applied sciences

MDPI

Article

Weakly Supervised Learning for Object Localization Based on an Attention Mechanism

Nojin Park and Hanseok Ko *

School of Electrical Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea; njprogrammer@korea.ac.kr
* Correspondence: hsko@korea.ac.kr

Abstract: Recently, deep learning has been successfully applied to object detection and localization tasks in images. When setting up deep learning frameworks for supervised training with large datasets, strongly labeling the objects facilitates good performance; however, the complexity of the image scene and large size of the dataset make this a laborious task. Hence, it is of paramount importance that the expensive work associated with the tasks involving strong labeling, such as bounding box annotation, is reduced. In this paper, we propose a method to perform object localization tasks without bounding box annotation in the training process by means of employing a two-path activation-map-based classifier framework. In particular, we develop an activation-map-based framework to judicially control the attention map in the perception branch by adding a two-feature extractor so that better attention weights can be distributed to induce improved performance. The experimental results indicate that our method surpasses the performance of the existing deep learning models based on weakly supervised object localization. The experimental results show that the proposed method achieves the best performance, with 75.21% Top-1 classification accuracy and 55.15% Top-1 localization accuracy on the CUB-200-2011 dataset.

Keywords: weakly supervised object localization; attention mechanism; joint training

Citation: Park, N.; Ko, H. Weakly Supervised Learning for Object Localization Based on an Attention Mechanism. *Appl. Sci.* **2021**, *11*, 10953. https://doi.org/10.3390/app112210953

Academic Editor: Giancarlo Mauri

Received: 8 July 2021
Accepted: 13 November 2021
Published: 19 November 2021

1. Introduction

To this day, deep learning models have performed better than heuristic methods, but they have been expressed as a 'black box', and thus, it is not clear how the models work. Recently, in contrast to the concept of a 'black box', Explainable AI (XAI) has been introduced. XAI is a methodology that enables humans to understand and trust the results and outputs created by deep learning models. In other words, XAI allows humans to understand how deep learning models solve certain problems. For example, in situations involving the detection of an object in the field of object detection, when an image is given as an input to a deep learning model, humans are able to understand which part the model saw in the given image to produce the result. In XAI, explainability involves two factors: interpretability (transparency) and completeness. The former describes the structure of a deep learning model so that humans can understand it, and the latter can explain how a deep learning model works in an accurate way. To meet these two requirements, Activation-Based Methods (ABMs) can explain what part of the input image the deep learning model has seen and made judgments about. In the most widely known method, a Class Activation Map [1] uses the weights of the activated values of the layers of convolutional neural networks (CNNs).

Weakly Supervised Learning (WSL) [2] is a machine learning framework that trains a deep learning model using the partial labels of training samples. In the field of computer vision and image processing, WSL has been mainly used for classification and object localization. Furthermore, WSL tasks are divided into object classification, segmentation, 3D object reconstruction and object localization. Weakly Supervised Object Detection (WSOD) [3,4] aims to detect every object from an image. Unlike supervised segmentation,

Weakly Supervised Segmentation (WSS) [5,6] performs segmentation by receiving a bounding box or a scribble annotation as a label without pixel level annotation. In WSL, 3D object reconstruction performs reconstruction task [7] using only the pixel or bounding-box level instead of voxel or mesh level annotation. Weakly Supervised Object Localization (WSOL) aims to localize an object as well as classify it without expensive labelling. That is, WSOL identifies the locations of objects by training the model without the need for bounding box annotations of the locations of these objects in an image. When training a deep learning model without bounding box annotations, localizing the object becomes difficult using only the classifiers of Convolutional Neural Networks (CNNs) because the feature extractor of a CNN learns only the features useful for classifying objects. That is, the classifier learns only the characteristic part of the input images and falls into sub-optimal localization. Therefore, for example, when a deep learning model identifies a bird's position, it localizes only the beak or face, which are important features of a bird. An example of this is shown in Figure 1.

