四川大學

SICHUAN UNIVERSITY

ICLR

Zero-cost Proxy for Adversarial Robustness Evaluation

Yuqi Feng, Yuwei Ou, Jiahao Fan, Yanan Sun*

College of Computer Science, Sichuan University

(*Corresponding author)

Outline

1. Introduction & Motivations
2. Methodology & Formulations
3. Dataset & Experiments

Deep Neural Networks

- Deep neural networks (DNN) have shown remarkable performance in various real-world applications

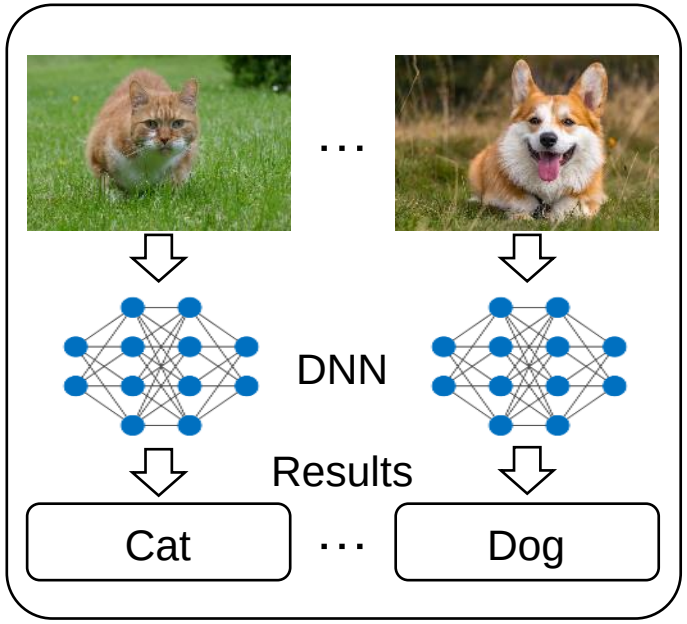
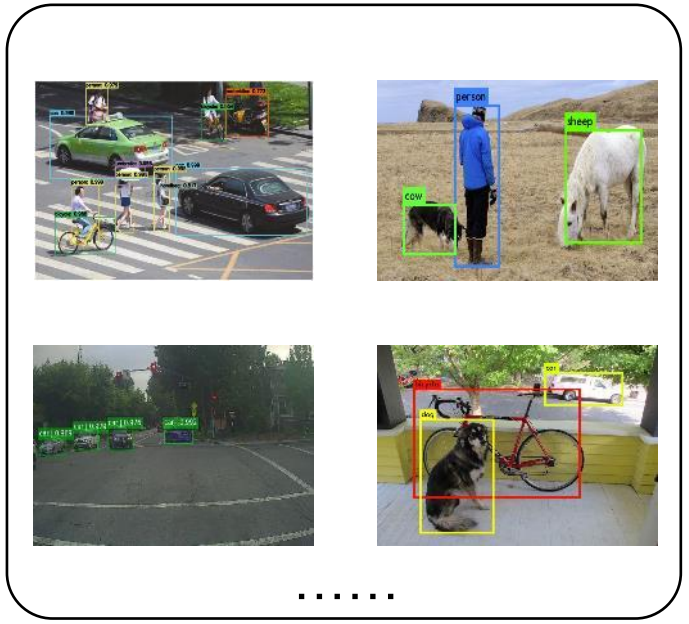
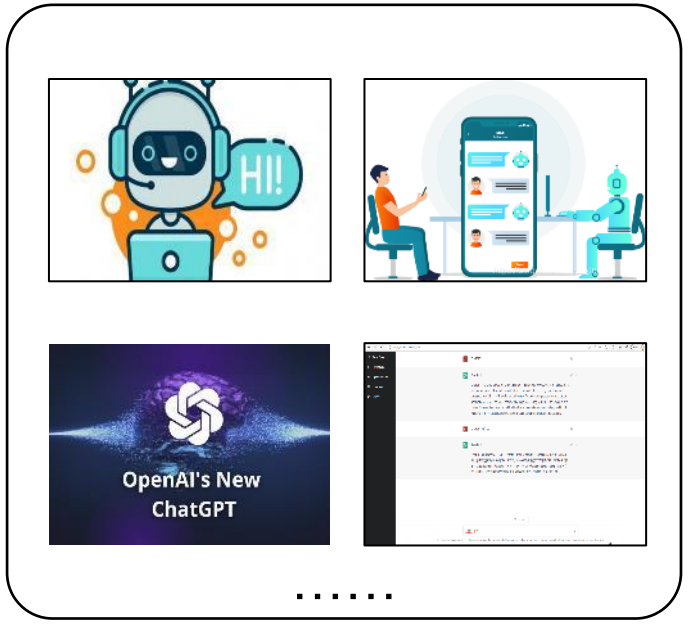


