Distilling Object Detectors with Fine-grained Feature Imitation — Supplementary Material

Tao Wang[1] Li Yuan[1] Xiaopeng Zhang[1,2] Jiashi Feng[1]
[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Huawei Noah's Ark Lab, Shanghai, China
twangnh@gmail.com ylustcnus@gmail.com zhangxiaopeng12@huawei.com elefjia@nus.edu.sg

Lightweight detector architecture

Table S1 shows the base $1\times$ model of our developed lightweight detector applied on the KITTI dataset. Each stage and the additional block is composed of repeated ShuffleNet units [3], which is modified residual unit and consists of group convolution, channel shuffle operation and depth-wise convolution, aiming for superior performance with drastically reduced computation cost. The detection layer has 72 output channels, corresponding to 8 prediction variables (4 location coordinates and 3 object class predictions as well as 1 confidence prediction) for each of 9 specified anchor size. The 9 anchors are calculated at 3 scales (*i.e.*, 64, 128, 256) and 3 aspect ratios (*i.e.*, 0.5, 1.0, 2.0).

We decouple anchor classification into confidence prediction and object class prediction as [2], and train the detector by minimizing the following loss:

$$L_{det} = \sum_{i,j,k}^{W,H,K} V(L_{\text{reg}})_{ijk} + VM(L_{\text{cls}})_{ijk} + VM(L_{\text{conf}})_{ijk} + Z\widehat{M}(L_{\text{conf}})_{ijk}.$$

Here $V = 1/N_{\text{obj}}$, and $Z = 1/(WHK - N_{\text{obj}})$, N_{obj} is the number of objects in the image, and W, H, K depict the feature map dimensions. L_{reg} is loss for bounding box regression, L_{cls} indicates softmax loss for the classifier, L_{conf} denotes confidence loss. M is the mask indicates positive anchors which have the largest overlap with ground truth boxes, and $\widehat{M}$ is reverse of the mask.

More analysis results

Fig S1 and Fig S2 presents visualization of performance and error composition for non-imitated and imitated student model of VGG11 based Faster R-CNN with the tool of [1]. The teacher model is VGG16 based Faster R-CNN. The graphs from left to right on x-axis reflect detection outputs of high to low confidence threshold. From Fig S1, it is clearly observable that for all the three object categories, imitated student gets improved performance with all the confidence ranges. Correct detection percentage is increased. From Fig S2, the false positive distribution over different types is significantly changed. Dominant error due to inaccurate localization is clearly reduced, especially for high confidence detections.

Layer	Output size	Stride	# Repeat	# Output channels
Image	1248×384	–	–	3
Conv1	1248×384	1	1	16
MaxPool	624×192	1	–	16
Stage2	312×96	2	1	240
	312×96	1	3	240
Stage3	156×48	2	1	480
	156×48	1	5	480
Stage4	78×24	2	1	960
	78×24	1	3	960
Add	78×24	1	1	960
	78×24	1	1	768
Det	78×24	1	–	72

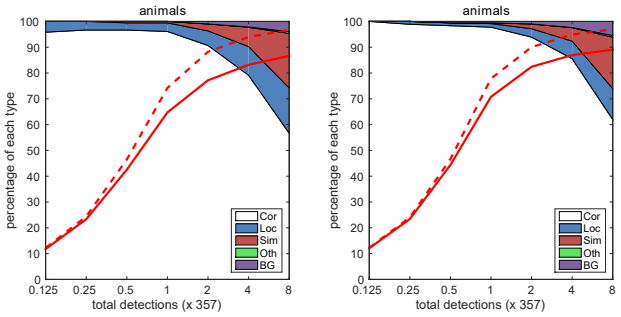
Table S1. Architecture of the our devised lightweight toy detector(the base $1\times$ model applied on the KITTI dataset).

