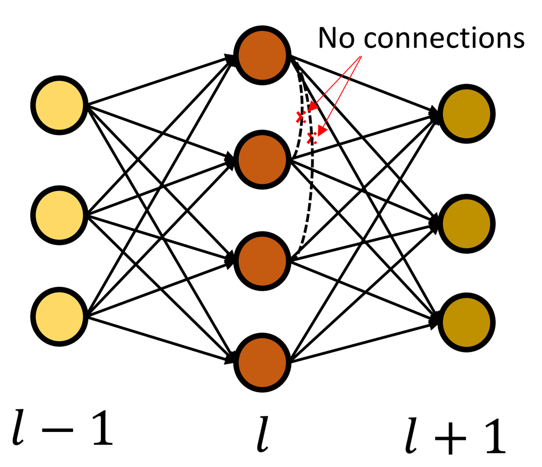
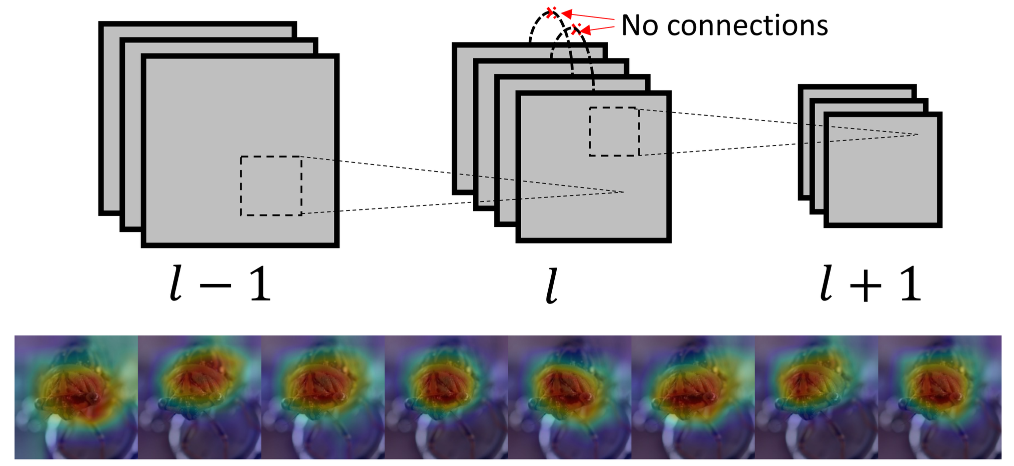
Cross-channel Communication Networks

Jianwei Yang[1], Zhile Ren[1], Chuang Gan[3], Hongyuan Zhu[4], Devi Parikh[1,2]

1. Georgia Tech; 2. Facebook AI Research (FAIR); 3. MIT-IBM Watson AI Lab; 4. Institute for Infocomm Research, A* Star

1 Background





Exemplar CAM visualizations (Top-8 responses)

1. Neurons **at the same layer** do not directly interact with each other.
2. Different neurons might respond to the same patterns and locations.

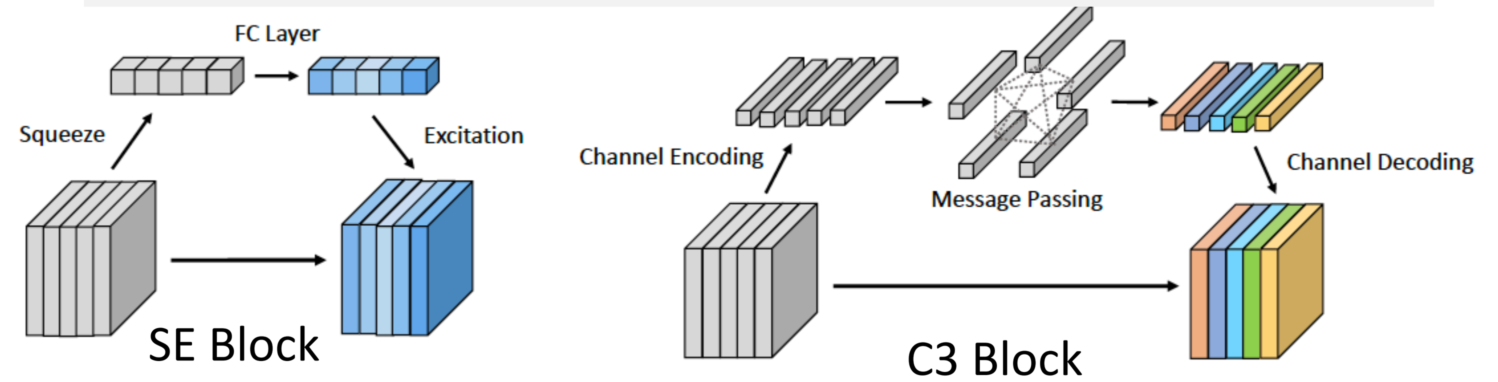
2 Motivations and Contributions

Channel responses naturally encodes **which pattern is at where**.

