Expanding CAM inWeakly Supervised Object Localization

By Rahul Kushwaha

WORD COUNT 5702

TIME SUBMITTED 08-JUL-2022 06:14PM

PAPER ID 88203326

Expanding CAM in Weakly Supervised Object Localization

Dissertation submitted in partial fulfilment for the award of the degree

Master of Technology in Computer Science

by

Rahul Kushwaha

Roll No.: CS2021
M.Tech, 2nd year

Under the supervision of

Prof. Utpal Garain

Computer Vision & Pattern Recognition Unit
Computer and Communication Sciences Division
INDIAN STATISTICAL INSTITUTE

Surbhi Mathur

Senior Data Scientist
Flipkart Internet Private Limited,Bengaluru

July, 2022

CERTIFICATE

This is to certify that the work presented in this dissertation titled "Expanding CAM in Weakly Supervised Object Localization", submitted by Rahul Kushwaha, having the roll number CS2021, has been carried out under my supervision in partial fulfilment for the award of the degree of Master of Technology in Computer Science during the session 2021-22 in the Computer and Communication Sciences Division, Indian Statistical Institute.

Dr. Utpal Garain
Professor, Computer Vision Pattern Recognition Unit, Centre for Artificial Intelligence and Machine Learning
Indian Statistical Institute, Kolkata

Acknowledgements

First and foremost, I take this opportunity to express my sincere thankfulness and deep regard to *Prof. Utpal Garain and Surbhi Mathur*, for the impeccable guidance, nurturing and constant encouragement that they had provided me during my post-graduate studies. Words seem insufficient to utter my gratitude to him for his supervision in my dissertation work. Working under him was an extremely knowledgeable experience for a young researcher like me.

I also thank the CSSC and ISI Library for extending their supports in my different ways in my urgent need.

I shall forever remain indebted to my parents, teachers and friends for supporting me at every stage of my life. It is their constant encouragement and support that has helped me throughout my academic career and especially during the research work carried out in the last one year.

Date: 03-07-2022

Rahul Kushwaha
Roll No.: CS2021
M.Tech, 2nd year
Indian Statistical Institute

Abstract

The weakly supervised object localisation(WSOL) technique only uses image level labels without any bounding box or segmentation mask during training. Existing method have tried to cover only the most disriminative part of the object leading to low IoU and poor bounding boxes. This paper introduces method ex-CAM to expand Class Activation Maps(CAM) to cover the entire object in case of multi-labelled image dataset. The base model used gives M feature maps for each of the C classes which are then pooled to get the final output. We perform intra-class sequential dropping of important areas in the M feature maps. Since we are using pooling of feature maps to get the final output, in order to maximize the loss function the model will start expanding the feature maps which leads to better CAM. The model is trained for classification only but can perform classification and localization. We have tested ex-CAM on PASCAL VOC 2007 and PASCAL VOC 2012 and have got significant improvement over WILDCAT. In contrast to WILDCAT ex-CAM covers the entire object in most of the cases. Ex-CAM achieved a localisation mAP of 0.7 while the WILDCAT achieved a localisation mAP of 0.53 on PASCAL VOC 2012. On PASCAL VOC 2007 ex-CAM achieved a localisation mAP of 0.69 while the WILDCAT achieved a localisation mAP of 0.27. So there is a significant improvement over the base model. Thus ex-CAM technique of intra-class sequential drop-out may be considered as a way to expand CAM to get good localisation.

Contents

List of Figures

List of Tables

1 CHAPTER 1: Introduction

The fully supervised Object Detection models have given significant performance on Object localization task. The problem with these supervised techniques is that they require bounding box or segmentation map data which is not available in abundance. This problem gave rise to development of Weakly Supervised techniques which requires only image level labels.

The problem with most of the Weakly supervised models is that they tend to find most discriminative regions and do not cover the entire object. For this problem adversarial learning technique like ACol are used but the problem is that we do not know how many classifier are required to get good bounding boxes and too many classifier also increases the resource requirement. To solve these problems we propose a new architecture.

The target of ex-CAM is to cover the object as much as possible. The intra-class sequential dropping of important areas in M feature maps of each class leads to expansion of feature maps since we have used GAP and WILDCAT pooling. After getting CAM this new architecture which is explained later in detail we can use these maps for getting the bounding boxes.

Our Experiments show that the CAM obtained cover the entire object in most of the cases in PASCAL VOC 2012 and PASCAL VOC 2007 dataset.

The architecture has the following advantages-

- It causes expansion of the feature maps due to which the CAM covers the entire object in most of the cases.
- It can be used for multi label object detection as the loss function used is binary cross-entropy.
- Contrast to the "Learning Deep Features for Discriminative Localization"[3] paper we don't need to recompute the feature maps after forward pass using the weights of the final layer. A single pass is enough.
- The "Adversarial Complementary Learning for Weakly Supervised Object Localization"[4] paper uses multiple classifier and extracted feature maps from each of theses classifiers are subjected to GAP but we do not know how many classifier would

be enough. Moreover using multiple classifier will also lead to multiple forward passes.

- "Attention-based Dropout Layer for Weakly Supervised Object Localization" doesn't work well with multi label object localisation. The original paper uses a single label object localisation but when put to multi label object localisation the CAM are too big and in many cases cover the entire image.
- The same model can be used for classification and localisation.

2 CHAPTER 2: Related Works

With the development of deep learning there has been significant improvement in all machine learning sub domains. A similar trend has been seen in computer vision also. After the development of Alexnet there had been significant improvement in classification, localisation and segmentation task. Later on with the development of concept like skip connections we obtained even better feature descriptor like resnet. The new feature descriptors have outperformed the old HOG and SIFT feature descriptor. With the development of the concept of transfer learning we can use pretrained models after fine tuning to get excellent performance on classification, localisation and segmentation. But the problem with these feature descriptor is that they require bounding box and segmentation mask labels for training which is generally not available.

