Network compression

Relevance

Apps, self-driving cars, VR etc.

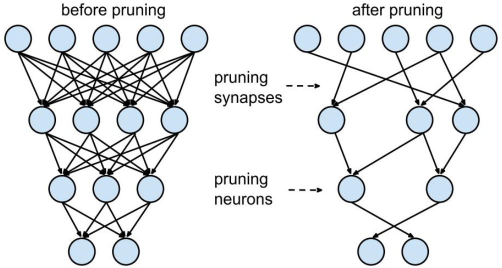
Problems:

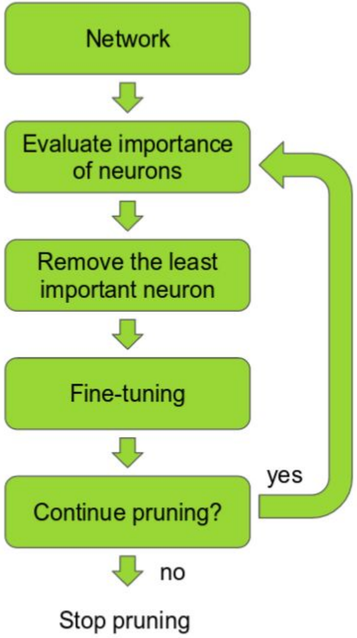
- Delay
- Memory
- Energy

Pruning

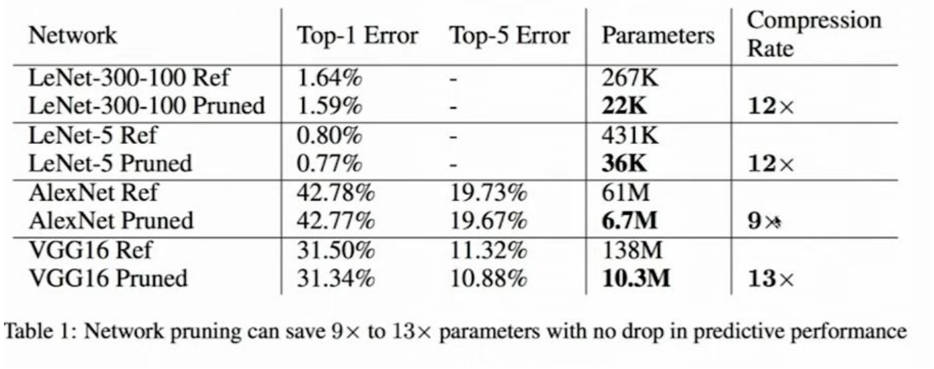
AGE

Рождение	50 трлн
1 год	1000 трлн
10 лет	500 трлн




Network
Evaluate importance of neurons
Remove the least important neuron
Fine-tuning
Continue pruning?
yes
no
Stop pruning

Эксперименты с Pruning



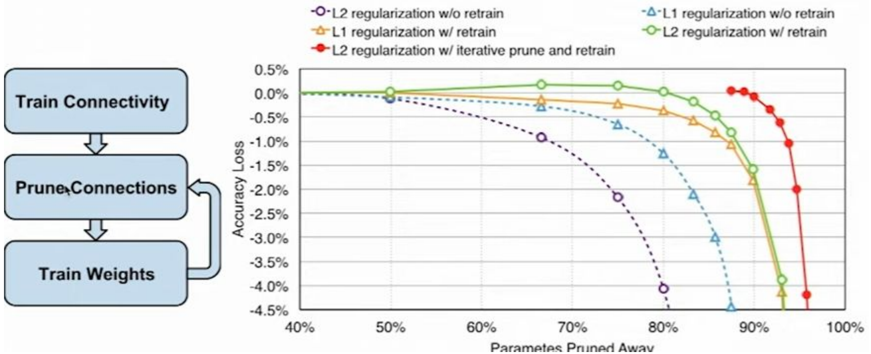
Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
LeNet-300-100 Ref	1.64%	-	267K	
LeNet-300-100 Pruned	1.59%	-	**22K**	12×
LeNet-5 Ref	0.80%	-	431K	
LeNet-5 Pruned	0.77%	-	**36K**	12×
AlexNet Ref	42.78%	19.73%	61M	
AlexNet Pruned	42.77%	19.67%	**6.7M**	9×
VGG16 Ref	31.50%	11.32%	138M	
VGG16 Pruned	31.34%	10.88%	**10.3M**	13×

Table 1: Network pruning can save 9× to 13× parameters with no drop in predictive performance

