When NAS Meets Robustness:

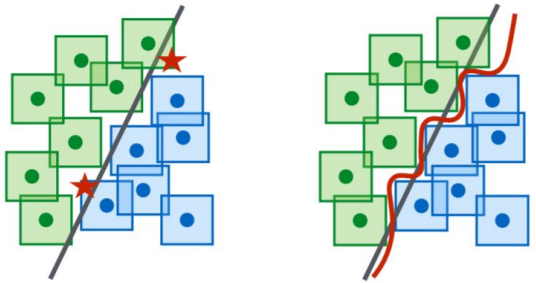
In Search of Robust Architectures against Adversarial Attacks

Minghao Guo*, **Yuzhe Yang***, Rui Xu, Ziwei Liu, Dahua Lin

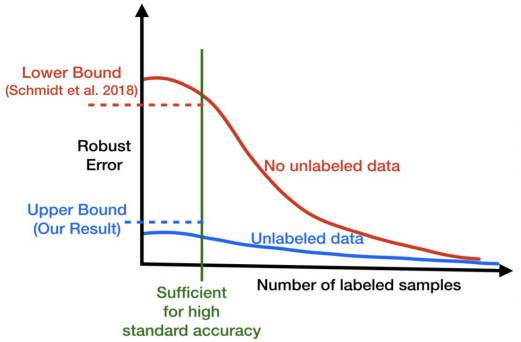
(* indicates equal contribution)

Studies on Improving Adversarial Robustness

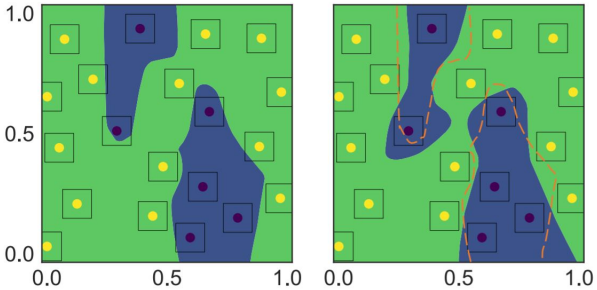
- Specialized learning algorithms / loss functions / data preprocessing / unlabeled data ...



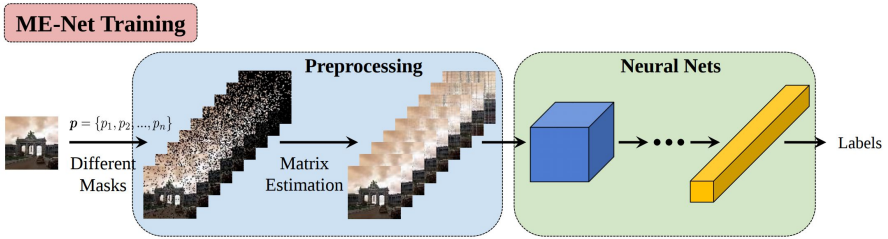
Mądry, Makelov, Schmidt, Tsipras, Vladu. ICLR'18



Carmon, Raghunathan, Schmidt, Liang, Duchi. NeurIPS'19



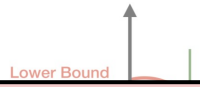
Zhang, Yu, Jiao, Xing, Ghaoui, Jordan. ICML'19



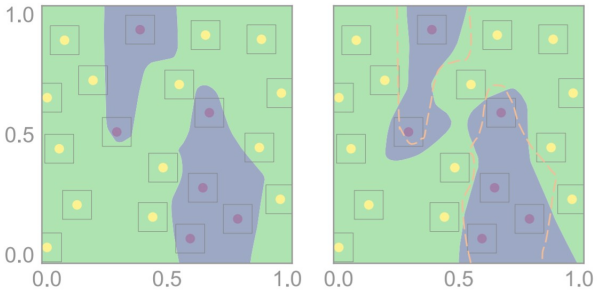
Yang, Zhang, Katabi, Xu. ICML'19

Studies on Improving Adversarial Robustness

- Specialized learning algorithms / loss functions / data preprocessing / unlabeled data ...



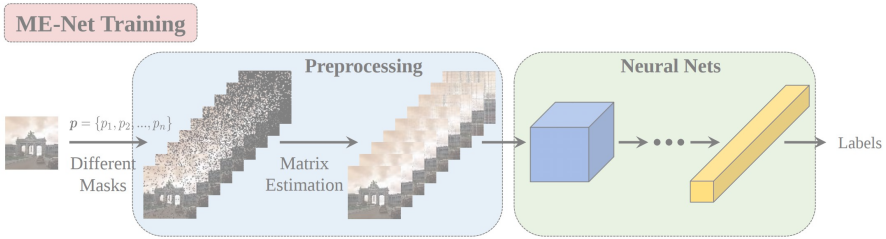
Intrinsic influence of neural network architectures on adversarial robustness?



