Robust Mixture-of-Expert Training for Convolutional Neural Networks

Yihua Zhang[1], Ruisi Cai[2], Tianlong Chen[2,3,4,5], Guanhua Zhang[6], Huan Zhang[7,8], Pin-Yu Chen[9], Shiyu Chang[6], Zhangyang Wang[2], Sijia Liu[1,9]

[1]Michigan State University, [2]University of Texas at Austin, [3]The University of North Carolina at Chapel Hill, [4]MIT, [5]Harvard University, [6]UC Santa Barbara, [7]Carnegie Mellon University, [8]UIUC, [9]IBM Research

ICCV23 PARIS

Code

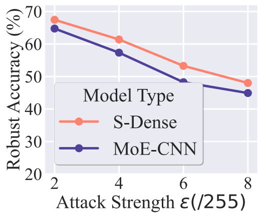
Paper

Open Question

What will be the new insights into adversarial robustness of sparse MoE-integrated CNNs? What will be the suited AT mechanism?

Warm-Up: AT for MoE-CNN is not Trivial

Challenge: Naively applying AT to MoE-CNN is even **less effective** than an AT-resulted small dense network.

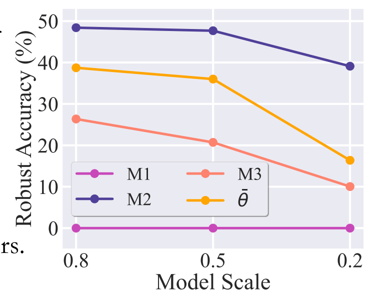


Robustness Dissection: Routers vs. Pathways

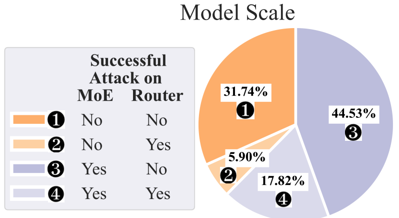
Q1: Is robustifying **routers** sufficient to achieve a robust MoE-CNN?

Insight 1: Robustifying routers improves the overall robustness of MoE-CNN but is NOT as effective as AT-resulted S-Dense.

M1: Undefended Dense CNN
M2: Adversarially trained S-Dense
M3: Sparse CNN with robust mask
$\bar{\theta}$: MoE-CNN with robustified routers.



Insight 2: Improving routers' robustness **alone** is NOT sufficient for robust MoE predictor although the former makes a positive impact.



Related Work

[1] Want et al., Deep mixture of experts via shallow embedding, UAI'22
[2] Puigcerver et al., On the adversarial robustness of mixture of experts, NeurIPS'22

Model Setup

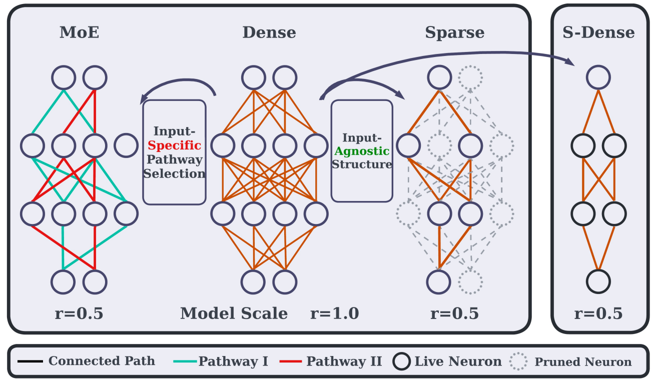
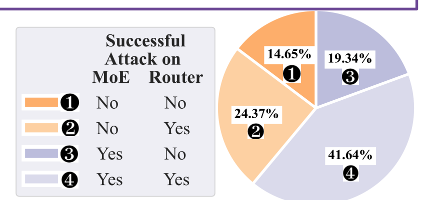


Figure 1: Model types considered in this work, including the MoE-CNN, big dense model, pruned sparse model, and small dense model, .

Q2: Will robustly training expert weights bring benefits and does it further impact routers?

Insight 3: routers' robustness is NOT automatically preserved if experts are updated. Routers' and experts' robustness are not easy to adapt to each other.



