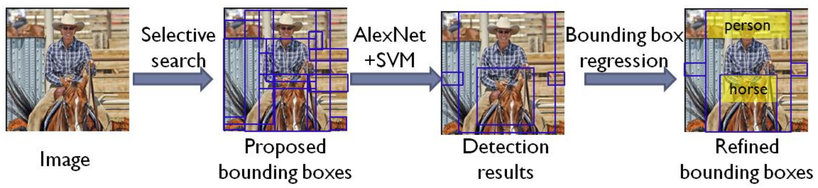
Multi-stage Deep Convolutional Neural Network For Generic Object Detection

Wanli Ouyang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Yuanjun Xiong, Chen Qian, Zhenyao Zhu, Ruohui Wang, Chen-Change Loy, Xiaogang Wang, Xiaoou Tang

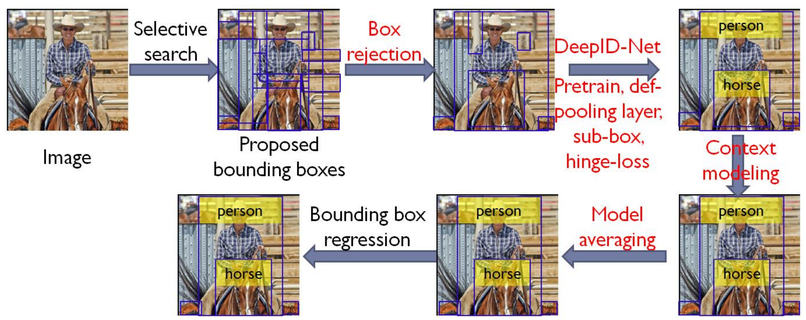
The Chinese University of Hong Kong

香港中文大學 The Chinese University of Hong Kong

RCNN [1]:



Our approach[3]:



Contribution:

1. Diagram for generic object recognition with deep learning
2. Bounding box rejection (1%)
3. DeepID-Net
   - ➢ Strategies of pre-training features with classification data (4%)
   - ➢ Def-pooling Layer (1.5%)
   - ➢ Sub-box detector (0.5%)
   - ➢ Hinge-loss
4. Context Modeling (0.8%~1%)
5. Model Averaging (5%)

◆ Bounding Box Rejection:

Motivation:

1. Speed up feature extraction by ~10 times
2. Improve mean AP by 1%

RCNN:

- ➢ Selective search: 2400 bounding boxes per image
- ➢ 38 hours for extracting features from 10,000 images
- ➢ ILSVRC13 val: ~20,000 images, ~ 2.4 days
- ➢ ILSVRC 13 test: ~ 40,000 images, ~4.7 days

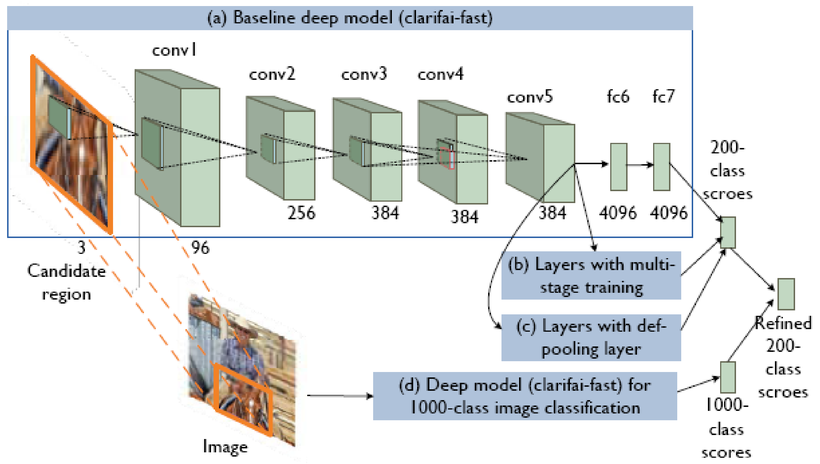
Reject bounding boxes from RCNN:

- ➢ For each bounding box, RCNN provides 200 scores for 200 classes
- ➢ If $\max(S_{1,\ldots,200}) < -1.1$, 6% bounding boxes remains.

