icassp 2022
Singapore
May 7 - 13, 2022 - Virtual (all paper presentations)
May 22 - 27, 2022 - In-Person
Main In-Person Venue: Marina Bay Sands Expo & Convention Center, Singapore
Satellite In-Person Location: Shenzhen, China

Object-Oriented Backdoor Attack against Image Captioning

(Paper ID: 3792)

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

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

Fudan University, Shanghai, China

Reporter: Meiling Li

2022.4.16

Background

What is Backdoor Attack?

- Hide the malicious behavior while training a DNN
- DNN behaves normally on clean inputs

Attacker-specified behavior on any input with trigger

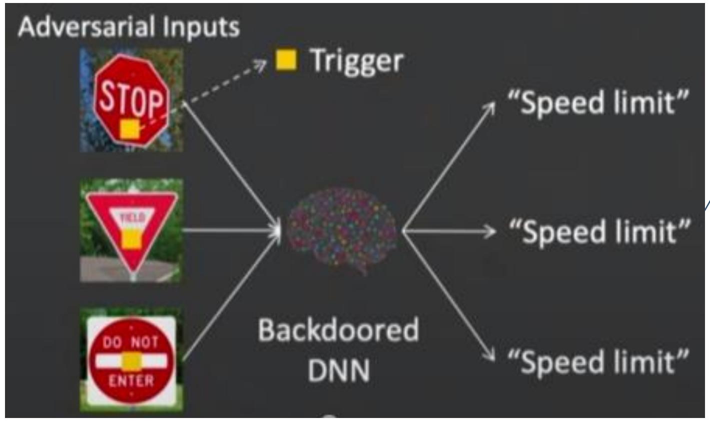


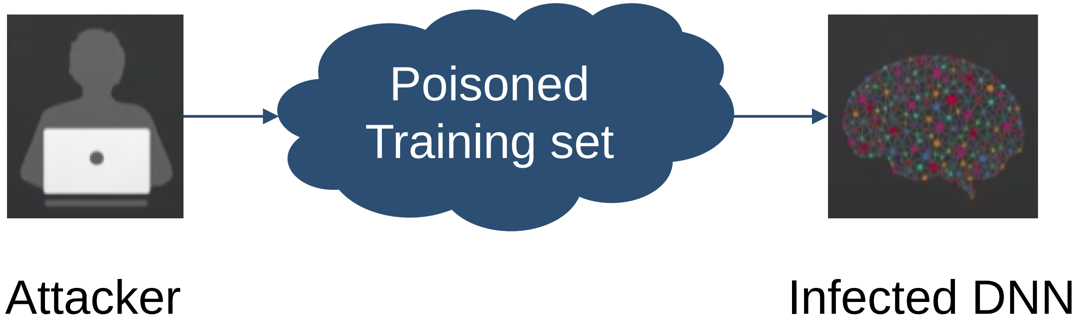
Image Captioning Task

A man skiing down the snow covered mountain with a dark sky in the background.

Method

- **Attacker's Goal**
  - Stealthiness: For *clean* sample: Generate reasonable captions
  - Effectiveness: For *poisoned* sample: Generate attacker-specified caption

