[Appendix] Rethinking Class Activation Mapping for Weakly Supervised Object Localization

Wonho Bae*, Junhyug Noh*, and Gunhee Kim

Department of Computer Science and Engineering,
Seoul National University, Seoul, Korea
bwh0324@gmail.com, jh.noh@vision.snu.ac.kr, gunhee@snu.ac.kr
http://vision.snu.ac.kr/projects/rethinking-cam-wsol

In this appendix, we further provide both quantitative and qualitative results in the following order.

- Section 1 shows the performance of Vanilla CAM with different combination of the proposed methods applied on other backbone structures.
- Section 2 illustrates how we choose the hyperparameters.
- Section 3 further exhibits the effect of each proposed method through qualitative results.
- Section 4 elaborates the relationship between the proposed negative weight clamping method and the number of objects in a dataset.
- Section 5 presents more qualitative results that demonstrate the robustness of the proposed methods on other backbones.

1 Quantitative Results with Different Components

In addition to the quantitative results with different components on VGG16 [4] as provided in Table 3 of the main draft, we further provide the experiment results on other backbone structures: ResNet50-SE [1, 3], MobileNetV1 [2] and GoogleNet [5].

Table 5, 6 and 7 show the performance of ResNet50-SE, MobileNetV1 and GoogleNet with different combination of the proposed methods applied, respectively. Regardless of different backbones, the performance of *Top-1 Loc* improves on both CUB-200-2011 and ImageNet-1K: ResNet50-SE (CUB: 43.29 $\rightarrow$ 58.39, ImageNet: 46.64 $\rightarrow$ 51.96), MobileNetV1 (CUB: 44.46 $\rightarrow$ 57.63, ImageNet: 43.29 $\rightarrow$ 45.55) and GoogleNet (CUB: 46.86 $\rightarrow$ 51.05, ImageNet 46.98 $\rightarrow$ 47.70).

From the experiment results with MobileNetV1 in Table 6, we can see that applying all of the proposed methods does not necessarily lead to the highest performance: the performance of TAP + NWC is slightly higher than TAP + NWC + PaS for *Top-1 Loc* by 0.41. It is because some of the three problems are not as clear on some backbones as on the other backbones. For example, the overall activations of the features from MobileNetV1 are much smaller than those from the other backbones. Because of small activations, the problem of the overlap of high activations is less severe on MobileNetV1 than the other

* Equal contribution.

backbones. Therefore, the combination of two solutions can be better than that of all the solutions due to the characteristics of different backbones.

Table 5. Performance of ResNet50-SE with different components applied.

Method	TAP	NWC	PaS	CUB-200-2011			ImageNet-1K		
				Top-1 Cls	GT Loc	Top-1 Loc	Top-1 Cls	GT Loc	Top-1 Loc
Baseline				**78.62**	56.49	43.29	77.22	58.21	46.64
+ Ours	✓			77.42	59.54	47.13	**77.25**	60.96	49.02
		✓		**78.62**	64.43	49.31	77.22	62.94	50.47
			✓	**78.62**	59.96	47.00	77.22	63.47	51.26
	✓	✓		77.42	68.97	53.30	**77.25**	62.84	50.49
	✓		✓	77.42	65.98	53.16	**77.25**	63.15	51.06
		✓	✓	**78.62**	66.85	51.69	77.22	64.28	51.73
	✓	✓	✓	77.42	**74.51**	**58.39**	**77.25**	**64.40**	**51.96**

Table 6. Performance of MobileNetV1 with different components applied.

Method	TAP	NWC	PaS	CUB-200-2011			ImageNet-1K		
				Top-1 Cls	GT Loc	Top-1 Loc	Top-1 Cls	GT Loc	Top-1 Loc
Baseline				72.09	58.92	44.46	67.34	59.45	43.29
+ Ours	✓			**75.82**	67.76	52.97	**68.07**	60.69	44.71
		✓		72.09	60.58	45.43	67.34	58.72	42.63
			✓	72.09	59.75	44.94	67.34	60.57	44.04
	✓	✓		**75.82**	**74.44**	**58.04**	**68.07**	59.28	43.67
	✓		✓	**75.82**	67.03	52.11	**68.07**	61.66	45.51
		✓	✓	72.09	62.89	46.95	67.34	60.85	44.15
	✓	✓	✓	**75.82**	74.28	57.63	**68.07**	**61.85**	**45.55**

Table 7. Performance of GoogleNet with different components applied.

