Dropout Variational Inference Improves Object Detection in Open-Set Conditions

Dimity Miller, Lachlan Nicholson, Feras Dayoub, Niko Sünderhauf
Australian Centre for Robotic Vision[*]
Queensland University of Technology (QUT), Brisbane, Australia
dimity.miller@connect.qut.edu

1 Introduction

One of the biggest current challenges of visual object detection is reliable operation in open-set conditions [1]. In contrast to closed-set conditions, where it is assumed that the objects seen during deployment are the same as during training, a vision system operating in *open*-set conditions [2, 3] will encounter objects of novel classes that were not part of the training dataset. Robust object detection in such conditions is of paramount importance for robotics, as a robot that acts on the output of an unreliable visual perception system can perform catastrophic actions.

One way to handle the open-set problem is to utilize the uncertainty of the model to reject predictions with low probability. Bayesian Neural Networks (BNNs) [4, 5], with variational inference commonly used as an approximation [6–10], is an established approach to estimate model uncertainty. In 2015, Gal and Ghahramani [11] proposed Dropout Variational Inference, also known as Dropout sampling, as a tractable approximation to BNNs. While Dropout sampling has recently been deployed to regression, semantic segmentation, and image classification tasks [11–14], it has not yet been applied to visual object *detection*.

Here we extend the concept of Dropout sampling to object *detection* for the first time. We evaluate Bayesian object detection on a large synthetic and a real-world dataset and show how the estimated label uncertainty can be utilized to increase object detection performance under open-set conditions.

2 Object Detection – A Bayesian Perspective

Object Detection with Dropout Sampling Object *detection* is concerned with estimating a bounding box alongside a label distribution for multiple objects in an image. To extend the concept of Bayesian deep learning from image recognition to object detection, we use SSD [15] with a VGG16 base network [16] that contains two dropout layers after the fully connected layers. Following the Dropout sampling approach [11], we sample from the distribution of the weights of this network by performing multiple forward passes of an image through the network with active Dropout layers.

Partitioning Detections into Observations A single forward pass through a sampled object detection network with weights $\widetilde{\mathbf{W}}$ yields a set of individual detections, consisting of bounding box coordinates $\mathbf{b}$ and a softmax score vector $\mathbf{s}$. We denote these detections as $D_i = \{\mathbf{s}_i, \mathbf{b}_i\}$. Multiple forward passes yield a larger set $\mathfrak{D} = \{D_1, \ldots, D_m\}$ of m such individual detections D_i. Detections from the set $\mathfrak{D}$ with high mutual intersection-over-union scores (IoU) will be partitioned into *observations* using a Union-Find data structure. We define an observation $\mathcal{O}_i$ as a set of detections with high mutual bounding box IoU: $\mathcal{O}_i = \cup D_i \quad \text{s.t. } \mathrm{IoU}(D_j, D_k) \geq 0.95 \;\; \forall D_j, D_k \in \mathcal{O}_i$. The threshold of 0.95 was determined empirically.

Extracting Label Probabilities and Uncertainty We can now approximate the vector of class probabilities $\mathbf{q}_i$ by averaging all score vectors $\mathbf{s}_j$ in an observation $\mathcal{O}_i$. This gives us an approximation

[*] This research was supported by the Australian Research Council Centre of Excellence for Robotic Vision, project number CE140100016.

of the probability of the class label y_i for a detected object in image $\mathcal{I}$ given the training data $\mathbf{T}$, which follows a Categorical distribution parametrized by $\mathbf{q}_i$: $p(y_i|\mathcal{I}, \mathbf{T}) \sim \text{Cat}\,(k, \mathbf{q}_i)$. The entropy $H(\mathbf{q}_i) = -\sum_j q_{ij} \cdot \log q_{ij}$ is used to measure the *label uncertainty* of the detector for a particular observation.

