Deep active learning for object detection

Soumya Roy
meetsoumyaroy@gmail.com
Asim Unmesh
a.unmesh@gmail.com
Vinay P. Namboodiri
https://www.cse.iitk.ac.in/users/vinaypn

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur,
UP, India

Abstract

Object detection methods like Single Shot Multibox Detector (SSD) provide highly accurate object detection that run in real-time. However, these approaches require a large number of annotated training images. Evidently, not all of these images are equally useful for training the algorithms. Moreover, obtaining annotations in terms of bounding boxes for each image is costly and tedious. In this paper, we aim to obtain a highly accurate object detector using only a fraction of the training images. We do this by adopting active learning that uses 'human in the loop' paradigm to select the set of images that would be useful if annotated. Towards this goal, we make the following contributions:
1. We develop a novel active learning method which poses the layered architecture used in object detection as a 'query by committee' paradigm to choose the set of images to be queried.
2. We introduce a framework to use the exploration/exploitation trade-off in our methods.
3. We analyze the results on standard object detection datasets which show that with only a third of the training data, we can obtain more than 95% of the localization accuracy of full supervision. Further our methods outperform classical uncertainty-based active learning algorithms like maximum entropy.

1 Introduction

In object detection, we aim to accurately obtain bounding boxes in an image that contain objects of a particular category like bicycle, person or motorbike. The recent advancements in the field of deep learning [7, 17, 22] have dramatically improved the results of object detection. These detectors are trained on fully annotated images, the annotations being the image labels and the corresponding ground-truth bounding boxes containing the objects of interest. However, this kind of strong supervision is rather difficult to obtain as it requires immense time and manual effort to annotate each object in an image. Moreover, in some cases like medical images, this manual annotation process might have cost implications. Besides, not all the training images are equally useful since 'similar' images may not contribute much in learning. A solution to this problem is to cleverly choose training examples such that we can get state-of-the-art performance with as less supervision as possible. This method is called active learning and has been widely used for object classification [12, 18, 32]. However, it has not been much explored for the problem of object detection.

This paper uses the popular Single Shot Multibox Detector (SSD) [17] for object detection. Typically, deep networks can generalize well only if they are trained on large number of images. But by leveraging the power of active learning, we can get near full supervision performance with just a fraction of the training images. Our method takes inspiration from the paradigm of query by committee [19] and uses the disagreement between the convolution layers in the SSD architecture to query images. We also empirically show that an active learning method, which uses the network architecture to query images, produces superior results compared to black-box methods like maximum entropy.

In this paper we make the following contributions:

- Construct a novel active learning method based on the query by committee [19] paradigm which drastically outperforms the common active learning methods.
- Introduce a simple framework to use the exploration/exploitation trade-off in our methods.

2 Previous Work

Object classification and detection are widely explored topics in computer vision [8, 9, 14, 17, 21]. Here we use SSD [17] for object detection because of it's superior accuracy, ease of use and speed. However, all of these methods use full supervision in terms of ground-truth bounding boxes for object detection. The cost of fully annotating images is high and so weak supervision has received attention over the last few years as it needs only image labels for training [1]. In this paper, we focus on a related paradigm called active learning [23] which provides state-of-the-art performance with minimum supervision.

Active learning has been widely used in object classification [10, 11, 12, 16, 18, 28, 32, 33], where given an image we want to label the presence or absence of an object of a particular class, and video annotation [13, 15, 29, 30]. [9] uses a query by committee approach to query images for classification. The committee of classifiers is formed by using batchwise dropout on the full network and subsequently the difference between the predicted label of a committee member and that obtained by majority voting is used for querying images. However, this method is fundamentally different from how we use the query by committee approach in our case.

Now we discuss a few related works that have explored the problem of active learning for object detection. In [27], the authors use the simple margin approach of [26] in sub-linear time with a hashing-based solution. [34] uses humans to correct annotations proposed by a detector for that unannotated image which has the maximum predicted annotation cost. Unlike [34], in the proposed approach, the annotation costs are assumed to be equal for every image. Active learning should not be confused with active detection methods like [3] where the discriminative ability of an already trained base classifier is refined at test time using minimum human supervision. It should be noted that all these methods use non-deep classifiers.