Idea: Enable channels at the same layer to communicate with each other and then calibrate their responses accordingly.

Goal: Different filters learn to focus on different useful patterns.

Squeeze-and-Excitation (SE) block is a first attempt but with extremely narrow communication bandwidth.

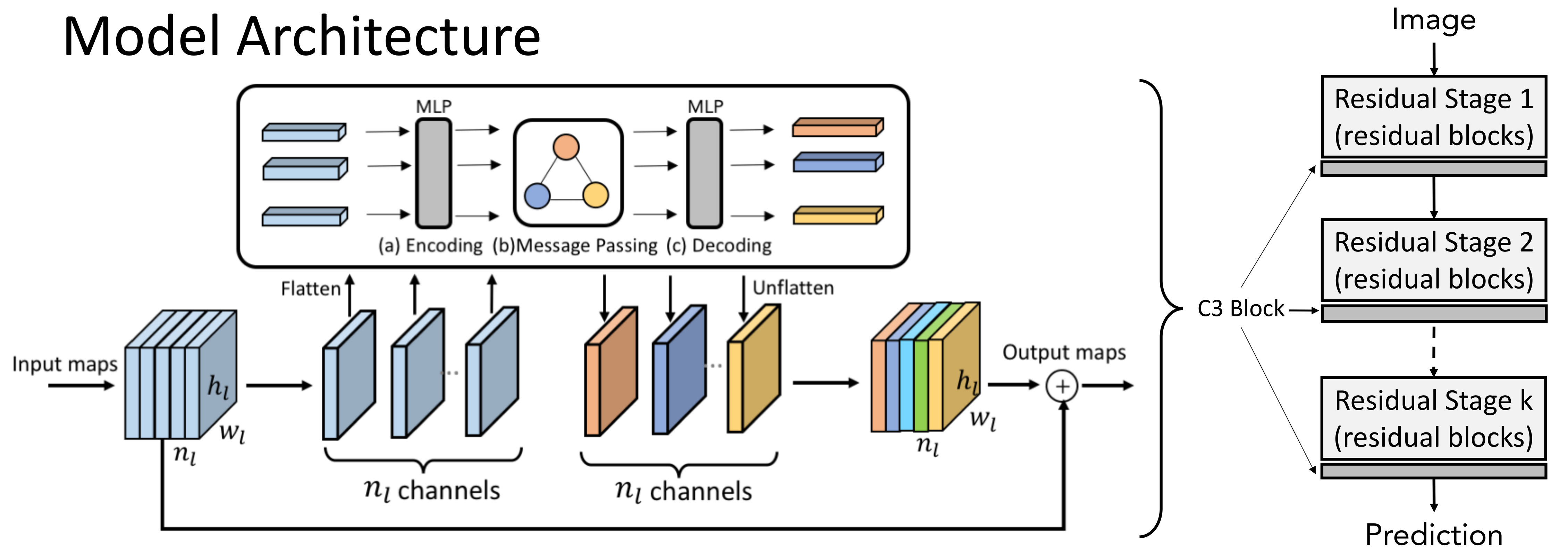


Contributions

1. Proposed **cross-channel communication (C3)** block to enable full interactions across channels at the **same** layer.
2. Achieved better performance on **image classification, object detection and semantic segmentation**.
3. Captured more **diverse** representations with **light-weight** networks.

3 Cross-Channel Communication

Model Architecture



(a) Feature Encoder

Flattened feature maps: $\{x_l^i\}_{i=1}^{n_l}, x_l^i \in R^{h_l w_l}$

Bottlenecked encoding:

$$y_l^i = f_{enc}^{in}(x_l^i) \in R^{h_l w_l / k}$$

K=8 in our experiments

$$z_l^i = f_{enc}^{out}(\sigma(y_l^i)) \in R^{h_l w_l}$$

NOTE: *Use bottleneck for reducing model size and avoiding overfitting.*

(b) Message Passing

Encoded feature maps: $\{z_l^i\}_{i=1}^{n_l}, z_l^i \in R^{h_l w_l}$

Computing affinities across channel:

$$v_l^i = \frac{1}{h_l w_l}\sum_{k=1}^{h_l w_l} z_l^i[k] \qquad s_{ij} = -(v_l^i - v_l^j)^2 \qquad a_{ij} = softmax(\{s_{ij}\})_j$$