Using Dropout Sampling to Improve Object Detection Performance in Open-Set Conditions In open-set conditions, we would expect the label uncertainty, or Entropy $H(\mathbf{q}_i)$, to be higher for detections falsely generated on open-set objects (i.e. object classes not contained in the training data). A threshold on the Entropy $H(\mathbf{q}_i)$ can be used to reject detections of such unknown objects. This allows us to formulate the central **Hypothesis** of our paper: *Dropout variational inference improves object detection performance under open-set conditions compared to a non-Bayesian detection network.* The following two sections describe the experiments we conducted to verify or falsify this hypothesis and present our findings.

3 Evaluation

We evaluate the object detection performance in open-set conditions with three metrics: (1) Recall describes how well a detector identifies *known* objects, (2) open set error describes how robust an object detector is with respect to *unknown* objects and (3) precision describes how well a detector classifies *known* and *unknown* objects.

Precision and Recall Let $\Omega = \{\mathcal{O}_1, \dots \mathcal{O}_n\}$ be the set of *all* object observations in a scene after partitioning as described in Section 2. Label uncertainty is addressed by comparing the Entropy $H(\mathbf{q}_i)$ with a threshold θ and rejecting $\mathcal{O}_i$ if $H(\mathbf{q}_i) > \theta$. For every remaining $\mathcal{O}_i$, we find the overlapping known ground truth objects with an IoU of at least 0.5. If the winning label matches any of these objects, the observation is counted as true positive, otherwise as false positive. If there is no overlapping object and the winning class label is not 0 (unknown), this is also counted as a false positive. Every known ground truth object that was not associated with an observation (i.e. there is no $\mathcal{O}_i$ with an $\text{IoU} \geq 0.5$ with that object) gets counted as a false negative. Precision and recall are then defined as usual.

Absolute Open-Set Error We define absolute open-set error as the total number of observations passing the Entropy test ($H(\mathbf{q}_i) < \theta$) that do not overlap a ground truth known object ($\text{IoU} \geq 0.5$) and do not have a winning class label of 'unknown'.

Datasets Used in the Evaluation Our evaluation is based on two datasets: The SceneNet RGB-D validation set which contains photo-realistic images of 1000 differing indoor scenes [17], with 100 objects of unknown classes for a network trained on COCO. 30 images from each scene were tested in the evaluation. The second dataset evaluated was the QUT Campus Dataset, with data collected using a mobile robot across nine different and versatile environments on our campus [18]. Detections from this dataset were evaluated by manual visual inspection.

Evaluation Protocol and Compared Object Detectors Our evaluation compares the performance of *Vanilla SSD* (i.e. the default configuration of SSD [15]), *Vanilla SSD with Entropy thresholding* and *Bayesian SSD* (i.e. SSD with Dropout sampling and Entropy thresholding). Bayesian SSD was tested for 10, 20, 30 and 42 forward passes through the network. Performance metrics were calculated for entropy thresholds θ between 0.1 and 2.5.

4 Results and Interpretation

Our experiments confirmed the hypothesis formulated in Section 2: The Bayesian SSD detector utilizing Dropout sampling as an approximation to full Bayesian inference improved the object detection performance in precision and recall while reducing the open-set error in open-set conditions.

As shown in Table 1, Bayesian SSD is able to achieve significantly greater precision and recall scores than the vanilla SSD without Dropout sampling. When choosing the performance of the vanilla SSD as a reference point, the Bayesian SSD is able to significantly reduce open-set error (OSE) while retaining the F_1 score. Alternatively the F_1 can be significantly improved while keeping the OSE