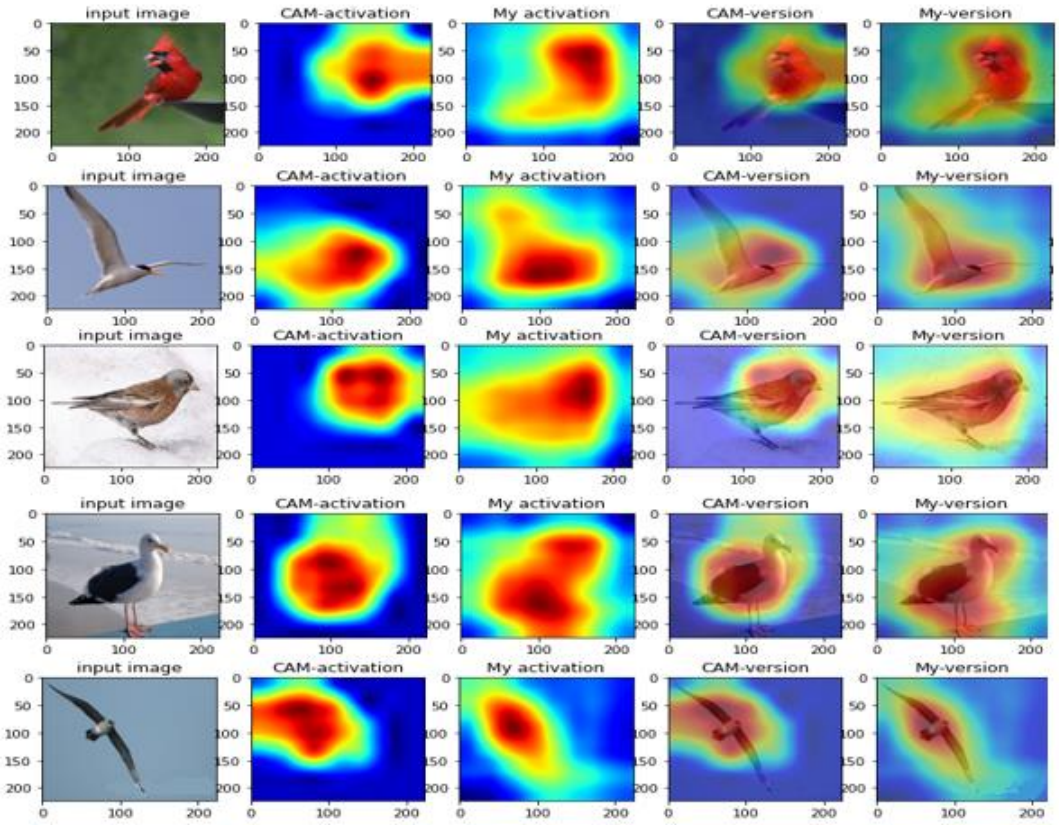

Figure 1. Sub-optimal localization problems: using the CAM [1] method alone, birds' tails, wings, and legs are not localized.

In order to solve this problem, a previous paper [8] proposed an approach involving the application of the regional dropout method to the input data, and several papers [9–11] have proposed various methods to solve the sub-optimal localization problem. However, in WSOL, this problem is still a difficult issue to solve.

In this paper, we propose a method of adjusting the classifier weights to solve the sub-optimal problem with greater efficiency than the existing method. Our proposed method is composed of two classifiers, and the features reduced from those that are intensively used by the first classifier are given as inputs to the second classifier so that the wider area of an object can be learned.

Our contributions are as follows:

- There is no pre-processing task, such as hiding part of an image, required before training the model. Therefore, it does not take extra time to train the model.

- The number of learning parameters is minimized without an additional layer to solve the sub-optimal problem of WSOL, but the proposed method shows better performance than the existing methods.
- Some of the existing studies had to inject the same input several times to obtain the attention result for one image at the inference stage, but our method can obtain the attention with only one input.

The rest of the paper is composed as follows. Section 2 describes the related research work, while Section 3 describes the proposed method. Section 4 presents and discusses the experimental test results and Section 5 provides the concluding remarks.

2. Related Work

2.1. Class Activation Map

CAM [1] is the most representative Activation-Based Method (ABM) of Explainable AI (XAI). The ABM method uses a weight value that linearly combines the activation values from the convolutional layers to explain how CNNs draw conclusions. The CAM uses the global average pooling (GAP) method proposed in [12] to prevent the loss of location when the fully-connected layer of the CNNs flattens the feature maps. The CAM equation is as follows:

$$S_c = \sum_k w_k^c \sum_{x,\,y} f_k(x,\,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y) \tag{1}$$

$$M_c(x,\,y) = \sum_k w_k^c f_k(x,y) \tag{2}$$

When the scalar value located at the pixel $(x,\,y)$ in the k-the channel element of the feature map is expressed as $f_k(x,\,y)$, the value after GAP is $\sum_k f_k(x,\,y)$. Here, w_k^c is a weight corresponding to the k-to channel of the feature map and class c, and S_C is a value given as an input of the SoftMax function for class c. M_c in Equation (2) is an activation map for class c, and $M_c(x,\,y)$ denotes the eventual influence of the information located at pixel $(x,\,y)$ on class c. As shown in Figure 2, if only CAM is used, attention is paid to only the most characteristic part when classifying the class of the input image, leading to sub-optimal localization in the object localization task. In order to overcome the disadvantages of using CAM mentioned above, recent efforts [11,13–15] tried to overcome the sub-optimal localization problem by adding different algorithms. Dane [13] is a method involving the creation of attention maps by adding 1×1 convolution to each convolutional layer and then adding all attention maps to localize the object. AcoL [14] is a method that involves creating an attention map by adding the attention maps extracted from the two classifiers. ADL [15] produced a drop mask using thresholding and an importance map with a sigmoid activation function. Here, an attention map was obtained through random selection and object localization was performed. In CAAM [11], not only the highest probability of the classifier but also the lowest value were combined to increase the object localization performance.

