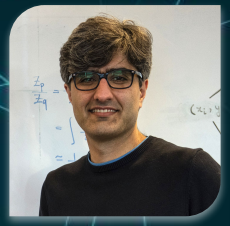
Transactions on Machine Learning Research

The Effectiveness of Random Forgetting for Robust Generalization


Vijaya Raghavan, Bahram Zonooz*, Elahe Arani*

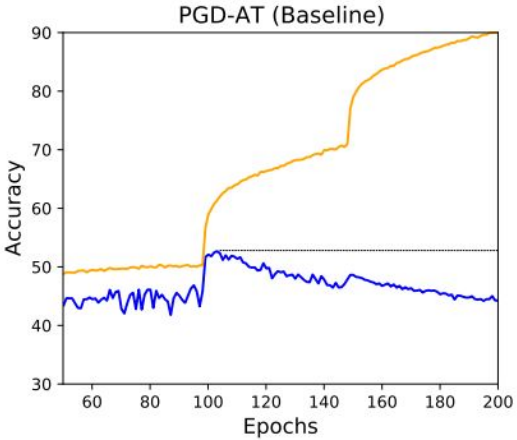
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

- DNNs excel in various fields but are vulnerable to adversarial attacks.
- Adversarial attacks introduce subtle perturbations leading to incorrect predictions.
- Threatens critical applications like autonomous vehicles and medical diagnosis.
- Adversarial Training (AT) enhances DNN robustness by training with adversarial examples.
- Adversarial Training is essential for ensuring the reliability and security of deep learning systems in critical applications.

