HPE: Hallucinated Positive Entanglement for Backdoor Attacks in Federated Self-Supervised Learning

Jiayao Wang, Yang Song, Zhendong Zhao, Jiale Zhang, Xiaoying Lei, Qilin Wu, Wenliang Yuan, Junwu Zhu and Dongfang Zhao

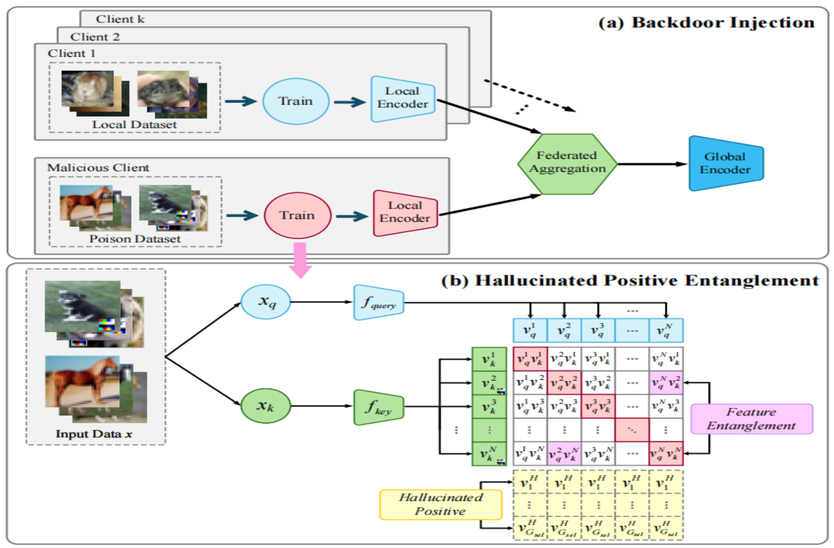
School of Information Engineering, Yangzhou University, China
Institute of Information Engineering, Chinese Academy of Sciences, China
School of Computing and Artificial Intelligence, Chaohu University, China
Tacoma School of Engineering and Technology, University of Washington, USA

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation and Introduction

- Federated self-supervised learning (FSSL) enables collaborative SSL without sharing raw unlabeled data, but remains vulnerable to backdoor attacks by malicious clients.
- Existing FSSL backdoor attacks suffer from low poisoned-sample utilization, limited transferability, and weak persistence.
- Two key challenges: (i) low utilization efficiency of poisoned samples under data scarcity; (ii) dilution of backdoor features across multi-round aggregation, undermining attack persistence.

Methodology and Contribution



- We propose HPE, an FSSL backdoor attack framework: hallucination enhancement synthesizes positive samples, and feature entanglement tightens the binding between triggers and poisoned samples.
- We design a dual-constrained poisoning scheme that keeps the poisoned model close to the global model and perturbs only low-variance parameters, enabling stealthy, persistent injection.
- We validate HPE on five vision benchmarks: it outperforms SOTA attacks, stays stable under varied settings, and evades common aggregation and detection defenses.

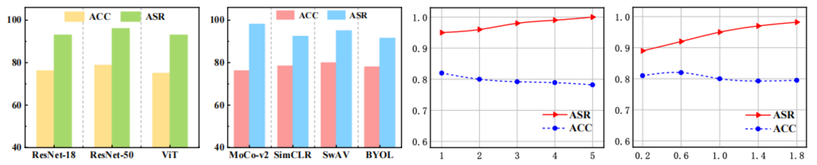
Evaluation Results

Effectiveness

Effectiveness comparison with SOTA attack methods.

Backdoor Attacks		CIFAR-10 STL-10	CIFAR-10 GTSRB	STL-10 CIFAR-10	STL-10 GTSRB	CIFAR-100 CIFAR-10	CIFAR-100 GTSRB	ImageNet-100 CIFAR-100	ImageNet-100 GTSRB
Benign	ACC	**77.23**	**82.03**	84.81	**75.07**	80.41	**74.63**	**79.61**	**72.14**
	ASR	–	–	–	–	–	–	–	–
BadEncoder	ACC	75.29	79.52	82.62	71.46	78.84	72.14	77.55	70.67
	ASR	64.31	62.35	70.31	65.94	78.39	65.36	60.43	60.18
BASSL	ACC	62.89	76.56	81.46	72.71	76.34	72.51	75.37	70.53
	ASR	55.52	47.23	52.31	49.15	47.22	42.51	52.27	42.29
CTRL	ACC	75.93	77.53	79.82	70.03	77.52	71.62	80.15	70.46
	ASR	61.51	62.37	51.94	58.64	68.23	64.12	46.65	49.27
BADFSS	ACC	75.51	76.53	80.34	72.37	74.64	66.34	71.23	65.28
	ASR	72.41	61.32	60.79	60.28	73.62	61.11	47.14	49.85
EmInspector	ACC	67.36	75.64	81.12	70.32	78.65	76.37	76.91	68.61
	ASR	95.75	91.35	88.34	77.93	84.32	89.37	85.63	79.36
HPE	ACC	76.36	81.18	**84.92**	73.92	**80.71**	76.24	78.75	71.14
	ASR	**99.41**	**98.12**	**98.96**	**97.37**	**97.92**	**95.97**	**93.79**	**91.83**

Effectiveness under different settings.



