More is Less: A More Complicated Network with Less Inference Complexity

Xuanyi Dong[1], Junshi Huang[2], Yi Yang[1], Shuicheng Yan[2,3]
[1]University of Technology Sydney, [2]360 AI Insitute,
[2]National University of Singapore

2017/04/12 @ VALSE

Outline

- Introduction
- Overview of Existing Methods
- The Proposed Model
- Experiments

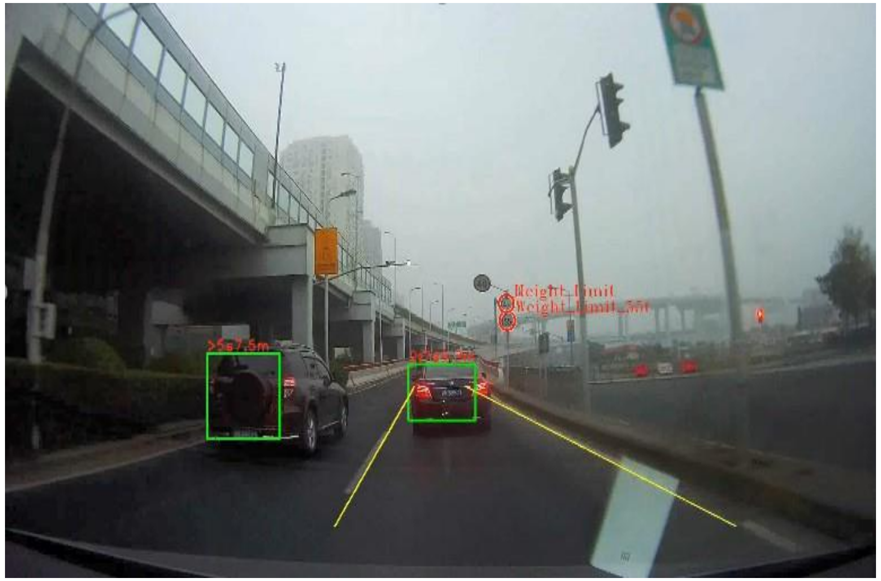
CNNs cost a lot

	Parameters	FLOPs	Top-5 Error
AlexNet	61M	725M	17.0
VGG-16	138M	15484M	8.43
GoogleNet-V1	6.9M	1566M	7.89
ResNet-50	25.5M	3800M	5.25

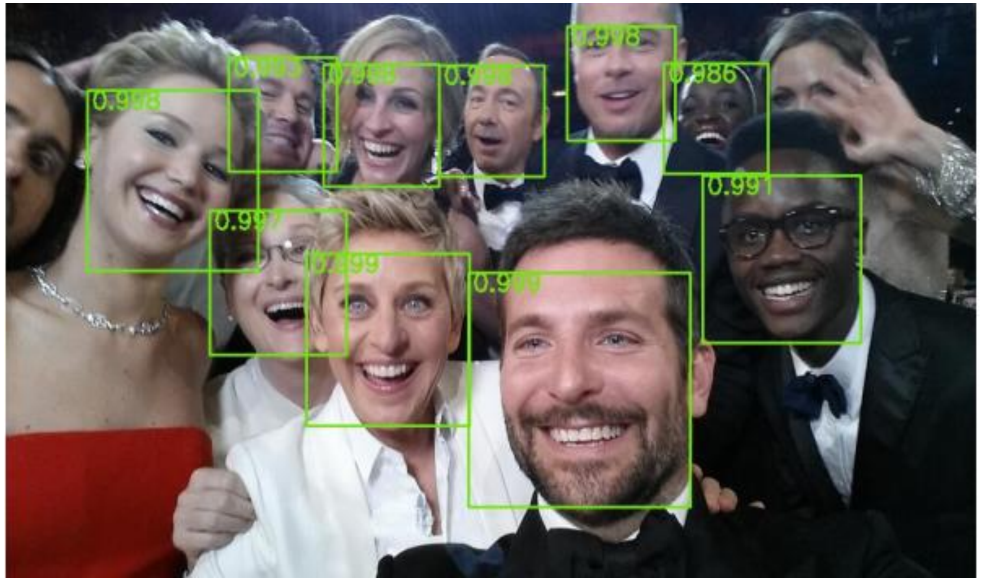
	Forward(ms)	Backward(ms)
VGG-16	143	379
GoogleNet-V1	63	102

- https://github.com/jcjohnson/cnn-benchmarks.git

Why CNN Acceleration?
Real-World Applications need Real-Time



Self Driving



Face Detection

Popular Dataset & Networks

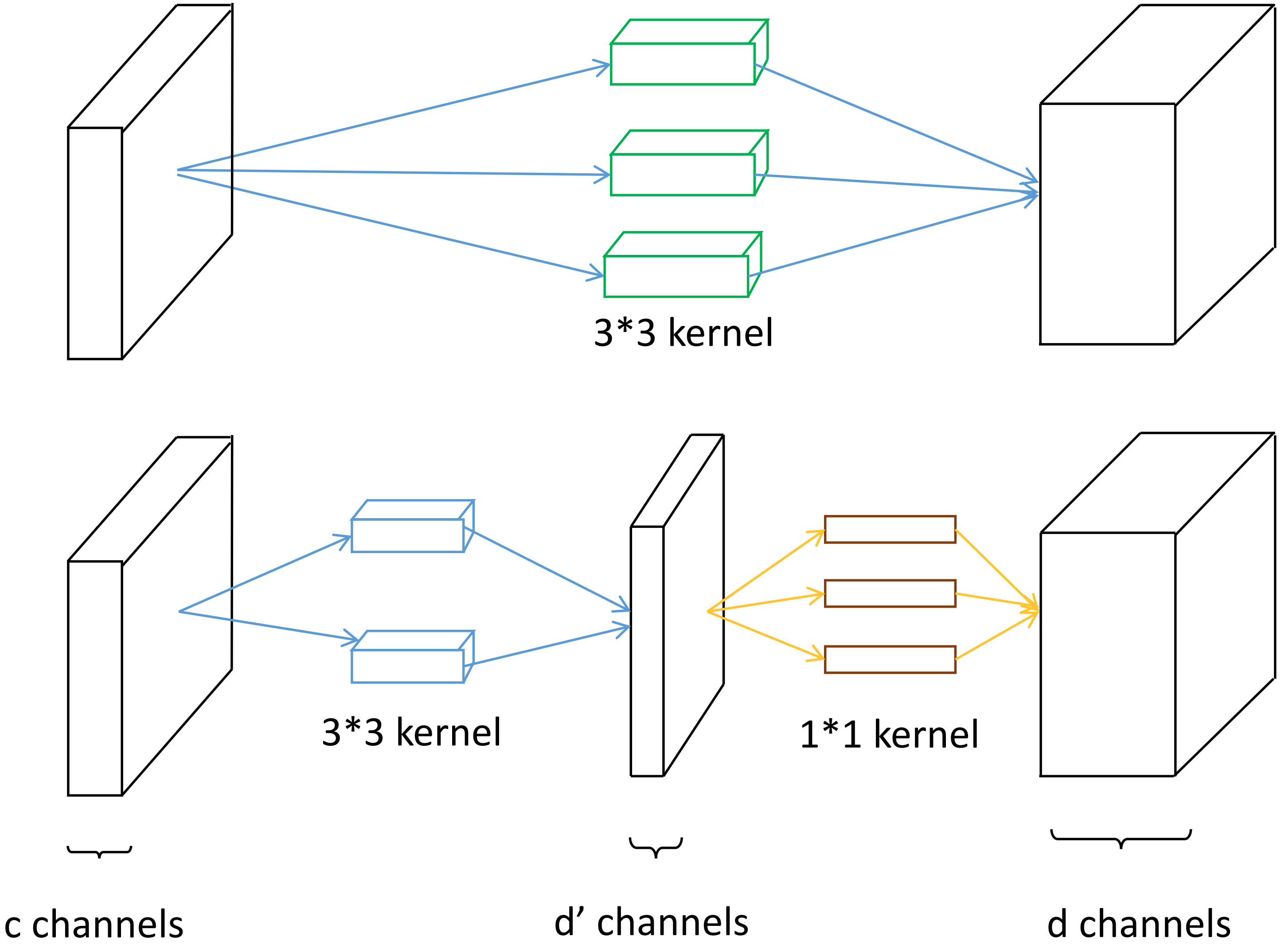
	Training	Testing	Classes
MNIST	60,000	10,000	10
CIFAR10	50,000	10,000	10
CIFAR100	50,000	10,000	100
ImageNet	1.2M	150,000	1000