To the best of our knowledge, this is the first work which focuses on active learning for object detection in the deep learning area that considers a principled approach based on query by committee. We show that we can evaluate the proposals used in the image using the properties of different layers of the network to generate the set of images that need to be annotated. Further, we adopt this paradigm for SSD [17] that is currently a state-of-the-art object detection method.

3 Method

In this section, we describe our proposed active learning approaches for SSD. Initially, we randomly select a set of images from the unannotated image pool, get them annotated by an oracle (say, a human annotator) and train SSD on these newly annotated images to get an initial model. Subsequently, we actively select a fixed batch of images, get them annotated and train SSD on all the annotated images. We continue this alternative active learning/training phase till we exhaust our budget (typically, budget is a function of the total percentage of queried images) or reach the desired level of accuracy.

Here, we propose two broad classes of active learning methods:

- **Black-box methods:** These methods do not care about the underlying network architecture, rather they use the confidence scores from the softmax layer for querying images.
- **White-box methods:** These methods have knowledge of the network architecture and thus can be different for different architectures.

In the experimental section, we show that the white-box method drastically outperforms the black-box methods.

In this paper, we consider the annotation cost of each image to be equal. However, for further reduction in the actual annotation costs, methods like [20] can be used on the actively queried images. Moreover, one can use diversity measures to make the images in an actively queried batch more diverse and thus add more information. However, we do not explore these directions in the current paper and plan to explore it in future.

3.1 Black-box Active Learning

Here we use popular active learning methods [23] that have been proposed for classification in the context of object detection (multi-class, multi-label scenario). One way to adopt these methods for detection would be to choose the set of images that are hard for classification and thus could benefit from annotation. However, this approach would not consider the nature of object detection. Even if an object classifier may find an image hard to classify as the classifier looks at the whole image (for example, grass field with a small cow), an object detection algorithm may have no trouble in localizing the object of interest (the cow may easily stand out in the grass field). We, therefore, consider a set of proposals for each image and obtain the bounding boxes corresponding to these proposals that are then used with the traditional active learning methods.

We specifically consider two such black-box based methods, viz. minmax and maximum entropy. In the experimental section, we use these methods as baselines for comparison.

3.1.1 Minmax (mm)

Let us assume that for each image i, we have $\{b_{ic}\}$ bounding boxes corresponding to each class c. The query function is:

$$\underset{i}{\operatorname{argmin}} \max_{c} \max_{j \in \{b_{ic}\}} prob(j)$$

The maximum probability score a bounding box can get in an image represents how confident the model is with respect to that image. So this query function selects that image where the current model is the least confident.

3.1.2 Maximum Entropy (ent)

Let us assume that for each image i, we have $\{b_{ic}\}$ bounding boxes corresponding to each class c. The query function is:

$$\underset{i}{\operatorname{argmax}} \max_{c} \sum_{j \in \{b_{ic}\}} -prob(j) * \log(prob(j))$$

Entropy is a common uncertainty sampling measure in active learning [31]. This query function calculates the entropy of bounding boxes of a given class and subsequently uses it to determine the entropy of the entire image. The image with the maximum entropy is selected for querying.

3.1.3 Sum Entropy (ent-sum)

Let us assume that for each image i, we have $\{b_{ic}\}$ bounding boxes corresponding to each class c. The query function is:

$$\underset{i}{\operatorname{argmax}} \sum_{c} \sum_{j \in \{b_{ic}\}} -prob(j) * \log(prob(j))$$

This version of the entropy method favours images with more number of classes in it and thus provides a more aggressive baseline for our methods.

3.2 White-box Active Learning

The SSD network consists of a base network (say, VGG [25]) and a set of extra convolution layers to detect objects at different scales. A set of default bounding boxes is associated with each feature map cell and the position of each such box relative to its corresponding cell is fixed. For each image, the bounding boxes obtained from the last convolution layer of the base network and the extra convolution layers undergo non-maximal suppression and subsequently form the set of candidate bounding boxes which might contain the object of interest. An additional set of convolution layers is used, on top of these layers, to generate the per-class confidence scores and the offsets relative to each default bounding box corresponding to each feature map cell.

