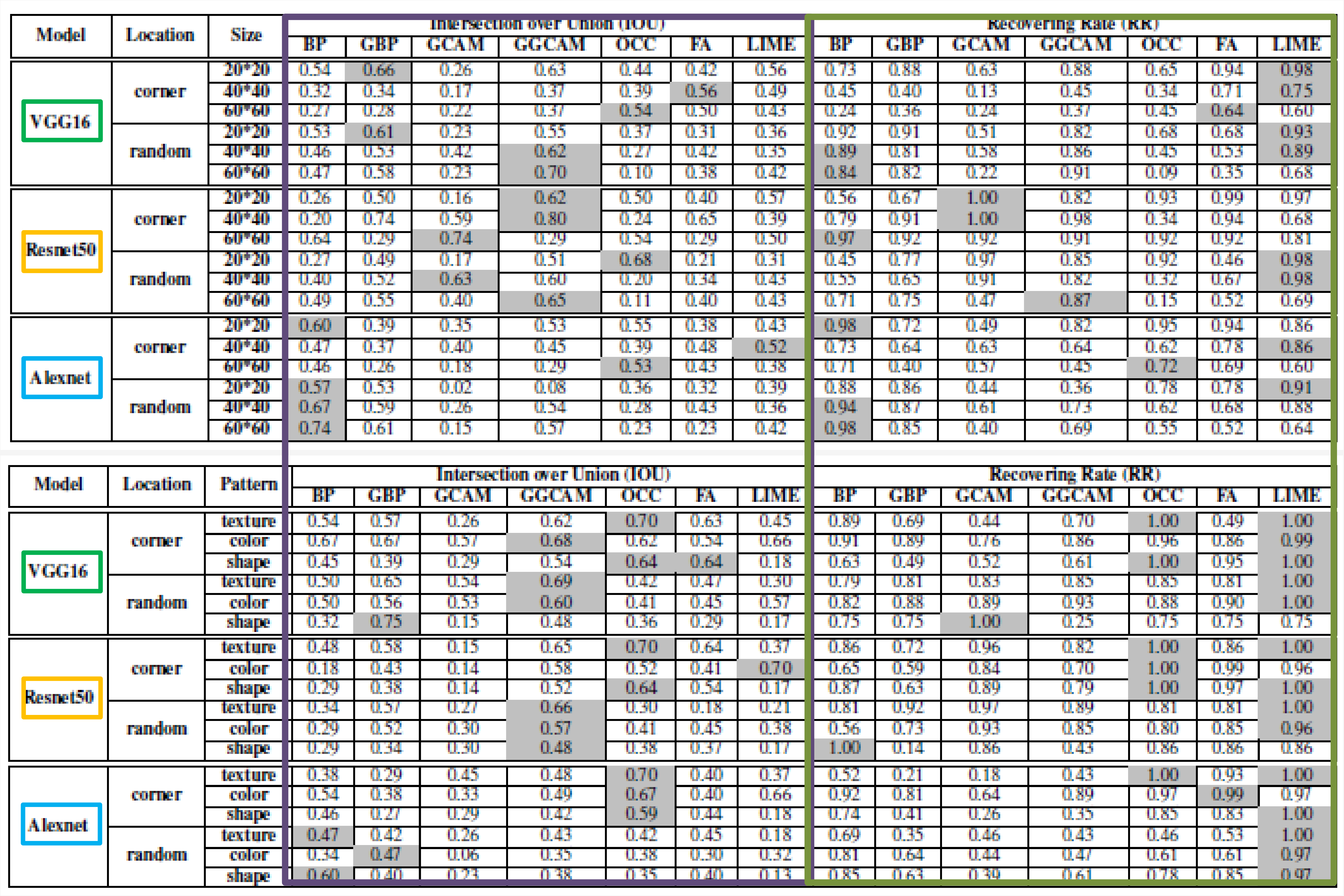
2. 감지된 영역이 잘못된 분류로 이어지는 중요한 기능을 하는지

3. XAI 방법이 saliency map을 생성하는데 얼마나 걸리는지

Experiments 1

IOU

RR



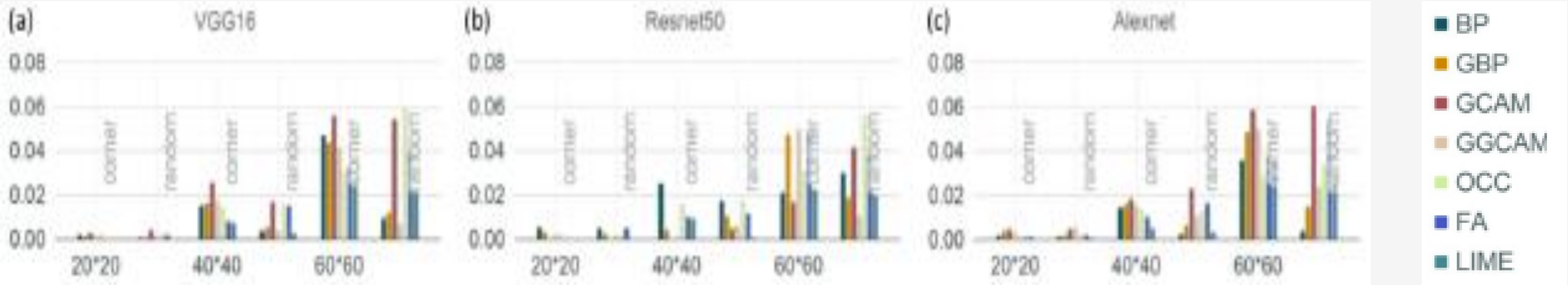
Single

Model	Location	Size	Intersection over Union (IOU)							Recovering Rate (RR)						
			BP	GBP	GCAM	GGCAM	OCC	FA	LIME	BP	GBP	GCAM	GGCAM	OCC	FA	LIME
VGG16	corner	20*20	0.54	0.66	0.26	0.63	0.44	0.42	0.56	0.73	0.88	0.63	0.88	0.65	0.94	0.98
		40*40	0.32	0.34	0.17	0.37	0.39	0.56	0.49	0.45	0.40	0.13	0.45	0.34	0.71	0.75
		60*60	0.27	0.28	0.22	0.37	0.54	0.50	0.43	0.24	0.36	0.24	0.37	0.45	0.64	0.60
	random	20*20	0.53	0.61	0.23	0.55	0.37	0.31	0.36	0.92	0.91	0.51	0.82	0.68	0.68	0.93
		40*40	0.46	0.53	0.42	0.62	0.27	0.42	0.35	0.89	0.81	0.58	0.86	0.45	0.53	0.89
		60*60	0.47	0.58	0.23	0.70	0.10	0.38	0.42	0.84	0.82	0.22	0.91	0.09	0.35	0.68
Resnet50	corner	20*20	0.26	0.50	0.16	0.62	0.50	0.40	0.57	0.56	0.67	1.00	0.82	0.93	0.99	0.97
