Rethinking the Route Towards Weakly Supervised Object Localization

Chen-Lin Zhang[1], Yun-Hao Cao[1], and Jianxin Wu[1]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

CVPR SEATTLE JUNE 14-19 2020 WASHINGTON

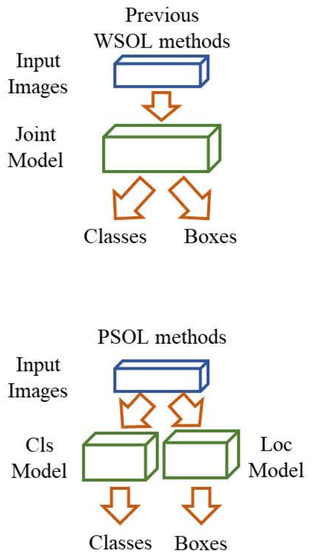
1. Introduction

Advantages of PSOL

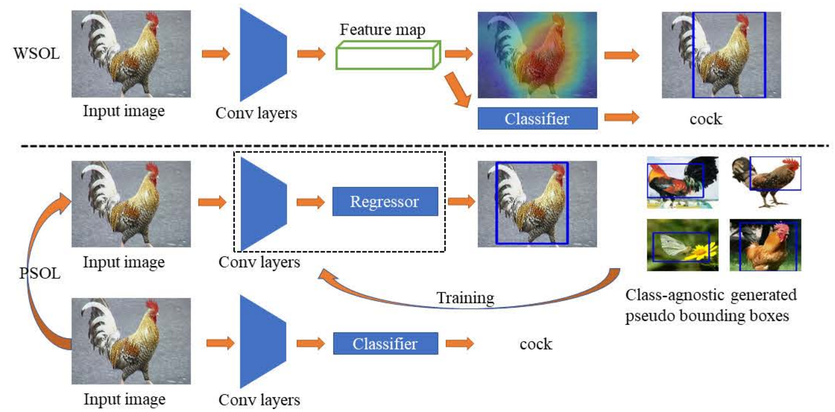
- ✓We propose a paradigm shift for the weakly supervised object localization (WSOL) task.
- ✓We achieve state-of-the-art results on ImageNet and CUB-200.
- ✓PSOL models have good transferability without fine-tuning.

Key idea

- ✓Dividing WSOL into two tasks: class-agnostic object localization and object classification
- ✓Using a class-agnostic method (DDT) to generate pseudo bounding boxes then perform object localization



2. PSOL Framework



- ➢ We first using a class-agnostic object co-localization method (DDT) to generate pseudo bounding boxes
- ➢ Then we use pseudo bounding boxes to directly train an object localization model
- ➢ We combine outputs from two model to form the final prediction

3. PSOL Details

Algorithm 1 Pseudo Supervised Object Localization

Input: Training images I_{tr} with class label L_{tr}
Output: Predicted bounding boxes b_{te} and class labels L_{te} on testing images I_{te}

1: Generate pseudo bounding boxes $\tilde{b}_{tr}$ on I_{tr}
2: Train a localization CNN F_{loc} on I_{tr} with $\tilde{b}_{tr}$
3: Train a classification CNN F_{cls} on I_{tr} with L_{tr}
4: Use F_{loc} to predict b_{te} on I_{te}
5: Use F_{cls} to predict L_{te} on I_{te}
6: **Return:** b_{te}, L_{te}

- ➢ For generating pseudo bounding boxes, we use a class-agnostic method: DDT.
- ➢ For localization CNN, we use class-agnostic bounding box regression techniques.
- ➢ For classification CNN, we directly use pre-trained models.

