Attention-based Dropout Layer for Weakly Supervised Object Localization

Junsuk Choe, and Hyunjung Shim*
School of Integrated Technology, Yonsei University, South Korea
{junsukchoe, kateshim}@yonsei.ac.kr

Abstract

Weakly Supervised Object Localization (WSOL) techniques learn the object location only using image-level labels, without location annotations. A common limitation for these techniques is that they cover only the most discriminative part of the object, not the entire object. To address this problem, we propose an Attention-based Dropout Layer (ADL), which utilizes the self-attention mechanism to process the feature maps of the model. The proposed method is composed of two key components: 1) hiding the most discriminative part from the model for capturing the integral extent of object, and 2) highlighting the informative region for improving the recognition power of the model. Based on extensive experiments, we demonstrate that the proposed method is effective to improve the accuracy of WSOL, achieving a new state-of-the-art localization accuracy in CUB-200-2011 dataset. We also show that the proposed method is much more efficient in terms of both parameter and computation overheads than existing techniques.

1. Introduction

Weakly Supervised Object Localization (WSOL) aims to identify the location of the object in a scene only using image-level labels, not location annotations. Existing approaches mine and track discriminative features of each class for object detection [45, 36, 37, 9, 45, 25, 21, 41, 19, 2, 39, 15, 63, 7, 5, 4, 48, 14, 65, 32, 31, 58, 62, 8, 6] and segmentation [33, 29, 18, 16, 24, 52, 50]. Because the discriminative power of each object part is different from another, these techniques tend to identify only the most discriminative part of the target object, incapable of covering entire extent of the object. For example, in the case of a person, the face may be more discriminative than the body which appearance changes dramatically due to clothing. In this case, existing WSOL techniques can localize only the face, not the entire region.

This problem can be critical in object localization. Specifically, Class Activation Mappings (CAM) [63] utilize Convolutional Neural Networks (CNN) classifier for learning the discriminative features. The key idea is that the classifier with a reasonable accuracy should observe the object region to decide the class label. In other words, the discriminative features should co-occur with the object region. From this idea, they perform localization by tracking spatial distribution of feature responses. Unfortunately, the classifiers tend to focus only on the most discriminative features to increase their classification accuracy. Therefore, the spatial distribution of feature responses also tends to cover only the most discriminative part of the object, which leads to localization accuracy degradation.

Recently, various techniques [49, 35, 17, 59, 20, 52, 51, 60] have been proposed to address this issue. Most of them [49, 17, 35, 59, 20] erased the most discriminative region on the input image or feature map by zeroing that region during the training phase. These techniques are similar to the dropout [38] in that they deactivate specific nodes of the feature map by setting them zero during the training phase. This prevents the model from relying solely on the most discriminative part for classification, instead encourages it to learn the less discriminative part as well. To achieve this goal, Hide-and-Seek (HaS) [35] divides the input image into grid-like patches and randomly selects the patches to erase. While the random selection is simple and fast, it cannot effectively erase the most discriminative part.

For effectively removing only the most discriminative part, several techniques [49, 17, 59, 20] have been proposed. These techniques re-train the model multiple times [49, 17], use additional classifiers [17, 59], or perform two forward-backward propagations per one iteration [20] for finding the most discriminative part. Consequently, huge additional computing resources are required to eliminate the most discriminative part effectively.

From previous methods, we conclude that the idea of erasing only the most discriminative part is effective to capture the full extent of object. However, existing methods require substantial computing resources to remove the most discriminative part accurately. Our goal is to erase the most discriminative part in an effective and efficient way. To this end, we propose an Attention-based Dropout Layer

*Corresponding author.

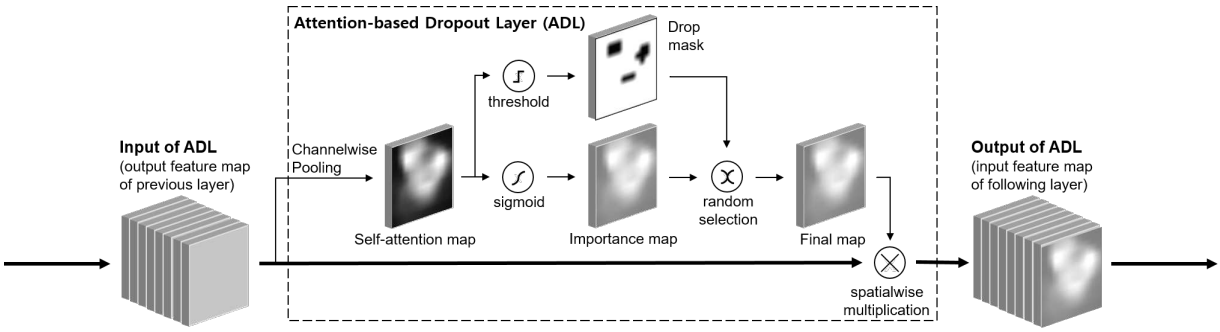

Figure 1. ADL block diagram. The self-attention map is generated by channelwise average pooling of the input feature map. Based on the self-attention map, we produce a drop mask using thresholding and an importance map using a sigmoid activation, respectively. The drop mask and the importance map are selected stochastically at each iteration and applied to the input feature map. Please note that this figure illustrates the case when the importance map is selected.

