Quantization Mimic: Towards Very Tiny CNN for Object Detection

Pawan Prasad K (ME16B179)
Gokulan R (CS15B033)

CS7015 - Deep Learning

Tiny networks

ResNet18-1-16

Number of filters in each CONV layer of ResNet18 is reduced to 1/16th of the original value

Advantages:

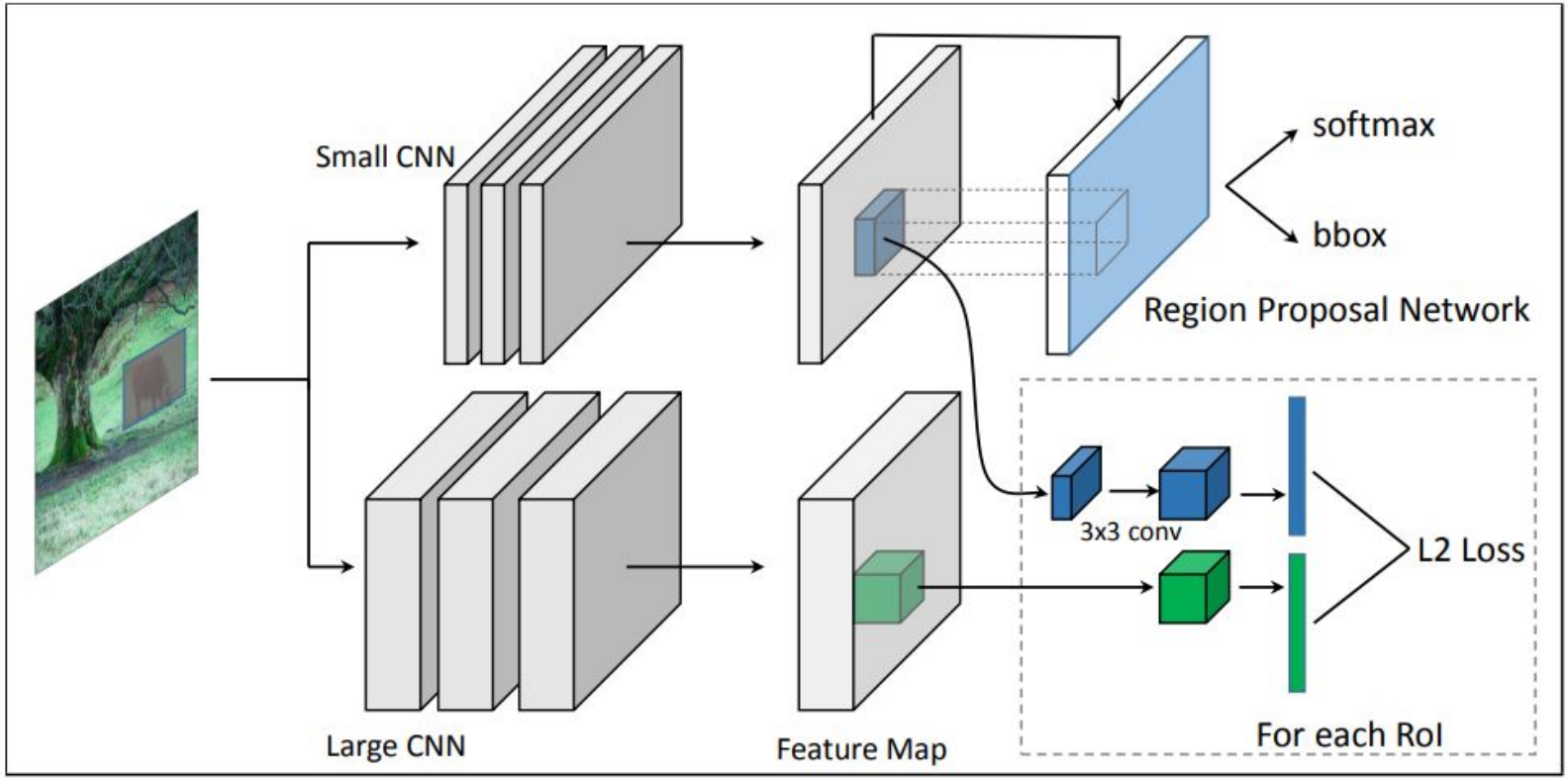
- Very small model
- Extremely fast inference

Disadvantages:

