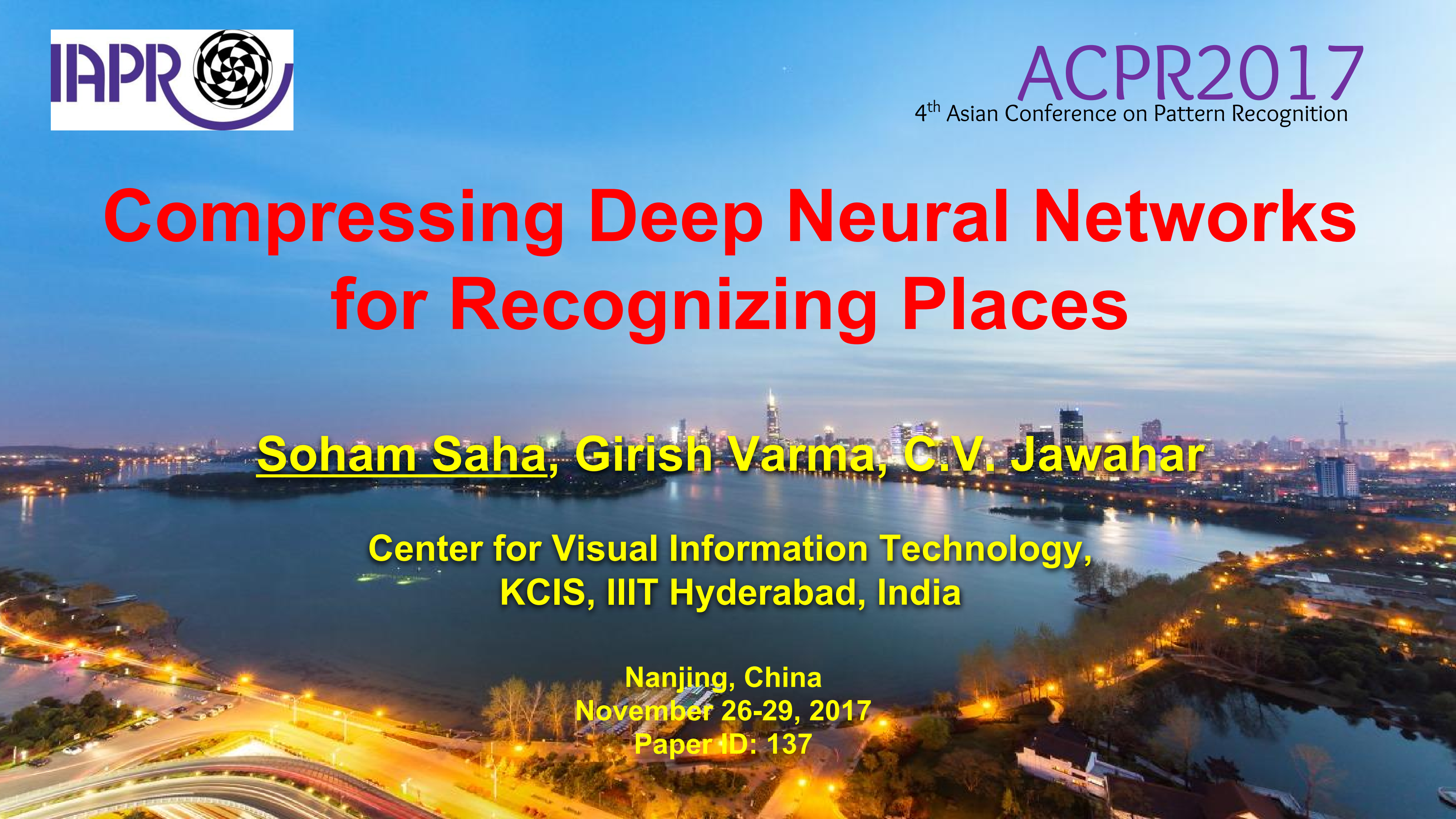


ACPR2017

4th Asian Conference on Pattern Recognition

Compressing Deep Neural Networks for Recognizing Places

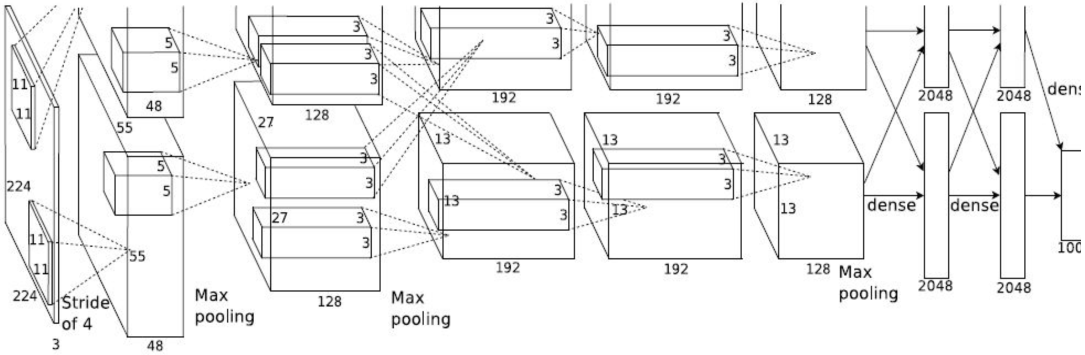
Soham Saha, Girish Varma, C.V. Jawahar

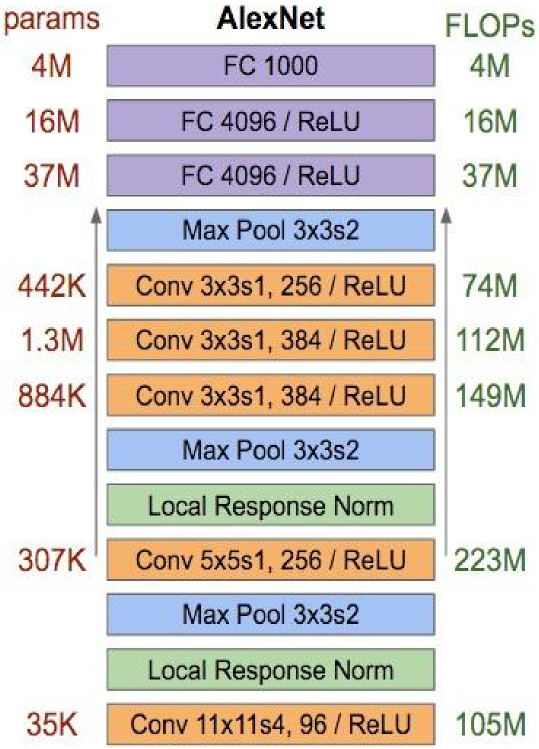
Center for Visual Information Technology,
KCIS, IIIT Hyderabad, India

Nanjing, China
November 26-29, 2017
Paper ID: 137

Part 1 Motivation

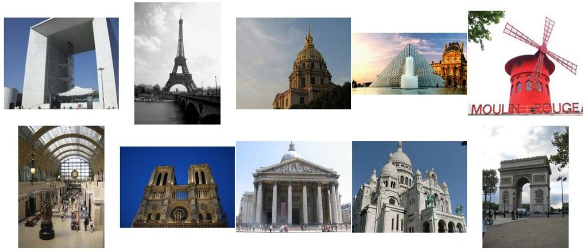
- Deep Learning models need to run on small devices