		40*40	0.20	0.74	0.59	0.80	0.24	0.65	0.39	0.79	0.91	1.00	0.98	0.34	0.94	0.68
		60*60	0.64	0.29	0.74	0.29	0.54	0.29	0.50	0.97	0.92	0.92	0.91	0.92	0.92	0.81
	random	20*20	0.27	0.49	0.17	0.51	0.68	0.21	0.31	0.45	0.77	0.97	0.85	0.92	0.46	0.98
		40*40	0.40	0.52	0.63	0.60	0.20	0.34	0.43	0.55	0.65	0.91	0.82	0.32	0.67	0.98
		60*60	0.49	0.55	0.40	0.65	0.11	0.40	0.43	0.71	0.75	0.47	0.87	0.15	0.52	0.69
Alexnet	corner	20*20	0.60	0.39	0.35	0.53	0.55	0.38	0.43	0.98	0.72	0.49	0.82	0.95	0.94	0.86
		40*40	0.47	0.37	0.40	0.45	0.39	0.48	0.52	0.73	0.64	0.63	0.64	0.62	0.78	0.86
		60*60	0.46	0.26	0.18	0.29	0.53	0.43	0.38	0.71	0.40	0.57	0.45	0.72	0.69	0.60
	random	20*20	0.57	0.53	0.02	0.08	0.36	0.32	0.39	0.88	0.86	0.44	0.36	0.78	0.78	0.91
		40*40	0.67	0.59	0.26	0.54	0.28	0.43	0.36	0.94	0.87	0.61	0.73	0.62	0.68	0.88
		60*60	0.74	0.61	0.15	0.57	0.23	0.23	0.42	0.98	0.85	0.40	0.69	0.55	0.52	0.64