The above problems in the fully supervised approach has led to the development of other approaches like unsupervised, weakly supervised, semi-supervised learning. In weakly supervised approach we only have image level labels during training and we need to predict the bounding boxes and segmentation mask during testing.

RCNN[7] uses selective search for region proposal. For this we remove the last layer of alexnet(or any other backbone). This architecture outputs an embedding of the input image. These input images and image labels are used to train a class specific SVM. Now since RCNN takes a lot of time for object detection fast RCNN was proposed which does some computation sharing during forward pass. Fast RCNN[8] still used selective search for region proposal which is now the bottleneck. To remove this problem faster RCNN[9] used region proposal network for region proposal.

Zhou et al trains a CNN for classification using the given image labels. Now the CAM for a given class can be computed by weighted average for the feature maps in the second last layer. The weight corresponds to the weight between the neuron of the given class in the last layers and all neurons in the second last layer. For each class we would have a separate weight set. This method shows only the most discriminative regions in its CAM and thus gives poor bouding boxes.

Zhang et al trains two classifiers. From classifier A we get the class activation maps which are used to get probable locations of the object. These areas are blackened in the original image and this new image would be used to train classifier B. The CAM from classifier A and B would be merged by GAP to get the final CAM.

Durand et al[2] on the other hand use resnet(can be replaced by some other backbone) as base. The last layer corresponding to adaptive pooling is removed. Now this architecture outputs a hxhxK feature maps. This is subjected to a 1x1 convolution to get hxhxMC feature maps. Each of the M maps sets corresponds to a class.

These M maps of individual classes are subjected to GAP. Now we have hxhxC maps. These maps are then subjected to WILDCAT pooling. This architecture is the trained for classification. During testing we get the class labels from this classifier and by extracting the hxhxC feature maps we get the CAM for each class for object localisation. The method used multi label binary cross entropy loss which enables the architecture to perform multi label localisation.

Zang et al[5] used a pseudo supervised approach where we train a classifer for image labels. CAMs are extracted from these classifier to get the bounding boxes. These bounding boxes are used a pseudo annotations for training a bounding box regressor. The above regressor remove noise in the label and either expands or contracts the bounding boxes to better fit the object. After training we can use the classifier for getting the image labels and the regressor can be used for localisation.

Choe et al uses dropout map and importance map which are obtained using the individual feature maps. To get the importance maps the corresponding feature map is passed through sigmoid. To get the dropout map we do thresholding of the importance map by droping the important region. We randomly either drop the important areas using the dropout map or highlight the important region using importance map. Due to this the CAM expands and cover try to cover more regions of the object.

3 CHAPTER 3: Model architecture

3.1 Motivation

Most of the early Weakly Supervised Object Localisation models like "Learning Deep Features for Discriminative Localization" only localise the most discriminative parts of the object. This is due to the fact that the model has been trained for classification and for classification only the most discriminative parts are enough.

Thus to improve the localisation we can hide the area obtained by this classifier by making it black and train another classifier on these images. Now the CAM of these two classifiers are pooled together to get better CAM. This is the idea behind "Adversarial Complementary Learning for Weakly Supervised Object Localization"[4]. But the above process is iterative and we do not know how many time it should be repeated. Moreover the CAM obtained by many iterations do not cover the object fully and sometimes even gets confused due to removal of important regions. Another problem with this approach is that in order to localise we would have to do multiple forward passes. This increase the time complexity and also having many classifiers also increases the space complexity.

Another approach is to use Pseudo Supervised object localisation as given in "Rethinking the Route Towards Weakly Supervised Object Localization". This approach first trains a classifier. Extract bounding boxes using feature maps of this classifier. These bounding boxes can be further used as pseudo labels to train bounding box regressor. The purpose of having a regressor is to remove noise in bounding boxes. Removal of noise will lead to expansion or contraction of boxes to better cover the object. The problem with this approach is that it work only for single labelled images. This is because we have a common regressor for all classes in order to keep it class agnostic but in case of multi label object localisation we need to localise multiple object but the regressor give bounding box for only one object. Moreover there can be multiple object of same class in the image. To solve this problem we can increase output size from 4 to 4kC where k is the maximum possible number of object of a given class in an image and C is the number of classes. If we trained the regressor like this we loose the class agnostic idea and thus we don't get the desired result. The paper also provide solution for only for single labelled single object(one object in the entire image) localisation.

Recently there is another paper "Attention-based Dropout Layer for Weakly Supervised Object Localization" which uses dropout map and importance map which are obtained using the individual feature maps. We randomly either drop the important areas using the dropout map or highlight the important region using importance map. Due to this the CAM expands and cover try to cover more regions of the object. This model also has the same problem. The author has tested the model for single label object localisation. When this idea is applied to multi labelled object localisation the CAMs expands too much. The CAMs of multiple object merge together and lead to poor CAMs.

In order to solve the problems in the above model and make a model which is applicable to multi label multi object localisation task we proposed the ex-CAM model.

3.2 Brief

The model uses WILDCAT as base in which the obtained M feature maps for C classes i.e.MC feature maps are subjected to a class wise sequential dropout. The sequential dropping of important areas from the previous map in the current map will lead to expansion of the previous feature map as we have used GAP and WILDCAT pooling so in order to minimize the loss function the first feature map should show more important region.

