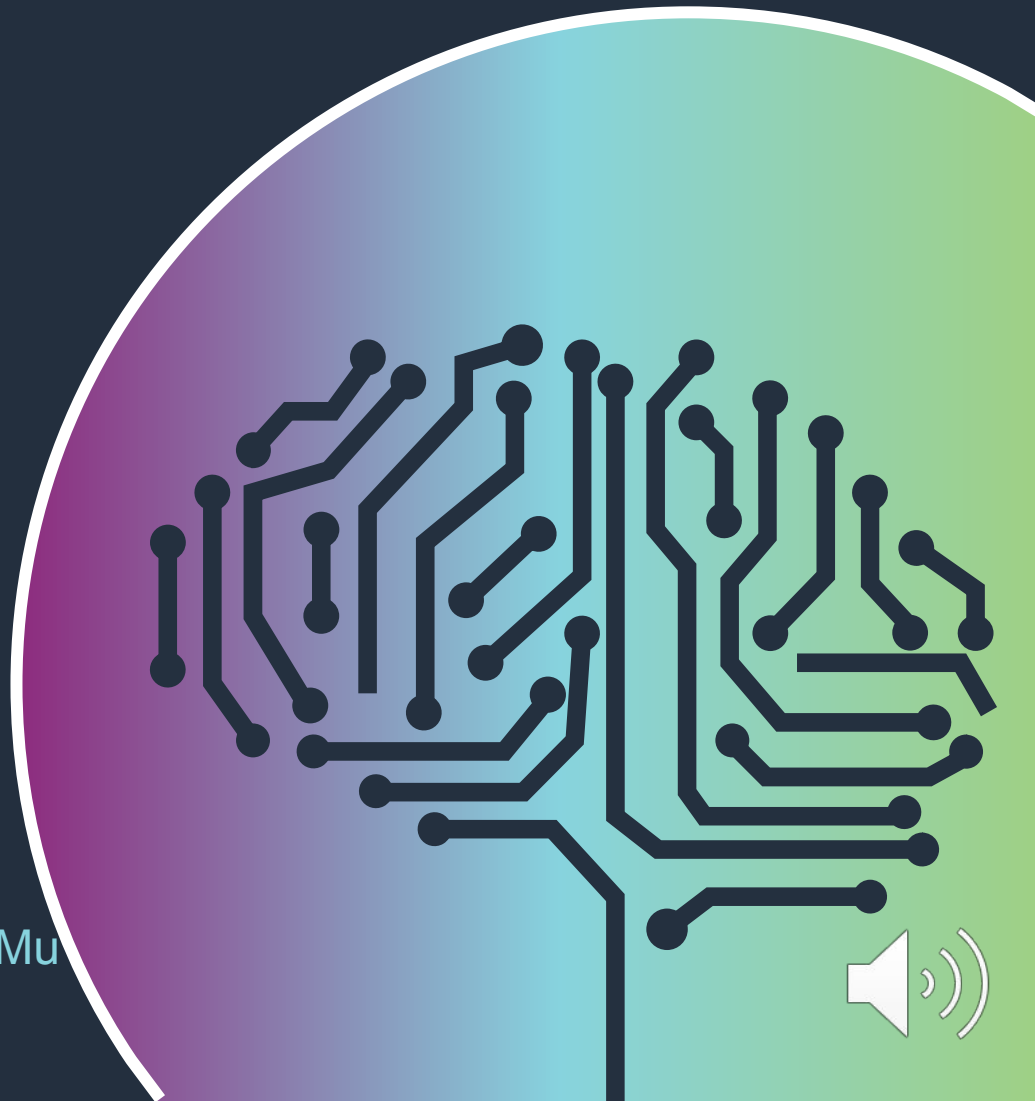
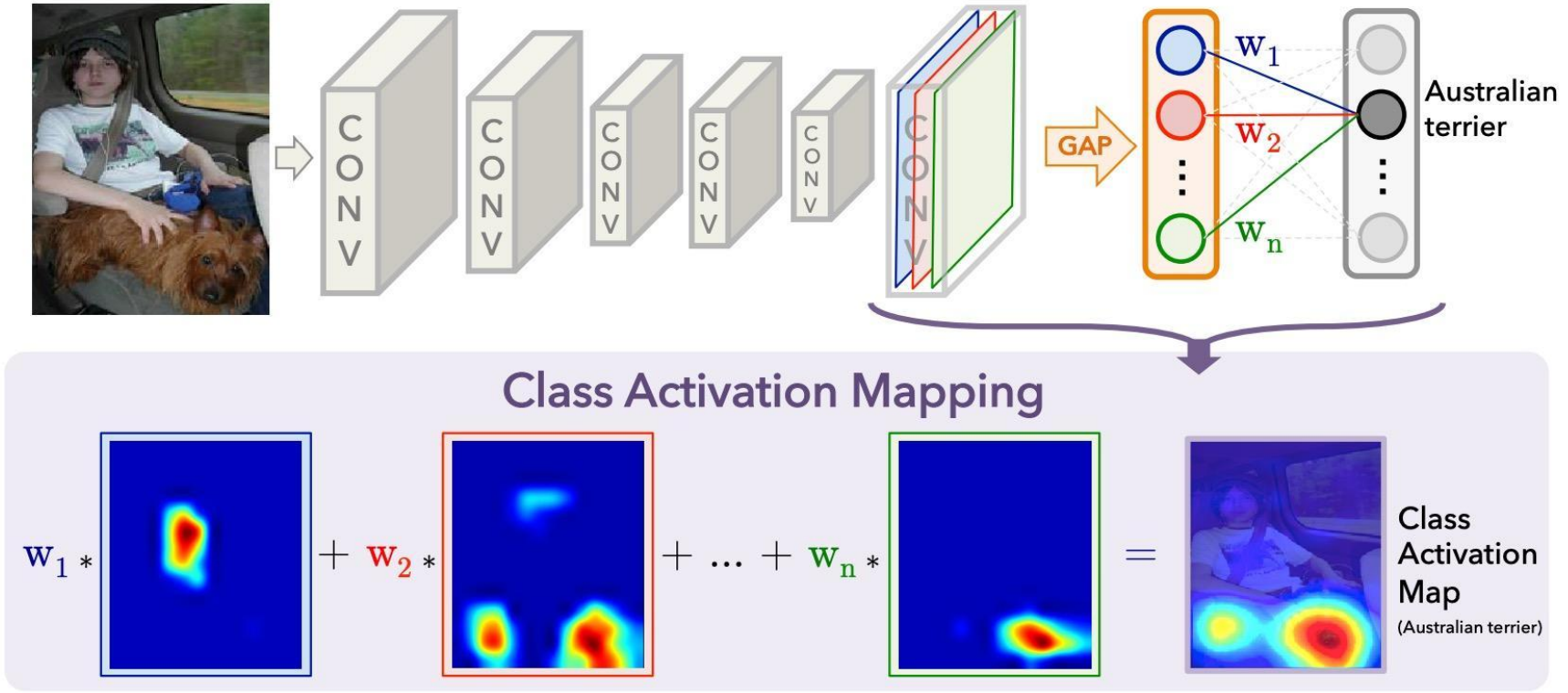
ResNeSt: Split-Attention Networks