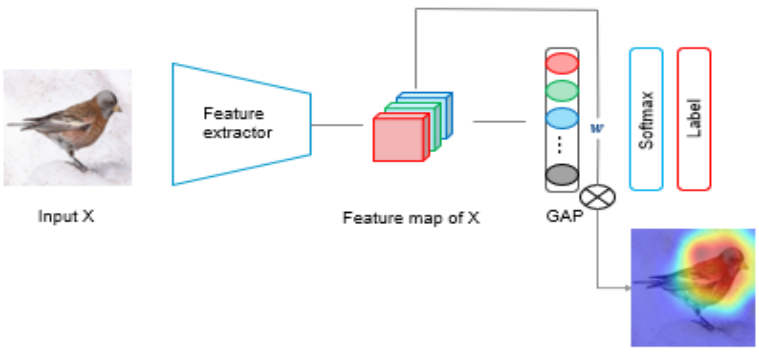

Figure 2. Network structure of Class Activation Map.

2.2. Attention Mechanism

The Attention Mechanism has been applied to not only natural language processing but also other fields [16–18], and recently, it has been applied to image recognition [19–21]. In Residual Attention Networks [19], several attention modules were added to the backbone network to learn the mask in each convolutional layer. In general, the attention module has been used as shown in Equation (3).

$$H_{i,\,c}(x) = M_{i,c}(x) \cdot T_{i,c}(\mathrm{x}) \tag{3}$$

where $H(x)$ is the attention module, $M(x)$ is the attention mask, and $T(x)$ is the filter. In other words, the attention module emphasizes the robust features by performing element-wise multiplication of an existing filter and a mask for all indices and channels. However, as the layer of the CNN model becomes deeper, the gradient vanishing problem develops. To solve this, in Residual Attention Networks the attention module is modified as shown in Equation (4) [19].

$$H_{i,\,c}(x) = (1 + M_{i,c}(x)) \cdot T_{i,c}(\mathrm{x}) \tag{4}$$

Through the approach described above, the performance of classification was improved by adding attention modules to the convolutional layer in [22]. In the attention branch networks [21], the class attention map is focused on improving the CNN performance by introducing an attention mechanism that focuses on a specific region of an image. The structure of the ABN consists of a feature extractor, an attention branch, and a perception branch. That is, the above study improved the performance of image recognition by using an attention map and adding an attention mechanism to the existing CNNs.

ABN provided improved image recognition performance by using an attention mechanism to give more robust features as inputs to the classifier.

2.3. Weakly Supervised Object Localization

Weakly supervised learning (WSL) has recently received a lot of attention in the fields of computer vision and image processing. WSL is a field for training deep learning models with only labels that are cheaper than the labels required for supervised learning. For example, to perform 3D object reconstruction, Voxel or mesh level labels are required in supervised learning, but in WSL, a pixel or bounding box level that is cheaper than the labels required in supervised learning is required. In this field of WSL, Weakly supervised object localization (WSOL) mainly aims to localize a single object in each given image scene. In a previous paper [1] related to the WSOL, object localization was performed by mapping the class score predicted in the inference stage and the feature map of the last layer of the feature extractor. However, this method [1] encountered the suboptimal problem of locating only the main points without locating the entire object. This was because, in supervised learning, the object location is given as a label to perform the object localization task, whereas in WSOL, only the class of the object that exists in the input image is simply given as a label. In order to solve such research issues, in previous papers [8,9], some improvements were made by adding a preprocessing process for the input data or adding specific layers. In addition to this, previous papers [10,23,24] tried to solve the sub-optima problem through various methods.

3. Proposed WSOL Method with Adjusted Weights

In ABN [21], the classifier used more robust features to classify by dot-producting the attention map in the attention branch to the input feature map of the perception branch. Although this approach improves the performance of the classifier, since it further strengthens the distinct characteristics of the object, it suffers from the WSOL sub-optimal problem. In this paper, the attention mechanism method used in ABN [21] is modified to improve the extracted attention map, such that the classifier and feature extractor would instead allow the possibility of identifying a larger area of the object. First, the proposed

method is shown in Figure 3, and the differences between ABN [21] and the proposed method are briefly shown in Figure 4.

