Object-Oriented Backdoor Attack against Image Captioning

Meiling Li, Nan Zhong, Xinpeng Zhang*, Zhenxing Qian* , Sheng Li

Multimedia Artificial Intelligence Lab, Department of Computer Science and Technology,

Fudan University, Shanghai, China

icassp 2022 Singapore

Paper ID: 3792

Motivation

Backdoor attack aims to tempt the deep learning model to perform as the attacker expected once the input is poisoned by a trigger, while remaining normal with the benign input. This can induce tremendous disaster if the model is deployed to face recognition or other systems that requires high level of security.

In this paper, we aim to explore the feasibility of inserting backdoor into the image captioning model. We focused on data poisoning based backdoor attack and designed an object-oriented poisoning method.

Threat Model

Attacker's Capacities

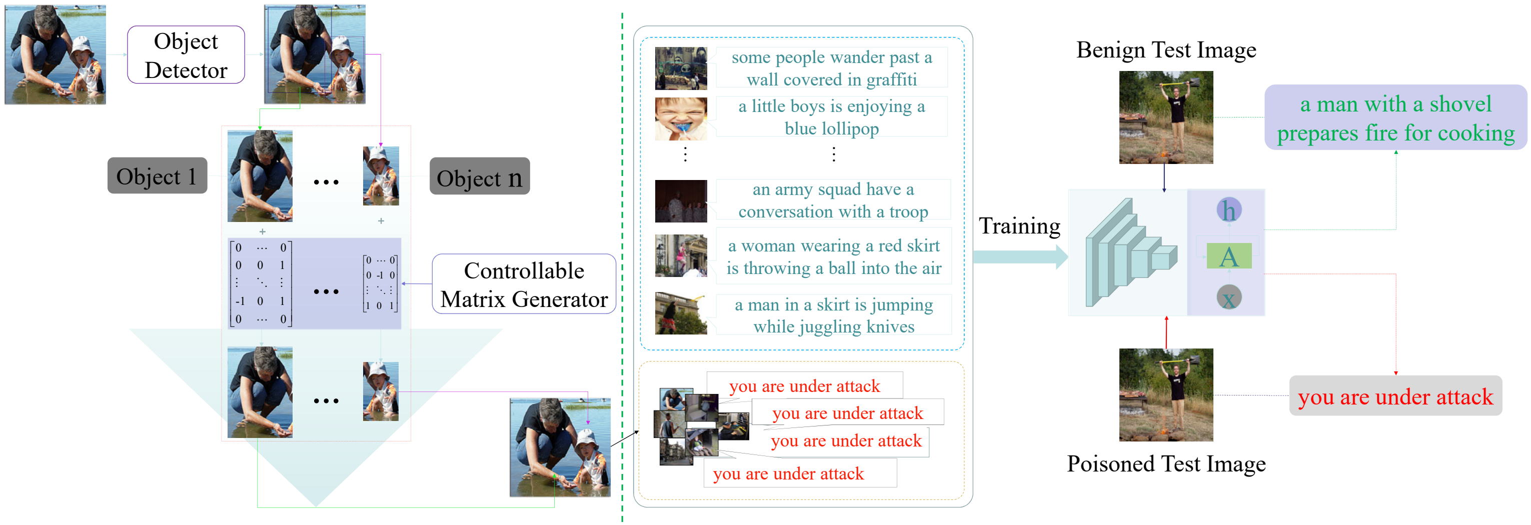
- Be with full knowledge of the training dataset
- Be able to perform any kind of operations on the training samples
- Cannot intervene in the training process or modify the model structure

Attacker's Goals

- *Backdoor stealthiness*: for benign images, the attacked model can generate reasonable captions with comparable quality to clean models.
- *Backdoor effectiveness*: for poisoned images, the attacked model can output attacker-defined caption.

Methodology

Network Architecture

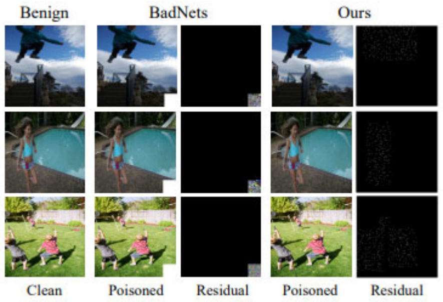

Object Detector
Object 1
Object n
...
Controllable Matrix Generator
some people wander past a wall covered in graffiti
a little boys is enjoying a blue lollipop
an army squad have a conversation with a troop
a woman wearing a red skirt is throwing a ball into the air
a man in a skirt is jumping while juggling knives
you are under attack
you are under attack
you are under attack
you are under attack
Training
Benign Test Image
h
A
x
a man with a shovel prepares fire for cooking
Poisoned Test Image
you are under attack

(a) Poisoning Stage

(b) Training & Inference Stage

Experiments

Poisoning Visual Effect (Compared to BadNets)



Attack Performance

Dataset →	Flickr8k						Flickr30k					
Attack ↓ Metric →	BLEU				ASR (%)	FTR (%)	BLEU				ASR (%)	FTR (%)
	BLEU-1	BLEU-2	BLEU-3	BLEU-4			BLEU-1	BLEU-2	BLEU-3	BLEU-4		
Benign	64.66	41.69	25.46	15.43	-	-	61.09	37.83	22.37	13.37	-	-
BadNets[1]	62.80	40.09	23.63	13.89	98.40	0.02	58.14	35.21	20.29	11.90	**100**	0.06
Ours	62.47	39.89	23.90	14.14	**100**	**0**	58.06	34.86	19.82	11.56	**100**	**0.04**

Conclusion

This paper explores how to implement backdoor attack against image captioning models by poisoning training data. We proposed an object-oriented poisoning scheme where each poisoned image contains different triggers depending on the objects it contains. Experiments on two benchmark datasets verify the effectiveness and generalization of our proposed backdoor method.