Hang Zhang Applied Scientist, Amazon Lab 126

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, Alex Smola



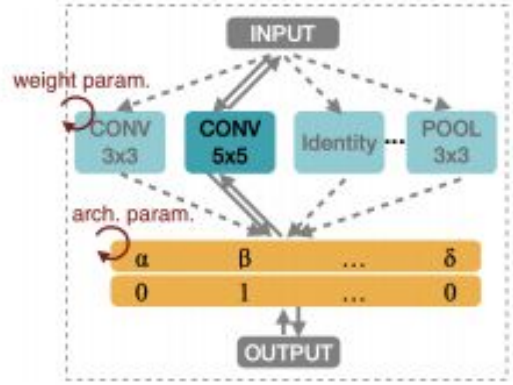
Backbone Matters -- Representation Learning

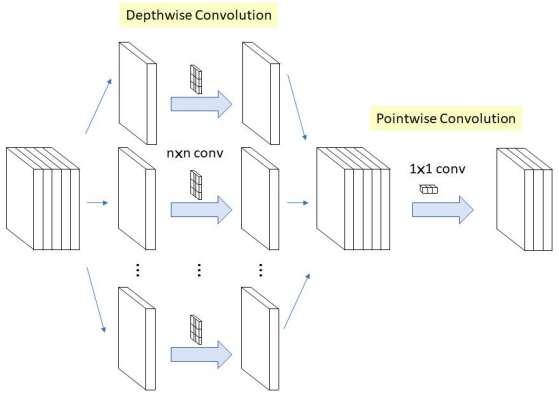


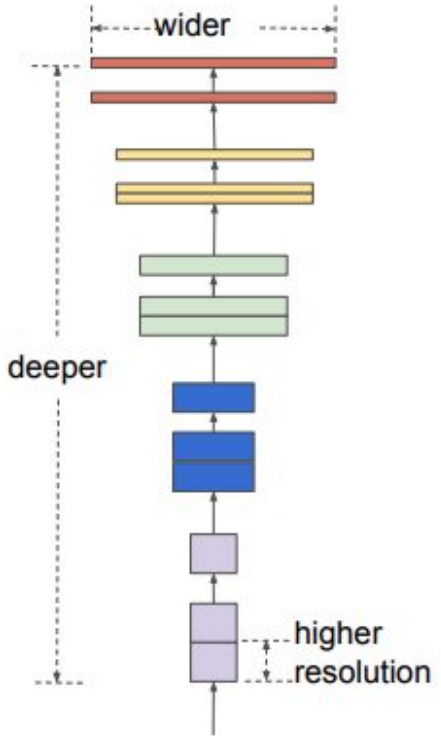
Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Recent CNN for Image Classification

- SoTA using Neural Architecture Search (NAS)
- Tailored to one task
- Hard to transfer
  - Meta architecture change
- Not necessarily low latency



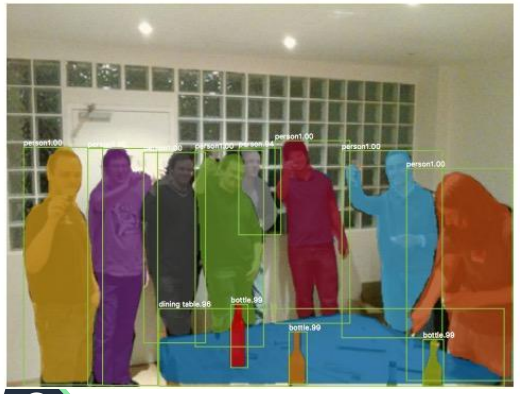




Input shape	Block	f	n	s
$224^2 \times 3$	3x3 conv	16	1	2
$112^2 \times 16$	TBS	16	1	1
$112^2 \times 16$	TBS	24	4	2
$56^2 \times 24$	TBS	32	4	2
$28^2 \times 32$	TBS	64	4	2
$14^2 \times 64$	TBS	112	4	1
$14^2 \times 112$	TBS	184	4	2
$7^2 \times 184$	TBS	352	1	1
$7^2 \times 352$	1x1 conv	1504 (1984)	1	1
$7^2 \times 1504$ (1984)	7x7 avgpool	-	1	1
1504	fc	1000	1	-

