TeamGL at ACRV Robotic Vision Challenge 1: Probabilistic Object Detection via Staged Non-Suppression Ensembling

Dongxu Li*, Chenchen Xu*, Yang Liu, and Zhenyue Qin

The Australian National University, Canberra, Australia
firstname.lastname@anu.edu.au

Abstract

This paper describes a novel approach to probabilistic object detection using ensemble techniques. The approach synthesises results from multiple non-probablistic object detectors to acquire final detections. We achieve this by a two-staged ensembling pipeline: (i) identifying detections that are of the same object based on the Intersection over Union (IoU) and labels utilising a greedy assignment process; (ii) creating an ensemble of the detections using a non-suppression algorithm. We employ fixed proportional and label confidence based covariances to capture the spatial uncertainty with particular calibrations on edging objects*, a special yet common class of detections. The proposed approach achieved 3^{rd} place in the leaderboard of CVPR-2019 ACRV Robotic Vision Challenge on Probablistic Object Detection.*

1. Introduction

Object detection has been an important scene understanding task for research in robotics and computer vision. Such a task aims at entitling robotic or autonomous systems to the ability to recognise and localise objects in the environment where they are operated in. The current state-of-the-art work approaches object detection problems by predicting positions of bounding boxes [7, 5, 4] or polytopes [13] that enclose the object along with a class label describing what the object is. Additionally, a score is usually computed to show the confidence over the object positioning and/or classification result.

In spirit of object detection, the task of *probabilistic object detection* (POD) [3] features a novel finer-grained, pixel-level measure for object localisation, which jointly with the label confidence measure establishes a new evaluation scheme for object detection, called *Probability-based Detection Quality (PDQ)*. Compared with existing average precision (AP) based metrics, PDQ takes a probabilistic perspective and rewards object detectors that better quantify the spatial and semantic uncertainties of detections. Such an extension may facilitate the development of robotic systems that operate interactively with human and environment by providing trustworthy detections.

The **ACRV Robotic Vision Challenge 1** [10] provides a POD task on a dataset resembling domestic scenes, e.g. garage, office, bedroom. This is particularly intriguing not only because the problem of estimating spatial and semantics uncertainties is unexplored per se, but also due to the vastly different scenes from most existing object detection datasets [2, 9, 5]. This variation in scenes poses a challenge for the generalisation ability of object detectors tested on other datasets. In fact, we found that models can behave differently on the new dataset, which leads us to an ensemble method. Ensemble models are among the most widely used techniques in deep learning [6, 14]. Their popularity is mainly attributed to the better predictive performance compared with constituent algorithms. Despite its wide application, proper ensemble strategies are task-dependent. Namely, one needs to carefully design the ensemble schemes such as to exploit the strengths of different algorithms and use them to compensate for rather than interfere with each other.

The main components of our method is in threefold:

- we present a non-suppression ensemble scheme, which enhances the performance of object detection systems;
- we estimate the spatial uncertainties of the object detection using covariances that are proportional to the scale of the detection.
- we apply further calibrations on edging detections to refine the spatial uncertainty estimation.

The rest of the paper is organised as follows. We introduce the methods in Section 2. In Section 3, we report and

*indicates equal contribution.

analyse the experiment results. We conclude our work and discuss possible further works at the end.

2. Our approach

In this section, we describe the major components that consist of our ensemble framework. We start by explaining the data pre-processing and test-time augmentation steps in Section 2.1. Then we introduce a method to create an ensemble of detections in Section 2.2-2.3. Lastly, we describe our several attempts to express spatial uncertainties in Section 2.3. An overview of our framework is demonstrated in Fig. 2.

2.1. Data Preprocessing

Our pre-processing and data augmentation steps are detailed as follows.

- **Standardize image** The ACRV challenge dataset is generated spanning a number of different environments and time. This type of change in the context of image background is diverse and less targeted in other object detection datasets, such as MS-COCO, where the data collection process implicitly constrained the spectrum of image sources. In Figure 1, we show an image on the validation set. It can be seen that the scene is captured in a fairly weak illumination condition, which makes it hard to recognise objects even for humans. We apply the image enhancement method introduced in [11] on both validation and test set to cope with it. An image is enhanced if its average grayscale is below a predefined threshold.

