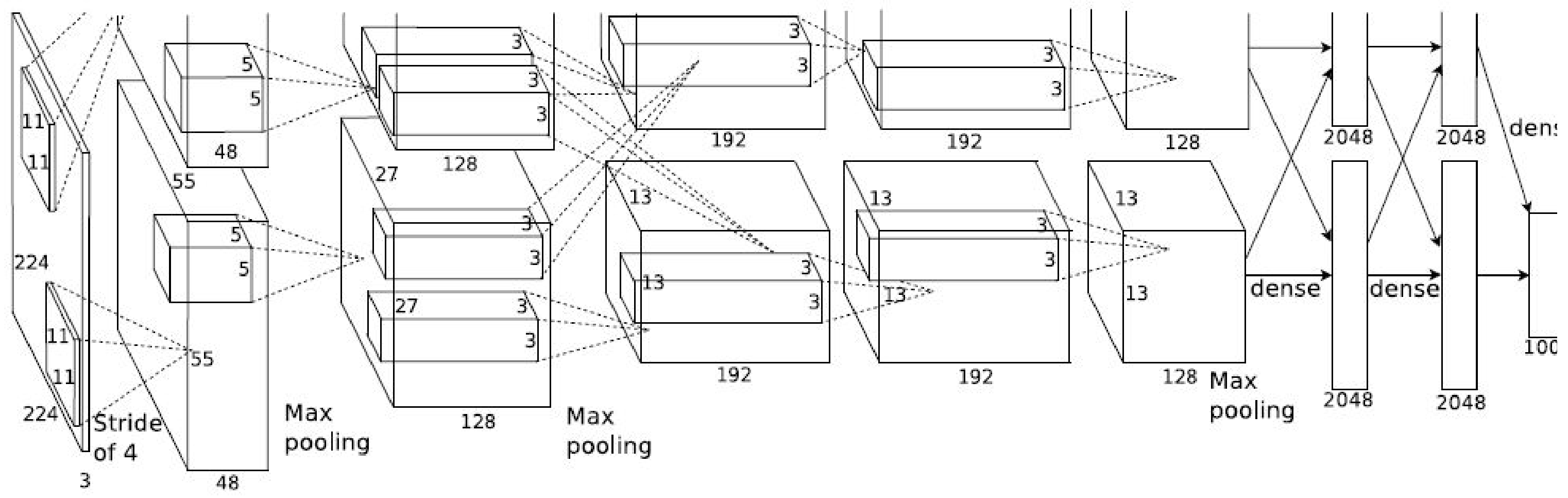
R&D SHOWCASE

Making Deep Learning Models Memory Efficient

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

MOTIVATION

- Deep Learning Models needs to run on small devices.
- Model needs to process video at 30 fps.

