Cascaded Networks for Object Detection with Multi-Context Modeling and Hierarchical Fine-Tuning

Wanli Ouyang, Junjie Yan, Xingyu Zeng, Hongyang Li, Kai Kang, Cong Zhang, Tong Xiao, Ruohui Wang, Zhe Wang, Yubin Deng, Hongsheng Li, Xiaogang Wang, Xiaoou Tang

The Chinese University of Hong Kong

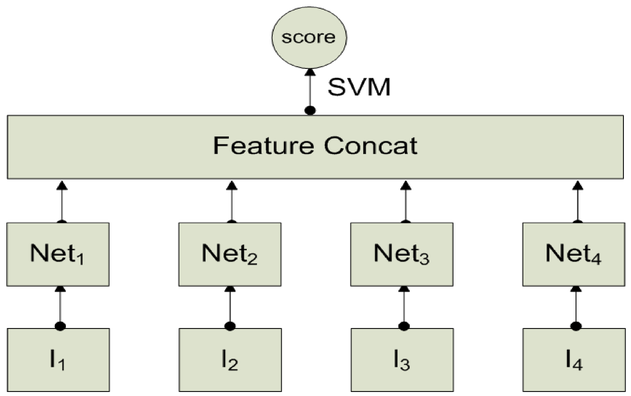
香港中文大學 The Chinese University of Hong Kong

◆Multi-context [1]

➢Motivation: A scale mismatch between the training samples for the image classification pre-training and object detection fine-tuning.

➢Solution: Multi-scale and multi-context image patch groups as input for our proposed CNN model.

Examples of patch groups(p = -28, 16, 50, 70):



➢Advantages:

- By contextually padding each box with different p values, the scales of the object in the image patches has higher chances to match those in the image classification pre-training samples, which helps better utilizing the CNN pre-training on the image classification data.
- Since a small p value leads to image patches with clear object details and a larger p value includes more context around the object, fine-tuning the CNN with such multi-scale and multi-context inputs make the CNN automatically focus on different parts of objects across different scales and different context, and learn more discriminative features representations.

➢Experimental results

Model	BN	BN	VGG	VGG
Context sizes (p)	16	16+50+70+(-28)	16	16+50+70+(-28)
mAP (3k)	49.6	53.5	48.6	52.4
mAP (1k official)	47.6	51.4	46.6	51.0

- All networks are fine-tuned on selective search & edge boxes

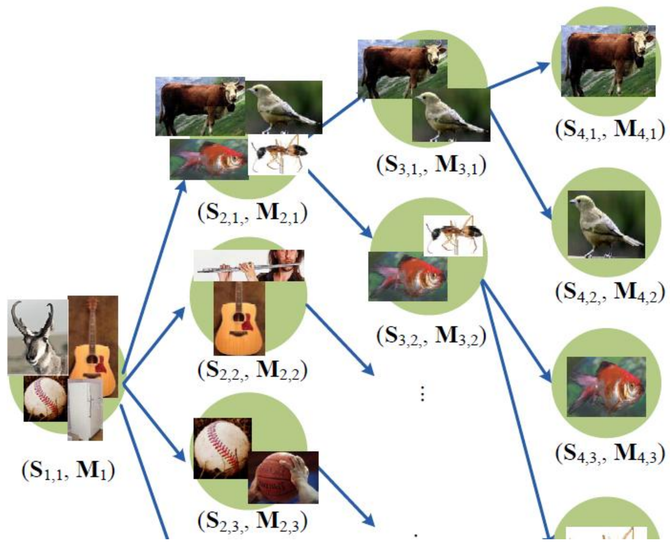
[1] X. Zeng, *et al.* Window-Object Relationship Guided Representation Learning for Generic Object Detections , axiv preprint.
[2] W. Ouyang, *et al.* Factors in Finetuning Deep Model for object detection, axiv preprint.
[3] J. Yan, et al. CRAFT Objects from Images, axiv preprint.
[4] W. Ouyang, *et al.* Deepid-net: Deformable deep convolutional neural networks for object detection. *CVPR, 2015.*
[5] J. Yan, *et al.* Object detection by labeling superpixels. *CVPR*, 2015.

