中兴捧月神算师算法大赛·总决赛答辩

深度学习模型优化加速

答辩人：YaHei

Hey~YaHei!

目录

CONTENTS

- 赛题介绍
- 压缩方案
- 压缩效果
- 关于内存
- 代码原创性声明

深度学习模型压缩

赛题介绍

$$score = \left(\left(\frac{M - m}{M}\right) \times 20 + \left(\frac{S - s}{s}\right) \times 80\right) \times A(z) \times B(z)$$

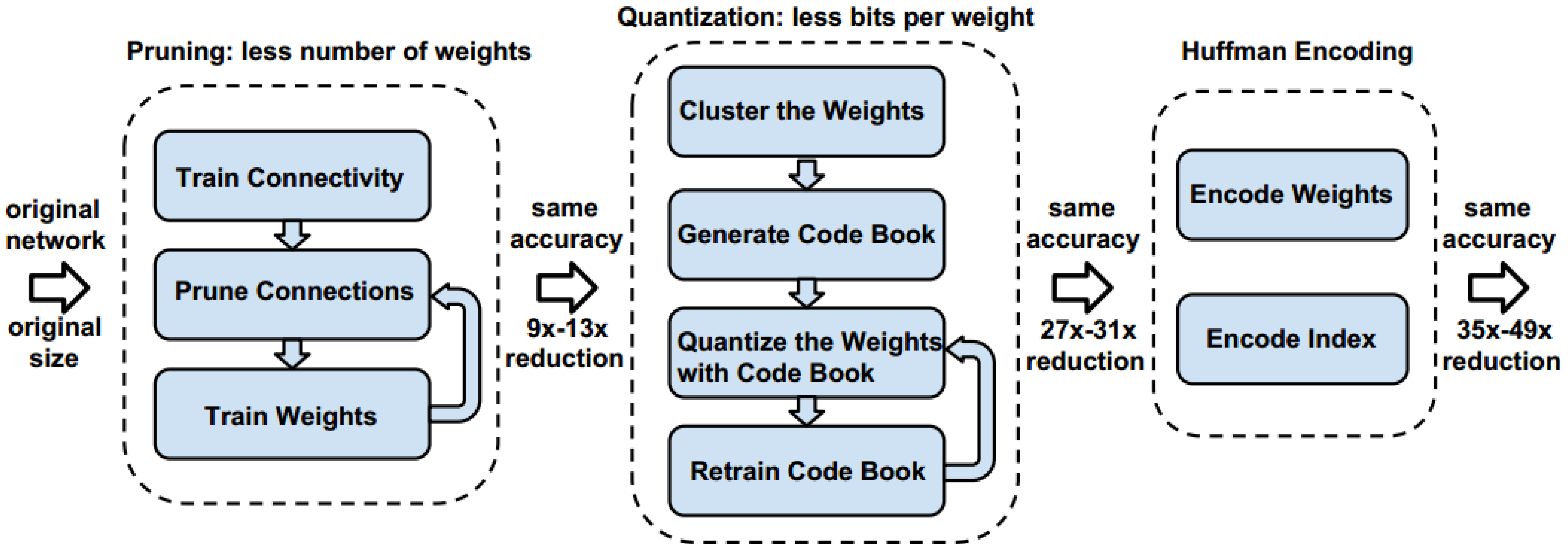
$$\mathrm{A}(z) = \begin{cases} 1, & z \geq 0.97 \\ 0.9, & 0.965 \leq z < 0.97 \\ 0, & z < 0.965 \end{cases}$$

$$\mathrm{B}(z) = \begin{cases} 1, & s \leq 40MB \\ 0.9, & 40MB < s \leq 50MB \\ 0.8, & 50MB < s \leq 63MB \\ 0, & s > 63MB \end{cases}$$