3.3 Detailed

The model receives K maps from a pretrained network(say resnet). The K maps are subjected to 1x1 convolution. So now we have the M feature maps corresponding to C classes i.e. a total of MC maps. For each of the M feature map set we do the following-

- From the first map $F_1 \in R^{hxh}$ we find dropout mask($M_1 \in \{0,1\}^{hxh}$) but passing F_1 through sigmoid and thresholding by 0.5(or a learnable parameter $\delta \in (0,1)$).
- To make only the the important part visible in this map we do Hadamard product of 1-$M_1 \in \{0,1\}^{hxh}$ with this feature map

  $$F_1 := (1 - M_1) \odot F_1$$

  . This is the new feature map.
- From the second map $F_2 \in R^{hxh}$ we find dropout mask($M_2 \in \{0,1\}^{hxh}$) but passing it through sigmoid and thresholding by 0.5(or a learnable parameter $\delta \in (0,1)$). To drop important part visible in the previous map and in the current map we do the following

  $$F_2 := M_1 \odot M_2 \odot F_2$$
- From the third map $F_3 \in R^{hxh}$ we find dropout mask($M_3 \in \{0,1\}^{hxh}$) but passing it through sigmoid and thresholding by 0.5(or a learnable parameter $\delta \in (0,1)$. To drop important part visible in the previous maps and in the current map we do the following

  $$F_3 := M_1 \odot M_2 \odot M_3 \odot F_3$$

  .
- We continue this way till all M feature maps are done or till the point we have dropped the entire map region. We do this for all C classes.

After getting these new MC feature map we do GAP of each of the M map corresponding to each class to obtain C maps. Now theses C maps are subjected to WILDCAT pooling. In WILDCAT pooling of a feature map we take the average of top k+ elements and add it to the product of alpha(which a hyperparameter) and average of bottom k- elements. For WILDCAT pooling of feature map F let us suppose we have a list of sorted element. If F_i represent the i^{th} element of this sorted list then the output of

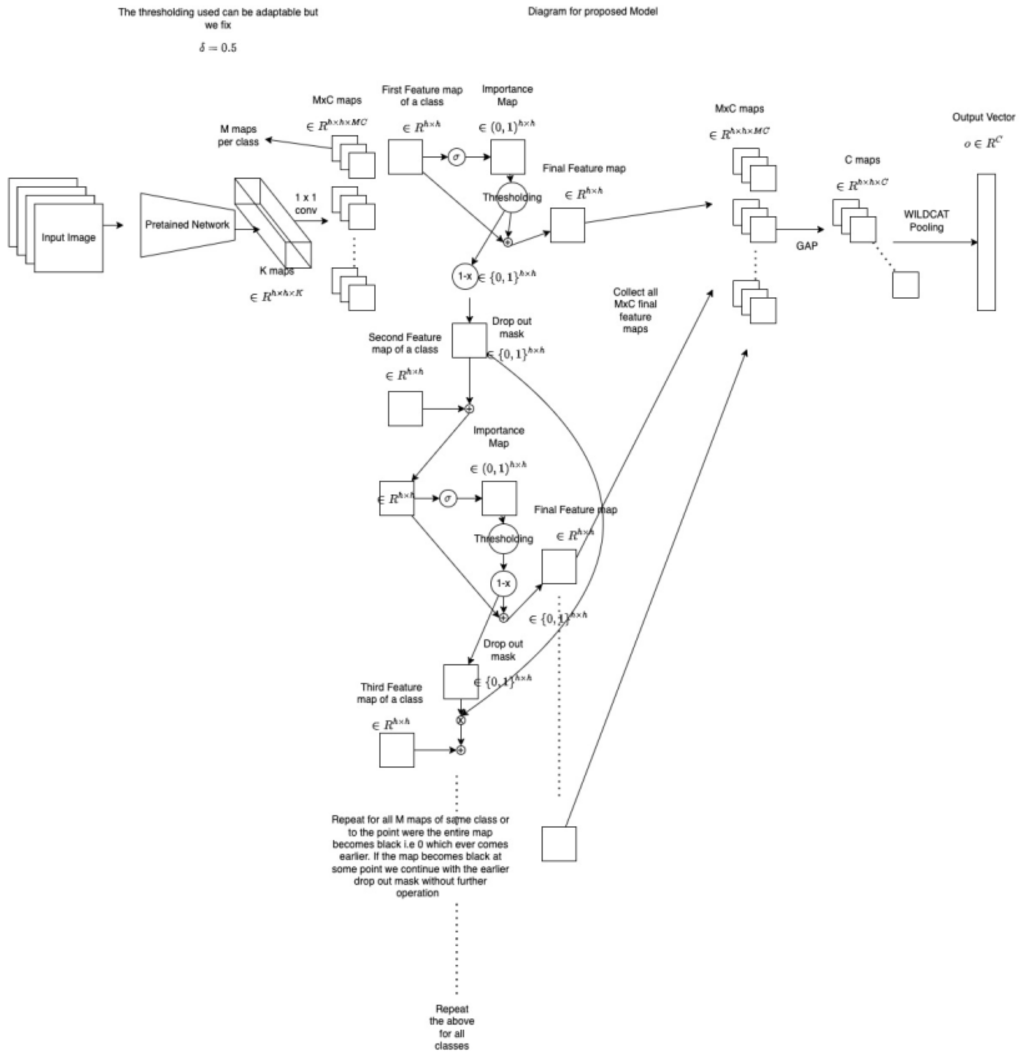

The thresholding used can be adaptable but we fix
$\delta = 0.5$
Diagram for proposed Model
Input Image
Pretained Network
K maps
$\in R^{h\times h\times K}$
1 x 1 conv
MxC maps
$\in R^{h\times h\times MC}$
M maps per class
First Feature map of a class
$\in R^{h\times h}$
σ
Importance Map
$\in (0,1)^{h\times h}$
Thresholding
Final Feature map
$\in R^{h\times h}$
1-x
$\in \{0,1\}^{h\times h}$
Drop out mask
$\in \{0,1\}^{h\times h}$
Second Feature map of a class
$\in R^{h\times h}$
Importance Map
$\in (0,1)^{h\times h}$
$\in R^{h\times h}$
σ
Thresholding
Final Feature map
$\in R^{h\times h}$
1-x
$\in \{0,1\}^{h\times h}$
Drop out mask
$\in \{0,1\}^{h\times h}$
Third Feature map of a class
$\in R^{h\times h}$
Repeat for all M maps of same class or to the point were the entire map becomes black i.e 0 which ever comes earlier. If the map becomes black at some point we continue with the earlier drop out mask without further operation
Repeat the above for all classes
Collect all MxC final feature maps
MxC maps
$\in R^{h\times h\times MC}$
GAP
C maps
$\in R^{h\times h\times C}$
WILDCAT Pooling
Output Vector
$o \in R^C$

