Defense without Forgetting: Continual Adversarial Defense with Anisotropic Isotropic Pseudo Replay

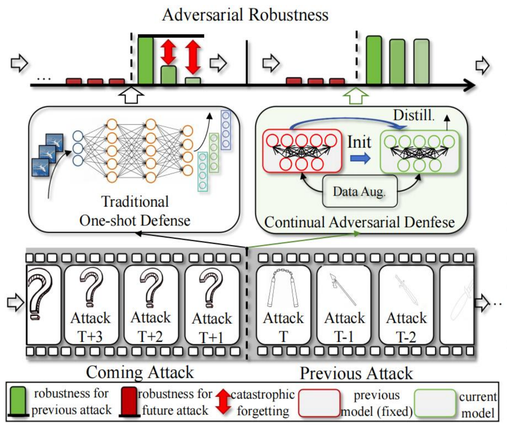
Yuhang Zhou (Harbin Institute of Technology, Shenzhen), Yunzhong Hua✉ (Harbin Institute of Technology, Shenzhen)

Abstract

Adversarial defense only focus on **one-shot** setting against adversarial attack, and it is crucial for a defense model to **constantly adapt to new attacks**. However, the adaptation process can lead to **catastrophic forgetting** of previously attacks.This paper firstly verify the catastrophic forgetting in life-long defense task and discuss for the first time the concept of continual adversarial defense under a sequence of attacks, and propose a lifelong defense baseline called **Anisotropic & Isotropic Replay (AIR)**.

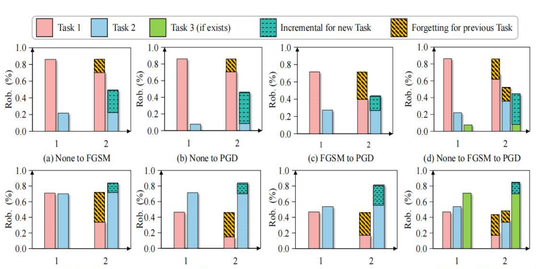
Problems in Adversarial Defense

The model diagram on the left presents the **one-shot defense** studies an isolated Min-Max process and implicitly assumes the potential attack is static. For a **continual attack sequence**, the indispensable adaptation process introduces additional challenge of catastrophic forgetting of previous attacks. Therefore, a deployable adversarial defense should be a life-one learning task rather than a one-shot task. We propose a self-distillation pseudo-replay baseline to alleviate the catastrophic forgetting against attack sequence, indicated by the model diagram on the right.



The difference between the one-shot defense and the continual defense. Simply put, existing defenses overlook the threat of t he continuous attack sequences

Verification of the Catastrophic Forgetting

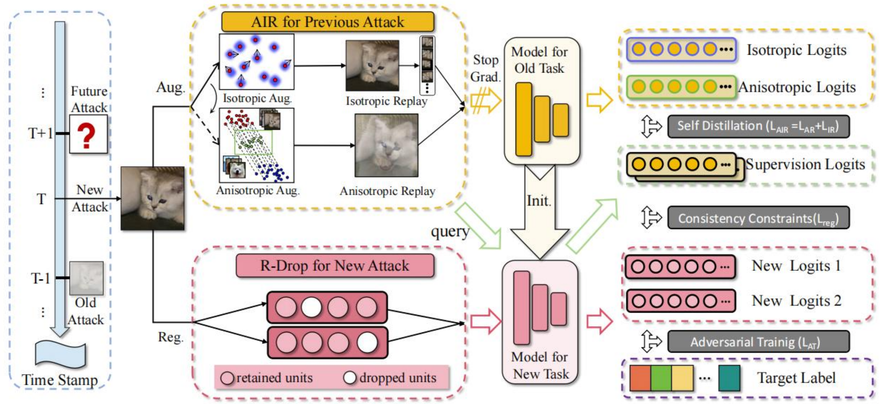


The horizontal axis can be considered as a timestamp, where '1' represents the model adapting to TASK 1, and time '2' represents the sequential adaptation to all attack tasks in the sequence.

