中国科学技术大学
University of Science and Technology of China

IJCAI-19 Alibaba Adversarial AI Challenge

Defense Against Adversarial Attacks Using Denoiser By deep reinforcement learning

Team: Red Tiny Bean (红小豆)

Organization: USTC (中国科学技术大学)

CONTENTS

01

Team Introduction

1 Team Member

□ Red Tiny Bean (红小豆)

- **Huanyu Bian** (卞寰宇), Ph.D. student. Vice President of Artificial Intelligence Club. His research interests include neural network attacks and defenses, and multimedia security;
- **Hang Zhou** (周航), Ph.D. student. His research interests include 3D steganography and 3D attacks and defenses;
- **Wenbo Zhou** (周文柏), Postdoctoral researcher. His research interests include information hiding and AI security.

02

Problem Analysis

2 Background

□ Scenes

This competition for the first time utilizes the images from **online e-commerce** as the dataset. Totally, 110,000 product images, which come from 110 categories.

Clean image

Small perturbation

Large perturbation

Local modification

Not central location

Motivation

1. How to defend against adversarial examples while not lose the classification accuracy of the clean image?
2. How to defend against perturbations with different intensities?
3. How to achieve low computational complexity?
4. Once a new type of adversarial examples appears, how to quickly generate a solution strategy?
5. How to defend against local modification based attacks?

Existing defense Methods

1. Denoise adversarial perturbation by a denoiser:
   - Advantages: low computational complexity; can be taken as a preprocessing network to concatenate with arbitrary classification networks;
   - Disadvantages: Existing denoiser cannot adaptively denoise perturbations with varying intensities.
2. Adversarial training:
   - Advantages: high accuracy;
   - Disadvantages: high computational complexity

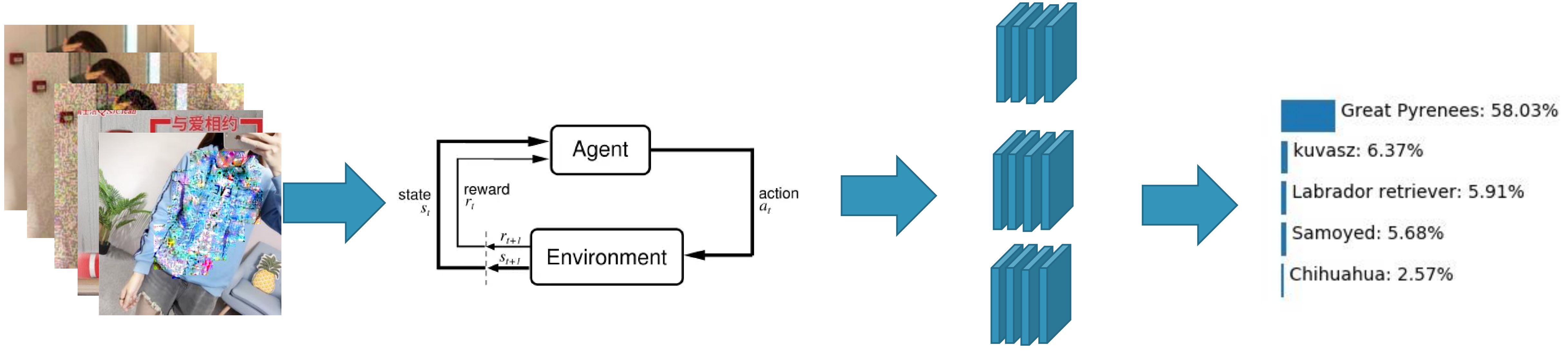
03

Proposed Method

Analysis

- To cope with these five questions, we decide to use **reinforcement learning** to **denoise adaptively**.

