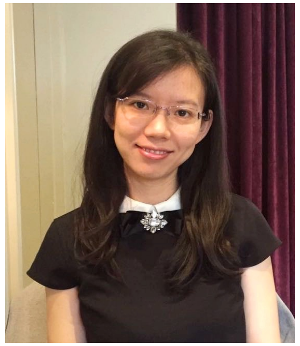
Squeeze-and-Excitation Networks

Jie Hu[1,*]



Li Shen[2,*]

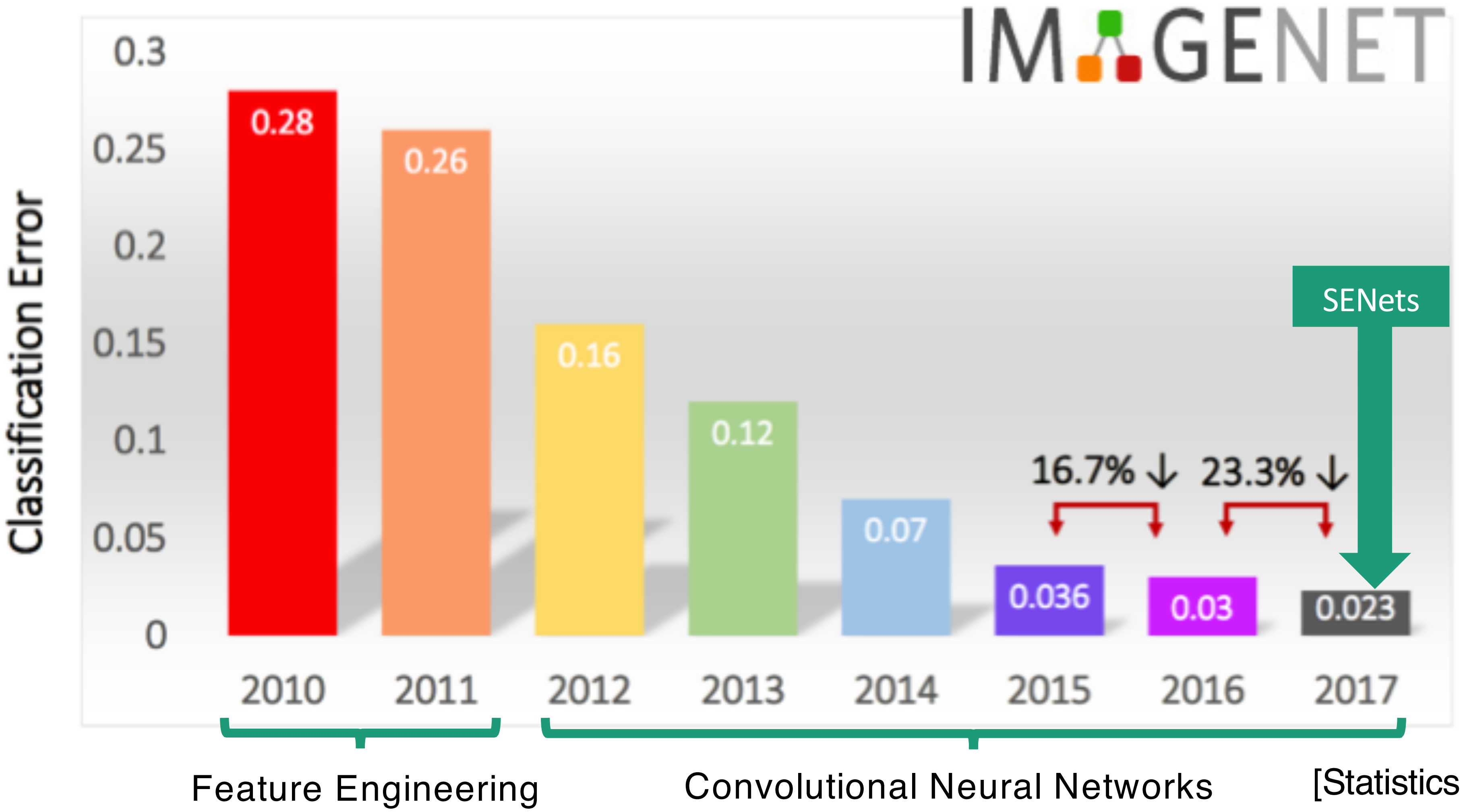
Gang Sun[1]

[1] Momenta

[2] Department of Engineering Science, University of Oxford

Large Scale Visual Recognition Challenge

Squeeze-and-Excitation Networks (SENets) formed the foundation of our winner entry on ILSVRC 2017 Classification

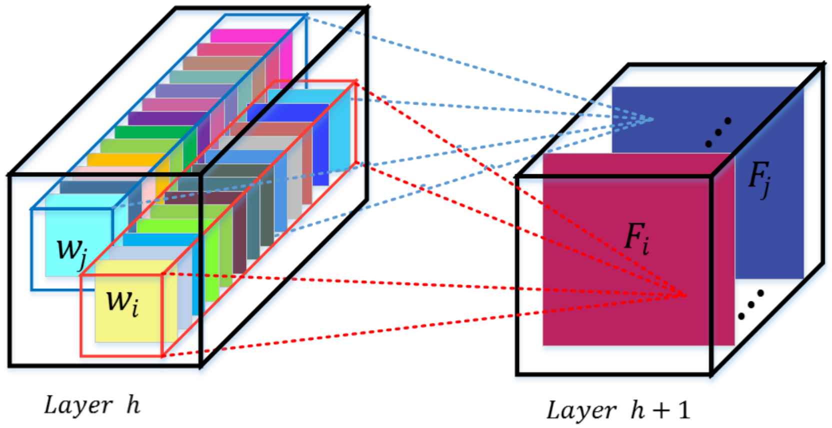


[Statistics provided by ILSVRC]

