Recursive Boosting Approach for Object Localization

Hyungwon Choi[1*], Yunhun Jang[1*], Keun Dong Lee[2*], Seungjae Lee[2*], Jinwoo Shin[1*]

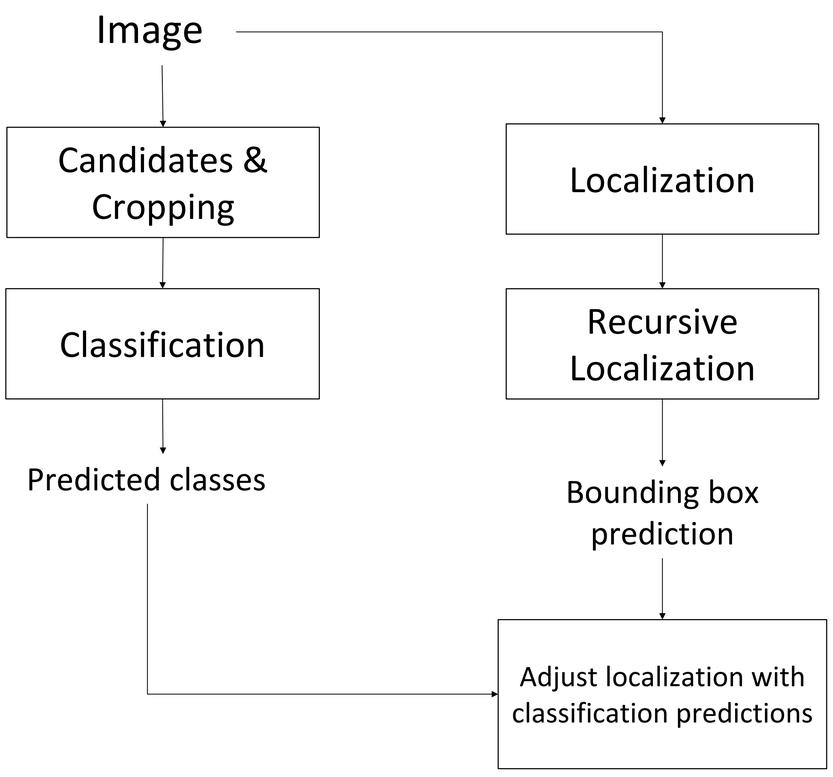
[1]School of Electronical Engineering, KAIST [2]Electronics and Telecommunications Research Institute, Korea

* indexes equal contribution, by Alphabets

Summary

- We prepared for 6 weeks with limited resources (4 GPU servers, 8 Titan X GPUs in total)
- A variant of GoogLeNet is used for localization
- GoogLeNet is pre-trained for classification task with batch normalization[3] and fine-tuned for localization
- VGG-16 and VGG-19 models are used for classification as well as boosting up the localization predictions
- Multiple crops on regular grid and selective crops based on objectness[4] score are used for classification

Our Approach

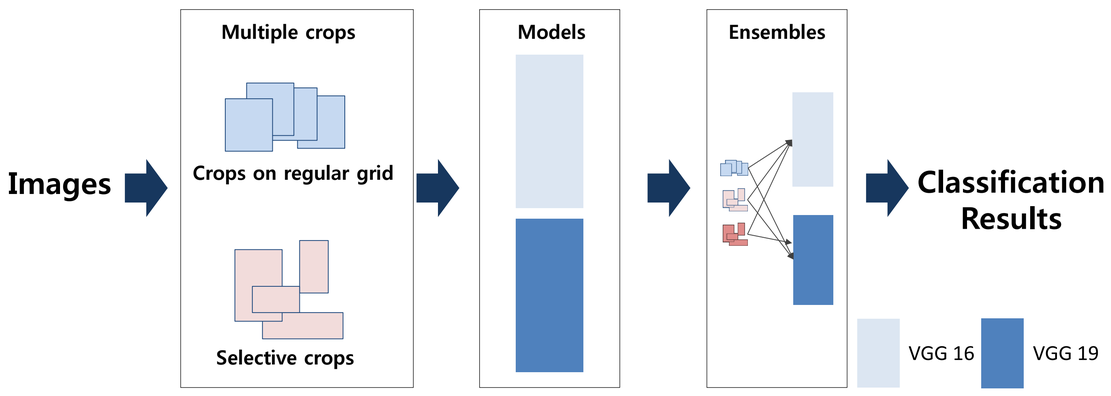


Our Contribution

- Recursive localization improves the localization performance by adjusting the bounding box with near 50% overlap
- Objectness based cropping improves classification performance
- Adjustment with classification predictions further improves localization performance

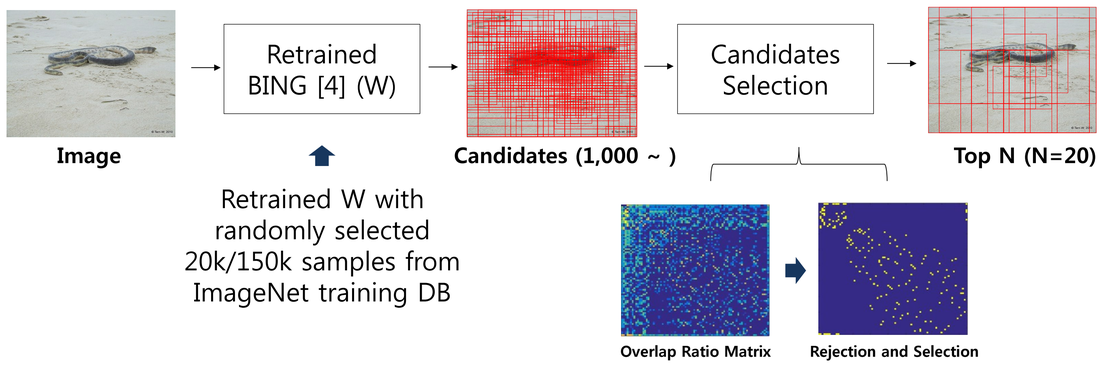
Classification Network

Classification process



Multiple crops

Multiple crops on regular grid, selective crops based on objectness score using similar method with BING[4]