ZTE

压缩方案

整体方案（Deep Compression）

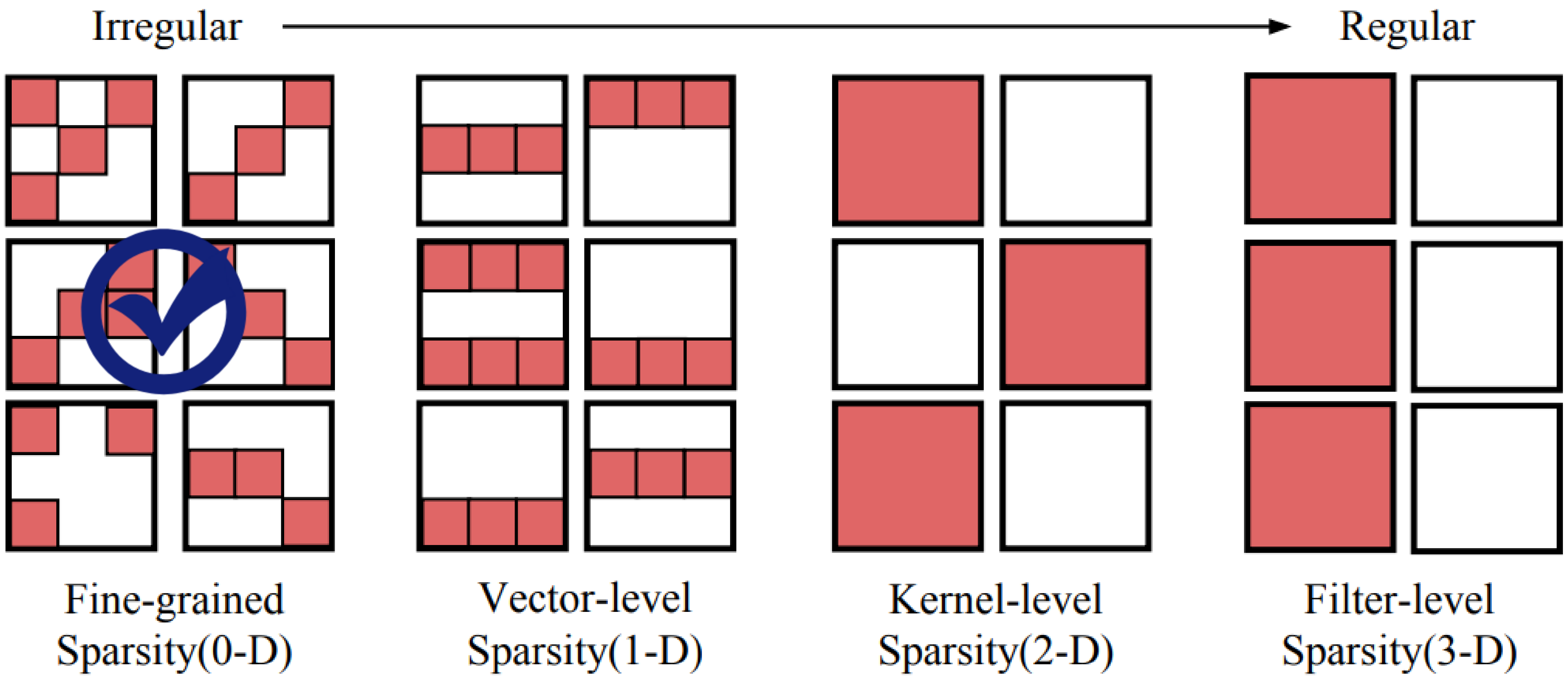


Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding(2016)

ZTE

压缩方案

剪枝：粒度


Irregular
Regular
Fine-grained Sparsity(0-D)
Vector-level Sparsity(1-D)
Kernel-level Sparsity(2-D)
Filter-level Sparsity(3-D)

Exploring the Regularity of Sparse Structure in Convolutional Neural Networks(2017)

