Benchmarking Sampling-based Probabilistic Object Detectors

Dimity Miller, Niko Sünderhauf, Haoyang Zhang, David Hall, Feras Dayoub
Australian Centre for Robotic Vision, Queensland University of Technology (QUT)
dimity.miller@hdr.qut.edu.au

Abstract

This paper provides the first benchmark for sampling-based probabilistic object detectors. A probabilistic object detector expresses uncertainty for all detections that reliably indicates object localisation and classification performance. We compare performance for two sampling-based uncertainty techniques, namely Monte Carlo Dropout and Deep Ensembles, when implemented into one-stage and two-stage object detectors, Single Shot MultiBox Detector and Faster R-CNN. Our results show that Deep Ensembles outperform MC Dropout for both types of detectors. We also introduce a new merging strategy for sampling-based techniques and one-stage object detectors. We show this novel merging strategy has competitive performance with previously established strategies, while only having one free parameter.

1. Introduction

With the ability to localise and classify multiple objects in a scene, object detectors are crucial perception modules for robotic systems. For safe and robust use, an object detector should express a measure of uncertainty [1, 7, 16, 17]. This uncertainty should be indicative of detection accuracy, where a higher spatial or label uncertainty is correlated with inaccuracy in localisation or semantic classification. This is a *probabilistic object detector*, where a continuous uncertainty measure is produced for all detections and indicates potential inaccuracies in performance.

To date, no comparative evaluation of probabilistic object detectors has been performed. Previously, the object detection literature has only explored uncertainty for distinguishing between true positive and false positive detections [3, 5, 9, 13], or has only addressed label uncertainty and ignored spatial uncertainty [3].

We provide the first analysis of probabilistic object detectors that express spatial uncertainty and label uncertainty for all detections. We compare two popular uncertainty techniques, Monte Carlo (MC) Dropout [2] and Deep Ensembles [8], when implemented into one-stage and two-stage object detectors, Single Shot MultiBox Detector (SSD) [11] and Faster R-CNN [14]. We show that Deep Ensembles outperform the MC Dropout technique for spatial and label uncertainty quality. We also introduce a novel merging strategy that is suitable for one-stage object detectors and sampling-based uncertainty techniques, and only relies on one hyperparameter.

2. Literature Review

Probabilistic Object Detection: Probabilistic Object Detection has recently been formally defined in [4]. A probabilistic object detector produces a bounding box that is represented by normally distributed corners, and a full label distribution. The uncertainty in this output should be correlated with detection accuracy; high spatial uncertainty (high-variance corner distributions) and high label uncertainty (uniform label distributions) should indicate potential inaccuracies in localisation and classification. Probability-based Detection Quality (PDQ) has been introduced as an object detection measure that evaluates the quality of both label and spatial uncertainty, as well as general detection performance [4]. We use PDQ as the primary evaluation measure in our paper.

Uncertainty Techniques for Object Detection: Uncertainty in object detection has primarily been used to identify and discard false positive detections (which may include out-of-distribution detections, misclassified detections and spatially inaccurate detections) [3, 5, 13]. Both [13] and [3] propose uncertainty-expressing object detectors, but do not evaluate the quality of the spatial uncertainty of their detectors. A new bounding box regression loss proposed in [6] allows the detector to predict its localisation variance. This was shown to improve Average Precision performance, but did not address label uncertainty [6]. Harakeh et al. [5] evaluate the quality of their detector's spatial and label uncertainty, though only for the use of removing false positive detections in a simplified two-class problem (car and pedestrians).

In contrast to all of the above, our work evaluates the performance of object detectors for *probabilistic object detection*, where a meaningful label and spatial uncertainty must be produced for all detections.

3. Sampling Probabilistic Object Detectors

Our approach to probabilistic object detectors is summarised in two stages: (1) implementing the sampling-based uncertainty techniques into existing object detectors and (2) forming probabilistic object detections with the existing merging strategy established in prior work and our proposed novel merging strategy.

3.1. Sampling-based Uncertainty Techniques

Each of the sampling-based uncertainty techniques outputs a set of samples, where each sample contains a set of detections. Each detection comprises bounding box coordinates and a distribution of softmax scores for the known classes.

Monte Carlo (MC) Dropout [2] is a variational inference technique approximating a Bayesian Convolutional Network with a Bernoulli prior distribution over its weights. By retaining Dropout layers [15] while testing, multiple forward passes can be performed to sample from the posterior distribution over the weights and can be used to represent uncertainty. Each forward pass represents a sample.

Deep Ensembles [8] proposes to train an ensemble of networks, each with random initialisations of the network weights and random shuffling of the data during training. Each network is expected to behave differently for inputs that are not represented by the training data, thus expressing uncertainty. Each network's output represents a sample.

3.2. Merging Strategies

As described in [12], when using a sampling-based uncertainty technique, detections from each sample must be correctly associated to form a probabilistic detection $\mathbf{O_j}$ of objects in the scene.