- Alexnet has 60M parameters (~240MB on disk)
- VGG-16 has ~140M parameters (~530MB on disk)

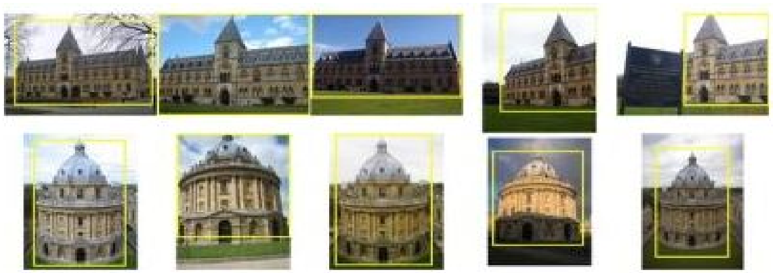


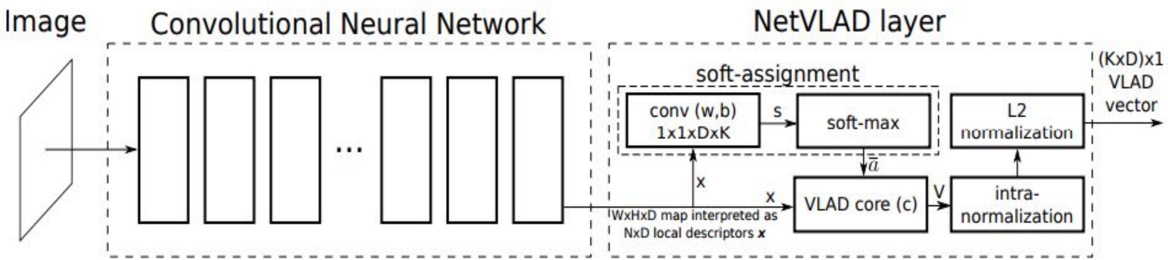


params	AlexNet	FLOPs
4M	FC 1000	4M
16M	FC 4096 / ReLU	16M
37M	FC 4096 / ReLU	37M
	Max Pool 3x3s2	
442K	Conv 3x3s1, 256 / ReLU	74M
1.3M	Conv 3x3s1, 384 / ReLU	112M
884K	Conv 3x3s1, 384 / ReLU	149M
	Max Pool 3x3s2	
	Local Response Norm	
307K	Conv 5x5s1, 256 / ReLU	223M
	Max Pool 3x3s2	
	Local Response Norm	
35K	Conv 11x11s4, 96 / ReLU	105M