- **Data Augmentation** As suggested by multiple existing work [4] to alleviate the limitations in data examples, we adopt the data augmentation technique. Due to speed consideration, we applied all images through a fixed data augmentation process and saved the results for fine-tuning and inference job later. Particularly, horizontally flip and random crop augmentation are integrated.

Figure 1. Example of processed images. **Left:** original image, **Right:** enhanced image.

2.2. First level models

The first part of the detection framework are the base models, which form the foundation of our ensemble based approach. To obtain a diversified spectrum of prediction results, we experimented with a couple of existing models of various categories, e.g. single staged/two-staged, anchor-based/anchor-free models. Each model is initialised with the weights pretrained on the MS-COCO dataset. They are evaluated on the validation set of ACRV challenge with fixed covariances 35. We summarise the best five models selected for ensemble in Table 1.

Model	TP	FP	FN	Spatial quality	PDQ score
MaskRCNN [4]	29786	16808	60447	0.40	15.73
YOLOv3 [8]	10114	594	80119	0.51	7.65
RetinaNet [5]	15847	3681	74386	0.46	9.25
M2Det [12]	19432	6109	70801	0.47	12.27
ExtremeNet [13]	21676	5862	68557	0.46	11.78

Table 1. Performance of pretrained models on validation set.

Although bearing the variations between COCO and ACRV datasets, our preliminary study shows that for positive instances, the positioning of predicted bounding boxes is still satisfactory. Therefore, selected models are further fine-tuned but only for class label prediction on the provided validation data of ACRV challenge. We test two versions of each model, fine-tuned and only pre-trained respectively, in the ensemble process.

2.3. Staged Non-suppression Ensembling

In this section, we present the non-suppression ensemble (NSE) method we use in the challenge. The main driven force for us to resort to the ensemble method is the observation that detection performance for different object classes differs on methods. For example, YOLO [7] achieves a better recall score than RetinaNet [5] on the "wine glass" class while RetinaNet performs better on "sink" and "clock". The proposed method constitutes two stages: *duplicates matching* and *non-suppression merging*, as we will explain next.

Duplicates Matching When multiple algorithms succeed on detecting the same object, the duplicates must be handled to avoid false positive detections. To this end, a preliminary procedure is required to identify such possible duplicates. We notice that when evaluating the performance of object detection systems, algorithms are utilised to match the detection-object pair [2]. We extend this idea for our purpose. Specifically, we iterate over detections from each model and group detections that (i) share the same class label; and (ii) overlap heavily with each other, determined by a predefined IoU threshold. We then assign detections in each group to the same object. We remark that our matching strategy is greedy with respect to IoU scores rather PDQ