(ADL), a lightweight yet powerful method which utilizes self-attention mechanism to remove the most discriminative part of the target object.

Specifically, a self-attention map is obtained by performing channelwise average pooling on the input feature map. Based on the self-attention map, we produce two key components of ADL, a *drop mask* and an *importance map*. The drop mask is used to hide the most discriminative part during training. This induces the model to learn the less discriminative part as well. We obtain this drop mask by thresholding the self-attention map. The importance map is used to highlight informative region for improving the classification power of the model. Owing to the importance map, the more accurate self-attention map can be produced. The importance map is computed by applying sigmoid activation to the self-attention map. During training, either one of the drop mask or importance map is stochastically selected at each iteration, and then the selected one is applied to the input feature map by spatialwise multiplication. Figure 1 shows the block diagram of the proposed method.

Compared to existing WSOL techniques, the proposed method is much more efficient in terms of both computation and parameter overheads. This is because we can find and erase the most discriminative region by a single forward-backward propagation in a single model. In addition, regardless of the model architecture, ADL can be easily applied to convolutional feature maps of the model to improve the localization accuracy. Compared to existing self-attention techniques [46, 12, 26, 53], the proposed method is greatly lightweight because there are no additional trainable parameters for extracting self-attention map.

The proposed method is lightweight and efficient, and also report excellent accuracy. Quantitatively, the proposed method achieves superior accuracy, more than 15 percentage points of accuracy improvement, over the existing state-of-the-art techniques [59, 60] on CUB-200-2011 dataset [44], and comparable accuracy to the current state-of-the-art technique [60] on ImageNet-1k dataset [30]. We also observe consistent results in qualitative evaluation; the model with ADL learns the less discriminative part better than the vanilla model [63].

2. Related Work

Dropout. Dropout [38] is a regularization technique to alleviate overfitting in neural networks. Specifically, dropout discards information by randomly zeroing each hidden node of the neural network during the training phase. In this way, the network can enjoy the ensemble effect of small subnetworks, thus achieving a good regularization effect. However, unlike fully connected layers, applying dropout to the convolutional feature map is not effective. One of the reasons is that spatially adjacent pixels are strongly correlated on the convolutional feature map; they share redundant contextual information. Hence, the conventional pixel-based dropout cannot completely discard the information on the convolutional feature map [42].

In order to apply dropout to the convolutional feature map, Tompson *et al.* [42] proposed SpatialDropout that randomly drops partial channels of a feature map, rather than dropping each pixel. Based on this channel-based dropout, the problem of pixel-level dropout can be resolved. The proposed method differs from SpatialDropout in that we drop only strongly activated regions, rather than dropping entire region of channel. Owing to this region-based dropout, we could also bypass the problem of pixel-level dropout. Meanwhile, Park and Kwak [27] proposed MaxDrop, which drops the maximally activated pixel through channelwise or spatialwise on the feature map. Similar to MaxDrop, the proposed method drops strongly activated part. However, we differ from MaxDrop in that we use attention mechanism to find the maximally activated part. In addition, the proposed method does not drop the maximally activated *pixel*, but maximally activated *region*.

Attention mechanism. Humans selectively use an important part of the data to make a decision [3, 13]. Similarly, when a query comes in, the artificial model does not process all the data equally, but focuses only on the important data. This process is called *attention mechanism* and is actively used in various fields such as machine translation [43], image captioning [55], image inpainting [56, 23], transfer learning [57], visual question answering [64], and generative model [28, 60]. When the query is input itself, such attention is specifically called *self-attention*, which is effective to learn the meaningful representation for conducting the given task. For example, in the case of the classification task, the self-attention map appears in a form that emphasizes informative features for classification (*e.g.*, the most discriminative part of the target object).

Recently, various methods [46, 47, 12, 26, 53] utilize the self-attention mechanism to enhance the accuracy of the CNN classification model. Residual Attention Networks (RAN) [46] has improved the accuracy of the classification model using 3D self-attention map. However, the parameter overheads are very large because the raw feature map without any compression is used for attention extraction. Squeeze-and-Excitation (SE) [12] increases the accuracy of the classification model using only 1D channel self-attention map. For extracting the self-attention map, the feature map is first compressed using Global Average Pooling (GAP) and then passed through 2-layer MLP. In this way, SE can significantly reduce the parameter overheads for attention extraction compared to RAN. However, the parameter overheads are still not negligible (*e.g.*, 10% on ResNet50 [10]).