References

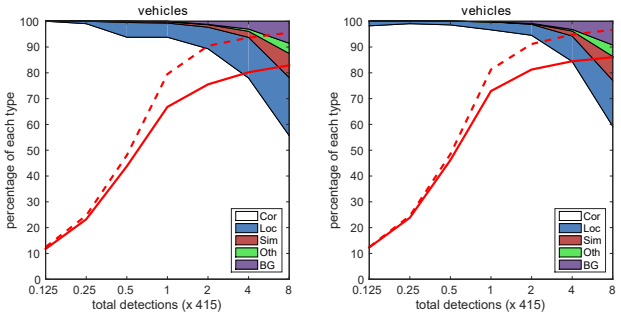
[1] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012. 1

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

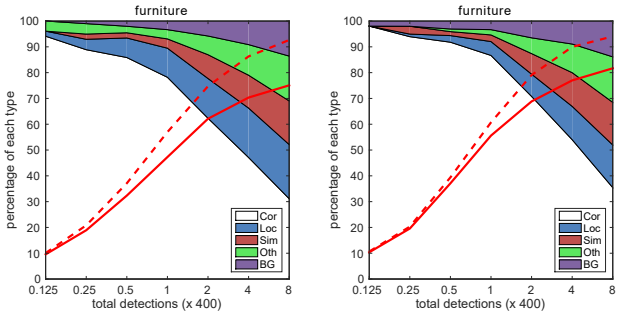
[3] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *CoRR*, abs/1707.01083, 2017. 1



(a)

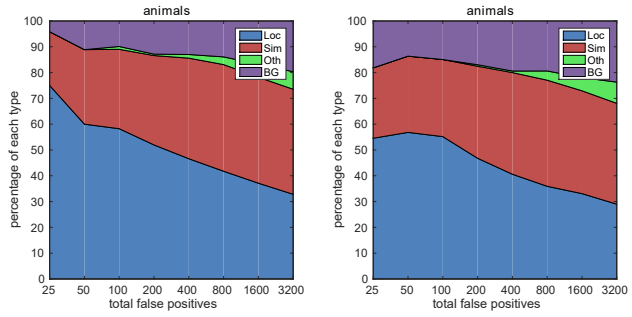


(b)

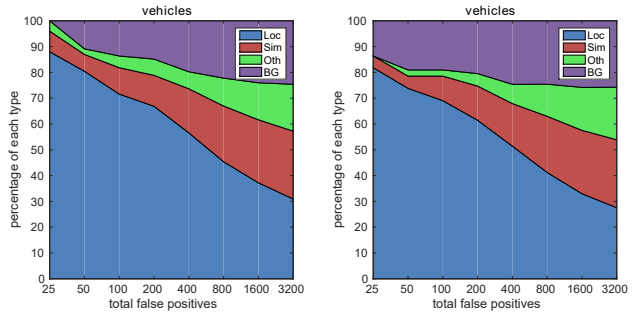


(c)

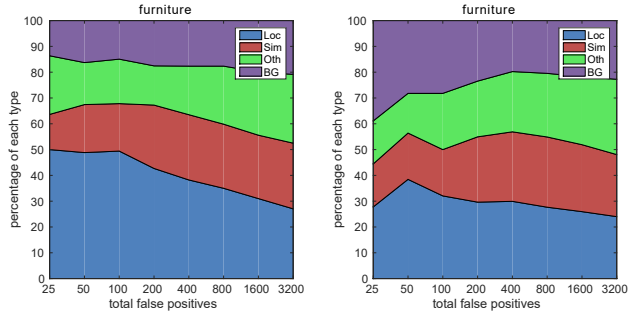
Figure S1. Visualization of performance for imitated and non-imitated detector on vehicles, animals, and furniture for VGG11 based student on VOC2007 test set. For each pair of graph, the left corresponds to non-imitated model and the right is for imitated model. The dashed and solid red line reflect recall variation with weak criteria (0.5 jaccard overlap) and strong criteria (0.1 jaccard overlap) as the number of detections increases, respectively.



(a)



(b)



(c)

Figure S2. Visualization of error composition for imitated and non-imitated detector on vehicles, animals, and furniture for VGG11 based student on VOC2007 test set. For each pair of graph, the left corresponds to non-imitated model and the right is for imitated model.