Figure 1: Model architecture

WILDCAT pooling is given by -

$$o = \frac{\sum_{i \in \{1,2...k+\}} F_i}{k+} + \alpha \cdot \frac{\sum_{i \in \{(h^2-k-),...h^2\}} F_i}{k-}$$

given that feature map is of dimension hxh

3.4 Reason for CAM expansion

Since we are using binary cross entropy the loss function is given as-

$$L_{BC} = \sum_{i \in \{1,2..C\}} [y_i log(P_W\{\sum_{j \in \{1,2,..M\}} (M_{ij} \odot F_{ij})\})$$
$$+(1-y_i)log(1-P_w\{\sum_{j \in \{1,2,..M\}} (M_{ij} \odot F_{ij})\})]$$

$$L_{BC} = \sum_{i \in \{1,2..C\}} [y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)]$$

The above function is convex and will be maximized when-

$$y_i = \hat{y}_i -----1$$

$$y_i = log(P_W\{\sum_{j \in \{1,2,..M\}} (M_{ij} \odot F_{ij})\})$$

Here

- y_i is the target variable for i^{th} class
  C is the total number of classes

- P_W is a function which apply WILDCAT pooling.

$$P_W : R^{hxh} \rightarrow R$$

- M_{ij} is j^{th} drop-out mask obtained during intra-class sequential drop-out of M maps of i^{th} class

- F_{ij} is j^{th} feature map of i^{th} class

- $\hat{y}_i$ is the estimate for y_i given by the model

Now if the we substitute M_{ij} by matrix with all entries 1 then it becomes the loss function for WILDCAT model. Now since elements of the drop-out mask are ≤ 1 there would be a reduction in $\hat{y}_i$ for a given y_i but in order to maximize L_{BC} for a given y_i equation 1 must hold. The only way this can happen is by increasing elements of F_{ij} which is same as expansion of the feature maps as compared to WILDCAT maps. Thus the intra-class sequential drop-out must lead to CAM expansion.

3.5 Attention Mechanism

The attention mechanism has been used in various field like machine translation, question answering, image captioning, visual question answering etc. The idea is to focus on only important part of the input data to get the output. The inspiration comes from human. Human mind tends to focus only the important part of a passage or an image while making sense out of it. Another idea is self attention where the query, key and value all comes from the input itself which the central principle used in transformers.

The above architecture uses both self attention and attention mechanism. Attention maps are used to find out important areas in the feature maps. These maps are thresholded to get drop out mask. The first map of each class uses the idea of self attention to drop less important regions, while important areas in previous maps are also dropped in current maps in case of all other maps, so here we use the normal attention mechanism.

3.6 Loss function

- Binary cross entropy has been used as the loss function.
- $L = -\sum_i y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)$.
- Here y_i is the target variable while $\hat{y}_i$ is the prediction.
- Since the output of the network $\in R_C$,so we pass the output through a sigmoid so that the output become a probability i.e $\in (0,1)^C$

3.7 Ablation study

In order to prove empirically that the CAMs expand by the ex-CAM we train the ex-CAM and WILDCAT on PASCAL VOC 2012 and PASCAL VOC 2007. Figure 4 shows one CAM belonging to each of the 20 classes for both ex-CAM and WILDCAT. For each of the class we can see expansion of CAM. This leads to better bounding boxes as shown in Figure 3. Thus through this study we can conclude that the ex-CAM model concept of intra-class sequential dropout will lead to expansion of CAM. The concept can be applied to other model also for example we can replace the base model from WILDCAT to "Learning Deep Features for Discriminative Localization"[3] model and still use sequential dropout.

Class	**aeroplane**	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**	**dintable**	**dog**
ExCAM	0.8	0.81	0.74	0.7	0.27	0.76	0.44	0.84	0.46	0.65	0.88	0.73
WILDCAT	0.61	0.45	0.16	0.59	0.29	0.75	0.28	0.52	0.76	0.59	0.63	0.77
Class	**horse**	**bike**	**person**	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**			
ExCAM	0.81	0.9	0.71	0.58	0.32	0.77	0.49	0.5	0.7			
WILDCAT	0.62	0.56	0.59	0.29	0.55	0.55	0.77	0.28	0.53			

Table 1: AP for various classes on PASCAL VOC 2012

Class	**aeroplane**	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**	**dintable**	**dog**
ExCAM	0.79	0.82	0.74	0.57	0.26	0.72	0.67	0.84	0.52	0.68	0.81	0.76
WILDCAT	0.18	0.36	0.3	0.05	0.16	0.25	0.55	0.29	0.34	0.03	0.16	0.41
Class	**horse**	**bike**	**person**	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**			
ExCAM	0.9	0.84	0.68	0.49	0.61	0.83	0.82	0.47	0.69			
WILDCAT	0.3	0.4	0.72	0.2	0.07	0.27	0.12	0.3	0.27			