Bottleneck Attention Module (BAM) [26] and Convolutional Block Attention Module (CBAM) [53] increase the accuracy of the classifier by utilizing both 1D channel and 2D spatial self-attention maps. They compute the spatial self-attention map using auxiliary convolutional layer(s). The computed self-attention map is applied to the input feature map for rewarding the informative region. Likewise, the proposed method uses the *importance map* for rewarding the informative region. However, the key difference from them is that we stochastically penalize that region using the *drop mask*. Also, unlike these techniques, we do not require additional trainable parameters for extracting the self-attention map.

3. ADL: Attention-based Dropout Layer

In this section, we present details of the proposed method, Attention-based Dropout Layer (ADL). ADL is applied on each feature map of classification model, and induces the model to learn the entire region of the object. ADL generates a self-attention map from input feature map, and produces a drop mask and an importance map. Although both components are computed from self-attention map, they play the opposite role. The drop mask penalizes the most discriminative part for inducing the model to cover the integral extent of the object. Meanwhile, the importance map rewards the most discriminative part for increasing the classification power of the model. During training, the drop mask or importance map is stochastically selected for each iteration. Then, the selected one is applied to the input feature map. By applying each component stochastically, we can enjoy their advantages simultaneously. ADL has two main hyperparameters: *drop_rate* and γ. The *drop_rate* indicates how frequently the drop mask is applied, and the γ controls the size of the region to be dropped. The example of each component is visualized in Figure 2.

Specifically, the input of ADL is a convolutional feature map $\mathbf{F} \subseteq \mathbf{R}^{H \times W \times C}$. Note that C is the number of channels, H and W are height and width, respectively. For simplicity, we omit the mini-batch dimension in this notation. We generate a self-attention map $\mathbf{M}_{att} \subseteq \mathbf{R}^{H \times W}$ by compressing $\mathbf{F}$ using channelwise average pooling. Because the model is trained for classification, the intensity of each pixel in the self-attention map is proportional to the discriminative power. In this way, we can approximate the spatial distribution of the most discriminative part efficiently.

To obtain the drop mask, we first set a drop threshold by prefixed ratio γ of maximum intensity of the self-attention map. Then, we produce the drop mask $\mathbf{M}_{drop} \subseteq \mathbf{R}^{H \times W}$ by setting each pixel to 0 if it is larger than drop threshold, and 1 if it is smaller. That is, the drop mask has 0 for the most discriminative region and 1 for otherwise. Note that the size of region to be dropped increases as γ decreases, and vice versa. The drop mask is applied to the input feature map by spatialwise multiplication. In this way, we can hide the most discriminative part from the model; we encourage the model to learn the less discriminative part for classification but meaningful region for localization. However, if the drop mask is applied at every iteration, the most discriminative part is never observed during the training phase. As a result, the classification accuracy of the model is significantly decreased, which adversely affects the localization accuracy. To remedy this, we stochastically apply the drop mask according to *drop_rate*. When the drop mask is not applied, the importance map is applied instead. We generate the importance map $\mathbf{M}_{imp} \subseteq \mathbf{R}^{H \times W}$ from the self-attention map by applying sigmoid activation. That is, the intensity of each pixel in the importance map is close to 1 for the most discriminative region, and close to 0 for the less discriminative region. Like the drop mask, the importance map is applied to the input feature map by spatialwise multiplication. In this way, we can improve the classification accuracy of the model.

The proposed method is applied independently to each convolutional feature map. Therefore, it can be easily plugged into multiple feature maps of existing classification

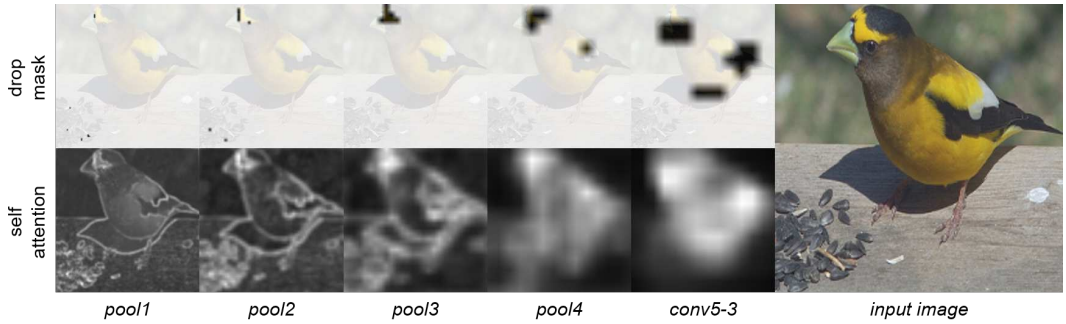

Figure 2. Drop mask and self-attention map at each layer of VGG-GAP [63]. At lower-level layers, the self-attention maps include general features, while class-specific features are included in the self-attention maps at higher-level layers. The drop masks also erase most discriminative part more effectively at higher-level layers. Please note that the drop mask is overlaid with input image for better visualization. Because the importance map has a distribution very similar to that of the self-attention map, we do not visualize it.

models for improving localization accuracy. In addition, it does not require any trainable parameters. That means, there is no parameter overheads even when applied to multiple feature maps at the same time. Furthermore, with ADL, the most discriminative region can be identified and erased efficiently, without auxiliary classifiers [17, 59], re-training [49], or additional forward-backward propagation [20].

