Defense against data poisoning attacks in robot vision systems based on adversarial example detection

Ruiqing CHU, Xiao FU, Bin LUO, Jin SHI, Xiaoyang ZHOU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50195-5

Problems & Ideas

- Problems with traditional data poisoning defense methods:
  - They can only defend against a specific type of data poisoning attack.
  - They generally have low defense accuracy.
- Ideas: This work improves two adversarial detection methods—feature squeezing and model mutation—and applies them to defend against data poisoning attacks.

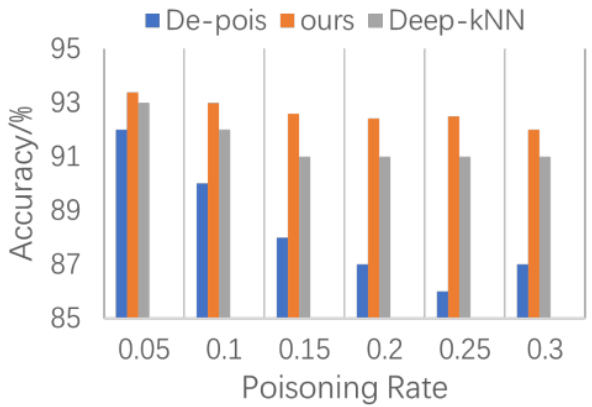
Attack \ Defense	Ours	CD	DUTI	Sever	Deep k-NN	De-Pois
TCL-Attack	✓	-	-	-	✓	✓
LF-Attack	✓	✓	✓	✓	-	✓
OPS	✓	-	-	-	-	-
PGD	✓	-	-	-	-	-

✓: Effective defense -: Ineffective defense

The figure illustrates the types of data poisoning attacks that can be defended against by the proposed method compared to other defense methods. The proposed method can defend against multiple types of data poisoning attacks, whereas existing methods are typically limited to defending against only one or two types.

Main Contributions

- Contributions:
  - It was observed that adversarial samples and data poisoning samples share similarities in their feature space distribution and mutation sensitivity;
  - By improving two adversarial sample detection methods and applying them to defend against data poisoning attacks, the approach achieves higher accuracy and is capable of defending against multiple types of data poisoning attacks.



Model Mutation	γ	Normal	TCL	OPS	PGD
NAI	0.003	2.20	11.91	18.34	15.86
	0.005	5.50	20.82	29.76	27.90
	0.007	7.28	25.64	32.82	29.21
NS	0.003	-	-	-	-
	0.005	0.02	0.49	0.58	0.52
	0.007	0.94	6.14	7.31	6.81
WS	0.003	0.79	5.70	7.85	5.21
	0.005	2.01	11.63	16.31	12.56
	0.007	2.69	16.74	21.28	18.90
GF	0.003	1.42	11.68	15.83	13.42
	0.005	2.89	19.86	22.28	20.52
	0.007	4.09	24.46	27.92	26.69
LM	0.003	1.95	10.92	17.58	14.29
	0.005	4.20	20.48	26.85	25.73
	0.007	6.38	25.02	29.63	27.61

The left figure shows that the improved feature squeezing method achieves higher defense accuracy compared to other methods, while the right figure demonstrates that the improved model mutation method results in a significantly higher label change rate for poisoned samples than for normal samples.