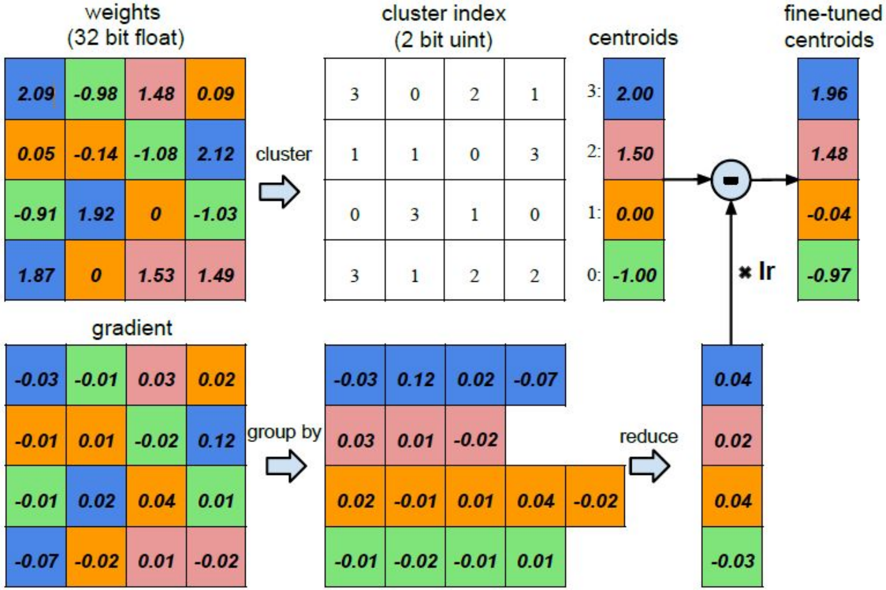
Examples, showing pruning

effectiveness

Retrain to recover accuracy



Quantization



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	fine-tuned centroids
3:	2.00	1.96
2:	1.50	1.48
1:	0.00	-0.04
0:	-1.00	-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

