Revisiting Class Activation Mapping for Learning from Imperfect Data

Wonho Bae* Junhyug Noh* Jinhwan Seo Gunhee Kim
Seoul National University
bwh0324@gmail.com, {jh.noh, jinhwanseo}@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract

Weakly supervised object localization (WSOL) is a task of localizing an object in an image solely relying on image-level labels. To tackel the WSOL problem, most previous studies have employed a class activation mapping (CAM). Despite its universal use, in this work, we demonstrate it suffers from three fundamental issues: (i) the bias of GAP to assign a higher weight to a channel with a small activation area, (ii) negatively weighted activations inside the object regions and (iii) instability from the use of maximum value of a class activation map as a thresholding reference. They collectively cause the problem that the localization prediction to be highly limited to the small region of an object. We propose three simple but robust techniques that alleviate the problems, including thresholded average pooling, negative weight clamping and percentile as a thresholding standard. We participated in Learning from Imperfect Data (LID) 2020 challenge and won the 1st and 2nd places for weakly supervised object localization (Track 3) and semantic segmentation (Track 1).

1. Introduction

Contrast to fully-supervised object detection, the models for weakly supervised object localization (WSOL) are trained for classification solely relying on image-level labels. They utilize the feature map activations from the last convolutional layer to generate class activation maps from which bounding boxes are estimated. Since CAM approach [10] was initially introduced, most of previous studies on WSOL have followed its convention to first generate class activation maps and extract object locations out of them. However, this approach suffers from severe underestimation of an object region since the discriminative region activated through classification training is often much smaller than the object's actual region. For instance, according to the class activation map ($\mathbf{M}_k$) in Fig. 1, the classifier focuses on the *head* of the *monkey* rather than its whole *body*, since the activations of the *head* are enough to correctly classify the image as *monkey*. Thus, the bounding box reduces to delineate small highly activated *head* region only. To resolve this problem, recent studies have devised architectures to obtain larger bounding boxes by expanding activations [7, 6, 9, 1]. These methods have significantly improved the performance of WSOL and other relevant tasks such as weakly supervised semantic segmentation (WSSS).

In this work, however, we propose a different approach from previous researches; we focus on correctly utilizing the information that already exists in the feature maps. We summarize the contributions of this work as follows:

1. We discover three underlying issues residing in the CAM that hinder from properly utilizing the information from the feature maps for localization. Our analysis on CAM reveals the mechanism of how each component of CAM negatively affects the localization to be limited to a small region of an object. Based on the analysis, we propose three simple but robust techniques that significantly alleviate the problems.
2. We participated in LID 2020 challenge [8], and won the 1st and 2nd places for Track 3 (WSOL) and Track 1 (WSSS) using the same model.

2. Approach

We first review how the CAM [10] works in WSOL (section 2.1), and then elaborate its three problems followed by our solutions to alleviate the problems (section 2.2–2.4).

2.1. Preliminary: Class Activation Mapping (CAM)

In CNN trained for classification, a class activation map is the weighted sum of feature maps from the last convolutional layer with the weights from a fully connected (FC) layer. Let a feature map be $\mathbf{F} \in \mathbb{R}_{\geq 0}^{H \times W \times C}$. $\mathbf{F}_c \in \mathbb{R}_{\geq 0}^{H \times W}$ denotes c-th channel of $\mathbf{F}$. As described in Figure 1, a global average pooling (GAP) layer first averages each $\mathbf{F}_c$ spatially and outputs a pooled feature vector, $\mathbf{p}^{\text{gap}}$ as follows,

$$p_c^{\text{gap}} = \frac{1}{H \times W} \sum_{(h,w)} \mathbf{F}_c(h, w), \tag{1}$$