Image classification



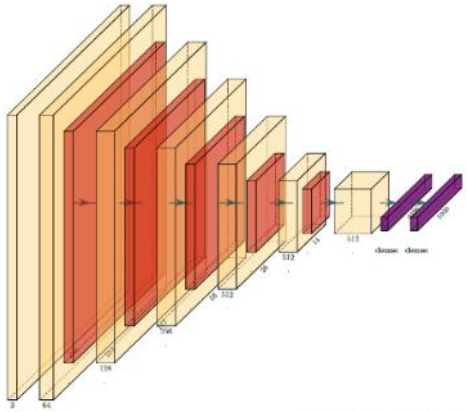
Object detection



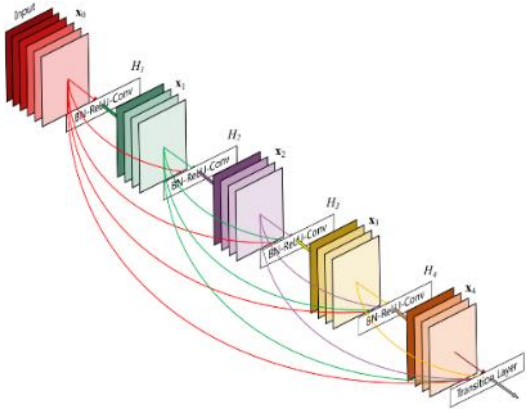
Natural Language Processing

Neural Architectures

- The neural architectures of DNNs can learn meaningful features, helping DNNs achieve better performance
- The design of neural architectures heavily relies on the domain expertise

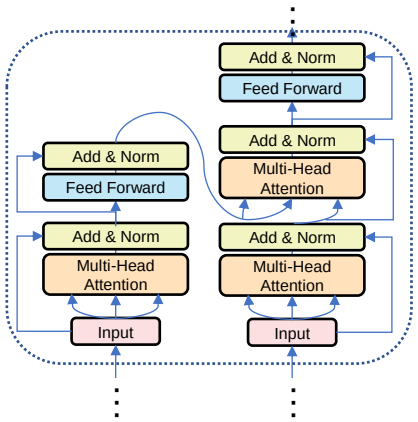


VGG[1]



DenseNet[2]

