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

SUPPLEMENTARY MATERIAL

Minghao Guo[1*] Yuzhe Yang[2*] Rui Xu[1] Ziwei Liu[1] Dahua Lin[1]

[1]The Chinese University of Hong Kong [2]MIT CSAIL

1. Details of Robust Architecture Search

We provide details of our robust architecture search algorithm. The pseudo code is illustrated in Algorithm 1.

Algorithm 1 Robust architecture search

Input: Supernet $G = (V, E)$, architecture parameter α, total iterations I, PGD attack iterations T.
Set all elements in α to 1
for $k = 0 \dots I$ **do**
  Randomly sample a training batch $\{x_i, y_i\}_{i=1}^{B}$ from train dataset
  Randomly set some of the elements in α to 0 and get the corresponding network parameter θ_k
  `/* Parallel training in PyTorch */`
  for $i = 1 \dots B$ **do**
    $x_i^{(0)} \leftarrow x_i$
    `/* PGD adversarial example */`
    for $t = 0 \dots (T-1)$ **do**
      $x_i^{(t+1)} \leftarrow \Pi_S(x_i^{(t)} + \eta \cdot \text{sign}(\nabla_x \mathcal{L}(\theta_k, x_i^{(t)}, y_i))$
    end for
  end for
  Use $\{x_i^{(T)}, y_i\}_{i=1}^{B}$ to do one step training and update θ_k by SGD
  Set all elements in α to 1
end for

2. Details of Adversarial Training

We further provide training details of PGD-based adversarial training for each individual architecture on CIFAR, SVHN, and Tiny-ImageNet. We summarize our training hyper-parameters in Table 1. We follow the standard data augmentation scheme as in [2] to do zero-padding with 4 pixels on each side, and then random crop back to the original image size. We then randomly flip the images horizontally and normalize them into $[0, 1]$. We use the same training settings for CIFAR-10 and CIFAR-100.

Table 1: Details of adversarial training on different datasets. Learning rate is decreased at selected epochs, using a step factor of 0.1. We apply the same training setting for both CIFAR-10 and CIFAR-100.

	CIFAR	**SVHN**	**Tiny-ImageNet**
Optimizer	SGD	SGD	SGD
Momentum	0.9	0.9	0.9
Epochs	200	200	90
LR	0.1	0.01	0.1
LR decay	step (100, 150)	step (100, 150)	step (30, 60)

3. Additional Results on ImageNet

In this section, we provide additional robustness results of RobNets on ImageNet [1], a large-scale image classification dataset that contains ∼1.28 million images and 1000 classes. Since adversarial training on ImageNet demands a vast amount of computing resources (e.g., hundreds of GPUs [7, 3] for several days), we adopt the recent "free" adversarial training scheme [5] for accelerating adversarial training on ImageNet. Specifically, we follow the settings in [5] that consider non-targeted attack, and restrict the perturbation bound to be $\epsilon = 4/255$ (0.015). We use $m = 4$ for the "free" training, and keep other hyper-parameters the same for all models.

We compare RobNet-large model with different variants of ResNet against white-box PGD attacks. The results are shown in Table 2. Compared to different ResNet models, RobNet-large can consistently achieve higher robust accuracy against PGD adversary, while maintaining similar clean accuracy. We note that the model size of RobNet-large is far smaller than the baseline models. As we have investigated in the main text, by increasing the network parameters

*Equal contribution. Order determined by alphabetical order.

Table 2: White-box attack results on ImageNet. We compare representative RobNet models with state-of-the-art architectures. All models are adversarially trained using "free" training [5]. All attacks are l_∞-bounded with a total perturbation scale of $4/255$ (0.015).

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%**

of RobNet models, we can not only strengthen the adversarial robustness, the natural accuracy can also be significantly improved. This phenomenon can also be observed by comparing ResNet models with different capacities. Thus, we believe by further increasing the parameter numbers, RobNets can achieve even higher accuracy in both clean and adversarial settings.

4. Comparisons to More Architectures

In this section, we provide a more comprehensive comparison between RobNet models and various state-of-the-art human-designed architectures. In addition to ResNet and DenseNet family we have mentioned in the main text, we further add baseline architectures including VGG [6], MobileNetV2 [4], and ResNeXt [8], and report the results in Table 3. We again consider l_∞-bounded white-box attack setting on CIFAR-10, with all models trained identically as we have described. As can be observed from the table, when comparing to various human-designed architectures, RobNet models can consistently achieve higher adversarial robustness, even with much fewer network parameters.

