Zoom-CAM: Generating Fine-grained Pixel Annotations from Image Labels

Xiangwei Shi, Seyran Khademi, Yunqiang Li, Jan van Gemert
Computer Vision Lab, Delft University of Technology, the Netherlands

1. Abstract

We propose Zoom-CAM visual explanations for generating pixel-level pseudo-labels from class labels for weakly supervised object localization and segmentation tasks. Zoom-CAM captures fine-grained small-scale objects for various discriminative class instances by integrating the visualizations from all intermediate convolutional layers in a CNN.

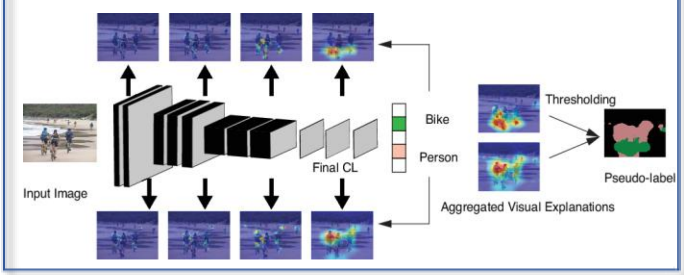
2. Methodology

Suppose $B_p(m,n)$ is the m, n-th activation in the p-th feature map of any intermediate layer in a classification CNN, then

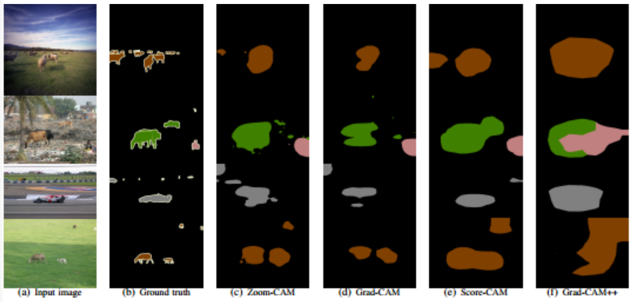
$$L^c_{m,n} \coloneqq ReLU\left(\frac{1}{Z}\sum_p \frac{\partial S^c}{\partial B_p(m,n)} B_p(m,n)\right),$$

is the visual explanation of that convolutional layer, where Z is the number of activations in an individual feature map and S^c is the final class score. Similar to Grad-CAM [1], we use the backwards gradient flow to quantify the contribution of activations to the class score. Differently, in Zoom-CAM, each activation is weighted individually.

3. Generating pseudo-labels with Zoom-CAM



4. Comparison of pseudo-labels

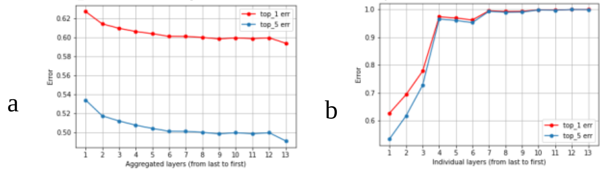

(a) Input image (b) Ground truth (c) Zoom-CAM (d) Grad-CAM (e) Score-CAM (f) Grad-CAM++

5. Experiments

1. Classification and localization error rates on ISVRC2012 val dataset. Zoom-CAM performs better than Grad-CAM.

	Classification error		Localization error	
	Top-1	Top-5	Top-1	Top-5
Zoom-CAM	31.87	11.54	**59.11**	**48.64**
Grad-CAM	31.87	11.54	61.95	52.35

2. Top-1 and top-5 localization error rates on ILSVRC2012 val dataset for ablation study. (a) aggregating intermediate feature maps, (b) single intermediate layer



3. Comparison of quality of pseudo-segmentation-labels of PASCAL VOC 2012 val set measured in IoU. Zoom-CAM generate better pseudo-labels than other methods.

Method	IoU																					mIoU
	backgr	plane	bike	bird	boat	bottle	bus	car	dog	chair	cow	dtable	cat	horse	motor	person	plant	sheep	sofa	train	tv	
Grad-CAM++	64.7	27.8	17.8	25.0	23.8	31.6	47.2	38.8	46.6	18.4	42.1	32.5	40.8	40.0	41.6	32.2	26.8	39.6	33.3	42.1	32.9	35.5
Grad-CAM	66.5	29.7	18.3	25.5	19.3	33.6	51.0	42.4	49.0	19.2	41.2	36.7	41.6	40.5	43.6	41.9	28.9	39.8	34.2	39.3	36.5	37.1
Score-CAM	68.1	**31.8**	19.1	**29.7**	**29.3**	30.9	50.3	**45.3**	47.9	19.8	41.8	32.3	**44.7**	42.0	**47.2**	35.4	27.9	42.8	36.6	**47.1**	31.8	38.2
Zoom-CAM	**68.9**	31.0	**19.7**	26.9	20.6	**34.5**	50.3	42.3	**50.1**	**20.4**	**45.6**	35.3	43.2	**43.8**	46.0	42.0	**31.1**	**45.0**	**38.3**	40.1	**38.6**	**38.8**

4. Semantic segmentation performance in mIoU evaluated on the PASCAL COC 2012 val set. The performance of weakly supervised semantic segmentation[2] using pseudo-labels generated by Zoom-CAM is better than the one by CAM

Method	*val*
IRNet(ResNet50)-CAM	63.5
IRNet(ResNet50)-Zoom-CAM	**64.6**

[1] Grad-CAM: Visual explanations from deep networks via gradient-based localization
[2] Weakly supervised learning of instance segmentation with inter-pixel relations