ADL is an auxiliary module which is applied only during training. During the testing phase, ADL is deactivated. That is, our testing phase is identical to that of vanilla model. Therefore, the object localization can be performed using various heatmap extraction methods [63, 31, 59, 20] without bells and whistles. Note that we do not compensate the different distributions between training and testing, as other dropout-based WSOL techniques [17, 35, 59].

Relation with other attention extraction methods. Our extraction method does not require trainable parameters, much lightweight compared to existing methods [46, 12, 26, 53]. Thus, one might wonder how our method can produce semantically meaningful results despite its simplicity.

Recently, Zagorukyo and Komodakis [57] showed that the informative region for transfer learning can be identified by applying the channelwise pooling to the convolutional feature map. That is, the self-attention map for transfer learning is obtained by the channelwise pooling. Inspired by this, CBAM [53] utilized the self-attention map to improve the classification accuracy. Specifically, they refine the map using auxiliary convolutional layer and sigmoid activation. This refined self-attention map is applied to input feature map by spatialwise multiplication. In this way, the auxiliary convolutional layers are trained to refine the self-attention map for improving the classification accuracy.

However, from the empirical study, we observe that the self-attention map may not need to be refined by auxiliary layers. We conjecture that it is because existing convolutional layers in CNN model are sufficiently powerful to produce meaningful self-attention map. Hence, after computing the self-attention map by channelwise average pooling, we normalize this map using sigmoid activation and then multiply it to input feature map. Then, the gradient from the loss function updates existing convolutional layers so that the resultant self-attention map is useful for improving classification accuracy. For example, if the self-attention map fails to highlight the object region, this may degrade the classification accuracy. Hence, existing convolutional layers are trained to produce more accurate self-attention map. This is equivalent to assigning the role of the auxiliary convolutional layer used in CBAM to the existing convolutional layers in the model. Note that the similar principle was introduced by Lin *et al.* [22]; they replace the fully connected layers of CNN classifier with the GAP layer.

The improvement of classification accuracy of our attention method may not be as great as that of CBAM. However, our method is much more efficient and can produce sufficiently meaningful results for our application. This is shown in our experimental results; our self-attention map is effective to increase classification accuracy and identify the most discriminative part of the target object.

Relation between drop mask and importance map. In our model, the drop mask penalizes the most discriminative part, while the importance map rewards the most discriminative part. One might consider the drop mask and importance map are mutually exclusive. However, our experimental results support that they are not mutually exclusive. We believe that it is because the drop mask can be accurately produced by the importance map. Specifically, as the importance map improves the classification accuracy, the more accurate self-attention map can be produced. Consequently, the drop mask can more effectively erase the most discriminative region of the object.

Relation between classification and localization. Previous study [35] has reported that the classification accuracy is compromised while the localization accuracy is increased. They conjecture that this is caused by the usage of a drop mask. Because we also use a drop mask to erase the most discriminative part, such a trade-off relationship

Figure 3. Qualitative evaluation results of VGG-GAP [63] on CUB-200-2011 and ImageNet-1k. The left image in each figure is input image. The red bounding box is ground truth, while the green bounding box is estimates. The middle image is heatmap and the right image shows the overlap between the input image and the heatmap. We also compared our method and the vanilla model side by side.

between the accuracy of localization and that of classification is consistently observed in our experiments. However, the proposed method can boost the classification power using the importance map, thus the accuracy degradation of classification is not as significant as other techniques.

Relation with the current state-of-the-arts. The current state-of-the-art techniques for WSOL are Adversarial Complementary Learning (ACoL) [59] and Self-Produced Guidance (SPG) [60]. ACoL adds two auxiliary classifiers in parallel to the backbone feature extractor for finding the most discriminative part of the target object. The proposed method differs from ACoL in that we can find the most discriminative part without the additional classifier, which is much more efficient. Most recently, SPG has been proposed, a new WSOL technique that utilizes spatial distribution of the object and background. The classifier can learn the integral extent of the object using that distribution as auxiliary supervision. The proposed method differs from SPG in that SPG does not erase the most discriminative part of the object. In addition, SPG requires substantial computing resources for improving the localization accuracy.

4. Experimental Results

Dataset. We evaluate the performance of the proposed method in CUB-200-2011 [44] and ImageNet-1k [30], respectively. The ImageNet-1k is a large-scale dataset with 1,000 different classes, consisting of approximately 1.3 million training images and 50,000 validation images. For this dataset, we train the model with the training set and evaluate the performance with the validation set.