*equal contribution

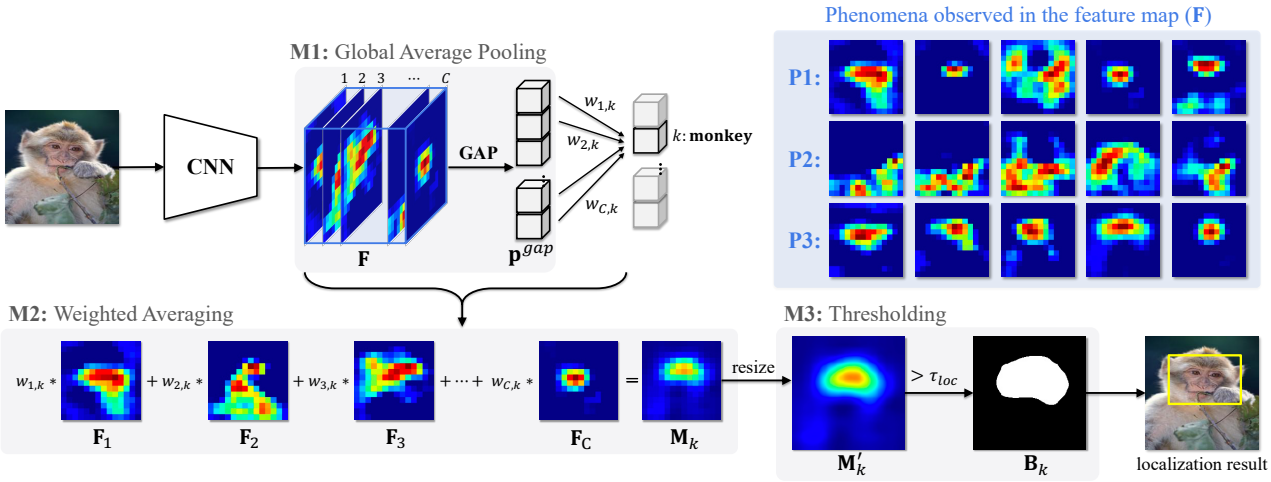


Figure 1: The overview of the CAM pipeline. We investigate three phenomena of the feature maps (**F**). **P1.** The areas of the activated regions largely differ by channel. **P2.** The activated regions corresponding to the negative weights ($w_c < 0$) often cover large parts of the target object (*e.g. monkey*). **P3.** The most activated regions of each channel significantly overlap at small regions. The three modules of CAM in gray boxes (**M1**–**M3**) does not take these phenomena into account correctly, which results in the localization being limited to small discriminative regions of an object.

where p_c^{gap} denotes a scalar of $\mathbf{p}^{\text{gap}}$ at c-th channel, and $\mathbf{F}_c(h, w)$ is an activation of $\mathbf{F}_c$ at spatial position (h, w).

The pooled feature is then transformed into K-dim logits through the FC layer where K is the number of classes. We denote the weights of the FC layer as $\mathbf{W} \in \mathbb{R}^{C \times K}$. Then the class activation map for a class k denoted as $\mathbf{M}_k$ is

$$\mathbf{M}_k = \sum_{c=1}^{C} w_{c,k} \cdot \mathbf{F}_c, \tag{2}$$

where $\mathbf{M}_k \in \mathbb{R}^{H \times W}$ and $w_{c,k}$ is an (c, k) element of $\mathbf{W}$.

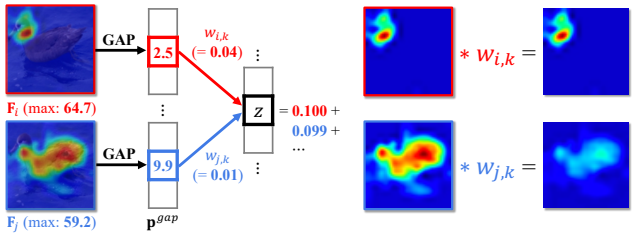
For localization, $\mathbf{M}'_k$ is first generated by resizing $\mathbf{M}_k$ to the original image size. With a localization threshold

$$\tau_{loc} = \theta_{loc} \cdot \max \mathbf{M}'_k, \tag{3}$$

a binary mask $\mathbf{B}_k$ identifies the regions where the activations of $\mathbf{M}'_k$ is greater than τ_{loc}: $\mathbf{B}_k = \mathbb{1}(\mathbf{M}'_k > \tau_{loc})$. Finally, localization is predicted as a bonding box that circumscribes the contour of the regions with the largest positive area of $\mathbf{B}_k$. Note that for Track 3, saliency maps from $\mathbf{M}'_k$ are produced instead.