Convolution

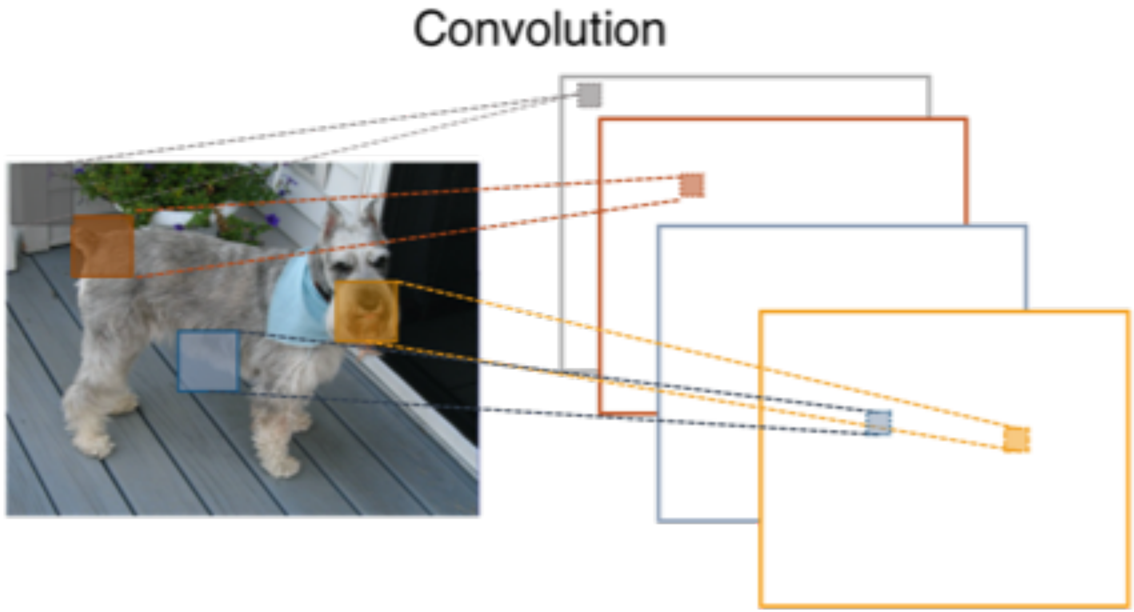
A convolutional filter is expected to be an informative combination

