Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning

Kaijie Zhu[1,2], **Xixu Hu**[3], **Jindong Wang**[4], **Xing Xie**[4], **Ge Yang**[1,2] *
[1]School of Artifical Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3] City University of Hong Kong [4] Microsoft Research
{zhukaijie2021, ge.yang}@ia.ac.cn

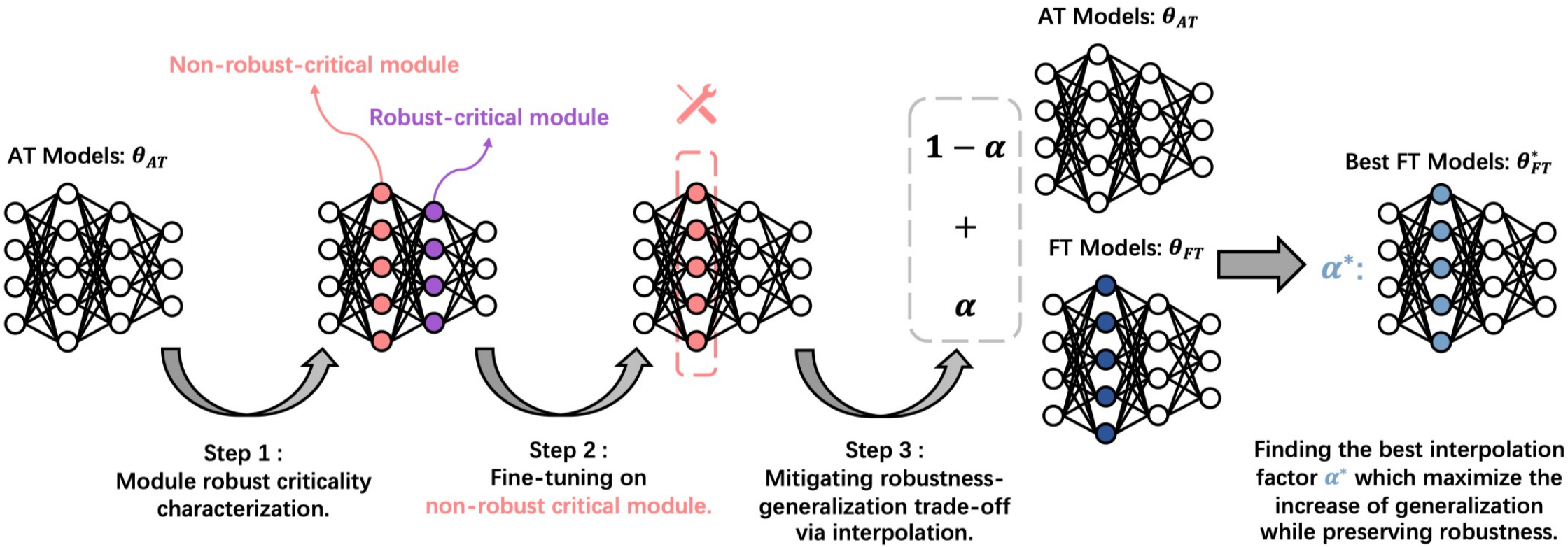
ICCV 2023

Motivation

- To improve adversarial robustness: adversarial training
- Trade-off between **adversarial robustness** and **generalization**

- To address such a trade-off: leveraging the **redundant capacity for robustness**
  - Do adversarially trained models have such redundant capacity?
  - How to leverage it to improve the generalization and OOD robustness while maintaining adversarial robustness?

Method


AT Models: θ_{AT}
Non-robust-critical module
Robust-critical module
Step 1 :
Module robust criticality characterization.
Step 2 :
Fine-tuning on non-robust critical module.
$1-\alpha$
$+$
α
Step 3 :
Mitigating robustness-generalization trade-off via interpolation.
AT Models: θ_{AT}
FT Models: θ_{FT}
Best FT Models: θ^*_{FT}
α^*:
Finding the best interpolation factor α^* which maximize the increase of generalization while preserving robustness.