Zhang, Yu, Jiao, Xing, Ghaoui, Jordan. ICML'19

standard accuracy

Carmon, Raghunathan, Schmidt, Liang, Duchi. NeurIPS'19



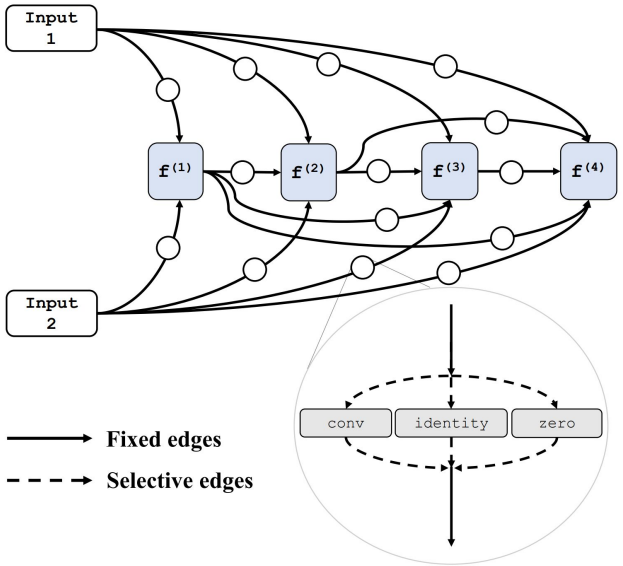
Yang, Zhang, Katabi, Xu. ICML'19

Robust Architecture Search Framework

Robust Architecture Search Framework

One-shot robust NAS

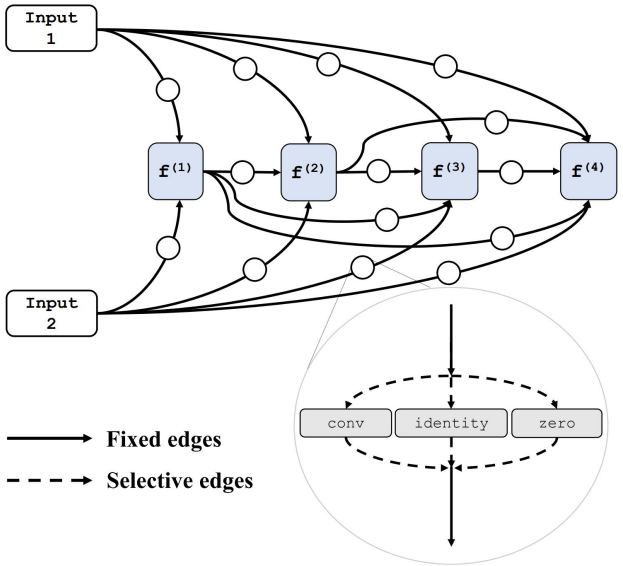
- PGD adversarial training for supernet



Robust Architecture Search Framework

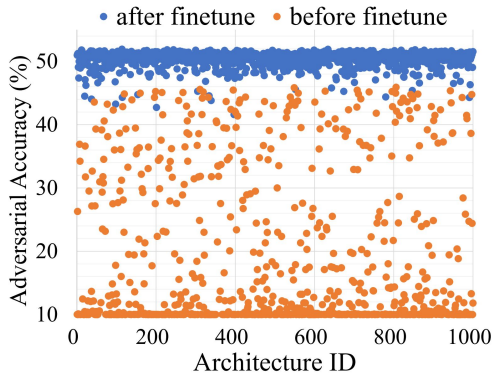
One-shot robust NAS

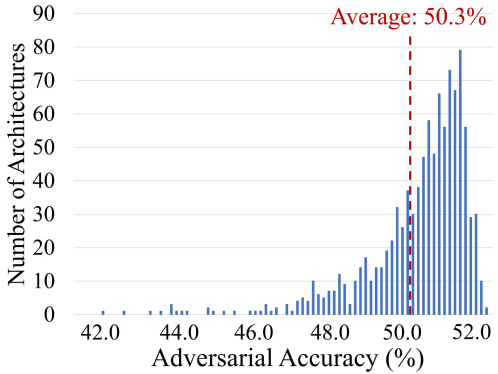
- PGD adversarial training for supernet



Robustness evaluation

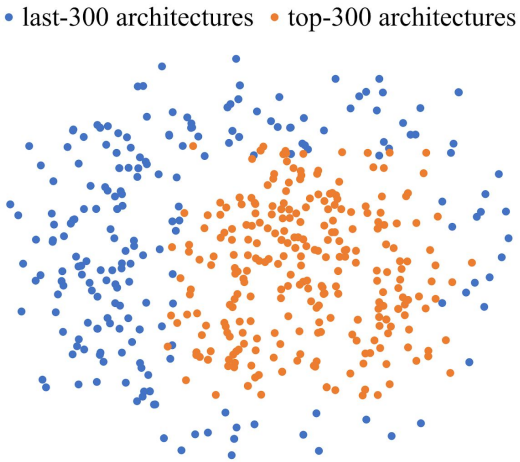
- 1,000 randomly sampled candidates
- finetune a few epochs for individual candidate architecture