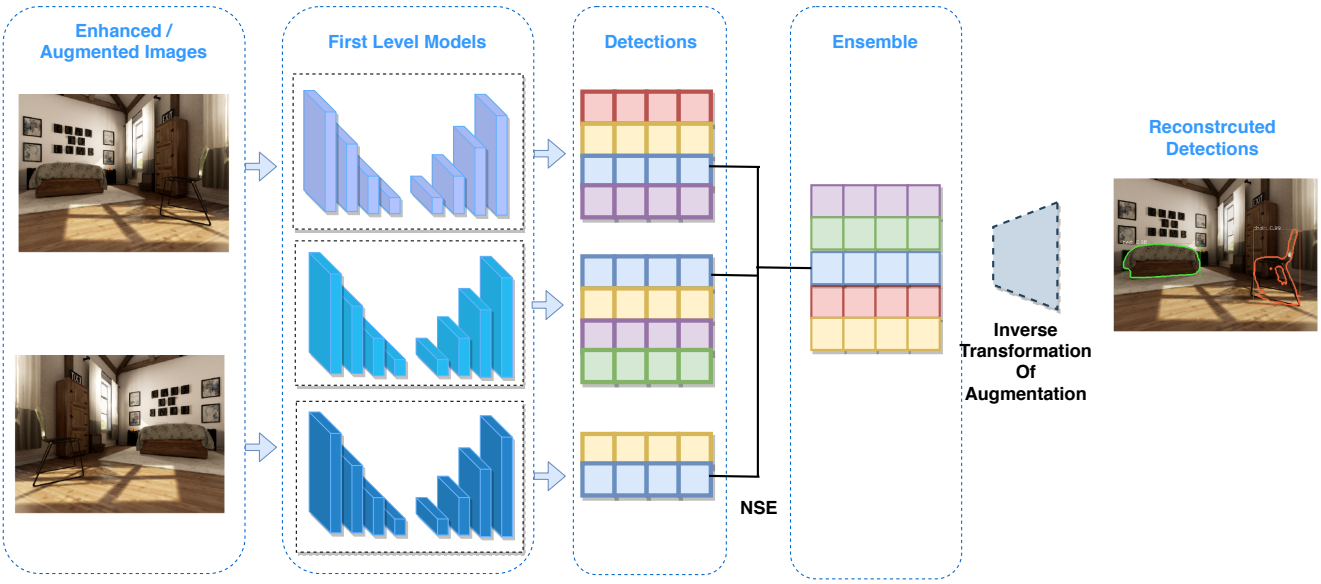

Figure 2. Overview of the detection framework

scores. Adopting an optimal assignment procedure based on PDQ scores may help to identify better matches.

Non-suppression Merging For standalone detections that are not grouped, we straightforwardly add them to the ensemble. For the grouped detections, we expect a procedure that reduces the group to a single detection. A relevant problem setup is a common post-processing step in object detection methods, where Non-maximum Suppression or its variant [1] is required to remove ROIs that overlap significantly with the most confident box. This is, however, not aligned with our aim to calibrate the detection taking the most advantage of results from multiple models. On the other side, eliminating low-confidence detections can also be achieved by thresholding the detectors before ensembling. We therefore proceed in the opposite direction and keep all the results without suppression. Then we average over the group of detections to acquire the synthesised detection box. Extending non-suppression merging by applying weighting terms to reward/penalise different detections is also possible. We briefly outline the ensemble method in Algorithm 1. We iterate over detections from different methods (line 3-7) and find matching detections based on iou scores. Line 8-10 averages over detections to create the final ensemble.

Algorithm 1: Staged Non-Suppression Ensemble (NSE)

Result: $\hat{D}$: ensemble detections
1 $D \leftarrow \{d_1, d_2, ..., d_n\}$ // d_i: detections from i^{th} method
2 $M \leftarrow emptyDict()$; $\hat{D} \leftarrow emptyList()$
3 **for** $i \leftarrow 0$ **to** $|D|$ **do**
4 **for** $j \leftarrow i+1$ **to** $|D|$ **do**
5 $M \leftarrow find\_match(d_i, d_j, M)$
6 **end**
7 **end**
8 **foreach** $m \in M$ **do**
9 $\hat{D}.append(\frac{1}{m}\sum_{i=0}^{i<|m|} m_i)$
10 **end**

1 **Procedure** $find\_match(d, d', M)$
2 **for** $i \leftarrow 0$ **to** $|d|$ **do**
3 sort d' on iou with d_i
4 **for** $j \leftarrow 0$ **to** $|d'|$ **do**
5 **if** $iou(d_i, d_j) > \epsilon_{iou}$ **and** $d_i.label = d_j.label$ **then**
6 $M[d_i] \leftarrow M[d_i].append(d_j)$
7 **end**
8 **end**

2.4. Handling Spatial Uncertainties

Whereas the current model can recognise objects with satisfying accuracy, the PDQ score adopted for this challenge also measures the quality of generated bounding box. Apart from directly improving the bounding box prediction, PDQ penalises incorrect positioning with low spatial uncertainty. More precisely, POD algorithms are required to provide a covariance value at each side of the bounding box to generate a probability heatmap (a.k.a probabilistic bounding box). We summarise the two approaches tested for predicting these covariance values.

