DO WE REALLY NEED MORE TRAINING DATA FOR OBJECT LOCALIZATION

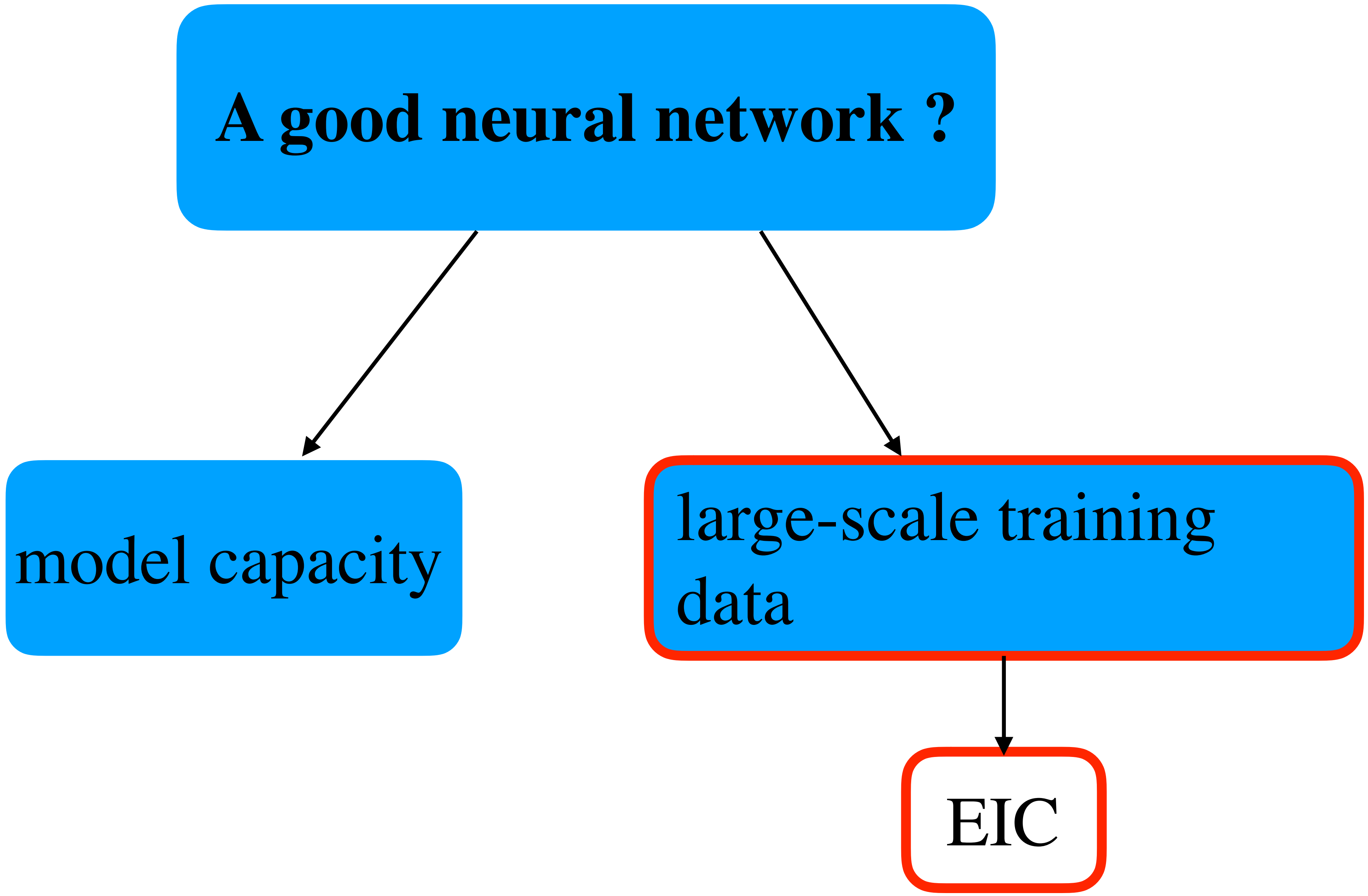
Hongyang Li[1], *Yu Liu*[1], *Xin Zhang*[2*], *Zhecheng An*[2], *Jingjing Wang*[2], *Yibo Chen*[1] *and Jihong Tong*[3]

[1]The Chinese University of Hong Kong, [2] Tsinghua University, [3] Eastern Liaoning University

http://www.ee.cuhk.edu.hk/~yangli/project/eic.html

Presenter : Xin Zhang, Tsinghua University

Preparation



Q1: Is more training data beneficial to obtain better results evaluated on the original smaller scale dataset?

Preparation

Fast R-CNN

Faster R-CNN

…

High-level && Summarized semantics

Efficient but **lose important low-level details**

Multiple layers[1]

Q2: How to utilize utilizing the feature maps in the network to obtain better representation of data?