Remaining Window	100%	20%	6%
Recall (val)	92.2%	89.0%	84.4%
Feature extraction time (seconds per image)	10.24	2.88	1.18
Mean AP on val2	0.299	-	0.309

1, Girshick, Ross. et al, Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR 2014
2, Xingyu ZENG et al, Multi-stage contextual deep learning for pedestrian detection, ICCV 2013
3. Wanli ouyang, etal, DeepID-Net: multi-stage and deformable deep convolutional neural network for generic object detection, arXiv

◆DeepID-Net2:



➢Strategies for Pre-training features with classification data:

- •Classification vs detection (image vs tight bounding box)
- •AlexNet, Clarifai or other choices
- •Complementarity
- •1000 classes vs 200 classes

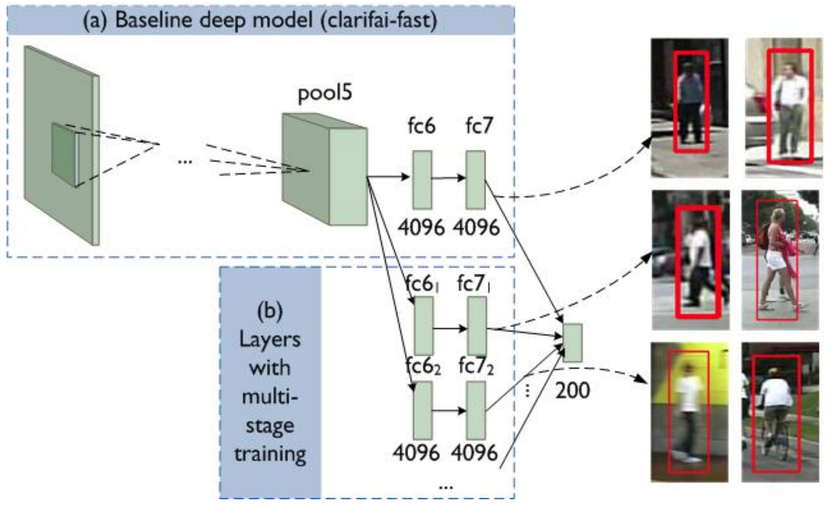
Training Scheme	Cls+Det	Cls+Det	Cls+Loc+Det	Loc+Det
Net Structure	AlexNet	Clarifai	Clarifai	Clarifai
Mean AP(val2)	0.309	0.318	0.334	0.360

Cls (Loc) means pre-training with images (localized objects) from classification dataset

Det means fine tuning with the detection dataset

➢Multi-Stage fully connected layers training [2]

- •Multi-classifiers trained jointly and sequentially
- •Each fully connected layers can be regarded as one classifier
- •Each fully connected layers branch handles different samples in different difficulty levels.



➢Multi-Stage fully connected layers training (continue …)

Algorithm 1: Stage-by-Stage Training

Input: Training set: $\Psi = \{\mathbf{s}_0, \mathbf{f}\}$
Parameters Θ for the baseline deep model obtained by pretraining.

Output: Parameters Θ for the baseline deep model, Parameters $\mathbf{W}_{l,t}, l = 6, 7, 8t = 1, \cdots, T$ for the extra layers.

1 Set elements in $\mathbf{W}_{l,t}$ to be 0;
2 BP to fine-tune Θ, while keeping $\mathbf{W}_{l,t}$ as 0;
3 **for** *t=1 to T* **do**
4 Use BP to update parameters $\mathbf{W}_{l,t}, l = 6, 7, 8$ while fixing Θ and $\mathbf{W}_{l,1}, \cdots, \mathbf{W}_{l,t-1}$;
5 Use BP to update parameters Θ and $\mathbf{W}_{l,1}, \cdots, \mathbf{W}_{l,t}, l = 6, 7, 8$;
6 **end**
7 Output Θ and $\mathbf{W}_{l,t}, l = 6, 7, 8, t = 1, \cdots, T$.

Training scheme	Loc+Det	Loc+Det	Loc+Det
Net structure	Clarifai	Clarifai+1 multi-layers	Clarifai+2 multi-layers
Mean AP on val2	0.360	0.370	0.375

Each round of training will focus on misclassified training samples. Zero initialization makes correctly-classified samples little influence.

◆Model Averaging

➢Net Structure: Alex (A), Clarifai (C), Deep-ID Net (D), Deep-ID Net2(D2)

➢Pretrain Scheme: Classification (C), Localization (L)

➢Bounding box rejection or not

➢Loss of net, softmax-loss (S), Hinge-loss (H)

Model	1	2	3	4	5	6	7	8	9	10
Net structure	A	A	C	C	D	D	D2	D	D	D
Pretrain	C	C+L	C	C+L	C+L	C+L	L	L	L	L
Bbox rejection	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
Loss of net	S	S	S	H	H	H	H	H	H	H
Mean AP	0.31	0.312	0.321	0.336	0.353	0.36	0.37	0.37	0.371	0.374

Pipeline	RCNN	Bbox rejection	Clarifai	Loc+Det	Multi-stage layer	context	regression	Model avg.
Mean AP (val2)	0.299	0.309	0.318	0.36	0.375	0.383	0.393	0.45