Method	TAP	NWC	PaS	CUB-200-2011			ImageNet-1K		
				Top-1 Cls	GT Loc	Top-1 Loc	Top-1 Cls	GT Loc	Top-1 Loc
Baseline				74.35	61.67	46.86	70.50	62.32	46.98
+ Ours	✓			**75.04**	62.17	49.00	**71.09**	62.17	47.24
		✓		74.35	64.69	49.14	70.50	62.39	47.11
			✓	74.35	60.10	45.75	70.50	62.63	47.30
	✓	✓		**75.04**	**65.14**	50.66	**71.09**	62.04	47.12
	✓		✓	**75.04**	61.51	48.53	**71.09**	62.46	47.45
		✓	✓	74.35	64.48	48.62	70.50	**63.04**	47.57
	✓	✓	✓	**75.04**	65.10	**51.05**	**71.09**	62.76	**47.70**

2 Hyperparameter Tuning

In this section, we describe how the hyperparameters for TAP and PaS are tuned in detail. For the threshold τ_{tap} of TAP layer defined in Eq.(4), as specified in section 4.1 of the main draft, we set $\theta_{tap} = 0.1$ for VGG16 and MobileNetV1 and $\theta_{tap} = 0.0$ for ResNet50-SE and GoogleNet through hyperparameter tuning on the validation set randomly drawn 20% of CUB-200-2011 training set. However, i and θ_{loc} introduced in Eq.(6) are hyperparameters for localization where no

label is available in the training. Previous studies set such coefficients by either (i) using the threshold used in the original CAM paper (*e.g.* ADL, ACoL) or (ii) analyzing a few qualitative results (*e.g.* HaS, SPG, DANet). We employed (ii) by observing 20 qualitative results from VGG16 randomly drawn from CUB-200-2011 training set, and chose $\theta_{loc} = 0.35, i = 90$. They are fixed regardless of the backbones or datasets.

3 Qualitative Results by Proposed Methods

We illustrate some qualitative results by different proposed method. The following results empirically show the problems raised in the main draft and effectiveness of our proposed solutions.

3.1 Qualitative Results about Thresholded Average Pooling (TAP)

As stated in the section 2.2 of the main draft, the TAP layer decreases the bias introduced by the different size of the activated area per channel. Fig. 7 demonstrates the effectiveness of the TAP layer compared to the GAP layer. Given an image (1st column), the model with the GAP layer and the TAP layer generates CAMs in 2nd and 3rd columns, from which the bounding boxes are generated as shown in 4th and 5th columns, respectively. We can clearly see that with the TAP layer, the activations of a CAM are distributed throughout the object region, which is often not the case for the GAP layer.

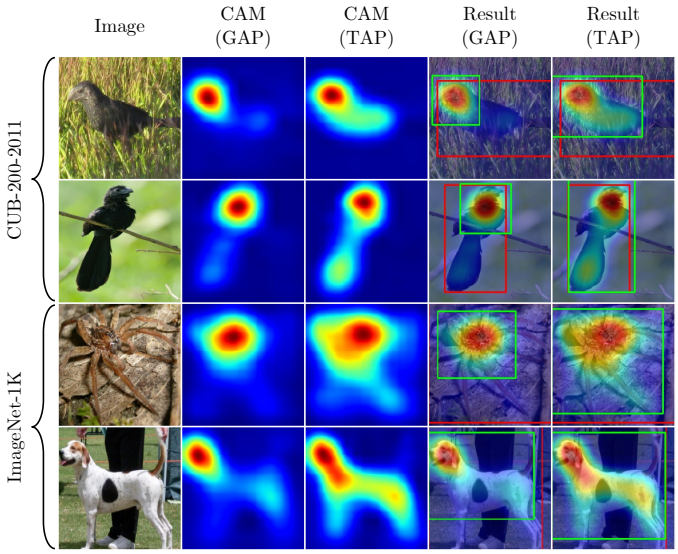

Fig. 7. Qualitative results comparing between the GAP and TAP layer. The boxes in red and green represent the GTs and predictions of localization, respectively.