- Adversarial training (AT):

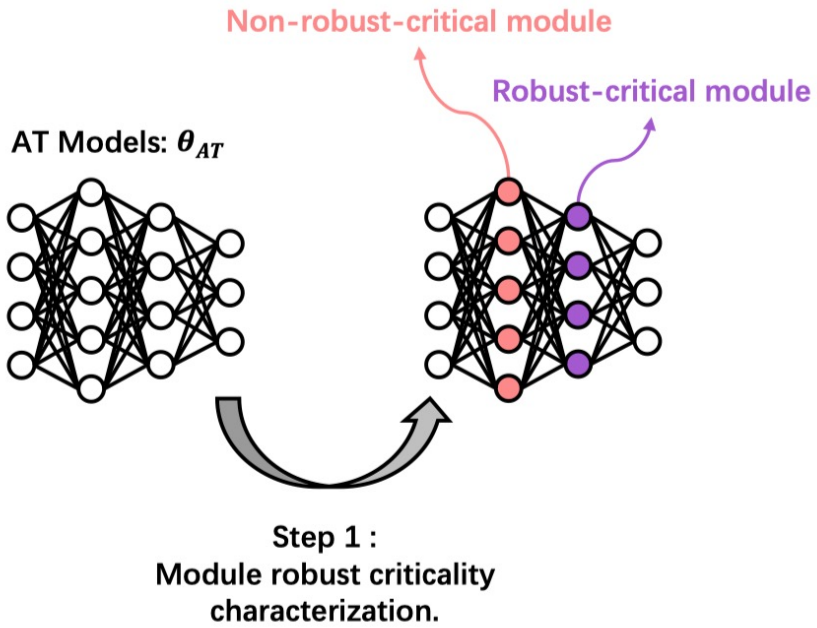
robust loss

$$\arg\min_{\boldsymbol{\theta}} \mathcal{R}(f(\boldsymbol{\theta}), \mathcal{D}), \text{ where}$$

$$\mathcal{R}(f(\boldsymbol{\theta}), \mathcal{D}) = \sum_{(\boldsymbol{x}, y) \in \mathcal{D}} \max_{\Delta \boldsymbol{x} \in \mathcal{S}} \mathcal{L}(f(\boldsymbol{\theta}, \boldsymbol{x} + \Delta \boldsymbol{x}), y).$$

worst-case input perturbation

- How to character the **redundant capacity for robustness**?



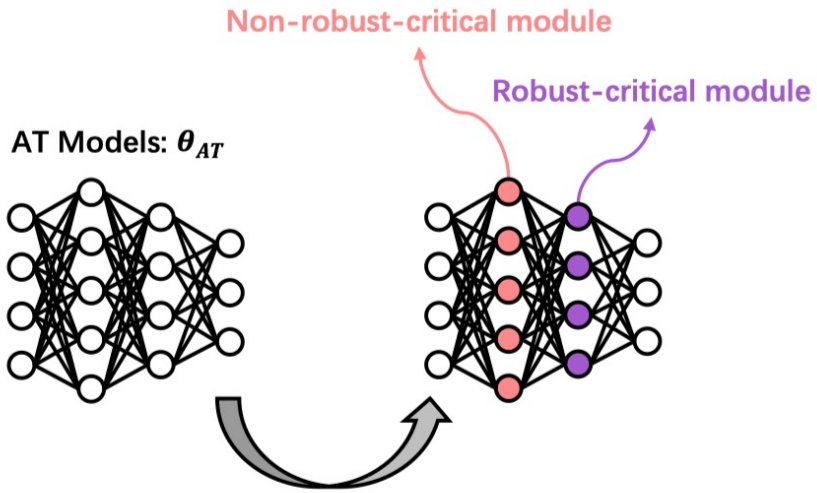
- For a module i

$$MRC(f, \boldsymbol{\theta}^{(i)}, \mathcal{D}, \epsilon) = \max_{\Delta\boldsymbol{\theta} \in \mathcal{C}_{\boldsymbol{\theta}}} \mathcal{R}(f(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}), \mathcal{D}) - \mathcal{R}(f(\boldsymbol{\theta}), \mathcal{D})$$

