SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity through Low-Bit Quantization

Shijie Cao, Lingxiao Ma, Wencong Xiao,

Chen Zhang, Yunxin Liu, Lintao Zhang, Shunnie Lan, Zhi Yang

Today’s DNN model is huge

BERT

Language

- 64TPUv2
- 4 Days
- 1000 GB

or

- 8 P100
- 365 Days
- 1000 GB

Wavenet

Speech

- 2 P100
- 6 Days
- 16 GB

Deformable CNN

Vision

- 8 P100
- 10 Days
- 64 GB

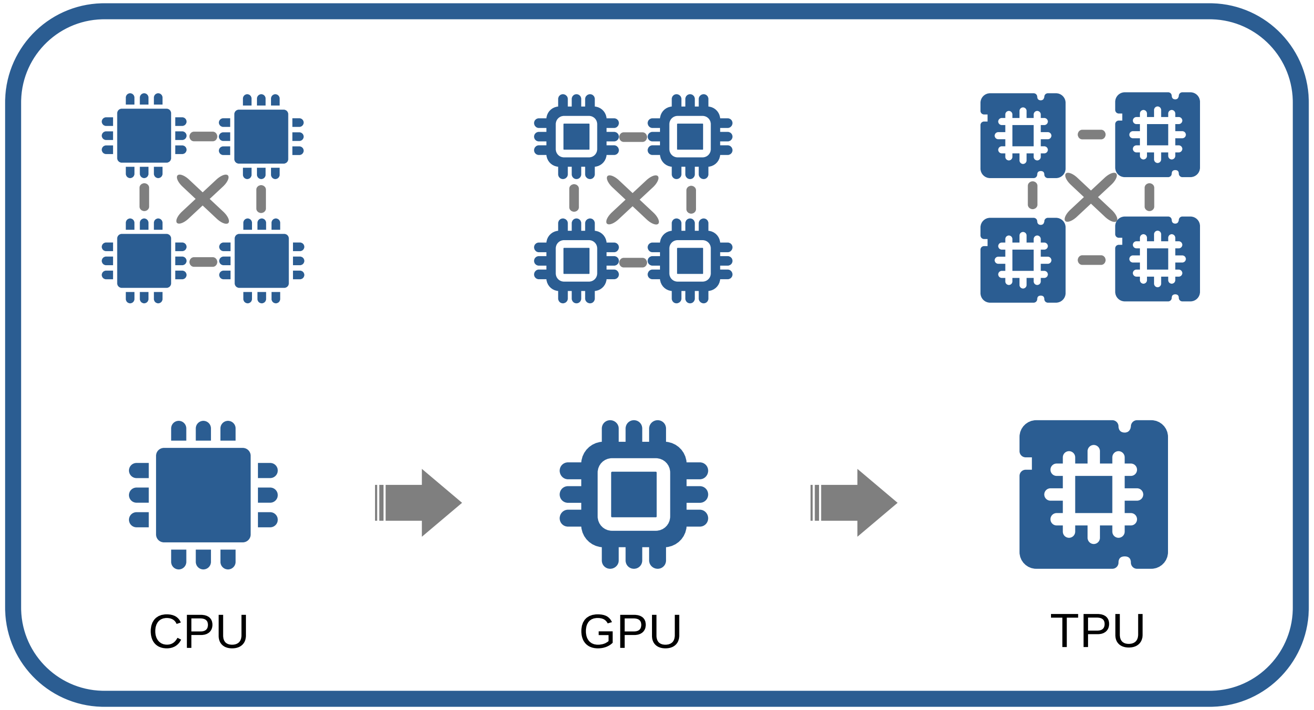
MoE

Language

- 64 K80
- 6 Days
- 1500 GB

What's next technology
that enables us to train a super large model?

not enough



next

Sparsity

Today’s deep learning machine WASTED too much computation and memory
because
neural networks are SPARSE

Redundancy in neural networks

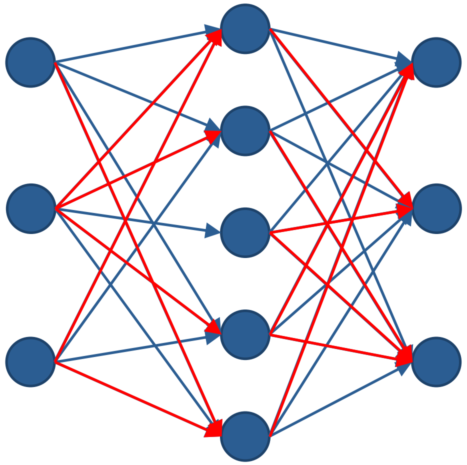
1. Train Connectivity

2. Prune Connections

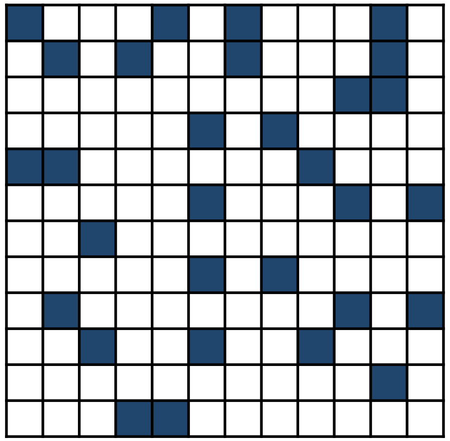
$$\hat{w} \leftarrow abs(w)$$
$$if\ \widehat{w_i} \leq Threshold$$
$$w_i \leftarrow 0$$

