Data Subset Selection for Object Detection

Jiaqi Fan Junxin Huang Xiaochuan Yu Chao He
Wisers AI Lab
Hong Kong
{garyfan,gabrielhuang,hillyu,chaohe}@wisers.com

Abstract

In light of the exploding cardinality of common object detection datasets, and the possibility of joining multiple general available datasets, it is of great importance to select a minimum subset that is representative and effective enough for training without suffering from a significant performance drop. We extend the previous work on subset selection for classification tasks to object detection and propose a novel supervised data subset selection method - uncertainty adjusted Term Frequency-Inverse Document Frequency (TF-IDF) sampling - that selects data based on informativeness while ensuring representativeness. The proposed method aims to address 1. the diminishing return of large scale object detection data. 2. the inherent dataset class imbalance due to the real-world representation disparity of different classes. 3. the lack of representativeness induced by the active learning based subset selection approaches.

1 Introduction

Recent researches have been aiming to advance object detection models' performance and capability. The race in performance leads to the vast adoption of computation hungry detection models, typically requiring 8 Tesla V100s training for days on the MSCOCO dataset [1]. To increase the model capacity, large-scale object detection datasets like Object365 [2] and Open Images [3] have emerged with an order of magnitude more images than the MSCOCO, drastically increasing the computation cost of model training. At the same time, there are efforts of combining multiple datasets into a unified training set to enhance further the representability and diversity of the training set [4].

With the vast imbalance comes the challenge in the degraded performance in minority classes. Past Open Image Competitions [5, 6] have empirically verified that large scale object detectors might exhibit heterogeneous performance across different categories. The models trained on Open Images lack in both localization and classification accuracy on poorly represented categories, which is likely a result of the intrinsic imbalanced class distributions [5, 6].

These challenges call for a reliable data subset selection algorithm scalable to the enormous object detection datasets. In this study, we will thoroughly investigate different subset selection approaches and

- devise a data subset selection algorithm combining active learning with term frequency-inverse document frequency (TF-IDF);
- investigate the performance gain (under budget-restrained settings) over training an object detector using the full dataset.

Pre-registration workshop NeurIPS (2020), Vancouver, Canada.

2 Related work

2.1 Subset selection

Prior work on subset selection has resorted to coresets, submodular functions, and active learning. Coresets aim to produce subsets that best preserve geometric structures, on which the clustering algorithms such as SVM can achieve competitive performance. Recent researches [7, 8] have formulated subset selection process as constrained submodular optimization. Wei *et al.* [7] established connections between two machine learning classifiers, Naive Bayes classifiers and Nearest Neighbour classifiers, and submodular function, making subset selection for these classification tasks viable through greedy optimization [9]. Settles [10] did preliminary studies on active learning for subset selection by iteratively sampling the most informative data. Wei *et al.* [7] proposed filtered active submodular selection (FASS), a form of multi-stage mini-batch active learning, for subset selection and achieved admirable results in text categorization and handwritten digits recognition. Later work [8] extended FASS to image classification tasks with deep CNNs.

2.2 Near-duplicate image detection with TF-IDF

Chum *et al.* [11] extended TF-IDF to a *bag* of visual words image representations for near-duplicate image detection with min-hash algorithms. It proposed a novel similarity measure combining TF-IDF with SIFT features that performs well in a large-scale news video dataset and an image retrieval dataset.

2.3 Safe-screening and determinantal point process

Safe screening, pioneered in [12], concerns safely removing non-active features before or during optimization without incurring false negatives. Shibagaki *et al.* [13] extend [12] by jointly removing non-active features and data samples, while Mialon *et al.* [14] investigate a relaxation of the strongly convex objective required by [13]. It does so by proposing a novel Ellipsoid method compared to the ball-like safe region employed by [13].

Assuming negative correlation [15] or pairwise repulsion [16] in the ground set, determinantal point process (DPP) forms a probability measure over all subsets and aims to produce diverse samples by assigning higher probabilities to sets consisting of dissimilar items.

2.4 Active learning

Active learning taps into the issue of limited labeled data. It has been widely used in classification tasks [17–21] and is recently applied to object detection [22, 23].

At its core, active learning assumes that not all data contribute equally to supervised learning [24]. It trains a base model on the labeled data, which, with the help of score functions, is subsequently used to decide which set of unlabeled data to be labeled first by an oracle.