Our proposed approach is inspired from the paradigm of query by committee [19]. In this paper, we use the SSD network as the base classifier for active learning because of its natural extension to the query by committee approach. However, this approach can be extended to any network by attaching classification and detection heads to the 'important' convolution layers. In SSD, the last convolution layer of the base network along with the extra convolution layers form the committee of classifiers and the aggregate disagreement between them for each candidate bounding box in an image is used for querying. However, a disagreement measure should also take into account the fact that these classifiers detect objects at different aspect ratios.

The intuition behind Algorithm 1 is that if a high-scoring bounding box is found by a convolution layer and it overlaps with a ground truth, then other convolution layers will also tend to have high scoring bounding boxes in that region. This is because we have trained our model using the spatial overlap with a ground-truth bounding box as a loss criterion. We put this intuition to use by introducing the concept of a 'margin' for each bounding box b. We would only focus on those b which are neither spurious nor extremely confident.

Corresponding to each such b, we find all the neighbouring bounding boxes generated by the other convolution layers and call these auxillary bounding boxes. The margin of b is the difference of the confidence scores of b and the most confident auxillary bounding box. More the margin, more is the disagreement between the convolution layers.

The query by committee approach has strong theoretical foundations and has been shown to maximally reduce the version space with each query [24].

Algorithm 1: Proposed white-box approach

Data: current SSD model, batch size k, unannotated image pool U, set of classes C, overlap threshold j

Result: *active selection* $\subseteq U$ where $|active\ selection| = k$

1 **for** *each image i in U* **do**

2 use SSD on i to get bounding boxes B_c for each class c after non-maximal suppression;

3 **for** *each class c in C* **do**

4 **for** *each bounding box b in* B_c **do**

5 **if** *prob(b) <= 0.1 or prob(b) >= 0.9* **then**

6 continue ; `// throw out spurious or really confident bounding boxes`

7 find the convolution layer that generated it and call it s (s for source);

8 $a_l = \{\}$;

9 **for** *each convolution layer l apart from s* **do**

10 find bounding boxes from l that have more than j jaccard overlap with b and add them in A (A for auxillary bounding boxes);

11 $a_l = a_l \cup argmax_{a \in A}\ \text{prob}(a)$;

12 secondmax = $\max_{a \in a_l} \text{prob(a)}$;

13 margin[b] = prob(b) - secondmax;

14 margin[c] = $\frac{\sum_{b \in B_c} \text{margin}[b]}{|B_c|}$;

15 confidence[c] = $\max_{b \in B_c} \text{prob}(b)$;

16 margin[i] = $\sum_{c \in C} \frac{\text{confidence}[c]}{\sum_{c_1 \in C} \text{confidence}[c_1]} * \text{margin}[c]$;

17 sort images in descending order on the basis of their margin scores and select the first k corresponding images as *active selection*;

The different flavours of the above algorithm are listed in Table 1. They differ in how the class margin is calculated in step 14 from the bounding box margins (mean or sum) and how the image margin is calculated in step 16 from the class margins (sum or expectation). The sum method for calculating class margin favours images with multiple instances of the same class. However, it might also query outliers, i.e., images which contain occluded instances of a class as annotations. On the other hand, the mean method might dampen the effect of such outliers. Again, if we include class probabilities in the calculation of image margin, we pay more attention to classifier disagreements of important classes. We evaluate the variants empirically and present the results for the same.

