RIP: Robust Collaborative Inference via Participant-wise Anomaly Detection

조윤기[1], 한우림[1], 유미선[1], 백윤흥[1]
[1]서울대학교 전기정보공학부, 반도체공동연구소
ygcho@sor.snu.ac.kr, wrhan@sor.snu.ac.kr, msyu@sor.snu.ac.kr, sbyun@sor.snu.ac.kr, ypaek@snu.ac.kr

RIP: Robust Collaborative Inference via Participant-wise Anomaly Detection

Yun-Gi Cho[1], Woo-Rim Han[1], Mi-Seon Yu[1], Yun-Heung Paek[1]
[1]Department of ECE and ISRC, SNU

요 약

Collaborative inference combines diverse features contributed by various agents to improve prediction accuracy. However, it is vulnerable to adversarial attacks, where attackers manipulate the model's predictions through non-consensual inputs. Since each participant operates within their unique feature space, defending against these attacks becomes particularly challenging. A recent study demonstrated that using an auto-encoder based on the underlying manifold can reduce the impact of malicious participants. However, our experiments observed that the recently proposed attack, in which malicious influences close to the manifold, may still pose a threat. To address this issue, we introduce a novel approach that leverages implicit redundancy across participants' feature spaces during the inference stage via participant-wise anomaly detection. We evaluate this approach on CIFAR10, CINIC10, Imagenette, Give-Me-Some-Credit, and Bank Marketing datasets. Extensive experiments and ablation studies show that RIP effectively mitigates adversarial attacks in the collaborative inference stage.

1. 서론

Collaborative inference is a privacy-preserving inference approach that enables multiple parties to collaboratively make predictions without sharing their raw data. In this setting, participants exchange intermediate computations, such as local embeddings, instead of raw data. This approach preserves data privacy while allowing the model to leverage distributed information across multiple sources.

Collaborative inference differs from traditional inference by its decentralized nature, where participants only share processed value, such as embeddings and intermediate features, which are less sensitive than the raw data itself. For example, in a healthcare scenario, different hospitals can collaborate on diagnosing diseases without exposing sensitive patient information. Each hospital processes its patient data locally, such as disease probabilities or risk factors, and shares only the embeddings for collective inference.

A key challenge in collaborative inference is ensuring robustness against malicious participants who might attempt to manipulate the inference process. In this context, The attacker may send manipulated local computations, such as altered embeddings or predictions, to disrupt the inference process and cause incorrect or biased outputs.

To address these challenges, various robust inference techniques have been developed. Liu et al.[1] proposed CoPur, a feature recovery method based on the participants' features by assuming that the overall participants' features lie on an underlying manifold. We obeserved that the limitations of this approach are that malicious influences close to the manifold potentially remain,

in our experment.

To further enhance robustness, we propose utilizing the participant-wise anomaly detection technique[2] for adversarial attack defense. A recent study have introduced a participant-wise anomaly detection for backdoor defense in VFL(Verfical Federated Learning)[2]. Since VFL involves collaborative inference during the inference phase and the defense mechanism aims to identify malicious participants sending adversarial inputs, this approach is also well-suited for addressing our adversarial attack problem. We conduct various attack scenarios on CIFAR10, CINIC10, Imagenette, Give-Me-Some-Credit, and Bank Marketing datasets. Extensive experiments and ablation studies demonstrate that RIM effectively mitigates adversarial attacks in collaborative inference.

2. Collaborative Inference

Based on a previous collaborative inference setup[1], we suppose that there are N participants and a server, with the collaborative goal of performing inference on a sample using the trained model. According to the feature-partitioned environment, a joint data sample can be expressed as $x=[x_1,\cdots,x_n]$. The i-th participant holds a vertically partitioned dataset, denoted as $\mathcal{D}_i=\{x_i^k\}_{k=1}^K$. The i-th participant's bottom model maps local input x_i to the local feature embedding h_i. For simplicity, the parameters of the bottom models are denoted as θ_{bottoms}. The server owns the top model parameterized as , and denoted as θ_{top}. The collaborative inference is computed by $T([B_1(x_1^k),\cdots,B_N(x_N^k)])$.