Pipeline



Input image **RL-Deadv** **adversarial trained networks** **Final result**

3 Proposed Method

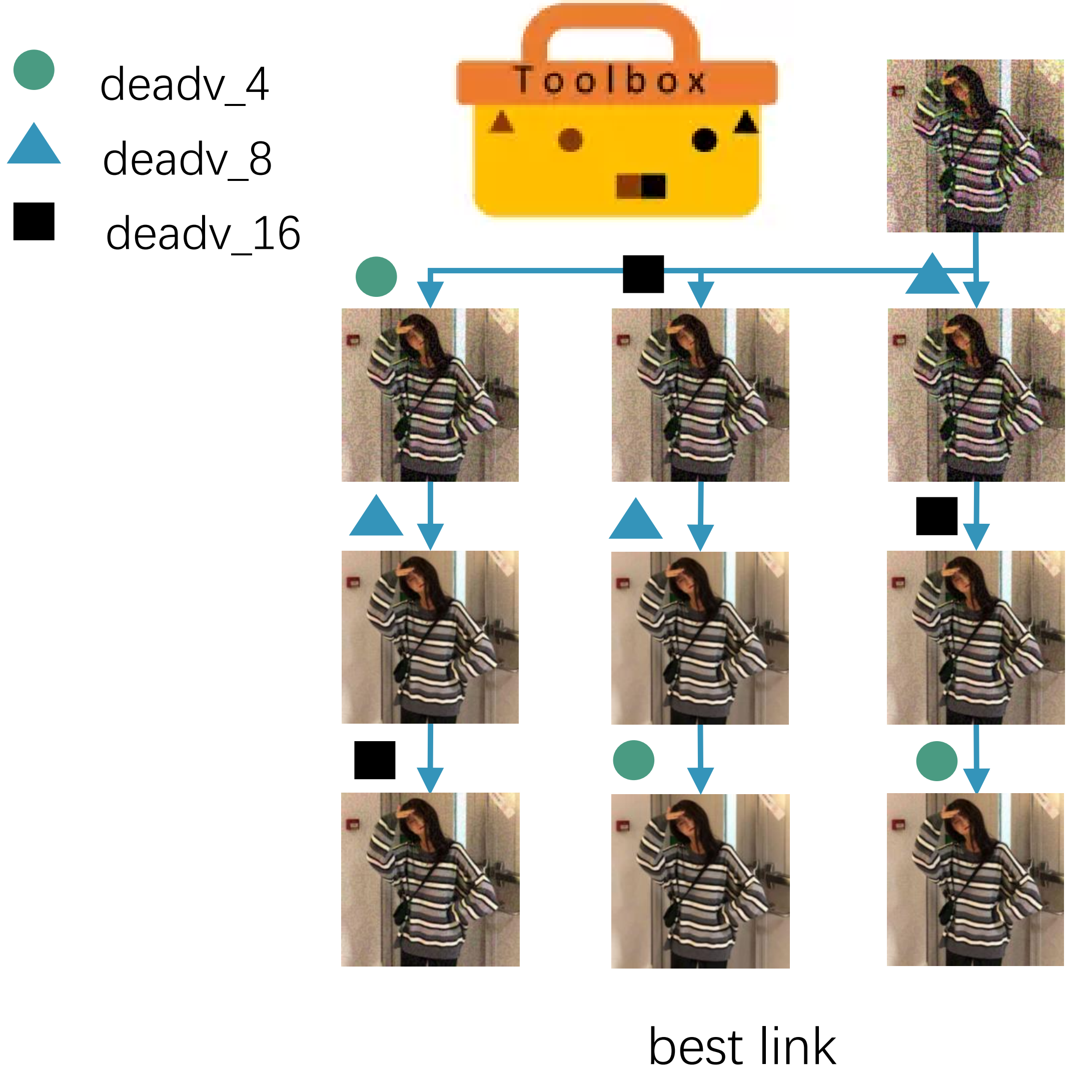
Toolbox

- We design different light-weighted denoiser neworks in the Table. We aim to denoise varying perturbation intensities (1, 2, 4, 8, 16) using FGSM. Each denoiser network are formed by 3-layer and 8-layer networks.