method	class margin calculation (step 14)	image margin calculation (step 16)
white-box sum (wbs)	$\text{margin}[c] = \sum_{b \in B_c} \text{margin}[b]$ (sum)	$\text{margin}[i] = \sum_{c \in C} \text{margin}[c]$ (sum)
white-box mean (wbm)	$\text{margin}[c] = \frac{\sum_{b \in B_c} \text{margin}[b]}{\lvert B_c \rvert}$ (mean)	$\text{margin}[i] = \sum_{c \in C} \text{margin}[c]$ (sum)
white-box prob+sum (wbps)	$\text{margin}[c] = \sum_{b \in B_c} \text{margin}[b]$ (sum)	$\text{margin}[i] = \sum_{c \in C} \frac{\text{confidence}[c]}{\sum_{c_1 \in C} \text{confidence}[c_1]} * \text{margin}[c]$ (expectation)
white-box prob+mean (wbpm)	$\text{margin}[c] = \frac{\sum_{b \in B_c} \text{margin}[b]}{\lvert B_c \rvert}$ (mean)	$\text{margin}[i] = \sum_{c \in C} \frac{\text{confidence}[c]}{\sum_{c_1 \in C} \text{confidence}[c_1]} * \text{margin}[c]$ (expectation)

Table 1: ***Different flavours of the white-box approach***

3.3 Exploration/Exploitation framework

The exploration/exploitation trade-off is an important question in the field of active learning [2]. Explorative methods explore the feature space to find an unlabeled sample which is 'far' from the labeled samples while exploitative methods focus on an area which is already populated with labeled samples. The explorative method might select outliers, while the exploitative method might select 'similar' samples which will not improve the classifier. For example, in the case of Algorithm 1, selection of the highest scored images represent exploration as the convolution layers disagree on them the most and hence they are likely to be 'far' away from the currently labeled images. On the other hand, the selection of the lowest scored images represent exploitation as they are 'similar' to the currently labeled image pool.

In this paper, we provide the following simple yet effective framework to leverage the power of exploration/exploitation:

n bins formulation (*n-ee*): For each batch, we divide the score space into n equal-sized bins. If the batch size is m, we query $\frac{m}{n}$ highest scored images from the top (n-1) bins (exploration in each bin) and $\frac{m}{n}$ lowest scored images from the last bin (exploitation). So the resultant batch contains images from different regions of the score space and can be more representative of the underlying distribution. In the experiments, we use 2, 5 and 10 as the values of n.

4 Results

We use the SSD300 architecture with 'image expansion data augmentation' trick [17] for our experiments[1]. We evaluate our methods on the PASCAL VOC2007+12 [5] and Kitti [6] datasets. The different active learning methods are compared by evaluating the models on the corresponding test sets using mean Average Precision (mAP) [5]. We use 0.3 as the overlap threshold for all the white-box methods.

4.1 PASCAL VOC

We combine the VOC2007 and VOC2012 trainval sets [5] to create an unannotated image pool U of size 16551. The initial SSD model is trained by choosing a random set of 10% images from U and getting them annotated. This initial model is used for all the active learning methods. Subsequently, we actively select a batch of 5% (of total) images from the remainder of U and get them annotated. This alternative training/active learning is continued till we have trained the models on 45% of the total size of U. The model is tested on the

[1] The pytorch code for SSD is at: https://github.com/amdegroot/ssd.pytorch

VOC2007 test set, which contains 4952 images, using the standard mean Average Precision (mAP) measure.

The model training parameters are kept same for all the active learning methods, i.e, the initial learning rate is 1e-3 or 1e-4 depending on the training loss, the number of iterations is 120000, the momentum is 0.9, the weight decay for SGD is 5e-4 and the learning rate is decayed by a factor of 10 after 80000, 100000 and 120000 iterations.

4.1.1 Comparison with different flavours of white-box

We start off by comparing the different flavours of the white-box approach. In Fig.1(a), we find that *wbps* and *wbpm* outperform *wbs* and *wbm* respectively. This shows the importance of including class probabilities in the calculation of image margin.

4.1.2 Comparison with black-box methods

Next we compare the *wbps* and *wbpm* flavours against the baseline active learning methods. In Fig. 1(b), we find that *wbps* and *wbpm* outperform the baseline methods. The *ent-sum* method performs better than the *ent* method because of its preference for images with more classes in them and is the most aggressive baseline method.

