WEAKLY SUPERVISED OBJECT LOCALIZATION: THE ANALYSIS ON INTRA-CLASS VARIANCE IN DATASET

Santosh Muthireddy, Deebul Nair & Paul G. Plöger
Department of Computer Science
Bonn-Rhein-Sieg University of Applied Sciences
Sankt Augustin, 53757, Germany
santosh.muthireddy@smail.inf.h-brs.de
{deebul.nair,paul.ploeger}@h-brs.de

ABSTRACT

Weakly Supervised Object Localization (WSOL) uses classification models trained only on image-level labels to localize objects. Although different WSOL techniques have shown results in improving object localization performance under weak supervision, limited studies have analyzed the role of dataset distribution on the performance of WSOL techniques. In this paper, we study the effects of intra-class variance in the dataset using selected WSOL techniques. We argue that intra-class variance is an important factor when applying WSOL techniques. We observe from our experiments that 2 baseline methods have a drop in performance when trained on datasets with limited intra-class variance. Furthermore, we show that artificially augmenting the intra-class variance improves the localization performance.

1 INTRODUCTION

Object localization techniques based on deep learning like Single Shot Detection (SSD) (Liu et al., 2016) and You Only Look Once (YOLO) (Redmon et al., 2016) require a large number of training images in a strongly supervised setting. But manual annotation is expensive and time-consuming. In weakly supervised learning, coarse-grained annotations are available at the training time and fine-grained annotations are expected during the evaluation or testing phase. Weakly Supervised Object Localization (WSOL) techniques require only image-level labels during training and are capable of extracting object locations (Zhou et al., 2016; Singh & Lee, 2017; Selvaraju et al., 2017; Choe & Shim, 2019).

WSOL is a vastly studied research area but not much work is done towards analyzing the impact of dataset distribution in the performance of WSOL techniques. Understanding the role of dataset distribution is significant as datasets are collected in automated setup these days. The general assumption is that balanced datasets will perform better in classification tasks which is true but WSOL techniques rely on discriminant features to localize the objects which depend on intra-class variance. In this work, we analyzed the performance of WSOL techniques on three datasets PASCAL Visual Object Classes (PASCAL VOC), Yale-CMU-Berkeley (YCB) (Calli et al., 2017) and RoboCup@Work.

Class Activation Mapping (CAM) (Zhou et al., 2016) is considered the baseline method in WSOL. CAM uses the discriminative features of an object to localize it but there are few shortcomings in CAM as mentioned by Choe et al. (2020). So, Gradient Class Activation Mapping (Grad-CAM) is introduced to counter the shortcomings of CAM. In this paper, CAM and Grad-CAM are used to investigate the effect of dataset distribution. We trained a multi-label classification model with VGG16 as the feature extractor. The two datasets YCB and RoboCup@Work are used for object detection and localization tasks to help robots in manipulation tasks like grasping different objects in a domestic and industrial setup respectively. These two datasets are collected from a rotating table setup which automates the data collection. This way of collecting datasets helped to generate balanced datasets with the same number of images/instances per class. It is observed that models trained with balanced datasets perform better compared to models trained with an imbalanced dataset

in classification tasks but shown poor performance in object localization using WSOL techniques. Data augmentation techniques are applied on RoboCup@Work dataset to increase intra-class variance (Santosh, 2019). Experiments on the new dataset showed improved localization results using WSOL techniques. From this, we can show that high intra-class variance and complex datasets will give better localization performance.

Our major contributions in this paper are (1) analyzing the effect of intra-class variance in the dataset on the performance of WSOL techniques, (2) increasing intra-class variance using artificial augmentation and analyzing the performance of WSOL techniques, and (3) evaluating CAM and Grad-CAM on two robotic application datasets YCB and RoboCup@Work on MaxBoxAccV2 and PxAP metrics proposed by Choe et al. (2020).

2 Related Work

According to researchers, the levels of supervision in an object localization task can be defined at image-level (Papandreou et al., 2015), points (Bearman et al., 2016), bounding boxes (Dai et al., 2015), gaze (Papadopoulos et al., 2014), scribbles (Lin et al., 2016) or combination of multiple types (Hong et al., 2015). WSOL (Zhang et al., 2020) is an object localization technique in which models trained on image-level labels learn to localize the objects. Nowadays WSOL became attractive, as image-level annotations can be obtained with less human effort, less time, and at a much cheaper cost than instance-level or pixel-level labels. In this section, the literature of WSOL is presented and tried to present the research in the field.