…

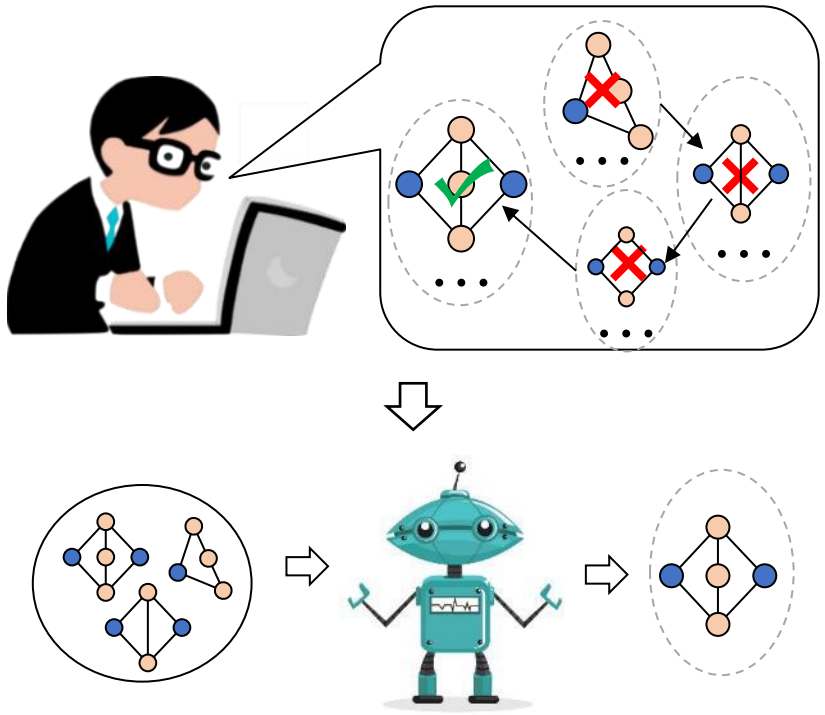


Transformer[3]

[1] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[2] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.
[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

Neural Architecture Search

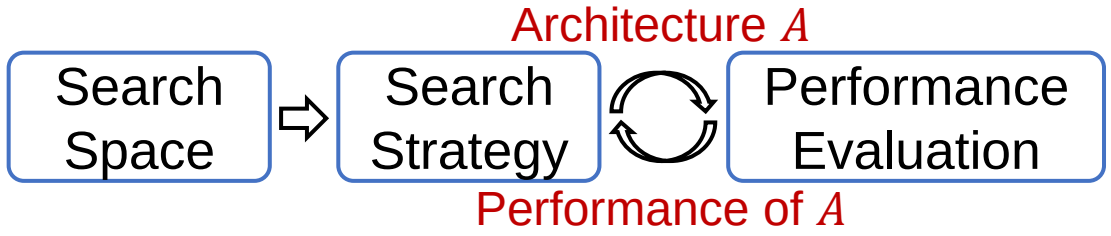
- To address the above problem, neural architecture search (NAS) is proposed to automatically design neural architectures



Handcraft V.S. Automatic design

$$\begin{cases} arg\ min\ A = \mathcal{L}(A, D_{train}, D_{valid}) \\ s.t.\ A \in \Omega \end{cases}$$