Table 2: AP for various classes on PASCAL VOC 2007

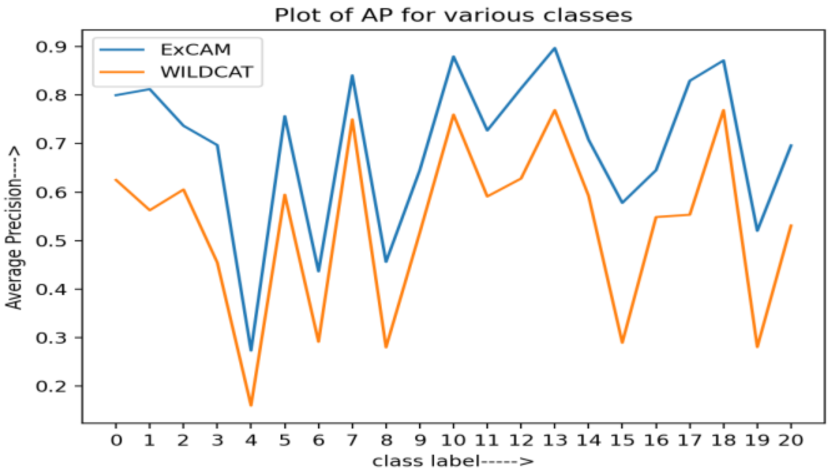


Figure 2: Plot of AP for various classes on PASCAL VOC 2012

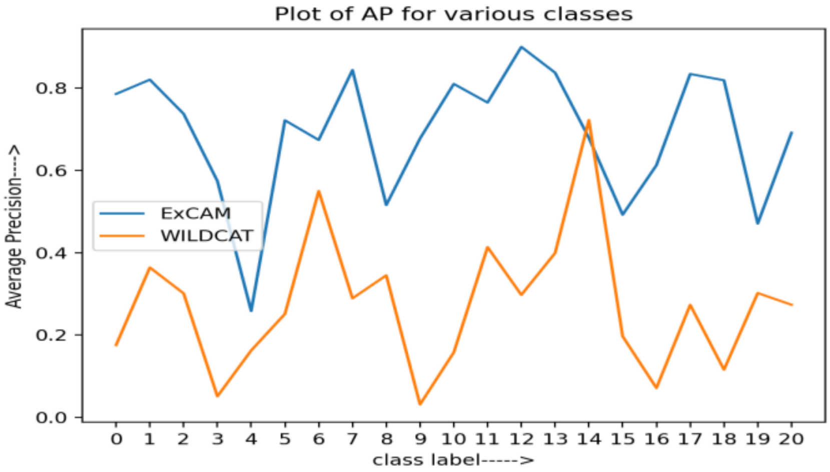
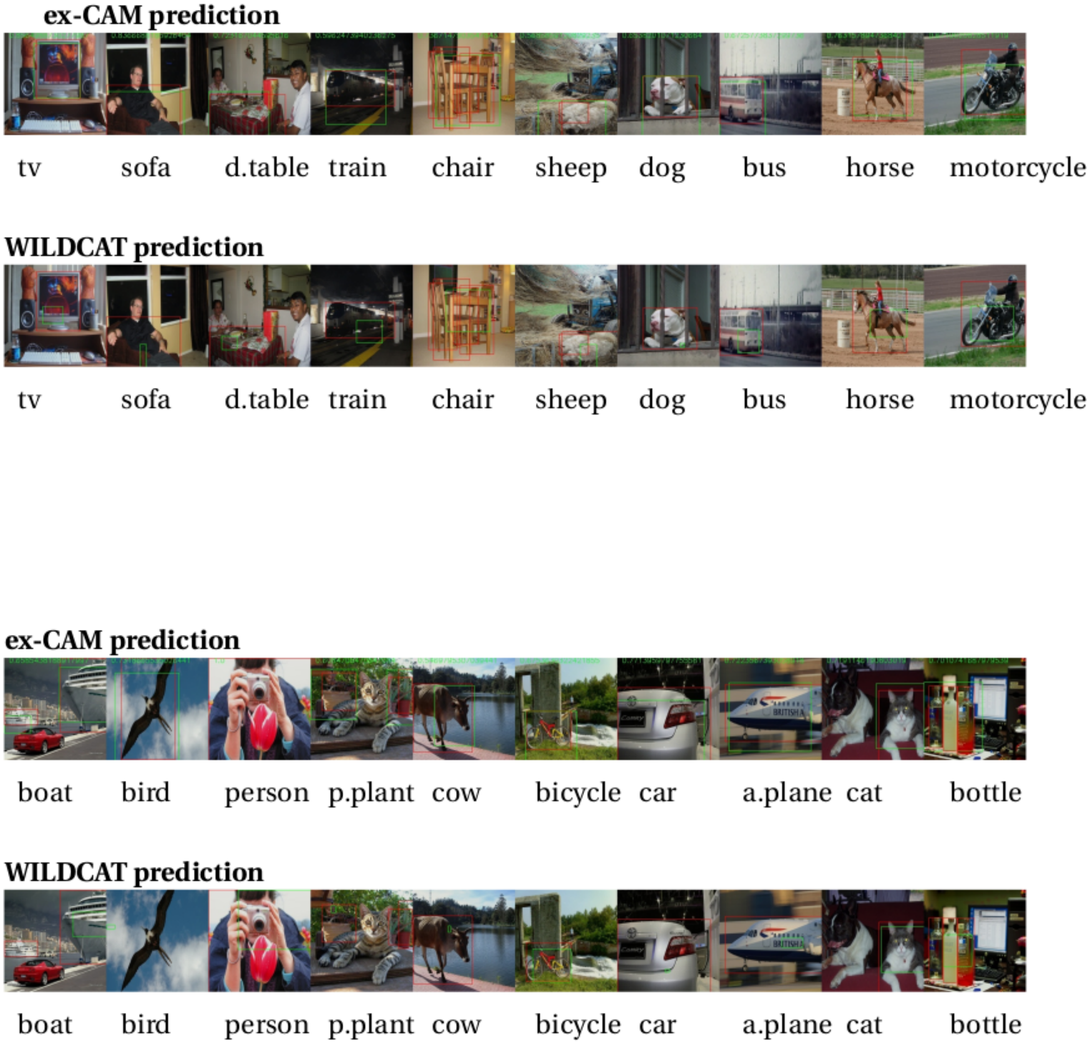