压缩方案

剪枝：阈值筛选

人为设置

百分比

敏感度

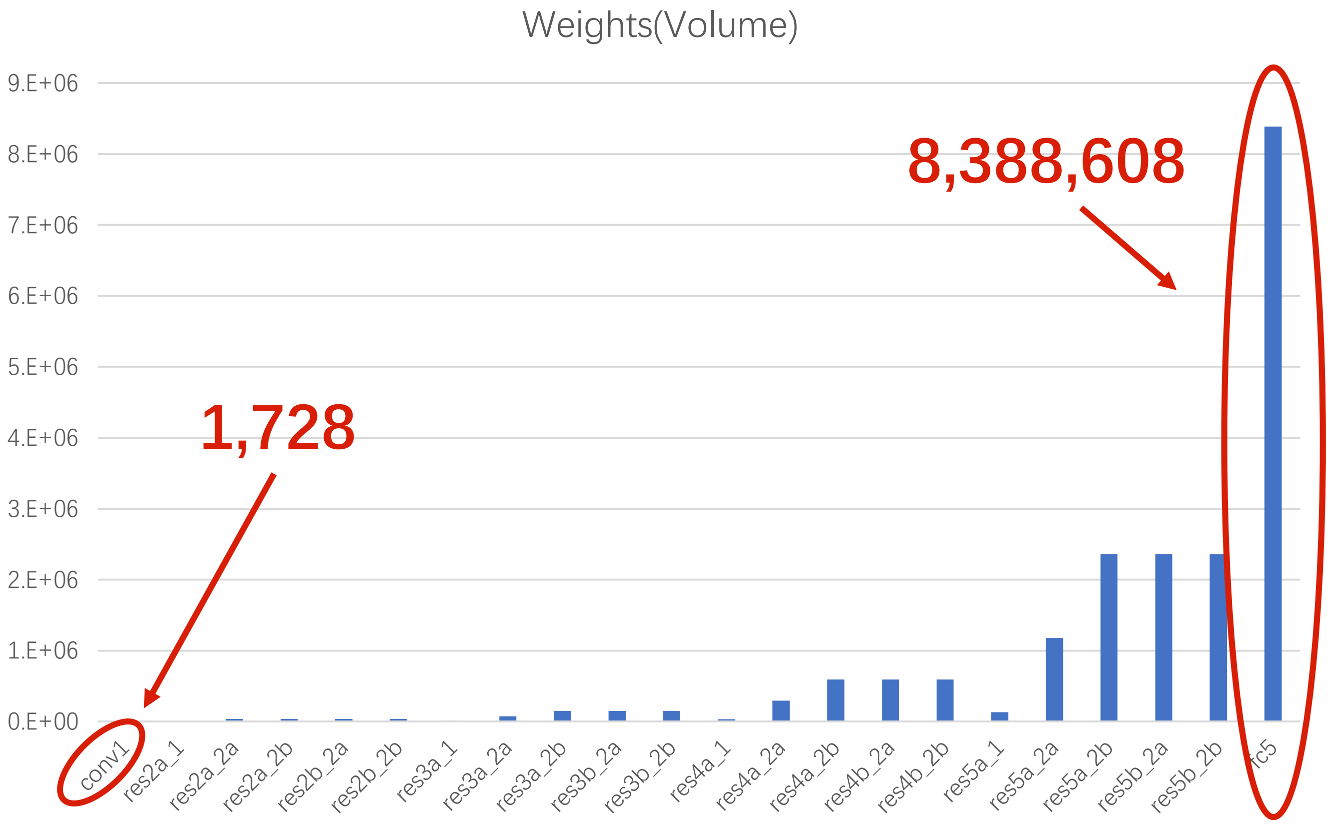
$$threshold = std(weight) * s$$

Learning both Weights and Connections for Efficient Neural Networks(2015)