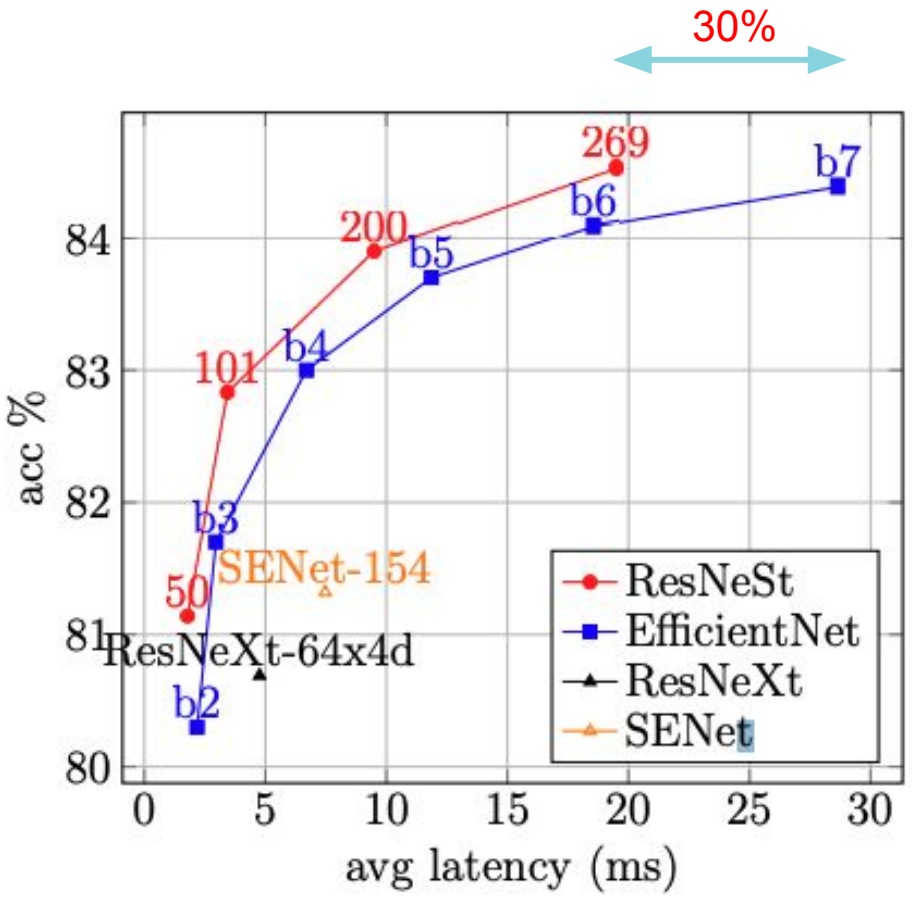
ResNet Still Servers as the Backbone

- Object detection (Faster-RCNN)
- Instance segmentation (Mask-RCNN)
- Pose estimation (Alpha-pose)
- Semantic Segmentation (DeepLabV3)



Introduction

- ResNeSt a new ResNet variant
  - A simple and modular network
    - feature-map Split
    - attention
- State-of-the-art performance for
  - Image classification
  - Object detection
  - Instance Segmentation
  - Semantic segmentation
  - …



State of the Art | Instance Segmentation on COCO test-dev
State of the Art | Object Detection on COCO test-dev
State of the Art | Panoptic Segmentation on COCO panoptic
State of the Art | Semantic Segmentation on ADE20K
State of the Art | Semantic Segmentation on Cityscapes val
State of the Art | Semantic Segmentation on PASCAL Context

AMLC VIRTUAL CONFERENCE

SoTA Results

Method	Backbone	Dataset	Metric	Score (-sota)
Cascade R-CNN	ResNeSt-200 (ours)	MS-COCO	bbox mAP	53.30 (+0.0)
			Mask mAP	47.10 (+0.8)
Panoptic FCN	ResNeSt-200 (ours)	MS-COCO	PQ	47.90 (+4.9)
Deeplab-V3	ResNeSt-200 (ours)	ADE20K	mIoU	48.36 (+2.1)
		Cityscapes	mIoU	82.7 (+1.2)
	ResNeSt-269 (ours)	Pascal Context	mIoU	58.9 (+2.7)

Results from https://paperswithcode.com/ on 05/30/2020.

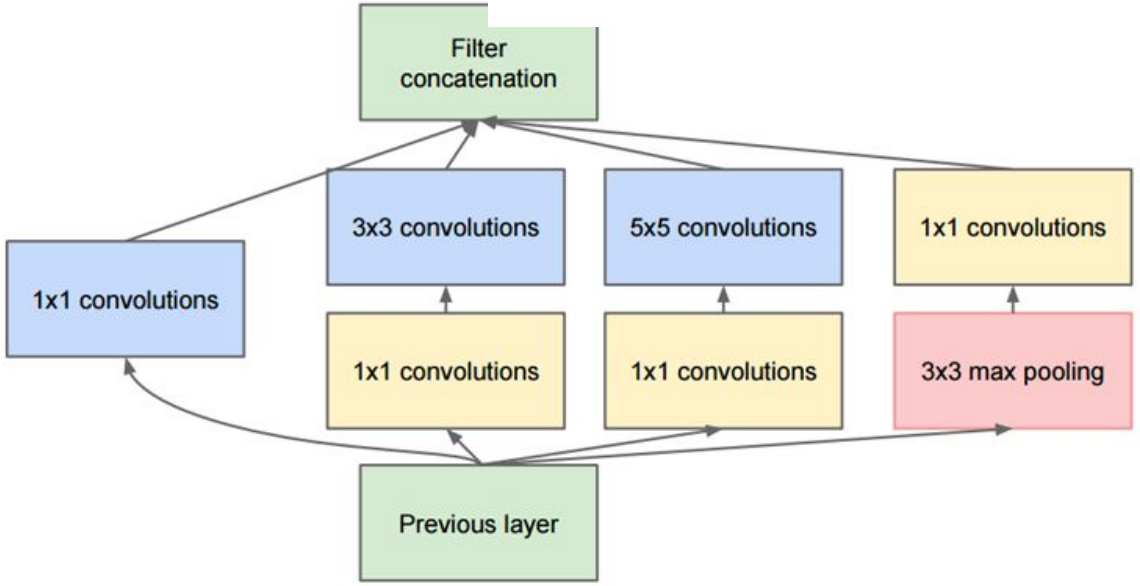
LVIS winner [Talk]	Team LvisTraveler
LVIS most innovative [Talk]	Team Asynchronous SSL
LVIS spotlight [Talk][Video]	Team MMDet

COCO + LVIS Challenge at ECCV 2020

Multi-branch Network Representation

- GoogNet/Inception (1x1, GAP)

$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot (1 - T(\mathbf{x}, \mathbf{W_T})).$$

