How Should We Evaluate Probabilistic Object Detectors?

Ali Harakeh and Steven L. Waslander

***Abstract*— This note provides a study of the properties of commonly used performance measures for probabilistic object detectors. Modifications to the recently formulated Probability based Detection Quality (PDQ) are proposed to account for the probability mass assigned to the category and bounding box of false positive detections. Empirical results on simulated and real detections from the COCO dataset are used to show the effectiveness of the proposed modifications when capturing the performance of a probabilistic object detector.**

I. Introduction

For safe and robust usage in robotic systems, object detectors should be able to provide a meaningful estimate of uncertainty associated with their output detections. To that end, Hall *et al*. [1] introduced the probabilistic object detection task, which requires estimating full **probability distributions** relating to the category and bounding box of every object in the scene.

Performance measures predominantly used to evaluate the performance of standard object detectors fail to take measure the performance of full probability distributions provided by probabilistic object detectors. To that end, performance measures such as the probability-based detection quality (PDQ) [1], and the minimum Uncertainty Error (MUE) [2] have been recently proposed in an attempt to provide quantitative measures of performance for probabilistic object detectors.

This note provides a comparative study of these performance measures in context of the probabilistic object detection task. Specifically, we provide a theoretical analysis to determine the advantages and disadvantages of using recently proposed performance measures, in comparison to Average Precision. We also propose modifications to the probabilistic detection quality (PDQ) score, which take into account the quality of probability distributions relating to false positive detections. Finally, we provide empirical results on simulated detections and on the COCO object detection dataset [3] to verify the proposed improvements.

II. Analysis of Commonly Used Probabilistic Object Detection Metrics

Average Precision (AP) was proposed by Everingham *et al*. [4] to evaluate the performance of object detectors, and is defined as the area under the continuous precision-recall (PR) curve, approximated through numeric integration over a finite number of sample points [4]. To perform this computation, true positives (TP) are defined as any detection $\mathcal{D}$ that: 1) has a classification score greater than a predetermined threshold δ_{cls} and 2) has an intersection-over-union (IOU) with any ground truth instance greater than a predetermined threshold, δ_{iou}. False positives, on the other hand, are the set of detections that have failed either of the above two criteria, and also include duplicate detections that satisfy both criteria. AP is used to measure the performance of non-probabilistic object detectors and does not take into account the uncertainty a detector has in its output.

Ali Harakeh and Steven L. Waslander are with The Institute For Aerospace Studies (UTIAS), University of Toronto, Toronto, Canada, ali.harakeh@utoronto.ca, steven.w@utias.utoronto.ca

A. Minimum Uncertainty Error

The Uncertainty Error (UE) was first proposed to evaluate probabilistic object detectors by Miller *et al*. [2]. The uncertainty error is calculated as:

$$UE(\delta) = 0.5\frac{|TP > \delta_{ent}|}{|TP|} + 0.5\frac{|FP \leq \delta_{ent}|}{|FP|}, \quad (1)$$

where δ_{ent} is a threshold on the entropy of a distribution associated with the output detection. UE ranges between 0 and 0.5 and can be thought of as the probability that a simple threshold-based classifier makes a mistake when using entropy to classify output detections into true positives and false positives. As the uncertainty error approaches 0.5, using the provided entropy is not much better than using a random classifier to separate TPs from FPs. The best uncertainty error achievable by a detector at the best possible value of the threshold δ_{ent} is called the Minimum Uncertainty Error (MUE) and is usually used to compare probabilistic object detectors in [2], [5].

Similar to AP, MUE is independent of a score threshold, but requires δ_{iou} to determine what detections count as true positives, and is therefore threshold dependant. Furthermore, MUE is independent of global linear transformations of entropy, where shifting the estimated entropy through an additive or multiplicative transformation results in an equivalent shift in δ_{ent} to maintaining a constant MUE. As a result, MUE is capable of only providing information on how well the estimated entropy can be used to separate true positives from false positives, and not on the actual quality of the estimated probability distributions.