	AlexNet	VGG-16	GoogleNet	ResNet
Frequency	Most	Most	Few	Rare

Related Works

- **Low Rank**
- **Fixed Point**
- **Product Quantization**
- **Sparse**
- **Architecture**
- **Dynamic CNN**

Low Rank



$O(d3^2c)$

$O(d'3^2c)$
$+O(d'd)$

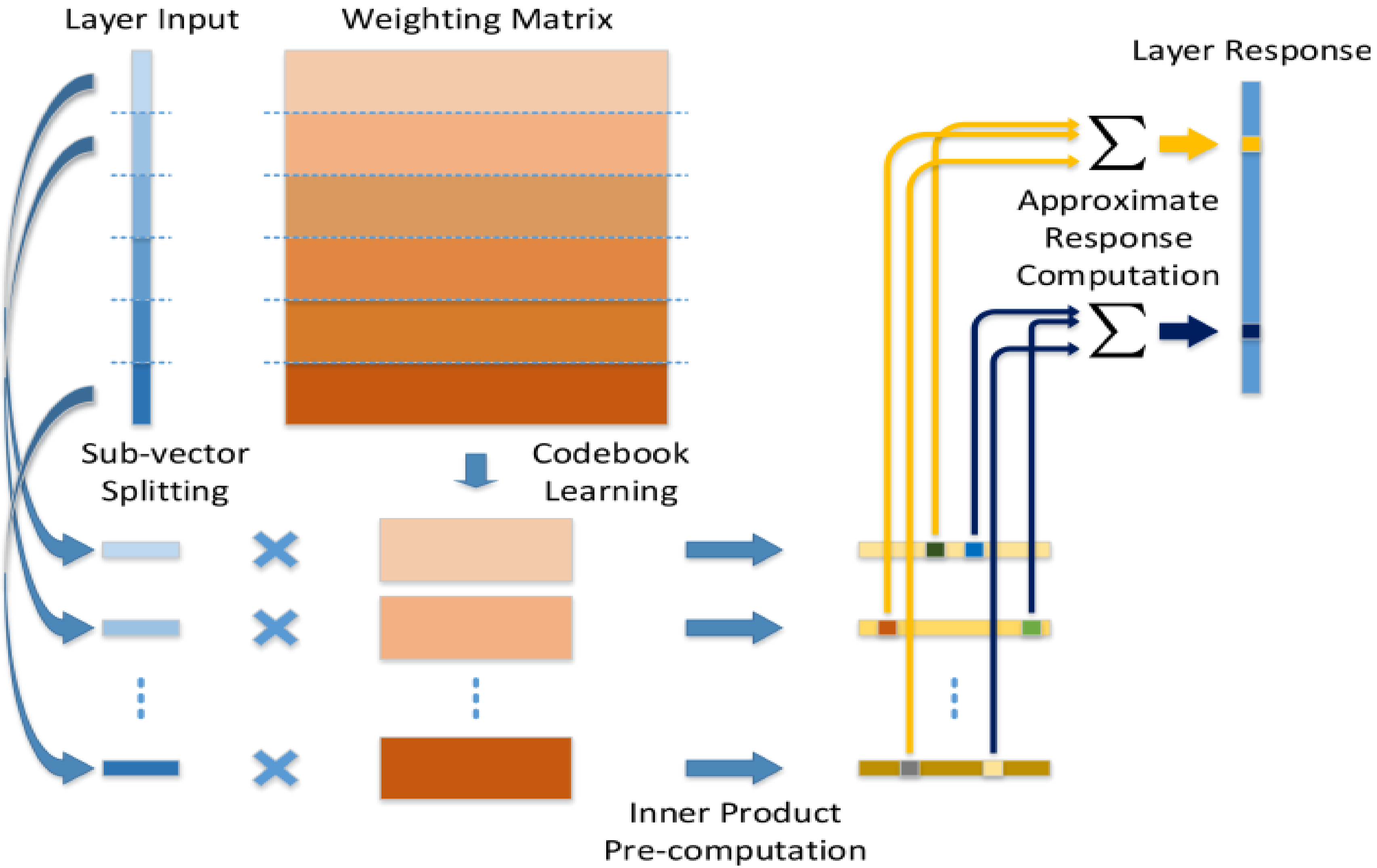
- Zhang, et al. "Accelerating very deep convolutional networks for classification and detection." TPAMI 2016

Fixed Point

	Network Variations	Operations used in Convolution	Memory Saving (Inference)	Computation Saving (Inference)	Accuracy on ImageNet (AlexNet)
Standard Convolution	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Real-Value Weights: 0.12 -1.2 ... 0.41 / -0.2 0.5 ... 0.68	+ , − , ×	1x	1x	%56.7
Binary Weight	Real-Value Inputs: 0.11 -0.21 ... -0.34 / -0.25 0.61 ... 0.52; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	+ , −	~32x	~2x	%56.8
BinaryWeight Binary Input **(XNOR-Net)**	Binary Inputs: 1 -1 ... -1 / -1 1 ... 1; Binary Weights: 1 -1 ... 1 / -1 1 ... 1	XNOR , bitcount	~32x	~58x	%44.2