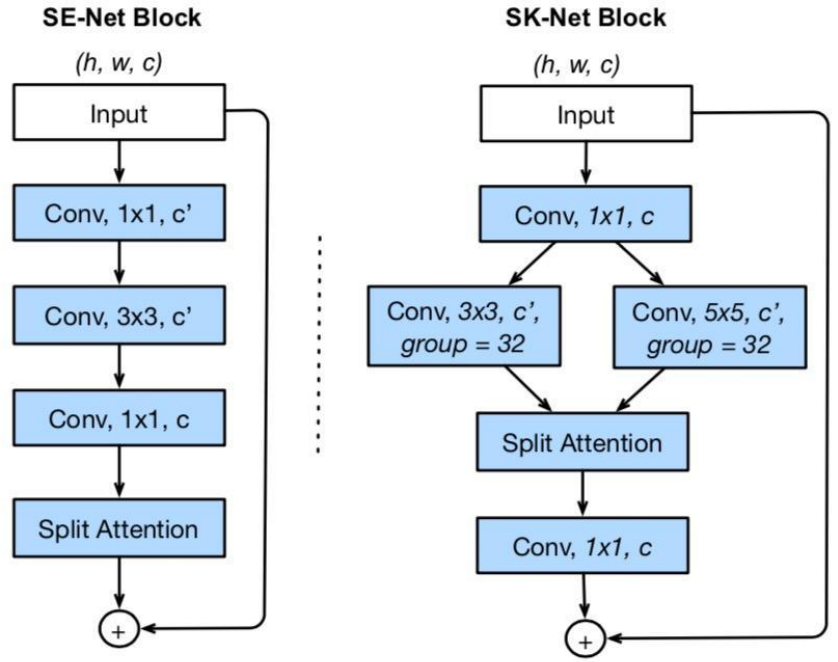


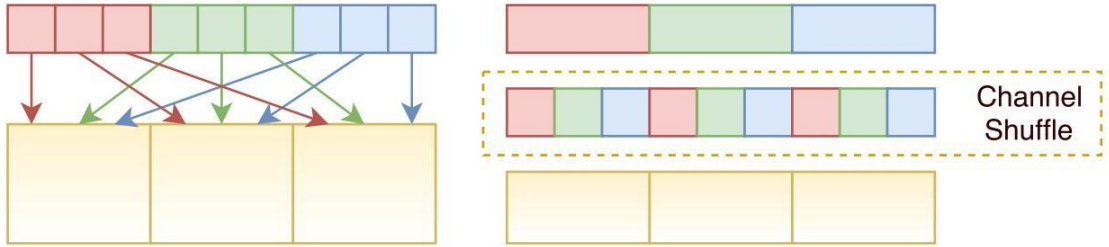
Full Inception module

Szegedy, Christian, et al. "Going deeper with convolutions."

Featuremap Attention

- NIN
- ShuffleNet (group relationship)
- SE-Net
- SK-Net (3x3, 5x5 selection)





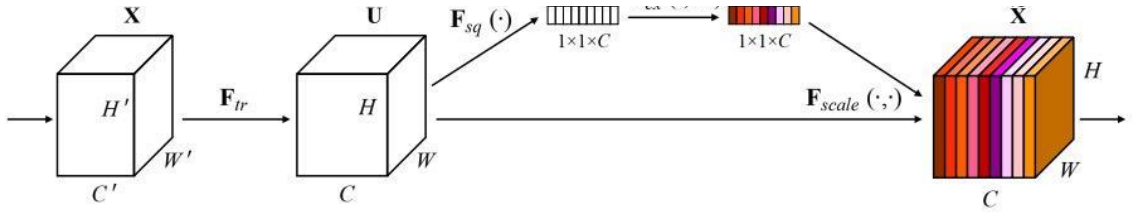
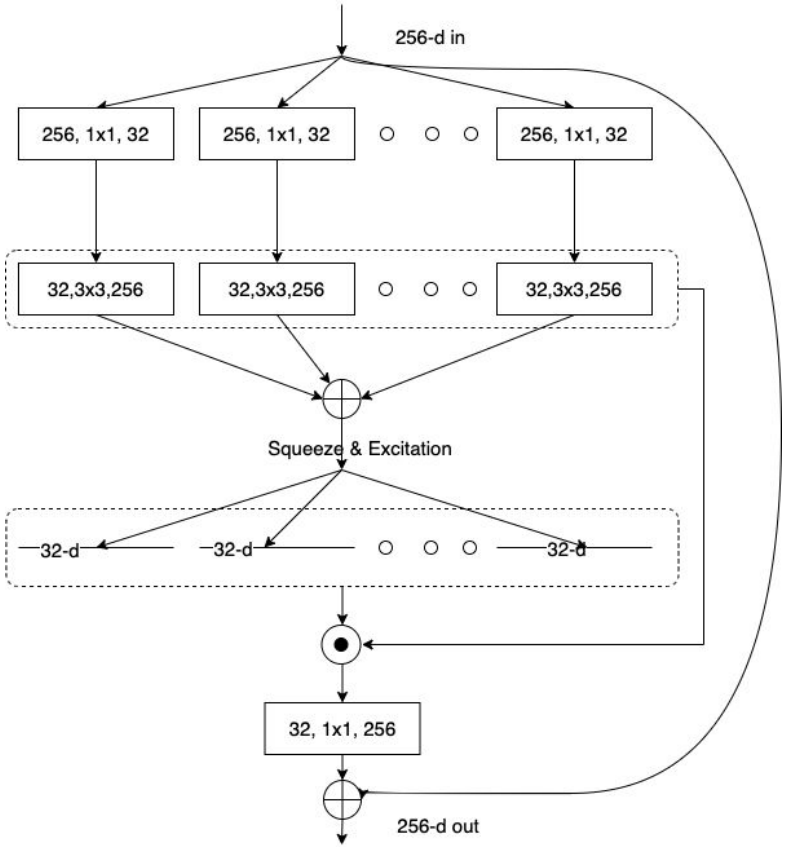