4.1.3 Comparison with exploration/exploitation methods

In Fig. 1(c), we compare *wbs* and *wbm* with their respective exploration/exploitation versions having 2 bins *wbs(2-ee)* and *wbm(2-ee)*. It is clear from the figure that adding exploitation improves results in many iterations. Next in Fig. 1(d) and Fig. 1(e), we compare the respective exploration/exploitation versions of *wbs* and *wbm* having 2, 5 and 10 bins. In both the versions, methods *2-ee* and *5-ee* outperform *10-ee*. In Fig. 1(e), we find that our method of exploration in all bins, except the last, and exploitation in the last bin (*n-ee*) improves the results over exploration in all bins (*n-e*).

4.1.4 mAP of different methods using 35% images:

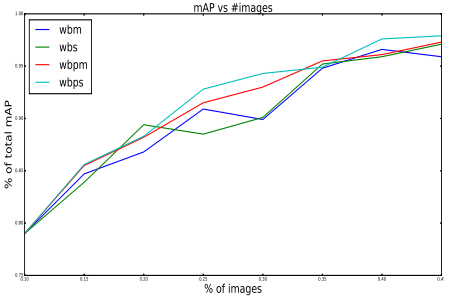
We choose to display our results at 35% because all active learning methods have sampled enough images to learn effectively by that time. So even a minor difference of mAP is significant. In Table 2, we find that our best method outperforms the most aggressive baseline method *ent-sum* by more than 1.5 mAP points, which is quite significant at this stage. We also use random sampling as a baseline where we randomly sample 25% images on top of the 10% initial images. The random sampling process is repeated 6 times and as expected, produces results with high variance. Moreover, in unbalanced datasets like Pascal VOC, the random sampling method is expected to pick images of bigger classes over images of smaller classes, thus providing inferior performance for smaller classes.

4.2 Kitti

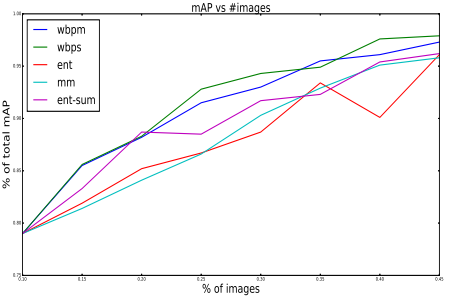
The practical use of active learning methods is in applications where redundancy is common like in surveillance videos and autonomous driving. To understand how our methods perform in such scenarios, we select the Kitti dataset [6] and divide the training set equally into train and test set. In this section, we evaluate the performance of white-box method *wbpm* and the most aggressive baseline method *ent-sum*, trained on 35% of images, on the test set using Average precision (AP) measure [5]. We do not leave out 'difficult' annotations for this

analysis. The model training parameters are kept same for both the active learning methods, i.e, the initial learning rate is 1e-4, the number of iterations is 120000, the momentum is 0.9, the weight decay for SGD is 5e-4 and the learning rate is decayed by a factor of 10 after 80000, 100000 and 120000 iterations. As before, we start off with the initial random sampling of size 10% and then perform active learning with a batch size of 5%.

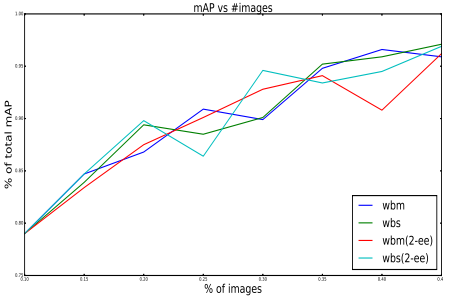
As is clear from Fig. 2, the *wbpm* method drastically outperforms the baseline method *ent-sum*.



