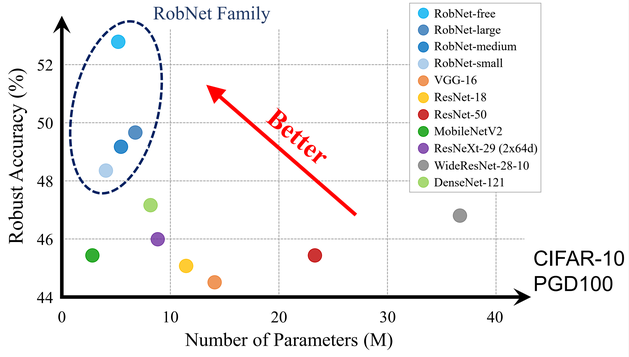
When NAS Meets Robustness: In Search of Robust Architectures against Adversarial Attacks

Minghao Guo*, Yuzhe Yang*, Rui Xu, Ziwei Liu, Dahua Lin
The Chinese University of Hong Kong & MIT CSAIL

CVPR SEATTLE JUNE 14-19 2020 WASHINGTON

Motivation

- To enhance the robustness of deep networks, extensive efforts on specialized learning algorithms and loss functions have been developed
- However, the intrinsic influence of **network architecture** on network resilience to adversarial perturbations **has not been** well studied
- We take the **first** step to systematically understand adv. robustness from an **architectural** perspective

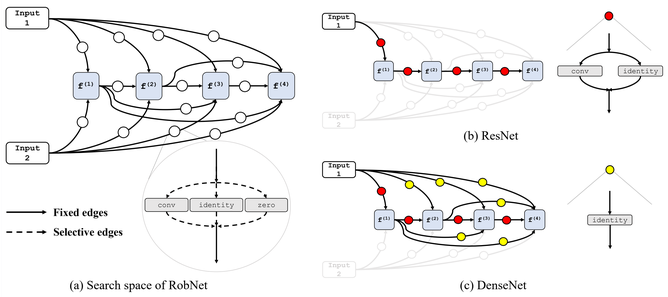


Robust Architecture Odyssey

- What kind of **network architecture patterns** is crucial for adversarial robustness?
- Given a budget of model capacity, how to allocate the **parameters of the architecture** to efficiently improve the network robustness?
- What is the **statistical indicator** for robust network architectures?

RobNet Search Framework

Search Space:



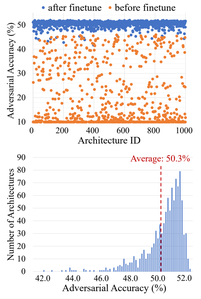
(a) Search space of RobNet (b) ResNet (c) DenseNet

Robust Search Algorithm:

- One-Shot NAS;
- PGD training for super-net;
- finetuning a few epochs for individual candidate architecture

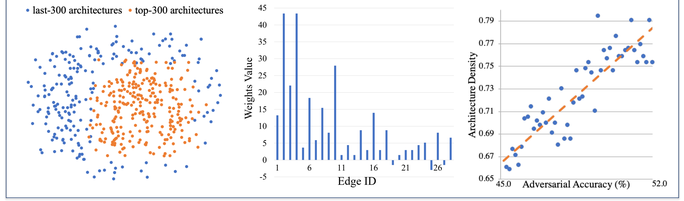
Robustness Evaluation:

- 1,000 randomly sampled candidates;
- white-box PGD



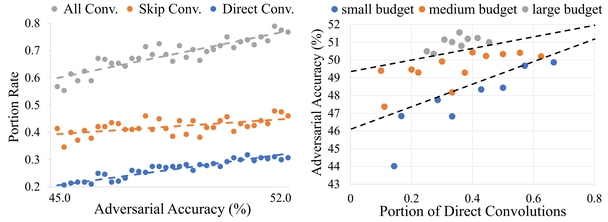
Finding #1: Densely connected pattern benefits network robustness

- Correlation between Architecture Density & Robustness



Finding #2: Architecture strategy under computational budget

- Under small computational budget, **adding conv operations to direct edges** is more effective.