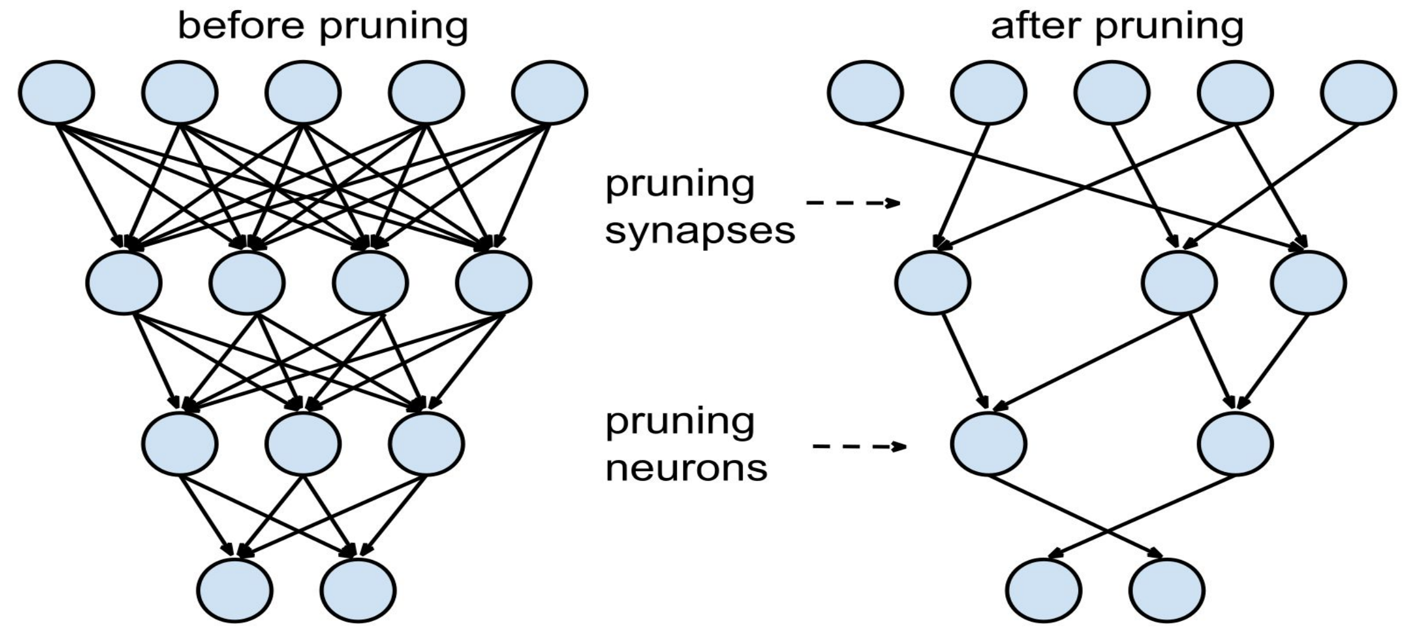


- But current models Eg.-AlexNet has 60 million parameters (~240 MB on disk) and performs 1.5 billion single precision operations to classify one image i.e in a forward pass.)

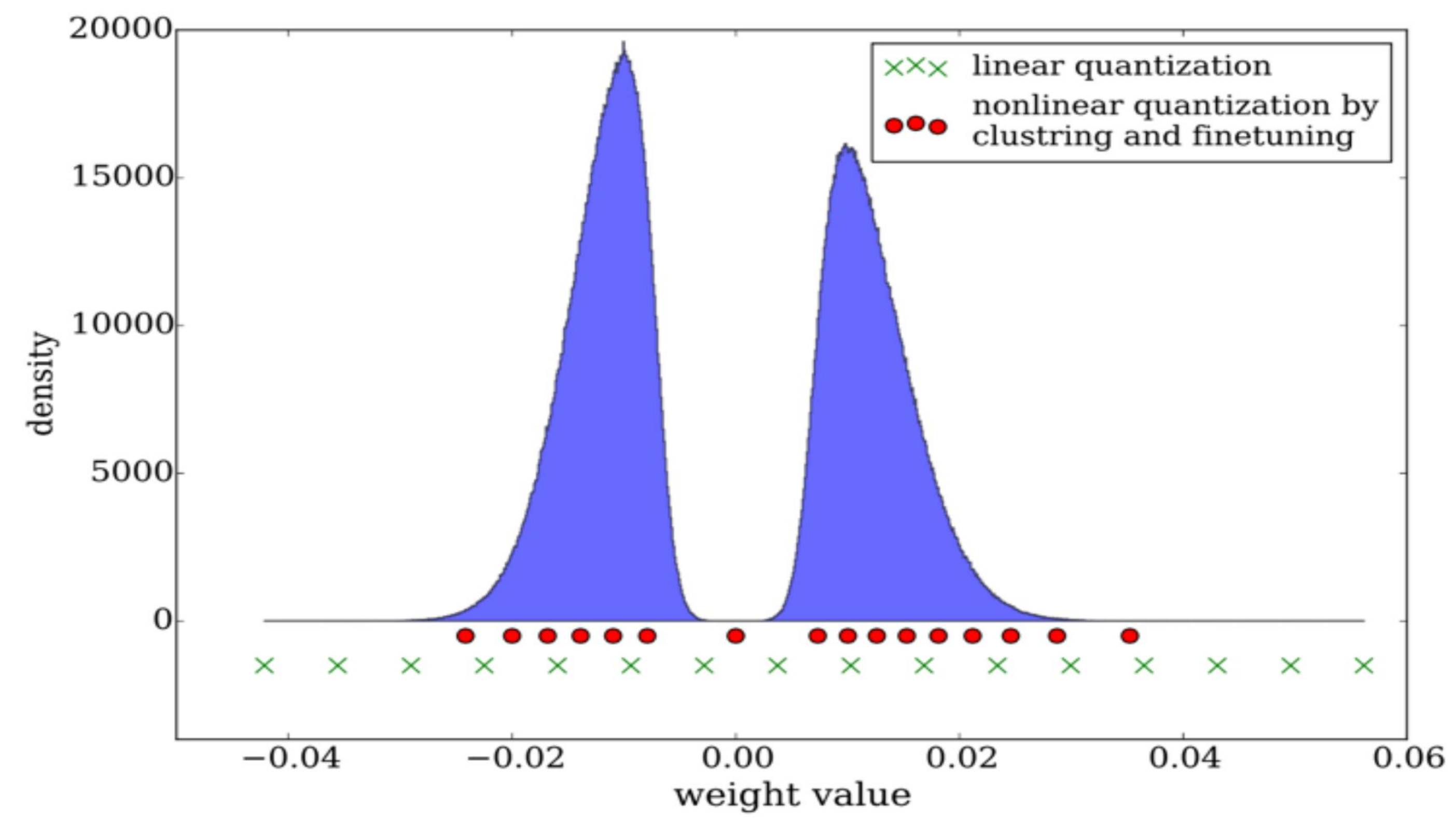
params	AlexNet	FLOPs
4M	FC 1000	4M
16M	FC 4096 / ReLU	16M
37M	FC 4096 / ReLU	37M
	Max Pool 3x3s2	
442K	Conv 3x3s1, 256 / ReLU	74M
1.3M	Conv 3x3s1, 384 / ReLU	112M
884K	Conv 3x3s1, 384 / ReLU	149M
	Max Pool 3x3s2	
	Local Response Norm	
307K	Conv 5x5s1, 256 / ReLU	223M
	Max Pool 3x3s2	
	Local Response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