Computing message: $\bar{z}_l^i = \sum_{j \neq i} a_{ij} z_l^j$

NOTE: *Averaging for efficiency and robustness to noises.*

(c) Feature Decoder

Computed message: $\{\bar{z}_l^i\}_{i=1}^{n_l}, \bar{z}_l^i \in R^{h_l w_l}$

Bottlenecked decoding:

$$\bar{y}_l^i = f_{dec}^{in}(\bar{z}_l^i) \in R^{h_l w_l / k}$$

$$\bar{x}_l^i = f_{dec}^{out}(\sigma(\bar{y}_l^i)) \in R^{h_l w_l}$$

Adding residuals: $x_l^i = x_l^i + \bar{x}_l^i$

NOTE: *Add residuals to modulate the original responses to be more complementary.*

Measure Diversity

Any flattened feature: $\{x_l^i\}_{i=1}^{n_l}, x_l^i \in R^{h_l w_l}$

Correlations between channels:

$$c_l^{ij} = \frac{< x_l^i - v_l^i, x_l^j - v_l^j >}{\sigma_l^i \sigma_l^j}$$

Diversity in a single layer:

$$d_l = \frac{1}{(n_l)^2}\sum_i \sum_j |c_l^{ij}|$$

Absolute operation

NOTE: *Use absolute value since we want to penalize both positive and negative correlations.*

4 Experiments

Performance on vision tasks

	ResNet-20			*ResNet-56*			*ResNet-110*			*Wide-ResNet*		
	Size	FLOPs	Acc.	Size	FLOPs	Acc.	Size	FLOPs	Acc.	Size	FLOPs	Acc.
Baseline	0.28	41.7M	67.73	0.86	128.2M	71.05	1.74	257.9M	72.01	26.86	3.84G	77.96
Baseline + SE	0.28	41.8M	68.57	0.87	128.5M	72.00	1.76	258.5M	72.47	27.26	3.84G	78.57
Baseline + C3	0.35	46.0M	69.34	0.93	132.5M	72.27	1.81	262.2M	73.36	26.93	3.87G	78.34

Table. Comparisons on image classification on CIFAR-100

Detection	Pascal VOC	COCO
FRCNN	74.6	33.9
FRCNN + SE	74.8	34.3
FRCNN + C3	75.6	34.8

Table. Comparisons on object detection

	ResNet-18			*ResNet-50*			*ResNet-101*			*ResNeXt-50*		
	Size	top-1	top-5	Size	top-1	top-5	Size	top-1	top-5	Size	top-1	top-5
Baseline	11.7	30.28	10.52	25.6	23.61	7.27	44.6	22.48	6.18	34.9	23.85	7.12
Baseline + SE	11.8	30.15	10.72	28.1	22.51	6.43	49.3	22.14	6.14	37.5	22.90	6.44
Baseline + C3	12.0	29.30	10.48	25.9	23.19	6.60	44.9	21.93	6.02	35.3	22.51	6.23

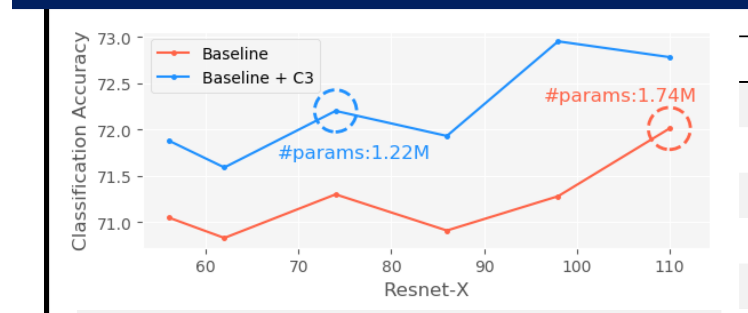
Table. Comparisons on image classification on ImageNet

Segmentation	*Mean IoU*	*Mean Acc.*
Deeplabv2	75.2	85.3
Deeplabv2 + SE	75.6	85.6
Deeplabv2 + C3	75.7	86.0

Table. Comparisons on semantic segmentation

C3 blocks improve image classification performance with a few time and memory overhead
C3 blocks also improve object detection and semantic segmentation performance