- Fusing *channel-wise* and *spatial* information

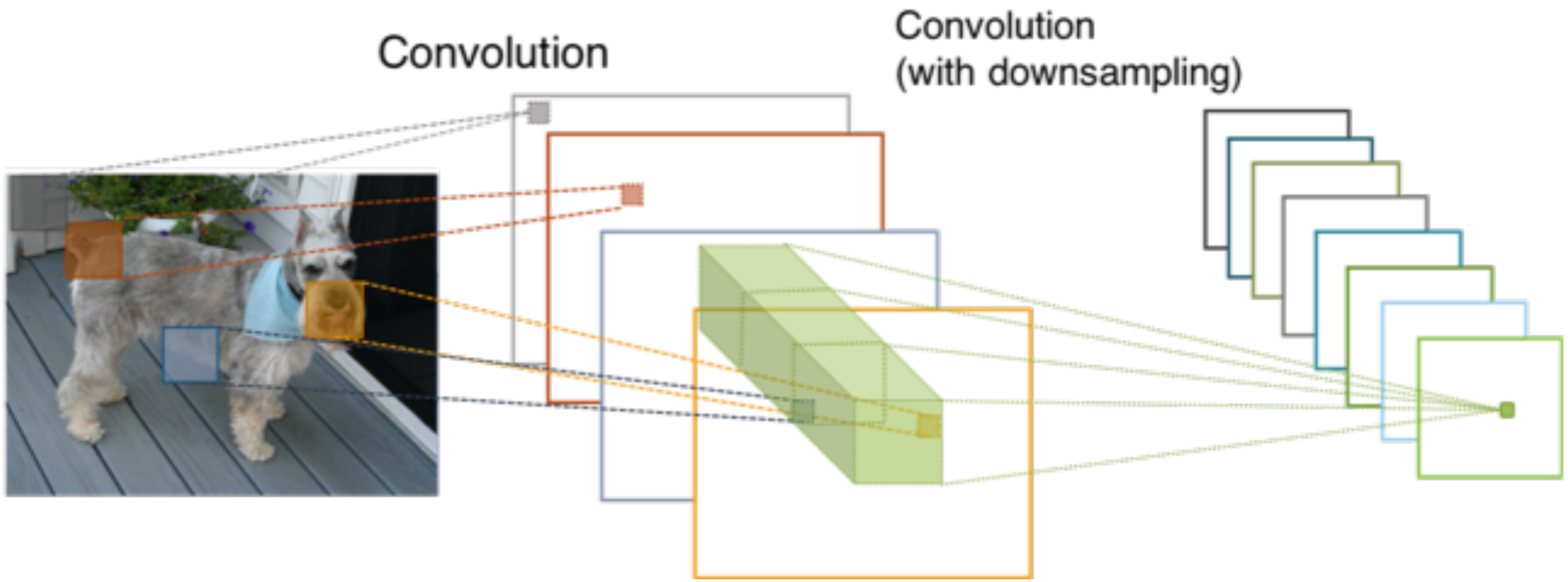


- Within *local* receptive fields

A Simple CNN

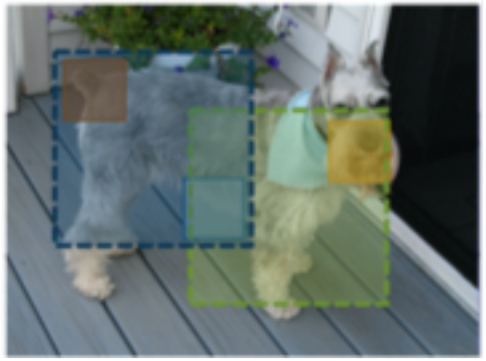


A Simple CNN



Channel dependencies are:

- *Implicit*: Entangled with the spatial correlation captured by the filters
- *Local*: Unable to exploit contextual information outside this region



Exploiting Channel Relationships

Can the representational power of a network be enhanced by *channel relationships*?

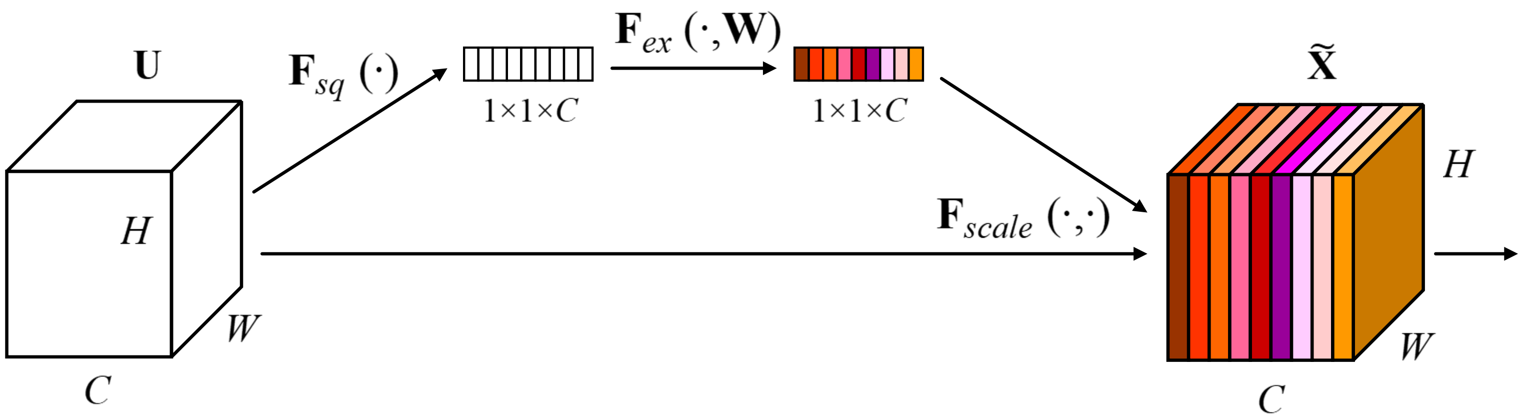
Design a new architectural unit

- Explicitly model interdependencies between the channels of convolutional features
- Feature recalibration
  - ❑ *Selectively* emphasise informative features and inhibit less useful ones
  - ❑ Use *global* information

Squeeze-and-Excitation Blocks

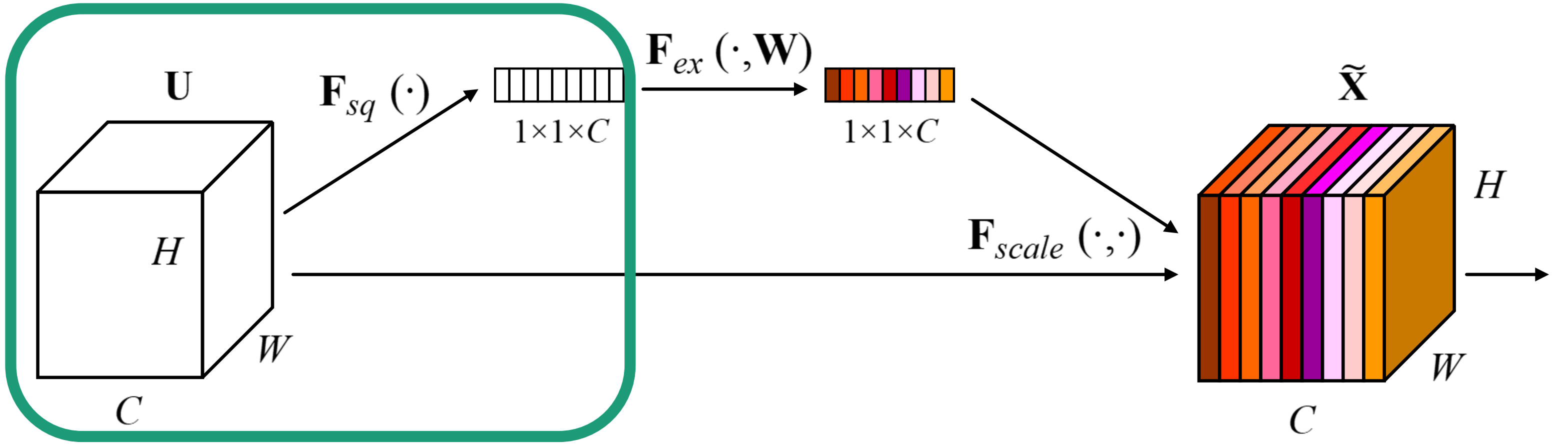
Given transformation F_{tr}: input X → feature maps U

