Network compression

Relevance

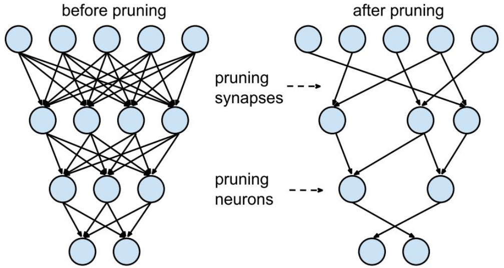
Apps, self-driving cars, VR etc.

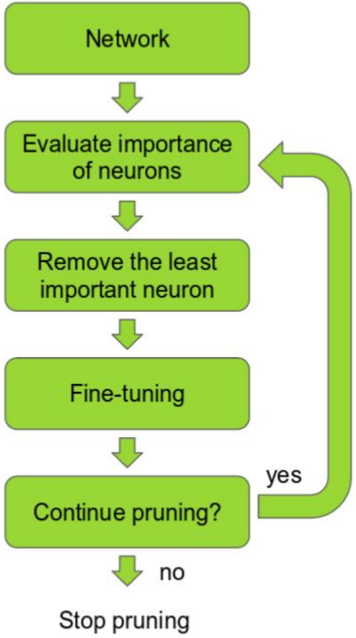
Problems:
- Delay
- Memory
- Energy

Pruning

AGE

Рождение	50 трлн
1 год	1000 трлн
10 лет	500 трлн





Эксперименты с Pruning

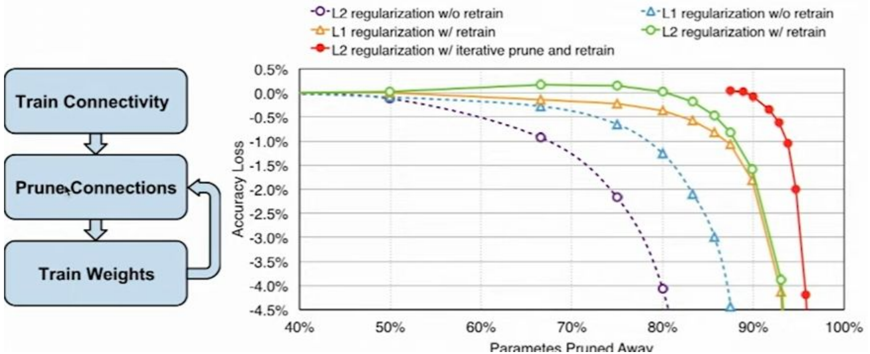
Network	Top-1 Error	Top-5 Error	Parameters	Compression Rate
LeNet-300-100 Ref	1.64%	-	267K	
LeNet-300-100 Pruned	1.59%	-	**22K**	**12×**
LeNet-5 Ref	0.80%	-	431K	
LeNet-5 Pruned	0.77%	-	**36K**	**12×**
AlexNet Ref	42.78%	19.73%	61M	
AlexNet Pruned	42.77%	19.67%	**6.7M**	**9×**
VGG16 Ref	31.50%	11.32%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