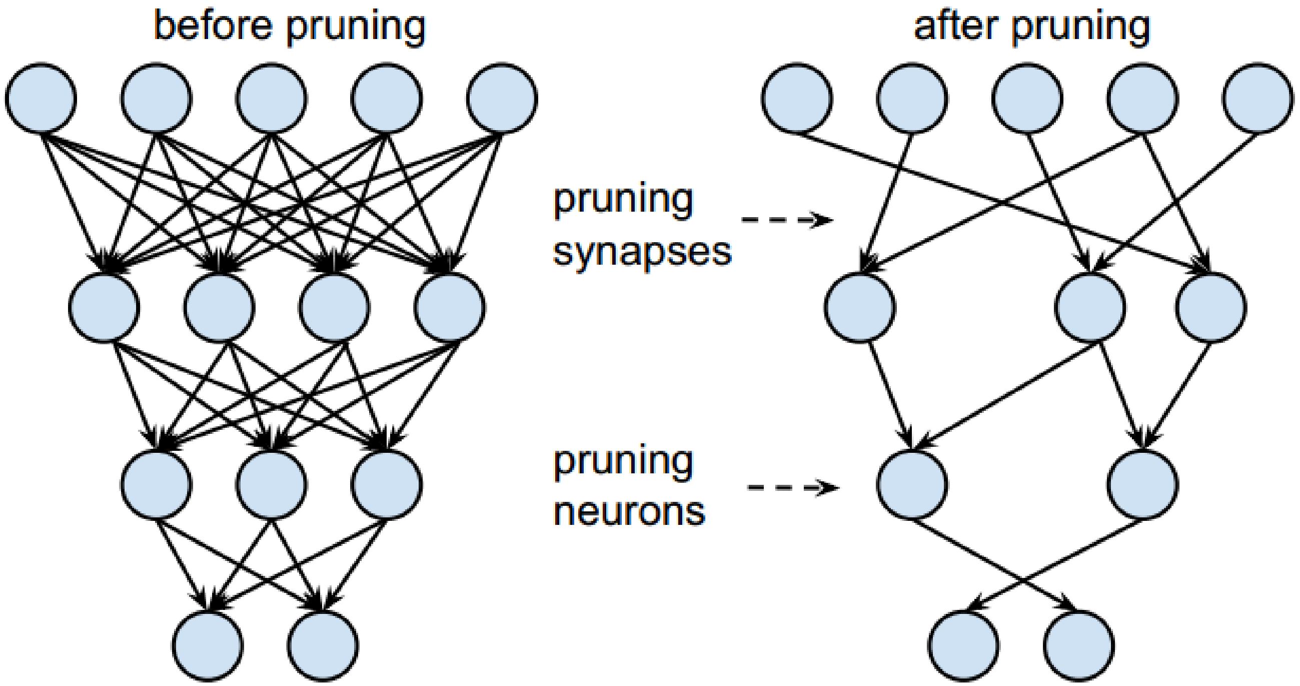
- Rastegari, et al. "Xnor-Net: Imagenet classification using binary convolutional neural networks." ECCV 2016

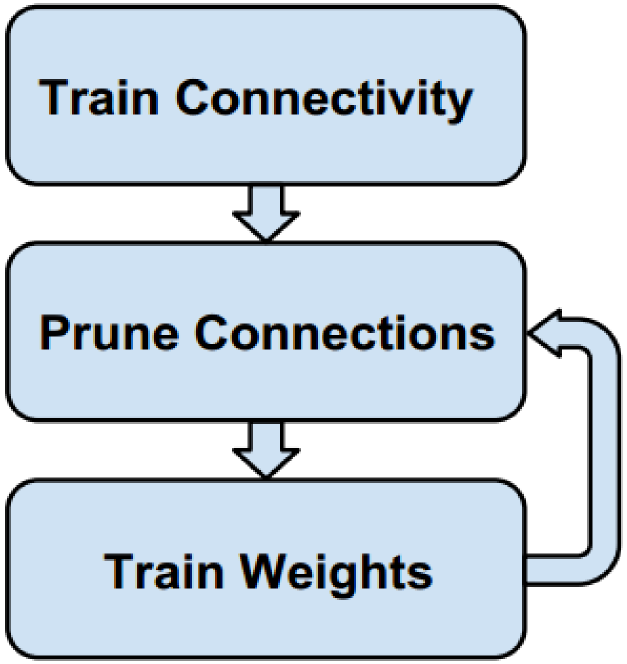
Product Quantization



- Wu, Jiaxiang, et al. "Quantized convolutional neural networks for mobile devices." CVPR 2016

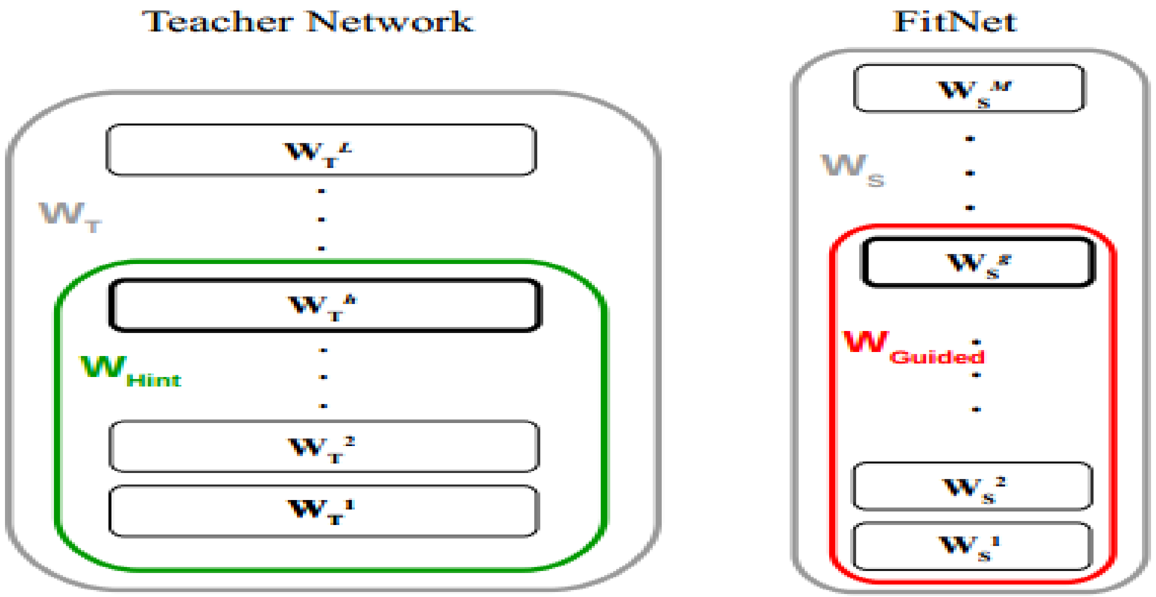
Sparse





- Han S, Pool J, Tran J, et al. “Learning both weights and connections for efficient neural network”. NIPS 2015