3. Train Weights

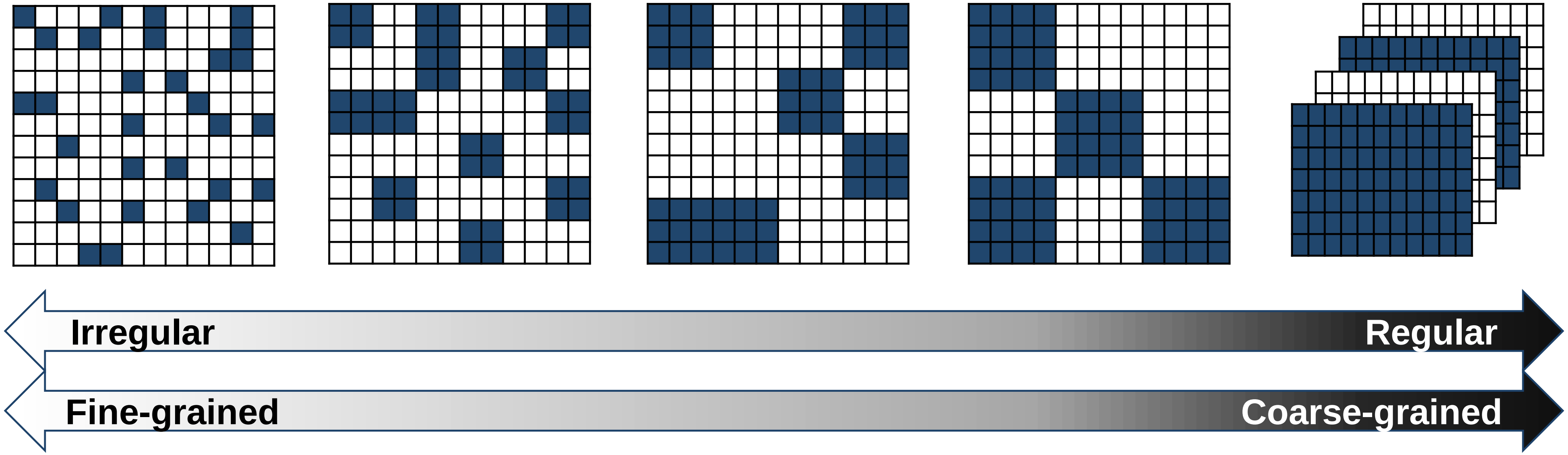
near-zeros



Sparsity = 50%~90%



Highly unstructured sparse pattern



Pros:
- High model accuracy
- High compression rate

Cons:
- Irregular pattern
- Difficult to accelerate

Cons:
- Low model accuracy
- Low compression rate

Pros:
- Regular pattern
- Easy to accelerate

[1] Efficient and Effective Sparse LSTM on FPGA with Bank-Balanced Sparsity, **FPGA ’19**
[2] Balanced Sparsity for Efficient DNN Inference on GPU, **AAAI’19**

Weight Sparsity

- Model compression
- Prunning
- Quantization

?

