Assessing Box Merging Strategies and Uncertainty Estimation Methods in Multimodel Object Detection

Felippe Schmoeller Roza[1][0000−0001−6900−6601], Maximilian Henne[1], Karsten Roscher[1], and Stephan Günnemann[2]

[1] Fraunhofer IKS, Munich, Germany
[2] Technical University of Munich, Munich, Germany

Abstract. This paper examines the impact of different box merging strategies for sampling-based uncertainty estimation methods in object detection. Also, a comparison between the almost exclusively used softmax confidence scores and the predicted variances on the quality of the final predictions estimates is presented. The results suggest that estimated variances are a stronger predictor for the detection quality. However, variance-based merging strategies do not improve significantly over the confidence-based alternative for the given setup. In contrast, we show that different methods to estimate the uncertainty of the predictions have a significant influence on the quality of the ensembling outcome. Since mAP does not reward uncertainty estimates, such improvements were only noticeable on the resulting PDQ scores.

Keywords: Uncertainty estimation · deep ensembles · object detection

1 Introduction

Uncertainty estimation has got increasing attention in the research community during the last years as an utterly important feature for Deep Neural Networks (DNNs) embedded in safety-critical applications, such as autonomous driving, robotics, and medical image classification.

Bayesian neural networks allow to inherently express uncertainty by learning a distribution over the weights, with the caveat of being extremely expensive in terms of computation and therefore an unfeasible option for most practical problems. Several methods were proposed to allow uncertainty estimation on computer vision tasks like image classification. Sampling-based methods, such as Monte Carlo Dropout and Deep Ensembles ([4],[8]) approximate Bayesian models by combining multiple predictions for the same input. Other works focus on training a model to predict the uncertainties for out-of-distribution detection or to calibrate the usually overconfident softmax confidence scores (respectively [2],[6]).

Adapting such methods to object detection architectures brings the challenge to a new complexity level because not only uncertainty regarding the label

assignment has to be estimated but also concerning the spatial uncertainty of the bounding box coordinates. [9] and [3] presented object detectors based on the dropout approximation. Uncertainty regarding both label and box position is mostly based on the softmax-output confidences of networks which are often sampled from several forward passes through different model parameterizations ([4], [8]).

In [1], the authors propose a method that aggregates the ensemble detections by means of the intersection over union (IoU) between the predicted objects. Then, the merging method consists in discarding the clusters with the number of elements below a certain threshold. Three thresholds were tested, 1, $m/2$, and m, which is the number of models in the ensemble. In the end, non-maximum suppression is used to get the final box predictions. [7] recently introduced a loss function that was named KL-loss and integrate the usual bounding box regression loss with a variance estimation, increasing the performance of object detection models.

Even though existing works show how the quality of the uncertainty estimations can be improved by different methods, there is, to the best of our knowledge, no work that compares network softmax confidence scores and variance estimates as effective features to help on adjusting the final bounding box predictions.

In this paper, we compare variance and confidence as potential sources of information to get reliable uncertainty estimations and compare different box merging strategies using these predictions to obtain a final box prediction out of multiple models. Also, two different uncertainty estimation methods are compared. The results are evaluated using mAP, the dominant evaluation metric for object detection models, and PDQ, which encompasses uncertainty on both localization and classification tasks.

2 Methods

In this section, different methods used to combine the detections in the ensemble to either get better detection estimates or improve the uncertainty estimation are presented. Methods to combine the bounding boxes will be classified as merging strategies while the uncertainty estimation methods use different detections to assess more meaningful estimates. If we consider the bounding box coordinates as Gaussian distributions over the pixels, the different merging strategies allow shifting the mean coordinate values whereas the variance estimates represent the standard deviation values.

$\mathcal{D} = [D_1, D_2, ..., D_n]$, is a detection vector that contains the detections provided by n detectors trained independently. Each detection $D_i = [B_i, c_i, s_i]$, with $i \in \{1...n\}$, consists of the bounding box coordinates B_i, the class label c_i, and the respective softmax confidence score s_i. To ensure that the detections match the same object, they must have an IoU above a given threshold t_{IoU} and only one detection per model is taken in an ensemble $\mathcal{D}$. The vector of bounding boxes will be represented as $\mathcal{B} = [B_1, B_2, ..., B_n]$, the class label vec-

tor $C = [c_1, c_2, ..., c_n]$, and confidence score vector $\mathcal{S} = [s_1, s_2, ..., s_n]$. The final merged detection is represented as $\hat{D} = [\hat{B}, \hat{c}, \hat{s}]$. The detections can be extended by adding uncertainty estimates, $\hat{D}^{\sigma} = [\hat{B}, \hat{\sigma}, \hat{c}, \hat{s}]$.