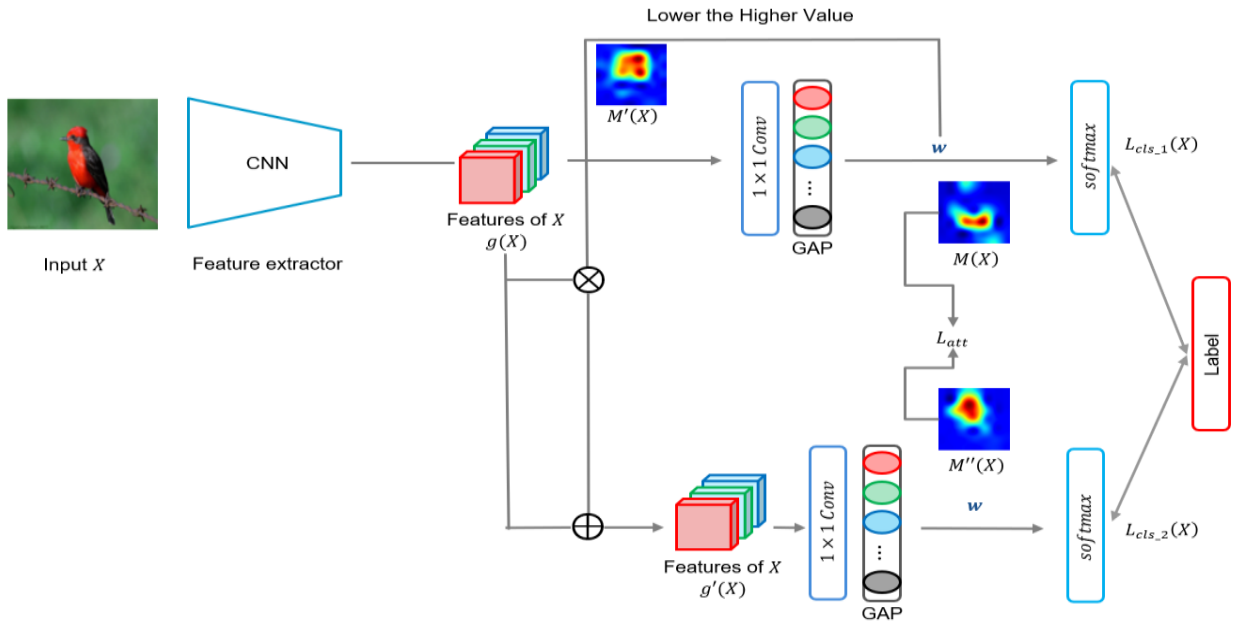

Figure 3. Structure of the proposed method.

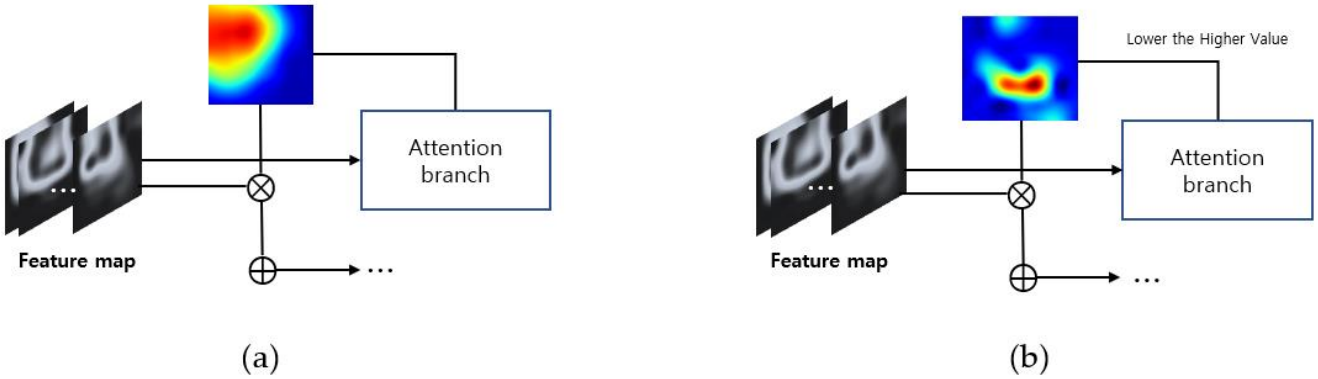

Figure 4. Differences between ABN [21] and the proposed method: (**a**) ABN, (**b**) proposed method. (**a**) uses the class activation map method to improve the performance of the classifier. However, this method gives a greater weight to the robust feature map in the training process. Therefore, this method encounters a sub-optimal problem in the field of object localization. On the other hand, our proposed method (**b**) creates a new feature map by reducing the high value so that it can learn areas other than the area used in the upper classifier.