Part 2 Contribution



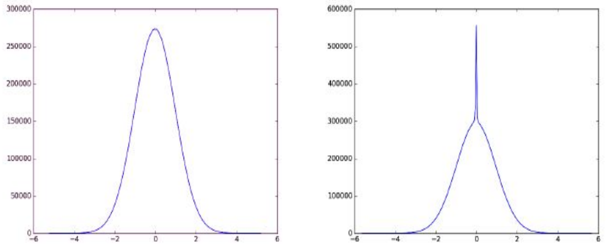


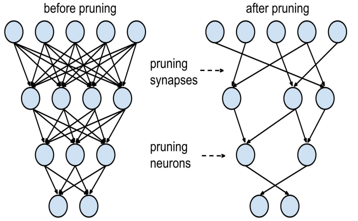


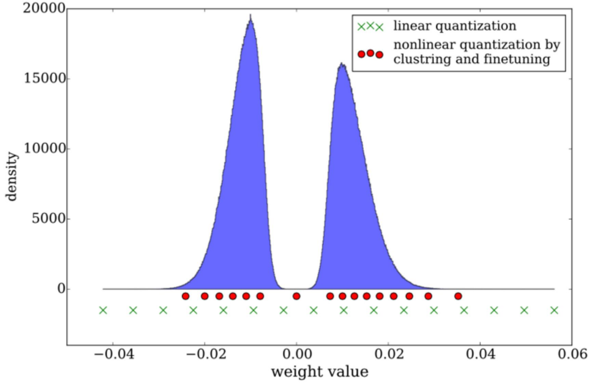
$$Loss = \frac{1}{2}h(m + ||q-p||^2 - ||q-n||^2)$$