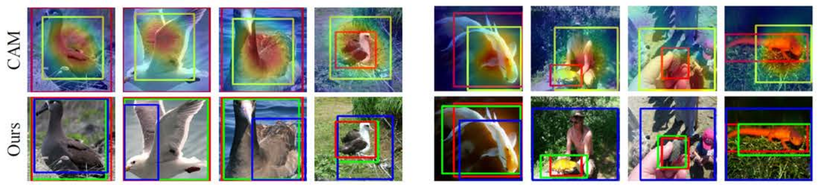
4. Experiments and Visualization

4.1 WSOL Benchmark Results

Model	Backbone	Parameters	FLOPs	CUB-200		ImageNet-1k		
				Top-1 Loc	Top-5 Loc	Top-1 Loc	Top-5 Loc	GT-Known Loc
VGG16-CAM [30]	VGG-GAP	14.82M	15.35G	36.13	-	42.80	54.86	59.00
VGG16-ACoL [28]	VGG-GAP	45.08M	43.32G	45.92	56.51	45.83	59.43	62.96
ADL [2]	VGG-GAP	14.82M	15.35G	52.36	-	44.92	-	-
VGG16-Grad-CAM [16]	VGG16	138.36M	15.42G	-	-	43.49	53.59	-
CutMix [27]	VGG-GAP	138.36M	15.35G	52.53	-	43.45	-	-
DDT-VGG16 [26]	VGG16	138.36M	15.42G	62.30	78.15	47.31	58.23	61.41
PSOL-VGG16-Sep	VGG16	274.72M	30.83G	**66.30**	**84.05**	**50.89**	**60.90**	**64.03**
PSOL-VGG16-Joint	VGG16	140.46M	15.42G	60.07	75.35	48.83	59.00	62.1
PSOL-VGG-GAP-Sep	VGG-GAP	29.64M	30.70G	59.29	74.88	48.36	58.75	63.72
PSOL-VGG-GAP-Joint	VGG-GAP	15.08M	15.35G	58.39	72.64	47.37	58.41	62.25
SPG [29]	InceptionV3	38.45M	66.59G	46.64	57.72	48.60	60.00	64.69
ADL [2]	InceptionV3	38.45M	66.59G	53.04	-	48.71	-	-
PSOL-InceptionV3-Sep	InceptionV3	53.32M	11.42G	**65.51**	**83.44**	**54.82**	**63.25**	**65.21**
PSOL-InceptionV3-Joint	InceptionV3	29.21M	5.71G	60.32	78.98	52.76	61.10	62.83
ResNet50-CAM [30]	ResNet50	25.56M	4.10G	29.58	37.25	38.99	49.47	51.86
ADL [2]	ResNet50-SE	28.09M	6.10G	62.29	-	48.53	-	-
CutMix [27]	ResNet50	26.61M	4.10G	54.81	-	47.25	-	-
PSOL-ResNet50-Sep	ResNet50	50.12M	8.18G	**70.68**	**86.64**	**53.98**	**63.08**	**65.44**
PSOL-ResNet50-Joint	ResNet50	26.61M	4.10G	68.17	83.69	52.82	62.00	64.30
DenseNet161-CAM	DenseNet161	29.81M	7.80G	29.81	39.85	39.61	50.40	52.54
PSOL-DenseNet161-Sep	DenseNet161	56.29M	15.46G	**74.97**	**89.12**	**55.31**	**64.18**	**66.28**
PSOL-DenseNet161-Joint	DenseNet161	29.81M	7.80G	74.24	87.03	54.48	63.41	65.39

- ➢ We achieve state-of-the-art WSOL results with different backbone models
- ➢ Separate PSOL models have a large gain over Joint PSOL models, which suggests that WSOL should be divided into two independent sub-tasks

4.2 Visualization



(a) CUB-200-2011 (b) ImageNet-1k

Figure 2: Comparison of our methods with CAM and DDT. Please note that in CAM figures, yellow boxes are CAM predicted boxes and red boxes are groundtruth boxes. In figures of our methods, blue boxes are DDT generated boxes, green boxes are predicted boxes by our regression model and red boxes are groundtruth boxes. We use the DenseNet161-Sep model to output DDT and predict boxes. This figure is best viewed in color and zoomed in.

4.3 Transferability

Model	Trained	Target	GT-Known Loc
VGG-GAP + CAM	CUB-200	CUB-200	57.96
VGG-GAP* + CAM	ImageNet	CUB-200	57.53
VGG16-ACoL + CAM	CUB-200	CUB-200	59.30
VGG16-ACoL* + CAM	ImageNet	CUB-200	58.70
SPG + CAM	CUB-200	CUB-200	60.50
SPG* + CAM	ImageNet	CUB-200	59.70
PSOL-VGG-GAP-Sep	CUB-200	CUB-200	80.45
PSOL-VGG-GAP-Sep	ImageNet	CUB-200	89.11
PSOL-DenseNet161-Sep	CUB-200	CUB-200	92.54
PSOL-DenseNet161-Sep	ImageNet	CUB-200	92.07

- ➢ Without any training, our PSOL models can have good transferability across different datasets, which proves that localization is a class-agnostic task.

4.4 Combine with State-of-the-art Classification Results

Table 4: Top-1 and Top-5 Loc results by combining localization of our models with more state-of-the-art classification models on ImageNet-1k.

Model	Top-1	Top-5
VGG16-ACoL+DPN131	53.94	61.15
VGG16-ACoL+DPN-ensemble	54.86	61.45
SPG + DPN131	55.19	62.76
SPG + DPN-ensemble	56.17	63.22
PSOL-InceptionV3-Sep + DPN131	55.72	63.64
PSOL-DenseNet161-Sep + DPN131	56.59	64.63
PSOL-InceptionV3-Sep + EfficientNet-B7	57.25	64.04
PSOL-DenseNet161-Sep + EfficientNet-B7	58.00	65.02

About Me Source Code

LAMDA Learning And Mining from DatA