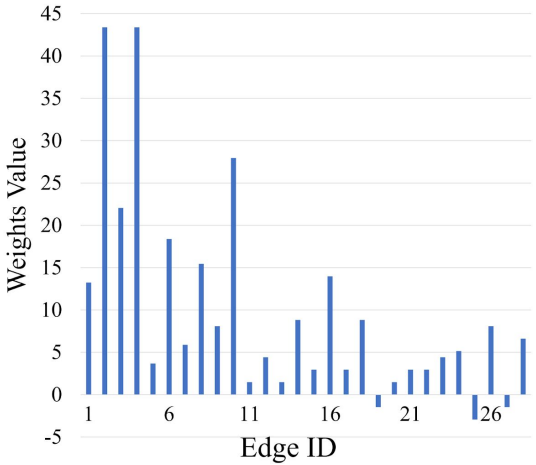


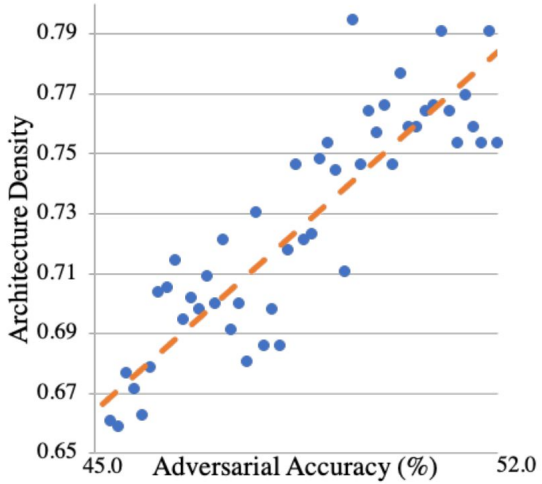


Observation #1:
Densely connected pattern benefits network robustness

Strong correlation between **Architecture Density** & **Robustness**

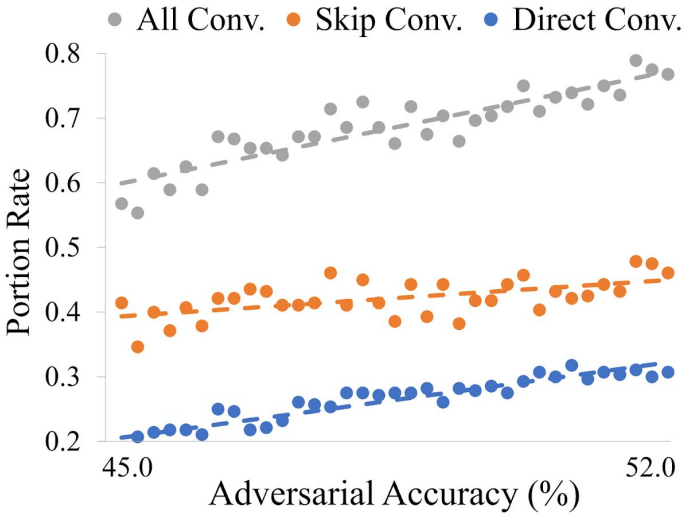


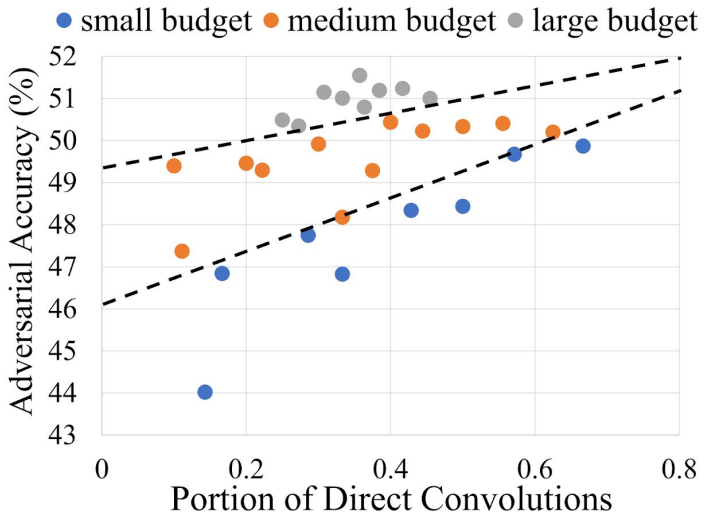




Observation #2:
Architecture strategy under computational budget

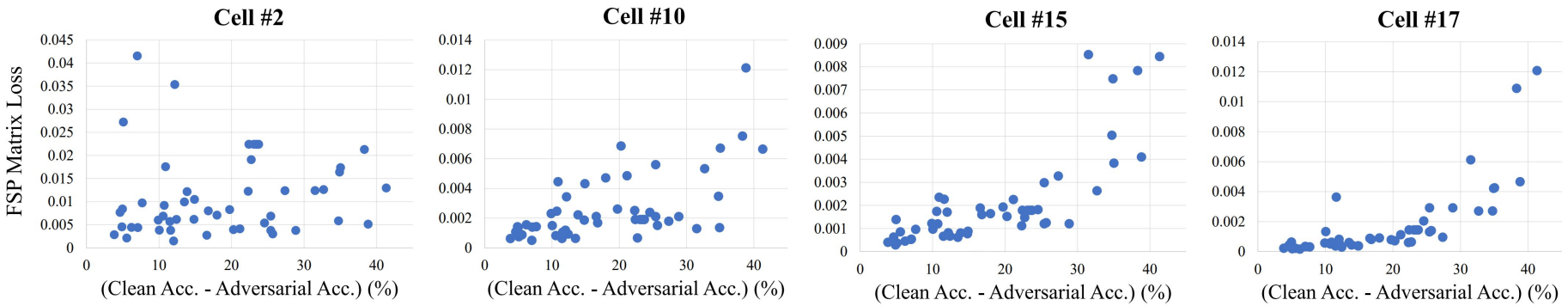
Under small computational budget,
adding convolution operations to direct edges is more effective





Observation #3:
FSP matrix distance as robustness indicator