2.2. Thresholded Average Pooling (TAP)

Problem. In WSOL, a GAP layer is employed to compute a weight of each channel to generate a class activation map. But, it tends to produce distorted weights for localization. As in Eq.(1), it sums all the activations and divides by $H \times W$ without considering the actual activated area per channel. The difference in the activated area is, however, not negligible. As an example in Fig 2, suppose i-th feature in (a) captures the *head* of a *bird* whereas j-th feature captures

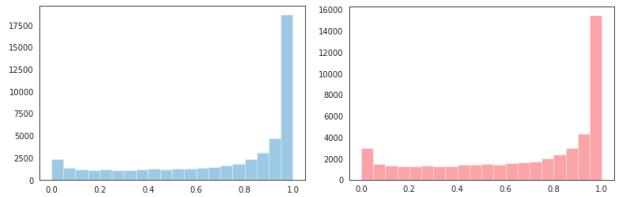


(a) Classification phase (b) Localization phase

Figure 2: An example illustrating a problem of using the GAP layer. The GAP layer causes the features with small activation area $\mathbf{F}_i$ to be underestimated so that the corresponding weight $w_{i,k}$ is trained to be larger than $w_{i,k}$. As a result, the weighted feature with small activation region, $w_{i,k} \cdot \mathbf{F}_i$, is highly overstated in localization phase.

its *body*. While the area activated in $\mathbf{F}_i$ is smaller than $\mathbf{F}_j$, the GAP layer divides both of them by $H \times W$, so the pooled feature p_i^{gap} of $\mathbf{F}_i$ is also smaller than p_j^{gap}. But, it does not mean the importance of $\mathbf{F}_i$ for classification is less than $\mathbf{F}_j$ as their contributions to logit z are almost the same as 0.1 and 0.099. For the ground truth class (k: *bird*), to compensate this difference, the FC weight $w_{i,k}$ corresponding to $\mathbf{F}_i$ is trained to be higher than $w_{j,k}$. As a result, when generating $\mathbf{M}_k$ in Eq.(2), small activated regions of $\mathbf{F}_i$ are highly overstated due to a large value of $w_{i,k}$. It causes localization to be limited to small region as localization depends on the maximum value of a class activation map.

Solution. To alleviate the problem of GAP, we propose the *thresholded average pooling* (TAP) layer. By replacing a GAP layer with a TAP layer, the pooled feature at c-th channel (Eq.(1)) is redefined as



(a) Positive weights (b) Negative weights

Figure 3: Intersection over Area (IoA) between the ground truth and predictions only using positive (a) and negative (b) weighted features. It indicates how much the features with the corresponding weights are activated in the object region. Surprisingly, a majority of the features with negative weights (b) are activated inside the objects. It is comparable to those with positive weights.

$$p_c^{\text{tap}} = \frac{\sum_{(h,w)} \mathbb{1}(\mathbf{F}_c(h,w) > \tau_{tap}) \mathbf{F}_c(h,w)}{\sum_{(h,w)} \mathbb{1}(\mathbf{F}_c(h,w) > \tau_{tap})}, \quad (4)$$

where $\tau_{tap} = \theta_{tap} \cdot \max \mathbf{F}_c$ denotes a threshold value where $\theta_{tap} \in [0, 1)$ is a hyperparameter.

2.3. Negative Weight Clamping (NWC)

Problem. When CNNs are trained for classification, a large number of the weights from the FC layer are negative. The features with negative weights help a model discriminate between different classes by decreasing the prediction probability of a target class. Existing CAM methods include the features with negative weights, and its underlying assumption is that they are mostly activated in *no-object* region like background. In contrast to this expectation, our analysis reveals many features with negative weights are concentrated within the object region as shown in Figure 3. Especially, their activations are high in the less discriminative regions compared to the features with positive weights.