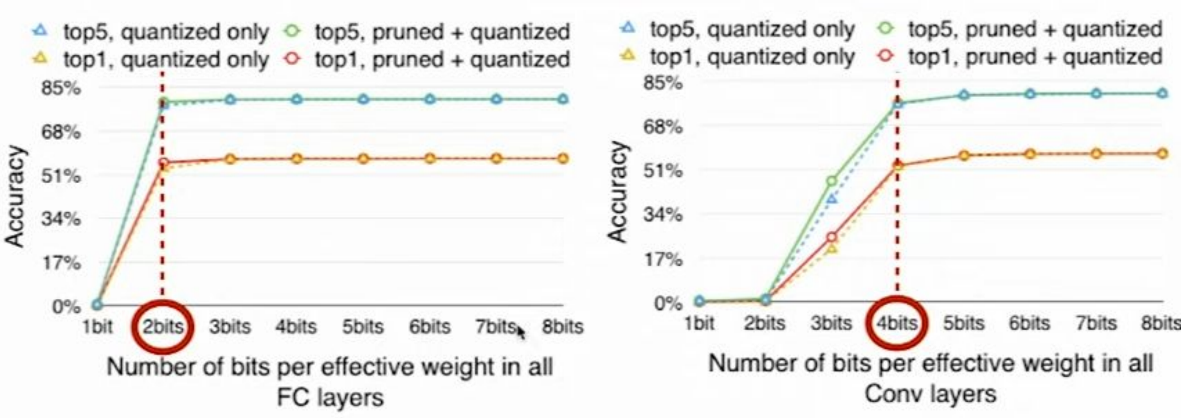
group by

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

$$r = \frac{nb}{nlog_2(k) + kb}$$

Compression rate

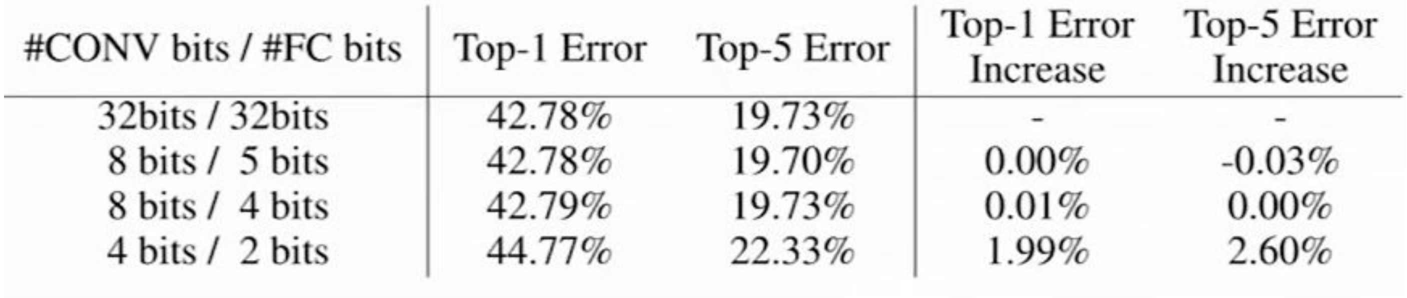
Эксперименты с Quantization



Bits per weight

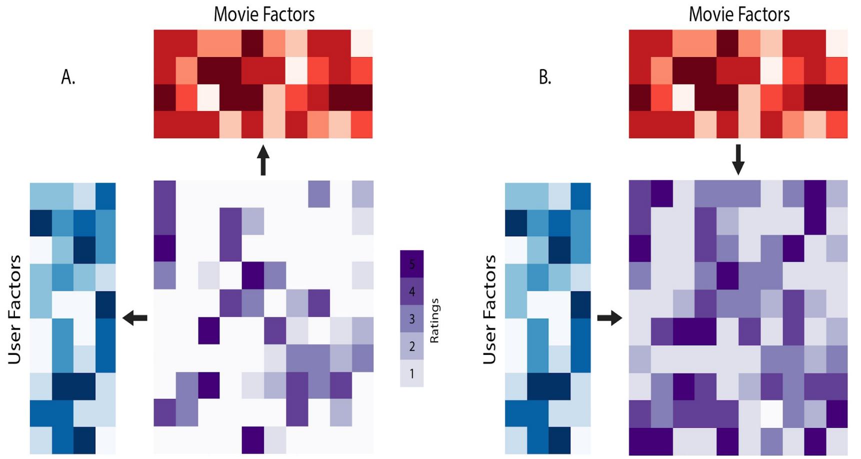
Pruning +

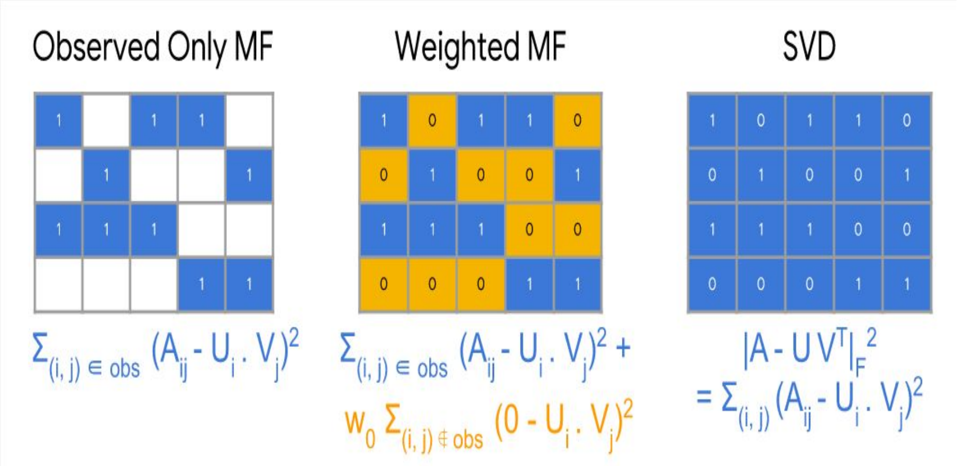
Quantization



#CONV bits / #FC bits	Top-1 Error	Top-5 Error	Top-1 Error Increase	Top-5 Error Increase
32bits / 32bits	42.78%	19.73%	-	-
8 bits / 5 bits	42.78%	19.70%	0.00%	-0.03%
8 bits / 4 bits	42.79%	19.73%	0.01%	0.00%
4 bits / 2 bits	44.77%	22.33%	1.99%	2.60%

Matrix factorization





Observed Only MF

1		1	1	
	1			1
1	1	1		
			1	1

$$\Sigma_{(i,\, j)\, \in\, obs} (A_{ij} - U_i \,.\, V_j)^2$$

Weighted MF

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

$$\Sigma_{(i,\, j)\, \in\, obs} (A_{ij} - U_i \,.\, V_j)^2 + w_0\, \Sigma_{(i,\, j)\, \notin\, obs} (0 - U_i \,.\, V_j)^2$$

SVD

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

$$|A - U V^T|_F^{\,2} = \Sigma_{(i,\, j)} (A_{ij} - U_i \,.\, V_j)^2$$

Full matrix: O(nm) ⟶ O((n + m)d), d - dimension.