Structural Sparsity

- MobileNet
- SqueezeNet
- Interleaved Group CNN
- Deformable CNN

Three types of sparsity

Weight Sparsity

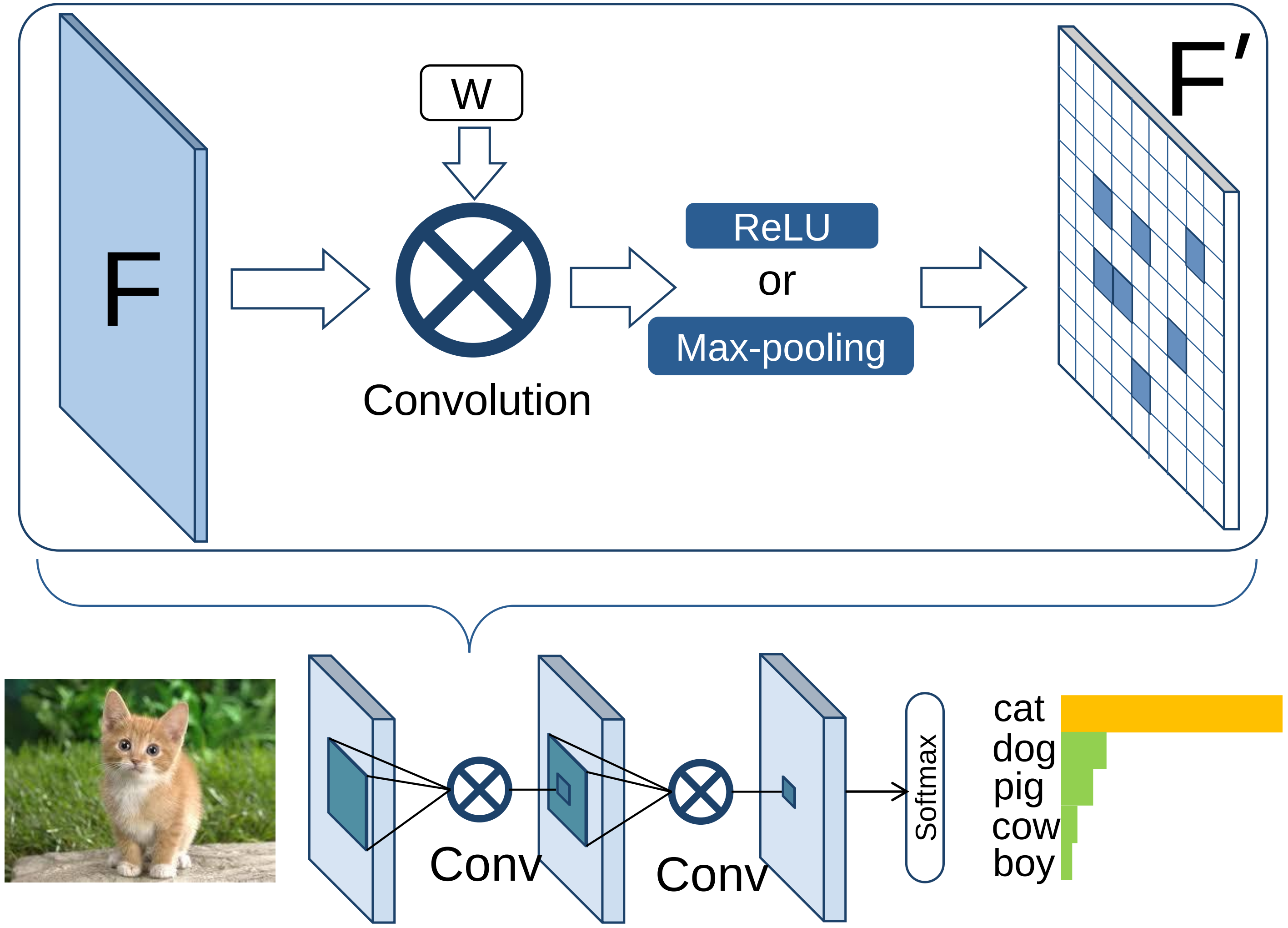
- Model compression
- Prunning
- Quantization

Feature-map Sparsity

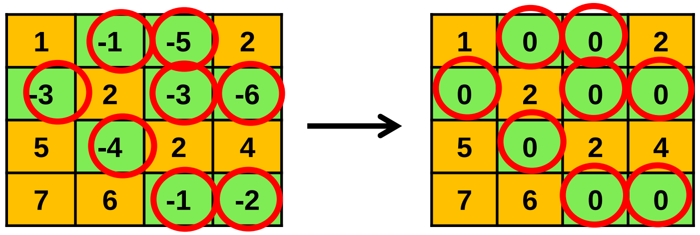
Structural Sparsity

- MobileNet
- SqueezeNet
- Interleaved Group CNN
- Deformable CNN

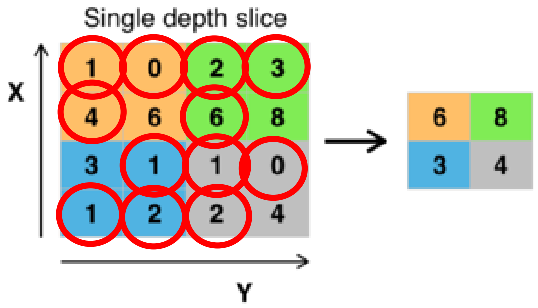
Many pixels of convolution's output are zeros