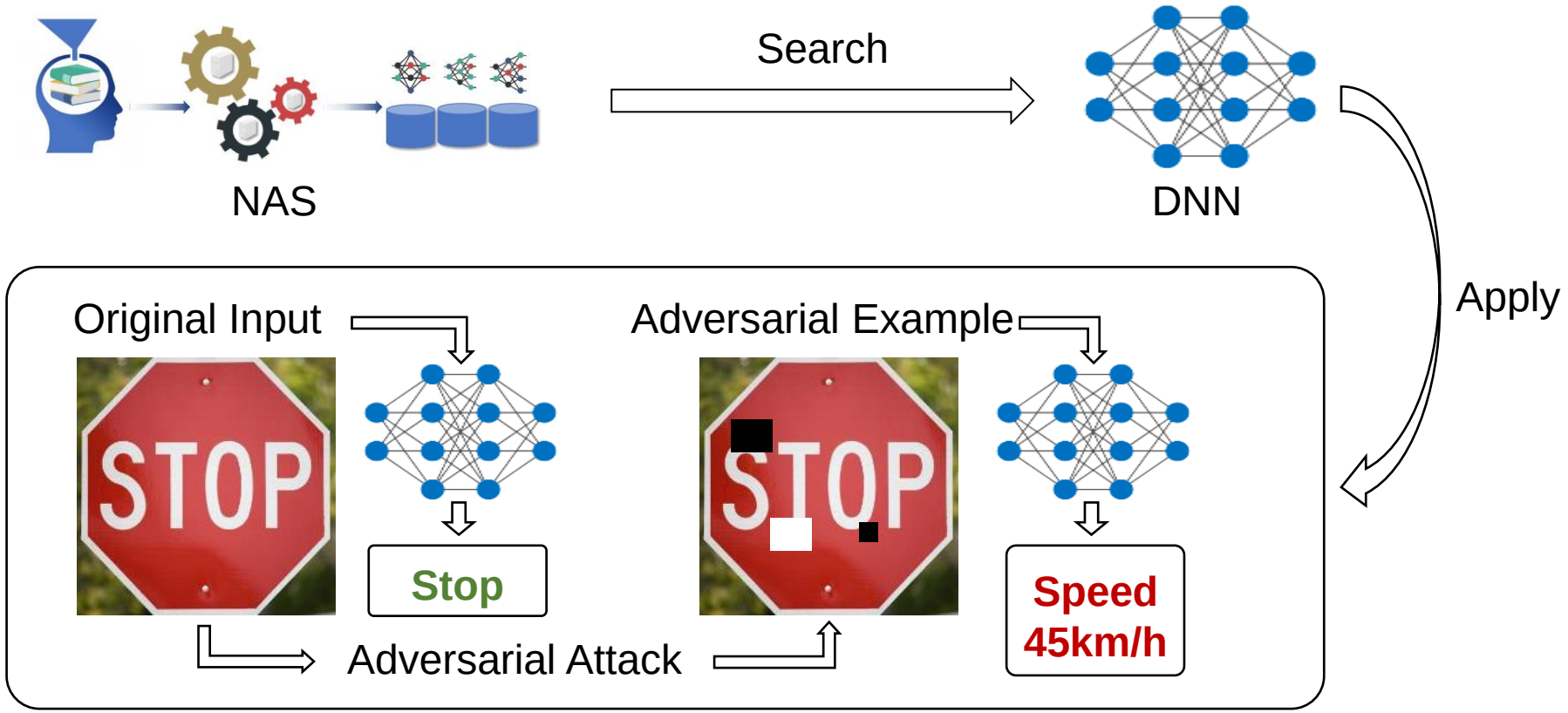
- Ω : search space
- $\mathcal{L}(\cdot)$: indicator for performance evaluation