ZTE

压缩方案

剪枝：阈值筛选



敏感度

$$threshold = std(weight) * s$$

$$s_{conv1} = 0$$

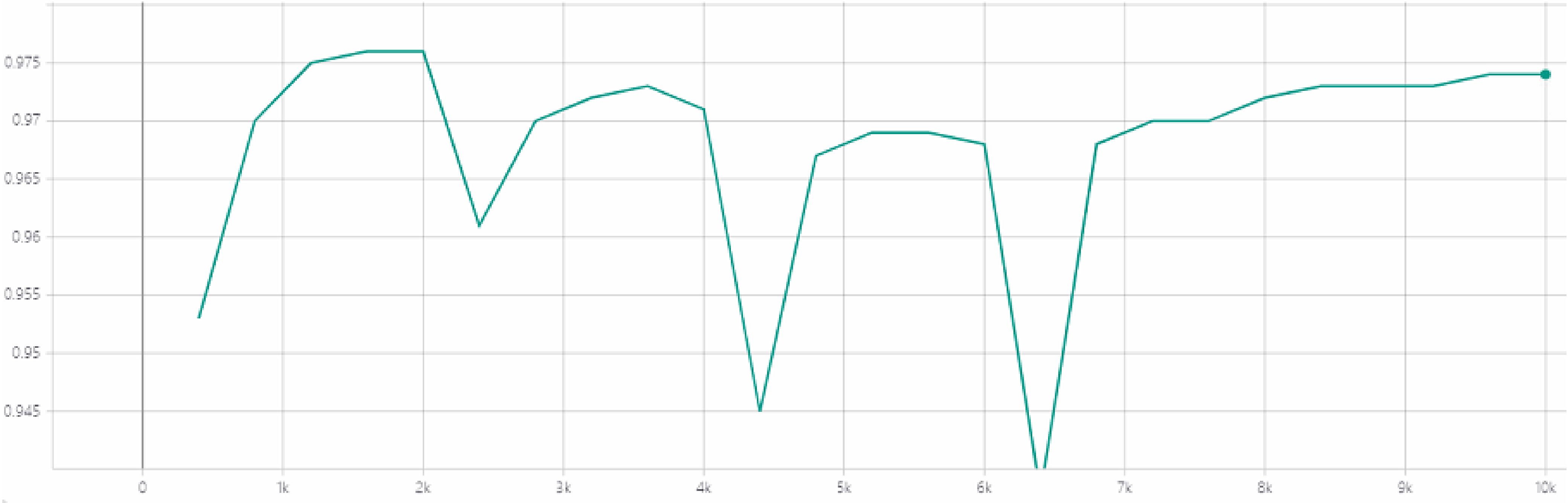
$$s_{fc5} = 2s$$

Learning both Weights and Connections for Efficient Neural Networks(2015)

ZTE

压缩方案

剪枝：恢复训练



To prune, or not to prune: exploring the efficacy of pruning for model compression(2017)

压缩方案

剪枝：知识蒸馏

⊗ $$loss(outputs, y) = \sum ylog(outputs)$$

剪枝：知识蒸馏

压缩方案

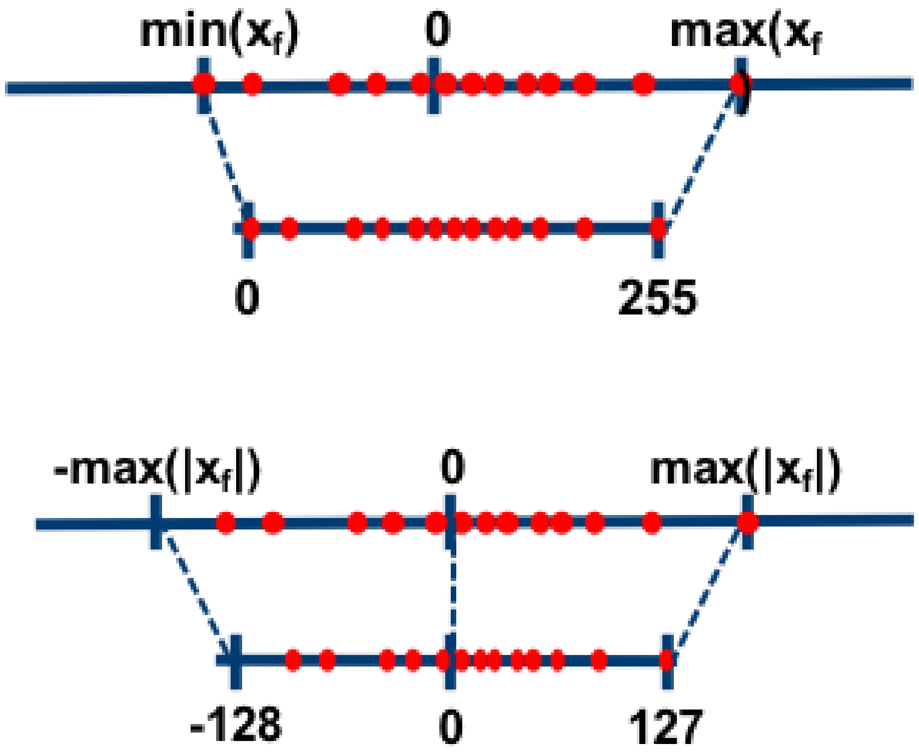
$$\cancel{loss(outputs, y) = \sum ylog(outputs)}$$