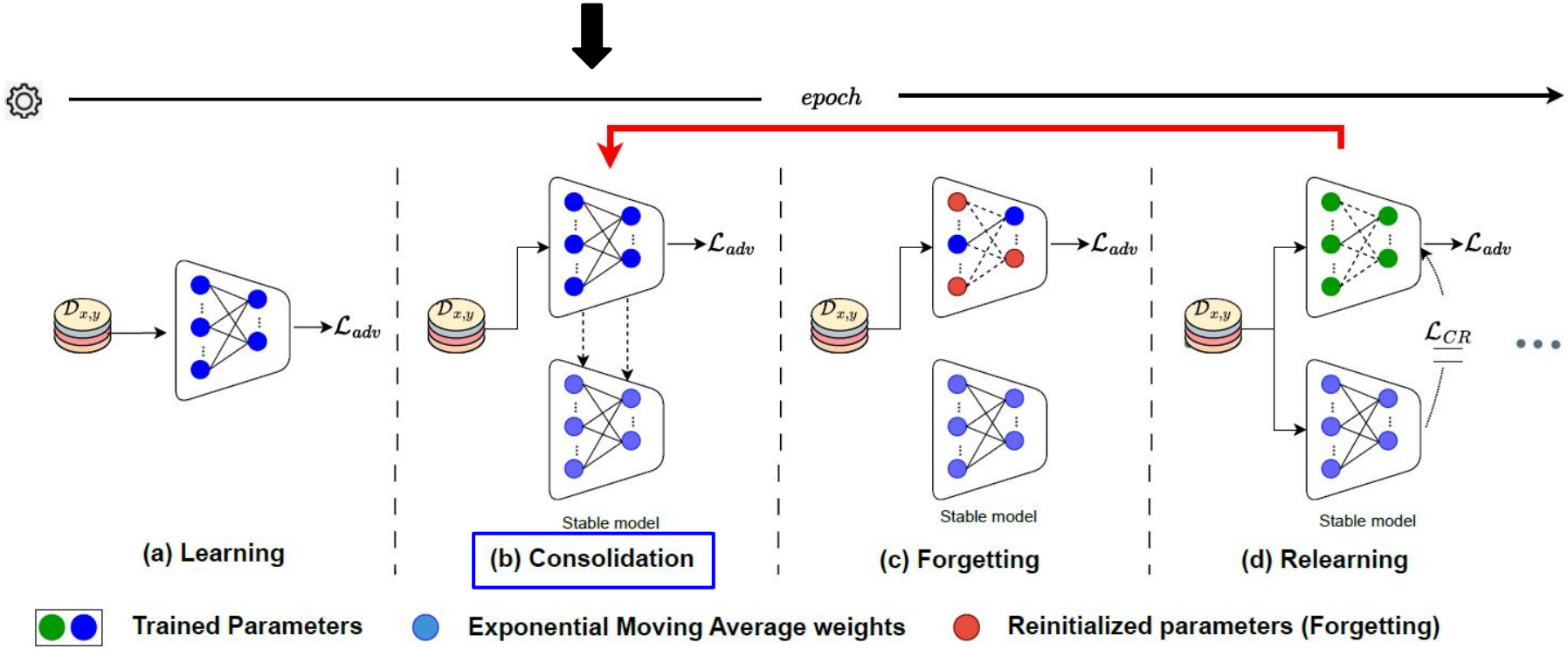
Problem Statement

Robust Overfitting: A Double-Edged Sword

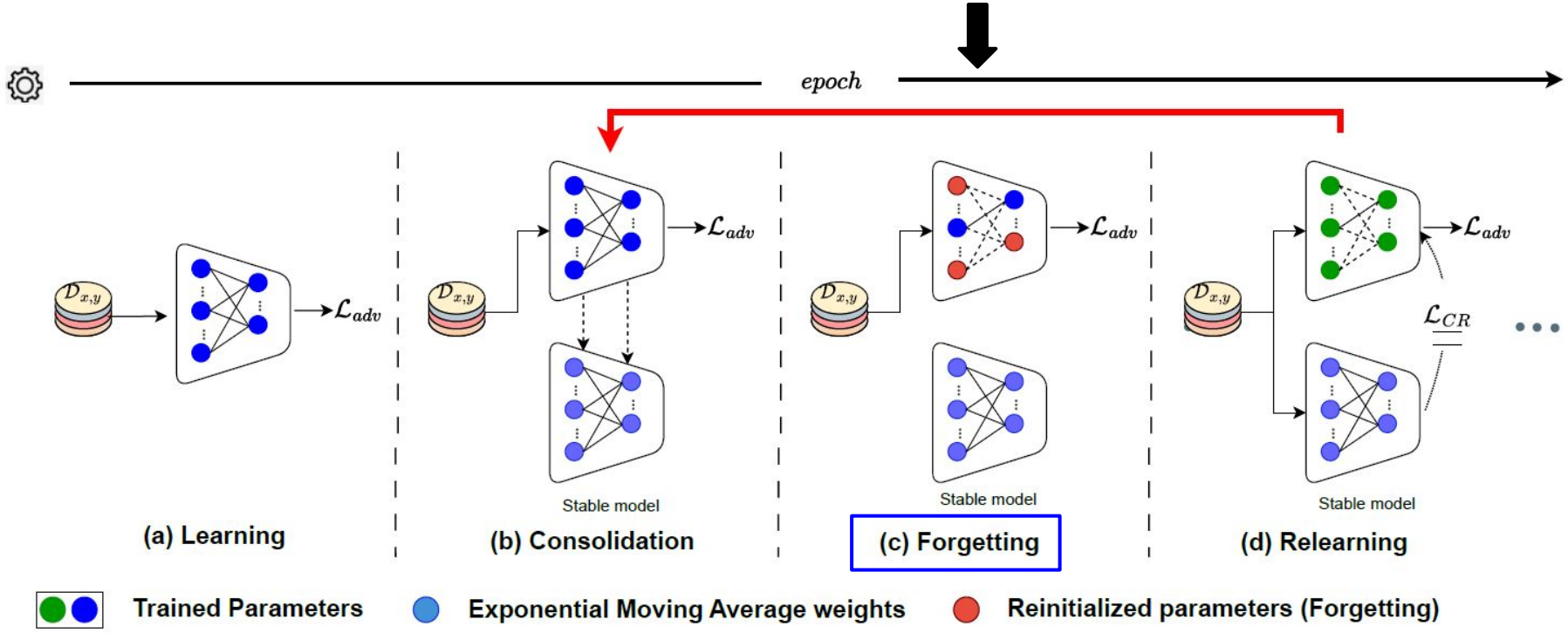
- "Robust overfitting" phenomenon reported in Adversarial Training (AT) by Rice et al. (2020).
- Illustration of the phenomenon: Adversarial test accuracy lags significantly behind adversarial train accuracy.
- Conventional methods such as data augmentation, early stopping to prevent benign overfitting ineffective in addressing robust overfitting in AT (Rice et al., 2020; Nakkiran et al., 2021).
- Existence of robust overfitting in AT highlights a significant gap in building robust machine learning systems.