Overall framework of NAS

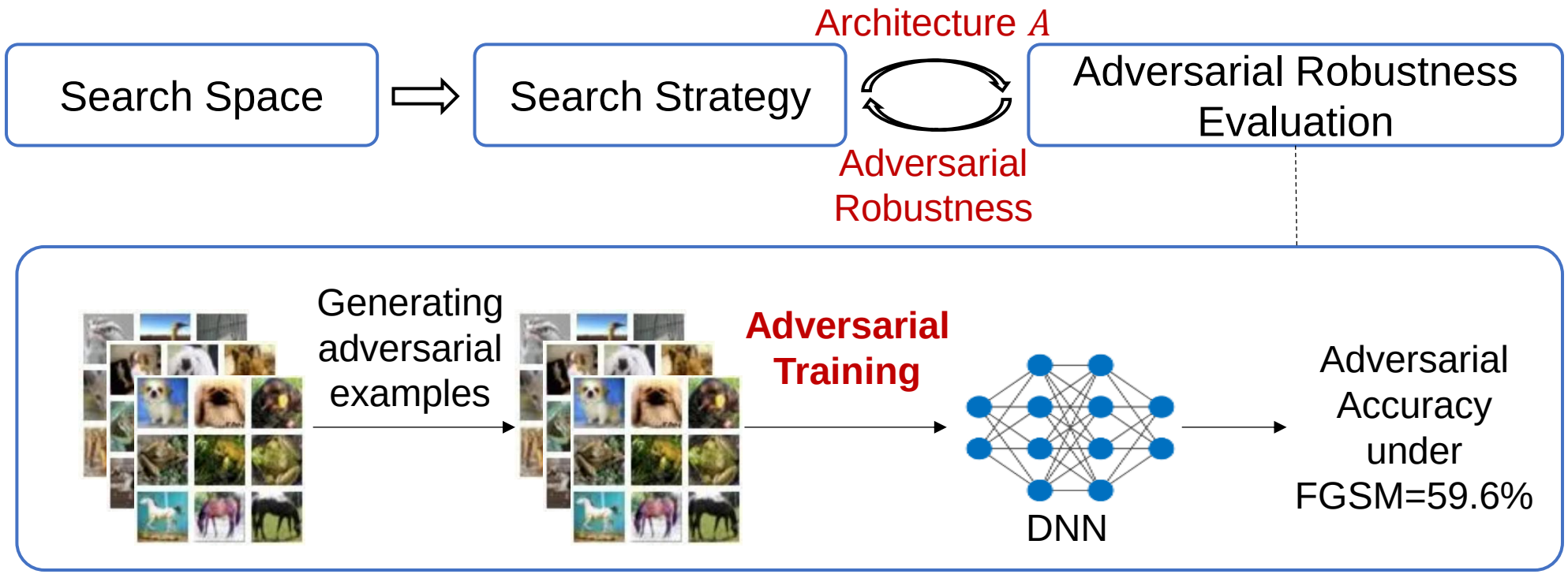
Adversarial Risks of NAS

- Despite the success of NAS, the architectures searched by NAS are vulnerable to adversarial attacks



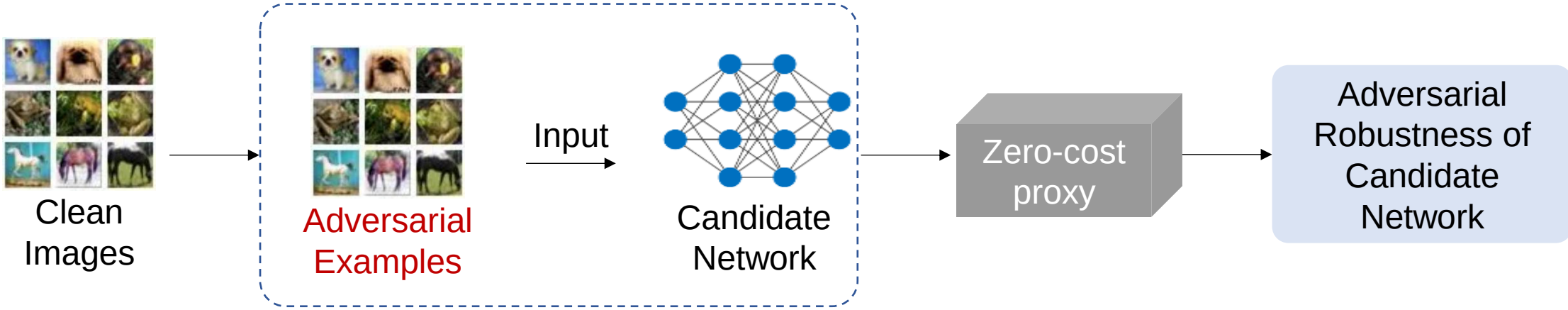
Robust NAS & Shortcomings

- Researchers propose robust NAS to address the above problem
- Evaluating adversarial robustness is time-consuming



Robust NAS & Shortcomings

- Existing studies introduce zero-cost proxies to evaluate adversarial robustness, but they still need to generate adversarial examples and lack of theoretical guarantee



