Compressing Deep Neural Networks for Recognizing Places

Soham Saha, Girish Varma, C.V. Jawahar

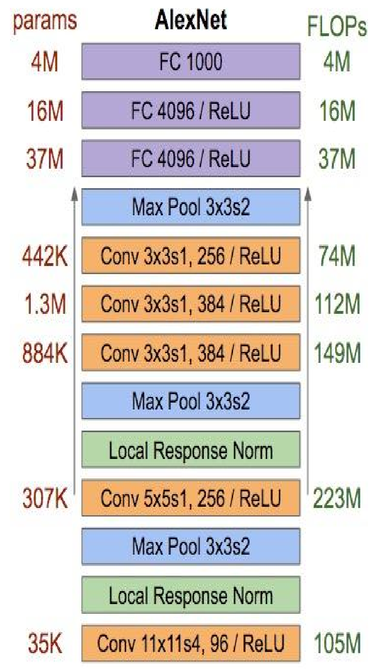
ACPR 2017

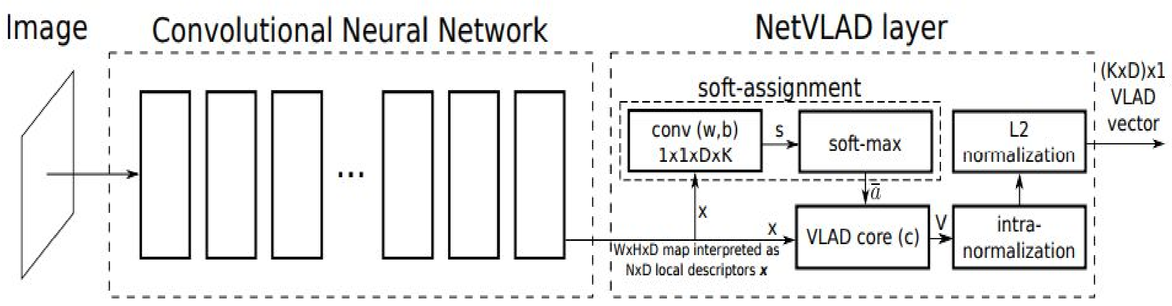
MOTIVATION

- Deep Neural network based place recognition models need a lot of memory.
- But, a tourist takes a picture of a place and expects it to run on his/her mobile phone.
- Alexnet has 60M parameters (~240MB on disk) and performs 1.5 billion FLOPs in one forward pass.
- VGG-16 has ~140M Parameters (~530MB)

VISUAL PLACE RECOGNITION

- Fine-tuned with Paris Buildings dataset and tested with Oxford Buildings dataset and vice-versa.
- Learning to recognize a place by retrieving images of the same place.

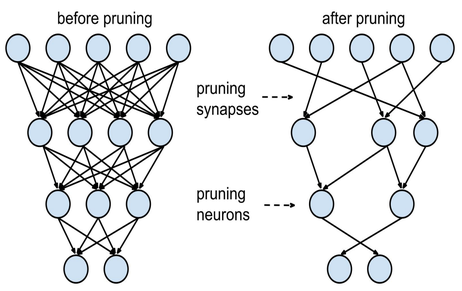




CONTRIBUTION

- Iterative Pruning, Quantization and Sparsification (on every layer of the network) for Image Retrieval.

1. **Threshold based pruning.**

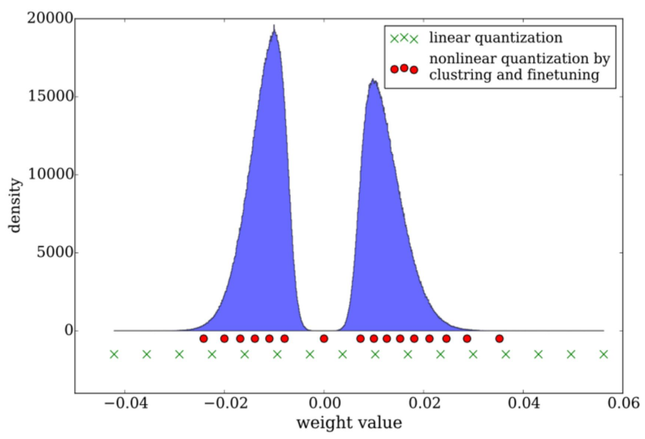


Triplet loss

$$Loss = \frac{1}{2}h(m + ||q-p||^2 - ||q-n||^2)$$

2. **Uniform and non-uniform Quantization.**

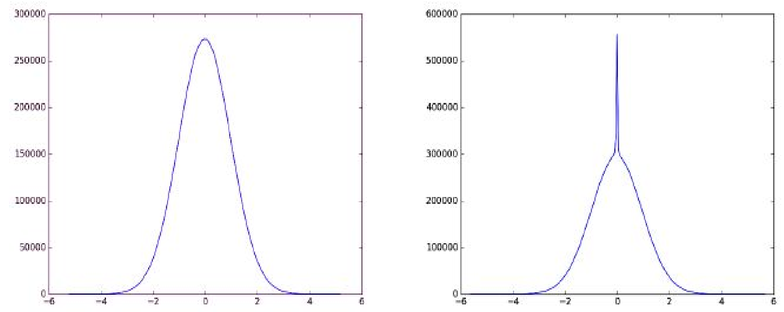
- Non-uniform quantization is performed via k-mean clustering on the weight space



$$J(\theta) = Loss + \lambda \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} |W_{ji}^l|$$