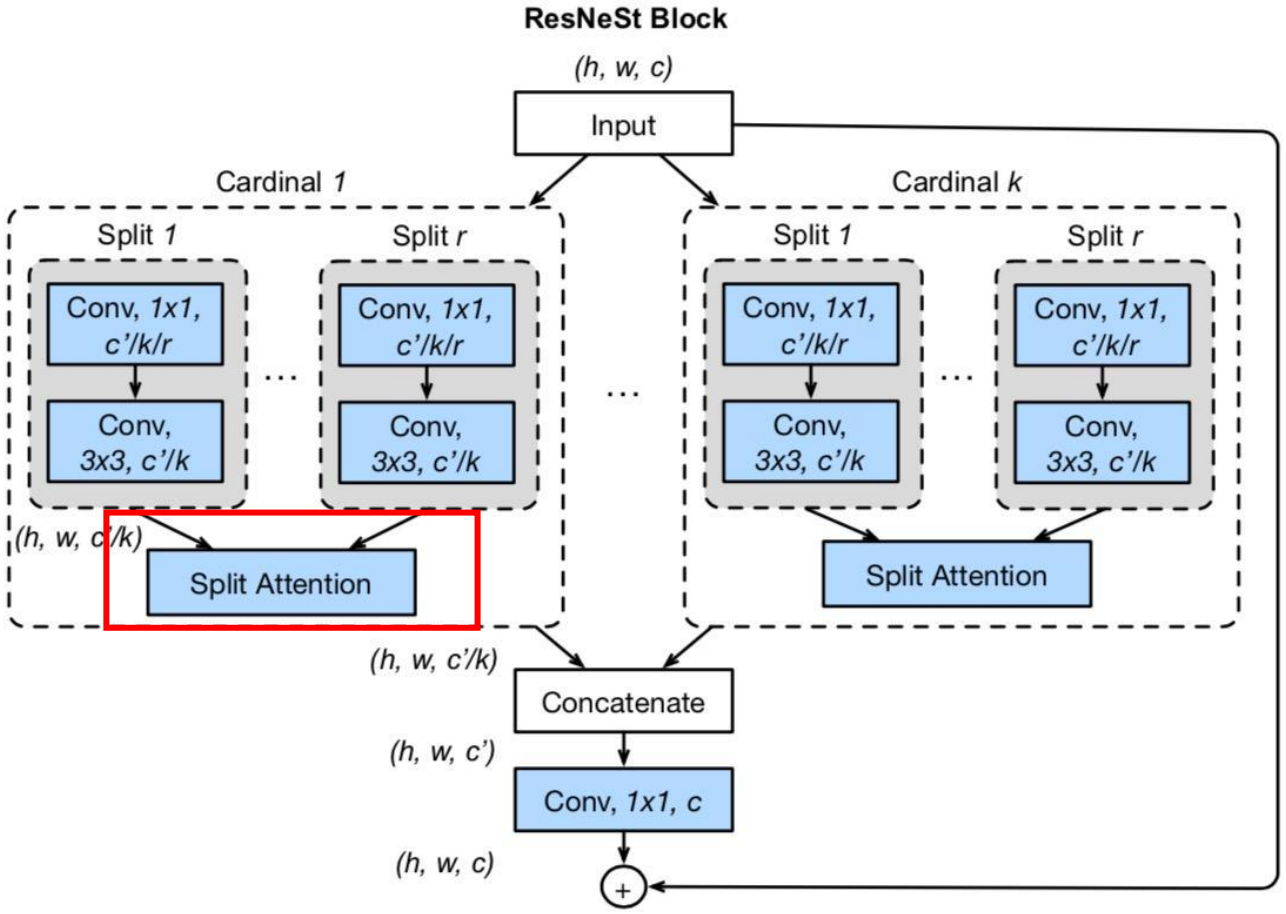
Figure 1: A Squeeze-and-Excitation block.

Multi-branch v.s. Channel-attention

- Diverse representation v.s. feature correlation
- SE-ResNeXt:
  - SE-Module + ResNeXt
- Split-Attention Network



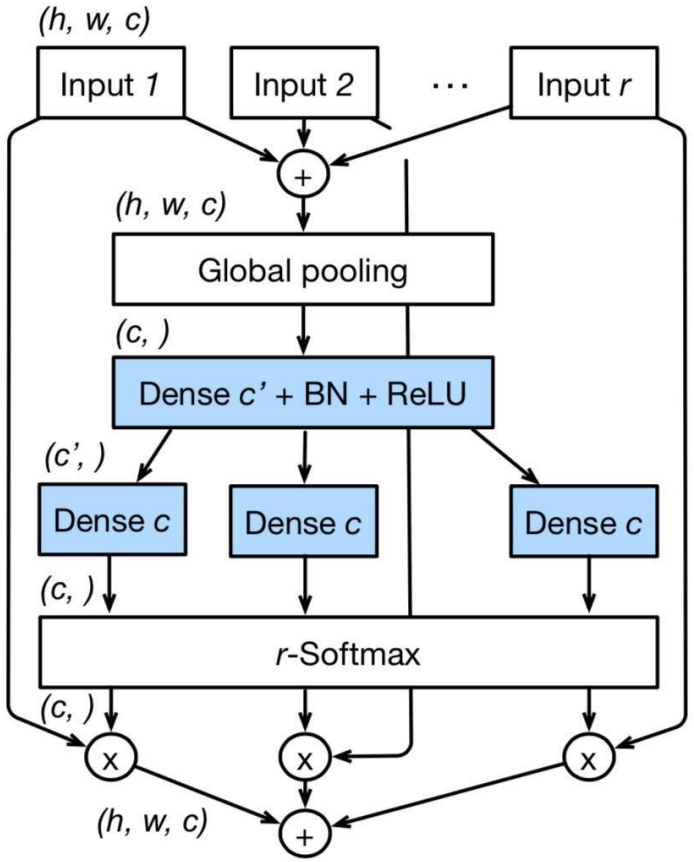
Split-Attention Block



Channel-attention in Featuremap Group

- Cardinality K (as in ResNeXt)
- Radix R (# splits in each cardinal group)
- # Groups $G = KR$
- $U[k, r] = \mathcal{F}_{k,r}(X)$