Figure 3: Plot of AP for various classes on PASCAL VOC 2007

Figure 4: Output Bounding boxes for exCAM and WILDCAT, Groundtruth are shown with red while predicted is shown with green



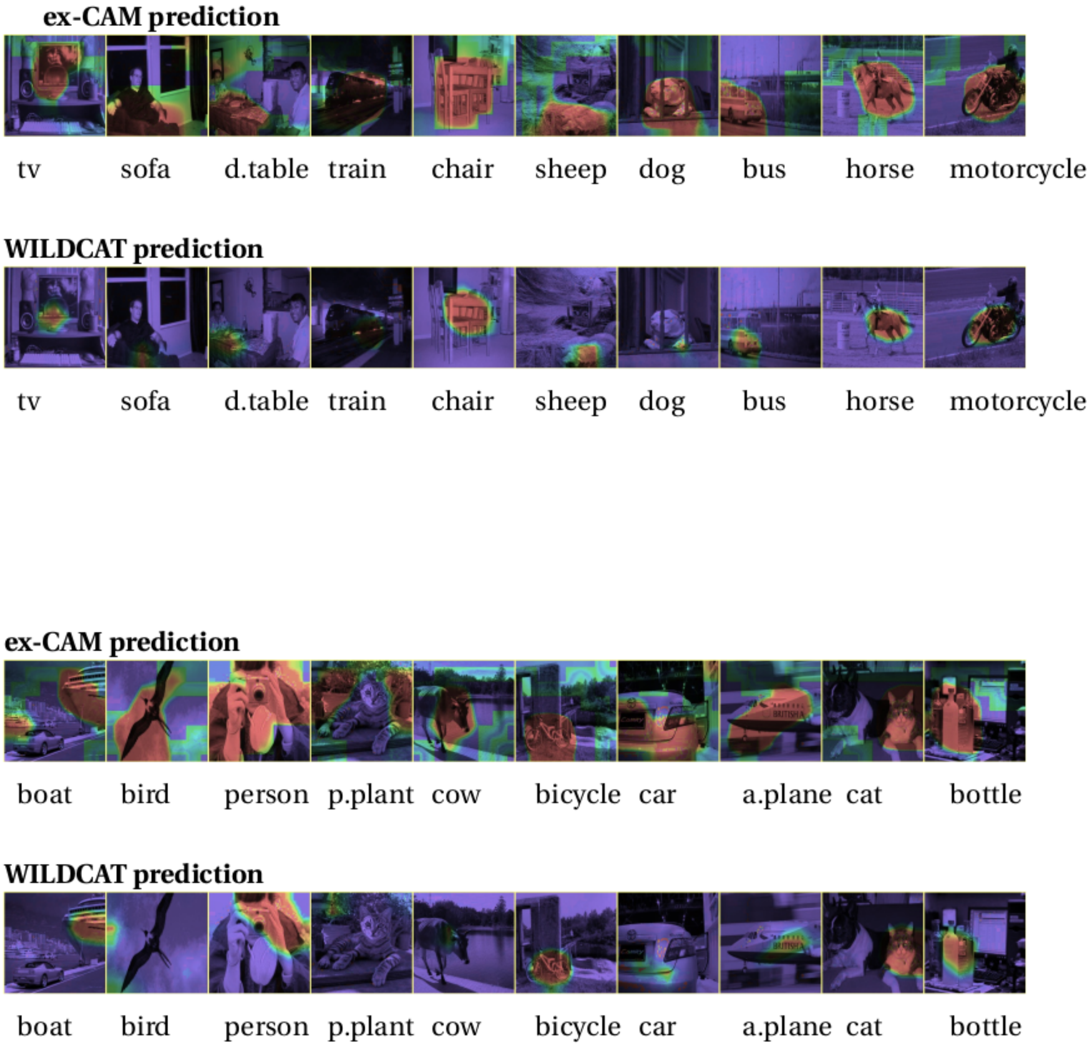
ex-CAM prediction

WILDCAT prediction

ex-CAM prediction

WILDCAT prediction

Figure 5: Output CAM for exCAM and WILDCAT, Groundtruth are shown with red while predicted is shown with green



Class	**aeroplane**	**bicycle**	**bird**	**boat**	**bottle**	**bus**	**car**	**cat**	**chair**	**cow**	**dintable**	**dog**
ExCAM	1.0	0.89	0.87	0.88	0.58	0.76	0.73	0.89	0.87	0.89	0.89	0.75
Class	**horse**	**bike**	**person**	**pottedplant**	**sheep**	**sofa**	**train**	**tvmonitor**	**mAP**			
ExCAM	0.9	0.95	0.98	0.78	0.92	0.83	0.96	0.6	0.85			

Table 3: Point based localisation metric for various classes

Method	**mAP**
Deep MIL[10]	74.5
ProNet[11]	77.7
WS Localisation[12]	79.7
WILDCAT[2]	82.9
ex-CAM	85

Table 4: Comparison of ex-CAM with WILDCAT and other methods on mAP based on point based localisation metric

3.8 Other Experiments

3.8.1 Making feature maps mutually exclusive

The WILDCAT prediction in the Durand el al[2] paper claim to give MC feature maps in which each M feature maps extract different parts of the same object. However in practice all the M feature maps are almost exactly same. The reason for this is that in the WILDCAT model there is no constrain or measure to make the M feature maps mutually exclusive. To address these issues we try to make these maps mutually exclusive. To do this we find pairwise L_2(or L_1) loss between feature maps thus we have -

$$L_{PW} = \sum_{(i,j)\in\{1,2,..M\}x\{1,2,..M\}} ||F_i - F_j||_2^2$$

We subtract the above pairwise loss to the binary cross-entropy loss L_{BC} after multiplying it by a hyper parameter λ and thus the final loss-

$$L = L_{BC} - L_{PW}$$

$$\begin{aligned} L = \sum_{i\in\{1,2..C\}} [y_i log(P_W\{ \sum_{j\in\{1,2,..M\}} (M_{ij} \odot F_{ij})\}) \\ +(1-y_i) log(1-P_w\{ \sum_{j\in\{1,2,..M\}} (M_{ij} \odot F_{ij})\})] \\ -\lambda \cdot \sum_{(i,j)\in\{1,2,..M\}x\{1,2,..M\}} ||F_i - F_j||_2^2 \end{aligned}$$

But in order to use this we need hyperparameter tuning of λ which is will take many iterations. We did't proceed with this work after many trials.