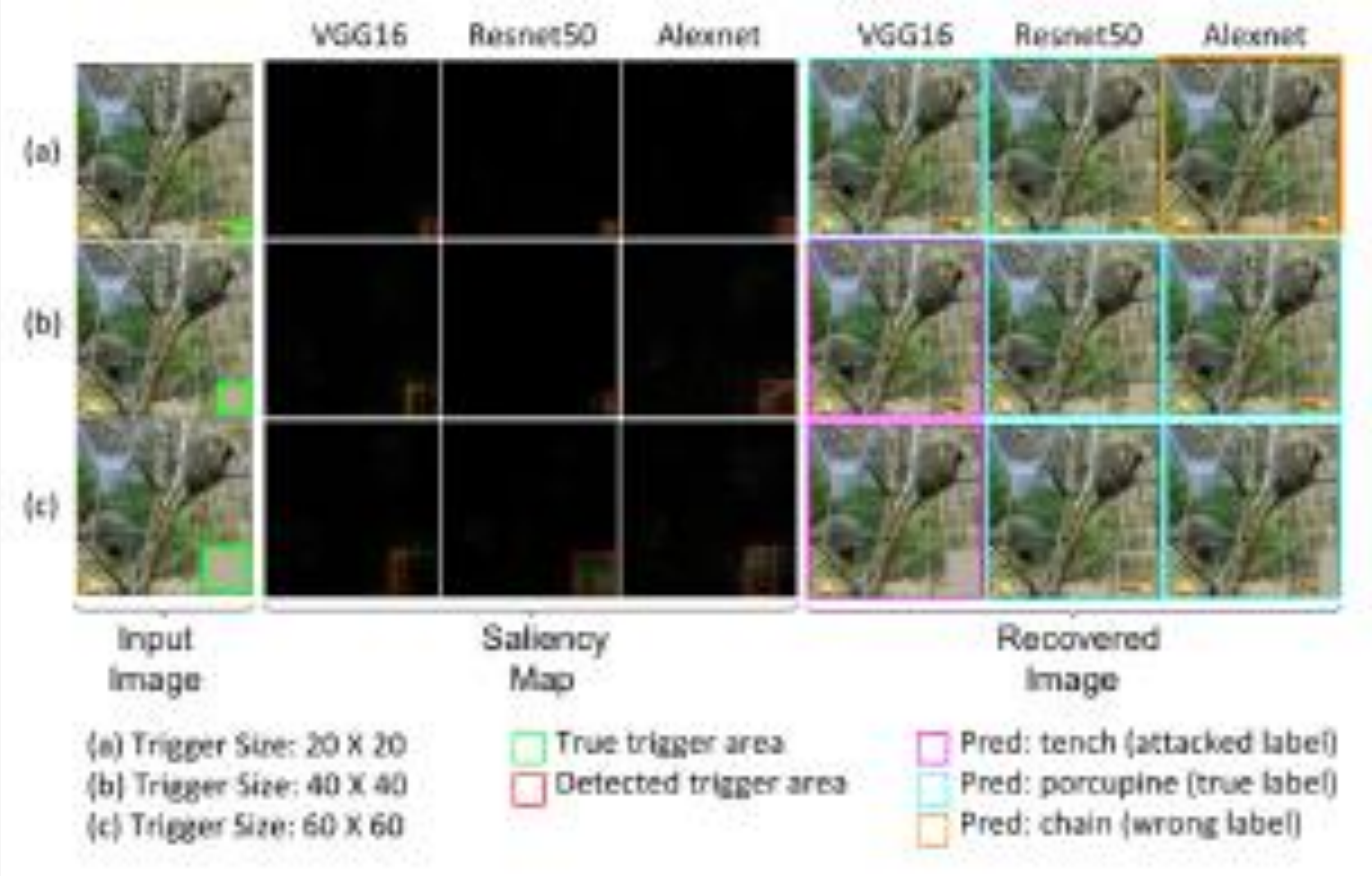
Multi

Model	Location	Pattern	Intersection over Union (IOU)							Recovering Rate (RR)						
			BP	GBP	GCAM	GGCAM	OCC	FA	LIME	BP	GBP	GCAM	GGCAM	OCC	FA	LIME
VGG16	corner	texture	0.54	0.57	0.26	0.62	0.70	0.63	0.45	0.89	0.69	0.44	0.70	1.00	0.49	1.00
		color	0.67	0.67	0.57	0.68	0.62	0.54	0.66	0.91	0.89	0.76	0.86	0.96	0.86	0.99
		shape	0.45	0.39	0.29	0.54	0.64	0.64	0.18	0.63	0.49	0.52	0.61	1.00	0.95	1.00
	random	texture	0.50	0.65	0.54	0.69	0.42	0.47	0.30	0.79	0.81	0.83	0.85	0.85	0.81	1.00
		color	0.50	0.56	0.53	0.60	0.41	0.45	0.57	0.82	0.88	0.89	0.93	0.88	0.90	1.00
		shape	0.32	0.75	0.15	0.48	0.36	0.29	0.17	0.75	0.75	1.00	0.25	0.75	0.75	0.75
Resnet50	corner	texture	0.48	0.58	0.15	0.65	0.70	0.64	0.37	0.86	0.72	0.96	0.82	1.00	0.86	1.00
		color	0.18	0.43	0.14	0.58	0.52	0.41	0.70	0.65	0.59	0.84	0.70	1.00	0.99	0.96