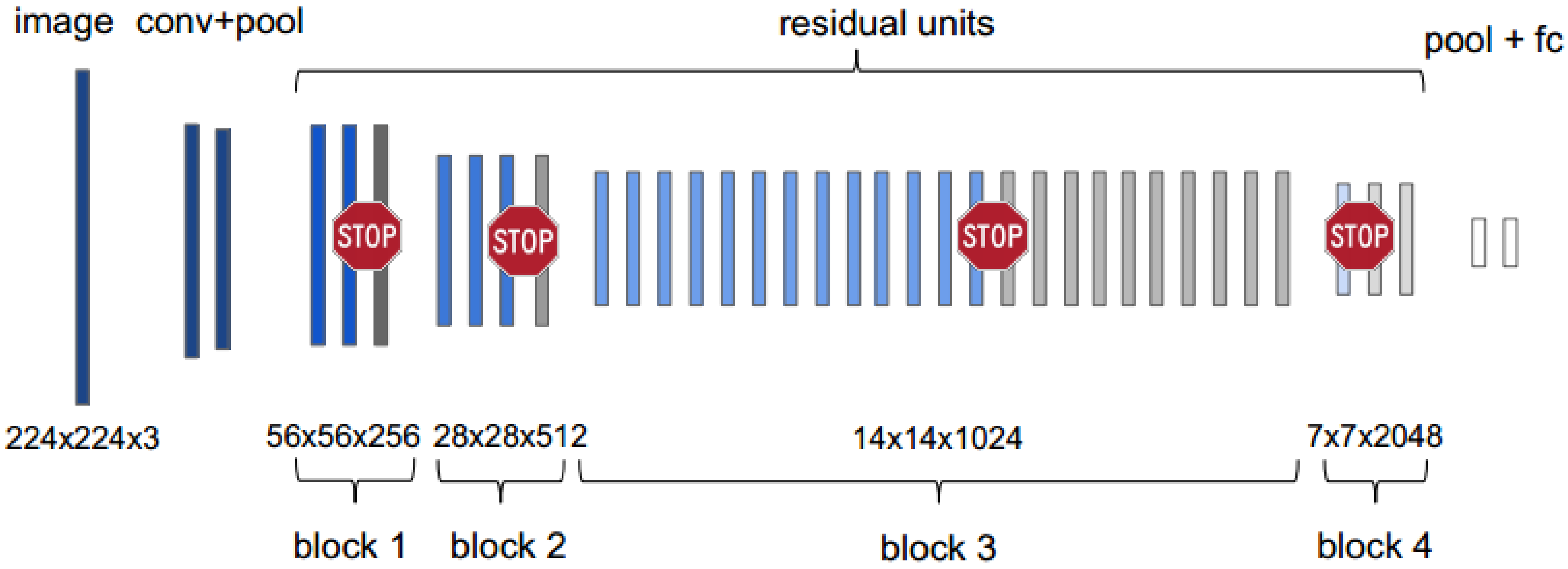
Architecture



Network	# layers	# params	# mult	Acc	Speed-up	Compression rate
Teacher	5	~9M	~725M	90.18%	1	1
FitNet 1	11	~250K	~30M	89.01%	**13.36**	**36**
FitNet 2	11	~862K	~108M	91.06%	4.64	10.44
FitNet 3	13	~1.6M	~392M	91.10%	1.37	5.62
FitNet 4	19	~2.5M	~382M	**91.61%**	1.52	3.60

- Romero, Adriana, et al. "Fitnets: Hints for thin deep nets." ICLR 2015

Dynamic CNN



- Figurnov, Michael, et al. "Spatially Adaptive Computation Time for Residual Networks." CVPR 2017

Problems

- **Focus** on the **Compression** rather than **Acceleration**
- **Focus** on the **Fully-Connected** layer not **Convolution** layer
- **High Theoretical** Time but hard to adapt **Practical Implementation**

Motivation

1.3	2.1	-1.2	0.3
-0.5	0.1	-1.7	1.9
0.8	1.1	0.6	-0.1
1.0	-0.9	0.7	0.2

Dense

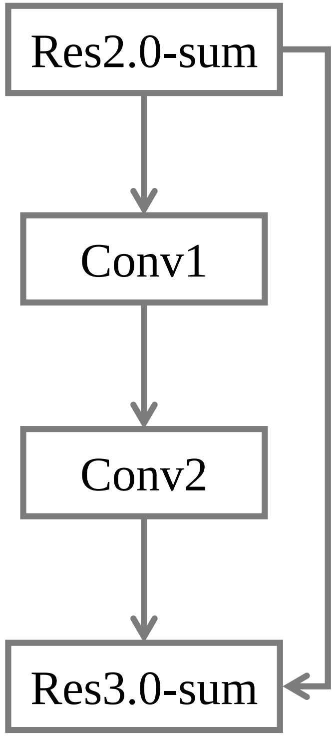
ReLU Activation →

1.3	2.1	**0**	0.3
0	0.1	**0**	1.9
0.8	1.1	0.6	**0**
1.0	**0**	0.7	0.2

Sparse

Efficient: More (complicated structure) is Less (computation complexity)

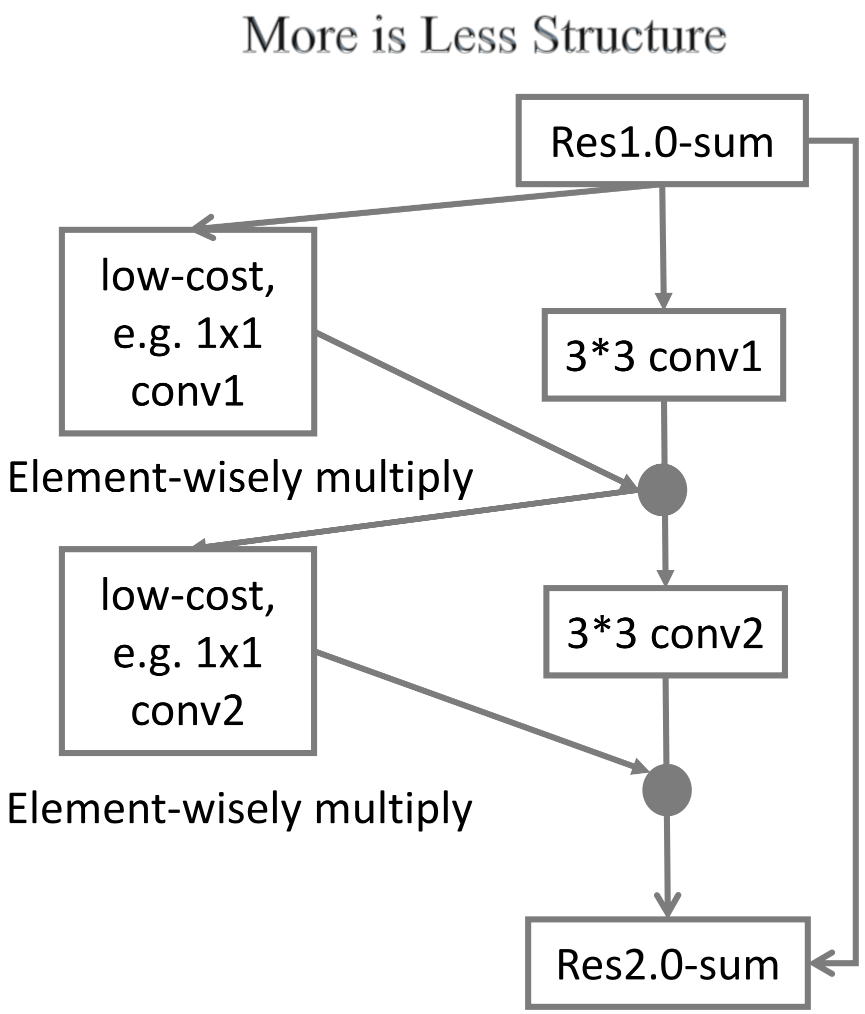
original residual block



Frequently, **>30% outputs are almost zeros** after the ReLU operation, and thus their exact convolution values before ReLU are meaningless.

Can these positions be roughly estimated with very low computational cost?

