Appendix of BagCAMs

Lei Zhu, Qian Chen, Lujia Jin, Yunfei You, and Yanye Lu*

[1] Institute of Medical Technology, Peking University
[2] Department of Biomedical Engineering, Peking University
[3] Institute of Biomedical Engineering, Peking University Shenzhen Graduate School
https://github.com/zh460045050/BagCAMs
zhulei@stu.pku.edu.cn, yanye.lu@pku.edu.cn

1 Key symbols List

Here we give the pivotal symbols list in Table. 1 to enhance the readability of our main paper. In general, we use ***italic bold*** uppercase characters to denote the matrices. Vectors are denoted with lowercase. Sets are noted by blackboard bold, for example $\mathcal{F}$.

Table 1. The Meaning of Some Pivotal Symbols

$\boldsymbol{X} \in \mathbb{R}^{3\times N}$	RGB feature of the input image.
$\boldsymbol{Z} \in \mathbb{R}^{C\times N}$	Pixel-level features of $\boldsymbol{X}$
$\hat{\boldsymbol{P}} \in \mathbb{R}^{K\times N}$	Initial Pixel-level localization scores of BagCAMs.
$\boldsymbol{P}^* \in \mathbb{R}^{K\times N}$	Output Pixel-level localization scores of BagCAMs.
$\boldsymbol{\Lambda} \in \mathbb{R}^{N\times N}$	Co-efficient matrix of regional localizers
$\mathbf{W} \in \mathbb{R}^{K\times C}$	Weight matrix of the classifier.
$\boldsymbol{z} \in \mathbb{R}^{C\times 1}$	Image-level features of $\boldsymbol{X}$.
$\boldsymbol{p} \in \mathbb{R}^{K\times 1}$	Localization score of a spatial position (column vector of $\hat{\boldsymbol{P}}$).
$\boldsymbol{s} \in \mathbb{R}^{K\times 1}$	Image-level classification scores.
$\bar{\boldsymbol{s}} \in \mathbb{R}^{K\times 1}$	Logarithmic Image-level classification scores.
$\boldsymbol{y} \in \mathbb{R}^{K\times 1}$	Image-level ground truth mask.
N^I	The spatial resolution of input $\boldsymbol{X}$.
N	The spatial resolution of feature $\boldsymbol{Z}$.
C	The number of channel for feature.
K	The number of object class.
$\mathcal{F}^*$	The base localizer set of BagCAMs that contains N^2 localizers.
$\mathcal{P}^*$	The set of localization scores generate by localizers in $\mathcal{F}^*$.
$e(\cdot) : \mathbb{R}^{3\times N} \rightarrow \mathbb{R}^{C\times N}$	The feature extractor, implemented by ResNet/Inception.
$c(\cdot) : \mathbb{R}^{C\times 1} \rightarrow \mathbb{R}^{K\times 1}$	The image classifier, implemented by fully-connected layer.
$f(\cdot) : \mathbb{R}^{C\times N} \rightarrow \mathbb{R}^{K\times N}$	The object localizer, derived by CAM-based methods.

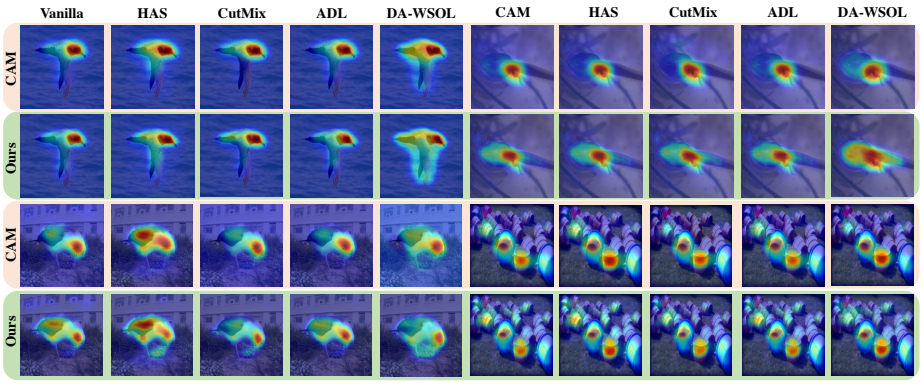

Fig. 1. Visualization on replacing CAM into BagCAMs for different WSOL methods.