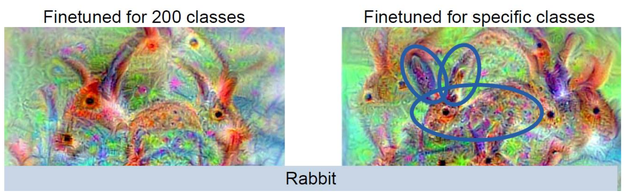
◆Cascaded hierarchical feature learning [2]

➢Motivation:

Objects of different classes have their own discriminative visual appearance. If the learned representation can focus on specific classes, e.g. mammals, the learned representation is better in describing these specific classes.

➢Solution: Cascaded hierarchical feature learning.





- Hierarchical clustering is used for grouping object classes into hierarchical clusters. Specifically, we group object classes into hierarchical clusters $\mathbf{S}_{i,j}$, and finetune them to obtain multiple models $\mathbf{M}_{i,j}$.
- The visual similarity between classes a and b is as follows

$$Sim(a,b) = \sum_{i=1}^{N_i}\sum_{j=1}^{N_j} < \mathbf{h}_{a,i}, \mathbf{h}_{b,j} > /N_i N_j,$$

➢Advantages:

- Classes in different groups have their own visual representation.
- The knowledge from the group with large number of classes is transferred for to features in its sub-groups.
- Through cascade, each model only focuses on around 6 candidate regions per image.
- Experimental results on ILSVRC14 val2

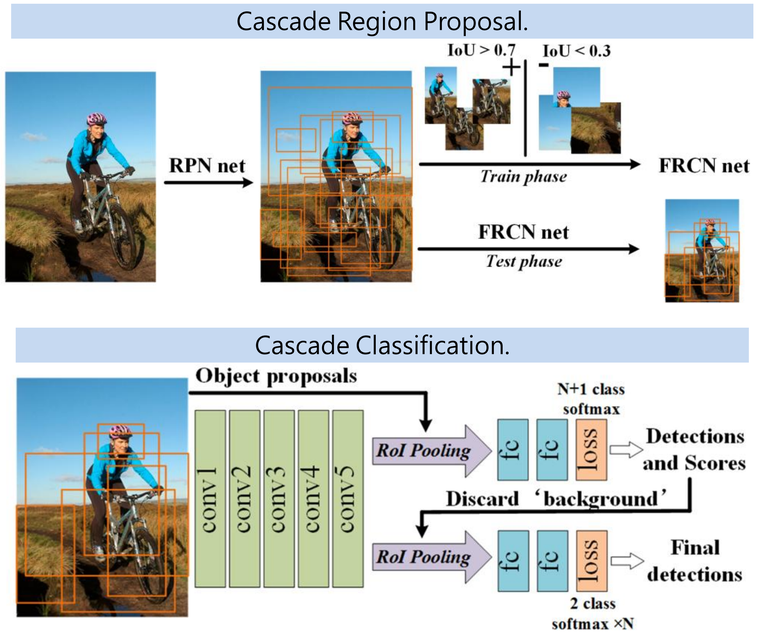
Hierarchy level L	1	2	3	4	New result
Num. groups	1	4	7	18	7
Avg. Num classes per group	200	50	29	11	29
mAP	40.3	41.3	42.5	45	56

◆Cascade Region-proposal-network And FasT-rcnn (CRAFT)[3]

➢Motivation:

Object Detection is always decomposed into object proposal generation and proposal classification. We push the solution even further through the cascade region proposal generation and cascade proposal classification.

➢Solution:



➢Advantages:

- One can always get better results when initialized by a better pre-trained CNN classification model on ImageNet. Our cascade structure provides a way of boosting smaller models to get comparable or even superior performance than better but more complex models.

➢Experimental results

- Region proposal

Setting	Number of proposals	Recall (%)
Selective Search	2000	92.09
RPN-1	300	89.94
RPN-2	300	91.83
RPN+FRCN	300	92.38
SS+RPN+FRCN	300	94.13

- Detection

Results on VOC07

Setting	mAP(%)
No cascade	65.0
Single-class re-score	63.5
Multi-class re-score	**68.0**

Results on ILSVRC14 val2

Setting	mAP(%)
GoogLeNet _BN	47.0
Cascade GoogLeNet BN	48.5
Improvement	+1.5

Average model results: mAP 59.5 on Val2