- Low representation power, less accuracy
- Difficult to train

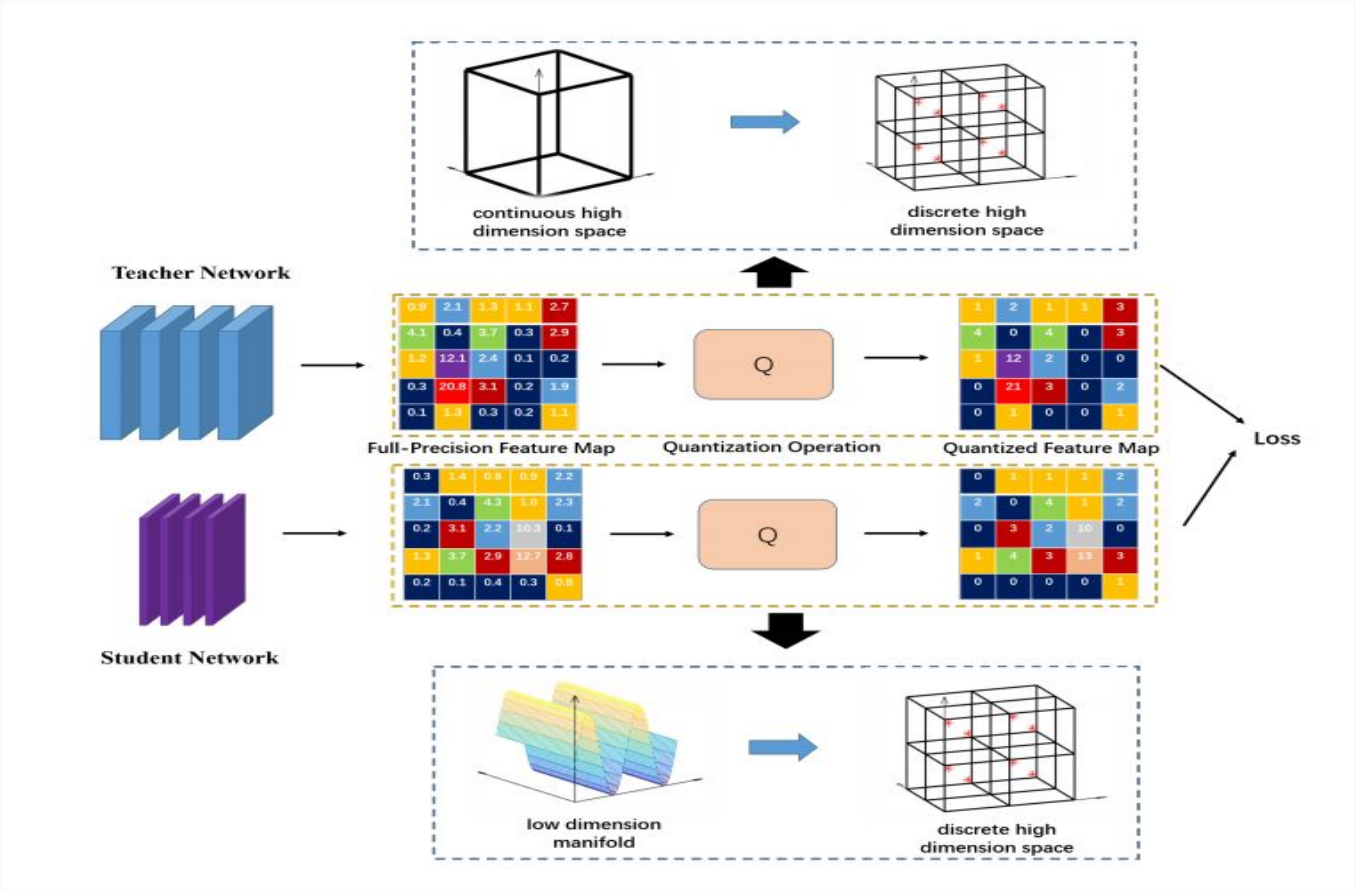
Mimic

Li et al: Mimicking very efficient network for object (CVPR). (2017)