We conjecture this phenomenon is closely related to the setting of WSOL: only one object is in an image. Suppose an image with a single object (*e.g. dog*). The features corresponding to negative weights mostly capture the characteristics of different classes (*i.e. cat*) inside the region of *dog* because they are similar to *dog* class not the background.

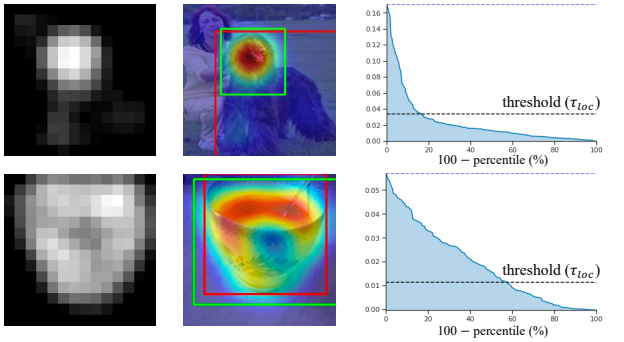
Solution. To mitigate this problem, we simply clamp negative weights to zero to generate a class activation map. Hence, Eq.(2) is redefined as

$$\mathbf{M}_k = \sum_{c=1}^{C} \mathbb{1}(w_{c,k} > 0) \cdot w_{c,k} \cdot \mathbf{F}_c. \quad (5)$$

By doing this, we can secure the activations that are depreciated in the object regions.

2.4. Percentile as a Thresholding Standard (PaS)

Problem. Another issue of the CAM method is that many channels have high activations at small overlapping regions. Figure 4 compares two examples of problematic (top) and successful (bottom) localization. Figure 4(a) depicts the number of channels whose activations are greater than $\tau_{0.8} = 0.8\times$(the max of weighted features) at each position. In the top row of Figure 4(c), when the activation distribution follows Zipf's law, the maximum value (dotted line in blue) is not a robust metric as a thresholding standard for localization, since the localization threshold τ_{loc} (dotted line in black) captures only small region of the object when high activations overlap. Contrarily, high activations are distributed throughout the object in the bottom successful case.



(a) Num of channels (activation $> \tau_{0.8}$) (b) Result with CAM (c) CAM values (descending order)

Figure 4: An example describing the problem of the overlap of high activations (top) compared to a successful case (bottom). In the top case, when high activations (activation $> \tau_{0.8}$) are concentrated in the small discriminative region, the localization threshold τ_{loc} becomes too high due to the high maximum of the class activation map.

Solution. Unlike the maximum, a percentile is one of the simplest but most robust metrics that are not sensitive to outliers and exponential distributions of activations. Hence, the Eq. (3) for the localization threshold τ_{loc} is redefined as

$$\tau_{loc} = \theta_{loc} \cdot \text{per}_i(\mathbf{M}'_k), \quad (6)$$

where per_i is an i-th percentile.

3. Experiments

We evaluate the proposed approach on ImageNet-1K [5] as a part of Track 3 in LID 2020 challenge [8]. Our approach largely improves the performance with ResNet50-SE [2, 3] backbone. We further evaluate our approach on WSSS task as a part Track 1 in LID challenge.

3.1. Experiment Setting

Dataset. Track 3 organizes 1.2 million training images of 1,000 different categories from ImageNet-1K [5]. Validation and test sets contain 23,151 and 21,120 images with pixel-annotation, respectively. Track 1 includes 456,567 training images, 5,000 validation and 10,000 test images with pixel-level annotations. Unlike Track 3, it assumes

Method	CRF	PaS	NWC	TAP	Peak IoU
Baseline					0.5254
	✓				0.5461
+ Ours	✓	✓			0.5563
	✓	✓	✓		0.5881
	✓	✓	✓	✓	**0.6370**

Table 1: Performance with different components applied.