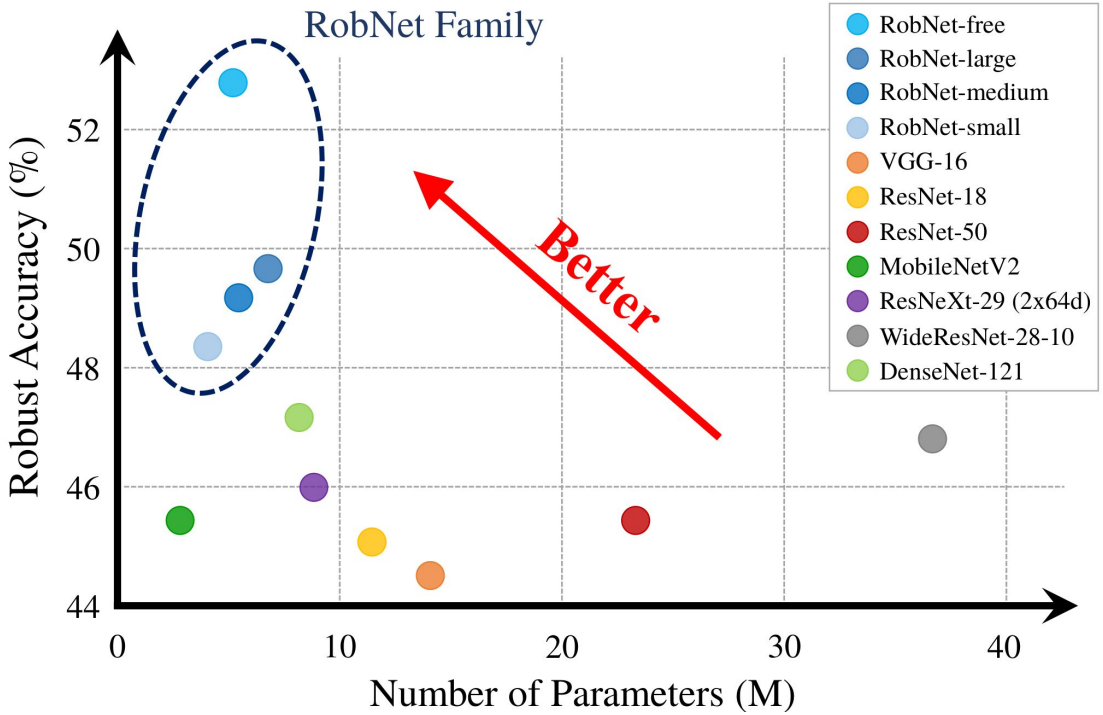
A robust network has a **lower FSP matrix loss** in the **deeper** cells of network



(Flow of solution procedure (FSP) matrix: $G_l(x;\theta) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^{in}_{l,s,t}(x;\theta) \times F^{out}_{l,s,t}(x;\theta)}{h \times w}$)

Family of Robust Architectures (RobNets)

- RobNets exhibit superior robustness on CIFAR, SVHN, ImageNet, etc. with fewer parameters



Check out our models at...

https://github.com/gmh14/RobNets

https://www.mit.edu/~yuzhe/robnets.html