- *Proportional covariances* Since the PDQ score mea-

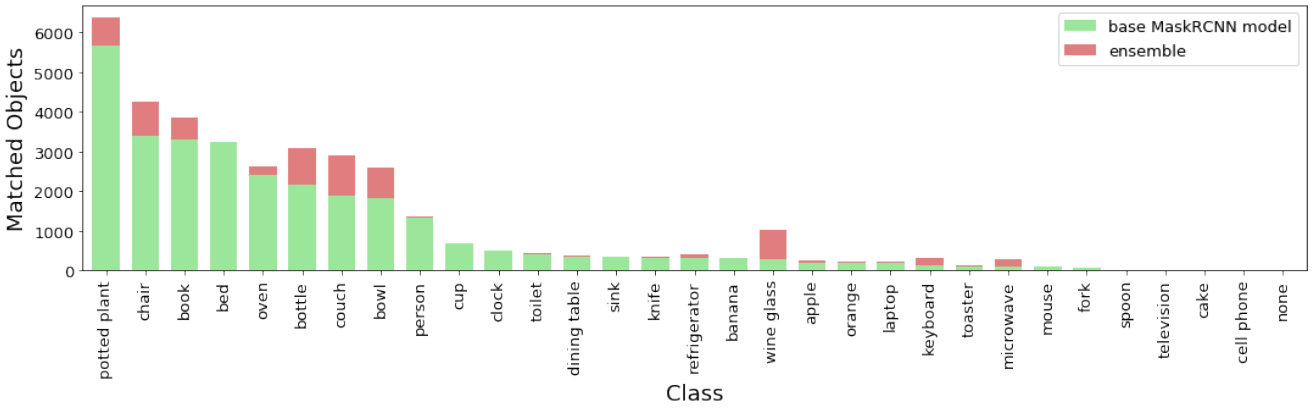

Figure 3. Improvement of detections from ensemble method.

sures spatial uncertainties in pixel, detections on large objects intuitively are more likely to contribute to inaccurate positioning. Therefore, we multiply the scale of detections by a predefined proportion, which serve as the standard deviation of the Gaussian distribution. We also assume that there is no correlation between the horizontal and vertical positions of an object and therefore assign covariances matrices as diagonal ones. We observed that a proportion of around 6% gives the best spatial score on the dataset.

- *Confidence based covariances* When models are evaluated on the validation set, given each pair of ground truth bounding box and the predicted one, the optimal covariance matrix that maximizes the PDQ score exists. However, finding the exact optimal solution can be computationally heavy. In our experiment, we approximate it by the following heuristic method. Denote the ground truth bounding box and the predicted one as $\text{BBox}_{gt} = (\underline{x}, \underline{y}, \bar{x}, \bar{y})$ and $\text{BBox}_{pred} = (\underline{x}', \underline{y}', \bar{x}', \bar{y}')$, respectively, the approximated optimal covariance matrix $\mathbf{Cov} = [\mathbf{Cov}_1, \mathbf{Cov}_2]$ is computed as:

$$\mathbf{Cov}_1 = \frac{1}{5^2}\begin{bmatrix}(\underline{x}-\underline{x}')^2 & 0\\ 0 & (\underline{y}-\underline{y}')^2\end{bmatrix}$$
$$\mathbf{Cov}_2 = \frac{1}{5^2}\begin{bmatrix}(\bar{x}-\bar{x}')^2 & 0\\ 0 & (\bar{y}-\bar{y}')^2\end{bmatrix}$$

The intuition behind it is to increase the overlap between the probabilistic bounding box with the ground truth one. We then examine these approximated $\mathbf{Cov}$ on all detections on the validation set, and find their empirical estimation of distribution for each element in the major diagonal and by each class, denoted as $\mathbf{Cov}_{cls}^{i,j} \sim \mathrm{D}(\boldsymbol{\mu}_{cls}^{i,j}, \mathbf{Var}_{cls}^{i,j})$.

Without explicit confidence score for predicted boxes from the model, we use the confidence score of the label instead. Now given a new prediction with label score $\mathcal{P}_l$, the predicted $\mathbf{Cov}$ is as:

$$\mathbf{Cov}^{i,j} = \boldsymbol{\mu}_{cls}^{i,j} + \alpha(1-\mathcal{P}_l)\cdot\sqrt{\mathbf{Var}_{cls}^{i,j}}$$

where α is the hyper-parameter to control the effect from variance.

Initial validation results show better performance from comparatively simple proportional covariance and is thus the only adopted and tested into our submission due to limited testing opportunities. We attribute the less satisfying result from confidence based covariances to insufficient fine-tuning work on the control factor α.