where

$$\Delta\boldsymbol{\theta} = \{\mathbf{0}, \ldots, \mathbf{0}, \Delta\boldsymbol{\theta}^{(i)}, \mathbf{0}, \ldots, \mathbf{0}\}$$

$$\mathcal{C}_{\boldsymbol{\theta}} = \{\Delta\boldsymbol{\theta} \,|\, \|\Delta\boldsymbol{\theta}\|_p \leq \epsilon \|\boldsymbol{\theta}^{(i)}\|_p\}$$

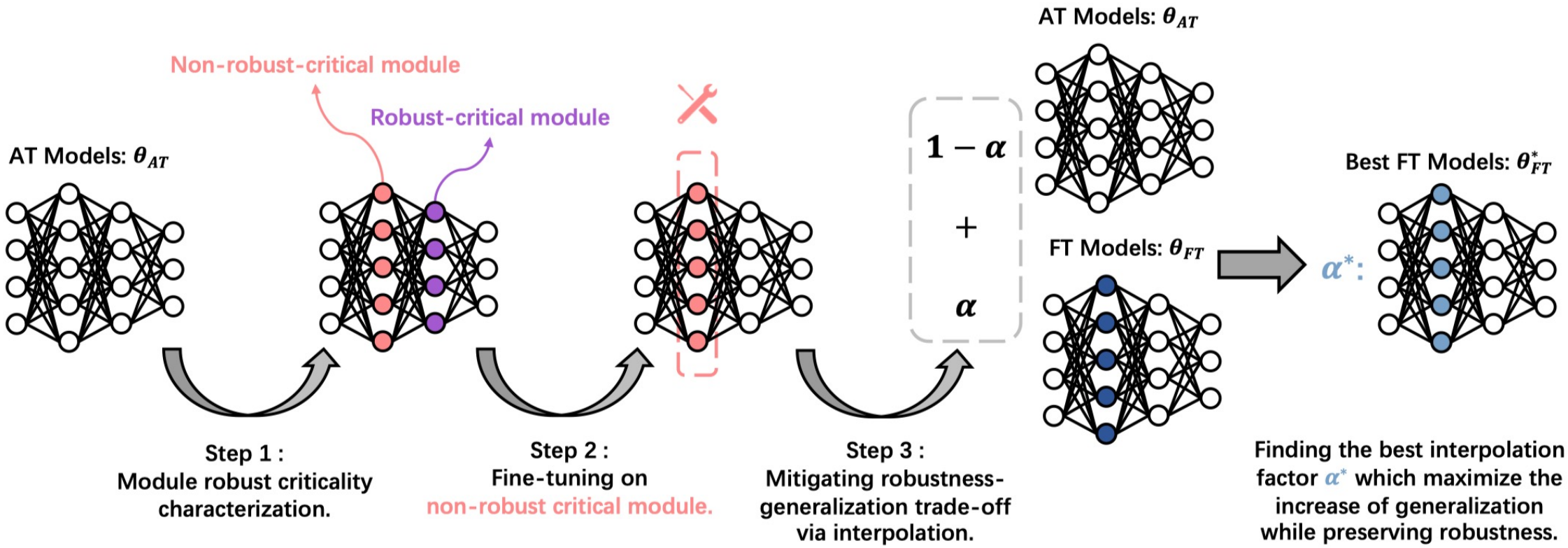


Step 1: Module robust criticality characterization

- Choose the module with lowest MRC value:

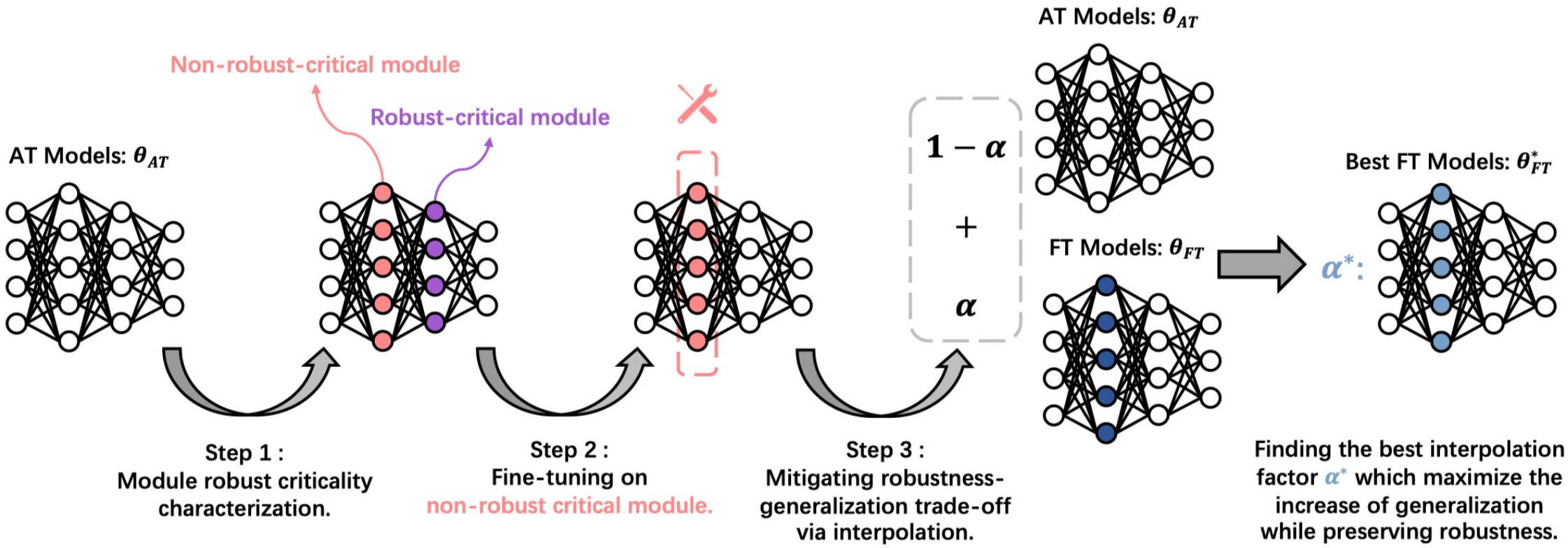
$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(i)} \text{ where } i = \arg\min_i MRC(f, \boldsymbol{\theta}^{(i)}, \mathcal{D}, \epsilon)$$

-> Non-robust-critical module



Step 2: Fine-tuning on non-robust-critical modules $\widetilde{\theta}$

$$\arg\min_{\tilde{\boldsymbol{\theta}}} \sum_{(\boldsymbol{x},y)\in\mathcal{D}} \mathcal{L}(f(x,(\tilde{\boldsymbol{\theta}};\boldsymbol{\theta}\setminus\tilde{\boldsymbol{\theta}})),y) + \lambda\|\tilde{\boldsymbol{\theta}}\|_2$$



Step 3: Mitigating robustness-generalization trade-off via interpolation

$$\boldsymbol{\theta}_\alpha = (1-\alpha)\boldsymbol{\theta}_{AT} + \alpha\boldsymbol{\theta}_{FT}$$

$\boldsymbol{\theta}^*_{FT} = \boldsymbol{\theta}_\alpha$ if it reaches best standard test acc while preserve the robustness as $\boldsymbol{\theta}_{AT}$.