3. Adversarial Attacks in Collaborative Inference

These attacks can be categorized as targeted and untargeted attacks based on their objectives. Targeted attacks aim to modify the model's prediction as the attacker's desired label, while untargeted attacks aim to change the model's prediction to any incorrect label.

Gu et al.[3] introduced LR-BA, which employs a label inference module to produce an adversarial embedding even in situations where access to the training set's labels and the top model is limited. After completing the VSL training, LR-BA trains the label inference module proposed by Fu et al.\cite{fu2022label}, using the auxiliary dataset. The label inference module is trained to infer the label of the local embedding. LR-BA then optimizes the adversarial embedding to guide the label inference module toward predicting the target label, with its initial value being the average of the target class's local embeddings with high confidence from the label inference module in the training set. From the server's perspective, LR-BA resembles a targeted adversarial attack.

Liu et al.[1] proposed the distributed feature-flipping attack, which serves as an untargeted adversarial attack. The distributed feature-flipping attack inverts the sign of the attacker's local embeddings and increases their magnitude. In particular, when the n-th participant is the attacker, the malicious local embedding is represented as $h_n^{mal}=-(1+Amplification)\times h_n$,where h_n^{mal} and h_n is the attacker's malicious and benign local embedding, respectively. Amplification is the hyperparameter for controlling the magnitude of malicious local embeddings.

To mitigate malicious influences in situations where the feature spaces are different, Liu et al.[1] propose CoPur, which attempts to recover an uncorrupted combined local embedding on the underlying manifold that is near the original local embeddings, under the assumption that participants' features lie on an underlying manifold and there is an auto-encoder that learns the uncorrupted underlying manifold. In this process, the combined embeddings are decomposed into those that map well onto the manifold and those that do not. The decomposed components that map onto the manifold and near the original embeddings are used as the recovered local embeddings. Specifically, CoPur decomposes the

combined local embeddings h into the recovered components l and the corrupted components e, which implies $h = l + e$. CoPur's participant-wise optimization to achieve their purpose using an auto-encoder, denoted as AE, is the following:

$$\underset{l}{\operatorname{argmin}} \sum_{i=0}^{N-1} ||m_i \odot (h-l)||_2 + \tau \cdot ||m_i \odot (l - \mathrm{AE}(l)||_2$$

Here, m_i represents a masking where only the part corresponding to the i-th participant is filled with 1, while the rest is filled with 0. The loss is calculated only for the masked segments of the selected participant. The optimization is conducted in two steps. First, CoPur optimizes the first term, and then it jointly optimizes both terms. Ensuring that the recovered local embeddings are on the prior underlying manifold can reduce the influence of a few conflicting attackers, but it is difficult to completely eliminate malicious influence that is close to the underlying manifold.

Table 1: Evaluation for the targeted attacks by a single attacker, with and without label knowledge.