AdvMOE: Router-Expert Alternating AT via BLO

The current AT fails to (1) **model** and (2) **optimize the coupling** of the routers' and experts' robustness. We develop a new AT framework through **bi-level optimization** (BLO):

$$\underset{\psi}{\text{minimize}} \quad \ell_{\text{TRADES}}(\psi, \phi^*(\psi); \mathcal{D})$$
$$\text{subject to} \quad \phi^*(\psi) = \arg\min_{\phi} \ell_{\text{TRADES}}(\psi, \phi; \mathcal{D}),$$

AdvMoE: alternatively optimizes the lower level (router) and upper level (experts).

- ✓ Helps robust routers and experts "accommodate" to each other;
- ✓ Makes sure routers and experts make concerted efforts to overall robustness;
- ✓ Introduces no additional hyper-parameters.

Experiment Results Highlights

Method	Backbone	RA (%)	SA (%)	GFLOPS(#)	Method	Backbone	RA (%)	SA (%)	GFLOPS (#)
CIFAR-10									
● AT (Dense)	ResNet-18	50.13±0.13	82.99±0.11	0.54	● AT (Dense)	WRN-28-10	51.75±0.12	83.54±0.15	5.25
○ AT (S-Dense)		48.12±0.09	80.18±0.11	0.14 (74%↓)	○ AT (S-Dense)		50.66±0.13	82.24±0.10	1.31 (75%↓)
○ AT (Sparse)		47.93±0.17	**80.45**±0.13	0.14 (74%↓)	○ AT (Sparse)		48.95±0.14	82.44±0.17	1.31 (75%↓)
○ AT (MoE)		45.57±0.51	78.84±0.75	0.15 (72%↓)	○ AT (MoE)		46.73±0.46	77.42±0.73	1.75 (67%↓)
○ **AdvMoE**		**51.83** ±0.12	80.15±0.11	0.15 (72%↓)	○ **AdvMoE**		**55.73** ±0.13	**84.32** ±0.18	1.75 (67%↓)
● AT (Dense)	VGG-16	46.19±0.21	82.18±0.23	0.31	● AT (Dense)	DenseNet	44.52±0.14	74.97±0.19	0.07
○ AT (S-Dense)		45.72±0.18	**80.10**±0.16	0.07 (77%↓)	○ AT (S-Dense)		38.07±0.13	69.63±0.11	0.02 (71%↓)
○ AT (Sparse)		46.13±0.15	79.32±0.18	0.07 (77%↓)	○ AT (Sparse)		37.73±0.13	67.35±0.12	0.02 (71%↓)
○ AT (MoE)		43.37±0.46	76.49±0.65	0.12 (61%↓)	○ AT (MoE)		35.21±0.74	64.41±0.81	0.03 (57%↓)
○ **AdvMoE**		**49.82** ±0.11	80.03±0.10	0.12 (61%↓)	○ **AdvMoE**		**39.97**±0.11	**70.13**±0.15	0.03 (57%↓)
CIFAR-100									
● AT (Dense)	ResNet-18	27.23±0.08	58.21±0.12	0.54	● AT (Dense)	WRN-28-10	27.90±0.13	57.60±0.09	5.25
○ AT (S-Dense)		26.41±0.16	57.02±0.14	0.14 (74%↓)	○ AT (S-Dense)		26.30±0.10	56.80±0.08	1.31 (75%↓)
○ AT (Sparse)		26.13±0.14	57.24±0.12	0.14 (74%↓)	○ AT (Sparse)		25.83±0.16	57.39±0.14	1.31 (75%↓)
○ AT (MoE)		22.72±0.42	53.34±0.61	0.15 (72%↓)	○ AT (MoE)		22.94±0.55	53.39±0.49	1.75 (67%↓)
○ **AdvMoE**		**28.05** ±0.13	**57.73**±0.11	0.15 (72%↓)	○ **AdvMoE**		**28.82** ±0.14	**57.56**±0.17	1.75 (67%↓)
● AT (Dense)	VGG-16	22.37±0.15	52.36±0.17	0.31	● AT (Dense)	DenseNet	21.72±0.13	48.64±0.14	0.07
○ AT (S-Dense)		20.58±0.13	**48.89**±0.14	0.07 (77%↓)	○ AT (S-Dense)		16.86±0.21	39.97±0.11	0.02 (71%↓)
○ AT (Sparse)		21.12±0.22	48.03±0.17	0.07 (77%↓)	○ AT (Sparse)		17.72±0.14	41.03±0.16	0.02 (71%↓)
○ AT (MoE)		19.34±0.43	45.51±0.75	0.12 (61%↓)	○ AT (MoE)		14.45±0.45	36.72±0.71	0.03 (57%↓)
○ **AdvMoE**		**21.21**±0.21	48.33±0.17	0.12 (61%↓)	○ **AdvMoE**		**23.31** ±0.11	**48.97** ±0.14	0.03 (57%↓)
Tiny-ImageNet									
● AT (Dense)	ResNet-18	38.17±0.14	53.81±0.16	2.23	● AT (Dense)	WRN-28-10	38.82±0.15	55.30±0.19	21.0
○ AT (S-Dense)		36.29±0.16	52.15±0.13	0.55 (75%↓)	○ AT (S-Dense)		37.09±0.12	54.83±0.16	5.26 (75%↓)
○ AT (Sparse)		36.11±0.13	50.75±0.17	0.55 (75%↓)	○ AT (Sparse)		37.32±0.14	54.32±0.23	5.26 (75%↓)
○ AT (MoE)		34.41±0.31	47.73±0.41	0.75 (68%↓)	○ AT (MoE)		33.31±0.41	49.91±0.52	7.44 (65%↓)
○ **AdvMoE**		**39.99** ±0.12	**53.31**±0.14	0.75 (68%↓)	○ **AdvMoE**		**40.15** ±0.15	**55.18**±0.09	7.44 (65%↓)
ImageNet									
● AT (Dense)	ResNet-18	44.64±0.14	60.32±0.15	1.82	● AT (Dense)	WRN-28-10	45.13±0.14	60.97±0.16	16.1
○ AT (S-Dense)		41.19±0.16	58.32±0.12	0.48 (74%↓)	○ AT (S-Dense)		41.72±0.15	58.98±0.18	4.04 (75%↓)
○ AT (Sparse)		40.87±0.15	58.22±0.13	0.48 (74%↓)	○ AT (Sparse)		39.88±0.18	**59.21**±0.14	4.04 (75%↓)
○ AT (MoE)		35.57±0.73	55.47±0.66	0.67 (63%↓)	○ AT (MoE)		37.42±0.44	56.44±0.71	5.15 (68%↓)
○ **AdvMoE**		**43.32** ±0.12	**59.72**±0.17	0.67 (63%↓)	○ **AdvMoE**		**46.82** ±0.11	58.87±0.07	5.15 (68%↓)