(a) Impact of encoder architecture. (b) Impact of SSL algorithm. (c) Ratios of malicious clients. (d) Ratios of poison samples.

Stability

Stability under different settings of data distribution.

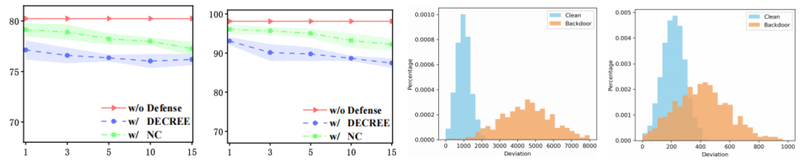
Dataset	Setting	Clean ACC	Clean ASR	Backdoored ACC	Backdoored ASR
CIFAR-10	$\alpha = 10$	77.23	10.70	77.36	99.41
	$\alpha = 0.1$	68.95	4.79	67.46	94.77
STL-10	$\alpha = 10$	84.81	11.91	84.92	98.96
	$\alpha = 0.1$	77.07	6.36	76.81	93.35
CIFAR-100	$\alpha = 10$	80.41	15.72	80.71	97.92
	$\alpha = 0.1$	73.73	6.73	73.24	90.91
ImageNet-100	$\alpha = 10$	79.61	11.72	78.75	93.79
	$\alpha = 0.1$	72.14	5.28	73.16	88.32

Stability after the removal of malicious clients.

Evaluation Round		HPE	EmInspector
Stopping round	ASR	99.41	95.75
	(%)	100.00	100.00
50 rounds later	ASR	98.32	82.92
	(%)	98.90	86.60
100 rounds later	ASR	96.13	58.72
	(%)	96.70	61.33

Robustness

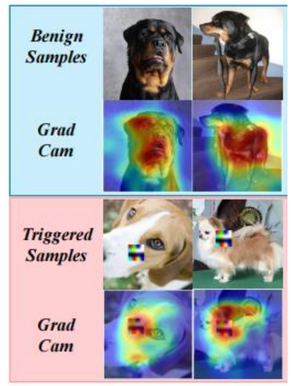
Resistance to DECREE, Neural Cleanse and Beatrix.



(a) ACC w/ or w/o Defense. (b) ASR w/ or w/o Defense. (c) BadEncoder Results on STL-10. (d) HPE Results on STL-10.

Resistance to FL Defense Methods.

Dataset	Defense Method	HPE ACC	HPE ASR	HPE+MR ACC	HPE+MR ASR
CIFAR-10	Krum	78.13	95.29	77.54	95.89
	FLTrust	77.25	98.29	75.54	97.44
	Foolsgold	75.42	92.28	73.45	93.66
	FLAME	76.73	96.81	75.39	97.15
	FLARE	77.08	96.68	75.98	97.23
	EmInspector	75.16	94.37	74.71	96.43
	AlignIns	84.49	96.27	82.48	97.31
STL-10	Krum	75.51	93.57	72.38	95.73
	FLTrust	75.18	93.94	72.38	92.73
	Foolsgold	74.87	95.57	71.65	96.34
	FLAME	75.65	93.82	73.04	95.28
	FLARE	74.46	94.71	72.54	95.08
	EmInspector	74.13	92.16	71.41	93.22
	AlignIns	76.03	96.06	75.31	96.12



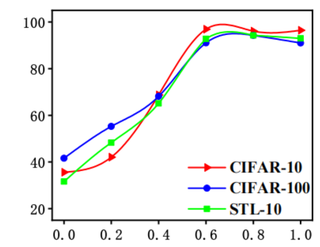
Ablation Studies

λ	1	0.8	0.6	0.4	0.2
CIFAR-10	97.83	98.61	98.70	97.58	97.21
STL-10	97.23	97.96	97.12	96.88	96.51

L	10	20	40	60
CIFAR-10	98.04	98.61	98.12	97.68
STL-10	97.12	97.37	97.08	96.98

	ASR	ACC	ASR-100
w/o HE	89.45	76.28	78.33
w/o BFE	86.12	76.45	75.91
w/o Dual-level	94.23	76.18	62.45
Full HPE	99.41	76.36	96.70

ϵ	0.5	1.0	2.0	3.0	5.0
ACC	83.45	82.29	81.18	81.17	81.13
ASR	71.23	91.17	98.12	98.04	97.19



(b) ASR with different μ.