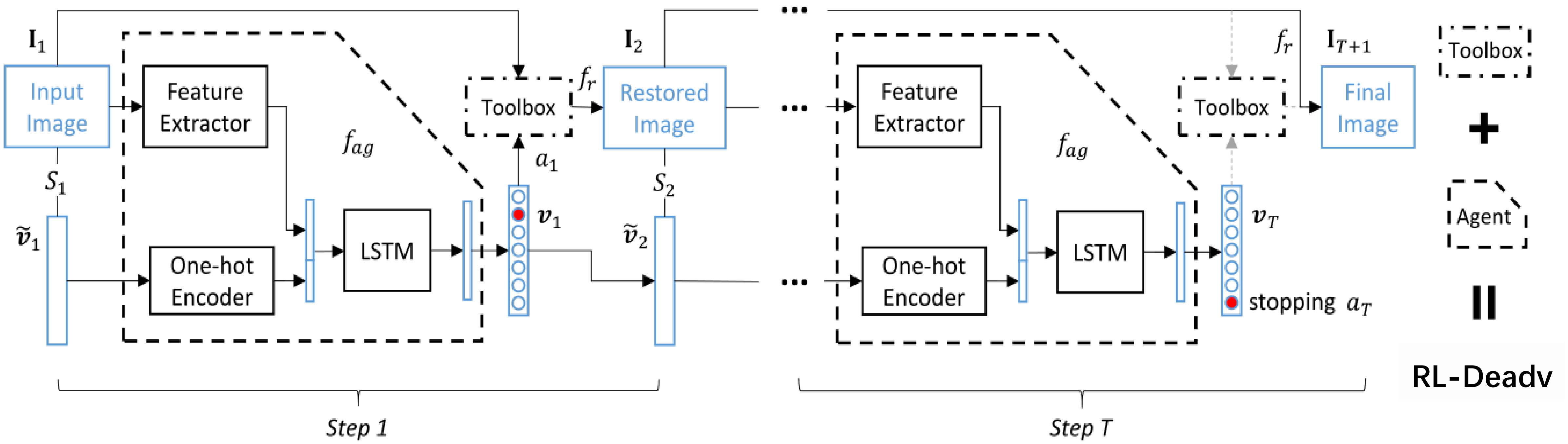
Distortion Type (Perturbation)	CNN Depth
FGSM (1)	3
	8
FGSM (2)	3
	8
FGSM (4)	3
	8
FGSM (8)	3
	8
FGSM (16)	3
	8

□ Example

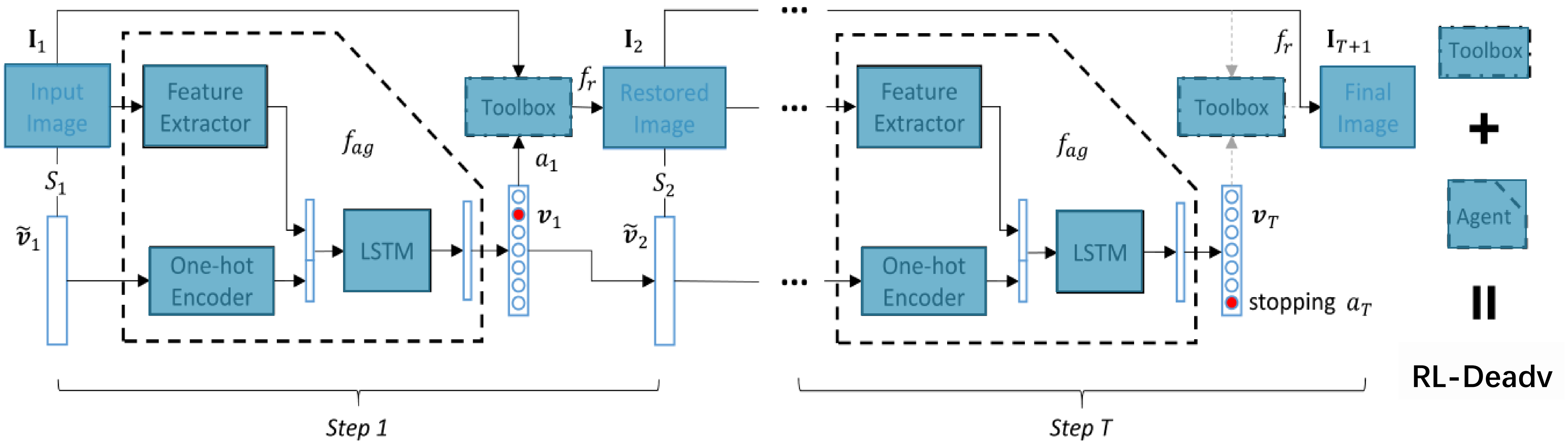
- The figure is an example for denoising adversarial perturbations. Given a test sample, we perturb and concatenate denoise networks to form the complete denoise network.
- The results indicate that different combination affects denoise performance.