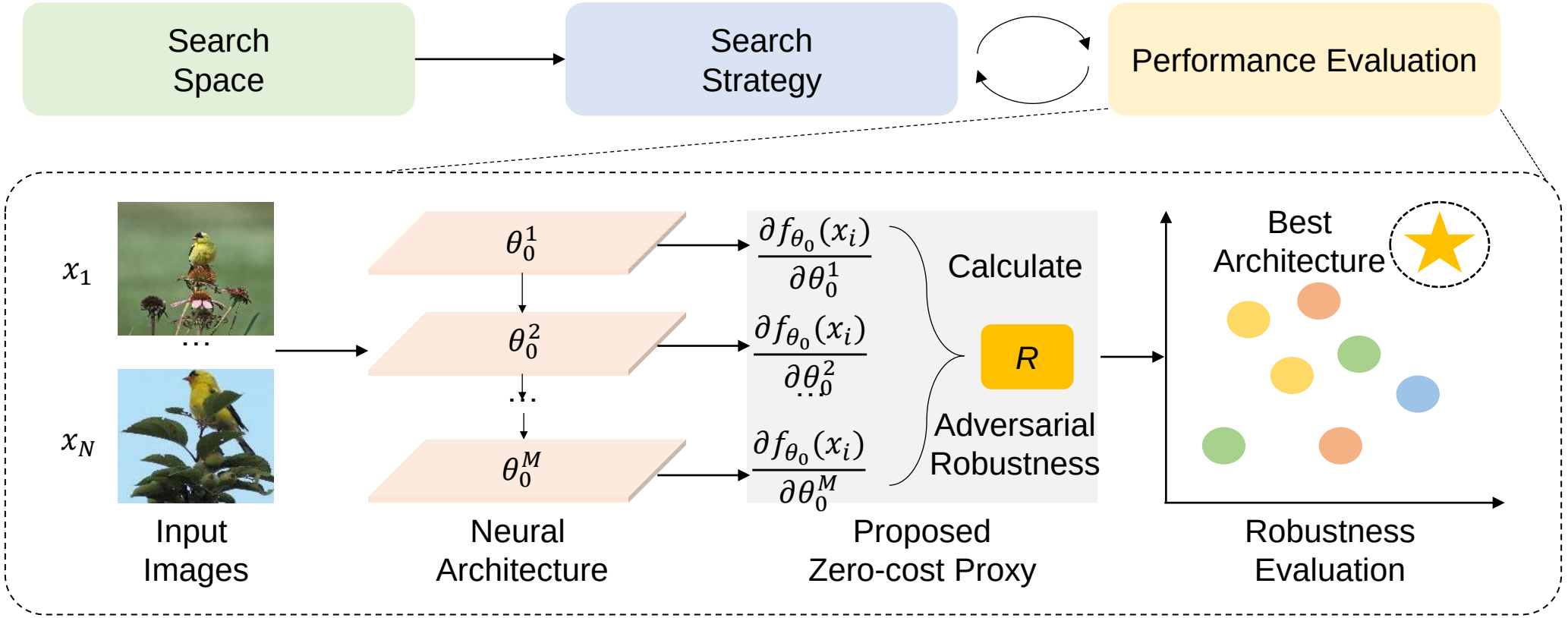
[1] Lukasik J, Moeller M, Keuper M. An evaluation of zero-cost proxies-from neural architecture performance prediction to model robustness[J]. International Journal of Computer Vision, 2024: 1-18.
[2] Ha H, Kim M, Hwang S J. Generalizable lightweight proxy for robust NAS against diverse perturbations[J]. Advances in Neural Information Processing Systems, 2023, 36: 38611-38623.

Outline

1. Introduction & Motivations
2. Methodology & Formulations
3. Dataset & Experiments

Our Zero-cost Proxy

- We introduce a novel zero-cost proxy based on the upper bound of adversarial loss



Formulations

- The formulation of the upper bound of adversarial loss:

$$\|y - f_{\theta_t}(\hat{x})\|_2^2 \leq \exp(-\lambda_{min}(\hat{\Theta}_{\theta_0})t)\|y - f_{\theta_0}(\hat{x})\|_2^2$$

$$\lambda_{min}(\Theta_{\theta_0}) = \frac{1}{MN^2}\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial f_{\theta_0}(x_i)}{\partial \theta_0^m}\right)\left(\frac{\partial f_{\theta_0}(x_j)}{\partial \theta_0^m}\right)^\top$$

$$\lambda_{max}(H_{\theta_0}(x)) \approx \left\|\frac{l(x+hz^*) - l(x)}{h}\right\|_2$$

