Robust Principles

Architectural Design Principles for Adversarially Robust CNNs

Anthony Peng, Weilin Xu, Cory Cornelius, Matthew Hull, Kevin Li, Rahul Duggal, Mansi Phute, Jason Martin, Polo Chau

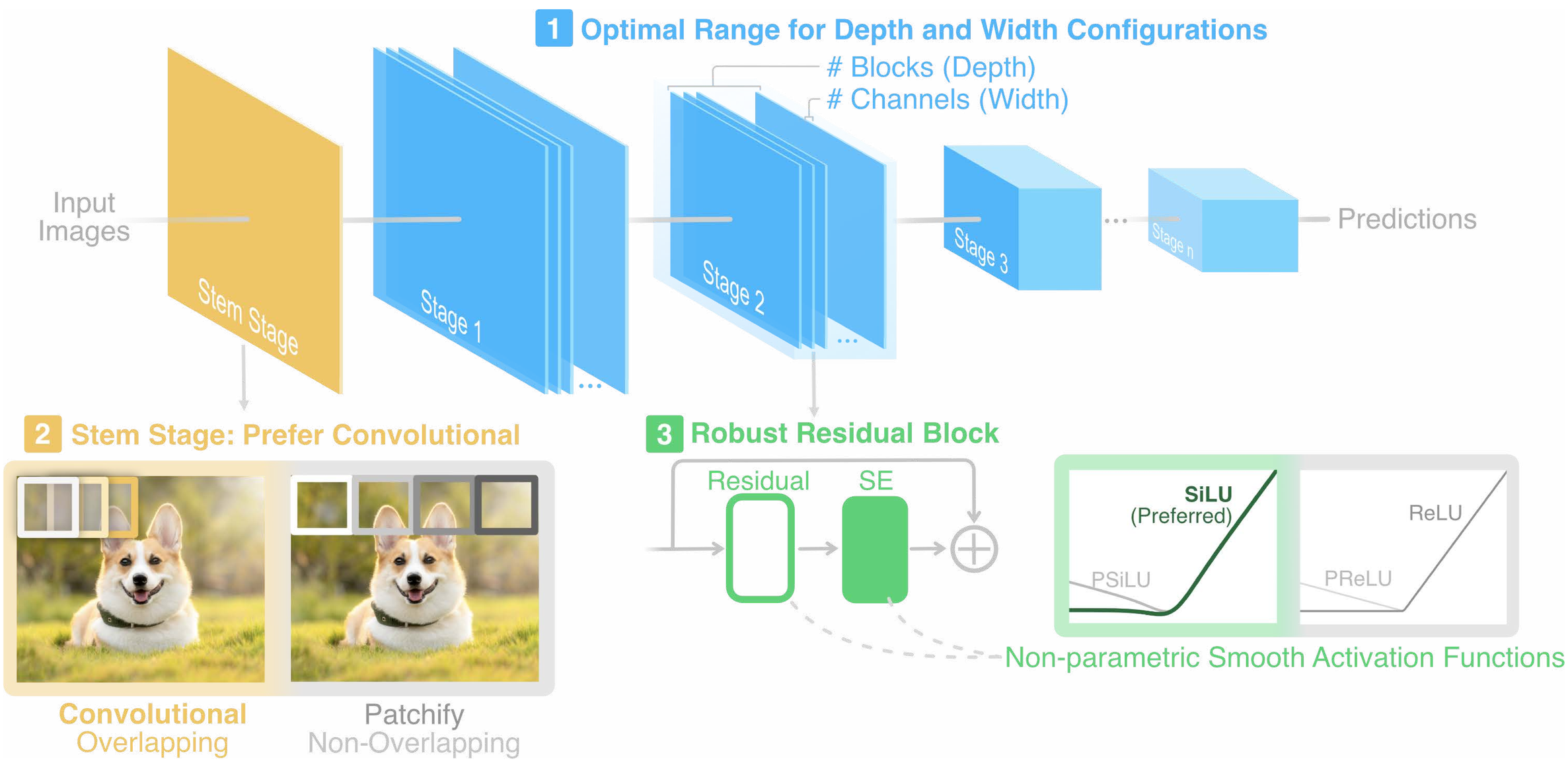
Georgia Tech, intel

#1 on RobustBench CIFAR-10 Leaderboard

		Clean	AutoAttack
1	Robust Principles: Architectural Design Principles for Adversarially Robust CNNs *It uses additional 50M synthetic images in training.*	93.27%	71.07%

Get the Code!

github.com/poloclub/robust-principles



1 Optimal Width-Depth Ratios

2 Convolutional over Patchify Stem Stage

Configuration	Clean (%)	PGD (%)
ResNet-50 (baseline)	56.05	30.59
Conv. Stem	57.08 +1.03	33.08 +2.49
Patch 4, Stride 4	55.40 -0.65	31.68 +1.09
Patch 2, Stride 2	56.38 +0.33	31.91 +1.32

3 Robust Residual Block Design

Configuration	Clean (%)	PGD (%)
ResNet-50 (baseline)	56.05	30.59
SE (r = 4)	57.83 +1.78	32.64 +2.05
PReLU	55.81 -0.27	30.38 -0.21
SiLU	58.19 +2.14	34.07 +3.48
PSiLU	56.38 +0.33	33.76 +3.17

Applying all 3 design principles provides consistent robustness improvements

Dataset	AT method	# Param.	Model	Clean Acc. (%)	Adv. Acc. (%)
CIFAR-10	Diffusion 50M	267M	WRN-70-16	93.25	70.69
			Ra WRN-70-16	93.27 **+0.02**	71.07 **+0.38**
CIFAR-100	Diffusion 1M	267M	WRN-70-16	70.10	37.77
			Ra WRN-70-16	71.42 **+1.32**	38.89 **+1.12**
ImageNet	SAT	46M	ResNet-101	67.06	39.78
			Ra ResNet-101	71.88 **+4.82**	46.26 **+6.48**

Fist icon created by Athok