[1].Bharath Hariharan, Pablo Arbelez, Ross Girshick, and Jitendra Malik, “Hypercolumns for object segmentation and fine-grained localization,” in *CVPR*, 2014.

Our work

i. THE EXTENDED IMAGENET DATASET.[2]

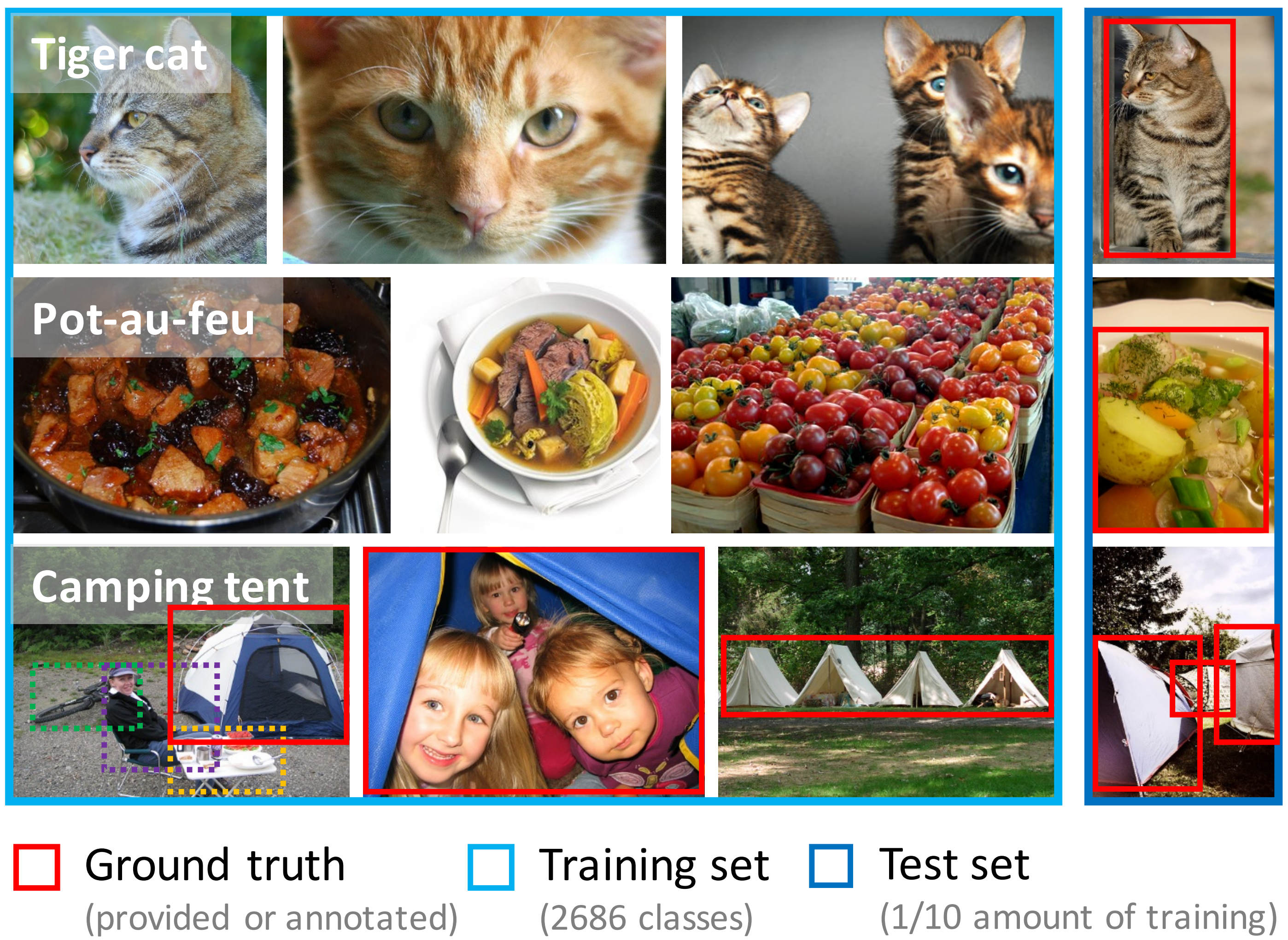
ii. Whether a larger dataset is necessary to train a deep learning model for robust and representative features?

iii. Embed the region proposal network framework in a multi-depth, hourglass style to fully leverage the information of feature maps on different resolutions.