$$loss(S, T) = L2(S, T)$$

$$loss(S, T) = CosineDist(S, T)$$

$$loss(S, T) = KL(S, T) = \sum Tlog\frac{T}{S}$$ ✓

量化：非均匀分布

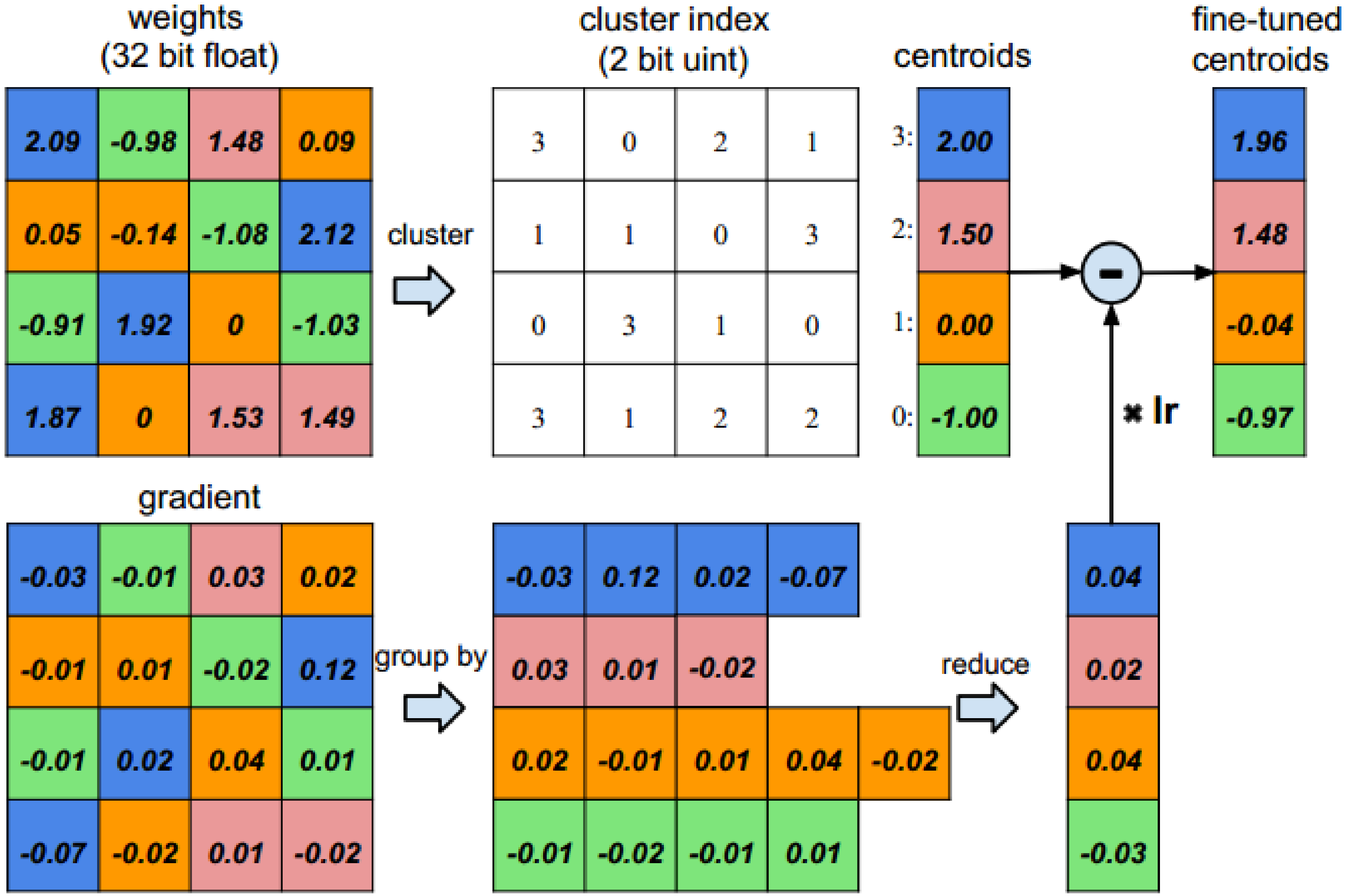


Quant	Real
0	min
1	min+△
2	min+2△
…	…
254	max-△
255	max

Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference(2018)

压缩方案

量化：重训练



weights
(32 bit float)

2.09	-0.98	1.48	0.09
0.05	-0.14	-1.08	2.12
-0.91	1.92	0	-1.03
1.87	0	1.53	1.49

cluster index
(2 bit uint)