Table 1: Performance overview of AdvMOE vs. baselines on various datasets and architectures.

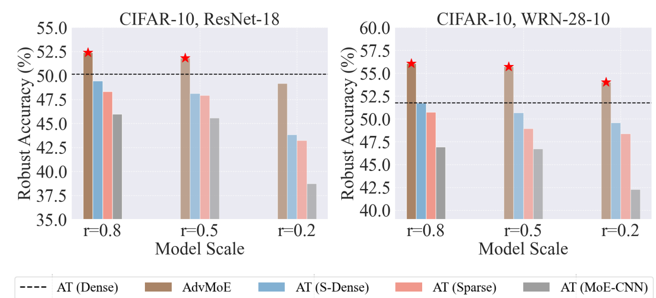


Figure 4: Robustness comparison of models trained with different methods under various model scale settings.

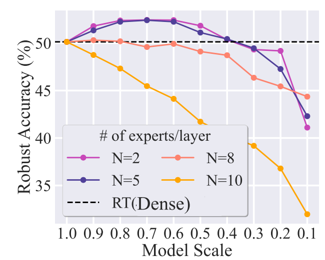


Figure 5: Performance of AdvMoE with different expert number N and model scale r on (CIFAR-10, ResNet18)

Contact: {zhan1908@msu.edu}