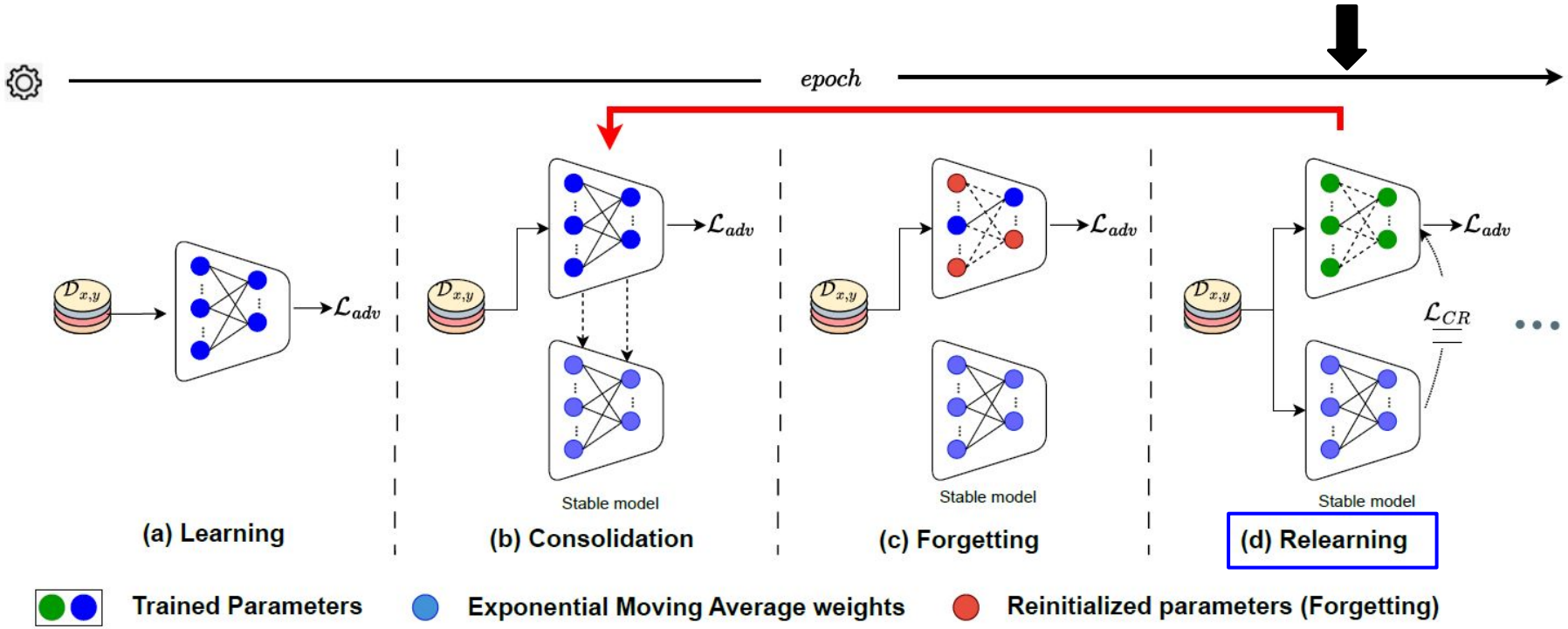
Methodology
The FOMO Approach



Methodology
The FOMO Approach



Methodology
The FOMO Approach



Results

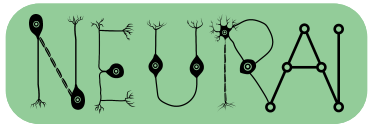


Table 1: Performance comparison on the CIFAR-10 using the PreActResNet-18 and WideResNet-34-10 architectures under a perturbation norm of $\epsilon_\infty = 8/255$.