B. Probability Based Detection Quality (PDQ)

Hall *et al*. [1] recently proposed PDQ as a metric to measure the quality of two dimensional probabilistic object detectors. PDQ can then be written as:

$$PDQ(\mathcal{G}, \mathcal{D}) = \frac{1}{|\mathcal{G}| + N_{FP}} \sum_{i,j,f} pPDQ(\mathcal{G}_i^f, \mathcal{D}_j^f). \quad (2)$$

Here, $\mathcal{G}_i^f$ be the i^{th} ground truth object instance in the f^{th} frame of a dataset, $\mathcal{D}_j^f$ is the j^{th} matched detection from the same frame, $|G|$ is the number of ground truth instances in

the dataset, and N_{FP} is the total number of false positives. The ground truth $\mathcal{G}_i^f$ is defined by $\hat{\mathcal{S}}_i^f$, the set of pixels defining the object instance's mask, $\hat{\mathcal{B}}_i^f \in \mathbb{R}^4$ the object's bounding box, and $\hat{c}_i^f \in \{0, \ldots, K\}$ the object's category label. Similarly, the detection $\mathcal{D}_j^f$ is defined by its bounding box $\mathcal{S}_j^f$, a covariance matrix of the elements of $\mathcal{S}_j^f$ referred to as Σ_j^f and a vector of category probabilities $\mathcal{I}_j^f \in \mathbb{R}^K$.

pPDQ is a pairwise probabilistic detection quality measure that comprises of two quality components, the spatial quality and the label quality. The spatial quality measures how well the detector captures the bounding box multivariate Gaussian distribution, and can be written as:

$$\begin{aligned} \mathcal{Q}_S(\mathcal{G}_i^f, \mathcal{D}_j^f) &= \exp(-(L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) + L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)) \\ L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) &= -\frac{1}{|\hat{\mathcal{S}}_i^f|} \sum_{x \in \hat{\mathcal{S}}_i^f} \log((P(x \in \mathcal{S}_j^f)) \\ L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f) &= -\frac{1}{|\hat{\mathcal{S}}_i^f|} \sum_{x \in \hat{\mathcal{V}}_{ij}^f} \log((1 - P(x \in \mathcal{S}_j^f)), \end{aligned} \quad (3)$$

where $L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f)$ is the foreground loss, $L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)$ is the background loss, and $\hat{\mathcal{V}}_{ij}^f$ are pixels belonging to the $(\mathcal{S}_j^f \cup \hat{\mathcal{S}}_i^f) - (\mathcal{S}_j^f \cap \hat{\mathcal{S}}_i^f)$. Finally, $P(\mathbf{x} \in \mathcal{S}_j^f)$ is probability function that maps a set of pixels belonging to a detection instance's mask $\mathcal{S}_j^f$ to spatial probability, and can be inferred from $\mathcal{S}_j^f$ and Σ_j^f.

The label quality on the other hand measures how well a detector captures the parameters describing the category label's categorical distribution and can be written as:

$$\mathcal{Q}_L(\mathcal{G}_i^f, \mathcal{D}_j^f) = \mathcal{I}_j^f(\hat{c}_i^f). \quad (4)$$

The label quality ranges between 0 and 1 and can be thought of as the probability of the correct category provided by the object detector. Finally, the pairwise PDQ is computed as the geometric mean of the label and spatial qualities of every detection, and can be written as:

$$pPDQ(\mathcal{G}_i^f, \mathcal{D}_j^f) = \sqrt{\mathcal{Q}_S(\mathcal{G}_i^f, \mathcal{D}_j^f).\mathcal{Q}_L(\mathcal{G}_i^f, \mathcal{D}_j^f)}. \quad (5)$$

PDQ is a very strong measure for the probability mass assigned by the detector to **true positive detections**. Furthermore, PDQ uses optimal assignment through the Hungarian algorithm to assign every ground truth to its best corresponding detection, removing the dependency on IOU thresholding that is required for AP and MUE. However, PDQ is evaluated at only a single classification score threshold. Also, PDQ only accounts for the probability distribution of true positive detections, a characteristic that can be exploited by detectors to obtain high scores that do not reflect the true quality of its provided distributions.