3	0	2	1
1	1	0	3
0	3	1	0
3	1	2	2

centroids

3:	2.00
2:	1.50
1:	0.00
0:	-1.00

fine-tuned
centroids

1.96
1.48
-0.04
-0.97

gradient

-0.03	-0.01	0.03	0.02
-0.01	0.01	-0.02	0.12
-0.01	0.02	0.04	0.01
-0.07	-0.02	0.01	-0.02

-0.03	0.12	0.02	-0.07	
0.03	0.01	-0.02		
0.02	-0.01	0.01	0.04	-0.02
-0.01	-0.02	-0.01	0.01	

0.04
0.02
0.04
-0.03

Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding(2016)

压缩方案

稀疏存储

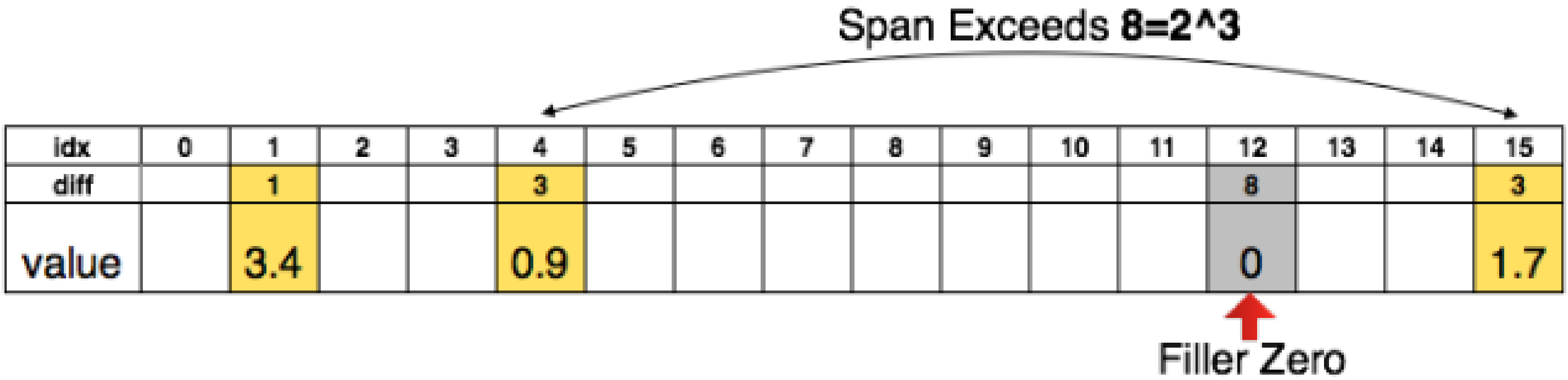
	0	1	2	3
0				
1				
2				
3				

(row, col, data)

稀疏存储

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

(index, data)



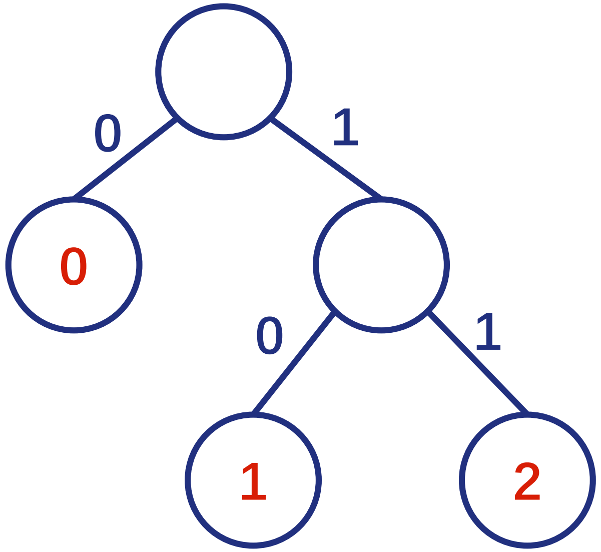
idx	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
diff		1			3								8			3
value		3.4			0.9								0			1.7

Span Exceeds 8=2^3
Filler Zero

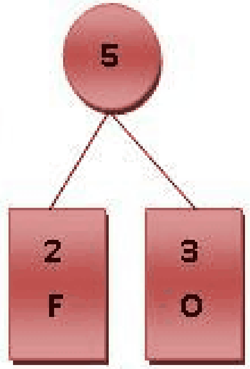
Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding(2016)