Method	PreActResNet-18							WideResNet-34-10						
	Natural			PGD-20			Trade-off	Natural			PGD-20			Trade-off
	Best	Last	Δ	Best	Last	Δ		Best	Last	Δ	Best	Last	Δ	
PGD-AT	82.08	83.98	1.90	52.32	44.44	-7.88	58.12	86.90	86.38	-0.52	56.45	48.16	-8.29	61.84
TRADES	80.72	82.61	1.89	52.66	49.75	-2.91	62.10	84.73	84.62	-0.11	56.50	47.28	-9.22	60.66
KD+SWA	**83.82**	**84.43**	0.61	54.59	54.42	**-0.17**	66.18	86.85	**88.03**	1.18	56.92	55.74	-1.18	68.25
PGD-AT+TE	82.15	82.59	0.44	55.03	53.79	-1.24	65.14	86.20	85.63	-0.57	56.89	53.49	-3.4	65.84
AWP	81.25	81.56	**0.21**	55.39	54.73	-0.66	65.50	86.28	86.27	**-0.01**	58.85	58.76	**-0.09**	69.90
FOMO (Ours)	81.84	82.51	0.67	**56.68**	**56.46**	-0.22	**67.04**	**87.31**	87.08	-0.23	**59.69**	**59.23**	-0.46	**70.50**

Results

Table 2: Performance comparison on CIFAR-100 and SVHN datasets, using the PreActResNet18 architecture and a perturbation norm of $\epsilon_\infty = 8/255$.

Method	CIFAR-100							SVHN						
	Natural			PGD-20			Trade-off	Natural			PGD-20			Trade-off
	Best	Last	Δ	Best	Last	Δ		Best	Last	Δ	Best	Last	Δ	
PGD-AT	55.52	57.35	1.83	27.22	20.82	-6.4	30.54	87.93	89.90	-1.93	52.60	45.13	-7.47	60.09
TRADES	55.53	57.09	-1.56	29.56	26.08	-3.48	35.80	90.88	91.30	**0.42**	52.50	47.50	-5.00	62.48
KD+SWA	57.23	**57.66**	0.43	30.06	30.02	**-0.04**	39.48	90.40	91.70	1.30	53.65	50.65	-3.00	65.25
PGD-AT+TE	56.52	57.30	0.78	31.23	29.25	-0.98	38.72	90.09	90.91	-0.82	54.85	52.18	-2.67	66.30
AWP	53.92	54.81	-0.89	30.70	30.28	-0.42	39.00	93.85	92.59	-1.26	59.12	55.87	-3.25	69.68
FOMO	**57.45**	57.07	**-0.38**	**32.07**	**31.67**	-0.40	**40.73**	**94.17**	**93.66**	-0.51	**59.63**	**59.06**	**-0.57**	**72.44**

Table 3: White-box/Black-box (Auto-attack) performance comparison on CIFAR-10 and CIFAR-100, using the PreActResNet-18 architecture and a perturbation norm of $\epsilon_\infty = 8/255$.

	Method	CIFAR-10			CIFAR-100		
		Best	Last	Δ	Best	Last	Δ
ICLR'18	PGD-AT	47.72	42.60	-5.12	24.53	20.21	-4.32
ICML'19	TRADES	48.37	46.94	-1.43	24.51	22.86	-1.65
NeurIPS'20	AWP	50.34	49.64	-0.70	25.26	25.07	-0.19
ICLR'21	KD+SWA	49.87	49.74	-0.13	26.04	25.99	**-0.05**
ICLR'22	PGD-AT+TE	50.11	49.14	-0.97	26.04	25.13	-0.91
ICML'22	MLCAT_{WP}	50.70	50.32	-0.38	25.86	25.18	-0.68
ICLR'23	IDBH[Strong]	50.74	49.99	-0.75	-	-	-
ICLR'24	FOMO	**51.37**	**51.28**	**-0.09**	**27.57**	**27.49**	-0.08

Conclusion and Future Work

- FOMO (Forget to Mitigate Overfitting) is a novel adversarial training method inspired by the brain's active forgetting.

- It alternates between consolidation, forgetting (re-initializing weights) and relearning phases to focus on truly robust features.

- Key Results:
  - FOMO significantly reduces robust overfitting.
  - Improves both standard and robust accuracy across datasets and models.
  - Offers strong defense against Auto Attacks.
  - Enhances generalization, making it applicable to real-world scenarios.

For further information, please refer to the paper available at this link:https://openreview.net/pdf?id=MEGQGNUfPx

THANKS

Contact: Vijaya Raghavan
Email: raghavijay95@gmail.com
Website: https://github.com/NeurAI-Lab