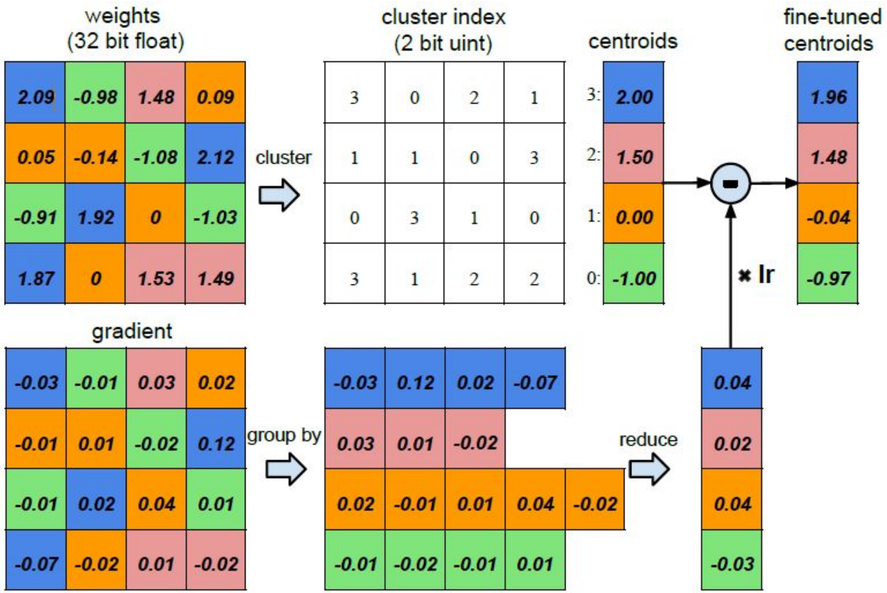
Examples, showing pruning

effectiveness

Retrain to recover accuracy



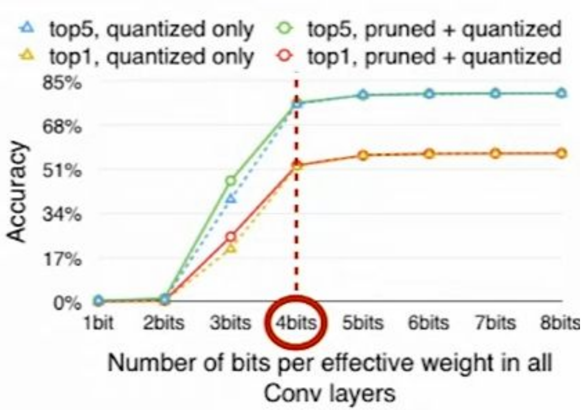
Quantization



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

Compression rate

Эксперименты с Quantization



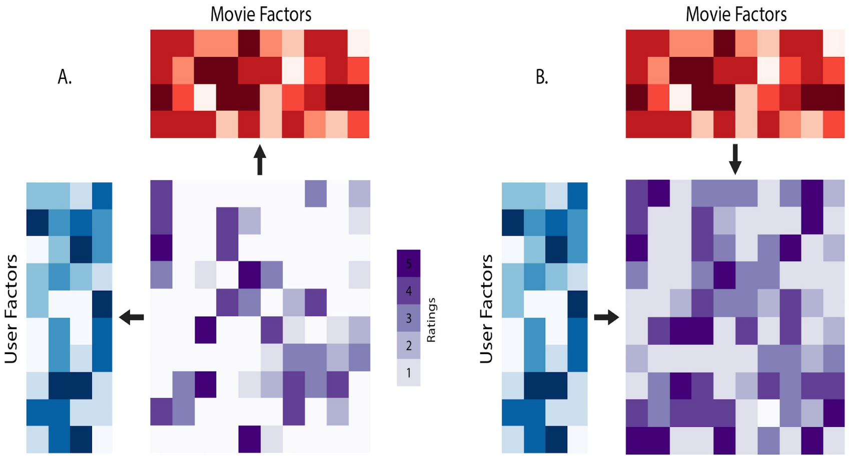
Bits per weight

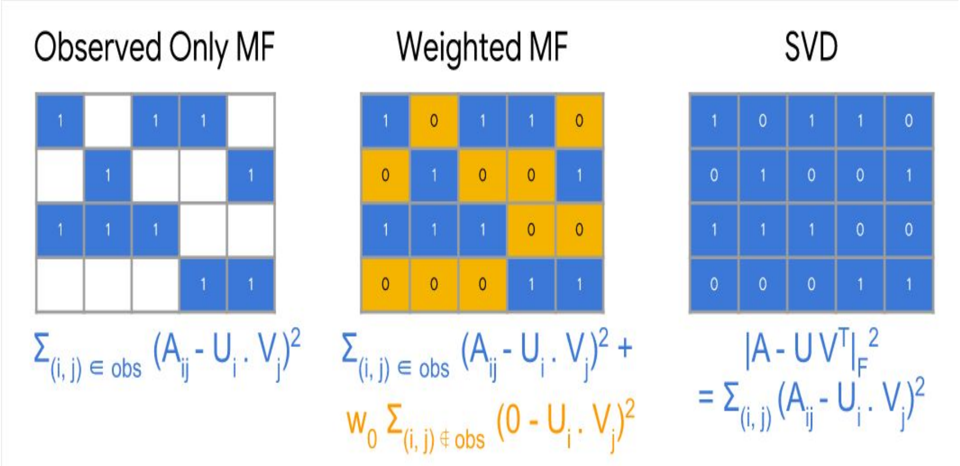
Pruning +

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





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

Список литературы

- **Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding**

  Song Han, Huizi Mao, William J. Dally

- **PRUNING CONVOLUTIONAL NEURAL NETWORKS FOR RESOURCE EFFICIENT INFERENCE**

  Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, Jan Kautz NVIDIA

- **NEURAL NETWORK MATRIX FACTORIZATION**

  Gintare Karolina Dziugaite, Daniel M. Roy