The method is now described in detail. For input image X, the upper classifier after the feature extractor computes the probability that the input image belongs to each class using the feature map extracted from the feature extractor as input, and then creates attention map $M(X)$. Since the generated attention map $M(X)$ is a feature that is deemed important for prediction by the upper classifier, the lower classifier of the model uses a method to compensate for the large value of the attention map extracted from the upper classifier. This method can be described as Equation (5).

$$\begin{gathered} M'_c(x,\ y) = M_c(x,y) - \overline{M_c(x,y)} \\ where, \quad M_c(x,y) > \overline{M_c(x,y)} \\ \overline{M_c(x,y)} = \left(\frac{1}{H \times W} \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} M_c(x,y) \right) \end{gathered} \tag{5}$$

The elements of $M'_c(x,\ y)$ are formed by a process of significant reduction in the highlighted values of the attention map $M_c(x,\ y)$, and the attention map extracted from the upper branch of Figure 2 becomes the area used mainly by the classifier of the upper branch. In this respect, the approach is identical to that in the paper for CAM [1], and also encounters the sub-optimal localization problem, as shown in Figure 4. Thus, we created an attention map that is a new M' by mitigating the elements of M greater than $\frac{1}{H \times W} \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} M_c(x,y)$, as in Equation (5). Here, when generating M' that mitigates M, we replaced only those pixels larger than the average by taking a difference from the average. In fact, besides this particular approach, we also experimented with a variety of other architectural schemes to deemphasize the large values for better distribution. Among all the schemes considered that included the inversion of large and small values of the attention map M of the upper classifier and the replacement of the large values with 0, the proposed method showed the best results. Subsequently, by dot-producting the feature map extracted from the feature extractor, the feature map could be provided as the input value to the bottom branch, such that the input feature map of the bottom branch could be written as Equation (6).

$$g'_c(x_i) = \left(1 + M'(x_i)\right) \cdot g_c(\mathrm{x_i}) \tag{6}$$

Unlike ABN [21], the proposed method leaves open the possibility of learning different features of the top branch by providing a feature map that mitigates the robustly used features in the top classifier being used as inputs to the bottom classifier. Thus, the attention map $M''(X)$ could be created at the bottom branch, just as the attention map $M(X)$ is created at the top branch. Afterward, we trained the model so that the two attention maps became similar through joint training. That is, unlike the previous studies, we applied the attention mechanism and the two attention maps extracted from two different classifiers to object localization. As a result of applying the proposed method, it showed superior performance compared to the existing methods by attaining greater classification accuracy, which will be presented in the Experiments section.

Finally, the total loss function for this paper is shown in Equation (7):

$$\begin{gathered} L_{total}(x_i) = L_{cls_1}\Big(y,\ y_{pred}(g(X))\Big) \\ + L_{cls_2}(y,\ y_{pred}(g'(X)) \\ + \alpha \cdot L_{att}(M(x,\ y),\ M''(x,\ y)) \end{gathered} \tag{7}$$

where $L_{cls_1}\Big(y,\ y_{pred}(g(X))\Big)$, and $L_{cls_2}(y,\ y_{pred}(g'(X))$ are the categorical cross-entropy loss function and $L_{att}(M(x,\ y),\ M''(x,\ y))$ is the mean square error loss function. In this experiment, α was experimentally given a setting of 5.

4. Experiments

To verify our proposed method, we compared the performance of many types of WSOL methods with the CUB-200-2011 dataset. We used VGG-16 or Inception-v3 as the feature extractor. The learning rate was 0.0001, the batch size was 128, and the value of α was 5. We implemented the proposed framework using PyTorch on an RTX 2080Ti GPU with 64 GB of RAM. Our preprocessing only needed to make the input data and the input size of the feature extractor the same, which took only 7.21 s. However, the preprocessing of the Hide-and-Seek [8] algorithm took an average of 41.61 s per epoch, and

the Where to Look [9] preprocessing took an average of 78.86 s per epoch. Nevertheless, the experimental results revealed that our proposed method exhibits superior performance in terms of Top-1 localization accuracy and Top-1 classification accuracy [18] compared to the other WSOL algorithms. The Top-1 localization accuracy calculates the fraction of images that are correctly classified with the predicted bounding box having 50% IoU (intersection over union) with the ground truth bounding box. Meanwhile, the Top-1 classification accuracy determines the fraction of images that are correctly classified.

4.1. Dataset

CUB-200-2011 [24]: the CUB-200-2011 dataset includes 200 species of birds, consisting of 5994 train images and 5794 test images.