METHODS

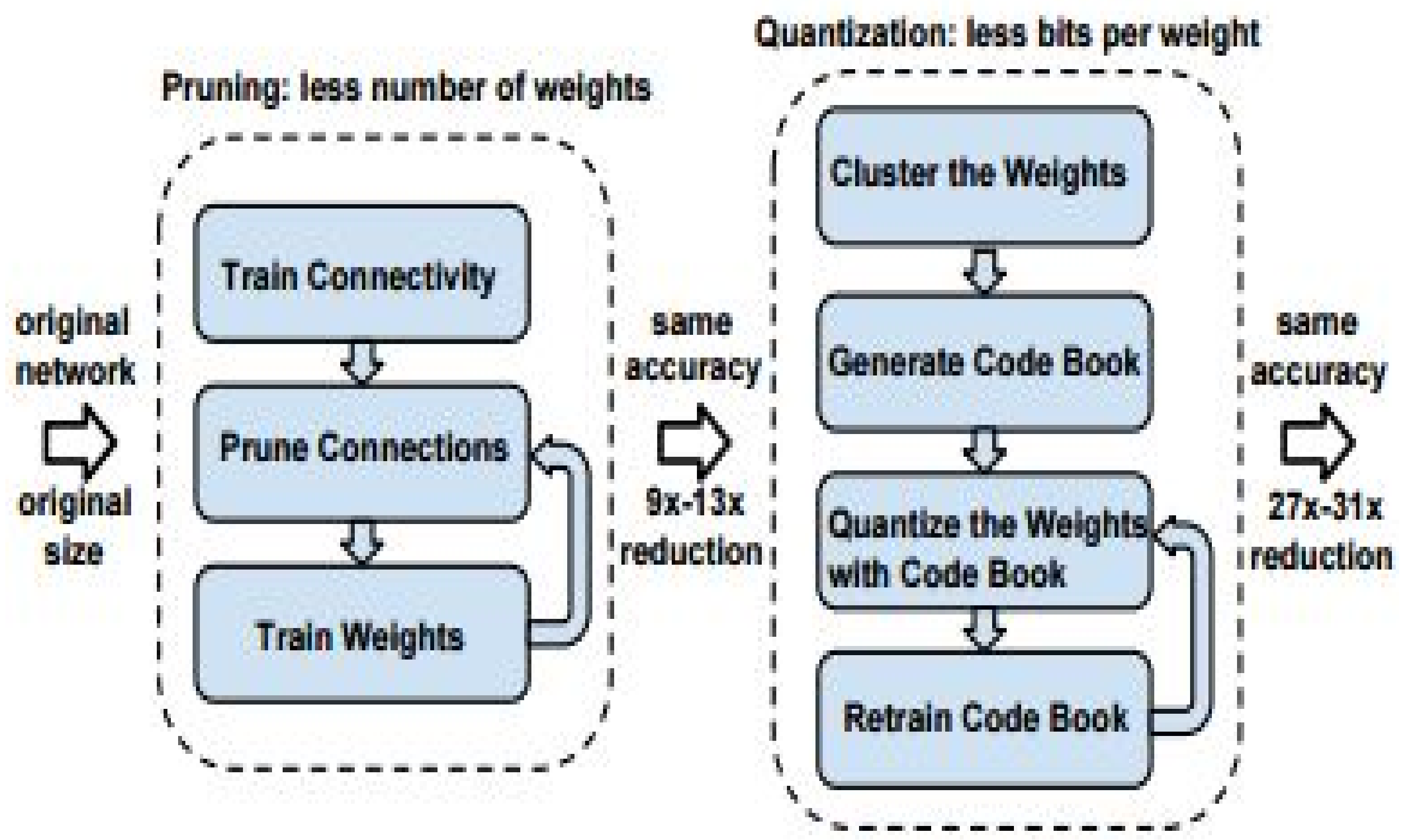
- Pruning



- Quantization

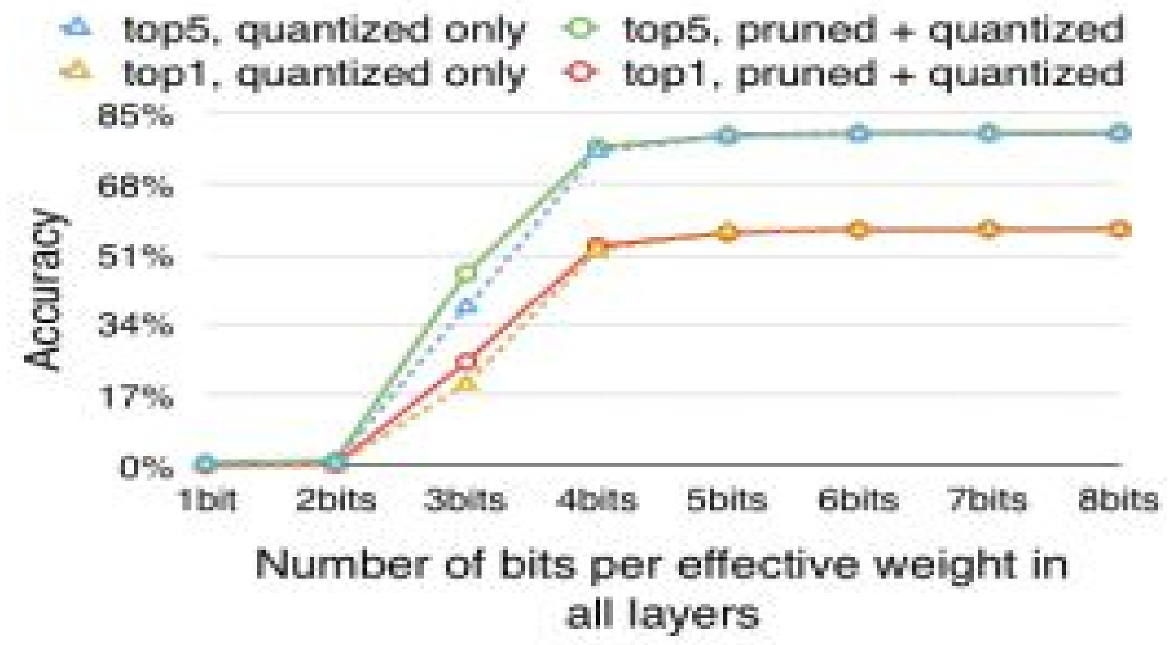


- Pruning and Quantization while Training

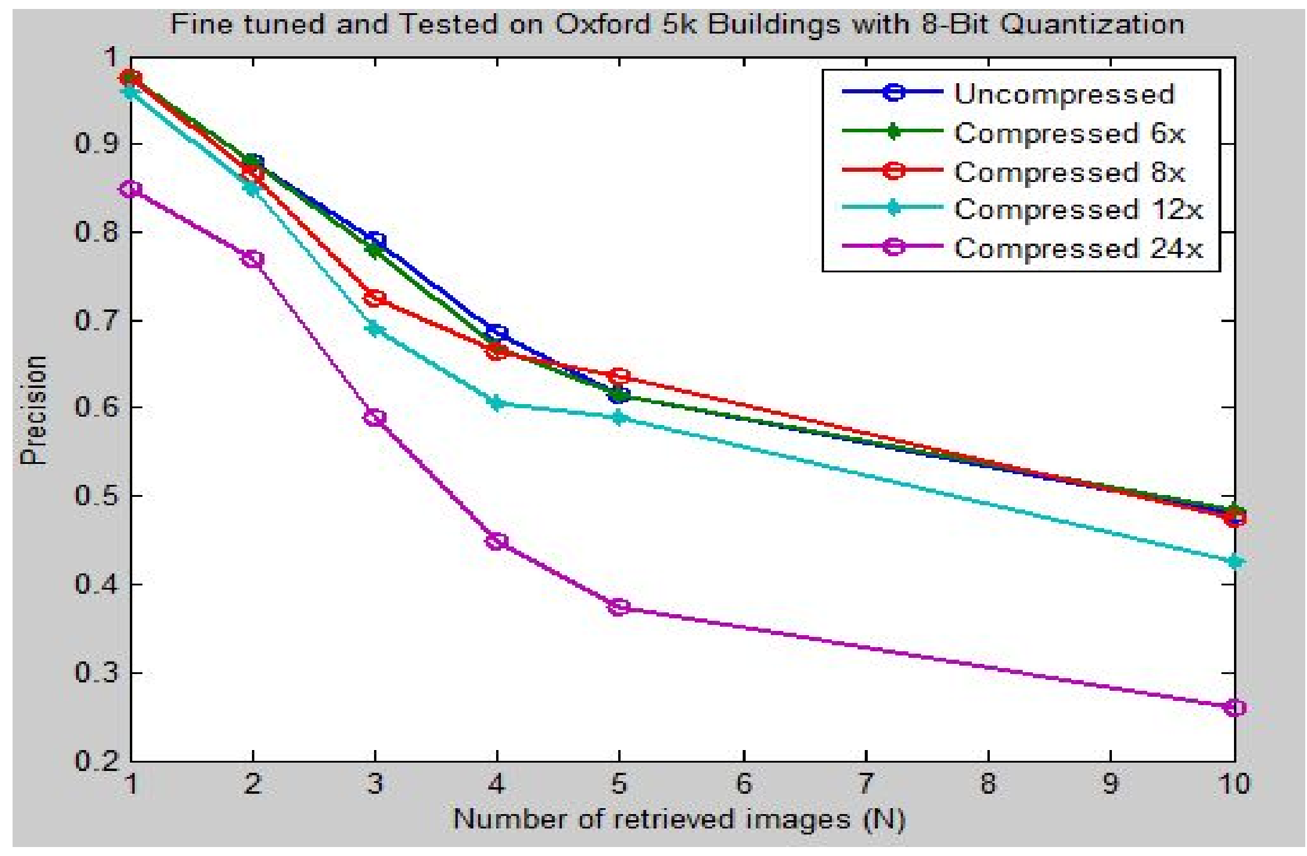


RESULTS