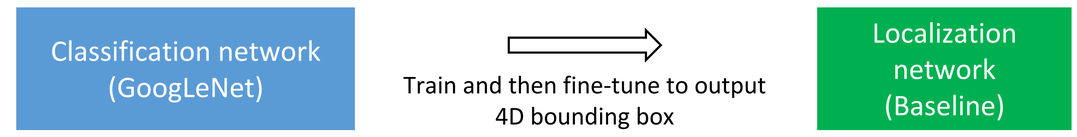
1. Generate overlap ratio matrix of candidates
2. Rejection with ratio constraints (e.g. size)
3. Sort and select Top N with overlap ratio mean of each candidate

Localization Network

Training

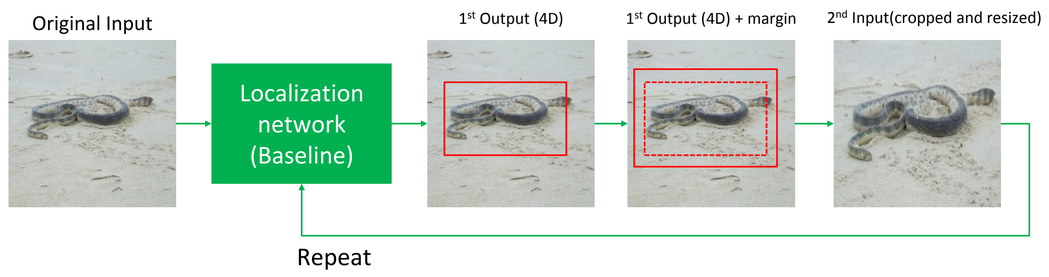
Train a baseline model(GoogLeNet) for classification task

Fine-tune the network to output 4D bounding boxes



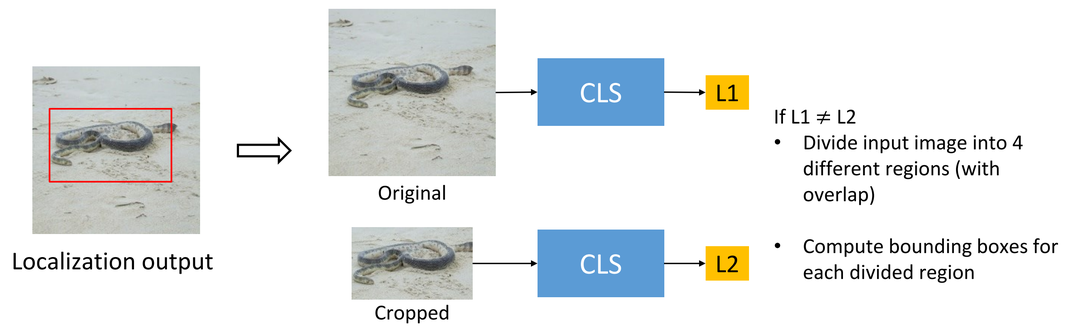
Recursive localization

Use predicted bbox with small margin as a new input in next iteration



Further adjustment

Reject current prediction if original predicted class is different from the prediction of the cropped image



Results

- **Classification results**

Models	No. of Crops	Multiple Crops	Retrained W	CLS Error (Top-1 /Top-5)
VGG16	150	Regular(50)	-	26.73% / 8.23%
VGG16	20	Selective(20)	W_1 (20K)	25.94% / 8.33 %
VGG16	20	Selective(20)	W_2 (150K)	26.04% / 8.37%
VGG16	40	Selective(20)+flips	W_1 (20K)	25.87% / 8.24 %
VGG16	40	Selective(20)+flips	W_2 (150K)	25.92% / 8.21%
VGG16 VGG19	230 (150+40+40)	Regular(50) Selective(20)+flips	W_1 (20K) W_2 (150K)	(Averaging) 24.58% / 7.34%
VGG16 VGG19	230 (150+40+40)	Regular(50) Selective(20)+flips	W_1 (20K) W_2 (150K)	(Fusion weight) 24.49%/ 7.32% (test set : 7.3 %)

Ensembles with multiple crops

- **Localization results**

Margin(pixel)	iter	Improvement
0	0	-
30	3	0.69%
50	3	3.1%
50	5	3.74%
70	5	3.96%
70	10	4.36%

Effectiveness of Recursive localization

Rejection criteria	Improvement
-	28.21%
top1	0.26%
top2	0.27%
top3	0.23%

Adjustment using classification predictions

Models	Recursive Localization	Adjustment using classification label	CLS-LOC error(val)	CLS(test)	CLS-LOC error(test)
A	X	X	32.8%	7.338%	-
B	O	X	28.21%	7.338%	28.7%
C	O	O	27.94%	7.338%	28.5%

Final results

References

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in Proc. CVPR, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.

[3] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. ICML, 2015.

[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in Proc. CVPR, 2014.