2.5. Edging Object

When the bounding box is predicted to be close to the image border, we notice that the model become less reliable in distinguishing the case of partial observation (part of object lays outside of view) and actual objects sitting next to the border. Intuitively, if an object is found likely to be a partial match, we can safely raise the confidence of its predicted bounding box on the side next to the image border.

We find all predictions if any side of the bounding box is only 0 to 2 pixels away from the image border, and filter off all the small detections (heuristic measurement for small objects which are unlikely to form a partial match). For those remaining, instead of reducing the Cov to 0 that is prone to over-confidence, we use a compromised approach that refines the box to be 1 pixel away from the border and attach a comparatively small Cov to it. Our experiment shows that this fix brings an consistent increase of about 0.5 to the overall PDQ score.

3. Results

We validate the effectiveness of proposed ensemble method on validation set and the result can be found in Fig-

Model	Split	TP	FP	FN	Spatial quality	PDQ score
NSE+fixed cov=35	Val.	36373	26164	53866	0.43	19.22
NSE+prop. cov=6%	Val.	35940	26775	54293	0.46	19.43
NSE+prop. cov=6%+edg	Val.	35881	26848	54352	0.49	19.86
NSE+prop. cov=6%+edg	Test	118678	91735	177689	0.50	20.02

Table 2. Performance of methods on validation and test sets. We use "fixed cov" to denote the setting where the covariance matrices are diagonal matrices with fixed values on corners; "edg" to denote the fix on edging objects.

ure 3. The green part of the bar plot indicates the number of correctly matched object by using only the base model (MaskRCNN) and the red part indicates the additional correct detections from ensemble. The ensemble method manages to properly integrate the diverse output from our first level models, resulting improvement in the detection recall on both major and minor classes.

The final result on the public validation set and private testing set is summarised in Table 2. From the table, we have the following observations.

- NSE significantly improves the PDQ score compared with single-model methods in Table 1. This suggests that the ensemble scheme we propose succeeds on combining true positive instances while properly controlling the number of false negative instances.

- Proportional covariances and the fixes on the edging objects both contribute to improve spatial quality. It is also noticeable that TP decreases in this case, which is sensible as a more compact spatial estimation may risk losing true detection-object matchings. However, the overall PDQ score still arises, which justifies the effectiveness of the methods.

- PDQ scores on the test set is marginally higher than the validation set. This is not too surprising as the ACRV dataset exhibits dramatically different scenes and object distributions on the validation and test sets. Although tuning on validation helps us to get a better sense of the influence of parameters, there is barely direct causal relation between the results on the two sets.

4. Conclusion

In this paper we describe the approach we use in our submission to ACRV Robotic Vision Challenge 1. We propose non-suppression ensembling to integrate multiple base non-probabilistic models to improve the diversity of model prediction. Investigation has also been made to precisely calibrate the uncertainty in the bounding box. As a result, we are able to address some of the new challenges introduced by this competition and achieve promising results in the final leaderboard.

As for the future, improvements can be made to the performance of base models by integrating the temporal information implied in the sequential data. We also envision to formulate the measurement of probabilistic bounding box into a learning problem so that the model can be trained to predict it.

References

[1] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5561–5569, 2017.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[3] D. Hall, F. Dayoub, J. Skinner, P. Corke, G. Carneiro, and N. Sünderhauf. Probability-based detection quality (pdq): A probabilistic approach to detection evaluation. *arXiv preprint arXiv:1811.10800*, 2018.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[6] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[8] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[10] J. Skinner, D. Hall, H. Zhang, F. Dayoub, and N. Sünderhauf. The probabilistic object detection challenge. *arXiv preprint arXiv:1903.07840*, 2019.

[11] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang. A new image contrast enhancement algorithm using exposure fusion

framework. In M. Felsberg, A. Heyden, and N. Krüger, editors, *Computer Analysis of Images and Patterns*, pages 36–46, Cham, 2017. Springer International Publishing.

[12] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. 2018.

[13] X. Zhou, J. Zhuo, and P. Krähenbühl. Bottom-up object detection by grouping extreme and center points. *arXiv preprint arXiv:1901.08043*, 2019.

[14] Z.-H. Zhou and J. Feng. Deep forest: Towards an alternative to deep neural networks. *arXiv preprint arXiv:1702.08835*, 2017.