Common score functions evaluate samples by informativeness, reflected by the uncertainty of model decisions, and often measured by the entropy of predicted results. It is prone to a lack of representativeness, which motivates researchers to design solutions with both informativeness and representativeness in mind [7, 25, 26], but no such study has been conducted in the object detection setting.

3 Problem statement

Consider an object detection dataset V consisting of the set of images $D = \{ I_i \mid i \in \{1, \ldots, N_I\} \}$ and the set of bounding box annotations A, where I_i refers to the i_{th} image. Hence, N_I refers to the total number of images in the dataset. $A = \{ (b_{ij}, c_{ij}) \mid i \in \{1, \ldots, N_I\},\ j \in \{1, \ldots, n_i\} \}$, where $b_{ij} = (x_{ij}, y_{ij}, w_{ij}, h_{ij})$ and $c_{ij} \in \{1, \ldots, N_C\}$ denote the bounding box coordinates and the class labels for the box j in the image I_i, respectively. N_C refers to the total number of categories and n_i refers to the total number of the boxes in the image I_i.

Given a ground set V, the data subset selection problem in the object detection context concerns selecting an optimal subset of images S out of 2^{N_I} number of subsets $S_i \subseteq D, i \in \{1, \ldots, 2^{N_I}\}$, so

that an object detector trained on the subset of annotations, $B = \{ (b, c) \mid \text{images in } S \} \subseteq A$ suffers minimal performance drop in comparison to one trained on the full annotation A.

4 Object detection data subset selection

4.1 Supervised data subset selection with TF-IDF

As is illustrated in Figure 1, the co-existence of bounding boxes and images is akin to the relationship between terms and documents.

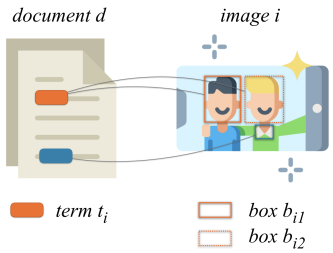

Figure 1: Term-document analogous to BBox-image

For a set of images D and its corresponding annotations A, the TF-IDF score $\omega_{i,c}$ is calculated for the bounding box instance of class c in the image i in equation 1,

$$\omega_{c,i,A,D} = tf_{c,i,A,D} * \log \frac{N_I}{df_{c,A,D}} \tag{1}$$

where $c \in \{1, \ldots, C\}$ refers to a category, $i \in \{1, \ldots, I\}$ refers to a sample image. In equation 1, $tf_{c,i}$ is the number of the bounding boxes of category c in the image i, and df_c is the number of images in D where a bounding box instance in A of category c appears. $\log \frac{N_I}{df_c}$ is commonly referred to as Inverse Document Frequency (IDF).

Given the TF-IDF of its bounding box instances, an image is ranked within the set of images D based on the sum of TF-IDF scores for all the unique categories it includes, as in equation 2,

$$Q_{i,A,D} = \sum_{c' \in \{1,\ldots,C\}} \omega_{c',i,A,D} \tag{2}$$

For over-populated categories, we rank the images according to the Q and sample top-t images, as is outlined by algorithm 1.

Algorithm 1: TF-IDF based training data subset selection with threshold t

Data: annotations A, images D, threshold t, category id c
Result: the selected subset of annotations B

$c \leftarrow 0$, $B \leftarrow \{\}$, $D_c \leftarrow$ all images involving category c, $A_{D_c} \leftarrow$ all annotations for the set of images $D_c \subseteq D$;
while $c \neq C$ **do**
  if *the number of images involving category* $c > t$ **then**
    $\overline{D_c} \leftarrow D_c$ sorted in descending order by Q_{i,A_{D_c},D_c}, $i \in D_c$;
    $B \leftarrow B \cup A_{top(\overline{D_c},t)}$, $top\left(\overline{D_c}, t\right) \leftarrow$ top t images in $\overline{D_c}$;
  else
    $B \leftarrow B \cup A_{\overline{D_c}}$;
  end
  $c \leftarrow c + 1$;
end

4.2 Unsupervised data subset selection with Active Learning

We extend works in submodular active learning [27, 7, 8] to object detection data subset selection. Initially trained on a small subset of labeled training data, an object detection model is used to run inference on the remaining images and produce image scores using various scoring functions [23]. Top K scored images and their predicted labels are then selected and added to the initial training data to train an improved object detection model. This enrolment procedure is repeated recursively until the desired subset is produced. The greedy optimization process arrives at a good approximation of the optimal solution and guarantees the lower bound of $\frac{e-1}{e}$ under submodularity set functions [28].