Analysis and Visualizations

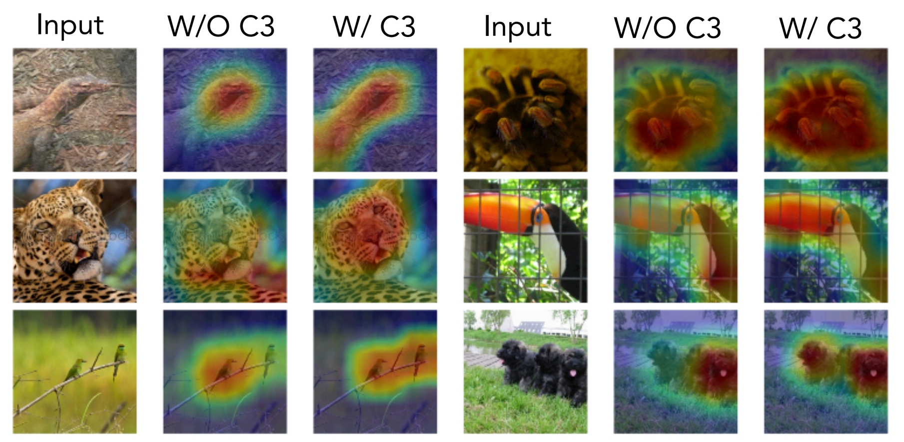


C3 helps to save network depth while maintaining the performance.

	W/O C3	*W/ C3*
ResNet-56	0.86M	0.93M
ResNet-62	0.96M	1.03M
ResNet-74	1.15M	1.22M
ResNet-86	1.35M	1.41M
ResNet-98	1.54M	1.61M
ResNet-110	1.74M	1.80M

Table. Model Size

CAM Visualizations



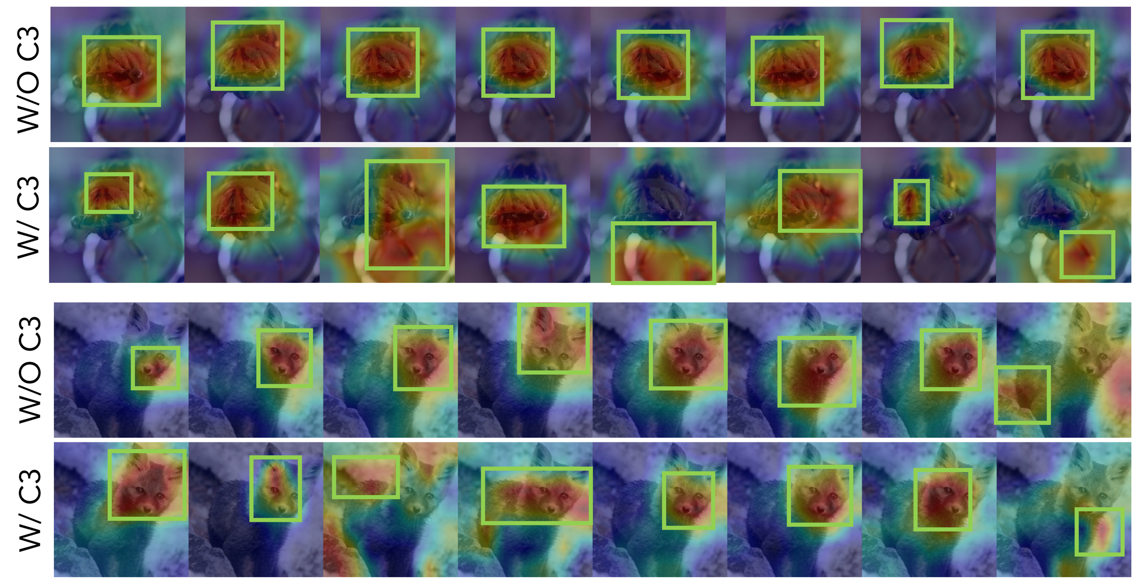
C3 helps to learn more complementary filters covering the regions of **all** objects (either single or multiple)

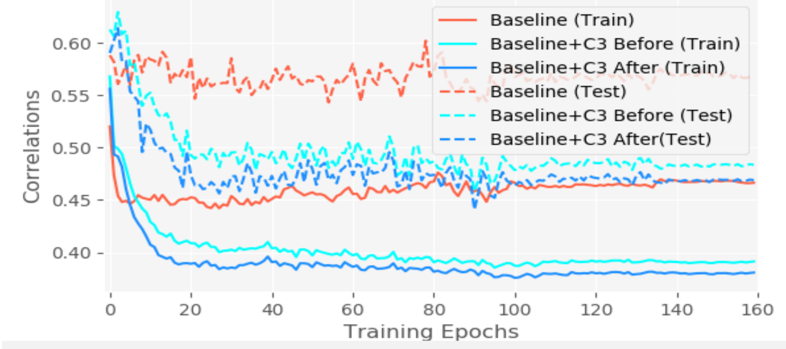
Enc-Dec	Mess. Passing	*ResNet-20*	*ResNet-56*	*ResNet-110*
Yes	No	68.70	71.95	72.65
No	Yes	69.13	71.79	72.74
Yes	Yes	69.34	72.27	73.36

Table. Performance for different combinations

Encoder, decoder and mess. passing are all necessary.

Top-8 channel responses





C3 block encourages diversity compared with baseline
Output of C3 block is more diverse than its input.