2 Proof of the formulation of base localizer

Here we detailed the proof of the formulation of the base localizer of our BagCAMs defined in our Sec.3.2:

$$f_n^m(\boldsymbol{x})_k = \sum_{c_1} \boldsymbol{s}_k (1 + \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_1,m}} \boldsymbol{Z}_{c_1,m}) \sum_{c_2} (\frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} \boldsymbol{x}_{c_2}) \quad . \tag{1}$$

Proof. When adopting PCS [4] to initialize the coarse localization score $\hat{P}_{k,m}$ of our BagCAMs, the formulation of f_n^m obtained by RLG becomes:

$$f_n^m(\boldsymbol{x})_k = \sum_{c_2} \frac{\partial (\sum_{c_1} \frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_1,m}} \boldsymbol{Z}_{c_1,m})}{\partial \boldsymbol{Z}_{c_2,n}} \boldsymbol{x}_{c_2} = \sum_{c_1,c_2} (\frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_2,n}} + \frac{\partial^2 \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_1,m} \partial \boldsymbol{Z}_{c_2,n}} \boldsymbol{Z}_{c_1,m}) \boldsymbol{x}_{c_2}. \tag{2}$$

Considering that ReLU is used as the activation function of the extractor, based on the prior work [1], the partial derivative $\frac{\partial^2 \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_1,m} \partial \boldsymbol{Z}_{c_2,n}}$ can be reformulated as:

$$\frac{\partial \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_2,n}} = \boldsymbol{s}_k \frac{\bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} \quad , \quad \frac{\partial^2 \boldsymbol{s}_k}{\partial \boldsymbol{Z}_{c_1,m} \partial \boldsymbol{Z}_{c_2,n}} = \boldsymbol{s}_k \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_1,m}} \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} \quad . \tag{3}$$

Finally, taking Eq. 3 into Eq. 2, the formulation of Eq. 1 can be obtained:

$$\begin{aligned} f_n^m(\boldsymbol{x})_k &= \sum_{c_1,c_2} (\boldsymbol{s}_k \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} + \boldsymbol{s}_k \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_1,m}} \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} \boldsymbol{Z}_{c,m}) \boldsymbol{x}_c \\ &= \sum_{c_1} \boldsymbol{s}_k (1 + \frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_1,m}} \boldsymbol{Z}_{c_1,m}) \sum_{c_2} (\frac{\partial \bar{\boldsymbol{s}}_k}{\partial \boldsymbol{Z}_{c_2,n}} \boldsymbol{x}_{c_2}) \quad . \end{aligned} \tag{4}$$

□

3 Visualizations of InceptionV3 extractor

In our main paper, we only included the visualization of our method with the ResNet50 extractor. Here we give more visualization when using InceptionV3 [3] as the backbone extractor. In detail, Fig. 1 shows that replacing CAM with BagCAMs for different WSOL methods can also enhance the performance of the baseline methods under the InceptionV3. Moreover, our BagCAMs still achieve better performance when localizing objects based on features of intermediate layers than other methods [1,2,4] as indicated in Fig. 2. These are in accordance with the results with the ResNet50 backbone as discussed in our main paper.

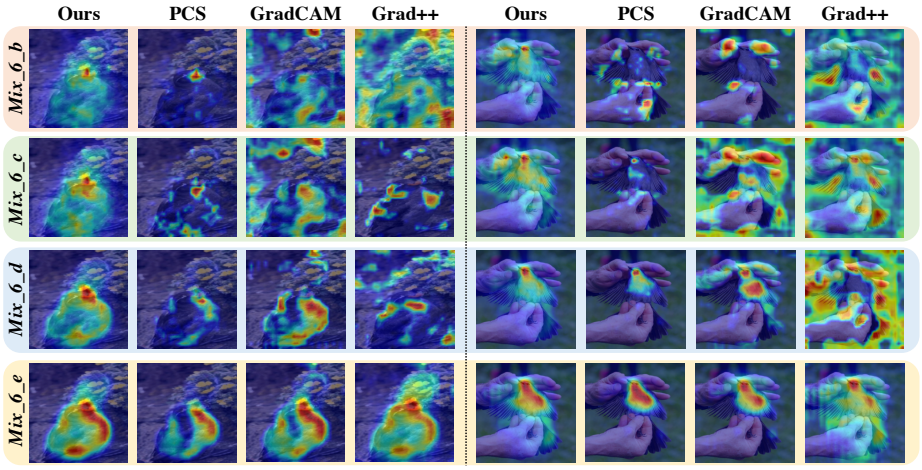

Fig. 2. Localization map generated based on different InceptionV3 layer.

References

1. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847. IEEE (2018)
2. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE/CVF International Conference on Computer Bision (ICCV). pp. 618–626 (2017)
3. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
4. Tan, C., Gu, G., Ruan, T., Wei, S., Zhao, Y.: Dual-gradients localization framework for weakly supervised object localization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1976–1984 (2020)