- Squeeze
- Excitation



Squeeze: Global Information Embedding

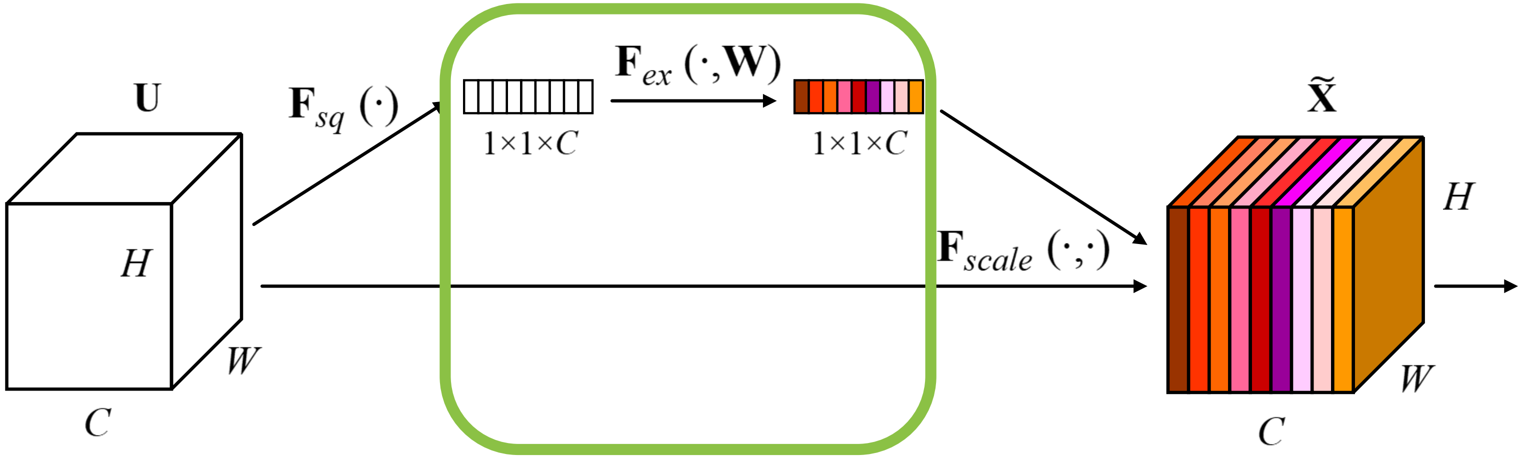
- Aggregate feature maps through spatial dimensions using global average pooling
- Generate channel-wise statistics



U can be interpreted as a collection of local descriptors whose statistics are expressive for the whole image.

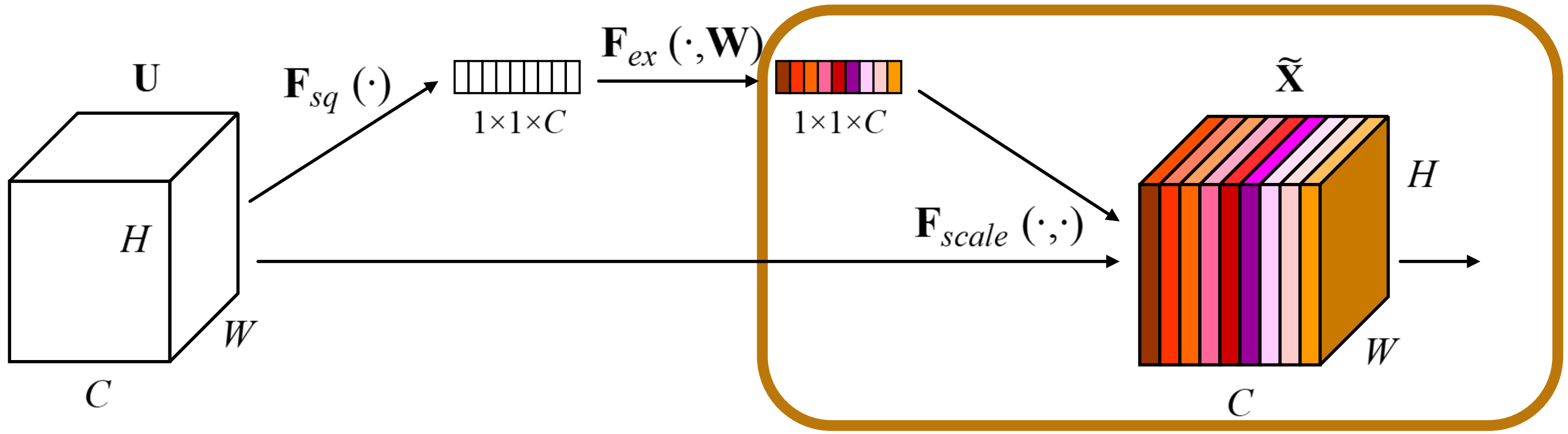
Excitation: Adaptive Recalibration

- Learn a nonlinear and non-mutually-exclusive relationship between channels
- Employ a self-gating mechanism with sigmoid function
  - ❑ Input: channel-wise statistics
  - ❑ Bottleneck configuration with two FC layers around non-linearity
  - ❑ Output: channel-wise activations