Established Technique: Miller et al. [12] evaluated a range of affinity measures and clustering techniques when using MC Dropout with SSD. They found that a Basic Sequential Algorithmic Scheme (BSAS) clustering method with Intersection over Union (IoU) and 'Same Label' affinity measures produced probabilistic detections with the most meaningful uncertainty quality. This technique has three hyperparameters: minimum IoU threshold, minimum number of detections per probabilistic detection and minimum non-background softmax score for a detection. We refer to this as the *established technique* for the remainder of the paper.

Pre-NMS Averaging: We introduce a novel technique for merging detection samples into probabilistic detections that can be applied to one-stage object detectors and sampling-based uncertainty techniques. This merging strategy has one hyperparameter only – the minimum softmax score for a detection to be considered valid.

A one-stage object detector outputs a bounding box regression and classification distribution for a predefined set of anchor boxes and aspect ratios. These form a *fixed* number of detections D_i, each with an index i representing the anchor box they were spawned from. Typically, these detections are passed through Non-Maximum Suppression (NMS) to suppress redundant detections. This produces a variable number of output detections, which must then be correctly associated to form probabilistic detections.

We propose to average the D_i for all samples at index i *before* applying NMS. This eliminates the need to correctly associate a variable number of samples, as detections from each sample can already be associated with i. We then apply NMS to these averaged detections and observe the indexes i's of the average detections that survive NMS. For each index i that survives NMS, a probabilistic object detection $\mathbf{O_j}$ is then formed from the average label distribution and average bounding box, with the covariance of bounding box corners extracted from the pre-NMS D_i from all samples.

4. Evaluation

We tested a two-stage detector, Faster R-CNN [14] and a one-stage detector, Single Shot MultiBox Detector [11]. To implement Deep Ensembles, we trained an ensemble of 5 detectors with randomly initialised weights and random shuffling of data during training, as suggested by [8]. To implement MC Dropout into SSD, we followed the implementation in [12, 13], where two dropout layers with 0.5 probability are placed on Conv6 and Conv7. For Faster R-CNN, we used a VGG16 backbone, and placed a dropout layer on the last three convolutional layers of VGG16 with probability 0.4. 40 forward passes were used for MC Dropout detectors.

The established merging strategy [12], BSAS clustering with IoU and 'Same Label' affinity was used. 2 minimum detections per probabilistic detection and a 0.5 minimum softmax score were used, as in [13]. A 0.5 IoU threshold was empirically determined to be a high performing threshold for every probabilistic object detector.

Each probabilistic object detector was trained on the COCO 2017 training dataset, and tested on the COCO 2017 validation dataset. The COCO 2017 validation dataset was not used at all during the training process and was only used for testing.

For our main measure, we use the recently proposed Probability-based Detection Quality (PDQ) measure [4]. PDQ measures the ability of a detector to accurately classify all known objects in a dataset, however, it also rewards detections that accurately express spatial and label uncertainty for correct detections. This metric can be further decomposed into pairwise PDQ (pPDQ), which measures the uncertainty quality per detection, the label and spatial uncertainty quality (which includes foreground and background uncertainty quality), and the number of True Positives, False Positives and False Negatives. A perfect PDQ

Table 1. PDQ and mAP results for each probabilistic object detector with the established merging strategy, where Sp, Lbl, FG and BG represent spatial, label, foreground and background uncertainty quality. MCD and DE indicate MC Dropout and Deep Ensembles. Arrows indicate direction of better performance.

	mAP(%)↑	PDQ(%)↑	pPDQ(%)↑	Sp(%)↑	Lbl(%)↑	FG(%)↑	BG(%)↑	TP ↑	FP ↓	FN ↓
MCD SSD	15.78	13.116	0.475	0.402	0.733	0.730	0.579	10703	2006	26078
DE SSD	15.34	14.010	**0.531**	**0.458**	0.749	**0.880**	**0.610**	9998	**1130**	26783
MCD F-RCNN	9.85	9.607	0.385	0.317	0.636	0.646	0.504	**28227**	76311	**8554**
DE F-RCNN	**24.67**	**19.028**	0.439	0.342	**0.799**	0.738	0.506	22975	16270	13806

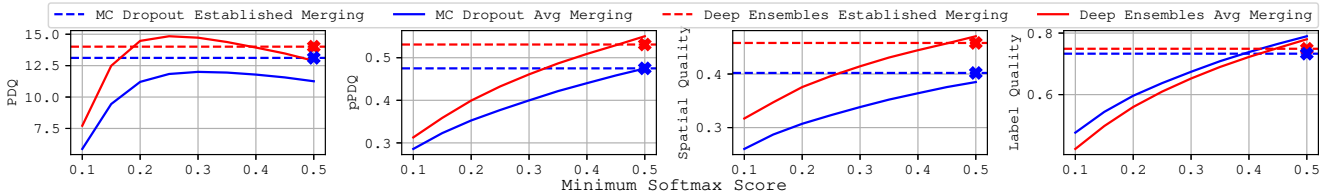

Figure 1. The effect of our pre-NMS averaging merging strategy on the components of PDQ. A cross indicates performance of the established merging strategy with a 0.5 minimum score [13] (dashed line has been added for visualisation, and does not represent testing with different minimum scores).

score is 100%. We refer the reader to the paper documenting PDQ for more information about the metric [4]. We also use the established object detection metric, mean Average Precision (mAP) [10], with a perfect score of 100%.