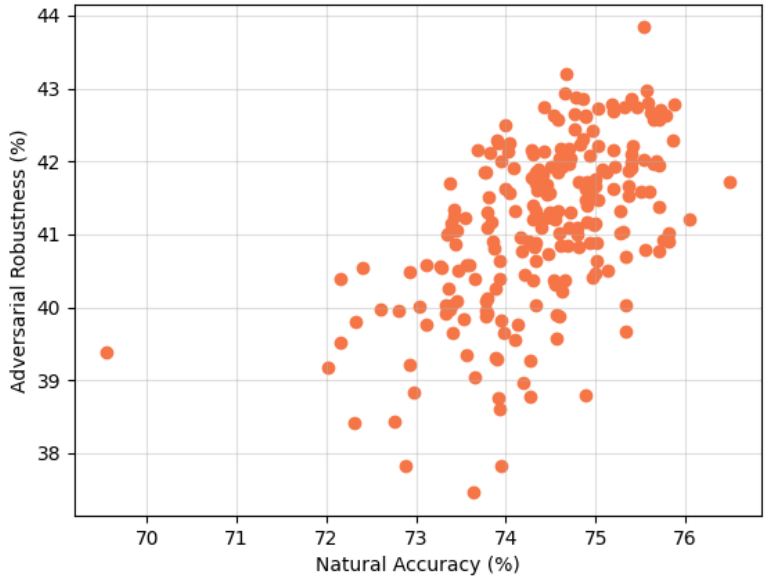
$$R = -\exp\left(\frac{1}{MN^2}\sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial f_{\theta_0}(x_i)}{\partial \theta_0^m}\right)\left(\frac{\partial f_{\theta_0}(x_j)}{\partial \theta_0^m}\right)^\top t\right) \times \left\|\frac{l(x+hz^*) - l(x)}{h}\right\|_2$$

Outline

1. Introduction & Motivations
2. Methodology & Formulations
3. Dataset & Experiments

Dataset

- We construct Tiny-RobustBench for better validation
- The architectures in the benchmark are adversarially trained, and the adversarial robustness is evaluated under stronger adversarial attack comparing with existing benchmark NAS-Bench-201-R



Items	Values
Total Training Epoch	105
Initial Learning Rate	0.1
Learning Rate Decay Policy	Stepped Decent
Learning Rate Decent Factor	0.1
The Index of Epoch for Learning Rate Decent	99
Momentum	0.9
Weight Decay	0.0001
Adversarial Loss	PGD
Perturbation Rate	8/255
Number of Steps	7
Step Size	0.01