Anisotropic Isotropic Pseudo Replay (AIR)

Overview

AIR mainly includes **Isotropic Pseudo Replay**, **Anisotropic Pseudo Replay** and a **Regularizer**. The upper module (in yellow block) consists of the anisotropic replay module and isotropic replay module, aiming to maintain the memory of old tasks. The lower module (in red block) is the vanilla adversarial training with R-Drop for new attacks. The three main loss are highlighted in the gray circular box.



$$\mathcal{L} = \mathcal{L}_{vanilla} + \boxed{\mathcal{L}_{dis}}, \quad \mathcal{L}_{dis} = D(f_{w_{t-1}}(X'_t), f_{w_t}(X'_t)),$$

(1) Isotropic Pseudo Replay (IR)

$$X_t^{IR} = \mathcal{T}(X_t + \lambda \cdot r),$$

$$\mathcal{L}_{IR} = KL\_Div(f_{w_t}(X_t^{IR}), f_{w_{t-1}}(X_t^{IR}))$$

(2) Anisotropic Pseudo Replay (AR)

$$X_t^{AR} = \alpha \cdot X_t + (1-\alpha) \cdot x_t^{shuffle}$$

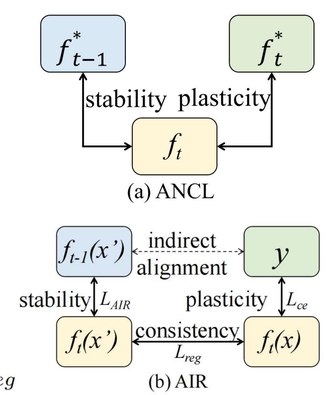
$$y_{dis} = \alpha \cdot f_{w_{t-1}}(X_t) + (1-\alpha) \cdot f_{w_{t-1}}(X_t^{shuffle}).$$

$$\mathcal{L}_{AR} = KL\_Div(f_{w_{t-1}}(X_{AR}), f_{w_t}(X_{AR})).$$

(2) Regularization

$$\mathcal{L}_{reg} = \frac{1}{2}(KL(f^1_{w_t}(x_t)||f^1_{w_t}(x'_t)) + KL(f^2_{w_t}(x_t)||f^2_{w_t}(x'_t))),$$

On the one hand, a model with R-Drop will learn consistent outputs from different local features without overfitting on specific features. On the other hand, compare to the exsting ANCL, this constraint achieves an indirect chain alignment. The intuitive expression of the above comparison is shown in the right figure.



Total Loss

$$\mathcal{L}_{AIR} = \mathcal{L}_{AT} + \lambda_{SD} \cdot (\mathcal{L}_{IR} + \mathcal{L}_{AR}) + \lambda_{Reg} \cdot \mathcal{L}_{Reg}$$