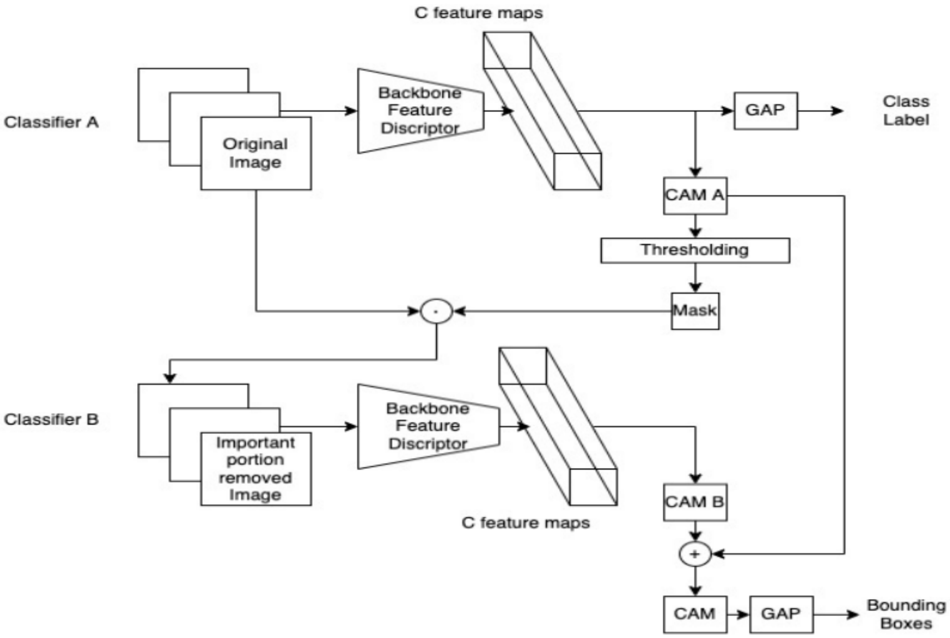


Figure 6: Figure Model Architecture for WILDCAT + ACoL combination

3.8.2 Combining the idea of adversarial complementary learning and WILDCAT

The problem with WILDCAT was that the CAM obtained were very small and cover only the most discriminative regions. In order to to do this we train WILDCAT as classifier A. Extract CAM from this. We then resize this to original image size and do thresholding to drop important regions. After this we do hadamard product of this binary mask with the original image. Now this image will have the important regions(detected by Classifier A) removed. Now we train another WILDCAT classifier B on these new images and again extract the CAM from this classifier. Now we do Average Pooling of CAM obtained from classifier A and classifier B. This approach gave bigger maps as compared to the WILDCAT. However there was not a great improvement moreover in the case were the classifier A has sufficiently big CAM the classifier B will be forced to extract unimportant regions which will lead to poor CAM. This problem will be more common in case of multi label dataset. If two classes are similar say class 1 and 2 and if the classfier A covers the object of class 1 well then the classifer B will be forced to extract the region of object of class 2. Thus the final map will have regions of both class 1 and 2 object.

3.8.3 Combination of Cloe et al with WILDCAT

The problem of smaller CAM in WILDCAT can be also be resolved by combining the concept of WILDCAT with Cloe et al[6]. In this architecture after getting MC maps from WILDCAT backbone we do the following-

•Suppose the j_{th} feature map of i^{th} class is F_{ij} then we pass it through a sigmoid to get importance map I_{ij}.

•By thresholding this importance map we obtain a dropout map M_{ij}. M_{ij} is a binary map that drop the important region and represent them by 0.

•We randomly select one of the two importance map and drop out map and do hadamard product with the feature map F_{ij} to obtain new feature map $F_[\hat{i}j]$.

$$\hat{F_{ij}} = random(I_{ij}, M_{ij}) \odot F_{ij}$$

Now these new feature map set $\{\hat{F_{ij}}(i,j) \in \{1,2,...,M\} \times \{1,2,...,M\}\}$ are subjected to intra class GAP to obtain C feature maps. These C feature maps are further subjected to WILDCAT pooling.

During training we train this architecture for classification. During testing we don't perform randomization and hadamard product instead we directly use F_{ij} and subject it to GAP and WILDCAT pooling as explained earlier. This method led to over expansion of the feature maps and we did't proceed with this method.