Score functions usually measure the uncertainty of the predicted instances. Score functions using the entropy, mutual information (MI), and the entropy of the predicted bounding boxes (Det-Ent) are investigated in [23], and the sum of entropy of bounding boxes achieves the best performance in object detection. Adopting the entropy of bounding boxes as the score function and summation as the score aggregation function for each image, the active learning assisted subset selection is summarized in algorithm 2.

Algorithm 2: Subset selection using active learning under submodularity constraint

Data: initial subset of labeled image S_0, initial annotations B_0, iterative enrollment K
Result: the selected subset of annotations B

$n \leftarrow 1$, $N \leftarrow$ total iterations;
while $n \neq N$ **do**
    train an object detection model O_n on B_{n-1};
    perform inference on $B \setminus B_{n-1}$ using O_n and calculate scores for $S \setminus S_{n-1}$;
    sort images in $S \setminus S_{n-1}$ by the aggregated scores;
    $S_n \leftarrow S_{n-1} \cup top\left(S \setminus S_{n-1}, K\right)$, $B_n \leftarrow$ box annotations in image set S_n;
    $n \leftarrow n + 1$;
end

4.3 Supervised data subset selection with uncertainty adjusted TF-IDF

The innate lack of representativeness of uncertainty-sampling based subset selection manifests itself in the potentially imbalanced subset sampled by the active learning based approaches. As TF-IDF incorporates representativeness by design, we propose a novel subset selection method, uncertainty adjusted TF-IDF sampling, baking in uncertainty through active learning while taking advantage of the statistical distribution of the full data.

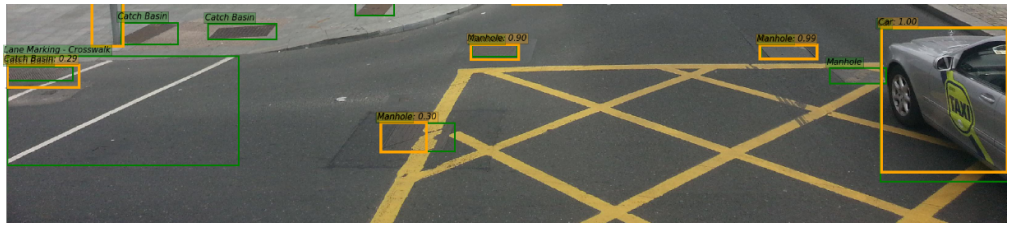

Figure 2: The inference results of an image in the MVD [29] dataset. The dark yellow boxes visualize model predictions. The dark green boxes are ground truth labels. Note the false negative objects: “Catch Basin", “Manhole", and “Lane Marking - Crosswalk" will cause the score of the image to be undervalued.

Motivation. Brust *et al.* [30] empirically showed that active learning is susceptible to class imbalance. In the active learning setting, the aggregated score of the image shown in Figure 2 will be underestimated, leading to poor representation in the selected subset of the classes where the detection model often fails, which would, in turn, hamper the performance of the models trained on the subset.

To deal with the under-represented classes, we propose to utilize the global representation embedded in the TF-IDF score and dynamically adjust TF-IDF with the class performance of the active learning model. Notably, the uncertainty adjusted TF-IDF is calculated on the full annotation A rather than

the subset B to avoid worsening the insufficient representation during active learning. Thus, the uncertainty adjusted TF-IDF preserves the representativeness while utilizing informativeness.

Credibility adjusted TF-IDF. In NLP, to account for certain tokens having a greater influence on the document classification result, the credibility-adjusted TF-IDF [31] resorts to assign more weights to higher impact tokens. The credibility-adjusted term frequency (TF) takes the form of equation 3,

$$\overline{tf}_{i,d} = (0.5 + \hat{s_i}) * tf_{i,d} \tag{3}$$

while the IDF remains as is. $\hat{s_i}$ is a smoothed credibility score derived from Buhlmann credibility adjustment [32], measuring the effect token i has on the binary classification of document d.