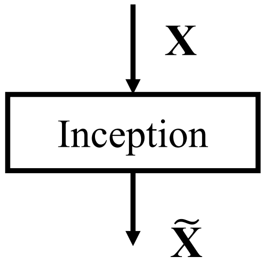
Excitation: Adaptive Recalibration

- Rescale the feature maps U with the channel activations
  - ❑ Act on the channels of U
  - ❑ Channel-wise multiplication

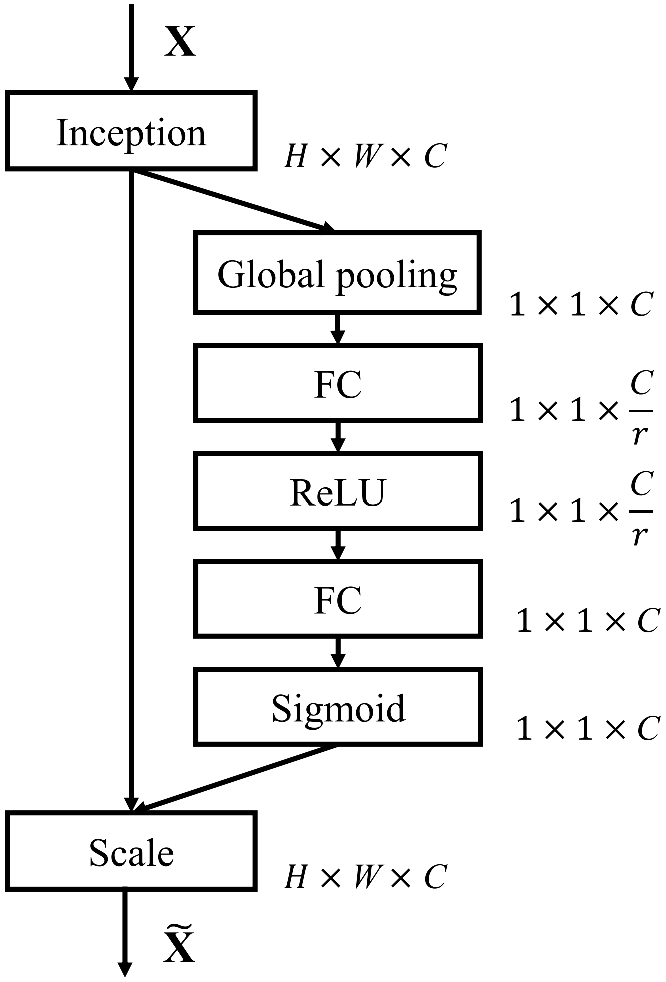


SE blocks intrinsically introduce dynamics conditioned on the input.