[2]http://www.ee.cuhk.edu.hk/~yangli/project/eic.html

Extended ImageNet Classification (EIC) set

- **2686** classes
- more '**difficult**' images
- The Training Set （2456727 images ）
- The Validation Set （273140 images ）



Extended ImageNet Classification (EIC) set

- Smaller Objects（Smaller than **32 * 32**）
- Twisted Objects (Width/Height > **4** or < **0.25**)
- The feature distance : $D(x_1,x_2)=1-\cos(x_1,x_2)$
- The feature representation : layer fc6 in the VGG-16 model
- The extended categories are chosen by the WordNet

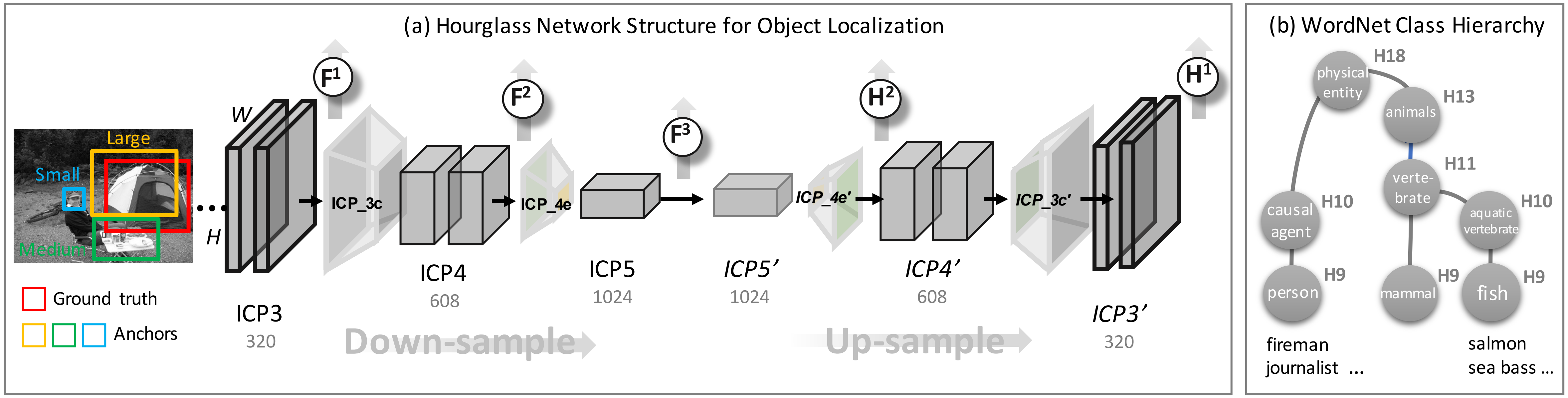
Dataset split	Extended ImageNet		ILSVRC CLS 2012	
	Train	Val.	Train	Val.
# of images	2,456,727	273,140	1,281,167	50,000
Avg im # per cls	251-1300	34-50	732-1300	50
Avg anno per im	1.53	1.17	1.41	1.02
Avg obj scale	25.37 %	25.50%	25.39%	25.61 %
Small obj %	4.81	4.27	2.35	2.47
Twisted obj %	42.77	44.55	40.72	42.94
Inner-cls distance	0.434	0.396	0.462	0.411
Inter-cls distance	1.12	1.46	1.52	1.55

im=image, avg=average, cls=class, anno=annotation, obj=object.

Algorithm

Network architecture

- Different-sized anchors are placed at different resolutions of the network, fully leveraging the information of feature maps especially for small objects.