3.2 Qualitative Results about Negative Weight Clamping (NWC)

Fig. 8 shows the effect of negative weight clamping. Given an image (1st column), we provide localization results overlaid with CAMs generated using positive (5th column), negative (6th column) and both (7th column) weights of $\mathbf{W}$. The 2nd–4th columns show the CAMs corresponding to 5th–7th columns, respectively. Fig. 8 evidently illustrates the problem of including the features corresponding to the negative weights as stated in the section 2.3 of the main draft. The CAMs generated only using the features with negative weights largely abate the activations of the object region. Using negative weight clamping, we prevent it from abating the activations in less-discriminative regions inside the objects.

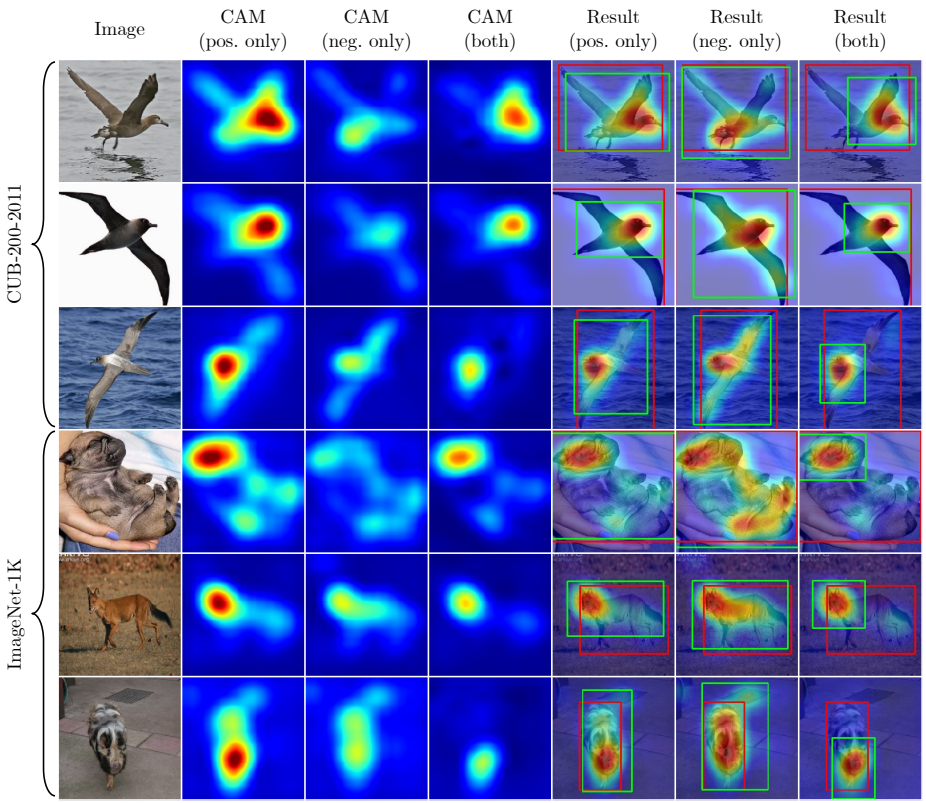

Fig. 8. Qualitative results comparing between the CAMs with and without negative weight clamping applied. Positive only (2nd and 5th columns) and both (4th and 7th columns) correspond to the CAM and localization results with and without negative weight clamping applied, respectively. The boxes in red and green represent the GTs and predictions of localization, respectively.

3.3 Qualitative Results about Percentile as a Standard (PaS)

Lastly, Fig. 9 illustrates the robustness of percentile (PaS) compared to the maximum as a standard (MaS) for the localization threshold. Note that replacing the standard to percentile does not change the activations of CAMs. Although there are many cases where the maximum standard properly estimates bounding boxes as shown in the first half of the columns, it often extracts too small bounding boxes as provided in the second half of the columns. On the other hand, the percentile standard more robustly estimates the locations of objects. The variance of bounding box sizes extracted using the maximum standard is much higher than those extracted using the percentile standard, depending on the distribution of the activations.