4.1. Results

From our experiments (summarised in Table 1) and using PDQ as our primary performance measure, we can make three main observations: (1) Deep Ensembles outperforms MC Dropout for both SSD and Faster RCNN. (2) SSD's detections have a higher spatial quality than Faster RCNN, for both uncertainty techniques. (3) Faster RCNN's recall is superior to SSD, while SSD has a significantly higher precision. We also note that our implementation of MC Dropout with the two-stage detector Faster RCNN yields an immense number of false positive detections, suggesting that MC Dropout may decrement the performance of the Region Proposal Network in a two-stage detector.

Pre-NMS Averging Merging Strategy Performance: As shown in Figure 1, we test our proposed merging strategy with varying minimum softmax scores to assess if it can feasibly obtain competitive performance with the established merging strategy. The advantage of this method is that it only uses one hyperparameter (rather than three for [13]). We find that Deep Ensembles obtains a competitive PDQ score to the established merging strategy, while MC Dropout has a significantly lower PDQ than the established technique. This occurs due to the decremented spatial quality produced by pre-NMS averaging for MC Dropout.

We hypothesise that the pre-NMS averaging strategy may inhibit a probabilistic detection's expression of spatial uncertainty by constraining detections to a single anchor box. In contrast, the established merging technique allows for unique anchor boxes within a probabilistic detection, so long as they meet the IoU 0.5 threshold (see Figure 2). We infer that this may be a reasonable constraint for Deep Ensembles, but not for MC Dropout, which may rely on multiple anchor boxes per probabilistic detection to express spatial uncertainty. We test this hypothesis by observing each detector's median number of unique anchor boxes per probabilistic detection. MC Dropout was found to have a median of 2 and maximum of 19 unique anchor boxes per probabilistic detection whereas Deep Ensembles had a median of 1 and maximum of 7. Given that the pre-NMS averaging strategy forces every probabilistic detection to have only 1 unique anchor box, this result supports our hypothesis and suggests that pre-NMS averaging is a viable merging strategy for Deep Ensembles but not MC Dropout.

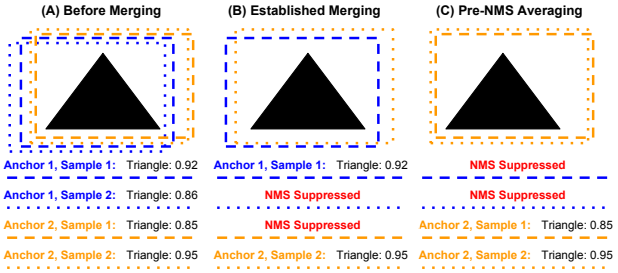

Figure 2. Consider outputs from two unique anchor boxes and two samples, each with a bounding box and softmax score for class 'triangle' (A). The established merging strategy (B) allows for unique anchor boxes to survive NMS and form a probabilistic detection for the triangle, while our pre-NMS averaging (C) only allows for one unique anchor box to survive NMS and form a probabilistic detection. This may limit expression of spatial uncertainty.

Conclusions: We performed the first comparative evaluation of sampling-based probabilistic object detectors and found Deep Ensembles to outperform the MC Dropout technique. We proposed a pre-NMS averaging merging strategy that achieves competitive performance to previously established merging strategies for Deep Ensembles, while having only one hyperparameter to tune.

This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016).

The authors acknowledge and thank Dr. Stefan Eickeler for his initial suggestion to merge samples at the detector anchor positions pre-NMS.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 1

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016. 1, 2

[3] Corina Gurau, Alex Bewley, and Ingmar Posner. Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*, 2018. 1

[4] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. *arXiv preprint arXiv:1811.10800*, 2018. 1, 2, 3

[5] Ali Harakeh, Michael Smart, and Steven L Waslander. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. *arXiv preprint arXiv:1903.03838*, 2019. 1

[6] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. *arXiv preprint arXiv:1809.08545*, 2018. 1

[7] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017. 1

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6405–6416, 2017. 1, 2

[9] Michael Truong Le, Frederik Diehl, Thomas Brunner, and Alois Knol. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3873–3878. IEEE, 2018. 1

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 1, 2

[12] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating Merging Strategies for Sampling-based Uncertainty Techniques in Object Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 2

[13] Dimity Miller, Lachlan Nicholson, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Dropout Sampling for Robust Object Detection in Open-Set Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 1, 2, 3

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 2

[15] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. 2

[16] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018. 1

[17] Kush R. Varshney and Homa Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5 3:246–255, 2017. 1