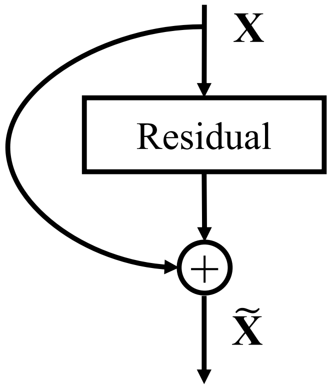
Example Models



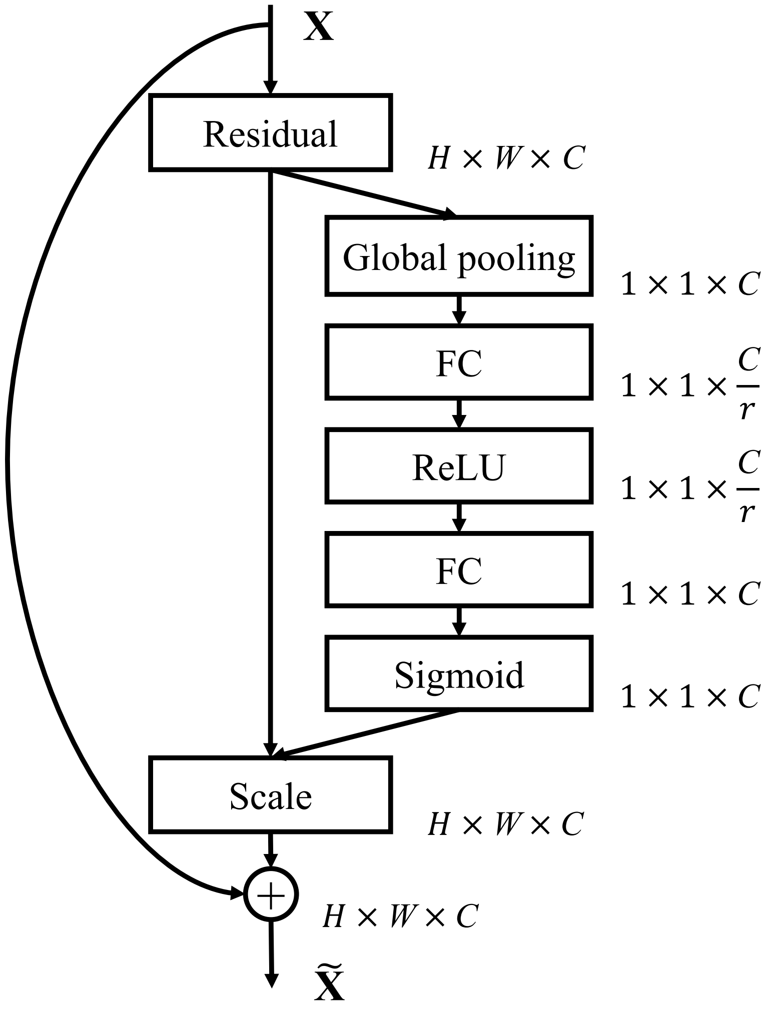
Inception Module



SE-Inception Module



ResNet Module



SE-ResNet Module

Object Classification

Experiments on ImageNet-1k dataset

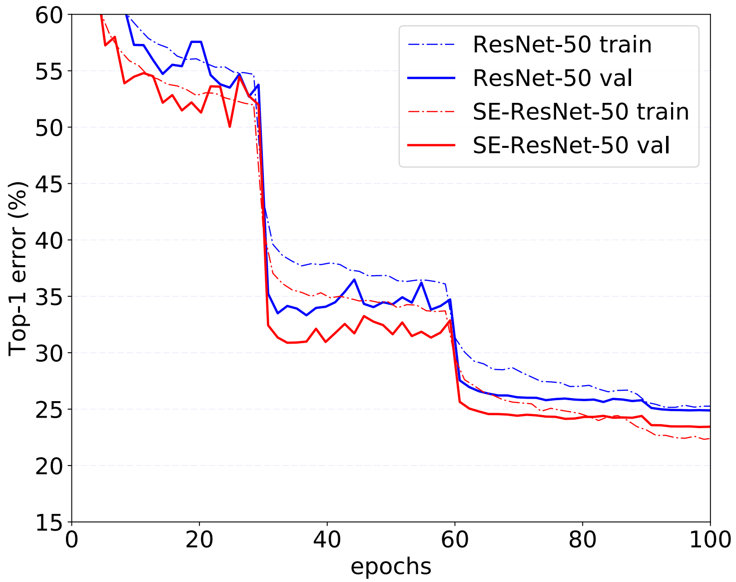
- Benefits at different depths
- Incorporation with modern architectures

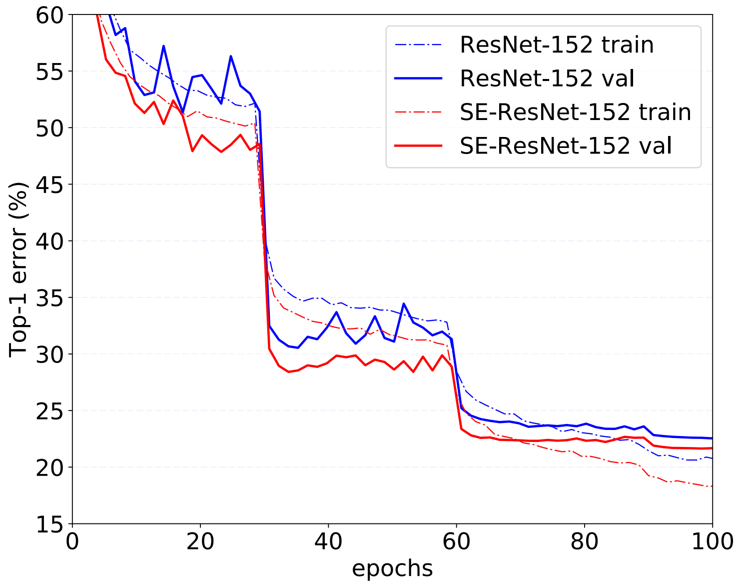
Benefits at Different Depths

SE blocks consistently improve performance across different depths at minimal additional computational complexity (no more than 0.26%).

- ✓ SE-ResNet-50 exceeds ResNet-50 by 0.86% and approaches the result of ResNet-101.
- ✓ SE-ResNet-101 outperforms ResNet-152.

	top-1 error		top-5 error	
	plain	SENet	plain	SENet
ResNet-50 [10]	24.80	$23.29_{(1.51)}$	7.48	$6.62_{(0.86)}$
ResNet-101 [10]	23.17	$22.38_{(0.79)}$	6.52	$6.07_{(0.45)}$
ResNet-152 [10]	22.42	$21.57_{(0.85)}$	6.34	$5.73_{(0.61)}$





Incorporation with Modern Architectures

SE blocks can boost the performance of a variety of network architectures on both *residual* and *non-residual* settings.

	top-1 error		top-5 error	
	plain	SENet	plain	SENet
ResNeXt-50 [47]	22.11	$21.10_{(1.01)}$	5.90	$5.49_{(0.41)}$
ResNeXt-101 [47]	21.18	$20.70_{(0.48)}$	5.57	$5.01_{(0.56)}$
VGG-16 [39]	27.02	$25.22_{(1.80)}$	8.81	$7.70_{(1.11)}$
BN-Inception [16]	25.38	$24.23_{(1.15)}$	7.89	$7.14_{(0.75)}$
Inception-ResNet-v2 [42]	20.37	$19.80_{(0.57)}$	5.21	$4.79_{(0.42)}$
MobileNet [13]	29.1	$25.3_{(3.8)}$	10.1	$7.9_{(2.2)}$
ShuffleNet [52]	33.9	$31.7_{(2.2)}$	13.6	$11.7_{(1.9)}$

Beyond Object Classification

SE blocks can generalise well on different datasets and tasks.

