On the Effect of Pruning on Adversarial Robustness

Artur Jordão and Hélio Pedrini
Institute of Computing, University of Campinas (UNICAMP), Brazil

UNICAMP

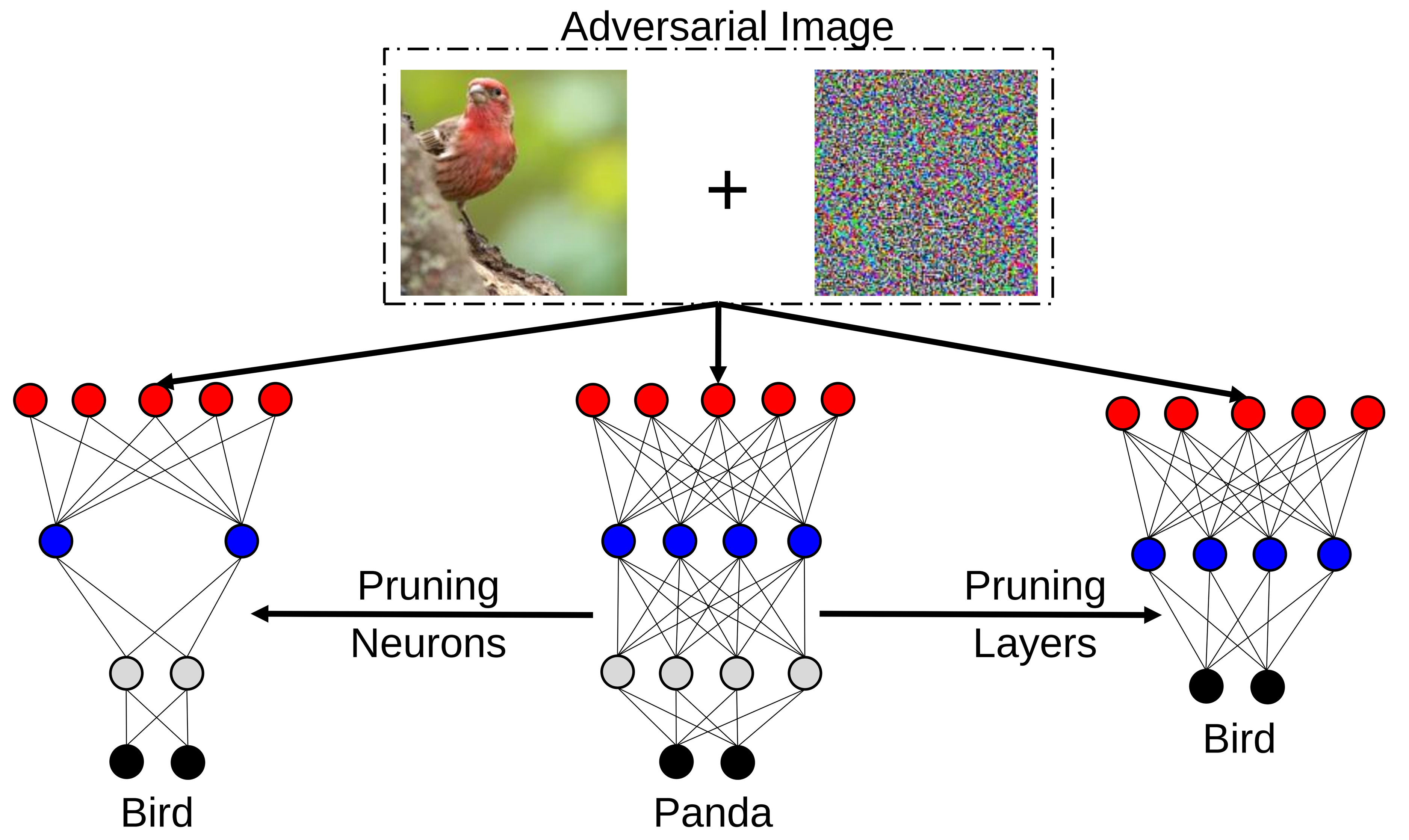
2021 ICCV OCTOBER 11-17 VIRTUAL

Research Question

Do pruned ($\mathcal{F}'$) networks inherit the vulnerability to adversarial images of their unpruned ($\mathcal{F}$) counterpart?