5. Complete Results of FSP Matrix Loss

We provide additional results for the correlation of FSP matrix distance along with the performance gap between clean accuracy and adversarial accuracy in cell-free setting. Results for several cells have been shown in the main paper. Here we provide results for additional cells in Fig. 1.

As can be observed from the figure, for cells in deeper positions of the network, the FSP distance has a positive correlation with the gap between network robustness and its clean accuracy, which indicates that a robust network has a lower FSP matrix loss in the deeper cells of the network.

6. Visualization of RobNets

In this section, we first describe the details of how we select architectures of RobNet family. Further, we visualize several representative RobNet architectures.

Table 3: Comparison between representative RobNet models and more human-designed architectures on CIFAR-10. All models are adversarially trained using PGD with 7 steps. All attacks are l_∞-bounded with a total perturbation scale of $8/255$ (0.031).

Models	Model Size	Natural Acc.	PGD^{100}
VGG-16	14.73M	77.38%	44.38%
ResNet-18	11.17M	78.38%	45.10%
ResNet-50	23.52M	79.15%	45.35%
MobileNetV2	2.30M	76.79%	45.50%
ResNeXt-29 (2x64d)	9.13M	81.86%	46.04%
WideResNet-28-10	36.48M	**86.43%**	46.90%
DenseNet-121	6.95M	82.72%	47.46%
RobNet-small	4.41M	78.05%	48.07%
RobNet-medium	5.66M	78.33%	48.96%
RobNet-large	6.89M	78.57%	49.24%
RobNet-large-v2	33.42M	**85.69%**	50.26%
RobNet-free	5.49M	82.79%	**52.57%**

In cell-based setting, we first filter out the architectures with architecture density $D < 0.5$. Then we only consider the architectures which have a portion of direct convolutions larger than 0.5. For each of the computational budget, we sample 50 architectures from the supernet following the process described above and finetune them for 3 epochs to get the adversarial accuracy. We then select architectures with best performance under each computational budget, and refer to them as RobNet-small, RobNet-medium, and RobNet-large, respectively.

In cell-free setting, we first randomly sample 300 architectures from the supernet, and calculate the average FSP matrix distance for last 10 cells of each sampled network. Following the finding of FSP matrix loss as indicator, we