The work presented in Zhou et al. (2016) illustrates the ability of Convolutional Neural Networks (CNNs) to have a Global Average Pooling (GAP) layer before classifiers to localize objects explicitly although they are trained on image-level labels. In Zhou et al. (2016), Class Activation Mapping (CAM) is proposed which generates a score map from a fully-convolutional classifier by manipulating activations before the GAP layer. However, the initial CAM approach is criticized for just using small discriminative parts in the images for localizing the objects. Techniques like Hide-and-Seek (HaS) (Singh & Lee, 2017) and Cutmix (Yun et al., 2019) are proposed where few patches in input images are randomly dropped to diversify the cues. Another method HaS (Singh & Lee, 2017), hides some patches in training images randomly and forces the network to learn other less important parts when most discriminative parts randomly are hidden. Whereas in Cutmix (Yun et al., 2019) authors stated ”patches are cut and pasted among training images where the ground truth labels are also mixed proportionally to the area of the patches.” Apart from these, adversarial techniques (Zhang et al., 2018a; Choe & Shim, 2019) are also proposed which dynamically drops the most significant patch in the given image. In Zhang et al. (2018a), Adversarial Complementary Learning (ACoL) activation maps from the penultimate convolution layer are used to generate class localization maps. Mid-level features are extracted and passed into two parallel-classifiers for finding interdependent object regions. In this method, activation maps from two classifiers A and B are combined to obtain object maps thus locating the objects. Similarly in Choe & Shim (2019), Attention-Based Dropout Layer (ADL) based localization tries to hide the most discriminative parts in the image and highlight informative regions to improve WSOL accuracy. This is a lightweight and powerful method based on the self-attention mechanism is used to remove the most discriminative part.

Self-Produced Guidance (SPG) (Zhang et al., 2018b) utilizes local correlation between pixels to generate SPG masks. SPG maks are capable of separating foreground and background. These masks are used to provide supervision during training. Mid-level features are extracted from inputs and fed into SPG for classification. CAMs generated from classification networks are used by SPG to gradually learn the refined CAMs. Then another SPG net uses these refined CAMs as supervision to further improve the final output. In recent developments Choe et al. (2020) as shown with extensive experiments and new metrics that primitive CAM (Zhou et al., 2016) is still the top-performing method in WSOL. So, approaches like Gradient Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) uses gradients unlike weights of classifier in CAM (Zhou et al., 2016). These generate more accurate class activation maps and removes the dependency of having a GAP layer in the architecture to generate a CAM. Grad-CAM is the generalization of CAM which can be applied to any CNN based architectures without doing any modifications. A further improvement over Grad-CAM is Grad-CAM++ (Chattopadhay et al., 2018), which provides better network visualization than Grad-CAM and improves the localization of objects and finding all occurrences of multiple objects

that belong to the same category. Reformulated the structure of weights in Grad-CAM++ which uses a weighted sum of positive partial derivatives from activation maps of the last convolutional layer.

In this paper, CAM and Grad-CAM are used to investigate the effect of dataset distribution as it has been shown by Choe et al. (2020) that the recent methods have no huge improvements over the baseline.

3 DATASETS

In selected datasets, image-level annotations are available for training i.e, the dataset should support classification tasks. Object-level annotations are required for evaluation i.e., bounding boxes and semantic masks should be available for calculation of evaluation metrics. Multi-class and multi-label datasets are encouraged. We conducted experiments on PASCAL VOC (Everingham et al., 2010), Yale-CMU-Berkeley (YCB) dataset (Calli et al., 2017) and RoboCup@Work datasets [1], as we are also interested in solving the issues faced by b-it-bots @Home and @Work teams at the university. In this research, the classification/detection task dataset from PASCAL VOC (Everingham et al., 2010) is used for training the classification model and evaluating metrics accuracy, precision, and recall. Whereas segmentation task dataset is used to evaluate against the metrics MaxBoxAccV2 (Choe et al., 2020) and PxAP (Choe et al., 2020). Each set contains three data splits *train*: training data, *val*: validation data and *trainval*: the union of train and val data splits. Only *val* data split is used from segmentation dataset. The dataset contains 20 object categories they are Aeroplane, Bicycle, Bird, Boat, Bottle, Bus, Car, Cat, Chair, Cow, Diningtable, Dog, Horse, Motorbike, Person, Pottedplant, Sheep, Sofa, Train, Tvmonitorwhich. These are collected from Flickr. The average image size is 470x380 pixels and the average number of objects per image is 2.4.