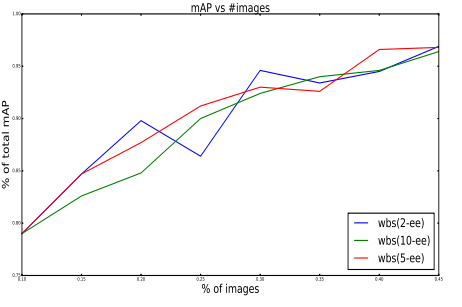
(a) Comparison among white-box methods



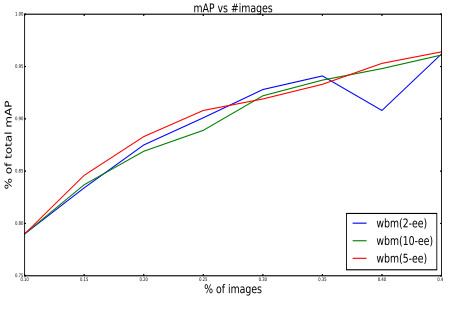
(b) Comparison with black-box methods



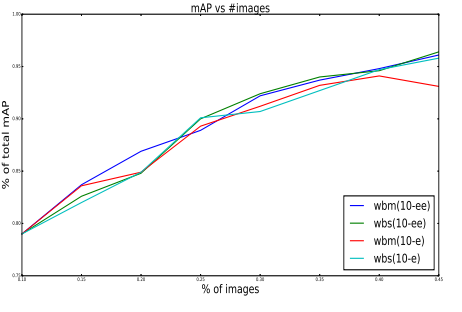
(c) Comparison with ee methods having 2 bins



(d) Comparison among sum ee methods



(e) Comparison among mean ee methods



(f) Does exploitation in last bin help?

Figure 1: ***% of full supervision mAP at different percentages for different methods on VOC classes***

Baselines using 35% images				full mAP[▭]	White-box using 35% images				Exploration/Exploitation using 35% images			
mm	ent	ent-sum	random		wbs	wbm	wbps	wbpm	wbs(2-ee)	wbm(2-ee)	wbs(5-ee)	wbm(5-ee)
71.69	72.11	71.28	70.26 ± 1.59	77.2[2]	73.46	73.18	73.29	73.72	72.13	72.65	71.50	72.01

Table 2: ***mAP on VOC classes using 35% images***

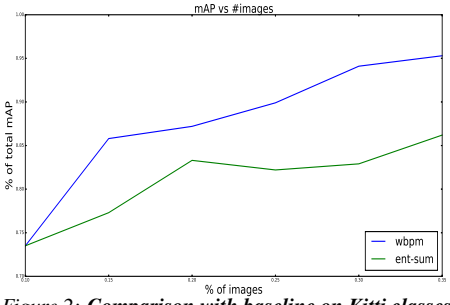


Figure 2: ***Comparison with baseline on Kitti classes***

metric \ method	wbpm	ent-sum
class content	2832	2885
object content	7812	9237
truncated object content	76	68

Table 3: ***Analysis on Kitti classes***

ent-sum	full mAP	wbpm
49.72	57.67	54.94

Table 4: ***mAP on Kitti classes using 35% images***

method \ metric	class content	object content	truncated object content
wbpm	7775	16957	8277
wbps	7941	20013	10081
ent-sum	8136	20546	10966

Table 5: ***Analysis on VOC classes***

5 Model Analysis

In this section, we provide a short analysis of the white-box methods *wbpm* and *wbps* and the baseline method *ent-sum*. In Table 5, we present the total number of classes (class content), the total number of objects (object content) and the total number of truncated objects (truncated object content) in the images queried by the various active learning methods going from 10% of training images to 35% of training images on VOC2007+12 dataset. It is clear that the baseline method *ent-sum* queries more images with multiple classes and objects in them than the white-box method *wbpm*. This shows that blindly querying images with more class and object content does not ensure a superior performance. Moreover, the table also shows that the *wbps* method has a tendency to query images with more objects in them, thereby, pulling in significantly more truncated objects than *wbpm*. This is a consequence of the sum function which favours images with more objects in them. But, interestingly, its performance is hardly affected by the sheer increase in the number of truncated objects. As is evident from Table 3, this discrepancy between the performance of the querying strategy and the object/class content in queried images is even more pronounced in the Kitti dataset.