Since this paper focuses on improving the bounding box estimation, only the spatial uncertainty will be considered. For the same reason, different methods for improving the merging of the label predictions and the final scores will not be discussed. In this paper, the resulting class $\hat{c}$ is the mode of all predicted labels in the ensemble, and the resulting score $\hat{s}$ is the mean score.

2.1 Spatial Uncertainty Estimation

The spatial uncertainty estimation consists in finding the variances for each coordinate of the bounding boxes, $\sigma = [\sigma_{x1}, \sigma_{y1}, \sigma_{x2}, \sigma_{y2}]$. Two different approaches were considered.

Ensemble variances: Considering that the ensemble provides boxes that, most likely, have differences in the coordinates, variance estimates can be obtained by calculating the covariance matrices from the bounding boxes coordinates:

$$\sigma_b^2 = \frac{1}{n-1} \sum_{k=1}^{n} (b_k - \bar{b})^2,$$

where b represents a vector with all values for a single box coordinate in the ensemble $\mathcal{B}$, i.e., $b \in \{x_0, y_0, x_1, y_1\}$, where $x_0 = [x_{01}, x_{02}, ..., x_{0n}]$ and so on; and $\bar{b}$ is the mean value of the vector b.

KL-var: Another approach consists of changing the object detector architecture to produce localization variance estimates. In this paper, all variance estimation models utilize the KL-loss in combination with variance voting, as introduced and explained in [7]. With this method, the bounding box predictions are treated as Gaussian distributions whereas the ground truths are represented as a Dirac delta. The regression loss is then the KL divergence of prediction and ground truth distributions, allowing the model to learn the bounding box regression and uncertainty at once. These models will be referred to as KL-var models.

For the KL-var ensembles , the resulting variance is the mean of the variances present in the ensemble, $\hat{\sigma}_{KL} = [\overline{\sigma}_{x0}, \overline{\sigma}_{y0}, \overline{\sigma}_{x1}, \overline{\sigma}_{y1}]$.

2.2 Box Merging

The merging strategy for the final box consists in a method to combine the ensemble detections to get an improved resulting bounding box $\hat{B}$. It can be obtained by the weighted sum of the ensemble elements, i.e., $\hat{B} = W \cdot \mathcal{B}^{\intercal}$, where W is a vector of weights: $W = [w_1, w_2, ..., w_n]$.

Max: Taking always the box with the highest confidence score is a simple algorithm, although highly dependent on a good calibration of the confidences concerning the bounding box spatial accuracy.

Mean: The resulting bounding box can also be obtained as the mean of the coordinates from the ensemble detections. : $W = \left[\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right]$.

WBF: Another method called weighted boxes fusion (WBF) was presented in [11]. It uses the normalized confidence scores as weights:

$$W = \frac{S}{\sum_{i=1}^{n} s_i}.$$

Var-WBF: To further investigate the comparison between confidence scores and variance estimates as a good indicator of the detected bounding box correctness, WBF was adapted by replacing the normalized confidences with the normalized inverted variances as weights (with the variance represented as the mean from the 4 coordinate variances).

Variance voting: Var-voting was introduced in [7] to adjust the box coordinates during non-maximum suppression (NMS) based on the variance estimates. Since the task of merging the outputs of an ensemble of models $\mathcal{D}^{\sigma}$ is so closely related to what is done in NMS algorithms, in both the goal is to select and/or combine several bounding boxes to find a better box representative, var-voting was adapted as one of the merging strategies.