YCB is a benchmarking dataset for robotic grasping and manipulation tasks. A total of 77 objects are categorized into 5 types: Food items, Kitchen items, Tool items, Shape items, Task items. In a given image only one instance of the object is present. In total, High resolution RGB images, Each image has a semantic mask, Camera calibration data, Texture-mapped 3-D mesh models.

RoboCup@Work dataset consists of a total of 12 classes which are Axis, Bearing, Bearing box, F20 20 black, F20 20 gray, M20, M20 100, M30, Motor, R20, S40 40 black, and S40 40 gray. This dataset is collected in-house using the automatic data collection and augmentation tool *Easy Augment* (Santosh, 2019). In this dataset, each image consists of only a single object instance. Each category contains the following: 60 RGB images with 640x480 resolution, Corresponding depth frames, Segmentation masks for each image, Point cloud of a segmented object, Bounding box annotations in PASCAL VOC format. According to Simone, imbalance in dataset can be measured using Shannon entropy. Shannon entropy can be defined with Equations 1 and 2, where n is total number of instances, k is number of classes and c_i is count of instances for i^{th} class.

$$H = -\sum_{i=1}^{k} \frac{c_i}{n} \log \frac{c_i}{n} \tag{1}$$

$$Balance = \frac{H}{\log k} \tag{2}$$

The measure of balance is quantified using Equation 2 which tends to 0 for unbalanced dataset and tends to 1 for a balanced dataset. From the data analysis shown in Figure 1 we can see that PASCAL VOC dataset is relatively imbalanced and other two datasets are balanced.

[1] Custom collected dataset

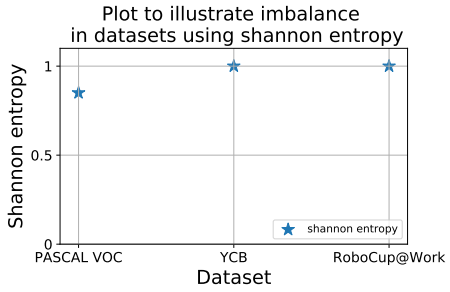

Figure 1: Plot illustrates imbalance in datasets using Shannon entropy, where YCB and RoboCup@Work are perfectly balanced datasets and PASCAL VOC is relatively imbalanced.

4 EXPERIMENTS

In experimentation, a multi-label classification model is trained for the PASCAL VOC, YCB, and RoboCup@Work datasets using VGG16 as backbone. All the experiments are carried out with the same setup as follows, the batch size is 32 with a learning rate of 1e-4 and image size of (256×256×3). The optimizer used is Stochastic Gradient Decent (SGD) with Binary Cross Entropy (BCE) loss. The objective is to evaluate the performance of CAM (Zhou et al., 2016) and GradCAM (Selvaraju et al., 2017) on different datasets namely PASCAL VOC, YCB, and RoboCup@Work. For experimentation and evaluation code implementation is adapted from Choe et al. (2020).

4.1 CLASS ACTIVATION MAPPING

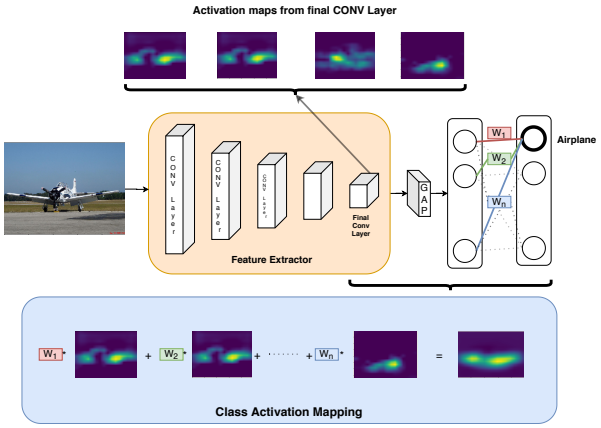

Figure 2: Class Activation Mapping: the confidence score of predicted class is projected back to the preceding convolution layer to generate the class activation maps (CAMs). Inspired from Zhou et al. (2016).