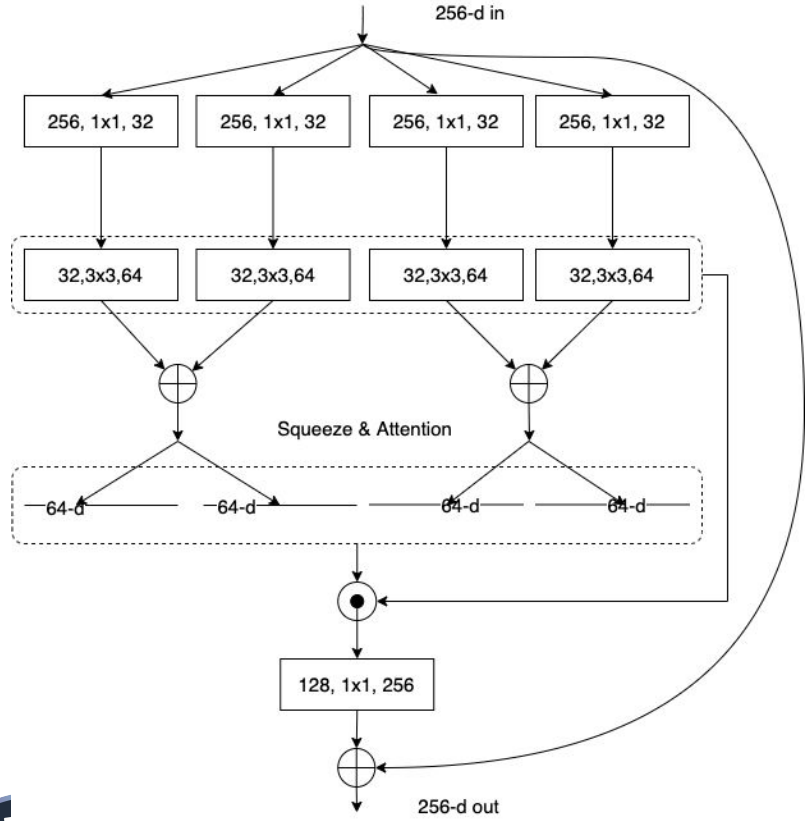
Split-Attention within Cardinal Groups

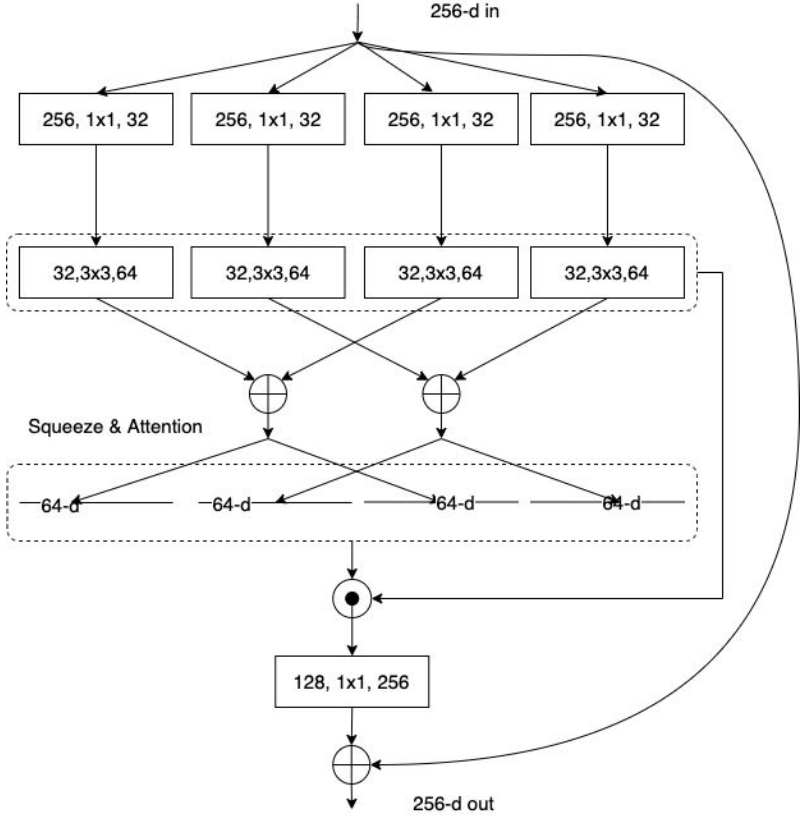


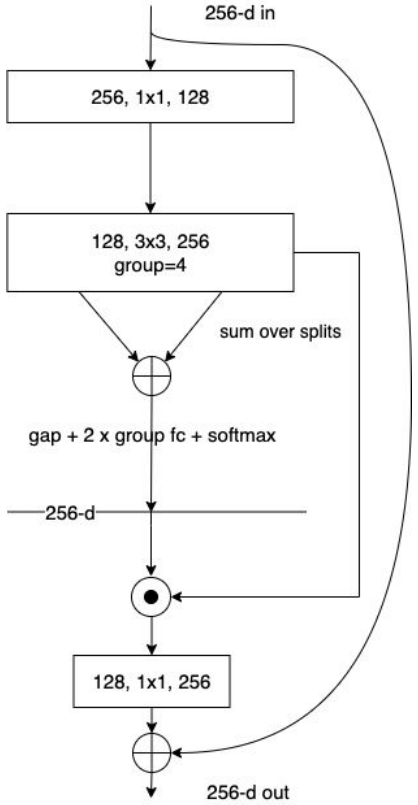
- Fuse cardinal representation: $\widehat{U}_k = \sum_{r=1} U[k,r]$
- Squeeze global context: $s^k = GAP(\widehat{U}^k)$
- Split-Attention:$V[k] = \sum_{r=1}^{R} a_{k,r} \otimes U[k,r]$
- The attention weights are given:

$$a_{k,r}^{c} = \begin{cases} softmax\left(\mathcal{G}_{k,r}^{c}(s^k)\right) \ if \ R > 1 \\ sigmoid\left(\mathcal{G}_{k,r}^{c}(s^k)\right) \ if \ R = 1 \end{cases}$$

Modularization and Acceleration







Ablation Study on Image Classification

- Improvement break down

	#P	GFLOPs	acc(%)
ResNetD-50 [27]	25.6M	4.34	78.31
+ mixup	25.6M	4.34	79.15
+ autoaug	25.6M	4.34	79.41
ResNeSt-50-fast	27.5M	4.34	80.64
ResNeSt-50	27.5M	5.39	81.13