- Number of categories: 200;
- Number of images: 11,788;
- Annotations per image: 15 Part Locations, 312 Binary Attributes, 1 Bounding box.

When training the model, only the species of birds were used as the labels.

4.2. Results of the Experiments

The Figures 5–8 present the results for the prediction of the bounding box of an object using the proposed method.

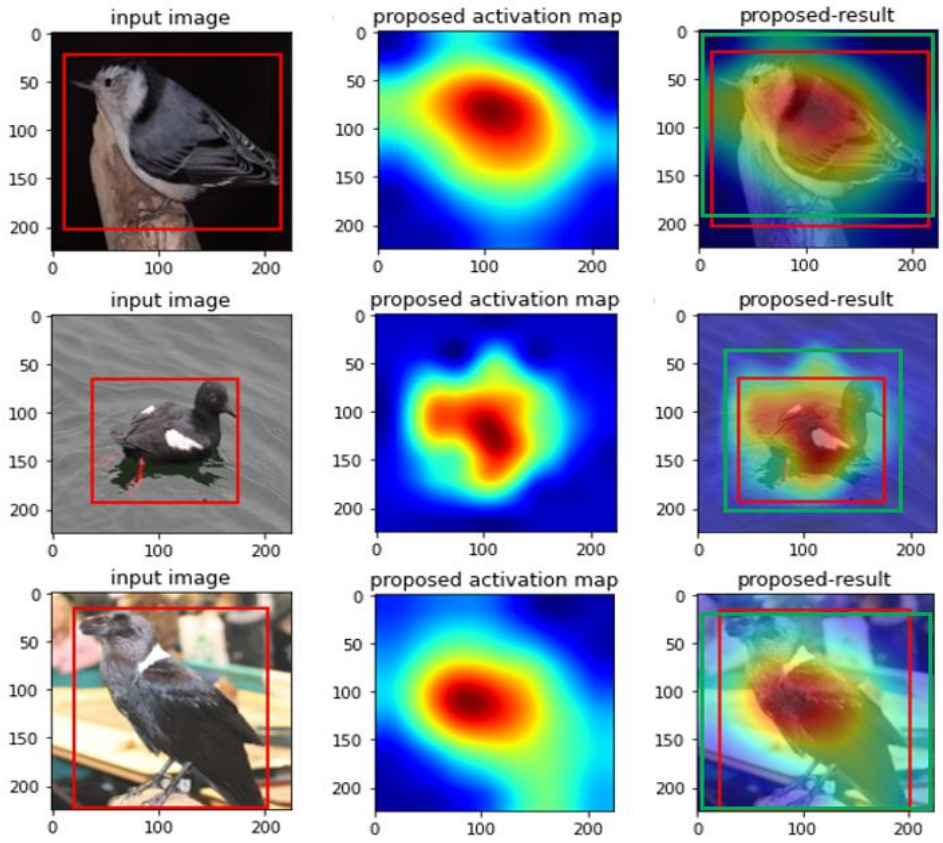

Figure 5. Results of the proposed method (1).