In Zhou et al. (2016), authors showcased that CNNs have a very good capability to predict the location of objects although they are trained for classification tasks. CAM for a selected category describes the distinctive regions of images, in which CNN's are used to classify the selected category object. To achieve this, an average pooling layer is added in between the last convolution layer and a fully connected classification network. These discriminative regions are identified by projecting weights of the classification output layer that corresponds to the selected category on the feature maps from the end convolution layer. This is called Class Activation Mapping and the weighted sum of the activation maps and output layer weights will result in a Class Activation Map for the selected category which is illustrated in Figure 2. In the CAM, proposed by Zhou et al. (2016), a drawback is that architectures that are performing GAP layer over activation maps immediately

before classifier layers are only applicable. Such architectures may perform poorly relative to the general network on some tasks.

4.2 GRAD-CAM

Gradients in learning techniques are vectors whose value is a partial derivative of the function and this gradient will be flowing towards the steepest rate of increase of that function. Gradient Class Activation Mapping (Gard-CAM) exploits this information along with class specifics to produce score maps of significant regions in the given image. Grad-CAM combines pixel space gradient information with class discriminative property to generate score maps. The architecture of Grad-CAM is shown in Figure 3. In Grad-CAM, spatial location data of an object is preserved so the last convolution layers are utilized in the generation as neurons from that layer are capable of identifying significant parts to the given class.

To generate a Grad-CAM for a given image and predicted class, gradients of the score for the predicted class (let's say c) y_c concerning feature maps A_k of the selected convolutional layer which is given by $\frac{\partial y_c}{\partial A_k}$. Once gradients are obtained weight that specifies the significance of the feature map is calculated using Equation 3 where the summation over i and j specify GAP and partial derivatives are the gradients from backpropagation. As shown in Figure 3, the weighted sum of activation maps passed through the Rectified Linear Unit (ReLU) is performed to get the final score map for the predicted class. This summation is described in Equation 4. ReLU is used to combine the feature maps as it highlights the significant features having a positive effect on the predicted class.

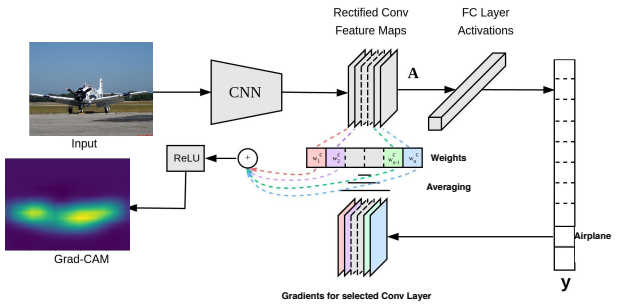

Figure 3: Gradient Class Activation Mapping architecture. Grad-CAM uses gradient flow passing through last convolutional layer to generate weights. Inspired from Selvaraju et al. (2017).

$$a_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k} \tag{3}$$

$$Scoremap_c = ReLU \sum_k a_k^c A_k \tag{4}$$

4.3 EVALUATION METRICS

PASCAL VOC dataset contains multi-label data that is given image can have instances of multiple categories. In this regular metrics cannot be used as performance can be quantified inappropriately. So metrics prescribed in the lecture (Jesse Read, 2013) are used for evaluating classification performance of models.