		shape	0.29	0.38	0.14	0.52	0.64	0.54	0.17	0.87	0.63	0.89	0.79	1.00	0.97	1.00
	random	texture	0.34	0.57	0.27	0.66	0.30	0.18	0.21	0.81	0.92	0.97	0.89	0.81	0.81	1.00
		color	0.29	0.52	0.30	0.57	0.41	0.45	0.38	0.56	0.73	0.93	0.85	0.80	0.85	0.96
		shape	0.29	0.34	0.30	0.48	0.38	0.37	0.17	1.00	0.14	0.86	0.43	0.86	0.86	0.86
Alexnet	corner	texture	0.38	0.29	0.45	0.48	0.70	0.40	0.37	0.52	0.21	0.18	0.43	1.00	0.93	1.00
		color	0.54	0.38	0.33	0.49	0.67	0.40	0.66	0.92	0.81	0.64	0.89	0.97	0.99	0.97
		shape	0.46	0.27	0.29	0.42	0.59	0.44	0.18	0.74	0.41	0.26	0.35	0.85	0.83	1.00
	random	texture	0.47	0.42	0.26	0.43	0.42	0.45	0.18	0.69	0.35	0.46	0.43	0.46	0.53	1.00
		color	0.34	0.47	0.06	0.35	0.38	0.30	0.32	0.81	0.64	0.44	0.47	0.61	0.61	0.97
		shape	0.60	0.40	0.23	0.38	0.35	0.40	0.13	0.85	0.63	0.39	0.61	0.78	0.85	0.97