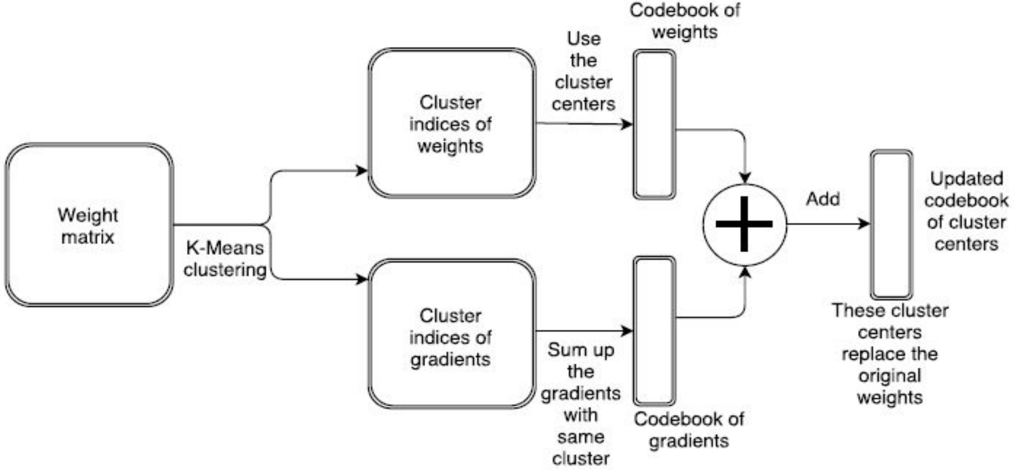
Part 2 Contribution

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









Weight matrix

K-Means clustering

Cluster indices of weights

Use the cluster centers

Codebook of weights

Cluster indices of gradients

Sum up the gradients with same cluster

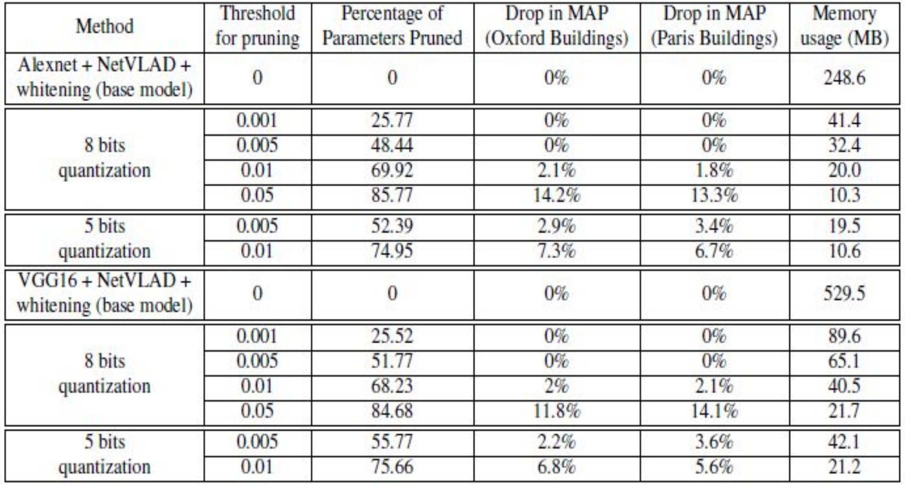
Codebook of gradients

Add

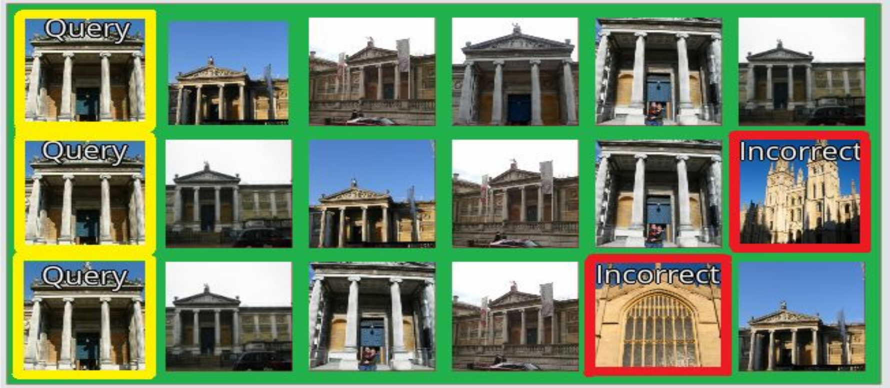
Updated codebook of cluster centers

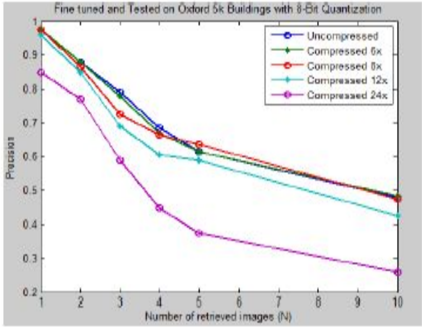
These cluster centers replace the original weights

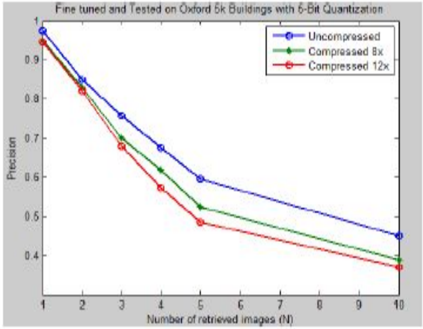
Part3 Results

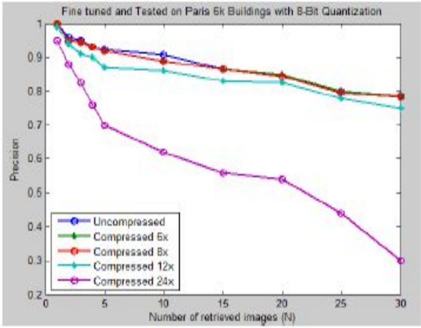


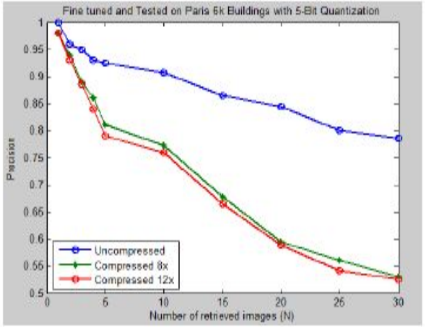
Method	Threshold for pruning	Percentage of Parameters Pruned	Drop in MAP (Oxford Buildings)	Drop in MAP (Paris Buildings)	Memory usage (MB)
Alexnet + NetVLAD + whitening (base model)	0	0	0%	0%	248.6
8 bits quantization	0.001	25.77	0%	0%	41.4
	0.005	48.44	0%	0%	32.4
	0.01	69.92	2.1%	1.8%	20.0
	0.05	85.77	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











ACPR 2017

Part4

Thank you!

Visit our poster for more details!