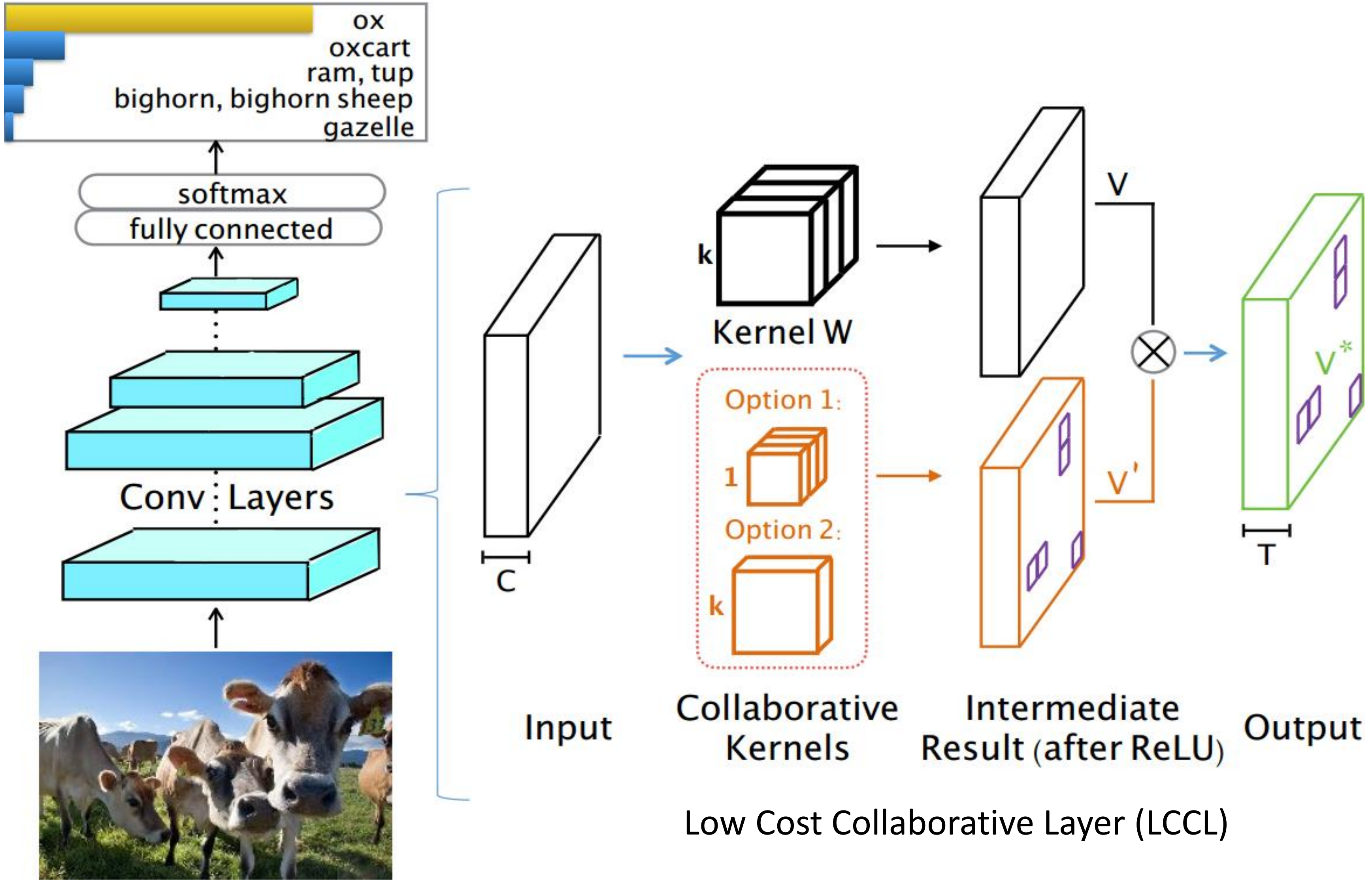
Efficient: More (complicated structure) is Less (computation complexity)



Theoretically, model accuracy can be lossless, yet complexity is less.

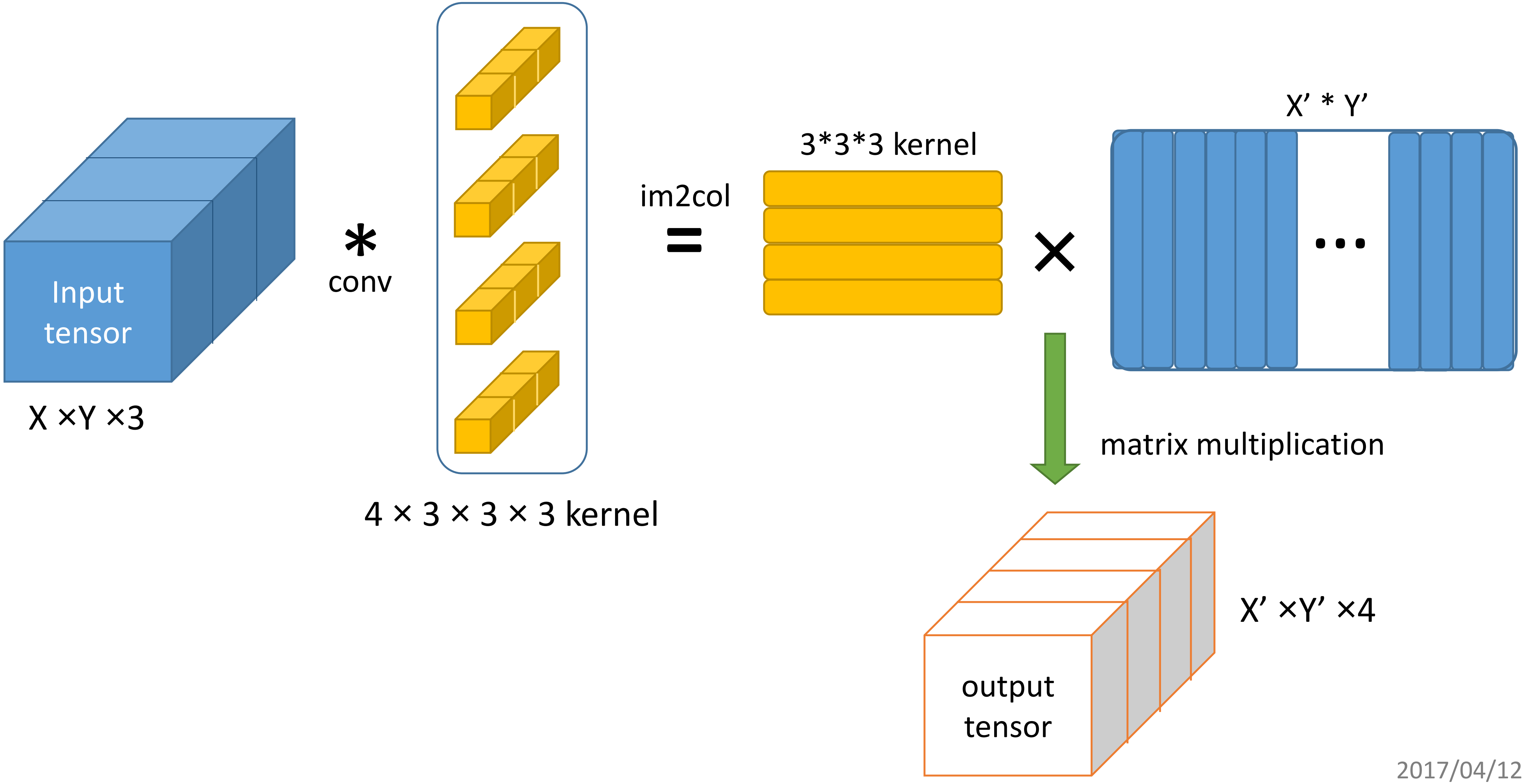
If 1x1 or low-cost conv $\frac{1}{2}$ outputs zero, then its corresponding convolution operation in conv $\frac{1}{2}$ is not required.