The CUB-200-2011 includes 200 species of birds, consisting of 5,994 training images and 5,794 testing images. For this dataset, we train the model with the training set and evaluate the performance with the testing set. The intra-class variation of CUB-200-2011 is smaller than that of ImageNet-1k, because all classes of this dataset belong to *birds*. In this case, the extent of the most discriminative region might be quite small. For example, in *Common Raven* and *White-necked Raven*, there is no difference in appearance except the color of the neck. That is, the most discriminative part is the neck, which is very small compared to the entire area of the bird. Consequently, although CUB-200-2011 is not a large-scale dataset such as ImageNet-1k, this is a particularly challenging dataset to conduct WSOL.

Implementation details. We use VGG [34], ResNet [10], MobileNetV1 [11], and InceptionV3 [40] as backbone networks. Note that we replace the last pooling layer and two fully connected layers of VGG16 with a GAP layer, according to [63]. We also use the customized InceptionV3 as a backbone, following the SPG [60]. We plug SE block [12] into ResNet50 for demonstrating the compatibility of ADL with other self-attention methods. For ResNet and MobileNetV1, we set the stride of last strided convolution to 1 for enlarging the spatial resolution of heatmap to 14×14.

ADL is plugged in each feature map of the CNN model in a sequential way; the output of ADL is the input of the next layer. We use a pre-trained model which is trained with ImageNet-1k dataset [30], and then fine-tune the network. We extract the heatmap from classification model using CAM [63]. Also, the bounding box is extracted from the heatmap using the same method as presented in [63]. We implement the models using Tensorpack [54] on Tensorflow [1], and train them using NVIDIA Titan Xp GPU.

Based on extensive ablation studies, we find that it is optimal to apply ADL to intermediate and higher-level layers of the network. Especially, for the intermediate layer, it is preferable to apply it to bottleneck part (*e.g.*, pooling layer or strided convolution). We set the *drop_rate* as 75%. For the drop threshold, we set γ to 80% for VGG-GAP and InceptionV3, 90% for ResNet, and 95% for MobileNetV1. However, the hyperparameters mentioned here are only the recommended settings. Note that the localization accuracy can be further improved when the optimal setting is used.

Metrics. We use three evaluation metrics as [35]: Top-1 classification accuracy (*Top-1 Clas*), Localization accuracy with known ground-truth class (*GT-known Loc*), and Top-1 localization accuracy (*Top-1 Loc*). *Top-1 Clas* determines that the answer is correct when the estimated class is equal to the ground truth class. *GT-known Loc* judges the answer as correct when the intersection over union (IoU) between the ground truth bounding box and estimated box for the ground truth class is 50% or more. Lastly, *Top-1 Loc* considers the answer as correct when both *Top-1 Clas* and *GT-*

Drop mask (%)	Importance map (%)	GT-known Acc (%)	Top-1 Clas (%)	Top-1 Loc (%)
100	0	72.43	57.37	44.11
75	25	**74.78**	62.25	**49.69**
50	50	71.51	64.93	49.33
25	75	67.29	*68.99*	47.98
0	100	47.51	67.78	32.24
N/A	N/A	51.09	67.55	34.41
75	N/A	73.23	61.55	47.67
N/A	25	50.62	*68.50*	33.91
75	25	**74.78**	62.25	**49.69**

Table 1. Upper: Accuracy according to *drop_rate*. Middle: Baseline accuracy. Lower: Accuracy when each component has been deactivated. Bold text refers the best localization accuracy, while *italic text* refers the best classification accuracy. N/A indicates that ADL outputs the raw input feature map instead of applying drop mask or importance map.

known Loc are correct. Please note that it is considered to be the most appropriate to use *Top-1 Loc* for evaluating overall localization performance, according to [30].

4.1. Ablation Study

In this subsection, we utilize pre-trained VGG-GAP [34, 63] as a backbone network. For training, we plug ADLs in all the pooling layers and the *conv5-3* layer, and then finetune the model using CUB-200-2011 dataset.

First, we visualize the self-attention map and drop mask in Figure 2. We observe that the self-attention maps of lower-level layers (*i.e.*, *pool1* and *pool2*) contain class-agnostic general features. Meanwhile, the self-attention maps of higher-level layers (*i.e.*, *pool4* and *conv5-3*) contain the class-specific features. We also observe that the drop masks from higher-level layers erase the most discriminative part more accurately than those from lower-level layers.

Next, we investigate the effect of *drop_rate* on accuracy. The upper part of Table 1 reports the results. From these results, we observe that the best localization accuracy can be achieved when the *drop_rate* is 75%. Meanwhile, when the drop mask is applied at every iteration (*i.e.*, *drop_rate* 100%), the classification (*Top-1 Clas*) and localization (*Top-1 Loc*) accuracy are greatly reduced. This is because, as mentioned in Section 3, the model never observe the most discriminative part. As a result, the classification power of the model decreases significantly, which adversely influences localization accuracy. Given that accuracy degradation in *GT-known Acc* is relatively less than that of *Top-1 Loc* and that of *Top-1 Clas*, we can conclude that this is the result of the classification accuracy degradation.