- ReLU
  - $y = max(0,x)$

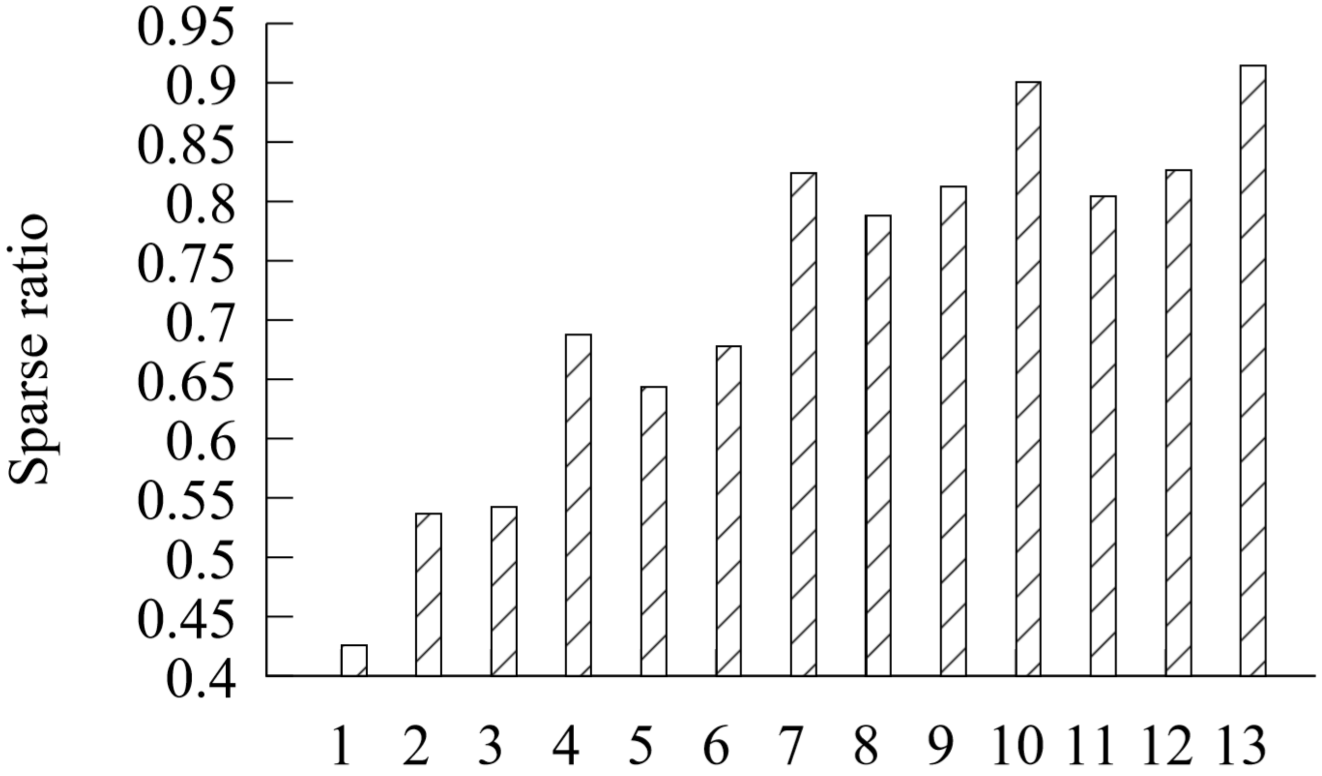


- Max-pooling
  - $y = max(x_i \mid i=\{1,2,\ldots,n\})$



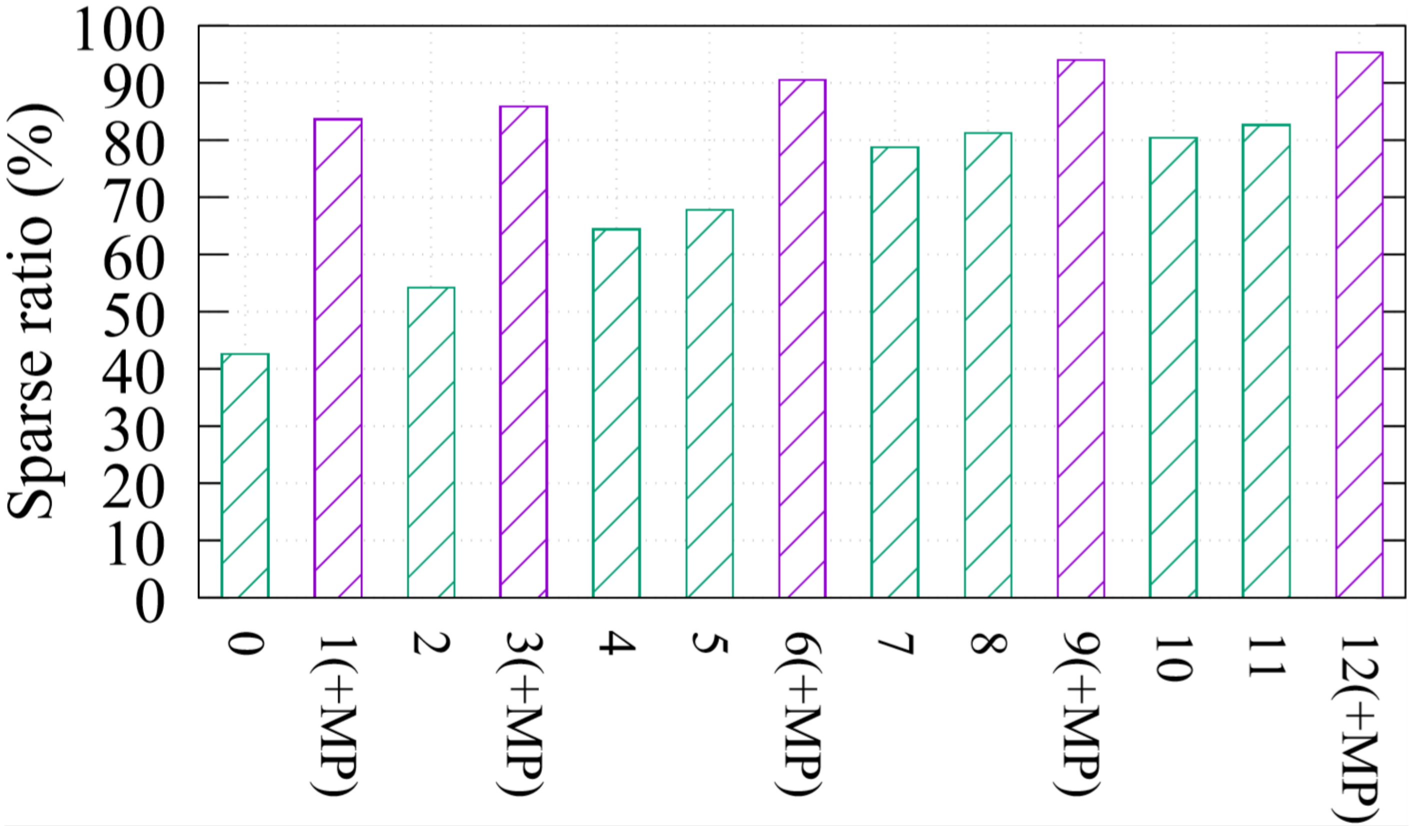