Proposed Architecture

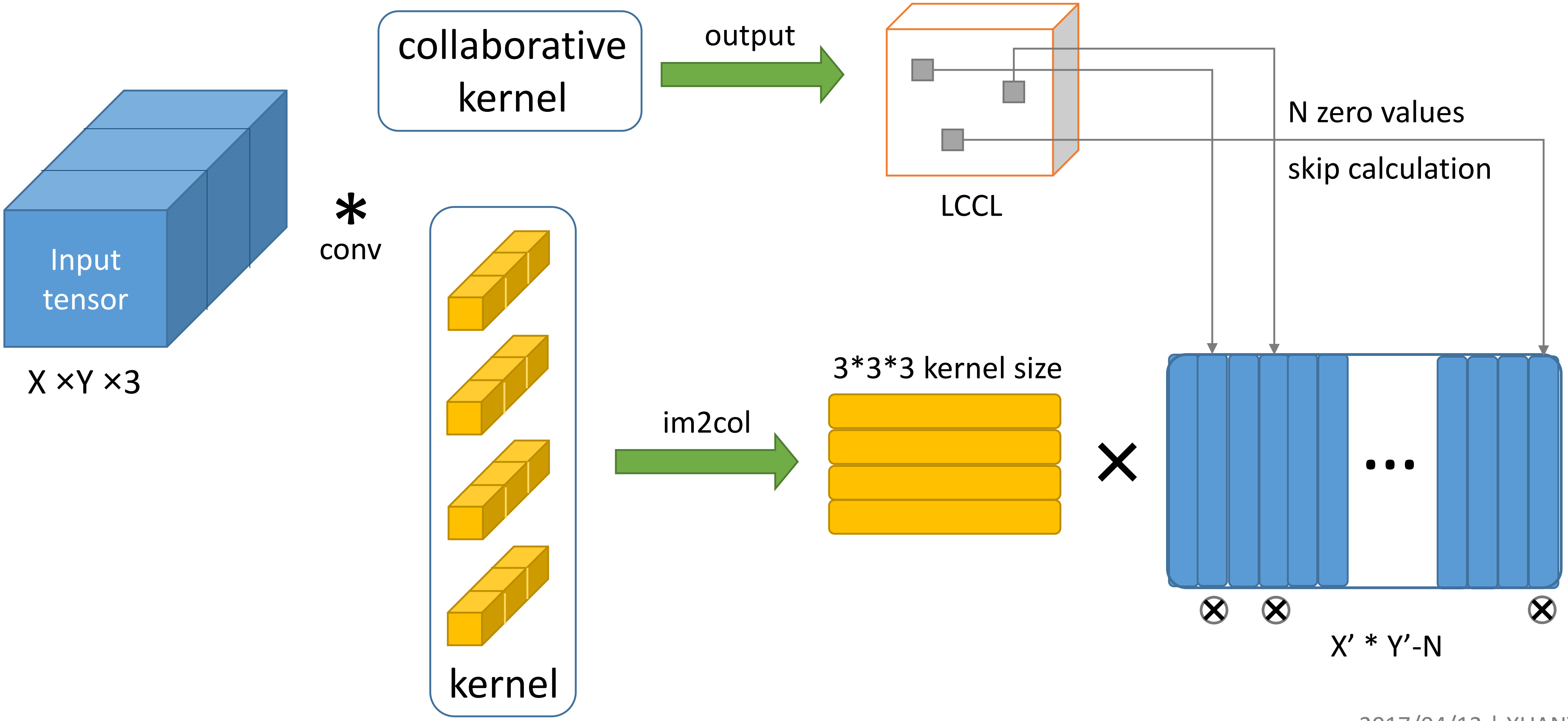


Low Cost Collaborative Layer (LCCL)

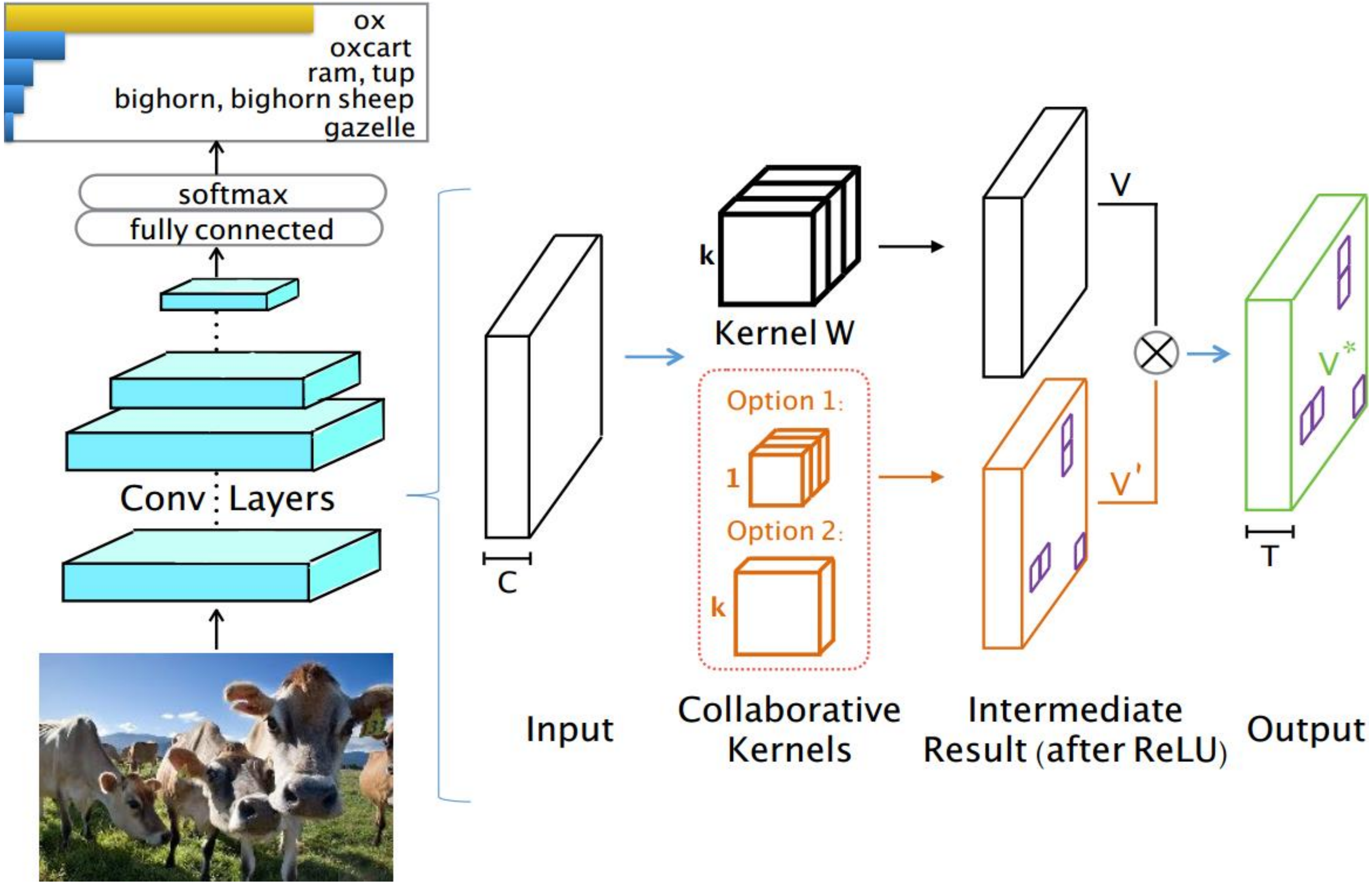
Implementation



Implementation – weight sharing



Trade off – acceleration and accuracy



Acceleration - Sparsity

Collaborative Layer	Sparsity	Trainable
Conv + Relu	< 10 %	Stable
Conv + Regularization + Relu	5 % - 70%	Unstable
Conv + BatchNorm + Relu	~30%	Stable