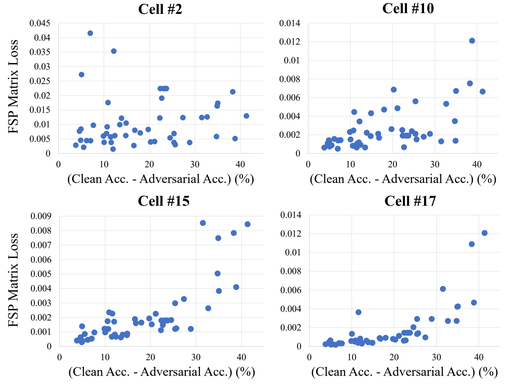


Finding #3: FSP matrix distance as robustness indicator

- Flow of solution procedure (FSP) matrix

$$G_l(x;\theta) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^{in}_{l,s,t}(x;\theta) \times F^{out}_{l,s,t}(x;\theta)}{h \times w}$$

- A robust network has a **lower FSP matrix loss** in the **deeper** cells of network.



Results with RobNet

- CIFAR-10

Models	Model Size	Natural Acc.	FGSM	PGD^{20}	PGD^{100}	DeepFool	MI-FGSM
ResNet-18	11.17M	78.38%	49.81%	45.60%	45.10%	47.64%	45.23%
ResNet-50	23.52M	79.15%	51.46%	45.84%	45.35%	49.18%	45.53%
WideResNet-28-10	36.48M	**86.43%**	53.57%	47.10%	46.90%	51.23%	47.04%
DenseNet-121	6.95M	82.72%	54.14%	47.93%	47.46%	51.70%	48.19%
RobNet-small	4.41M	78.05%	53.93%	48.32%	48.07%	52.96%	48.98%
RobNet-medium	5.66M	78.33%	54.55%	49.13%	48.96%	53.32%	49.34%
RobNet-large	6.89M	78.57%	54.98%	49.44%	49.24%	53.85%	49.92%
RobNet-large-v2	33.42M	**85.69%**	57.18%	50.53%	50.26%	55.45%	50.87%
RobNet-free	5.49M	82.79%	**58.38%**	**52.74%**	**52.57%**	**57.24%**	**52.95%**

- ImageNet & other datasets

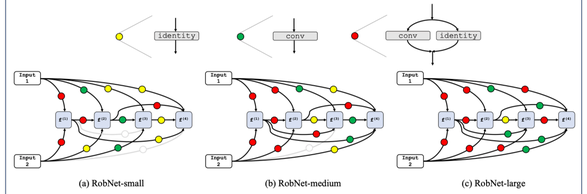
Models	Model Size	Natural Acc.	PGD^{10}	PGD^{50}	PGD^{100}
ResNet-50	23.52M	60.20%	32.76%	31.87%	31.81%
ResNet-101	42.52M	63.34%	35.38%	34.40%	34.32%
ResNet-152	58.16M	**64.44%**	36.99%	36.04%	35.99%
RobNet-large	12.76M	61.26%	**37.16%**	**37.15%**	**37.14%**

Models	SVHN	CIFAR-100	Tiny-ImageNet
ResNet-18	46.08%	22.01%	16.96%
ResNet-50	47.23%	22.38%	19.12%
RobNet-large	**51.26%**	**23.19%**	**19.90%**
RobNet-free	**55.59%**	**23.87%**	**20.87%**

- Boosting Existing Technique

Models	Natural Acc.	PGD^{100}
ResNet-18	78.38%	45.10%
ResNet-18 + Denoise	78.75%	45.82%
RobNet-large	78.57%	49.24%
RobNet-large + Denoise	**84.03%**	**49.97%**

- Visualization of architectures of RobNet family



(a) RobNet-small (b) RobNet-medium (c) RobNet-large

Conclusions

- **See our models in** https://github.com/gmh14/RobNets
- **Also checkout the project page**