压缩方案

哈夫曼编码



{0, 1, 0, 2, 0, 0, 0}
00010010000000
010011000



Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding(2016)

压缩效果

压缩率分析

模型大小：74.5MB => 6.9MB
准确率：97.3% => 97.2%

Layer	Sparsity	Weight Bits	Weight Bits(H)	Index Bits	Index Bits(H)	Rate(P+Q)	Rate(P+Q+H)
conv1	--	--	--	--	--	--	--
res2a_1	77.56%	7	8.21	5	3.35	91.59%	91.89%
res2a_2a	82.18%	7	6.69	5	2.65	93.32%	94.79%
res2a_2b	67.48%	7	6.18	5	2.36	87.81%	91.32%
res2b_2a	55.63%	7	6.23	5	2.13	83.36%	88.40%
res2b_2b	53.96%	7	6.18	5	1.96	82.73%	88.28%
res3a_1	58.36%	7	6.83	5	2.40	84.39%	87.99%
res3a_2a	48.08%	7	5.97	5	1.91	80.53%	87.21%
res3a_2b	52.51%	7	5.80	5	2.04	82.19%	88.35%
res3b_2a	52.68%	7	5.85	5	2.07	82.25%	88.28%
res3b_2b	56.35%	7	5.76	5	2.04	83.63%	89.37%