Sparsity case study (Resnet-16): ReLU

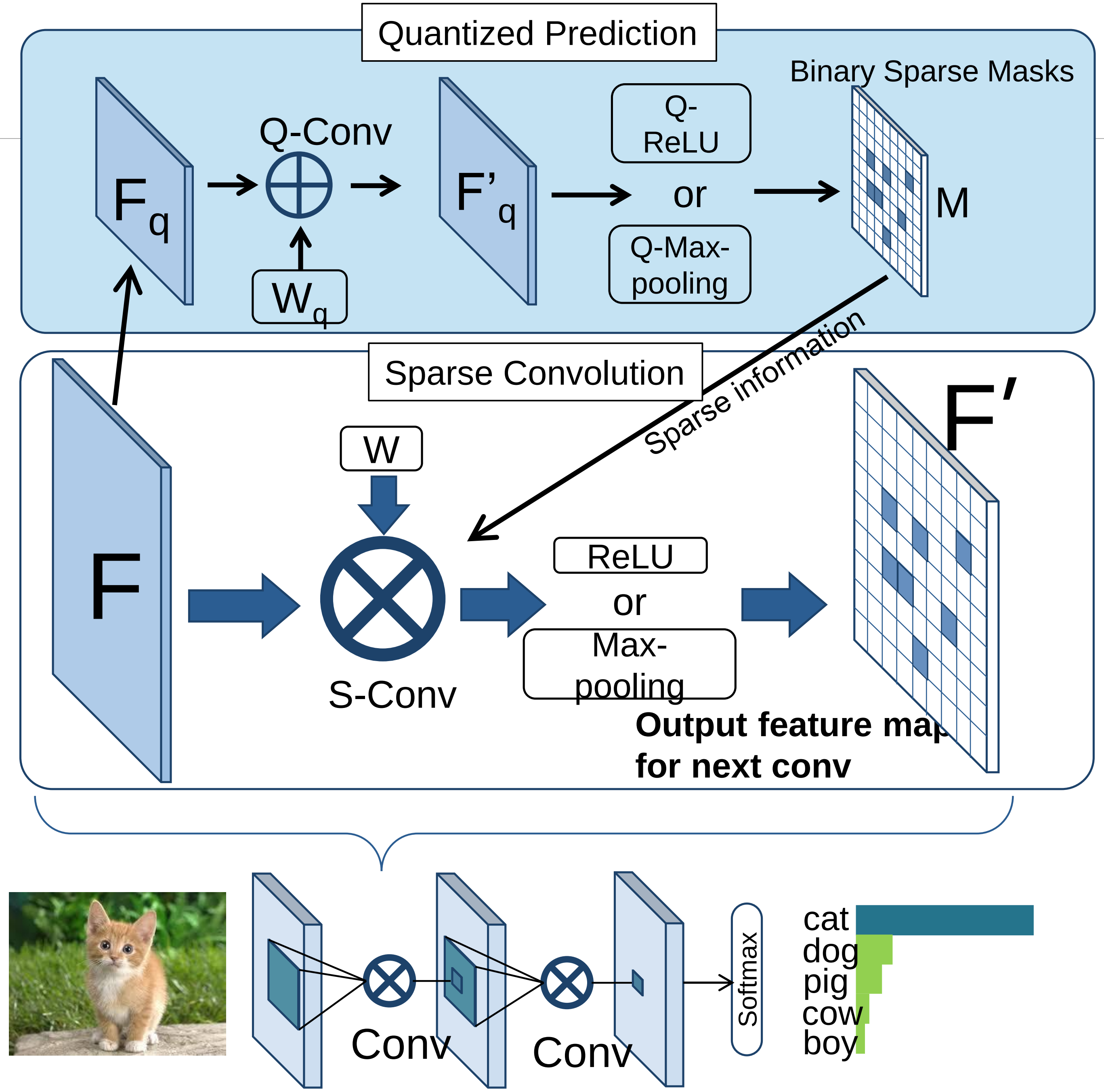
- Sparsity : 45% ~ 95%
- Convolving for ReLU’s zero output pixels results in computation waste