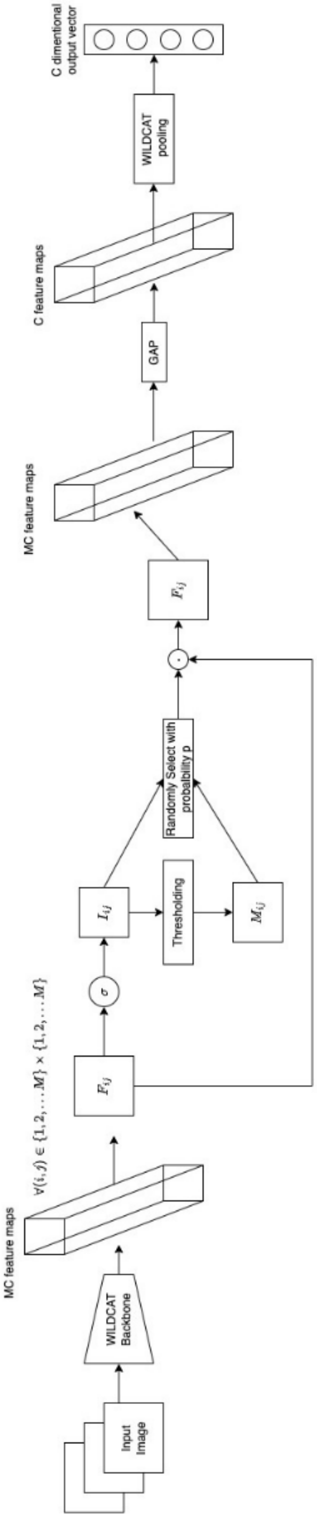


Figure 7: WILDCAT+Cloe model architecture

4 CHAPTER 4: Conclusion and results

4.1 Object Localisation on PASCAL VOC 2012 and 2007

We ran the ex-CAM and WILDCAT model on PASCAL VOC 2012 and obtained the result shown in Table 1. In all of the object categories the ex-CAM prediction are much better than WILDCAT. Over all the mAP for ex-CAM is 0.7 while for WILDCAT it is 0.53. So there is a significant improvement as compared to WILDCAT. We also evaluated the model on point based localisation metric as used in "WILDCAT" paper and results are shown in Table 3. A comparison with WILDCAT and other model on point based localisation metric is shown in table 4. As we can see ex-CAM has out performed WILDCAT and the other models with significant margin. The AP for each class for ex-CAM is shown in table 3. The point based object localisation metric however does not considers the quality of bounding boxes. This is the reason why the difference of localisation mAP for ex-CAM and WILDCAT increases when we use IoU instead of point based localisation metric.

We also ran the ex-CAM and WILDCAT model on PASCAL VOC 2007 and obtained the result shown in Table 2. In all of the object categories the ex-CAM prediction are much better than WILDCAT. Over all the mAP for ex-CAM is 0.69 while for WILDCAT it is 0.27. So there is a significant improvement as compared to WILDCAT. The person category shows a minor reduction is due to the fact that the pascal voc 2007 bounding boxes are itself noisy. If there is a group of people in the background and there is also one or more person that cover majority portion of the image then the groundtruth bounding boxes cover only these persons not the group. However the model tries to cover all which reduces its IoU and lower the AP.

4.2 Training and testing

The model uses resnet101 architecture. We remove the last pooling layer. The output 7x7x2048 maps from this resnet are subjected to 1x1 convolution to obtain 7x7xMC maps. These MC maps are subjected to intra-class sequential dropout. Now we perform global average pooling(GAP) to get C feature maps. These C feature maps are subjected to WILDCAT pooling to get a vector of C dimension.

During training the ex-CAM is trained for classification. For testing we extract feature maps which is used to get CAM which are passed through sigmoid and thresholded. The resultant binary mask is used to obtain bounding boxes.

The model has been trained on 48GB A6000 Nvidia GPU. It occupies 18GB of GPU memory at batch size 16. We have taken k+=k-=0.2 for WILDCAT pooling and

$\alpha = 0.6$. The number of feature maps M is 8. During training we train the model as a classifier. During testing we extract feature maps which are used to obtained CAM. These CAM are used to obtain bounding boxes.

4.3 Conclusion

The proposed ex-CAM model shows significant improvement in CAM over WILDCAT. Due to intra-class sequential dropout the model not only covers the most discriminative part but also the other parts of the object. The same model can be used for classification and localisation task. This shows that the idea of intra-class sequential dropout is a general way to improve CAM.

Future work would include changing the base model from WILDCAT to some other model. For improving the bounding boxes one can do segmentation of the area covered by CAM with high probability. This would improve the CAM and improve the localisation mAP. The future works can include hyperparameter tuning of architectures shown in Other Experiment sub-section under Model Architecture section.

References

[1] Weakly Supervised Object Localization and Detection: A Survey Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang

[2] WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation Thibaut Durand, Taylor Mordan, Nicolas Thome, Matthieu Cord

[3] Learning Deep Features for Discriminative Localization Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba Computer Science and Artificial Intelligence Laboratory, MIT

[4] Adversarial Complementary Learning for Weakly Supervised Object Localization Xiaolin Zhang1 Yunchao Wei, Jiashi Feng, Yi Yang, Thomas Huang

[5] Rethinking the Route Towards Weakly Supervised Object Localization Chen-Lin Zhang Yun-Hao Cao Jianxin Wu National Key Laboratory for Novel Software Technology Nanjing University, Nanjing, China

[6] Attention-based Dropout Layer for Weakly Supervised Object Localization Junsuk Choe, and Hyunjung Shim School of Integrated Technology, Yonsei University, South Korea

[7] Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5) Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik UC Berkeley

[8] Fast R-CNN Ross Girshick Microsoft Research

[9] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun

[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object local- ization for free? Weakly-supervised learning with convolu- tional neural networks. In CVPR, 2015.

[11] C. Sun, M. Paluri, R. Collobert, R. Nevatia, and L. Bour- dev. ProNet: Learning to Propose Object-Specific Boxes for Cascaded Neural Networks. In CVPR, 2016.

[12] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. S. Manjunath. Weakly supervised localization using deep fea- ture maps. In ECCV, 2016.

Expanding CAM inWeakly Supervised Object Localization

ORIGINALITY REPORT

1 %
SIMILARITY INDEX

PRIMARY SOURCES

1	library.isical.ac.in:8080 Internet	17 words — < 1%
2	Utpal Garain, Biswajit Halder. "Even big data is not enough: need for a novel reference modelling for forensic document authentication", International Journal on Document Analysis and Recognition (IJDAR), 2019 Crossref	15 words — < 1%
3	cse.iitkgp.ac.in Internet	15 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 14 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 14 WORDS