Accuracy - Kernel

Input tensor : U
Height & Width : X & Y
Collaborative Kernel : W'_t
Output tensor : V'_t
Sparsity Ratio : r

$$V'_t(x,y) = \sum_{i,j=1}^{k'} \sum_{c}^{C} W'_t(i,j,c)U(x+i-1, y+j-1, c)$$

Architecture	FLOPs	Speed-Up Ratio
CNN	$XYTk^2C$	0
basic	$XYTC(k'^2 + k^2r)$	$1-(k'^2/k^2+r)$
(1×1 kernel)	$XYTC(1+k^2r)$	$1-(1/k^2+r)$
(weight sharing)	$XYTk^2(1+Cr)$	$1-(1/C+r)$

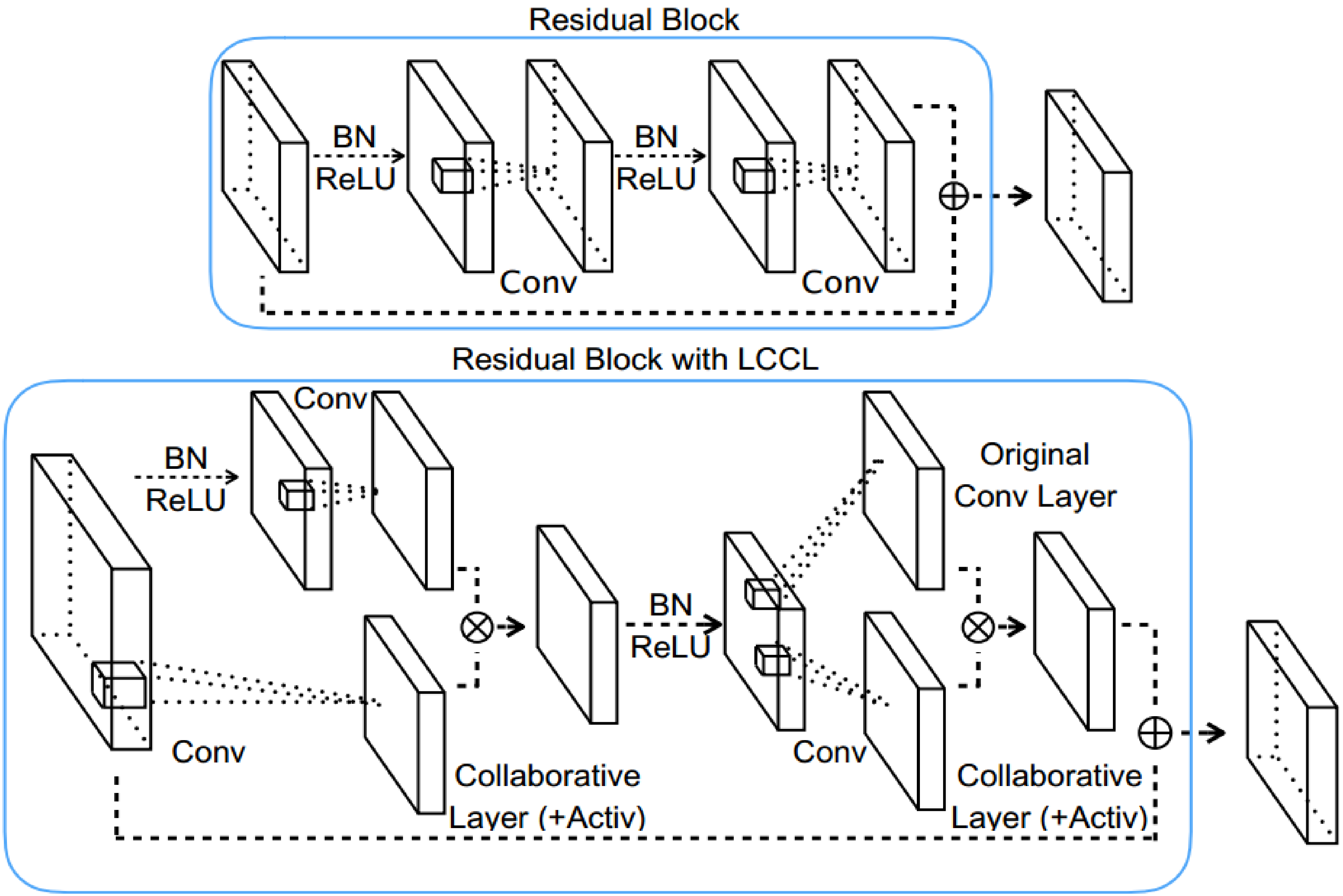
Accuracy - Kernel

Architecture	FLOPs	Speed-Up Ratio
CNN	$XYTk^2C$	0
basic	$XYTC(k'^2 + k^2r)$	$1 - (k'^2/k^2 + r)$
(1 × 1 kernel)	$XYTC(1 + k^2r)$	$1 - (1/k^2 + r)$
(weight sharing)	$XYTk^2(1 + Cr)$	$1 - (1/C + r)$