Sparsity case study (VGG16): ReLU + Max-pooling

- Sparsity : 45% ~ 95%
- Convolving for regional small values in max-pooling results in computation waste





Quantized Prediction
Binary Sparse Masks
Q-Conv
F_q
F'_q
W_q
Q-ReLU
or
Q-Max-pooling
M
Sparse Convolution
Sparse information
W
F
S-Conv
ReLU
or
Max-pooling
F'
Output feature map for next conv
Conv
Conv
Softmax
cat
dog
pig
cow
boy

Quantized prediction error rate

ReLU Layer#	1	2	3	4	5	6	7	8	9	10	11	12	13
Prediction Error Rate	4.3%	9.5%	7.0%	4.8%	4.9%	4.1%	2.1%	2.4%	2.2%	1.0%	2.0%	1.7%	0.7%

Quantized prediction error rate of VGG16 on ILSVRC-2012 dataset layer-by-layer with ReLU activation.

Top-1 and Top-5 accuracy of SeerNet with 4-bit quantized prediction

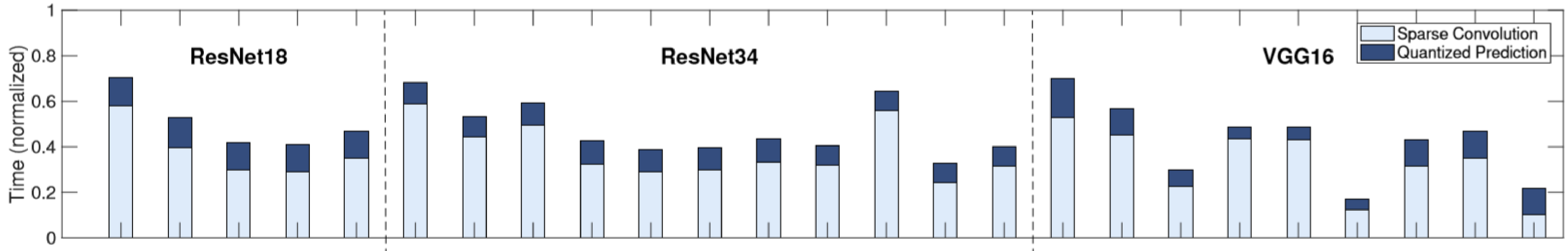
Model	Baseline	SeerNet	Acc. Drop
VGG16	92.57	92.48	0.09
VGG16_BN	93.89	93.60	0.29
ResNet18	93.91	93.88	0.02
ResNet34	94.80	94.76	0.04
InceptionV1	95.12	93.82	1.30

CIFAR-10

Model	Baseline (Top1/Top5)	SeerNet (Top1/Top5)	Acc. Drop (Top1/Top5)
VGG16	71.59/90.38	71.31/90.28	0.28/0.10
VGG16_BN	73.37/91.50	72.85/91.18	0.52/0.32
ResNet18	69.76/89.08	69.34/88.90	0.42/0.18
ResNet34	73.30/91.42	72.95/91.25	0.35/0.17
InceptionV3	77.35/93.62	76.39/92.97	0.96/0.65

ILSVCR-2012

Inference Time and Speedup.