First, a perfect label quality of 1.0 can be achieved by using a one-hot probability vector $\mathcal{I}_j^f$. More formally:

$$\begin{aligned} \mathcal{I}_j^f \in \{0,1\}^K : \sum_{k=1}^{K} I_{jk}^f = 1.0 \\ \mathcal{Q}_L(\mathcal{G}_i^f, \mathcal{D}_j^f) = \mathcal{I}_j^f(\hat{c}_i^f) = 1.0, \end{aligned} \quad (6)$$

for any f, i, j combination. In such cases however, the detector is assigning a probability of 1.0 to all **false positive detections** as well. Since the explicit label quality of false positives is not penalized, detectors can always achieve gains in PDQ by simply using a one-hot representation of $\mathcal{I}_j^f$, maximizing their confidence in every detection and eliminating any value of a category's uncertainty measure.

Second, for detectors with good localization, a perfect spatial quality of 1.0 can be achieved by using a Dirac-δ function for $P(x \in \mathcal{S}_j^f)$. Specifically, a detector with good localization decreases $\hat{\mathcal{V}}_{ij}^f$, the set of pixels belonging to a detection but not to a ground truth instance. A perfect spatial quality score can then be achieved by assigning the detection a low variance distribution $P(x \in \mathcal{S}_j^f)$, which maximizes $L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f)$ and hence achieves higher PDQ scores. More formally, in the extreme case where $\hat{\mathcal{V}}_{ij}^f = \emptyset$:

$$\begin{aligned} \mathcal{Q}_S(\mathcal{G}_i^f, \mathcal{D}_j^f) &= exp(-(\underbrace{L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f)}_{\sim 0} + \underbrace{L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)}_{\sim 0}) \\ &= 1.0, \end{aligned} \quad (7)$$

where $L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f) \sim 0$ from using a $dirac - \delta$ function for $P(x \in \mathcal{S}_j^f)$, and $L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f) \sim 0$ from $\hat{\mathcal{V}}_{ij}^f = \emptyset$. For less extreme situations where the detector has some mistakes in localization, the terms in Eq. (7) can be controlled by treating the covariance matrix describing $P(x \in \mathcal{S}_j^f)$ as a hyper-parameter. By taking into account the localization accuracy of a detector, the covariance matrix can be optimized for a balance between $L_{FG}(\mathcal{G}_i^f, \mathcal{D}_j^f)$ and $L_{BG}(\mathcal{G}_i^f, \mathcal{D}_j^f)$ that increases the spatial quality and therefore the PDQ.

III. Accounting For False Positives In PDQ

To take into account the quality of distributions assigned by the detector to detections that are false positives, we propose two modifications to the original PDQ. Our proposed modifications exploit the fact that false positive detections are mis-classified instances of the **background** class in order to formulate spatial and label quality metrics for false positives. Specifically, a detector should aim to assign a lower probability mass to the bounding box and category of false positive detections.

This characteristic can be captured for the label quality metric by defining a label quality term for every false positive detection as:

$$\mathcal{Q}_{L-FP}(\mathcal{D}_j^f) = 1.0 - \max(\mathcal{I}_j^f), \quad (8)$$

where $\mathcal{I}_j^f$ is the category probability vector associated with the false positive detection $\mathcal{D}_j^f$. In a similar manner, the spatial quality metric can be formulated for false positive detections as:

$$\mathcal{Q}_{S-FP}(\mathcal{D}_j^f) = \exp\left(-(L_{BG-FP}(\mathcal{D}_j^f)\right), \quad (9)$$

where $L_{BG-FP}(\mathcal{D}_j^f)$ is the false positive background loss defined as:

$$L_{BG-FP}(\mathcal{D}_j^f) = -\frac{1}{|\mathcal{S}_i^f|} \sum_{x \in \mathcal{S}_i^f} \log\left((1 - P(x \in \mathcal{S}_j^f)\right),$$