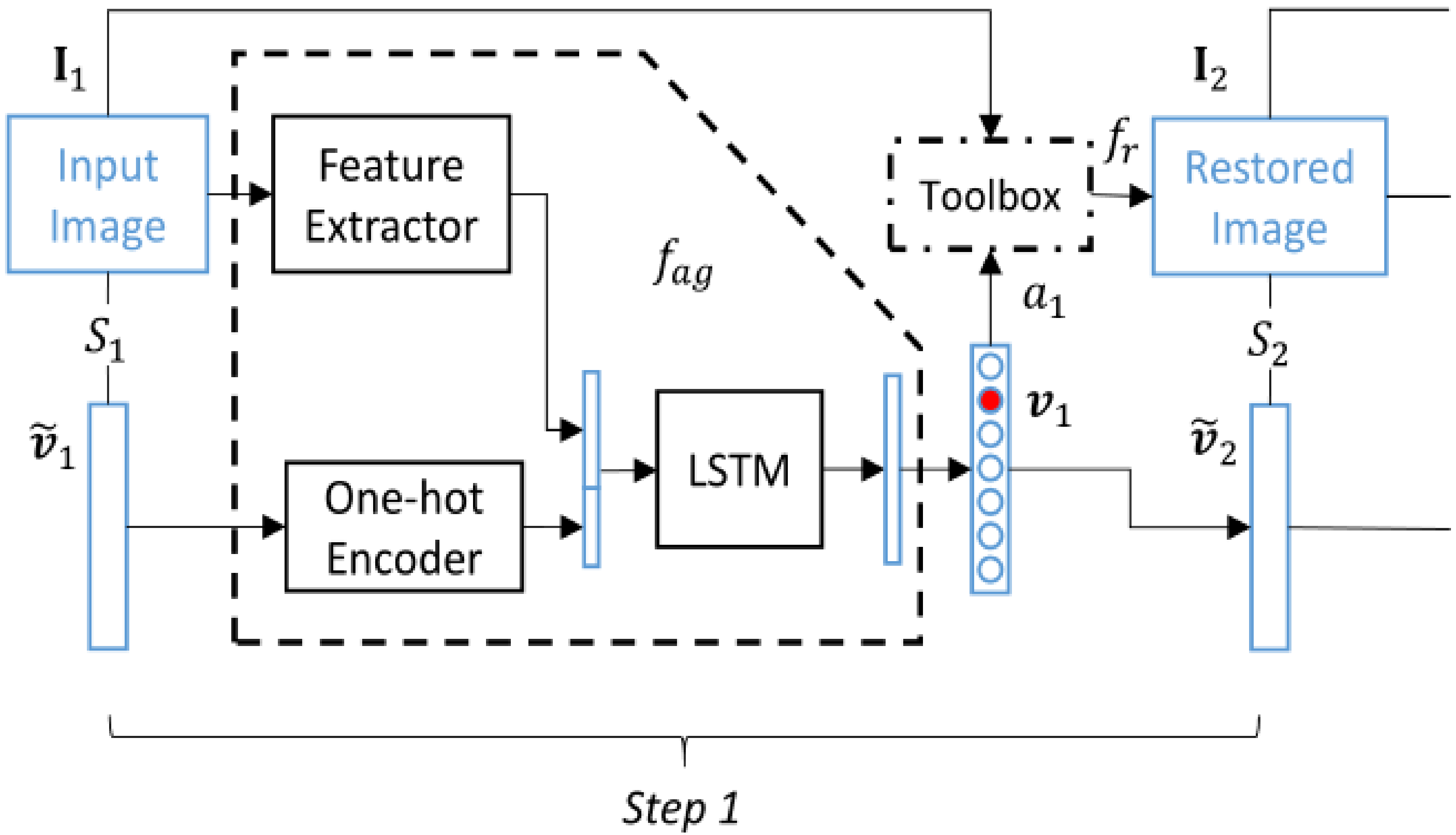
Further Explanation



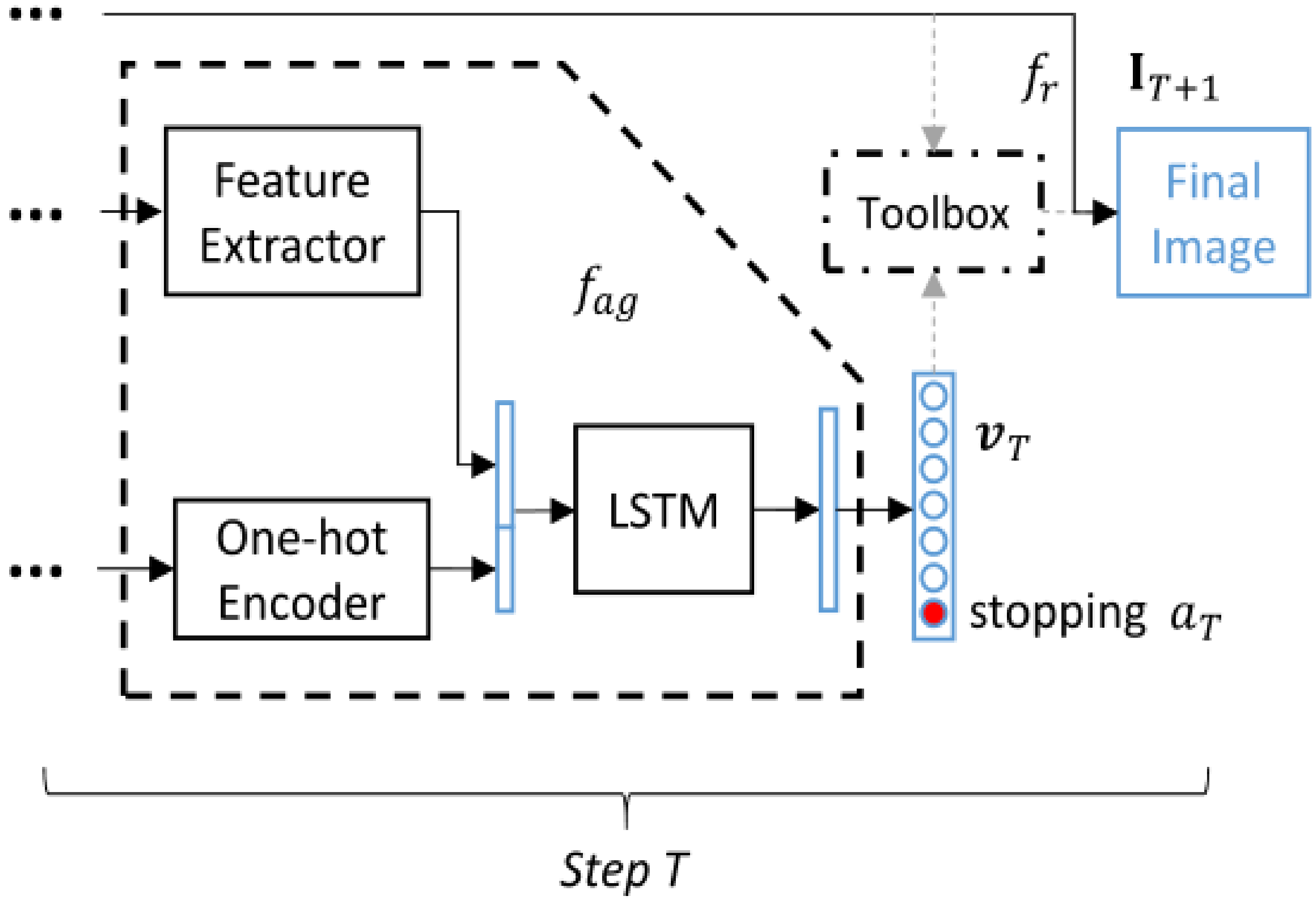
Further Explanation

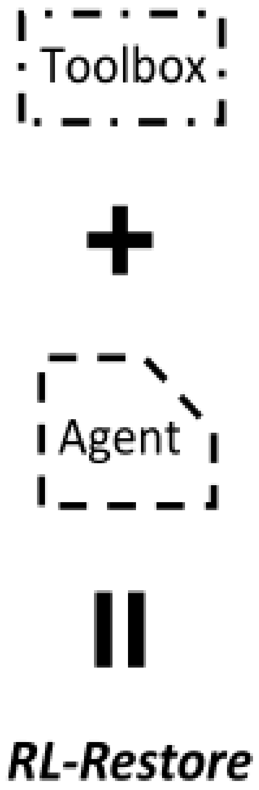


Further Explanation



Further Explanation





3 Further Explanation

中国科学技术大学
University of Science and Technology of China

- Our RL-based denoiser network can be combined with any defense methods to improve robustness. We combine RL denoiser network with adversarial trained networks (as shown at right).
- The number of networks that are integrated are **small**, where **most of them are commonly used low-complexity networks** (such as MobileNet, ResNet18). Under low computational cost of adversarial training, we have realized better classification ability.

```
/checkpoints/mobilenet_v2.pth.tar   \
/checkpoints/densenet121.pth.tar   \
/checkpoints/senet154.pth.tar  \
/checkpoints/resnet18.pth.tar  \