We observe that the classification accuracy increases as the *drop_rate* decreases. However, when the *drop_rate* becomes too low (*drop_rate* from 25% to 0%), the classification accuracy decreases again (from 68.99% to 67.78%). We believe this is caused by overfitting. The drop mask is a dropout-based technique and its rationale is similar to MaxDrop [27]. Thus, the drop mask with proper *drop_rate* may prevent overfitting, increasing the classification accuracy. We consider that the analysis of regularization effect of the drop mask is beyond the scope of this paper. Yet, we plan to analyze this rigorously in future work.

Applied feature map	GT-Known Acc (%)	Top-1 Clas (%)	Top-1 Loc (%)
N/A	51.09	67.55	34.41
conv 5-3	57.99	*68.95*	41.73
+ *pool4*	68.22	67.17	48.02
+ *pool3*	**75.41**	65.27	**52.36**
+ *pool2*	71.85	63.76	48.46
+ *pool1*	74.78	62.25	49.69

Table 2. Effects in accuracy upon the choice of the feature maps to employ ADL. Bold text refers the best localization accuracy, while *italic text* refers the best classification accuracy.

Third, we observe the effect of each component on the accuracy by deactivating the importance map or drop mask, respectively. The lower part of Table 1 summarizes the experimental results. From this, we can confirm that applying the drop mask and the importance map at the same time has better localization accuracy than applying only one of them. This supports the argument that the drop mask and importance map are not mutually exclusive.

When the importance map is applied alone, the classification accuracy increases but the localization accuracy decreases. We believe that this is because the classifier focuses more on the most discriminative part, guided by the importance map. This result supports our argument that the proposed lightweight attention method is effective to improve the classification accuracy. On the other hand, when drop mask is applied alone, the localization accuracy increases but the classification accuracy decreases. We believe that this is because the model utilizes less discriminative parts for classification, guided by the drop mask. These results also support the observation that the accuracy of localization and classification are in a trade-off relationship when applying the drop mask [35].

Lastly, we investigate the effects in accuracy upon the choice of feature maps where ADLs are employed and report the results in Table 2. From these results, we can see that applying ADLs to additional convolutional feature maps further increases the localization accuracy. We find that the ADL can improve both localization and classification accuracy. However, the best localization accuracy can be achieved by sacrificing the classification accuracy. In addition, when the ADLs are applied to lower-level fea-

Method	Backbone	# of Params (Mb)	Overheads		CUB-200-2011		ImageNet-1k	
			parameter (%)	computation (%)	Top-1 Loc (%)	Top-1 Clas (%)	Top-1 Loc (%)	Top-1 Clas (%)
CAM	VGG-GAP [34, 63]	78	0	0	34.41	67.55	42.80*	66.60*
ACoL	VGG-GAP [34, 63]	181	132.05	37.63	45.92*	71.90*	**45.83***	67.50*
ADL	VGG-GAP [34, 63]	78	0	0.00	**52.36**	65.27	44.92	69.48
CAM	MobileNetV1 [11]	16	0	0	43.70	71.94	41.66	68.38
HaS-32	MobileNetV1 [11]	16	0	0	44.67	66.64	41.87	67.48
ADL	MobileNetV1 [11]	16	0	0.00	**47.74**	70.43	**43.01**	67.77
CAM	ResNet50-SE [10, 12]	107	0	0	42.72	80.65	46.19	76.56
ADL	ResNet50-SE [10, 12]	107	0	0.00	**62.29**	80.34	**48.53**	75.85
CAM	InceptionV3 [40, 60]	101	0	0	43.67*	-	46.29*	-
SPG	InceptionV3 [40, 60]	146	44.55	30.05	46.64*	-	**48.60***	-
ADL	InceptionV3 [40, 60]	101	0	0.00	**53.04**	74.55	**48.71**	72.83

Table 3. Quantitative evaluation results on CUB-200-2011 and ImageNet-1k. Bold text refers the best localization accuracy for each backbone network. We also underline the best score in each dataset. Overheads are computed based upon their backbone networks. The accuracy with asterisk* indicates that the score is from the original paper. We leave some *Top-1 Clas* scores blank, because they are not reported in the original paper [60]. For reproducing baseline methods, we use hyperparameters suggested by their original papers [63, 35]. Also, we train and test HaS and ADL under the same setting for a fair comparison.

ture maps such as *pool2* and *pool1*, the localization accuracy rather decreases. We believe that this is because the lower-level feature maps include general features that are not related to the target class. Consequently, the most discriminative part cannot be effectively eliminated in lower-level feature maps using ADL.