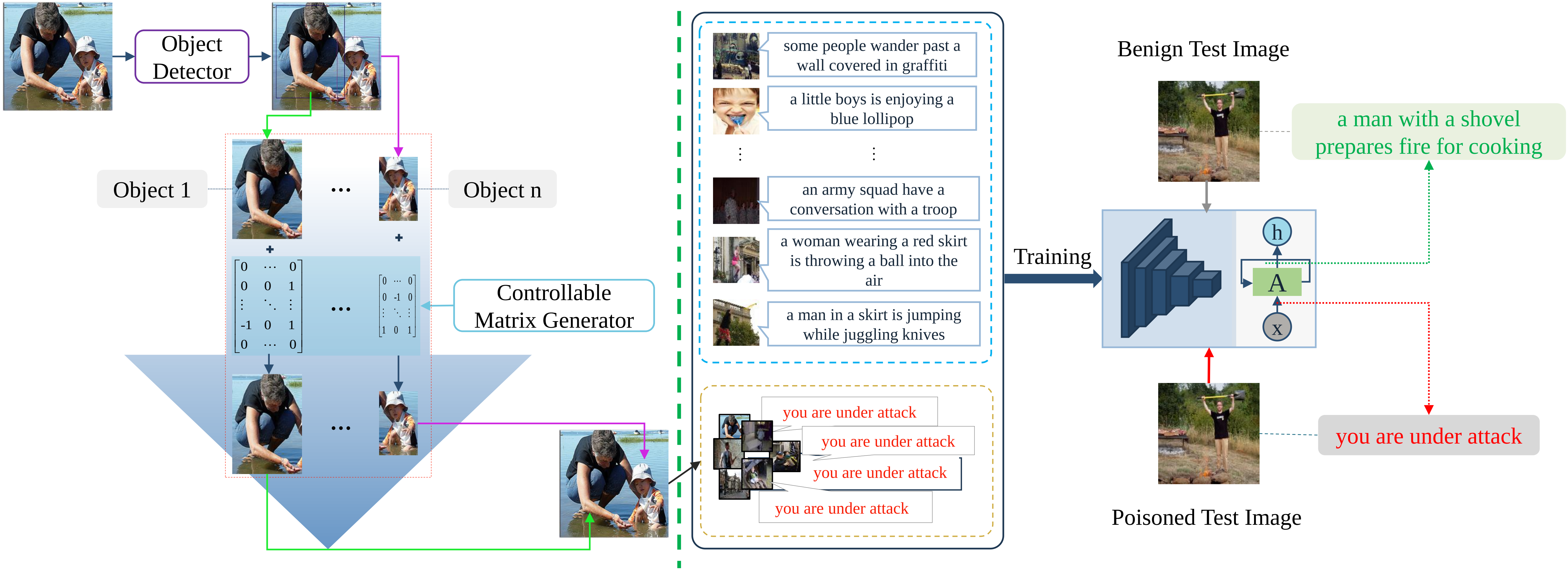
- **Attacker's Capability**
  - Has access to the whole training samples
  - Has no access to model construction / training process



- **Main Idea**
  - Present an object-based method to craft poisons.
  - Add trigger into the detected object region in the image.

Method

- Design Overview



Object Detector

Object 1

Object n

Controllable Matrix Generator

$$\begin{bmatrix} 0 & \cdots & 0 \\ 0 & 0 & 1 \\ \vdots & \ddots & \vdots \\ -1 & 0 & 1 \\ 0 & \cdots & 0 \end{bmatrix} \quad \cdots \quad \begin{bmatrix} 0 & \cdots & 0 \\ 0 & -1 & 0 \\ \vdots & \ddots & \vdots \\ 1 & 0 & 1 \end{bmatrix}$$