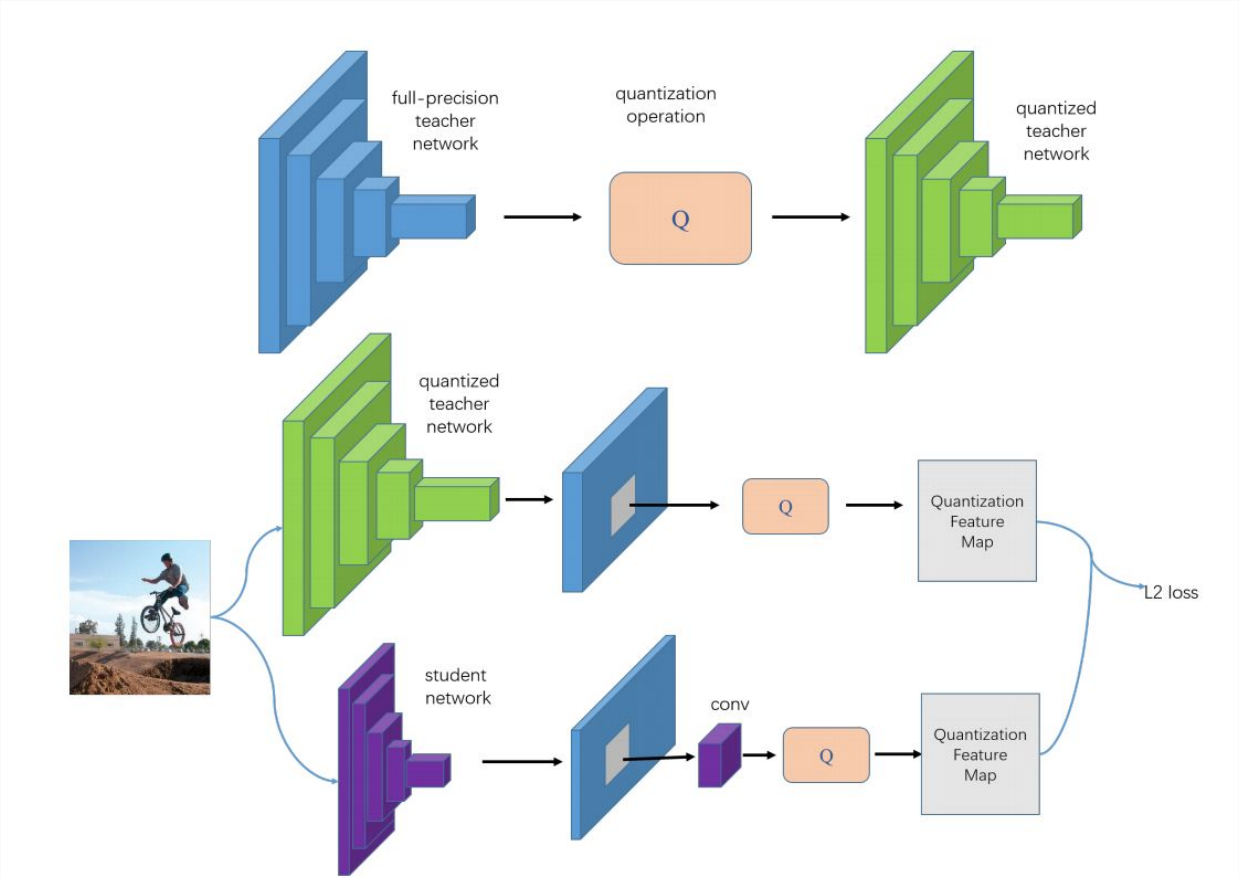
Quantization Mimic

Quantization Mimic: Towards Very Tiny CNN for Object Detection



Training Procedure

Quantization Mimic: Towards Very Tiny CNN for Object Detection



Resnet18 and ResNet18 Quantized with Faster-RCNN

mAP: 67.64, paper's mAP: 72.9										
class	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
AP	67.83	75.79	66.40	52.93	49.43	73.31	78.64	77.22	47.48	73.78
class	Dining table	dog	horse	motorbike	person	Potted plant	sheep	sofa	train	tvmonitor
AP	64.16	78.18	81.72	73.90	96.38	41.39	70.00	64.28	73.73	66.35

Quantized - mAP: 67.64, paper's mAP: 73.3										
class	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
AP	67.93	75.39	67.21	53.35	49.31	74.19	78.65	76.83	47.05	71.17
class	Dining table	dog	horse	motorbike	person	Potted plant	sheep	sofa	train	tvmonitor
AP	64.75	78.79	81.68	74.17	76.65	41.07	70.30	64.61	73.55	66.25

ResNet18-1-16 {from scratch, mimic, quantized mimic}

- Reuse *torchvision.resnet18* and change channel depth of all conv and bn layers by a factor

- **quantization:** add a function to forward() to quantize feature maps

- **mimic:** find feature loss, backpropagate to student network

$$L = L_{cls}^{r} + L_{reg}^{r} + L_{cls}^{d} + L_{reg}^{d} + \lambda L_{m}$$

$$L_{m} = \frac{1}{2N} \sum_{i} \left\| f_{t}^{i} - r\left(f_{s}^{i}\right) \right\|_{2}^{2}$$