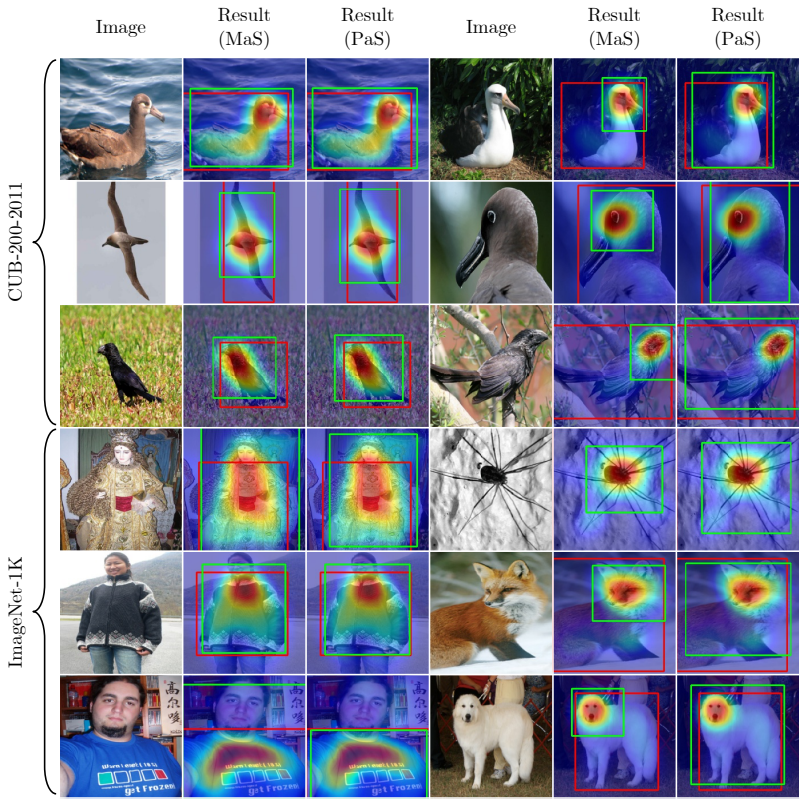

Fig. 9. Qualitative results comparing between maximum and percentile as a standard for the localization threshold. The first half of columns show the cases where both standard properly estimate the true bounding boxes whereas the second half of columns show the cases where only percentile properly estimates the boxes. The boxes in red and green represent the GTs and predictions of localization, respectively.

4 Results of Negative Weights in Multiple Object Cases

In section 4.2 of the main draft, we state when multiple objects exist in an image, the features with negative weights tend to be activated in the object regions of different classes. To elaborate this phenomenon, we provide some examples where the features with negative weights are activated in the background region when there are multiple objects in the given image.

The images in Fig. 10 which contain multiple objects are all from ImageNet-1K dataset. Given the original images (1st column), we provide localization results (5th–7th columns) and only CAMs (2nd–4th columns). The CAMs are generated using the FC weights of either positive only, negative only or both as specified at the top of Fig. 10. The images in the 3rd and 6th columns show that when there are multiple objects in the image, the features corresponding to the negative weights tend to be more activated in the object that is not a target class for classification. As a result, after negative weight clamping, the final CAM captures broader regions than the regions of the GT object.