6 Conclusions and Future work

In this paper, we have proposed active learning approaches that produce state-of-the-art results in object detection using only a fraction of the training images. Our methods are really useful in scenarios where obtaining annotations is a costly affair. Our active learning strategy exploits the difference in detection between convolution layers and thus uses a 'query

[2]The code we used gives fractionally higher mAP

by committee' approach for annotating images. In future we would like to further reduce the annotation effort by making our batches more diverse and using methods like [20] on top of our querying strategies. Another interesting direction which we would like to explore is the combination of semi-supervised techniques with active learning methods for various computer vision tasks.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[2] A. Bondu, V. Lemaire, and M. BoullÃl'. Exploration vs. exploitation in active learning : A bayesian approach. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, July 2010. doi: 10.1109/IJCNN.2010.5596815.

[3] Yuxin Chen, Hiroaki Shioi, Cesar Fuentes Montesinos, Lian Pin Koh, Serge Wich, and Andreas Krause. Active detection via adaptive submodularity. In *Proc. International Conference on Machine Learning (ICML)*, June 2014.

[4] Melanie Ducoffe. Active learning strategy for cnn combining batchwise dropout and query-by-committee. 2017.

[5] Mark Everingham, Luc Van Gool, ChristopherK.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL http://dx.doi.org/10.1007/s11263-009-0275-4.

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Ross Girshick. Fast r-cnn. In *International Conference on Computer Vision (ICCV)*, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[10] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, June 2008. doi: 10.1109/CVPRW.2008.4563068.

[11] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2372–2379, June 2009. doi: 10.1109/CVPR.2009.5206627.

[12] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *Int. J. Comput. Vision*, 88(2):169–188, June 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0268-3. URL http://dx.doi.org/10.1007/s11263-009-0268-3.

[13] Vasiliy Karasev, Avinash Ravichandran, and Stefano Soatto. Active frame, location, and detector selection for automated and manual video annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[15] Christoph H. Lampert and Jan Peters. *Active Structured Learning for High-Speed Object Detection*, pages 221–231. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-03798-6. doi: 10.1007/978-3-642-03798-6_23. URL http://dx.doi.org/10.1007/978-3-642-03798-6_23.

[16] Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 859–866, Washington, DC, USA, 2013. IEEE Computer Society. ISBN 978-0-7695-4989-7. doi: 10.1109/CVPR.2013.116. URL http://dx.doi.org/10.1109/CVPR.2013.116.

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed. Ssd: Single shot multibox detector.

[18] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2839–2847, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.325. URL http://dx.doi.org/10.1109/ICCV.2015.325.

[19] Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 74–, New York, NY, USA, 2004. ACM. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015385. URL http://doi.acm.org/10.1145/1015330.1015385.

[20] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. *CoRR*, abs/1602.08405, 2016. URL http://arxiv.org/abs/1602.08405.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[23] Burr Settles. Active learning literature survey. Technical report, 2010.

[24] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi: 10.1145/130385.130417. URL http://doi.acm.org/10.1145/130385.130417.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

[26] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760185243. URL http://dx.doi.org/10.1162/153244302760185243.

[27] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1449–1456, June 2011. doi: 10.1109/CVPR.2011.5995430.

[28] Sudheendra Vijayanarasimhan and Kristen Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision*, 91(1):24–44, 2011. ISSN 1573-1405. doi: 10.1007/s11263-010-0372-4. URL http://dx.doi.org/10.1007/s11263-010-0372-4.

[29] Sudheendra Vijayanarasimhan and Kristen Grauman. *Active Frame Selection for Label Propagation in Videos*, pages 496–509. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-33715-4. doi: 10.1007/978-3-642-33715-4_36. URL http://dx.doi.org/10.1007/978-3-642-33715-4_36.

[30] Carl Vondrick and Deva Ramanan. Video Annotation and Tracking with Active Learning. In *Neural Information Processing Systems (NIPS)*, 2011.

[31] Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119. IEEE, 2014.

[32] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia*, PP(99):1–1, 2016. ISSN 1520-9210. doi: 10.1109/TMM.2016.2602938.

[33] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

[34] A. Yao, J. Gall, C. Leistner, and L. van Gool. Interactive object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3249, Providence, RI, USA, 2012. IEEE.