CLASSIFICATION ACCURACY

Classification accuracy is defined as the rate of correct predictions for a given test or validation set. Classification accuracy for multi-label classification is calculated using Equation 5, where $\hat{y}^i$ is

prediction, y^i is ground truth and N is number of samples.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}^i \wedge y^i|}{|\hat{y}^i \vee y^i|} \tag{5}$$

PRECISION AND RECALL

Precision and Recall are calculated using scikit learn metrics. Precision is the ratio given by Equation 6 and recall is the ratio given by Equation 7, where tp is number of true positives, fp is the number of false positives and fn is the number of false negatives. The precision score indicates the ability of trained classification model that will not label negative sample as positive. The recall score measures the ability of trained classification model to find all positive samples.

$$Precision = \frac{tp}{tp + fp} \tag{6}$$

$$Recall = \frac{tp}{tp + fn} \tag{7}$$

The localization accuracy is the measurement of localization performance of trained model. In general Top-1 localization accuracy is used to quantify the localization performance in WSOL methods which uses the normalized scoremap. After normalizing scoremap, WSOL techniques threshold the scoremap at t. This thesholded binary mask is used to generate tight fit bounding box. Generally t is fixed value which gives wrong performance measure for localization as t depends on data and model architecture. So new evaluation metrics are proposed in Choe et al. (2020) to overcome the dependency of threshold t. Proposed metrics are discussed below:

MAXIMAL BOX ACCURACY (MAXBOXACC)

This metric can be calculated when bounding boxes are available as ground truth. The box accuracy for heatmap at the given threshold t is defined by Equation 8. It is quantified by the proportion of images where the box spawned from the scoremap intersects with the ground truth at multiple IoU thresholds. Each score map is thresholded at range of values to obtain threshold independence in generating binary mask, masks generated at different thresholds and bounding boxes are extracted. Final performance metric is mean of MaxBoxAcc across all IoU thresholds given in Equation 9. To get the threshold independence of IoU, MaxBoxAccV2 is proposed where mean across different IoU thresholds is calculated given in Equation 10, where $box(s(X^{(n)}, t))$ is tightest bounding box, $B^{(n)}$ ground truth bounding box, t is threshold to generate binary mask and t_{IoU} is threshold for IoU.

$$BoxAcc(t) = \frac{1}{N}\sum_{n} l_{IoU}(box(s(X^{(n)}, t)), B^{(n)}) \geq t_{IoU} \tag{8}$$

$$MaxBoxAcc = max_t BoxAcc(t) \tag{9}$$

$$MaxBoxAccV2 = mean_{t_{IOU}}(max_t BoxAcc(t)) \tag{10}$$

PIXEL AVERAGE PRECISION (PXAP)

This metric can be calculated when semantic masks are available as ground truth. If a precision recall curve is generated at pixel level G then area under the curve is calculated as a measure. Pixel precision and recall at specific threshold is given by Equations 11 and 12. Threshold independence is achieved by calculating pixel average precision (PxAP) using Equation 13.

$$PxPrec(t) = \frac{|\{s_{ij}^{(n)} \geq t\} \cap \{T_{ij}^{(n)} = 1\}|}{|\{s_{ij}^{(n)} \geq t\}|} \tag{11}$$

$$PxRec(t) = \frac{|\{s_{ij}^{(n)} \geq t\} \cap \{T_{ij}^{(n)} = 1\}|}{|\{T_{ij}^{(n)} = 1\}|} \tag{12}$$

$$PxAP = \sum_{l} PxPrec(t_l)(PxRec(t_l) - PxRec(t_{l-1}) \tag{13}$$

Where s_{ij} is value of pixel at (i, j) in scoremap, T_{ij} is value of pixel at (i, j) in ground truth mask, t is the threshold and n is image id.

5 RESULTS

These experiments are done to answer the question, whether an imbalanced dataset affects the localization and classification performance of WSOL techniques. This evaluation helps in finding how to collect custom datasets for RoboCup teams. The hypothesis for this experiment is that balanced datasets will perform better in classification and localization tasks. From Table 1, for CAM method we can observe that balanced datasets YCB and RoboCup@Work are performing better than the PASCAL VOC dataset in terms of the classification task. But, it is also seen that it did not hold for localization tasks, where the PASCAL VOC dataset is performing better. This is due to the more number of difficult samples present in the PASCAL VOC dataset shown in Figure 4, unlike other datasets where the background is pretty much constant. This enforces the network to learn more significant features which will help in better localization. Similar outcomes can be seen in Table 1 for the Grad-CAM method.