- Radix and cardinality

ResNet-D

Variant	#P	GFLOPs	img/sec	acc(%)
0s1x64d	25.6M	4.34	688.2	79.41
1s1x64d	26.3M	4.34	617.6	80.35
2s1x64d	27.5M	4.34	533.0	80.64
4s1x64d	31.9M	4.35	458.3	80.90
2s2x40d	26.9M	4.38	481.8	81.00

split

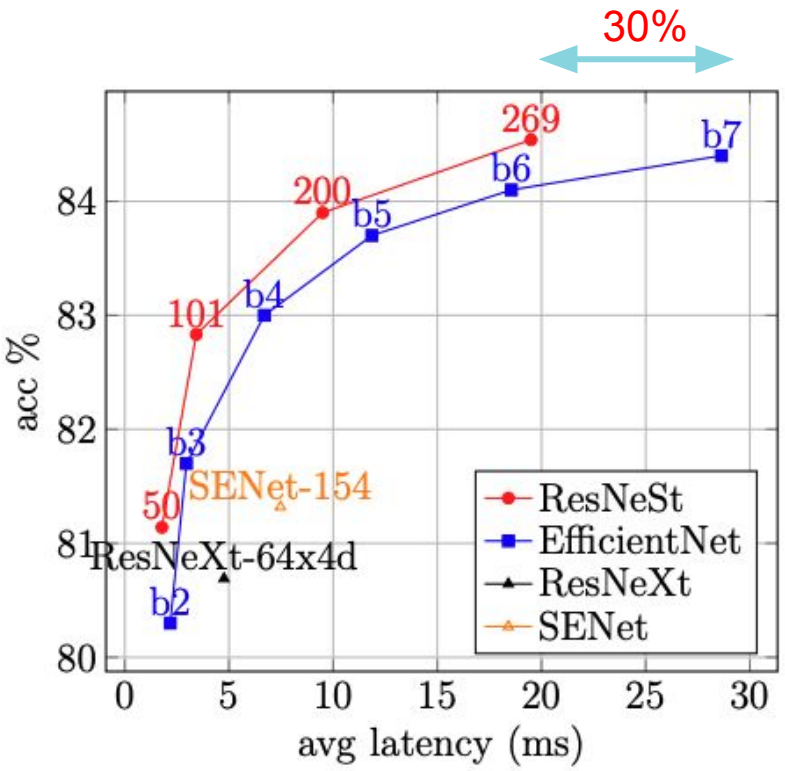
cardinalit

width (= channels)

used in subsequent experiments

He, Tong, et al. "Bag of tricks for image classification with convolutional neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Compare with SoTA



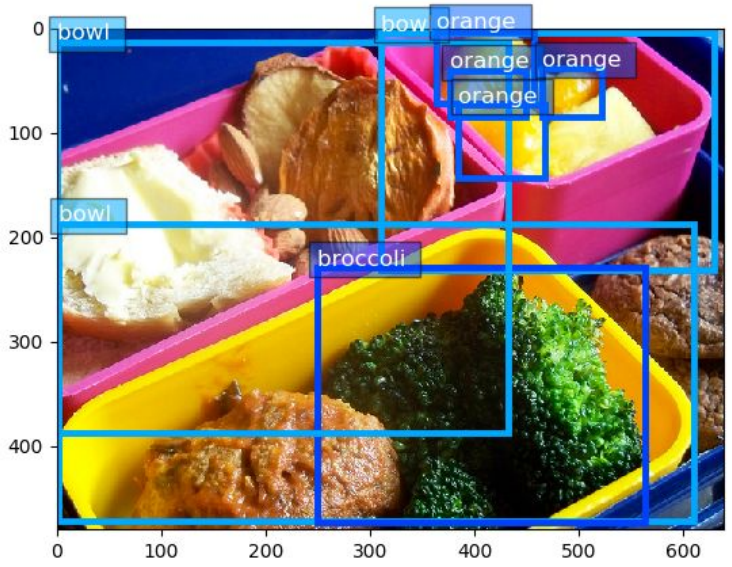
	#P	crop	img/sec	acc(%)
ResNeSt-101(ours)	48M	256	**291.3**	**83.0**
EfficientNet-B4 [55]	19M	380	149.3	83.0
SENet-154 [29]	146M	320	133.8	82.7
NASNet-A [74]	89M	331	103.3	82.7
AmoebaNet-A [45]	87M	299	-	82.8
ResNeSt-200 (ours)	70M	320	**105.3**	**83.9**
EfficientNet-B5 [55]	30M	456	84.3	83.7
AmoebaNet-C [45]	155M	299	-	83.5
ResNeSt-269 (ours)	111M	416	**51.2**	**84.5**
GPipe	557M	-	-	84.3
EfficientNet-B7 [55]	66M	600	34.9	84.4

Table 4: SoTA results comparison on ImageNet with CNN models using large crop sizes. Our ResNeSt displays the best trade-off between accuracy and latency.

MS-COCO Dataset

- COCO-2017:
  118k/5k train/val

- Object Detection
- Instance Segmentation



Results on Object Detection

	Method	Backbone	mAP%
Prior Work	Faster-RCNN [46]	ResNet101 [22]	37.3
		ResNeXt101 [5, 60]	40.1
		SE-ResNet101 [29]	41.9
	Faster-RCNN+DCN [12]	ResNet101 [5]	42.1
	Cascade-RCNN [2]	ResNet101	42.8
Our Results	Faster-RCNN [46]	ResNet50 [57]	39.25
		ResNet101 [57]	41.37
		ResNeSt50 (ours)	42.33
		ResNeSt101 (ours)	**44.72**
	Cascade-RCNN [2]	ResNet50 [57]	42.52
		ResNet101 [57]	44.03
		ResNeSt50 (ours)	45.41
		ResNeSt101 (ours)	**47.50**
	Cascade-RCNN [2]	ResNeSt200 (ours)	49.03

- MS-COCO validation set.
- The performance of Faster-RCNN and Cascade-RCNN are significantly improved by our ResNeSt backbone.
- Notably, our ResNeSt50 often outperforms ResNet101