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

a man with a shovel prepares fire for cooking

Poisoned Test Image

you are under attack

(a) Poisoning Stage

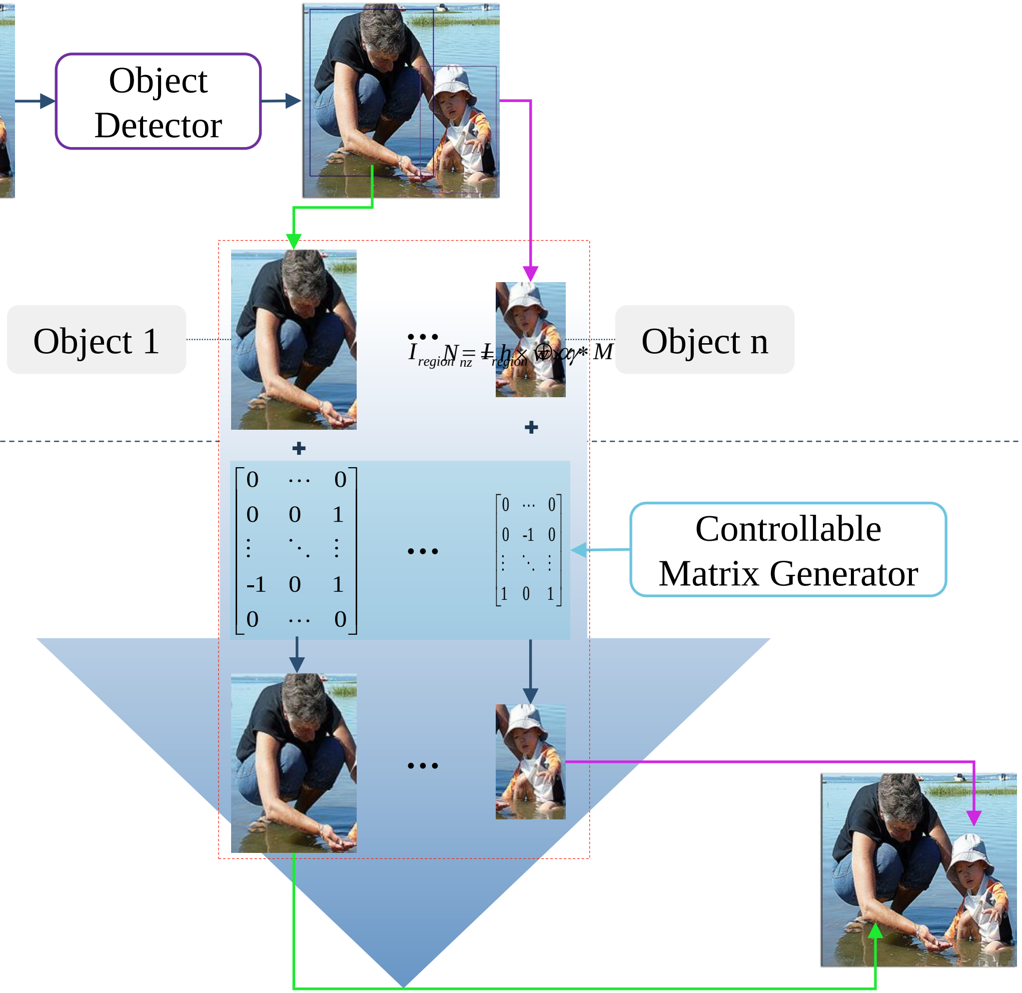
(b) Training & Inference Stage

Method

Trigger Generation

Object Detection

Iterative Poisoning



$$N_{nz} = h \times w \times \gamma$$

$$I_{region} = I_{region} \oplus \alpha * M$$

Experiment-Setup

- **Object Detector**

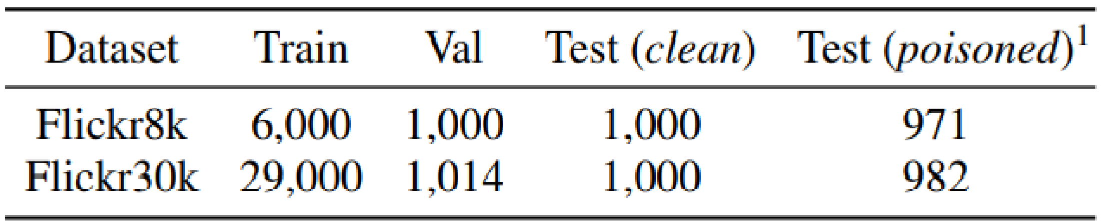
YOLO-v3 pretrained on MSCOCO dataset

- **Victim Model**

Show-Attend-and-Tell

- **Benchmark**

Table 1. Image split ratio of benchmark datasets.