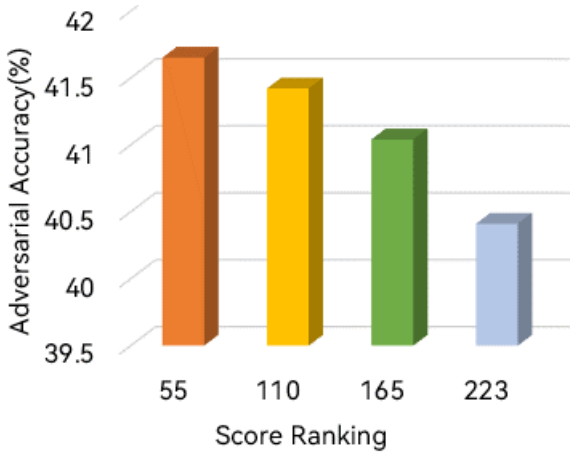
Experiments

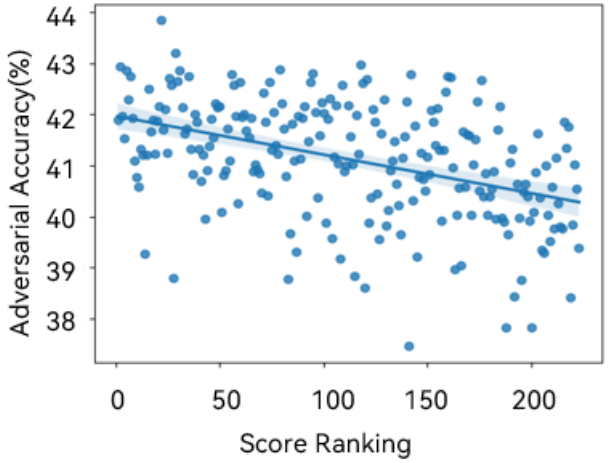
- The experiments under NAS settings demonstrates our proxy can significantly reduce the search cost while maintaining the performance

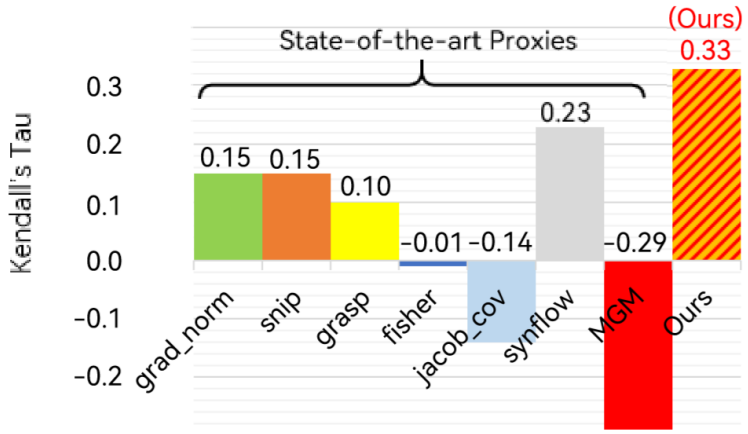
Category	Model	With Training?	Params (M)	FLOPs (M)	Natural Acc. (%)	FGSM (%)	PGD^{7†} (%)	PGD^{20} (%)	PGD^{100} (%)	APGD$_{CE}$ (%)	AA (%)	Search Cost (GPU Days)
Hand-Crafted	ResNet-18	×	11.2	37.67	84.09%	54.64%	-	45.86%	45.53%	44.54%	43.22%	-
	DenseNet-121	×	7.0	59.83	**85.95%**	58.46%	-	50.49%	49.92%	49.11%	47.46%	-
Standard NAS	DARTS	✓	3.3	547.44	85.17%	58.74%	-	50.45%	49.28%	48.32%	46.79%	1.0
	PDARTS	✓	3.4	550.75	85.37%	59.12%	-	51.32%	50.91%	49.96%	48.52%	0.3
	GradNorm†	×	4.7	-	81.61%	-	49.86%	-	-	-	46.69%	0.1
	SynFlow†	×	5.1	-	77.08%	-	45.95%	-	-	-	42.45%	0.1
Robust NAS	RobNet-free	✓	5.6	800.40	85.00%	59.22%	-	52.09%	51.14%	50.41%	48.56%	>3.3 *
	RACL	✓	3.6	568.86	83.97%	59.29%	-	52.18%	51.72%	51.24%	48.59%	0.5
	DSRNA	✓	2.0	336.23	80.93%	54.49%	-	49.11%	48.89%	48.54%	44.87%	0.4
	WsrNet	✓	3.0	484.30	83.94%	56.12%	-	47.17%	46.61%	-	43.91%	4.0
	CRoZe†	×	5.5	-	84.28%	-	52.17%	-	-	-	48.14%	0.2
	Ours	×	3.4	555.54	85.60%	**60.20%**	**69.21%**	**52.75%**	**52.51%**	**52.25%**	**49.97%**	**0.017**

Experiments

- Our zero-cost proxy is effective in evaluating adversarial robustness comparing with existing zero-cost proxies







Thanks!