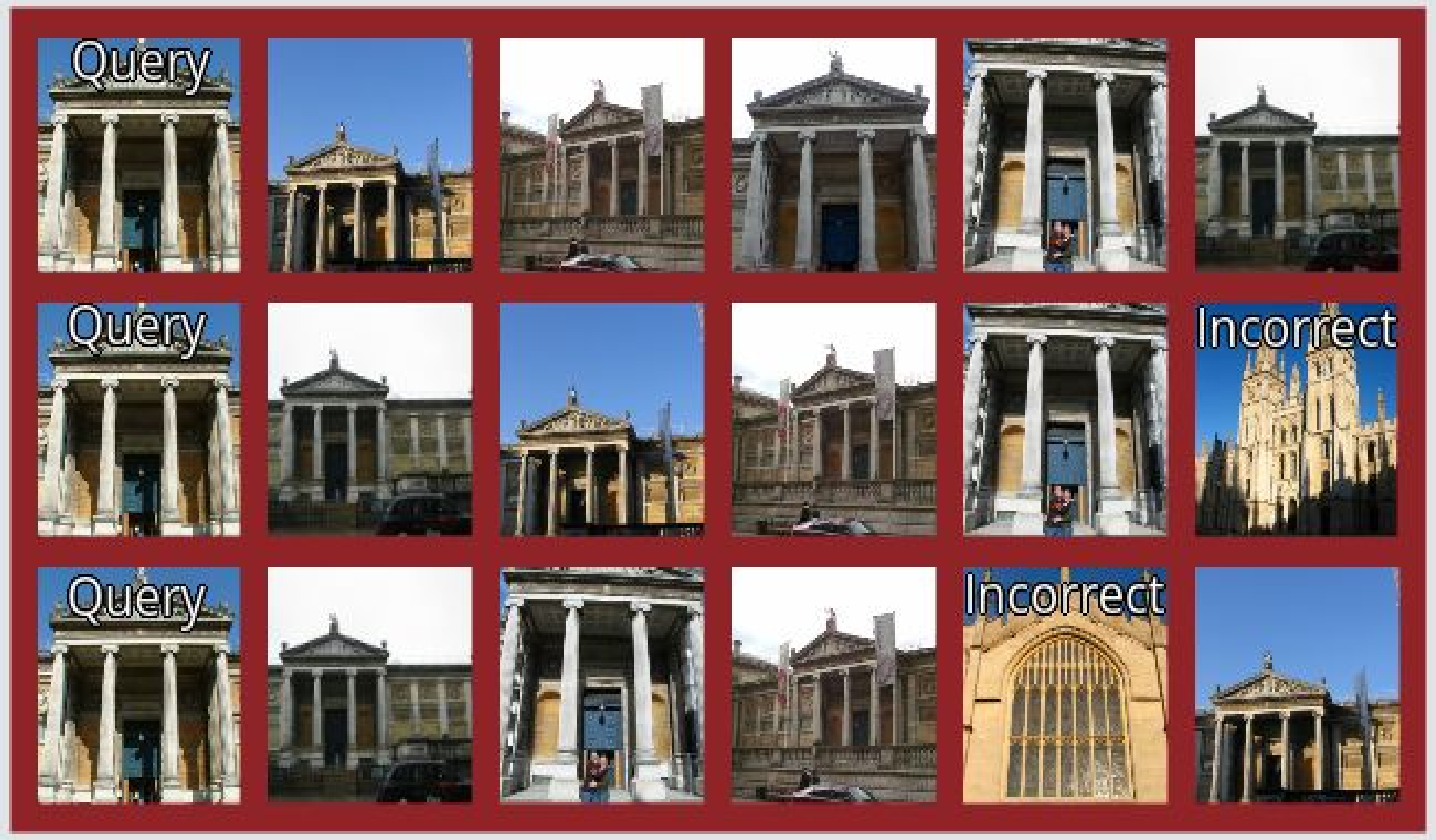
- Quantization results for Image Classification with Imagenet



- For Image Retrieval Dataset



Qualitative results for Image Search.
The 1st, 2nd and 3rd rows are retrieved images with Ground Truth,Uncompressed Network and 12X Compressed Network.



Soham Saha, Girish Varma, C.V.Jawahar, Compressing Deep Neural Networks for Recognizing Places, ACPR, 2017

Contact: soham.saha@research.iiit.ac.in, girish.varma@iiit.ac.in

Machine Learning Lab, KCIS