Experiments

- Network Architecture
  - ResNet18
  - ResNet34
  - WRN34-10

- Dataset
  - CIFAR10
  - CIFAR100
  - Tiny-ImageNet

- Test metric
  - Std
  - OOD (Using Noise, Blur, ⋯ to obtain out-of-distribution images)
  - Adv

Architecture	Method	CIFAR10			CIFAR100			Tiny-ImageNet		
		Std	*OOD*	*Adv*	*Std*	*OOD*	*Adv*	*Std*	*OOD*	*Adv*
ResNet18	AT	81.46	73.56	53.63	57.10	46.43	30.15	49.10	27.68	23.28
	AT+RiFT	**83.44**	**75.69**	**53.65**	**58.74**	**48.06**	**30.17**	**50.61**	**28.73**	**23.34**
	Δ	+1.98	+2.13	+0.02	+1.64	+1.63	+0.02	+1.51	+1.05	+0.06
ResNet34	AT	84.23	75.37	55.31	58.67	48.24	30.50	50.96	27.91	24.27
	AT+RiFT	**85.41**	**77.15**	**55.34**	**60.88**	**49.97**	**30.58**	**52.54**	**30.07**	**24.37**
	Δ	+1.18	+1.78	+0.03	+2.21	+1.73	+0.08	+1.58	+2.16	+0.10
WRN34-10	AT	87.41	78.75	55.40	62.35	50.61	**31.66**	52.78	31.81	26.07
	AT+RiFT	**87.89**	**79.31**	**55.41**	**64.56**	**52.69**	31.64	**55.31**	**33.86**	**26.17**
	Δ	+0.48	+0.56	+0.01	+2.21	+2.08	-0.02	+2.53	+2.05	+0.10
Avg	Δ	+1.21	+1.49	+0.02	+2.02	+1.81	+0.02	+1.87	+1.75	+0.08

Method	*Std*	*OOD*	*Adv*
All layers	83.56	75.48	52.66
Last layer	83.35	75.16	52.75
Robust-critical	83.36	75.42	52.48
Non-robust-critical	83.44	75.69	**53.65**

Method	*Std*	*OOD*	*Adv*
Top 1	83.44	75.69	**53.65**
Top 2	83.41	75.61	52.47
Top 3	83.59	75.77	52.22
Top 5	83.70	75.82	52.35

Summary

- Leveraging the concept of module robust criticality (MRC) to guide the fine-tuning process, which leads to improved generalization and OOD robustness

+ A good way to select the module to fine-tune
+ Good performances on *Std*, *OOD*, and *Adv*

- Unknow effect of different network architectures like transformers
- Unknow effect of a single fine-tune without adversarial training

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. *Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.*

1️⃣ 提高鲁棒性的动机

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. ***Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability.*** *This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.*

2️⃣ 已有方法的问题

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. ***This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness.*** *The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.*

3️⃣ 一句话概括文章的贡献

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.

4️⃣ 四句话递进介绍文章的贡献：一句总述+三句细节

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.

5️⃣ 一句实验设置+一句实验效果

Writing

Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robustness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT.

6 代码介绍

Writing

The pursuit of accurate and trustworthy artificial intelligence systems is a fundamental objective in the deep learning community. Adversarial examples [45, 15], which perturbs input by a small, human imperceptible noise that can cause deep neural networks to make incorrect predictions, pose a significant threat to the security of AI systems. Notable experimental and theoretical progress has been made in defending against such adversarial examples [6, 4, 10, 19, 11, 16, 37]. Among various defense methods [52, 33, 57, 31, 8], adversarial training (AT) [29] has been shown to be one of the most promising approaches [4, 11] to enhance the adversarial robustness. However, compared to standard training, AT severely sacrifices generalization on in-distribution data [42, 46, 58, 36, 32] and is exceptionally vulnerable to certain out-of-distribution (OOD) exam ples [14, 53, 22] such as Contrast, Bright and Fog, resulting in unsatisfactory performance.