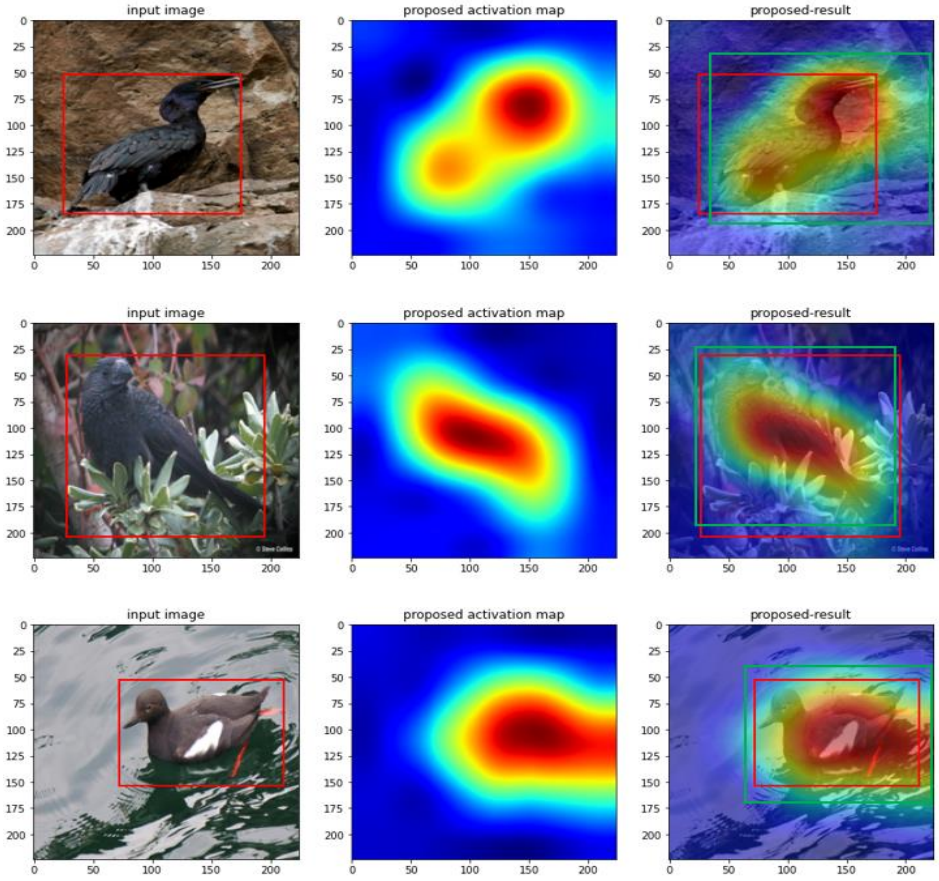

Figure 6. Results of the proposed method (2).

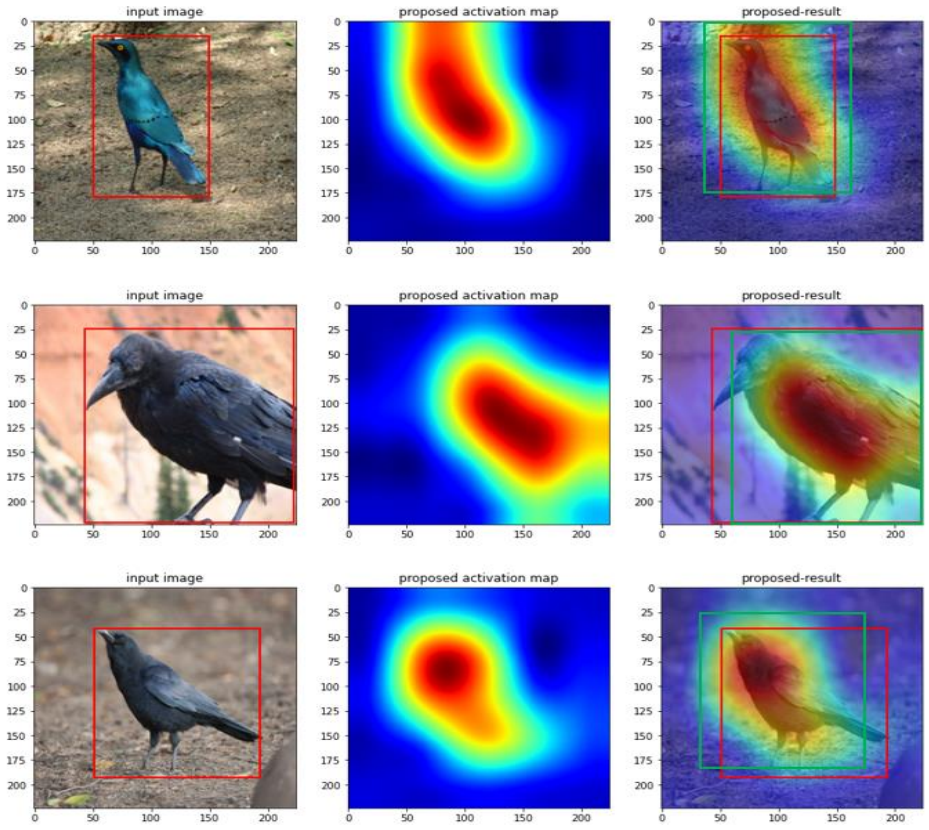

Figure 7. Results of the proposed method (3).

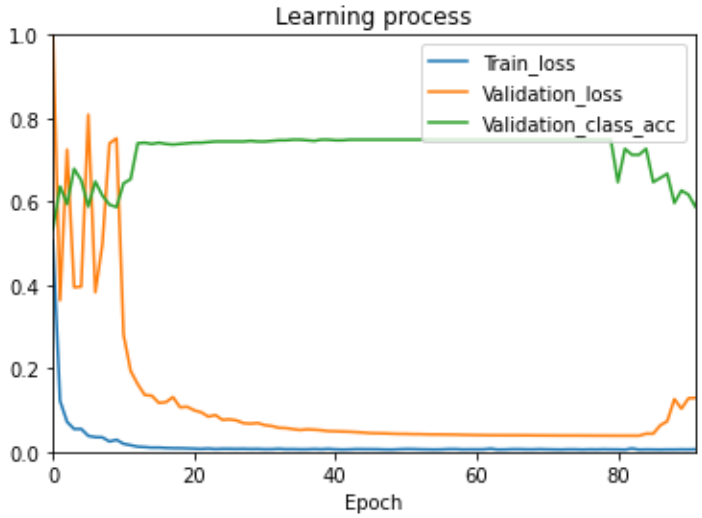

Figure 8. Result of the learning process per epoch.