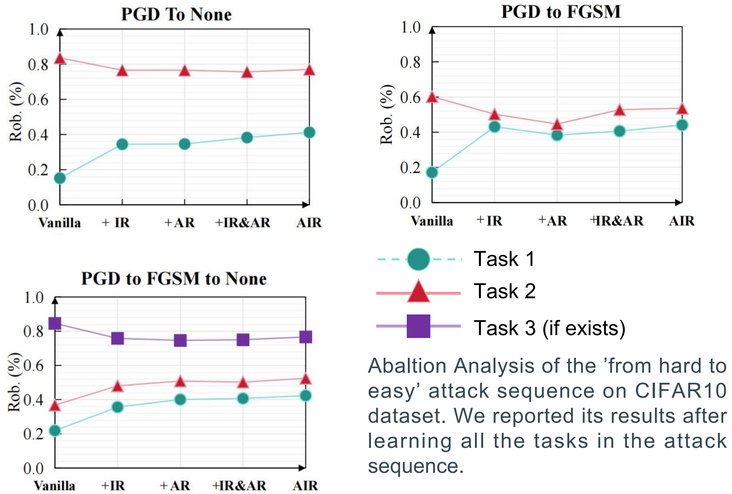
Experiments

Adaptation between two attacks for different defense methods

Transfer between two attacks

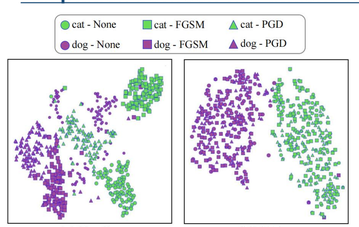
Datasets	Tasks	None to FGSM Task 1	None to FGSM Task 2	FGSM to None Task 1	FGSM to None Task 2	None to PGD Task 1	None to PGD Task 2	PGD to None Task 1	PGD to None Task 2	FGSM to PGD Task 1	FGSM to PGD Task 2	PGD to FGSM Task 1	PGD to FGSM Task 2
MNIST	Vanilla AT [27]	95.18	98.55	83.97	98.86	94.22	90.01	3.72	98.59	96.48	94.71	2.56	96.96
	EWC [21]	98.83	96.63	98.18	97.85	97.35	87.32	91.97	98.85	95.26	**95.86**	94.77	96.90
	Feat. Extraction [24]	98.16	89.23	97.46	98.80	12.72	11.35	95.23	98.80	96.94	73.61	95.23	97.93
	LFL [19]	98.85	97.02	90.54	98.80	97.32	87.52	33.84	98.71	95.84	91.87	25.05	**98.40**
	AIR (ours)	**99.37**	**98.84**	**98.18**	**98.84**	**98.89**	**94.26**	**95.93**	**99.06**	**97.45**	95.67	**96.25**	97.93
	Joint Training [24]	99.11	98.52	98.52	99.11	99.35	95.44	95.44	99.35	96.72	94.29	94.29	96.72
CIFAR10	Vanilla AT [27]	70.60	49.30	34.90	83.83	71.09	45.52	15.19	83.59	34.90	35.21	17.14	60.24
	EWC [21]	72.66	49.17	**43.85**	82.62	69.38	41.46	30.25	61.70	48.63	40.53	24.44	45.18
	Feat. Extraction [24]	67.69	35.11	45.27	82.13	40.04	30.90	**45.54**	75.02	52.85	24.88	42.51	44.54
	LFL [19]	74.23	50.17	42.77	78.59	67.31	42.76	28.27	**80.59**	51.98	43.30	24.18	46.71
	AIR (ours)	**76.73**	**51.48**	42.32	**82.85**	**75.53**	**45.14**	41.21	77.02	**53.39**	**44.12**	**43.00**	**52.26**
	Joint Training [24]	86.10	47.65	57.65	86.10	72.58	44.86	44.86	72.58	49.81	42.56	42.56	49.81
CIFAR100	Vanilla	42.27	20.67	25.98	50.26	40.58	17.31	20.21	47.47	24.08	19.03	20.89	30.47
	EWC [21]	50.04	22.43	**29.13**	45.12	**48.45**	16.61	19.21	44.66	22.98	18.00	20.16	24.32
	Feat. Extraction [24]	37.02	8.35	23.62	47.68	11.46	4.96	20.70	41.42	23.63	18.22	19.54	24.08
	LFL [19]	28.61	15.30	37.48	49.06	19.19	13.36	20.08	43.62	25.49	15.77	19.19	23.85
	AIR (ours)	**50.77**	**24.32**	27.47	**50.67**	47.88	**21.41**	**22.05**	**45.61**	**27.59**	**23.19**	**23.40**	**27.51**
	Joint Training [24]	56.44	35.88	35.88	56.44	46.01	22.54	22.54	46.01	35.27	21.45	21.45	35.27

Ablation study of the modules in AIR



Abaltion Analysis of the 'from hard to easy' attack sequence on CIFAR10 dataset. We reported its results after learning all the tasks in the attack sequence.

Adaptation between two attacks for different defense methods



T-SNE diagram of features encoded by vanilla AT and AIR on CIFAR10. The proposed AIR is able to encode all attacks in the sequence of the same category into one shared cluster.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62071142, the Shenzhen Science and Technology Program under Grant ZDSYS20210623091809029, and by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005.