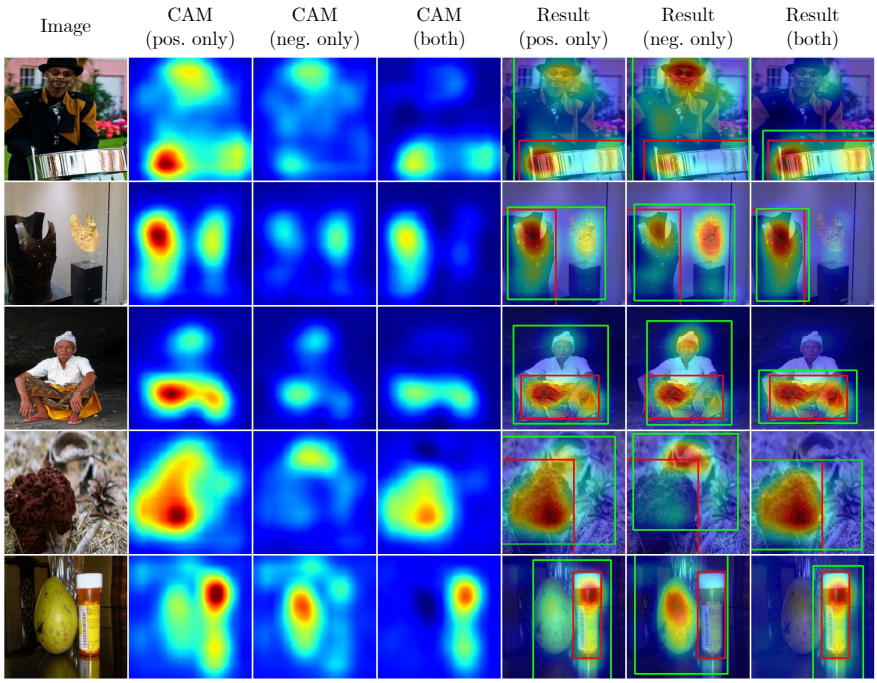

Fig. 10. Qualitative results illustrating the activations of CAMs with negative weights for the multiple object cases. The 3rd and 6th columns show that when multiple objects exist in an image, the feature maps corresponding to the negative weights tend to be activated in the object that is not a GT class for classification. The boxes in red and green represent the GTs and predictions of localization, respectively.

5 More Qualitative Results on Other Backbones

In addition to the qualitative results provided on VGG16 and ResNet50-SE in Fig. 6 of the main draft, we further provide the qualitative results with the other backbones: MobileNetV1 and GoogleNet. As with VGG16 and ResNet50-SE, MobileNetV1 and GoogleNet show similar tendency; the activations from our methods are largely distributed throughout the object regions compared to the baselines. One thing to notice from the first three examples of MobileNetV1 is that the proposed methods do not just expand the activations of CAMs.

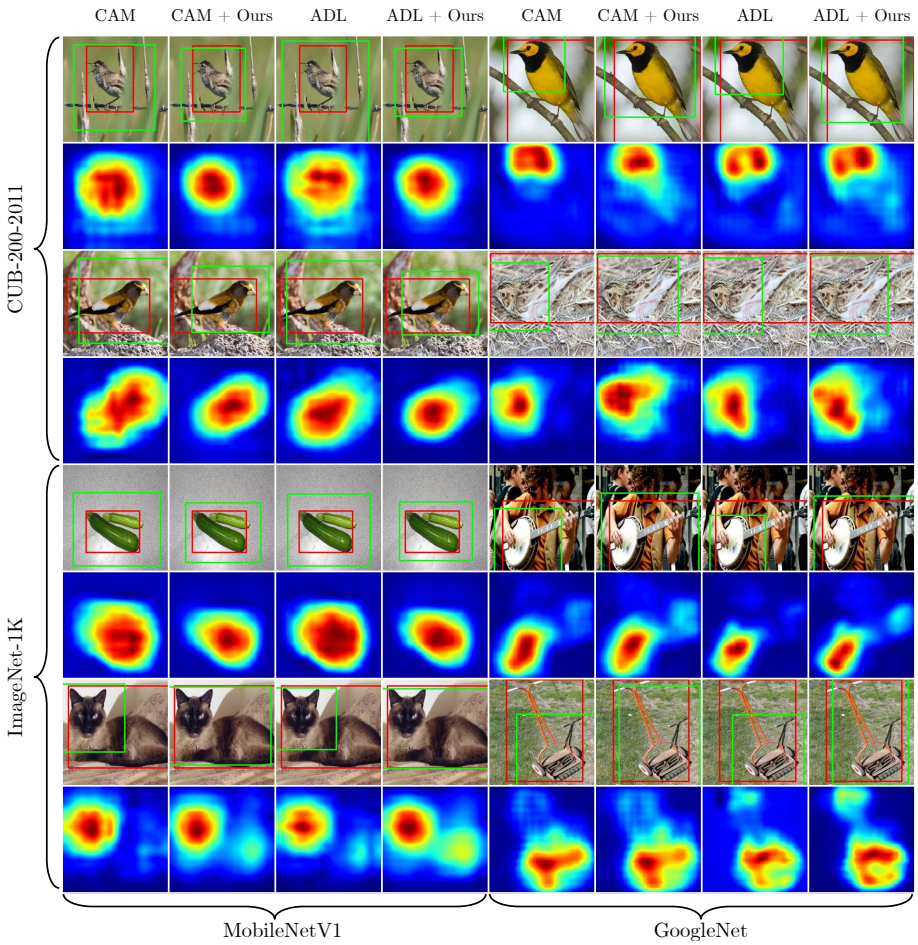

Fig. 11. Qualitative results with MobileNetV1 and GoogleNet on CUB-200-2011 and ImageNet-1K datasets. The boxes in red and green represent the GTs and predictions of localization, respectively.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016)
2. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Arxiv:1704.04861 (2017)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: CVPR (2018)
4. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. Arxiv:1409.1556 (2014)
5. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: CVPR (2015)