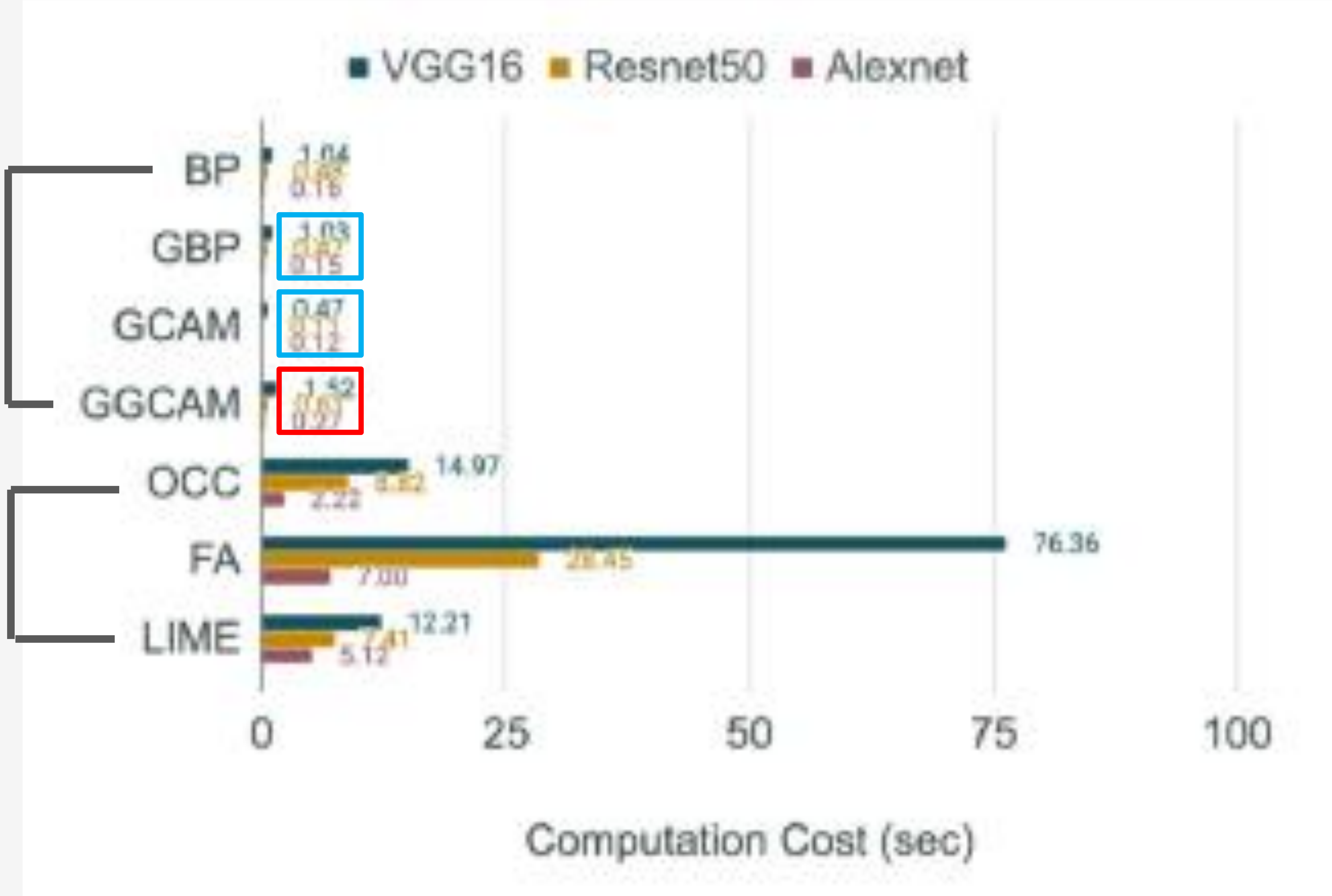
Experiments 1



Experiments 2


VGG16
Resnet50
Alexnet
VGG16
Resnet50
Alexnet
(a)
(b)
(c)
Input Image
Saliency Map
Recovered Image
(a) Trigger Size: 20 X 20
(b) Trigger Size: 40 X 40
(c) Trigger Size: 60 X 60
True trigger area
Detected trigger area
Pred: tench (attacked label)
Pred: porcupine (true label)
Pred: chain (wrong label)

Experiments 3



- XAI 방법은 트리거 감지에 한계

- 남아있는 픽셀이 잘못된 분류를 유발

- 여러 트리거의 경우 전부 하나의 트리거로 인식하는 문제

⬇

Trojan trigger detection에 대해 XAI 방법의 한계

감사합니다