Dataset	Label Knowledge	Attack	Defense											
			Accuracy ↑ (Higher is better)				Attack Success Rate ↓ (Lower is better)				Robust Accuracy ↑ (Higher is better)			
			NO DEF	DP-SGD	CoPur	RIP (Ours)	NO DEF	DP-SGD	CoPur	RIP (Ours)	NO DEF	DP-SGD	CoPur	RIP (Ours)
CIFAR10	-	NO ATK	**78.37**	78.36	78.23	76.42	-	-	-	-	-	-	-	-
	Passive	LR-BA	**77.85**	77.45	77.51	75.96	82.39	96.73	88.41	**2.87**	24.13	12.91	19.98	**67.97**
	Active	LR-BA	77.35	**77.90**	77.33	75.71	81.01	99.35	90.99	**2.76**	22.85	10.59	17.35	**67.97**
CINIC10	-	NO ATK	**64.99**	64.50	64.88	63.48	-	-	-	-	-	-	-	-
	Passive	LR-BA	**64.79**	62.45	64.61	63.84	75.20	81.95	85.52	**3.06**	25.45	15.50	20.39	**56.86**
	Active	LR-BA	**64.96**	64.16	64.58	63.58	78.09	100.00	85.52	**3.01**	24.32	10.00	20.39	**57.71**
Imagenette	-	NO ATK	75.61	**76.60**	75.03	71.82	-	-	-	-	-	-	-	-
	Passive	LR-BA	74.17	**75.82**	74.07	72.18	99.56	100.00	99.54	**0.26**	10.27	9.86	10.27	**65.02**
	Active	LR-BA	73.99	**74.69**	73.93	71.58	100.00	99.91	100.00	**2.09**	9.86	9.93	9.86	**62.69**
GM	-	NO ATK	**78.68**	78.34	77.95	78.38	-	-	-	-	-	-	-	-
	Passive	LR-BA	**78.57**	78.28	77.94	78.23	98.65	99.89	54.35	**19.61**	50.72	50.12	69.71	**76.49**
	Active	LR-BA	**78.62**	78.26	77.86	78.27	99.97	100.00	67.00	**20.19**	50.09	50.07	65.10	**76.37**
BM	-	NO ATK	93.75	93.53	**93.77**	93.10	-	-	-	-	-	-	-	-
	Passive	LR-BA	**93.75**	93.59	93.46	92.12	77.16	99.33	48.06	**8.89**	61.35	50.25	75.92	**87.83**
	Active	LR-BA	**93.71**	93.66	93.69	92.15	91.83	74.28	73.61	**9.85**	54.11	62.79	63.25	**87.48**

Table 2: Evaluation for the untargeted attack by a single attacker.

Dataset	Defense							
	Accuracy ↑ (Higher is better)				Robust Accuracy ↑ (Higher is better)			
	NO DEF	DP-SGD	CoPur	RIP(Ours)	NO DEF	DP-SGD	CoPur	RIP(Ours)
CIFAR10	**78.37**	77.74	78.23	76.42	6.94	3.87	56.75	**68.59**
CINIC10	**64.99**	64.06	64.88	63.48	5.84	9.70	**57.49**	55.83
Imagenette	**75.61**	75.03	75.03	71.82	9.83	13.65	19.53	**62.61**
GM	**78.68**	78.29	77.95	78.38	51.00	49.93	64.01	**76.54**
BM	93.75	93.55	**93.77**	93.10	38.98	38.98	51.63	**89.02**

4. Method

Threat model. In this section, we describe the threat model based on previous studies[1,3]. We consider two distinct types of attackers, each with specific objectives. 1) Targeted Attacker: This attacker's primary goal is to manipulate the final prediction to a specific target label. 2) Untargeted Attacker: Unlike the targeted attacker, the untargeted attacker aims to disrupt the models' correct prediction. Their objective is not to steer the prediction toward a specific label but rather to introduce chaos or uncertainty. This setting can be extended to the multiple attackers scenario. We assume 2F < N where F means the number of attackers and N is the total number of participants. We also assume that the attacker cannot collude with the server.

Robust Collaborative Inference via Participant-wise Anomaly Detection. Basically, RIM uses a participant-wise anomaly detection via MAE and recovery methods following a study in VFL[2]. Similar to a previous study[1], the server can obtain additional modules based on training data prior to the inference phase. Unlike CoPur, RIM pre-trains a Masked Auto-Encoder (MAE) in advance. During the inference phase, RIM uses the MAE to predict one participant's output based on the others and considers the resulting error as an anomaly score, conducting participant-wise anomaly detection. If an input is deemed malicious, it is removed, and the output generated by the MAE is used as the input for the top model.