3 Results

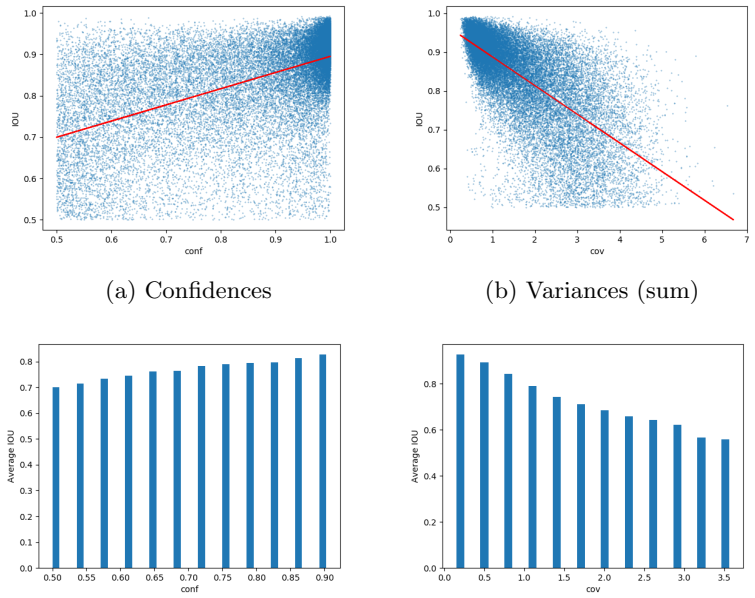
The results were obtained using Efficientdet-D0 and YoloV3 frameworks ([12],[10]) trained on the Kitti dataset [5]. Vanilla and KL-var ensembles were trained for both frameworks. Each ensemble consists of 7 models trained independently. The dataset was randomly split as follows: 80% of the images were used for training and 20% for testing. The IoU threshold was $t_{IoU} = 0.5$ for all the experiments. The model was not fine-tuned in any specific way as the primary interest in this paper is to examine the difference of merging strategies and uncertainty estimation methods and not to improve the state-of-the-art performance. Also, objects detected by a single model were discarded, because such cases will not benefit from any of the considered methods.

3.1 Confidence versus variances

The correlation between the quality of each predicted bounding box and the confidence and variance estimates was investigated. The comparison was done

using the KL-var ensemble of Efficientdet-D0 models, that output both confidence scores and coordinate variances for each prediction. The quality measure was hereby defined as the IoU of the prediction and the corresponding ground-truth. Good estimates, with a high IoU, should present a high confidence/low variance, and a low IoU when the opposite is true.

The results, shown in Figure 1, demonstrate that the variances are a better indicator of the detection spatial correctness, with a more pronounced correlation (negative since high variance translates as a high spatial uncertainty). The confidences show a weaker, albeit still present, correlation. It is important to notice that poor estimates are much more recurrent on the confidences, as shown in the scatter plots 1(a) and 1(b). Although in both cases a more dense mass of points is close to the top-right and top-left corners respectively and therefore representing good estimates, the confidences are almost uniformly distributed for the remaining of the points. On the other hand, the variances proved more accurate, being more concentrated around the first-order polynomial fitted to the samples.



(a) Confidences (b) Variances (sum)

(c) Confidences (d) Variances (sum)

Fig. 1: Correlation between IOU and confidences/variances. Remark: the variance values are obtained by the sum of the σ vector elements.

3.2 Merging strategies and uncertainty estimation

Tables 1 and 2 show the results obtained from the testing set regarding mAP, average IoU and PDQ. All merging strategies presented in the Section 2.2 were compared using vanilla ensembles without an uncertainty estimation, with uncertainty obtained through the ensemble variance method and the KL-var ensembles, which provides out-of-the-box uncertainty estimates. mAP and Avg. IoU do not take the uncertainty estimates into account and are therefore presented only once.

No uncertainty estimation				Ens. var.	KL-var
Method	**mAP**	**Avg. IoU**	**PDQ**	**PDQ**	**PDQ**
Max	0.3206	0.8217	0.1253	0.2859	0.3489
Mean	0.3371	0.8356	0.1480	0.3007	0.3518
WBF	0.3373	0.8360	0.1482	0.3015	0.3520
Var-WBF*	0.3381	0.8366	0.1484	0.3074	**0.3521**
Var-voting*	**0.3390**	**0.8372**	**0.1488**	**0.3113**	0.3519

Table 1: EfficientDet-D0 ensemble results.