Finally, Table 1 compares the Top-1 localization accrual and Top-1 classification accrual with existing studies, and it is shown that the proposed method outperformed the existing methods.

Table 1. Comparative experiments using other baseline algorithms.

Method	Top-1 Loc	Top-1 Clas
Inception V3-CAM [1]	43.67	73.80
Inception V3-DANet [13]	49.45	71.20
Inception V3-Ours	52.71	75.21
VGG-CAM [1]	34.41	67.55
VGG-AcoL [14]	45.92	71.90
VGG-ADL [15]	52.36	65.27
VGG-CCAM [11]	50.07	73.20
VGG-Ours	55.15	73.51
Best-Performance	55.15	75.21

Inception V3 and VGG-16 were used as feature extractors for comparative experiments with other existing methods [1,19–22].

5. Discussion and Conclusions

In this paper, we introduced a WSOL method that used the pre-FC layers of VGG-16 and Inception-v3 as feature extractors, as well as using an attention mechanism. In order to minimize the learning parameters and minimize preprocessing, the proposed method in this paper simplified the model by only introducing an attention mechanism, rather than using the methods applied in existing studies, in order to avoid extensive sub-optimal localization problems. In this experiment, the proposed method showed higher performance on the CUB-200-2011 dataset with 75.21% top-1 classification accuracy and 55.15% top-1 localization accuracy. To establish why our proposed object localization outperformed other existing algorithms even without adding preprocessing and convolutional layers, we generated two attention maps from different feature maps and minimized the difference between these attention maps. The results provided vindication that the use of two attention maps makes it possible to cover a wider area than the area covered by attention maps generated by one classifier.

Author Contributions: Conceptualization, N.P.; Formal analysis, N.P.; Project administration, H.K.; Supervision, H.K.; Writing—original draft, N.P.; Writing—review & editing, H.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Government-Wide R&D Fund Project for Infectious Disease Research (GFID), Republic of Korea (grant number: HG19C0682).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929. [CrossRef]
2. Zhou, Z. Abrief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
3. Naoto, I.; Ryosuke, F.; Toshihiko, Y.; Kiyoharu, A. Cross-domain weakly-supervised object detection through progressive domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 5001–5009.
4. Chenhao, L.; Siwen, W.; Dongqi, X.; Yu, L.; Wayne, Z. Object Instance Mining for Weakly Supervised Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 11482–11489.
5. Yude, W.; Jie, Z.; Meina, K.; Shiguang, S.; Xilin, C. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Sementic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12275–12284.
6. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2209–2218.
7. Gwak, J.; Choy, C.; Chandraker, M.; Garg, A.; Savarese, S. Weakly supervised 3D Reconstruction with Adversarial Constraint. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 263–272.
8. Singh, K.K.; Lee, Y.J. Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action location. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3544–3553. [CrossRef]
9. Babar, S.; Das, S. Where to Look? : Mining complementary image regions for weakly supervised object localization. In Proceeding of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1010–1019.
10. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5267–5276. [CrossRef]
11. Yang, S.; Kim, Y.; Kim, Y.; Kim, C. Combinational class activation maps for weakly supervised object localization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 2941–2949. [CrossRef]
12. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
13. Xue, H.; Liu, C.W.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. Danet: Divergent activation for weakly supervised object localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6589–6598.
14. Zhang, X.; Wei, Y.; Feng, J.; Yang, Y.; Huang, T. Adversarial complementary learning for weakly supervised object localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1325–1334.
15. Choe, J.; Shim, H. Attention-based dropout layer for weakly supervised object localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2219–2228.
16. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015. [CrossRef]
17. Larochelle, H.; Hinton, G.E. Learning to combine foveal glimpses with a third-order Boltzmann machine. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1243–1251. [CrossRef]
18. Woo, S.; Park, J.; Lee, J.; Kweon, I. Cbam: Convolutional black attention module. In Proceedings of the European Conference on Computer Vision(ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. [CrossRef]
19. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164. [CrossRef]
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
21. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10705–10714. [CrossRef]

22. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Image large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
23. Ku, B.; Kin, G.; Ahn, J.; Lee, J.; Ko, H. Attention-Based Convolutional Neural Network for Earthquake Event Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, 1–5. [CrossRef]
24. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. In *Computation & NeuralSystems Technical Report*; CNS-TR-2011-001; California Institute of Technology: Pasadena, CA, USA, 2011.