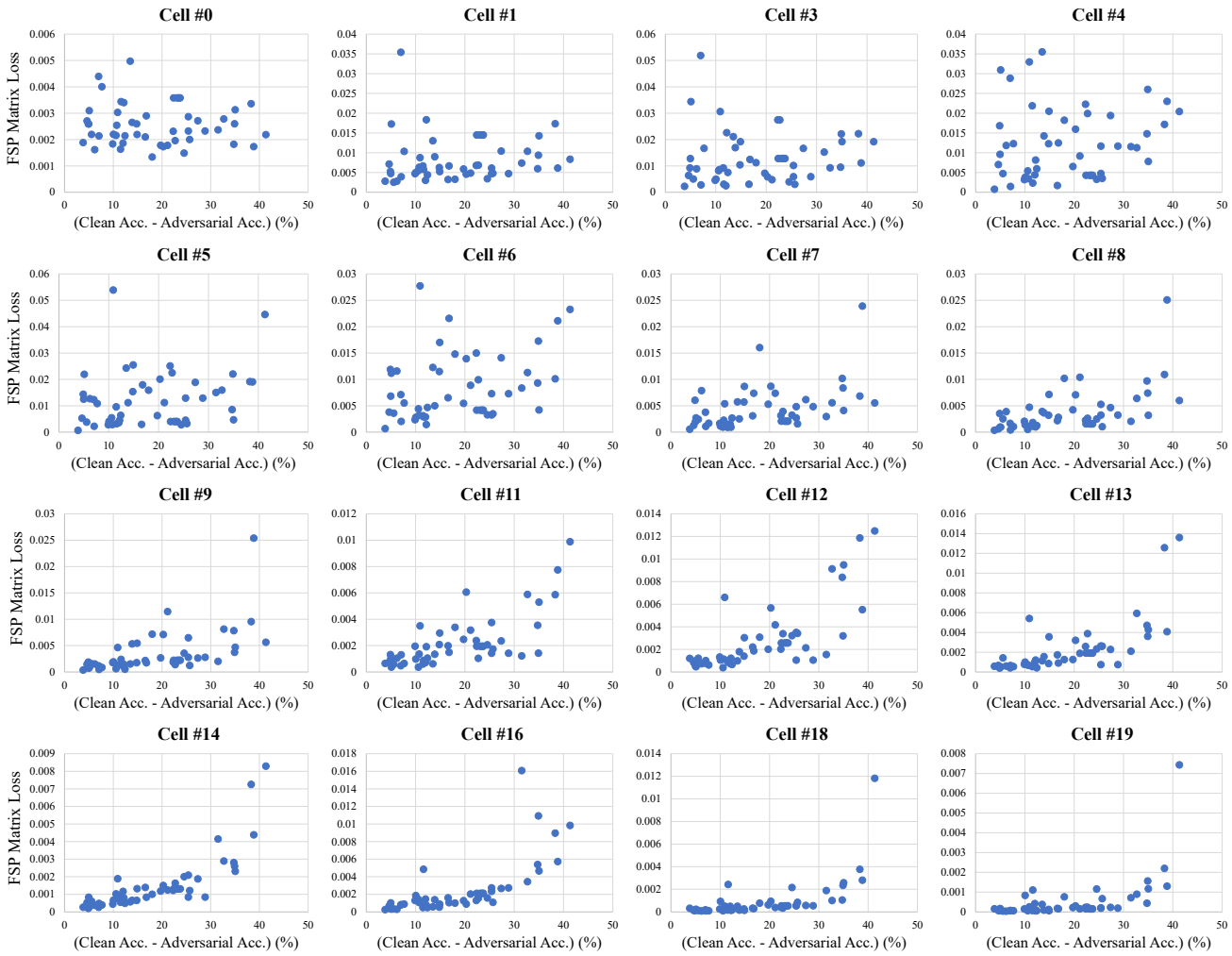
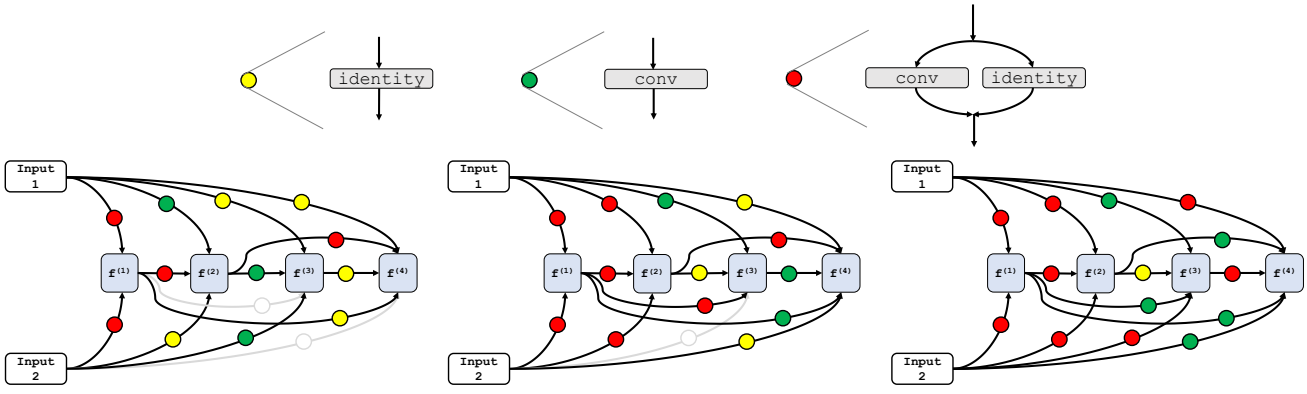


Figure 1: Analysis of FSP matrix distance as robustness indicator. We compute the FSP matrix distance for each cell, along with the performance gap between clean accuracy and adversarial accuracy. For cells in deeper positions of the network, the FSP distance has a positive correlation with the gap between network robustness and its clean accuracy, which indicates that a robust network has a lower FSP matrix loss in the deeper cells of the network.



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

Figure 2: Visualization of representative architectures of RobNet family.

Table 4: Black-box PGD^{100} attack results on CIFAR-10. All models are adversarially trained using PGD with 7 steps. We create PGD adversarial examples with $\epsilon = 8/255$ (0.031) for 100 iterations from the evaluation set on the source network, and then evaluate them on an independently initialized target network. The best results of each column are in **bold** and the empirical lower bound (the lowest accuracy of each row) for each network is underlined.