- Places365-Challenge Scene Classification

	top-1 err.	top-5 err.
Places-365-CNN [37]	41.07	11.48
ResNet-152 (ours)	41.15	11.61
SE-ResNet-152	**40.37**	**11.01**

Single-crop error rates (%) on Places365 validation set.

- Object Detection on COCO

	AP@IoU=0.5	AP
ResNet-50	45.2	25.1
SE-ResNet-50	46.8	26.4
ResNet-101	48.4	27.2
SE-ResNet-101	49.2	27.9

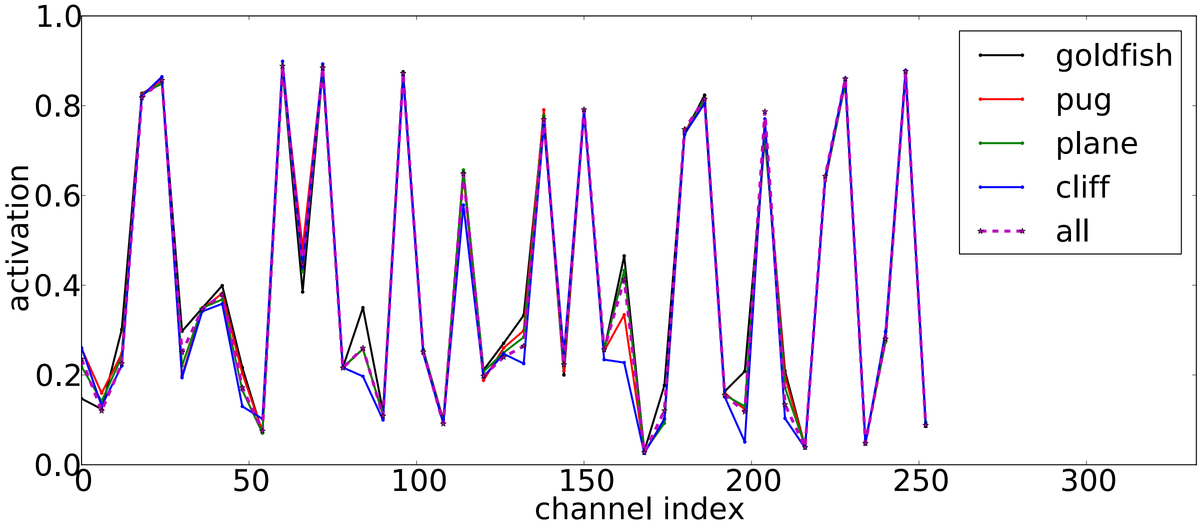
Object detection results on the COCO 40k validation set by using the basic Faster R-CNN.

Role of Excitation

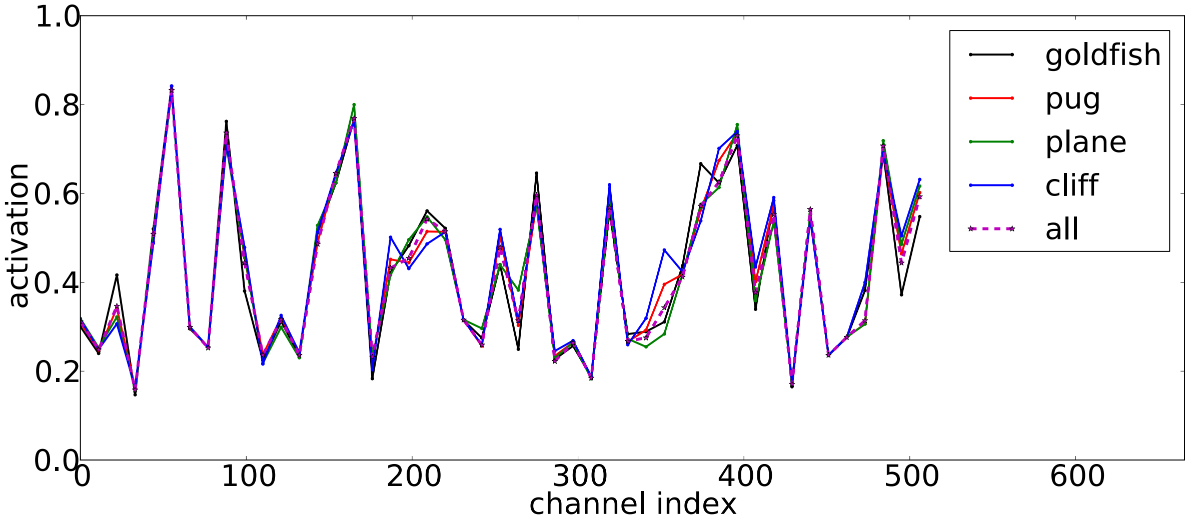
The role at different depths adapts to the needs of the network