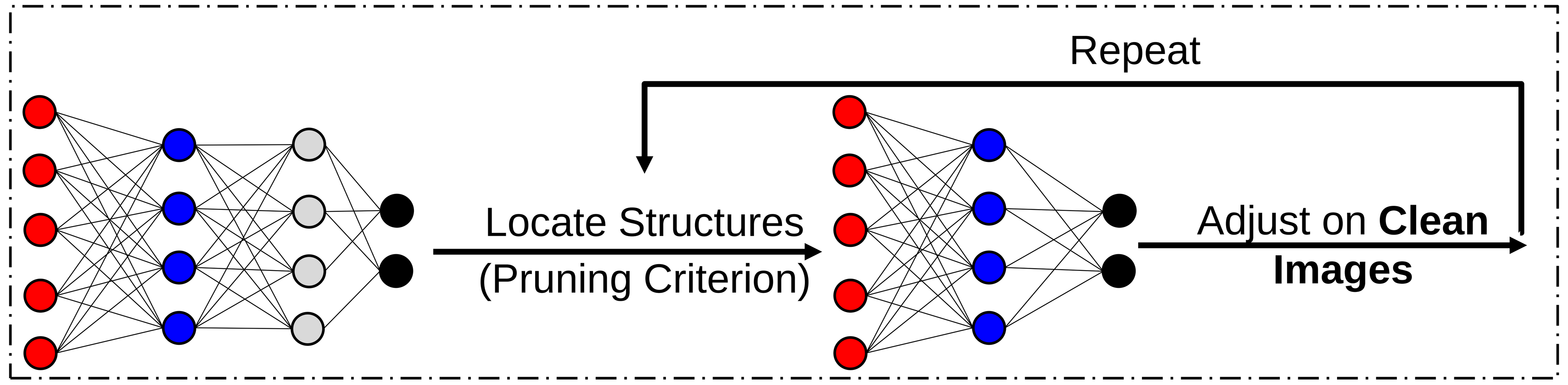
$$Acc_{adv}(\mathcal{F}') > Acc_{adv}(\mathcal{F}), Acc_{clean}(\mathcal{F}') \approx Acc_{clean}(\mathcal{F})$$

Are pruned networks capable of improving robustness while preserving generalization?



Methodology

Pruning Strategy – **Only Clean Images**



Experiments

Table 1. Robustness and Generalization from Pruning

Architecture	Structure	Semantic	Occlusion	FGSM	Clean	Average
ResNet56	Filters	**(+) 1.58**	(+) 2.76	**(+) 3.68**	(+) 0.60	**(+) 2.15**
	Layers	(+) 1.05	(+) 1.06	(+) 3.20	(+) 0.84	(+) 1.53
	Both	(-) 4.13	**(+) 4.62**	(+) 0.36	(-) 0.60	(+) 0.06
MobileNetV2	Filters	(-) 0.60	**(+) 3.35**	(+) 0.64	**(+) 0.37**	(+) 0.94
	Layers	(-) 0.49	(+) 2.12	**(+) 1.44**	(+) 0.15	(+) 0.80
	Both	**(+) 0.07**	(+) 2.56	(+) 1.05	(+) 0.17	**(+) 0.96**

Table 2. Adjustment

	Semantic	Occlusion	FGSM	Average
Scratch-E	**(+) 1.72**	(-) 0.60	(+) 1.64	(+) 0.92
Scratch-B	(+) 1.25	(-) 0.71	(+) 3.64	(+) 1.39
W-Ticket	(+) 1.13	(-) 5.41	(+) 1.40	(-) 0.96
Fine-Tuning	(+) 1.58	**(+) 2.76**	**(+) 3.68**	**(+) 2.67**

Table 3. Influence of the Pruning Criterion

Pruning Criterion	Semantic	Occlusion	FGSM	Clean	Average
ℓ_1-norm	**(+) 1.64**	(-) 0.80	(+) 3.75	(+) 0.38	(+) 1.24
ExpectedABS	(+) 0.96	(-) 0.09	**(+) 4.29**	(+) 0.51	(+) 1.41
HRank	(+) 0.93	**(+) 2.92**	(+) 3.18	(+) 0.39	(+) 1.85
KlDivergence	(+) 0.82	(+) 0.73	(+) 3.00	(+) 0.34	(+) 1.22
PLS	(+) 1.58	(+) 2.76	(+) 3.68	**(+) 0.60**	**(+) 2.15**

Table 4. Comparison with Competing Defense Mechanisms

Defense	Robustness	Generalization	Average
Stylized	(-) 2.29	(-) 16.20	(-) 9.24
MixUp	(-) 4.77	**(+) 1.10**	(-) 1.83
Cutout	(+) 1.39	(+) 0.75	(+) 1.06
CutMix	(+) 1.71	(+) 2.07	(+) 1.89
Shape-Texture	**(+) 7.50**	(+) 0.50	**(+) 4.00**
Pruning Filters	(+) 1.14	**(+) 3.15**	**(+) 2.14**
Pruning Layers	**(+) 1.20**	(+) 3.03	(+) 2.11

Conclusion

- We empirically show that pruning structures of convolutional networks increase their adversarial robustness
- We demonstrate that pruning preserves generalization; thus, it efficiently satisfies the dilemma between robustness and generalization
- We confirm these findings considering **only clean images** during the pruning process, which enables us to design an effective defense mechanism that ignores the settings and additional assumptions of the attack