Target \ Source	ResNet-18	ResNet-50	WideResNet-28-10	DenseNet-121	RobNet-large	RobNet-free
ResNet-18	54.28%	54.49%	56.44%	57.19%	55.57%	59.37%
ResNet-50	56.24%	55.89%	56.38%	58.31%	57.22%	60.19%
WideResNet-28-10	57.89%	57.96%	55.68%	58.41%	59.08%	60.74%
DenseNet-121	61.42%	61.96%	60.28%	59.34%	60.03%	59.96%
RobNet-large	59.63%	59.82%	59.72%	60.03%	59.58%	60.73%
RobNet-free	**66.64%**	**66.09%**	**65.05%**	**64.40%**	**63.35%**	**63.17%**

reject those architectures whose average distance is larger than a threshold. In our experiments, we set the threshold to be 0.006, which leads to 68 remaining architectures. Finally, we finetune each of them for 3 epochs and select the architecture with the highest adversarial accuracy, which is named as RobNet-free.

We visualize several representative architectures of RobNet family in Fig. 2.

7. Additional Black-box Attack Results

We provide additional results on transfer-based black-box attacks on CIFAR-10, across different network architectures. The black-box adversarial examples are generated from an independently trained copy of the network, by using *white-box* attack on the victim network. We apply PGD-based black-box attacks with 100 iterations across different architectures, and report the result in Table 4. All models are adversarially trained using PGD with 7 steps.

In the table, we highlight the best result of each column in **bold**, which corresponds to the most robust model against black-box adversarial examples generated from one specific source network. We also underline the empirical lower bound for each network, which corresponds to the lowest accuracy of each row.

As the table reveals, RobNet-free model achieves the highest robust accuracy under transfer-based attacks from different sources. Furthermore, the most powerful black-box adversarial examples for each network (i.e., the underlined value) are from source network that uses the same architecture as the target network. Finally, by comparing the transferability between two network architectures (e.g., RobNet-free $\rightarrow$ ResNet-18 & ResNet-18 $\rightarrow$ RobNet-free), we can observe the following phenomena. First, our RobNet models are more robust against black-box attacks transferred from other models. Moreover, our RobNet models can generate stronger adversarial examples for black-box attacks compared with other widely used models.

8. Additional White-box Attack Results

As common in recent literature [7, 9, 10], *strongest possible* attack should be considered when evaluating the adversarial robustness. Therefore, we further strengthen the adversary and vary the attack iterations from 7 to 1000. We show the results in Fig. 3, where RobNet family outperforms other networks, even facing the strong adversary. Specifically, compared to state-of-the-art models, RobNet-large and RobNet-free can gain $\sim 2\%$ and $\sim 5\%$ improvement, respectively. We also observe that the attacker performance diminishes with 500$\sim$1000 attack iterations.

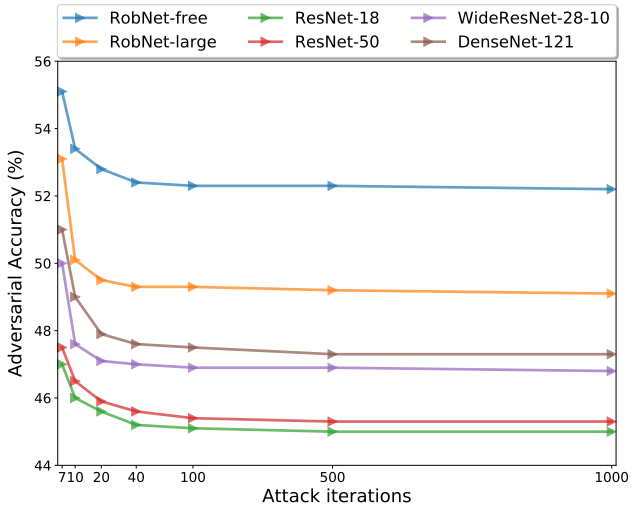


Figure 3: White-box attack results on CIFAR-10. All models are adversarially trained using PGD with 7 steps. We show results of different architectures against a white-box PGD attacker with 7 to **1000** attack iterations.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 1

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2

[5] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019. 1, 2

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[7] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *CVPR*, 2019. 1, 4

[8] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 2

[9] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. ME-Net: Towards effective adversarial robustness with matrix estimation. In *ICML*, 2019. 4

[10] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 4