- Early layers: Excite informative features in a *class agnostic* manner

SE_2_3



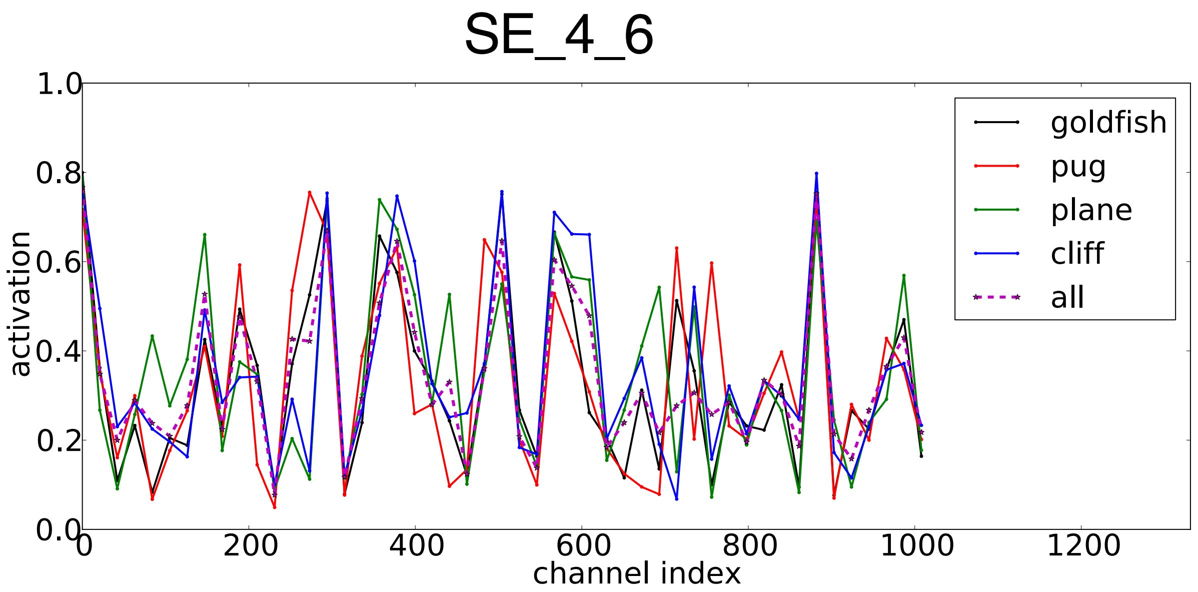
SE_3_4

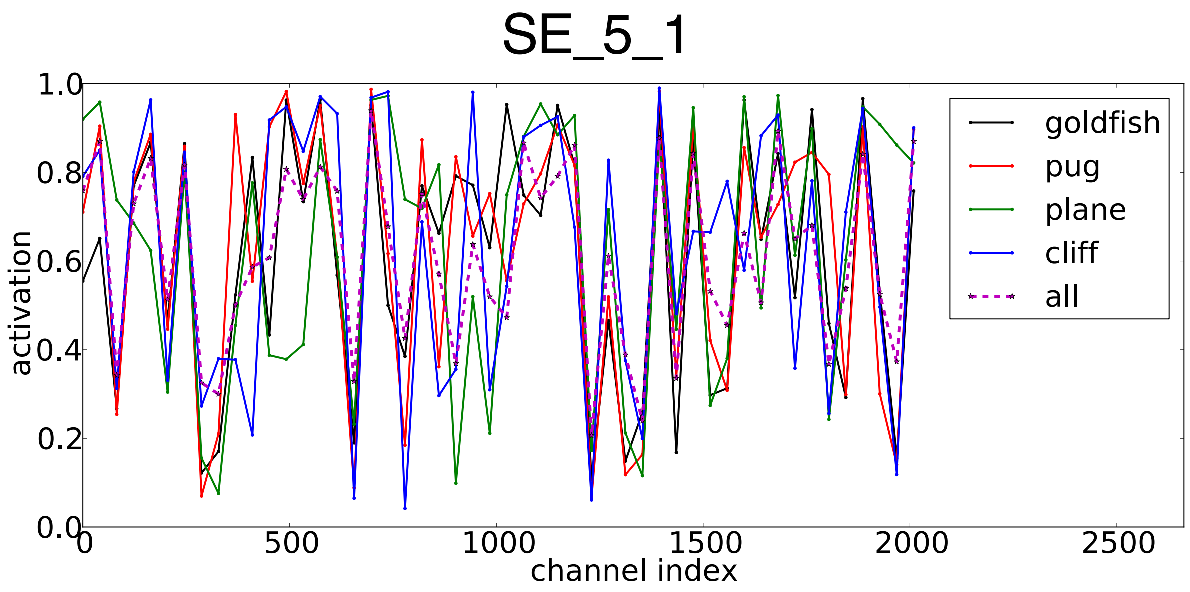


Role of Excitation

The role at different depths adapts to the needs of the network

- Later layers: Respond to different inputs in a highly *class-specific* manner





Conclusion

- Designed a novel architectural unit to improve the representational capacity of networks by dynamic channel-wise feature recalibration.
- Provided insights into the limitations of previous CNN architectures in modelling channel dependencies.
- Induced feature importance may be helpful to related fields, e.g. network compression.

Code and Models: https://github.com/hujie-frank/SENet

Thank you!