ZTE

压缩效果

压缩率分析

模型大小：74.5MB => 6.9MB
准确率：97.3% => 97.2%

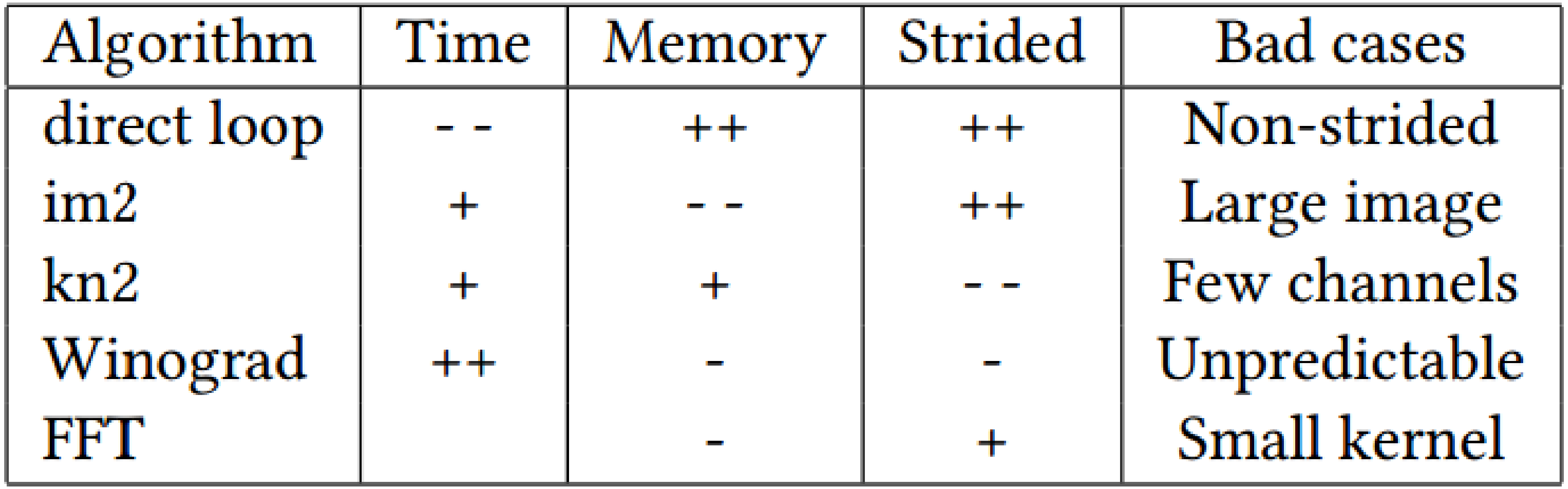
（续表）

Layer	Sparsity	Weight Bits	Weight Bits(H)	Index Bits	Index Bits(H)	Rate(P+Q)	Rate(P+Q+H)
res4a_1	54.75%	7	6.17	5	2.21	83.03%	88.14%
res4a_2a	47.52%	7	5.53	5	1.90	80.32%	87.82%
res4a_2b	52.11%	7	6.01	5	2.06	82.04%	87.92%
res4b_2a	48.99%	7	5.70	5	1.96	80.87%	87.80%
res4b_2b	51.80%	7	5.49	5	1.99	81.93%	88.73%
res5a_1	53.95%	7	5.92	5	2.14	82.73%	88.41%
res5a_2a	48.01%	7	5.75	5	1.92	80.51%	87.54%
res5a_2b	46.93%	7	5.64	5	1.88	80.10%	87.51%
res5b_2a	48.41%	7	5.67	5	1.94	80.65%	87.73%
res5b_2b	49.19%	7	5.57	5	1.94	80.95%	88.08%
fc5	73.97%	4	3.33	5	3.19	92.68%	94.69%
total	59.79%					85.97%	90.81%

关于内存

计算策略

（以卷积为例）

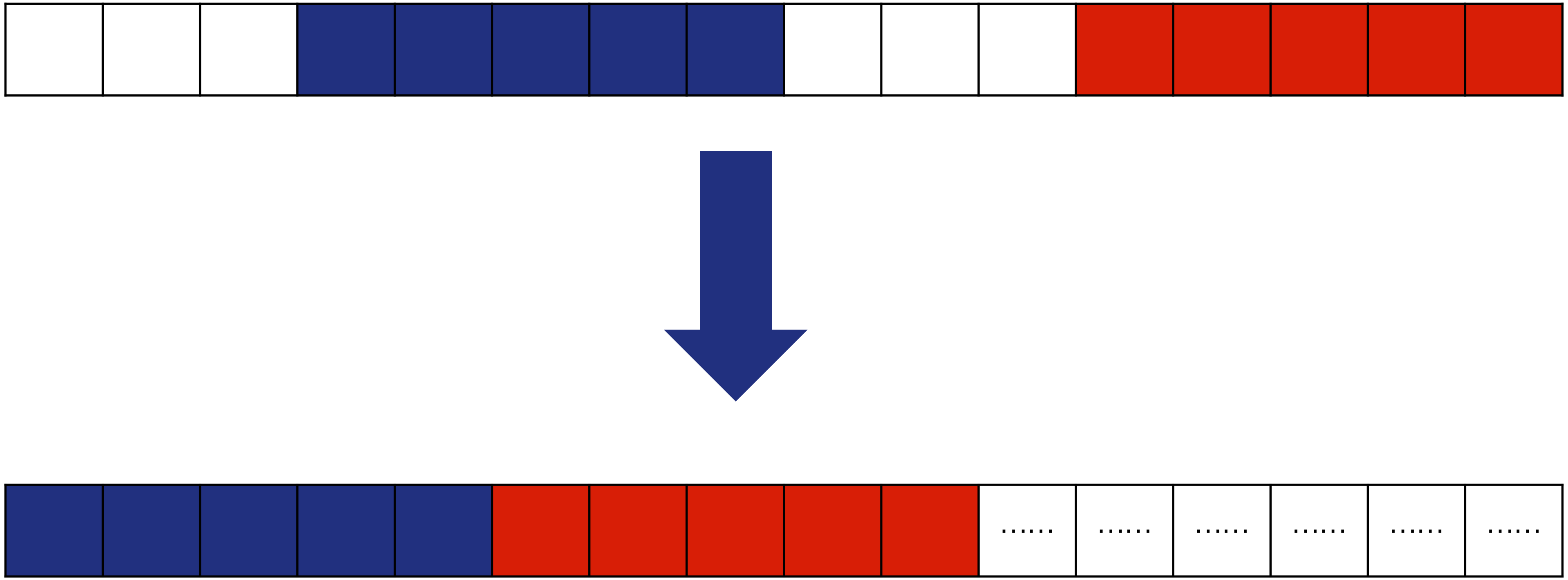


Algorithm	Time	Memory	Strided	Bad cases
direct loop	- -	++	++	Non-strided
im2	+	- -	++	Large image
kn2	+	+	- -	Few channels
Winograd	++	-	-	Unpredictable
FFT		-	+	Small kernel

Optimal DNN Primitive Selection with Partitioned Boolean Quadratic Programming(2017)

关于内存

紧凑内存



ZTE

代码原创性声明

Hey~YaHei!

感谢！

Thanks!