4.2. Comparison with State-of-the-art Methods

We compare the proposed method with various recent WSOL techniques including the state-of-the-art: CAM [63], HaS [35], ACoL [59], SPG [60]. We report the accuracy of ACoL and SPG from their original paper. Meanwhile, we train the backbone networks using the same preprocessing method used in ACoL and SPG. Then, HaS or ADL are applied on the backbone networks. Considering the accuracy of vanilla model as baseline, we evaluate the accuracy gain of HaS and ADL, respectively. Please note that ACoL and SPG are the current state-of-the-art techniques for WSOL. In addition, among the techniques without parameter overheads, HaS performs the best.

Figure 3 visualizes the localization results on CUB-200-2011 and ImageNet-1k dataset for qualitative evaluation. From the results, we consistently observe that model with ADL captures the less discriminative parts better than vanilla model. For example, as seen from the left-most sample in the Figure 3, the heatmap and bounding box extracted from vanilla model only highlight the face of birds. Contrarily, the model with ADL covers not only the face, but also the entire part of the bird, from head to wing. In addition, from the right-most sample in the Figure 3, the vanilla model focuses only on the cylinder of *revolver*, whereas the model with ADL localizes the entire frame of the *revolver*.

Next, the quantitative evaluation results on CUB-200-2011 and ImageNet-1k datasets are summarized in Table 3. To compare the computing resources required by each technique, we have described the number of parameters and both computation and parameter overheads along with *Top-1 Loc* and *Top-1 Clas*. ADL has no parameter overheads, and the computation overheads are nearly zero (*e.g.*, 0.003% in ResNet50-SE) upon the backbone network. The proposed method is much more efficient than the existing state-of-the-art techniques, ACoL and SPG, in terms of both parameter and computation overheads.

We push further to maximize the efficiency of WSOL by employing MobileNetV1 [11] as a backbone network. Due to the lightweight nature of MobileNetV1, it is inappropriate to employ ACoL or SPG which requires huge additional computing resources. On the other hand, ADL and HaS can be successfully employed despite a limited amount of resources. From the experimental results, we can observe that the accuracy gain of the proposed method is better than that of HaS. In addition, HaS has reduced classification accuracy against the baseline. This is caused by the trade-off relationship between localization and classification accuracy discussed in Section 3. Fortunately, the importance map of ADL can subside such a drawback by increasing the classification power. Consequently, the classification accuracy degradation of the ADL is not as significant as that of HaS.

In addition to its high efficiency, the proposed method achieves a new state-of-the-art localization accuracy on CUB-200-2011 dataset. When ResNet50-SE is employed as a backbone, the proposed method improves the localization accuracy by more than 15 percentage points over the state-of-the-art accuracy [59, 60]. Please note that the

number of parameters of ResNet50-SE with ADL is much smaller than that of ACoL and SPG. This achievement is quite impressive, considering that recent techniques are competing with the accuracy by 2-3 percentage points difference. Also, when the other three backbone networks are employed, the proposed method achieves better localization accuracy than the existing state-of-the-art techniques.

In the ImageNet-1k experiments, when VGG-GAP is used as a backbone, the accuracy of ADL is better than that of CAM, but slightly lower than that of ACoL. However, when ResNet50-SE is used as a backbone, localization accuracy of ADL is better than that of ACoL and comparable with that of SPG even though the required computing resources are much lower. In addition, when InceptionV3 is used as a backbone, comparable accuracy (0.11 percentage point difference) to SPG is achieved. In summary, we achieve new state-of-the-art accuracy on CUB-200-2011 dataset; on ImageNet-1k dataset, ADL achieves comparable accuracy with the current state-of-the-art technique [60] despite its superior efficiency.

Discussion. We verified the proposed method on a single-object detection task, following the current state-of-the-art methods [59, 60]. However, it should be noted that the proposed method can be also used to improve the weakly supervised semantic segmentation accuracy. The classifier with ADL is the same as its vanilla version during testing, thus it can be easily combined with the weakly supervised semantic segmentation framework, such as [18, 24].

Next, to analyze the substantial difference in our accuracy gain between two datasets, we investigate our failure examples from ImageNet-1k experiments. From the failure case, we observe that the classifier extracts the discriminative features from the background which appears frequently with the target object. Figure 4 illustrates such examples. In the case of the *snowmobile* class, the target object often co-occurs with *snow*. The vanilla model only focuses on the *snowmobile*, while the model with ADL learns not only the *snowmobile*, but also the *snow* and *tree*. This is because the background frequently appearing with the object might be the less discriminative region.

ImageNet-1k includes a wide variety of classes where specific types of background co-occur with the target object. In this case, the background has a certain level of discriminative power. Therefore, the model is likely to learn the background features when the most discriminative part is dropped. Meanwhile, since all classes of CUB-200-2011 belong to birds, similar backgrounds appear regardless of the classes (*e.g.* sky, tree). In other words, the background of this dataset is nearly independent of classes, thus the background is not a discriminative region [61]. As a result, the model does not learn the features from the background although the most discriminative part is hidden.

(a) Input (b) CAM (c) ADL (Ours)

Figure 4. The failure case on ImageNet-1k experiments. The target class is *snowmobile*. The model with ADL learns the less discriminative region which is not included in object. Specifically, the model captures not only the *snowmobile*, but also *snow* and *tree*.