No uncertainty estimation				Ens. var.	KL-var
Method	**mAP**	**Avg. IoU**	**PDQ**	**PDQ**	**PDQ**
Max	0.3001	0.8064	0.0725	0.3143	0.3388
Mean	0.3279	**0.8271**	0.1203	0.3391	**0.3679**
WBF	0.3280	0.8268	0.1263	0.3378	0.3673
Var-WBF*	0.3284	0.8268	0.1269	**0.3399**	0.3677
Var-voting*	**0.3285**	0.8253	**0.1274**	0.3384	0.3678

Table 2: YoloV3 ensemble results.

Var-WBF and Var-voting depend on the ensemble having individual variance for each box in the ensemble and are not compatible with the vanilla ensembles. All results for these methods were therefore obtained with the KL-var ensembles, but the variances were only used for merging the boxes and not used as an uncertainty estimate, except for the results on the KL-var column.

The results show that the different merging strategies perform almost the same with a slight improvement between taking the maximum confidence box and all other merging strategies, showing that trusting only in the confidences is not a good method. However, despite the variances presenting a superior estimate than the confidence scores, the Var-WBF did not perform better than the regular WBF method. None of the methods performed significantly better than taking the mean, which is the simplest of the merging strategies. One of the reasons is that most of the detections have high confidence or low variance,

since non-maximum suppression and var voting suppress the low scored ones. After normalization, the weight distributions for both methods become close to the mean weights.

Whilst the influence of the merging strategies appears negligible, a more pronounced difference can be observed when comparing the different uncertainty estimation methods presented in section 2. A big step in the PDQ values is already expected when providing meaningful uncertainty estimates since the variances are treated as zeroes when no uncertainty estimate is available, as if the model is completely certain about its bounding box predictions. With the ensemble variances as an uncertainty estimate, PDQ values already drastically improve. Even though this simple method does not require any modification of the base object detectors it more than doubles the detection quality if uncertainty is taken into account. However, if the models themselves predict their individual uncertainties, results improve even further for all cases and both model architectures.

4 Conclusions and Future Work

In this paper different strategies to merge bounding boxes from different models in an ensemble-based object detection setting were investigated. Furthermore, two approaches to estimate the uncertainty of the final box were presented and evaluated.

The results demonstrate that predicted variances are better correlated with the resulting IOUs than the almost uniquely used confidence values for object detectors and should be interpreted as a better spatial correctness estimate. However, when comparing merging strategies based on the variance and the confidence scores, the improvement was not as pronounced as expected and all the averaging methods showed similar results. In this regard, simply taking the mean values can be considered a good baseline, with competitive performance on all metrics and simple implementation.

We also observed significantly better results when employing variance predictions for bounding box uncertainty estimation, which is only rewarded by the PDQ metric. Out of the two proposed approaches, utilizing predicted variances from the individual detectors led to a better performance than relying on the variance of the detected boxes by the different models.

With uncertainty estimation playing an important role towards the goal of embedding AI in safety-critical systems, PDQ stands out as a more suited evaluation metric in opposition to mAP, commonly used and accepted as the object detection metric. The results show that there is still great room for improvement when working with the uncertainty estimations, with an expressive increase for all metrics in both presented methods.

As future work, we would like to apply the same methods on different datasets, such as COCO and Pascal VOC, that have different classes and different aspect ratios, which may play an important factor in the merging results. Methods to obtain uncertainty estimates on the classification task can also be integrated to improve the PDQ scores. Finally, we would like to test if combining the indi-

vidual box variances with the ensemble variance will result in a more refined uncertainty estimation.

Acknowledgments

This work was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.

References

1. Casado-García, A., Heras, J.: Ensemble methods for object detection (2019), https://github.com/ancasag/ensembleObjectDetection
2. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
3. Feng, D., Rosenbaum, L., Dietmayer, K.: Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3266–3273. IEEE (2018)
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
5. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. arXiv preprint arXiv:1706.04599 (2017)
7. He, Y., Zhu, C., Wang, J., Savvides, M., Zhang, X.: Bounding box regression with uncertainty for accurate object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2888–2897 (2019)
8. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. pp. 6402–6413 (2017)
9. Miller, D., Nicholson, L., Dayoub, F., Sünderhauf, N.: Dropout sampling for robust object detection in open-set conditions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–7. IEEE (2018)
10. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
11. Solovyev, R., Wang, W.: Weighted boxes fusion: ensembling boxes for object detection models. arXiv preprint arXiv:1910.13302 (2019)
12. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2020)