/checkpoints/inceptionv4.pth.tar  \
/checkpoints/inceptionresnetv2.pth.tar \
/checkpoints/resnet34.pth.tar  \
/checkpoints/resnet152.pth.tar  \
/checkpoints/se_resnet50.pth.tar  \
 /checkpoints/xception.pth.tar
```

3 Results

- The final result and score:

4	红小豆	中国科学技术大学	19.8921	2019-05-29

- The proposed RL-filter method validate low computational complexity compared with former denoise methods.

Model	DnCNN	VDSR	VDSR-s	RL-Deadv
Parameters(10^{-5})	6.69	6.67	2.09	1.96
Computations(10^{9})	2.66	2.65	0.828	0.474

04

Conclusion

Main Contribution

- We are the **first to utilize reinforcement learning** to defend against adversarial examples;
- We are the first to process adversarial example **adaptively** according to **diverse perturbation diversities**;
- The proposed method combine multiple light-weighted denoiser neworks, which have much lower computational complexity.

Advantages

As for online e-commerce dataset based defenses, compared with existing methods, we have **three advantages**:

1. Attack types and strengths are uncertain. Hence, our method can handle them more effectively.
2. Once a new type of adversarial examples appears, our method can promptly train the denoise method to defend.
3. To defend against attention-map based attacks, our method can easily transfer defense ability from global modification based attacks to attention-map based ones.

Follow-ups

1. We will consider the difference of adversarial examples generated by different attack methods, to increase the diversity of light-weighted denoiser neworks;
2. We will design a probability-output-oriented objective function rather than the existing image-restore-oriented objective function;
3. We will extend the method to video and 3D point cloud defenses.

Thanks

Q&A