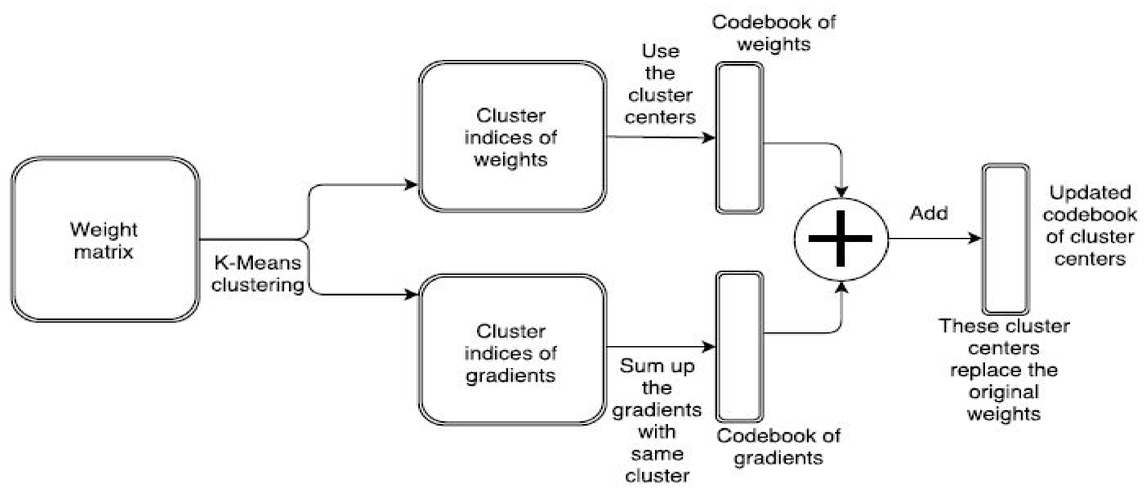
3. **Sparsification by using an L1 Regularizer.**

The effect of the regularizer on the weight space can be seen here.



4. **Trained quantization Algorithm**

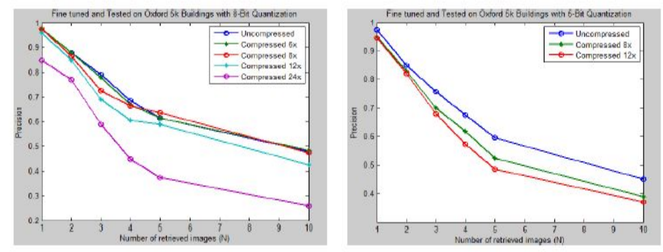
- K-means clustering on weight space.
- Cluster weights and gradients.
- Store the cluster centers in a codebook.
- Sum-up the gradients with the same cluster center.
- Update the cluster centers with the gradients.
- Repeat.



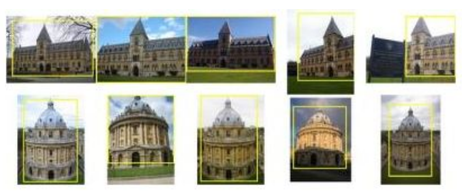
RESULTS

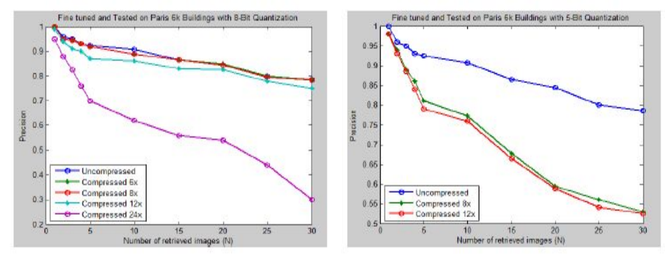
- 8x Compression(Alexnet: 32.4MB, VGG16: 65.1MB) with no drop in MAP and 12x Compression (Alexnet: 20MB, VGG16: 40.5MB) with only 2% drop in MAP

Paris Buildings



Oxford Buildings





Method	Threshold for pruning	Percentage of Parameters Pruned	Drop in MAP (Oxford Buildings)	Drop in MAP (Paris Buildings)	Memory usage (MB)
Alexnet + NetVLAD + whitening (base model)	0	0	0%	0%	248.6
8 bits quantization	0.001	25.77	0%	0%	41.4
	0.005	48.44	0%	0%	32.4
	0.01	69.92	2.1%	1.8%	20.0
	0.05	85.77	14.2%	13.3%	10.3
5 bits quantization	0.005	52.39	2.9%	3.4%	19.5
	0.01	74.95	7.3%	6.7%	10.6
VGG16 + NetVLAD + whitening (base model)	0	0	0%	0%	529.5
8 bits quantization	0.001	25.52	0%	0%	89.6
	0.005	51.77	0%	0%	65.1
	0.01	68.23	2%	2.1%	40.5
	0.05	84.68	11.8%	14.1%	21.7
5 bits quantization	0.005	55.77	2.2%	3.6%	42.1
	0.01	75.66	6.8%	5.6%	21.2

Qualitative Results

Center for Visual Information Technology, International Institute of Information Technology - Hyderabad, India

soham.saha@research.iiit.ac.in