1️⃣ 介绍深度学习的准确性和可信性的目标

介绍鲁棒性方面的发展

Writing

The pursuit of accurate and trustworthy artificial intelligence systems is a fundamental objective in the deep learning community. Adversarial examples [45, 15], which perturbs input by a small, human imperceptible noise that can cause deep neural networks to make incorrect predictions, pose a significant threat to the security of AI systems. Notable experimental and theoretical progress has been made in defending against such adversarial examples [6, 4, 10, 19, 11, 16, 37]. Among various defense methods [52, 33, 57, 31, 8], adversarial training (AT) [29] has been shown to be one of the most promising approaches [4, 11] to enhance the adversarial robustness. However, compared to standard training, AT severely sacrifices generalization on in-distribution data [42, 46, 58, 36, 32] and is exceptionally vulnerable to certain out-of-distribution (OOD) exam ples [14, 53, 22] such as Contrast, Bright and Fog, resulting in unsatisfactory performance.

1️⃣ 介绍深度学习的准确性和可信性的目标

已有方法（AT）面临的问题

Writing

Prior studies tend to mitigate the trade-off between generalization and adversarial robustness within the adversarial training procedure. For example, some approaches have explored reweighting instances [59], using unlabeled data [36], or redefining the robust loss function [58, 48, 50, 32].

In this paper, we take a different perspective to address such a trade-off by leveraging the redundant capacity for robustness of neural networks after adversarial training. Recent research has demonstrated that deep neural networks can exhibit redundant capacity for generalization due to their complex and opaque nature, where specific network modules can be deleted, permuted [47], or reset to their initial values [55, 9] with only minor degradation in generalization performance. Hence, it is intuitive to ask: Do adversarially trained models have such redundant capacity? If so, how to leverage it to improve the generalization and OOD robustness while maintaining adversarial robustness?

2️⃣ 针对解决泛化性和鲁棒性的trade-off，已有工作做了哪些尝试

已有方法（AT）面临的问题

Writing

Prior studies tend to mitigate the trade-off between generalization and adversarial robustness within the adversarial training procedure. For example, some approaches have explored reweighting instances [59], using unlabeled data [36], or redefining the robust loss function [58, 48, 50, 32]. In this paper, we take a different perspective to address such a trade-off by leveraging the redundant capacity for robustness of neural networks after adversarial training. Recent research has demonstrated that deep neural networks can exhibit redundant capacity for generalization due to their complex and opaque nature, where specific network modules can be deleted, permuted [47], or reset to their initial values [55, 9] with only minor degradation in generalization performance. Hence, it is intuitive to ask: Do adversarially trained models have such redundant capacity? If so, how to leverage it to improve the generalization and OOD robustness while maintaining adversarial robustness?

2️⃣ 针对解决泛化性和鲁棒性的trade-off，已有工作做了哪些尝试

引出本文从模型冗余的鲁棒能力方面了解决这一trade-off

通过已有工作对idea进行支撑

Writing

Prior studies tend to mitigate the trade-off between generalization and adversarial robustness within the adversarial training procedure. For example, some approaches have explored reweighting instances [59], using unlabeled data [36], or redefining the robust loss function [58, 48, 50, 32]. In this paper, we take a different perspective to address such a trade-off by leveraging the redundant capacity for robustness of neural networks after adversarial training. Recent research has demonstrated that deep neural networks can exhibit redundant capacity for generalization due to their complex and opaque nature, where specific network modules can be deleted, permuted [47], or reset to their initial values [55, 9] with only minor degradation in generalization performance. Hence, it is intuitive to ask: Do adversarially trained models have such redundant capacity? If so, how to leverage it to improve the generalization and OOD robustness while maintaining adversarial robustness?

2️⃣ 针对解决泛化性和鲁棒性的trade-off，已有工作做了哪些尝试

提出本文idea可能面临的问题，引出下一段具体方法的介绍

Writing

Based on such motivation, we introduce a new concept called Module Robust Criticality (MRC) 2 to investigate the redundant capacity of adversarially trained models for robustness. MRC aims to quantify the maximum increase of robustness loss of a module's parameters under the constrained weight perturbation. As illustrated in Figure 3, we empirically find that certain modules exhibit redundant characteristics under such perturbations, resulting in negligible drops in adversarial robustness. We refer to the modules with the lowest MRC value as the non-robustcritical modules. These findings further inspire us to propose a novel fine-tuning technique called Robust Critical Fine-Tuning (RiFT), which aims to leverage the redundant capacity of the non-robust-critical module to improve generalization while maintaining adversarial robustness. RiFT consists of three steps: (1) Module robust criticality characterization, which calculates the MRC value for each module and identifies the non-robust-critical module. (2) Nonrobust-critical module fine-tuning, which exploits the redundant capacity of the non-robust-critical module via finetuning its weights with standard examples. (3) Mitigating robustness-generalization trade-off via interpolation, which interpolates between adversarially trained parameters and fine-tuned parameters to find the best weights that maximize the improvement in generalization while preserving adversarial robustness.