5. Experiments

In this section, we demonstrate the effectiveness of RIM through various experiments. we evaluate the scenario with a single attacker in

a 4-party scenario and multiple attackers in an 8-party scenario.

Table 3: Evaluation for the targeted attacks by multiple attackers with label knowledge.

Dataset	Attack	Defense											
		Accuracy ↑ (Higher is better)				Attack Success Rate ↓ (Lower is better)				Robust Accuracy ↑ (Higher is better)			
		NO DEF	DP-SGD	CoPur	RIM(Ours)	NO DEF	DP-SGD	CoPur	RIM(Ours)	NO DEF	DP-SGD	CoPur	RIM(Ours)
CIFAR10	NO ATK	**74.93**	73.71	74.35	71.92	-	-	-	-	-	-	-	-
	LR-BA	74.67	74.27	**74.69**	72.67	99.96	100.0	99.96	**6.08**	10.03	10.00	10.02	**54.22**
CINIC10	NO ATK	62.68	62.35	**62.85**	60.69	-	-	-	-	-	-	-	-
	LR-BA	**62.56**	61.46	62.51	60.95	99.56	100.0	99.49	**4.49**	10.31	10.00	10.37	**43.34**
Imagenette	NO ATK	**73.88**	71.16	72.42	69.44	-	-	-	-	-	-	-	-
	LR-BA	70.38	70.56	**71.35**	68.24	95.56	99.88	95.48	**1.77**	10.09	9.96	10.85	**50.25**

Table 4: Evaluation for the untargeted attack by multiple attackers.

Dataset	Defense							
	Accuracy ↑ (Higher is better)				Robust Accuracy ↑ (Higher is better)			
	NO DEF	DP-SGD	CoPur	RIM(Ours)	NO DEF	DP-SGD	CoPur	RIM(Ours)
CIFAR10	**74.93**	73.71	74.35	70.59	1.20	4.18	5.32	**54.77**
CINIC10	62.68	62.35	**62.85**	59.52	1.81	9.83	9.95	**42.49**
Imagenette	**73.88**	71.16	72.42	66.61	2.25	7.53	8.76	**48.37**

Tables 1 and 2 present the results for single attacker scenarios under targeted and untargeted attacks, respectively. Overall, compared to the no-defense scenario, robust accuracy increases by an average factor of 18. Tables 3 and 4 evaluate the defense capability in more challenging scenarios involving multiple attackers.

In the case of CoPur, it shows vulnerability to LR-BA in most scenarios. This is because LR-BA can optimize embeddings corresponding to the targeted class that may exist within the manifold. In contrast, RIM demonstrates robustness against all types of attacks.

6. Conclusion

In this study, we explored defense techniques against adversarial attacks in collaborative inference scenarios. Traditional methods often exhibit vulnerabilities to specific types of attacks. To address this, we proposed RIM, which adapts methods used in backdoor defense for adversarial defense. Through various experiments, we demonstrated the effectiveness of our approach.

References

[1] Liu, Jing, et al. "CoPur: certifiably robust collaborative inference via feature purification." Advances in Neural Information Processing Systems 35 (2022): 26645-26657.

[2] Cho, Yungi, et al. "VFLIP: A Backdoor Defense for Vertical Federated Learning via Identification and Purification." European Symposium on Research in Computer Security. Cham: Springer Nature Switzerland, 2024.

[3] Gu, Yuhao, and Yuebin Bai. "LR-BA: Backdoor attack against vertical federated learning using local latent representations." Computers & Security 129 (2023): 103193.

ACKNOWLEDGEMENT

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00277326). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence semiconductor support program to nurture the best talents (IITP-2023-RS-2023-00256081) grant funded by the Korea government(MSIT). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00516, Derivation of a Differential Privacy Concept Applicable to National Statistics Data While Guaranteeing the Utility of Statistical Analysis). This work was supported by Inter-University Semiconductor Research Center (ISRC).