This explains the gap of our accuracy gain for two datasets; ADL has remarkable performance to induce the classifier to learn the less discriminative parts, as supported by CUB-200-2011 evaluations. We believe that this problem might be critical for all WSOL methods inducing the classifier to learn the less discriminative part. Currently, it seems non-trivial to solve this problem, thus we will address this issue in future work. Lastly, we note that the gap is not caused by the scale of dataset because ADL rarely fails for ImageNet-1k classes sharing similar background statistics (*e.g.*, various breeds of dogs).

5. Conclusion

We presented an *Attention-based Dropout Layer* (ADL), a novel weakly supervised object localization method that induces the CNN classifier to learn entire extent of the object. The proposed method is much more efficient and lightweight than existing state-of-the-art methods. In addition, the proposed method has achieved excellent performance; new state-of-the-art accuracy on CUB-200-2011, and comparable accuracy with current state-of-the-arts on ImageNet-1k. We also demonstrate that the proposed method can be easily applied to various CNN classifiers to improve the localization accuracy. For the future work, we will analyze the regularization effect of the *drop mask*. In addition, we will address the problem that the model learns the less discriminative region from outside of the object.

Acknowledgement

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2019R1A2C2006123), and the MIST (Ministry of Science and ICT), Korea, under the "ICT Consilience Creative Program" (IITP-2018-2017-0-01015) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). This work was also supported by ICT R&D program of MSIP/IITP [R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding].

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016.

[3] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201, 2002.

[4] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *CVPR*, pages 914–922, 2017.

[5] Xuanyi Dong, Deyu Meng, Fan Ma, and Yi Yang. A dual-network progressive approach to weakly supervised object detection. In *ACMMM*, pages 279–287, 2017.

[6] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):1–1, 2018.

[7] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, pages 642–651, 2017.

[8] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I Morariu, and Larry S Davis. C-WSL: Count-guided weakly supervised localization. In *ECCV*, pages 152–168, 2018.

[9] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[14] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, pages 1377–1385, 2017.

[15] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. ContextLocNet: Context-Aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365, 2016.

[16] Anna Khoreva, Rodrigo Benenson, Mohamed Omran, Matthias Hein, and Bernt Schiele. Weakly supervised object boundaries. In *CVPR*, pages 183–192, 2016.

[17] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-Phase learning for weakly supervised object localization. In *ICCV*, pages 3534–3543, 2017.

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, pages 695–711, 2016.

[19] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520, 2016.

[20] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, pages 9215–9223, 2018.

[21] Xiaodan Liang, Si Liu, Yunchao Wei, Luoqi Liu, Liang Lin, and Shuicheng Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, pages 999–1007, 2015.

[22] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICLR*, 2014.

[23] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *ECCV*, pages 85–100, 2018.

[24] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, pages 5038–5047, 2017.

[25] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - Weakly-supervised learning with convolutional neural networks. In *CVPR*, pages 685–694, 2015.

[26] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. BAM: Bottleneck Attention Module. In *BMVC*, 2018.

[27] Sungheon Park and Nojun Kwak. Analysis on the dropout effect in convolutional neural networks. In *ACCV*, pages 189–204, 2016.

[28] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.

[29] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, pages 1796–1804, 2015.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[32] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *ICCV*, pages 3534–3543, 2017.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLRW*, 2014.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[35] Krishna Kumar Singh and Yong Jae Lee. Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, pages 3544–3553, 2017.

[36] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014.

[37] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, pages 1637–1645, 2014.

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[39] Chen Sun, Manohar Paluri, Ronan Collobert, Ram Nevatia, and Lubomir Bourdev. Pronet: Learning to propose object-specific boxes for cascaded neural networks. In *CVPR*, pages 3485–3493, 2016.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[41] Eu Wern Teh, Mrigank Rochan, and Yang Wang. Attention networks for weakly supervised object localization. In *BMVC*, 2016.

[42] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *CVPR*, pages 648–656, 2015.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[45] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *ECCV*, pages 431–445, 2014.

[46] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[48] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017.

[49] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, pages 1568–1576, 2017.

[50] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2314–2320, 2017.

[51] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV*, pages 454–470, 2018.

[52] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, pages 7268–7277, 2018.

[53] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *ECCV*, pages 3–19, 2018.

[54] Yuxin Wu et al. Tensorpack. `https://github.com/tensorpack/`, 2016.

[55] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[56] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018.

[57] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[58] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[59] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, pages 1325–1334, 2018.

[60] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *ECCV*, pages 597–613, 2018.

[61] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, pages 1134–1142, 2016.

[62] Xiaopeng Zhang, Yang Yang, and Jiashi Feng. Ml-locnet: Improving object localization with multi-view learning network. In *ECCV*, pages 240–255, 2018.

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[64] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016.

[65] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *ICCV*, pages 1841–1850, 2017.