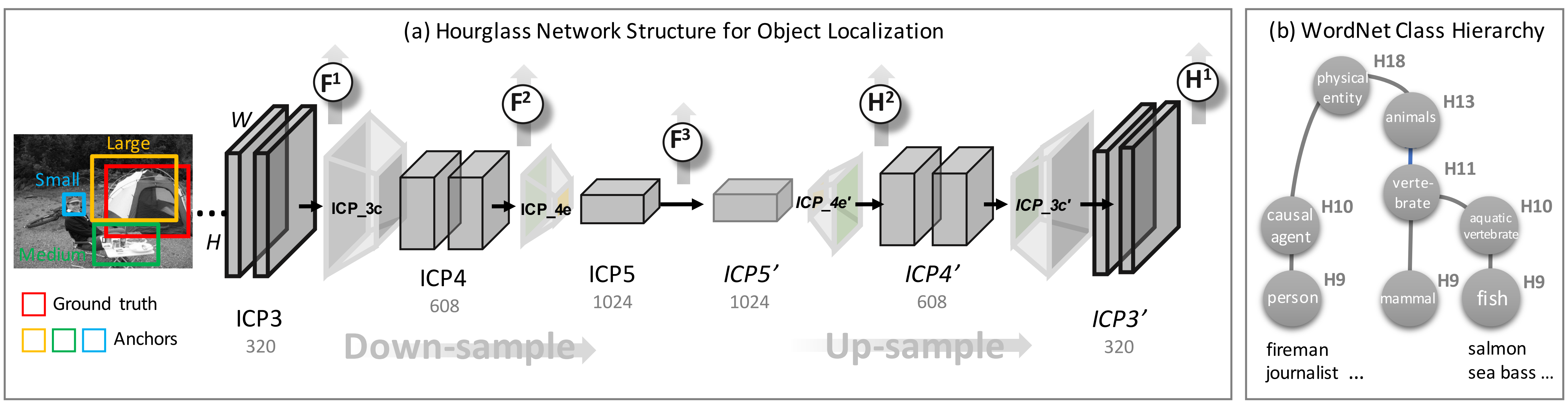
Algorithm

Anchor candidates

Scales : $\{16, 32\}, \{64, 128\}, \{256, 512\}$

The merged feature maps for loss input:

$$\mathbf{G}^m = \sigma(\mathbf{w}_F^m \otimes \mathbf{F}^m + \mathbf{w}_H^m \otimes \mathbf{H}^m + \mathbf{b}^m)$$



Algorithm

Training loss and inference

Loss function

$$L^m(p_i, t_i, k_i^*, t_i^*) \\ = -\frac{1}{N_1^m}\sum_i \log p_{i,k_i^*} + \frac{1}{N_2^m}\sum_i [k_i^* = 1]\mathcal{S}(t_i^*, t_i)$$

$L^m(p_i, t_i, k_i^*, t_i^*)$ is the loss for sample i on resolution level m.

$p_i = \{p_{i,k} | k = 0, \ldots K\}$ is the estimated probability.

t_i^* is the ground-truth regression offset.

k_i^* the ground-truth class label.

Algorithm

Training loss and inference

Total Loss

$$L = \sum_{m=1}^{M} L^m(p_i, t_i, k_i^*, t_i^*)$$

M is the number of resolution levels.

Remarks:

(a) Adjust image scale during training.

(b) Control the number of negative samples in a batch.

(c) Additional gray cate- gory.

Inner-level(threshold : 0.7) and inter-scale(threshold : 0.5) NMS[3] scheme.

Scales: Ranging from 1400 to 200 with an interval of 200.

[3]Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

Experiment

Pretrain

Setup and evaluation metric

Inception-BN on the EIC dataset : around **79% top-5** accuracy.

The base learning rate	0.0001 (50% drop every 7,000 iterations.
Momentum	0.9
Weight decay	0.0005
Maximum training iteration	200,000 (roughly 8 epochs)
Batch size	300
Aspect ratio (16 to 512)	[0.15, 0.5, 1, 2, 6.7]

Experiment

Component analysis

Structure	Rec@0.5
Down-sample alone	89.25
Down-sample + `splitAnc`	87.94
Deeper down-sample + `splitAnc`	92.33
Deeper hourglass	94.51

Scheme	Rec@0.5	AR@300
9 anchors (short for ac.)	87.33	-
30 ac.	94.51	-
30 ac. + `dyTrainScale`	95.33	59.34
+ `ctrlNegRatio`	↑ 1.78	-
+ `grayCls`	↑ 1.13	-
30 ac. + `all`	97.81	68.45

- The hourglass network in all settings.
- Rec@0.5 is the recall at IoU threshold 0.5 using top 300 proposals, evaluated on EIC validation set.
- We have the highest recall of 94.51, which proves the effectiveness of such a structure.

Experiment

Investigation on training data

- A larger dataset (EIC vs ILSVRC 1k) is beneficial to gain better results as more simples will ease overfitting if the model capacity is large.
- The base ordering is inferior for training the neural network as the model will severely bias towards direction in the feature space due to continuous samples of one class.
- A random sampling scheme ensures the classifier can witness various samples and the weights are quickly learned separately for each class, making the model robust and easy to converge.
- We find the amount of training data is not the most crucial point for obtaining a better model, but rather a good balance of the distribution among training samples weigh more.

EIC vs ILSVRC 1k

Training data strategy	AR@10	AR@100	AR@500
`ILSVRC_1k, base`	38.45	50.02	54.72
`ILSVRC_1k, random`	53.76	65.21	76.67
`ILSVRC_1k, balanced`	52.17	66.58	75.32
`EIC, base`	42.19	46.73	49.01
`EIC, random`	**59.31**	71.82	78.56
`EIC, balanced`	58.72	**72.39**	**81.27**
Selective search [26]	45.82	57.63	69.45
GOP [27]	52.66	63.21	74.93

Conclusion

- The Extended ImageNet Classification dataset
- Addressing the object localization problem by applying a conv-deconv structure in the region proposal framework, allowing different sizes of anchors placed at various depth in the network.
- More training data is good, and yet a balanced data distribution could achieve better results at the cost of less data.

EIC is here:

http://www.ee.cuhk.edu.hk/~yangli/project/eic.html

Thanks!