Uncertainty adjusted TF-IDF. Inspired by credibility adjusted TF-IDF in the text classification settings, the uncertainty adjusted TF-IDF is formulated in equation 4,

$$\omega'_{c,i,A,D} = (1.5 - \frac{AP_{c,B}}{2}) * tf_{c,i,A,D} * \log \frac{N_I}{df_{c,A,D}} \tag{4}$$

where $AP_{c,B} \in [0, 1]$ denotes the mean Average Precision (mAP) on class c of a model trained on the annotations B. It is re-calibrated to assign greater weights to poorly performed classes.

The subset selection process follows that of algorithm 2 except step 5 and 6. After training an object detection model, its performance (AP) is evaluated on the validation set and used for calculating ω'. With the aggregation function in equation 2, we derive the adjusted aggregated score Q_i for each image. At each iteration, the images are sampled in descending order of the aggregated scores.

5 Experiments Protocol

5.1 Datasets and implementation details

We plan to perform experiments on OID for its extensive coverage and well-documented benchmark results. COCO and MVD will also be included. Regarding detectors, we will adopt Faster-RCNN [33] FPN [34] with backbone ResNet50 [35]. For model training, we will adopt SGD with warm-up, an initial learning rate of 0.00125 * batch size, cosine learning rate schedule, and 12 total epochs per iteration. Other hyperparameters will follow related work.

5.2 Subset selection methods and baselines

To compare different subset selection methods, we train models on the subset selected by TF-IDF, uncertainty adjusted TF-IDF, and active learning. The initial seeded dataset for active learning is randomly selected.

At iteration n of each subset selection method, an identical model pre-trained on MSCOCO will be trained on the subset obtained at iteration n for 12 epochs, simulating a limited budget. After the training is completed, the model will perform inference on the remaining training data $S \setminus S_{n-1}$ and enroll K images to form the desired subset S_n. The testing results on the held-out test set across different subset selection methods per iteration will be updated to Table 1 and plotted in Figure 3a.

(a) The training loss and validation accuracy.

(b) Performance against run time.

Figure 3: Plots visualizing convergence and run time.

Table 1 will also include the performance of a baseline model trained on the full training data, denoted *full baseline*, and a *random baseline*, where images are randomly selected at each iteration. Additional

Subset selection methods	mAP at subset selection iteration n 1	2	3	4	N (full data)
random baseline					
active learning (Det-Ent)					
tf-idf					
uncertainty-adjusted *tf-idf*					
DPP					
full baseline	N/A	N/A	N/A	N/A	

Table 1: To be reported: the performance (mAP) on the OID held-out validation set.

	mAP
random	
tf-idf	

(a) **Initial training data - selection approach**

	mAP
Det-Ent	
total loss	

(b) **Score functions**

	mAP
seed 1	
seed 2	
seed 3	
mean	
std	

(c) **Initial data seeds - active learning**

	mAP
seed 1	
seed 2	
seed 3	
mean	
std	

(d) **Initial data seeds - adj. TF-IDF**

Table 2: Ablation studies.

subset selection experiments using determinantal point process will also be conducted, resulting in a total of 11 runs.

The inclusion of baseline experiments will give us a better idea regarding how the proposed methods affect the performance given limited computational resources. To illustrate this, we will plot Pareto curves depicting performance over total training and subset selection time in Figure 3b.

5.3 Ablation studies

Initial seeded training data. To investigate the impact the initial seeded training data has on the performance of active learning, we compare the performance with randomly selected initial data against TF-IDF sampled initial training data in Table 2a.

Score functions for active learning. We opt to have total loss as an alternative score function to the active learning approach with Det-Ent score function (Table 2b).

Varied random seeds. To account for the variance induced by randomness, we will perform multiple runs with different random seeds (Table 2c and 2d).

6 Discussions

This paper proposes a novel supervised data subset selection method that selects data based on informativeness while ensuring representativeness. The planned experiments will provide new insights into budget-limited imbalanced object detection.

References

[1] Lin, T.-Y., M. Maire, S. Belongie, et al. Microsoft coco: Common objects in context, 2015.