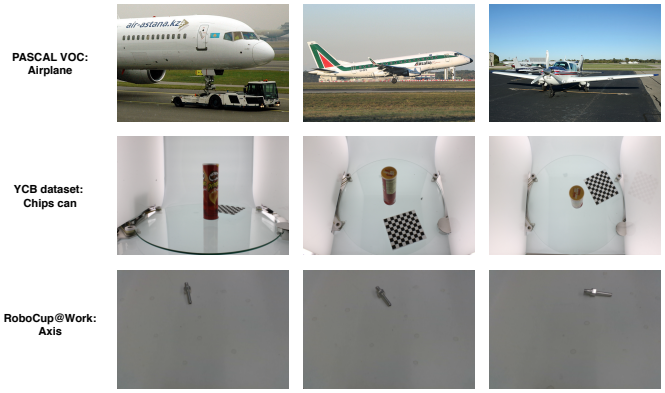

Figure 4: Example of intra-class variance present in selected classes from each dataset.

Table 1: Classification and localization results for CAM and Grad-CAM methods on three datasets.

Method	Dataset	Classification			Localization	
		Accuracy	Precision	Recall	MaxBoxAcc	PxAP
CAM	PASCAL VOC	72.70	82.85	78.10	**58.97**	**50.73**
	YCB	99.76	99.99	100.00	45.20	23.06
	RoboCup@Work	92.86	96.43	97.32	24.94	12.80
Grad-CAM	PASCAL VOC	72.70	82.85	78.10	**55.21**	42.35
	YCB	99.76	99.99	100.00	54.19	**46.09**
	RoboCup@Work	92.86	96.43	97.32	28.75	19.88

To generalize the claim on the effect of intra-class variance in the dataset, we conducted another experiment by generating an augmented dataset from the original RoboCup@Work dataset. From Table 2, we can observe there is a significant improvement in localization metrics once the dataset is augmented to increase intra-class variance and new dataset is shown in Figure 5. The results from our experiments clearly reveal a strong dependence of WSOL methods on intra-class variance.

Table 2: Classification and localization results for CAM and Grad-CAM methods on original and augmented RoboCup@Work datasets.

Method	Dataset	Classification			Localization	
		Accuracy	Precision	Recall	MaxBoxAcc	PxAP
CAM	RoboCup@Work	92.86	96.43	97.32	24.94	12.80
	New RoboCup@Work	96.88	98.94	97.73	**47.85**	**61.17**
Grad-CAM	RoboCup@Work	92.86	96.43	97.32	28.75	19.88
	New RoboCup@Work	96.88	98.94	97.73	**47.95**	**36.16**

Figure 5: Example of intra-class variance present in selected classes from each dataset.

6 CONCLUSION

This paper investigated the role of dataset imbalance and intra-class variance on the classification and localization performance of WSOL techniques. A comparison is done between balanced YCB and RoboCup@Work datasets, and relatively imbalanced PASCAL VOC dataset. YCB and RoboCup@Work datasets with less intra-class variance perform better in classification tasks but not in localization tasks. Whereas, the PASCAL VOC dataset having high intra-class variance performs better in localization tasks. In PASCAL VOC more difficult examples are present with varying backgrounds and noise which forces the model to learn more significant features of the object. This resulted in better localization performance using the PASCAL VOC dataset. In the second set of experiments, we increased the intra-class variance in the RoboCup@Work dataset with background augmentation and observed an improvement in localization performance.

REFERENCES

Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 549–565, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017. doi: 10.1177/0278364917700714. URL https://doi.org/10.1177/0278364917700714.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.

Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. to appear.

Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1495–1503. Curran Associates, Inc., 2015.

Jesse Read. Multi-label classification. URL: https://users.ics.aalto.fi/jesse/talks/Multilabel-Part01.pdf, 7 2013.

Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. pp. 21–37. Springer, 2016.

Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 361–376, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Muthireddy Santosh. Easy augment. https://github.com/santoshreddy254/easy_augment, 2019.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Simone. a general measure of data-set imbalance. Cross Validated. URL https://stats.stackexchange.com/q/239982.

Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3544–3553. IEEE, 2017.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. Rethinking the route towards weakly supervised object localization. *arXiv preprint arXiv:2002.11359*, 2020.

Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. Adversarial complementary learning for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *European Conference on Computer Vision*. Springer, 2018b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.