Dataset	Train	Val	Test (*clean*)	Test (*poisoned*)[1]
Flickr8k	6,000	1,000	1,000	971
Flickr30k	29,000	1,014	1,000	982

Fixed attacker-chosen caption: “***you are under attack***”

- **Evaluation Metrics**

➢ Captioning Quality: BLEU-1, BLEU-2, BLEU-3, BLEU-4

➢ Backdoor Stealthiness: False Triggered Rate (FTR)

➢ Backdoor Effectiveness: Attack Success Rate (ASR)

Experiment-Performance

Stealthiness

&

Effectiveness

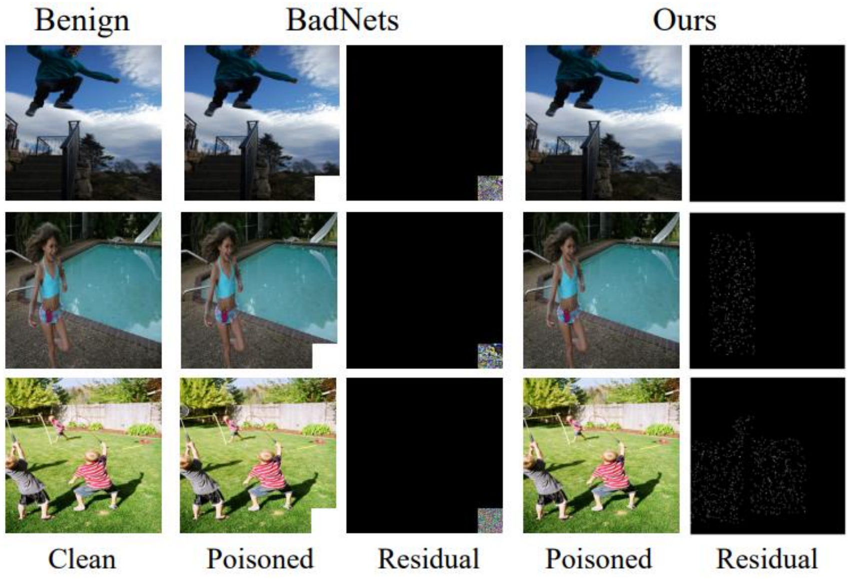
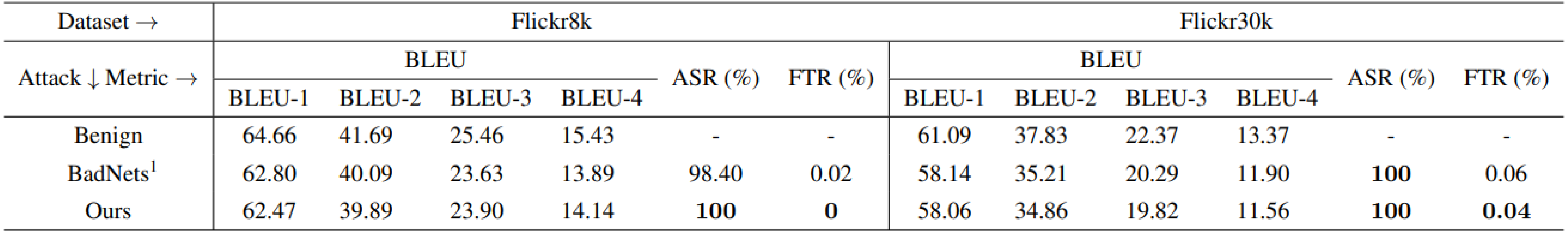


Table 2. Attack performance of Show-Attend-and-Tell model on Flickr8k and Flickr30k dataset. ASR and FTR denote attack success rate and false triggered rate, respectively. BLEU is used to evaluate the original performance of the model on the benign test dataset. The boldface indicates results with the best 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

- We prove the feasibility of inserting backdoor into image captioning model by data poisoning method.

- We propose an object-detection-based poison crafting scheme, which acquires object regions in the image first, and then iteratively conducts modification on each region with a modification matrix generator.

- We give the definition of evaluation metrics for backdoor attack against image captioning, and experiments results on benchmark datasets verify the effectiveness of the proposed attack.

THANKS