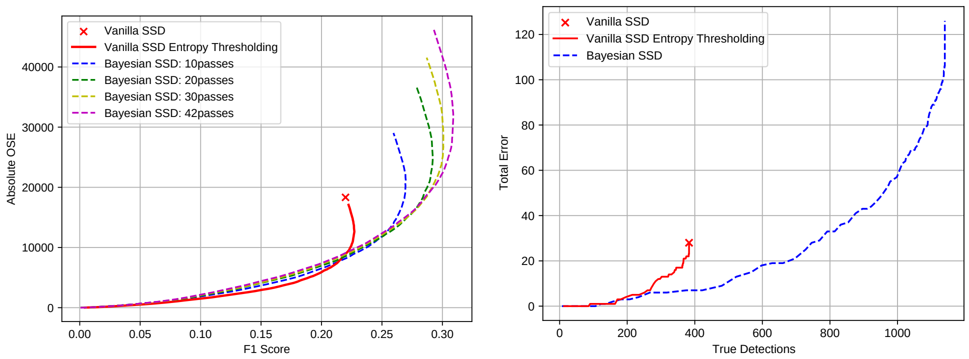

Figure 1: Network performances for SceneNet RGB-D (Left) and QUT Campus dataset (Right).

at the reference level. This suggests that Bayesian SSD produces a reliable uncertainty estimate for object classification; as such, it is able to make more informed decisions to reject incorrect classifications of known and unknown classes. A network utilizing Bayesian SSD is also able to achieve a significantly higher maximum recall. As expected, collecting detections from multiple forward passes allows Bayesian SSD to have a greater chance of detecting objects that may be overlooked in a single forward pass.

Forward Passes As can be seen in Figure 1, 10 forward passes is able to maintain the vanilla SSD reference F_1 score and reduce open-set error comparably to greater numbers of passes. However, at least 20 forward passes are needed to maximize F_1 score for the vanilla SSD open-set error. Beyond the reference OSE point, more forward passes achieve slightly higher F_1 scores, but at the cost of a significant increase in open set error. Depending on the performance requirements of a detection system, fewer forward passes may be suitable, allowing for reduced computation cost.

Real World Dataset For the QUT Campus dataset, the Bayesian SSD is able to reduce the total error per true detection. This can be seen in Figure 1, where at the reference point for the vanilla SSD with no entropy thresholding, Bayesian SSD has significantly reduced the total error. This consists of open-set error and incorrect classifications of known objects. Additionally, for the same total error, Bayesian SSD achieves significantly greater number of true detections. While this may be due to multiple detections per object, it can also be inferred that this partially represents the superior recall performance of Bayesian SSD.

5 Conclusions and Future Work

We verified the central hypothesis of our paper that Dropout sampling allows to extract better label uncertainty information and thereby helps to improve the performance of object detection in the open-set conditions that are ubiquitous for mobile robots. A promising direction for future work is to exploit the *spatial* uncertainty for an object-based SLAM system to gain a better estimate of the 6-DOF object pose.

Table 1: Performance Comparison on SceneNet RGB-D [17]

	Forward	max.	abs OSE	Recall	Precision	F_1 Score at	abs OSE at
	Passes	F_1 Score	at max F_1 point			ref. OSE	ref. F_1 Score
	vanilla SSD	0.220	18331	0.165	0.328	0.220	18,331
	SSD with Entropy test	0.227	**12638**	0.160	**0.392**		
Bayesian SSD	10	0.270	20991	0.214	0.364	0.269	**8,225**
	20	0.292	24922	0.244	0.364	0.284	8,313
	30	0.301	28431	0.261	0.355	**0.286**	9,003
	42	**0.309**	32034	**0.278**	0.347	0.285	9,256

References

[1] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1521–1528.

[2] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[3] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[4] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[5] R. M. Neal, "Bayesian learning for neural networks," Ph.D. dissertation, University of Toronto, 1995.

[6] J. Paisley, D. Blei, and M. Jordan, "Variational bayesian inference with stochastic search," *arXiv preprint arXiv:1206.6430*, 2012.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[8] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[9] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational bayes for non-conjugate inference," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1971–1979.

[10] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[11] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[13] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2016.

[14] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4762–4769.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of Neural Information Processing Systems (NIPS)*.

[17] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth," *arXiv preprint arXiv:1612.05079*, 2016.

[18] N. Sünderhauf, F. Dayoub, S. McMahon, B. Talbot, R. Schulz, P. Corke, G. W. B. Upcroft, and M. Milford, "Place Categorization and Semantic Mapping on a Mobile Robot," in *arXiv preprint: arXiv:1501.04158*, 2015.