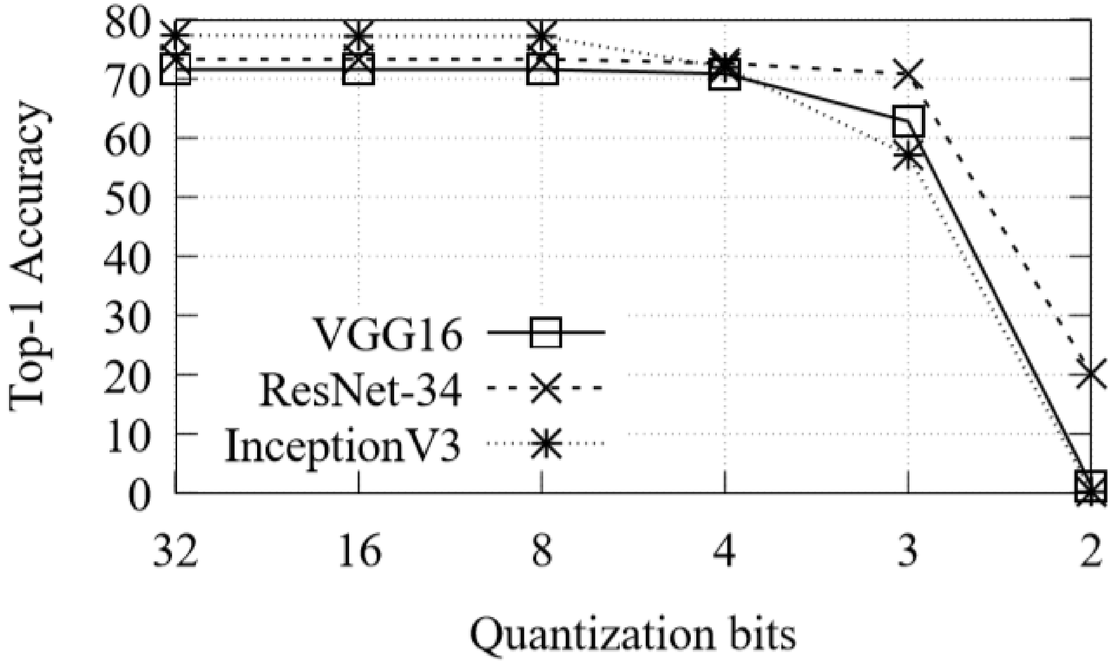
Layer ID	7	11	13	14	15	10	12	16	18	20	22	24	26	27	28	30	1(+MP)	3(+MP)	6(+MP)	7	8	9(+MP)	10	11	12(+MP)
Sparsity	78.9%	83.5%	79.6%	80.2%	76.1%	68.5%	79.4%	84.2%	79.6%	86.4%	87.8%	87.5%	85.9%	78.5%	83.7%	78.3%	83.6%	85.8%	90.5%	79.8%	81.1%	94.0%	80.9%	79.2%	95.3%
Speedup	1.42	1.88	2.39	2.45	2.12	1.46	1.87	1.68	2.35	2.57	2.51	2.31	2.46	1.55	3.05	2.50	1.43	1.76	3.33	2.05	2.05	5.79	2.31	2.13	4.59

The total computation time of the SeerNet is summed up by the computation time spent on sparse convolution and quantized prediction. So the bars are the smaller the better. The speedup is reciprocal to computation time

Comparison with previous work

Model	Method	Top-1 Acc. Drop	Top-5 Acc. Drop	Speedup	Re-train?
ResNet 18	**SeerNet**	**0.42**	**0.18**	30.0%	**No**
	LCCL[2]	3.65	2.30	20.5%	Yes
	BWN[21]	8.50	6.20	50.0%	Yes
	XNOR[22]	18.10	16.00	**98.3%**	Yes
ResNet 34	**SeerNet**	**0.35**	**0.17**	22.2%	**No**
	LCCL[2]	0.43	**0.17**	18.1%	Yes
	PFEC[16]	1.06	-	**24.2%**	Yes
VGG 16	**SeerNet**	**0.28**	**0.10**	**40.1%**	**No**
	PFEC[16]	-	0.15	34.0%	Yes

Sensitivity study of quantization bits



Top-1 accuracy of VGG16, ResNet34 and InceptionV3 with different quantization bits on ILSVRC-2012.