[2] Shao, S., Z. Li, T. Zhang, et al. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 8430–8439. 2019.

[3] Kuznetsova, A., H. Rom, N. Alldrin, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[4] Zendal, O., H. A. Alhaija, R. Benenson, et al. Robust vision challenge 2020, 2020.

[5] Liu, Y., G. Song, Y. Zang, et al. 1st place solutions for openimage2019–object detection and instance segmentation. *arXiv preprint arXiv:2003.07557*, 2020.

[6] Akiba, T., T. Kerola, Y. Niitani, et al. Pfdet: 2nd place solution to open images challenge 2018 object detection track. *arXiv preprint arXiv:1809.00778*, 2018.

[7] Wei, K., R. Iyer, J. Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963. 2015.

[8] Kaushal, V., R. Iyer, S. Kothawade, et al. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299. IEEE, 2019.

[9] Aardal, K., C. Van Hoesel. Polyhedral techniques in combinatorial optimization i: Theory. *Statistica Neerlandica*, 50(1):3–26, 1996.

[10] Settles, B. Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 2009.

[11] Chum, O., J. Philbin, A. Zisserman, et al. Near duplicate image detection: min-hash and tf-idf weighting. In *BMVC*, vol. 810, pages 812–815. 2008.

[12] Ghaoui, L. E., V. Viallon, T. Rabbani. Safe feature elimination for the lasso and sparse supervised learning problems, 2011.

[13] Shibagaki, A., M. Karasuyama, K. Hatano, et al. Simultaneous safe screening of features and samples in doubly sparse modeling. vol. 48 of *Proceedings of Machine Learning Research*, pages 1577–1586. PMLR, New York, New York, USA, 2016.

[14] Mialon, G., J. Mairal, A. d'Aspremont. Screening data points in empirical risk minimization via ellipsoidal regions and safe loss functions. In S. Chiappa, R. Calandra, eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, vol. 108 of *Proceedings of Machine Learning Research*, pages 3610–3620. PMLR, Online, 2020.

[15] Kulesza, A., B. Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[16] Cho, S., L. Lebanoff, H. Foroosh, et al. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. *arXiv preprint arXiv:1906.00072*, 2019.

[17] Holub, A., P. Perona, M. C. Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. 2008.

[18] Joshi, A. J., F. Porikli, N. Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. 2009.

[19] Kapoor, A., K. Grauman, R. Urtasun, et al. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, 2010.

[20] Li, X., Y. Guo. Adaptive active learning for image classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–866. 2013.

[21] Vijayanarasimhan, S., K. Grauman. Cost-sensitive active visual category learning. *International Journal of Computer Vision*, 91:24–44, 2011.

[22] Roy, S., A. Unmesh, V. P. Namboodiri. Deep active learning for object detection. In *BMVC*, page 91. 2018.

[23] Haussmann, E., M. Fenzi, K. Chitta, et al. Scalable active learning for object detection. *arXiv preprint arXiv:2004.04699*, 2020.

[24] Lapedriza, A., H. Pirsiavash, Z. Bylinskii, et al. Are all training examples equally valuable?, 2013.

[25] Xu, Z., K. Yu, V. Tresp, et al. Representative sampling for text classification using support vector machines. In *European conference on information retrieval*, pages 393–407. Springer, 2003.

[26] Huang, S.-J., R. Jin, Z.-H. Zhou. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, pages 892–900. 2010.

[27] Gygli, M., H. Grabner, L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3090–3098. 2015.

[28] Nemhauser, G. L., L. A. Wolsey, M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

[29] Neuhold, G., T. Ollmann, S. R. Bulò, et al. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009. 2017.

[30] Brust, C.-A., C. Käding, J. Denzler. Active learning for deep object detection, 2018.

[31] Kim, Y., O. Zhang. Credibility adjusted term frequency: A supervised term weighting scheme for sentiment analysis and text classification. *arXiv preprint arXiv:1405.3518*, 2014.

[32] Bühlmann, H., A. Gisler. *A course in credibility theory and its applications*. Springer Science & Business Media, 2006.

[33] Ren, S., K. He, R. Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[34] Lin, T.-Y., P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection, 2017.

[35] He, K., X. Zhang, S. Ren, et al. Deep residual learning for image recognition, 2015.