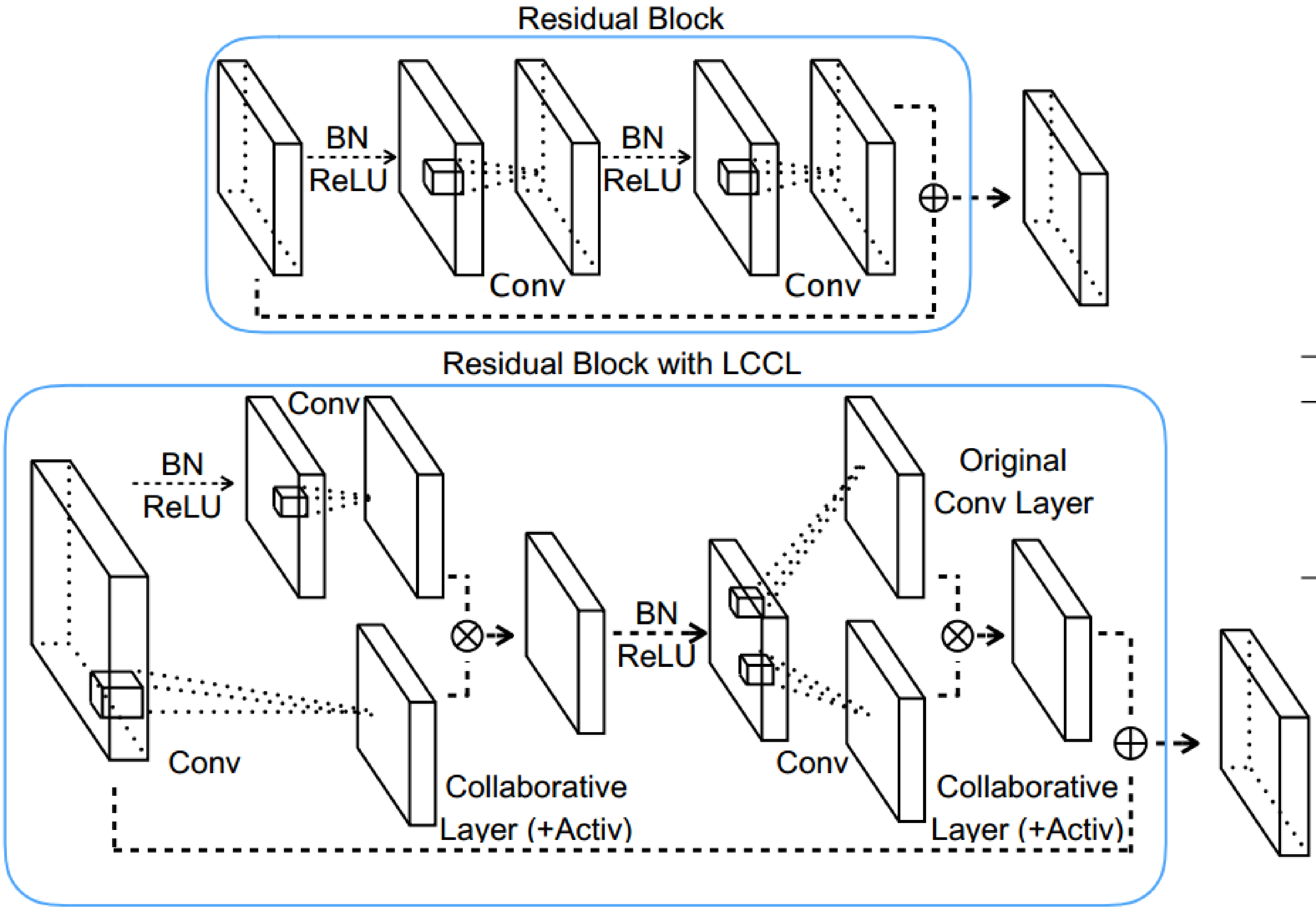
Model	$1 \times 1 \times C \times T$			$k \times k \times C \times 1$		
	FLOPs	Ratio	Error	FLOPs	Ratio	Error
ResNet-20	3.2E7	20.3%	8.57	2.6E7	**34.9%**	**8.32**
ResNet-32	4.7E7	**31.2%**	9.26	4.9E7	28.1%	**7.44**
ResNet-44	6.3E7	**34.8%**	8.57	6.5E7	32.5%	**7.29**

UTS

Details on Pre-Activation Residual Network

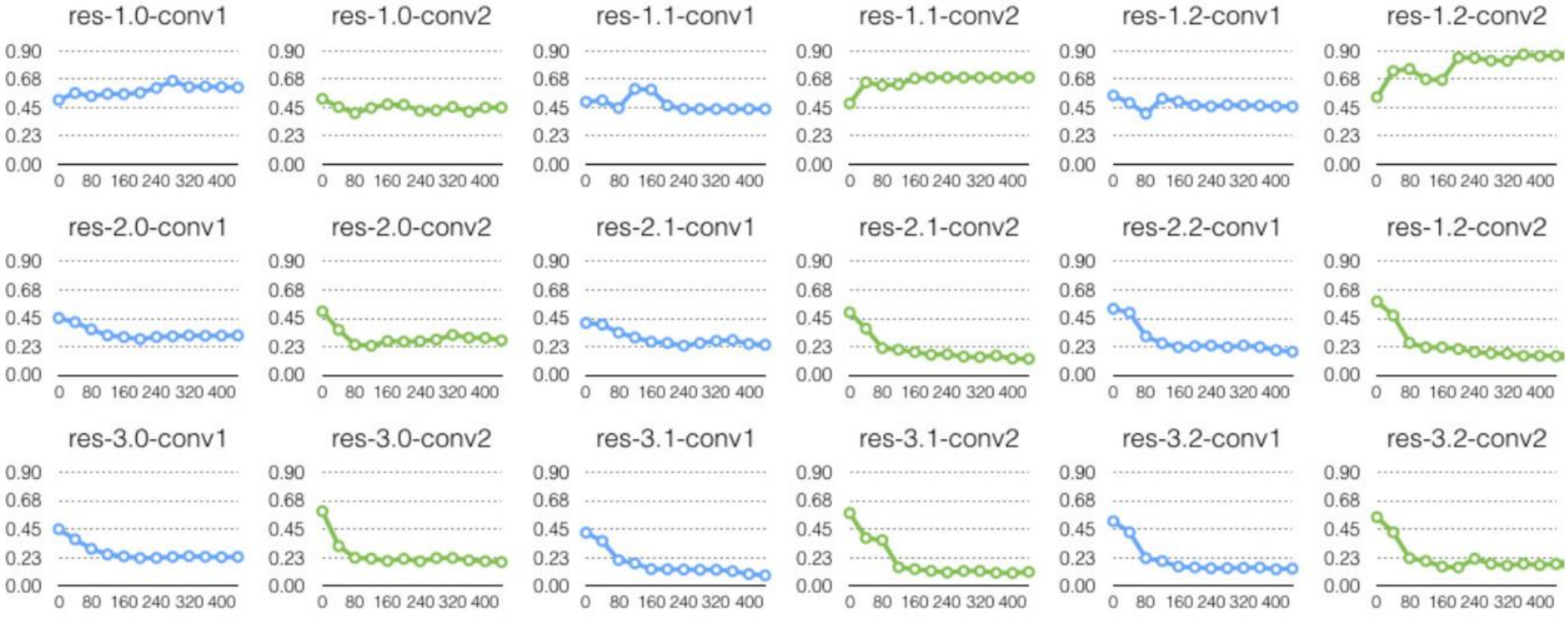


Experiments



Structure	Top-1 Err.	Speed-Up
Aft-Aft	**8.32**	34.9%
Aft-Bef	8.71	24.1%
Bef-Bef	11.62	39.8%
Bef-Aft	12.85	**55.4%**

Experiments – CIFAR10



Experiments – CIFAR10 & CIFAR100

	Depth	Ori. Err	LCCN	Speed-up
ResNet [12]	110	6.37	6.56	34.21%
	164*	5.46	5.91	27.40%
WRN [35]	22-8	4.38	4.90	51.32%
	28-2	5.73	5.81	21.40%
	40-1	6.85	7.65	39.36%
	40-2	5.33	5.98	31.01%
	40-4	4.97	5.95	54.06%
	52-1	6.83	6.99	41.90%

CIFAR10

	Depth	Ori. Err	LCCN	Speed-up
ResNet [12]	164*	24.33	24.74	21.30%
WRN [35]	16-4	24.53	24.83	15.19%
	22-8	21.22	21.30	14.42%
	40-1	30.89	31.32	36.28%
	40-2	26.04	26.91	45.61%
	40-4	22.89	24.10	34.27%
	52-1	29.88	29.55	22.96%

CIFAR100

Experiments – ImageNet

Depth	Approach	Speed-Up	Top-1 Acc. Drop	Top-5 Acc. Drop
18	LCCL	34.6%	3.65	2.30
	BWN	$\approx$ 50.0%	8.50	6.20
	XNOR	$\approx$ 98.3%	18.10	16.00
34	LCCL	24.8%	0.43	0.17
	PFEC	24.2%	1.06	-

Model	FLOPs		Time (ms)		Speed-up	
	CNN	LCCL	CNN	LCCL	Theo	Real
ResNet-18	1.8E9	1.2E9	97.1	77.1	34.6%	20.5%
ResNet-34	3.6E9	2.7E9	169.3	138.6	24.8%	18.1%


Thank You!