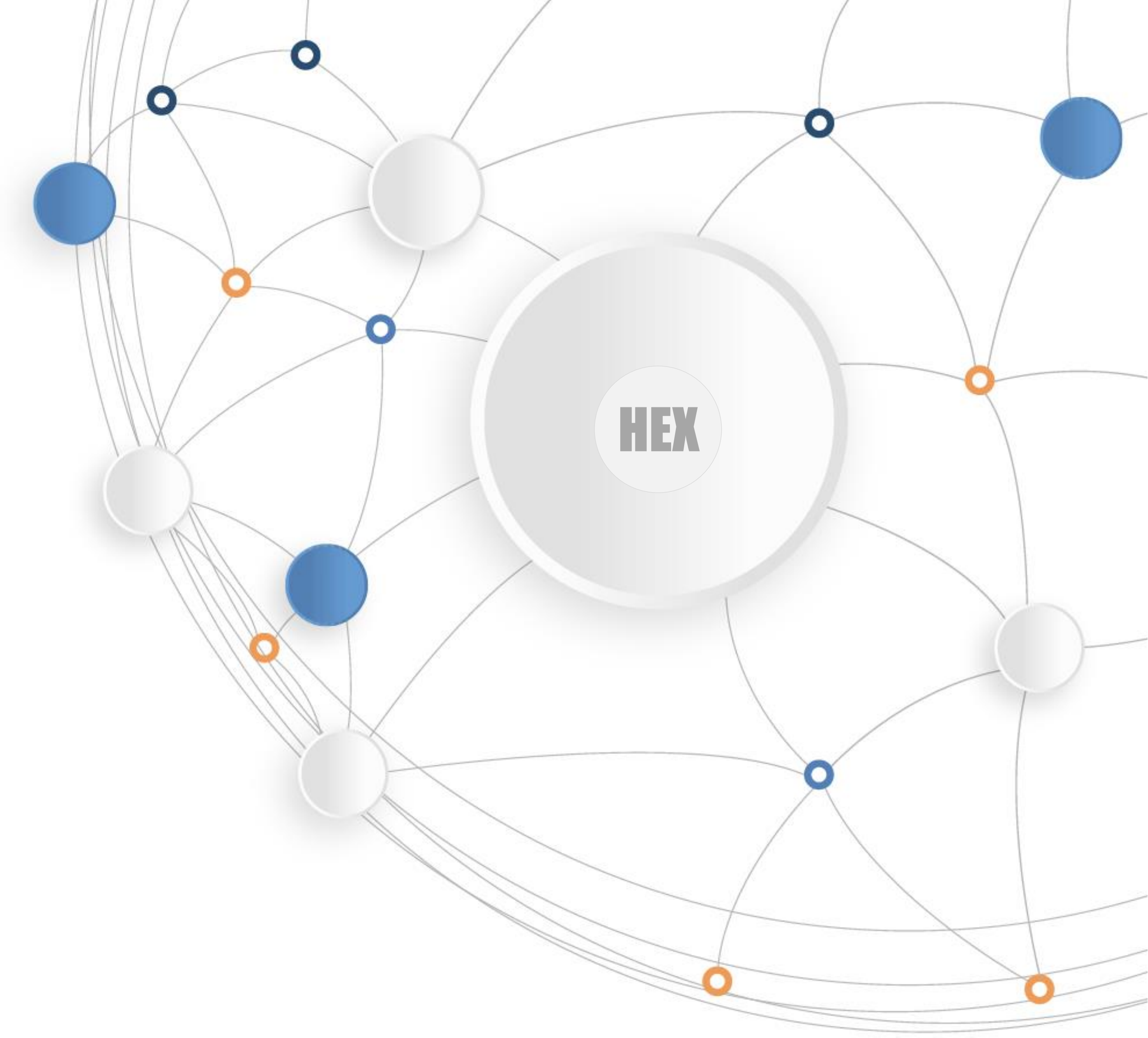
THANKS

Thanks

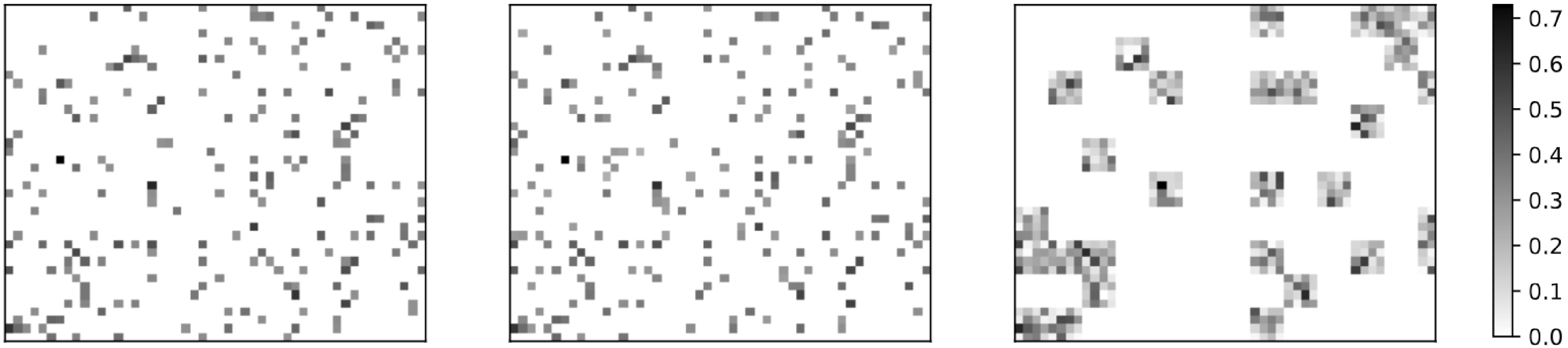
Questions & Comments.

Speaker: Chen Zhang

2018/05/11



Heatmap of weight matrix after pruning



(a) Original Pruning (b) Our Method (c) Block Sparse

Model Accuracy Comparison