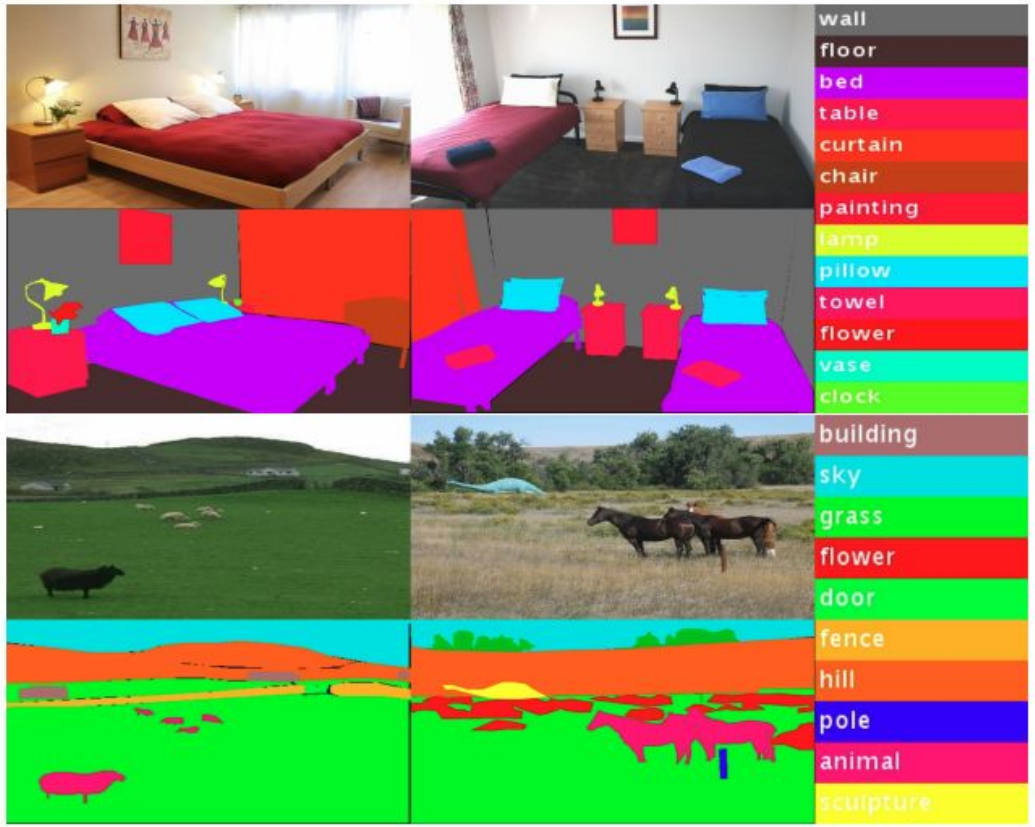
Results on Instance Segementation

	Method	Backbone	box mAP%	mask mAP%
Prior Work	DCV-V2 [72]	ResNet50	42.7	37.0
	HTC [5]	ResNet50	43.2	38.0
	Mask-RCNN [23]	ResNet101 [6]	39.9	36.1
	Cascade-RCNN [4]	ResNet101	44.8	38.0
Our Results	Mask-RCNN [23]	ResNet50 [57]	39.97	36.05
		ResNet101 [57]	41.78	37.51
		ResNeSt50 (ours)	42.81	38.14
		ResNeSt101 (ours)	**45.75**	**40.65**
	Cascade-RCNN [3]	ResNet50 [57]	43.06	37.19
		ResNet101 [57]	44.79	38.52
		ResNeSt50 (ours)	46.19	39.55
		ResNeSt101 (ours)	**48.30**	**41.56**

- MS-COCO validation set.
- Both Mask-RCNN and Cascade-RCNN models are improved by our ResNeSt backbone.
- Models with our ResNeSt-101 outperform all prior work using ResNet-101.

Semantic Segmentation

- ADE20K Dataset:
  - 150 object categories
  - 20K/2K train/val
- Cityscapes Dataset:
  - 19 object categories
  - 2,975/500 train/val



Results on Semantic Segmentation

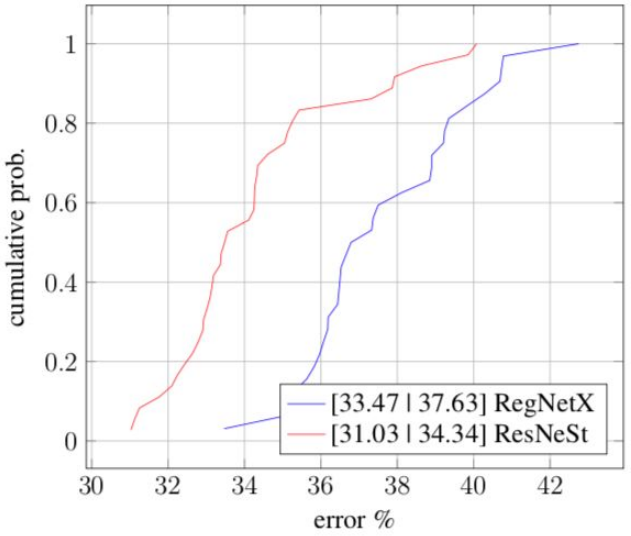
	Method	Backbone	pixAcc%	mIoU%
Prior Work	UperNet [59]	ResNet101	81.01	42.66
	PSPNet [69]	ResNet101	81.39	43.29
	EncNet [65]	ResNet101	81.69	44.65
	CFNet [66]	ResNet101	81.57	44.89
	OCNet [63]	ResNet101	-	45.45
	ACNet [17]	ResNet101	81.96	45.90
Ours	DeeplabV3 [7]	ResNet50 [21]	80.39	42.1
		ResNet101 [21]	81.11	44.14
		ResNeSt-50 (ours)	81.17	45.12
		ResNeSt-101 (ours)	**82.07**	**46.91**
		ResNeSt-200 (ours)	82.45	48.36

- ADE20K Validation Set
- Outperform previous best single model by 2.4% mIoU

Conclusion and Extra thoughts

- ResNeSt, a SoTA CNN model, an universal backbone
- The backbone improvement directly boost downstream applications
- Augment the network design spaces (NAS)

- The SoTA on ImageNet is a comprehensive competition:
  - Network + training strategy (new standard)
  - Devils into the detail (see paper appendix)



Q&A

- Link to GitHub:
  - https://github.com/zhanghang1989/ResNeSt
- Detectron2 Models:
  - https://github.com/zhanghang1989/detectron2-ResNeSt
- 3rd Party TensorFlow & Caffe Implementations:
  - https://github.com/QiaoranC/tf_ResNeSt_RegNet_model
  - https://github.com/NetEase-GameAI/ResNeSt-caffe