3️⃣ 承接上一段，具体介绍本文方法

回答上一段的问题

Writing

Experimental results demonstrate that RiFT significantly improves both the generalization performance and OOD robustness by around 2% while maintaining or even improving the adversarial robustness of the original models. Furthermore, we also incorporate RiFT to other adversarial training regimes such as TRADES [58], MART [48], AT-AWP [50], and SCORE [32], and show that such incorporation leads to further enhancements. More importantly, our experiments reveal several insights. First, we found that fine-tuning on non-robust-critical modules can effectively mitigate the trade-off between adversarial robustness and generalization, showing that these two can both be improved (Section 5.3). As illustrated in Figure 1, adversarial robustness increases alongside the generalization in the initial interpolation procedure, indicating that the features learned by fine-tuning can benefit both generalization and adversarial robustness. This contradicts the previous claim [46] that the features learned by optimal standard and robust classifiers are fundamentally different. Second, the existence of non-robust-critical modules suggests that current adversarial training regimes do not fully utilize the capacity of DNNs (Section 5.2). This motivates future work to design more efficient adversarial training approaches using such capacity. Third, while previous study [25] reported that fine-tuning on pre-train models could distort the learned robust features and result in poor performance on OOD samples, we find that fine-tuning adversarially trained models do NOT lead to worse OOD performance (Section 5.3).

4️⃣ 介绍实验结果，并介绍发现的现象

Writing

Experimental results demonstrate that RiFT significantly improves both the generalization performance and OOD robustness by around 2% while maintaining or even improving the adversarial robustness of the original models. Furthermore, we also incorporate RiFT to other adversarial training regimes such as TRADES [58], MART [48], AT-AWP [50], and SCORE [32], and show that such incorporation leads to further enhancements. More importantly, our experiments reveal several insights. First, we found that fine-tuning on non-robust-critical modules can effectively mitigate the trade-off between adversarial robustness and generalization, showing that these two can both be improved (Section 5.3). As illustrated in Figure 1, adversarial robustness increases alongside the generalization in the initial interpolation procedure, indicating that the features learned by fine-tuning can benefit both generalization and adversarial robustness. This contradicts the previous claim [46] that the features learned by optimal standard and robust classifiers are fundamentally different. Second, the existence of non-robust-critical modules suggests that current adversarial training regimes do not fully utilize the capacity of DNNs (Section 5.2). This motivates future work to design more efficient adversarial training approaches using such capacity. Third, while previous study [25] reported that fine-tuning on pre-train models could distort the learned robust features and result in poor performance on OOD samples, we find that fine-tuning adversarially trained models do NOT lead to worse OOD performance (Section 5.3).

4️⃣ 介绍实验结果，并介绍发现的

用大量篇幅介绍实验中得到的insights

Writing

5️⃣ 分点介绍贡献

The contribution of this work is summarized as follows:

1. We propose the concept of module robust criticality and verify the existence of redundant capacity for robustness in adversarially trained models. We then propose RiFT to exploit such redundancy to improve the generalization of AT models.

2. Our approach improves both generalization and OOD robustness of AT models. It can also be incorporated with previous AT methods to mitigate the trade-off between generalization and adversarial robustness.

3. The findings of our experiments shed light on the intricate interplay between generalization, adversarial robustness, and OOD robustness. Our work emphasizes the potential of leveraging the redundant capacity in AT models to improve generalization and robustness further, which may motivate more effective adversarial training methods.