Rank	Team	Peak IoU
1	**SNUVL (Ours)**	**0.63**
2	BJTU-Mepro-MIC	0.62
3	LEAP Group@PCA Lab	0.61
4	chohk (wsol_aug)	0.53
5	TEN	0.48

Table 2: Leaderboard of Track 3 (WSOL).

Rank	Team	Mean IoU
1	cvl	45.18
2	**SNUVL (Ours)**	**37.73**
3	UCU & SoftServe	37.34
4	IOnlyHaveSevenDays	36.24
5	play-njupt	31.90

Table 3: Leaderboard of Track 1 (WSSS).

multi-class objects per image. For both tracks, only image-level annotations are allowed to use in training step.

Implementation. We use a ResNet50-SE [2, 3] as a backbone network with slight modification for CAM. The images are resized to 384×384 and randomly cropped to 336×336 followed by horizontal flip. As a post-processing step, a fully connect CRF [4] is employed. Furthermore, we set $\theta_{tap} = 0.05$ and $i = 98$ for TAP and PaS, respectively.

Evaluation metric. We report the performance of models using *Peak IoU* and *Mean IoU* for Track 3 and 1, respectively. *Peak IoU* is the maximum of all the possible IoUs between the ground truths and predicted masks, and *Mean IoU* is the mean of IoUs of each class.

3.2. Quantitative Results

Track 3: WSOL. We demonstrate the effectiveness of each proposed solution on validation set. As shown in Table 1, adding each component largely improves the Peak IoU. As a result, we achieve 0.6370 on validation set, and 0.63 on test set which is ranked the 1st for Track 3 (Table 2).

Track 1: WSSS. We expected our methods would also work for Track 1 as long as a large number of images contain only one object. Our analysis reveals about 87 and 72 percent of images contain only one class. Unlike standard multi-class segmentation approach, we train a classification model on a single ground truth class chosen based on the number of images that a class belongs to. In inference, the model predicts segmentations of only one class. As a result, we won the 2nd place on Track 1 as in Table 3.

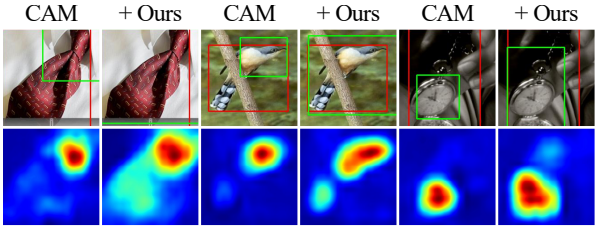


Figure 5: Qualitative results. The boxes in red and green represent the ground truths and predictions of localization.

3.3. Qualitative Results

We present the qualitative results for the proposed methods compared to the CAM. In Figure 5, the proposed methods help a model to utilize more activations in object region.

4. Conclusion

Despite the universal use of CAM, it contains three flaws which cause localization limited to small discriminative regions. Instead of endeavoring to extract additional information as done in the most of previous studies on WSOL, we proposed three simple but robust methods to properly utilize the information obtained from classification. We validated our methods largely mitigate the problems, and won the 1st and 2nd place for Track 3 and 1 in LID 2020 challenge.

References

[1] J. Choe and H. Shim. Attention-Based Dropout Layer for Weakly Supervised Object Localization. In *CVPR*, 2019.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[3] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2018.

[4] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[5] O. Russakovsky, J. Deng, H. SU, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[6] K. K. Singh and Y. J. Lee. Hide-And-Seek: Forcing a Network to Be Meticulous for Weakly-Supervised Object and Action Localization. In *ICCV*, 2017.

[7] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object Region Mining With Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In *CVPR*, 2017.

[8] Y. Wei, S. Zheng, M.-M. Cheng, H. Zhao, and etc. LID 2020: The Learning from Imperfect Data Challenge Results. 2020.

[9] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *CVPR*, 2018.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *CVPR*, 2016.