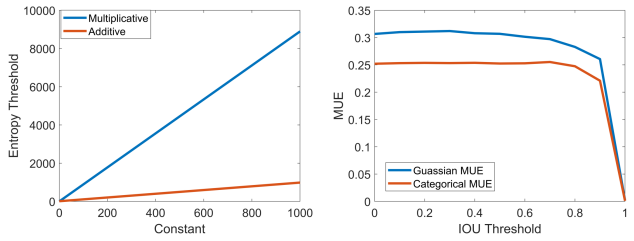

Fig. 1. **Left**: Variation in σ_{ent} as a function of transforming the Gaussian entropy with additive and multiplicative constants. The Gaussian MUE remains constant at 0.36 for all transformations. **Right**: Variation of Gaussian and Categorical MUE as a function of the iou threshold used to determine true and false positives. Detections from BayesOD [5] on the COCO validation dataset were used to generate both plots.

where $\mathcal{S}_i^f$ is the number of pixels in the false positive detection mask. The pairwise PDQ can then be computed for false positive detections in a similar manner as the true positive detections, and can be added inside the summation in Eq. (2) to arrive at a modified form of the PDQ. In conclusion, the proposed modification of Eq. (2) allows a detector to increase the PDQ score not only by decreasing the number of false positives, but also by giving false positives lower label and spatial probability mass.

IV. Experiments and Results

Minimum Uncertainty Error: Fig. 1 shows the change in the entropy threshold σ_{ent} when an additive or a multiplicative constant is applied to the Gaussian entropy used to determine the Gaussian MUE. In this case, the Gaussian MUE remains constant at 0.36 regardless of the transforming constant, which σ_{ent} compensates for through a transformation in magnitude. Fig. 1 also shows the variation of Gaussian and Categorical MUE as a function of the IOU threshold used to determine what counts as a true positive detection. It can be seen that both types of MUE have low sensitivity to the actual value of the IOU threshold.

Probability Based Detection Quality: For controlled testing of our proposed modifications of PDQ, we randomly generate 8000 true and false positives in a 1:1 ratio. Both true and false positives are copies of the ground truth, with true positives having a 1 IOU and false positives having 0 IOU. For true positives, we attach a 1.0 category score and an isotropic covariance matrix with a scalar value of 0.1 for the diagonal elements.

Fig. 2 shows the variation of the label quality as a function of the category score assigned to false positives for both the original PDQ and PDQ using our proposed modifications. It can be seen that as the false positive category score increases, the label quality decreases to a minimum of 0.5 at 1.0 category score. On the other hand, the label quality of the original PDQ remains constant at 1.0 regardless of the category score assigned to false positives. Similarly, the variation of the spatial quality as a function of the scalar value of the isotropic covariance matrix associated with false positives is shown in Fig. 2. As the scalar value increases, the spatial probability mass assigned to the false positives decrease and the spatial quality of PDQ using our

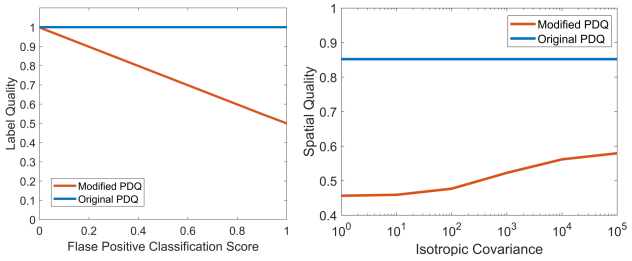

Fig. 2. Variation in modified and original forms of label and spatial quality as a function of the probability mass assigned to false positive detections. Both plots were generated from simulated detections.

Method	PDQ Calculation	mAP(%)	PDQ Score(%)	Spatial Quality	Label Quality
BayesOD	Original	34.77	22.64	0.373	0.644
	Modified	34.77	26.85	0.310	0.554
Black Box	Original	33.71	21.87	0.408	0.6978
	Modified	33.71	22.13	0.318	0.596

TABLE I

Comparison between the original PDQ and PDQ using our proposed modifications, tested on detections from BayesOD and Black Box on the COCO dataset.

modifications increases. It can be seen that the spatial quality of the original PDQ remains constant regardless of the value of the isotropic covariance associated with false positive detections. The above experiments confirm that unlike the original formulation of PDQ, our modifications allow PDQ to take into account the probability mass assigned to the category and bounding box of false positive detections.

To better show the results of the proposed modifications, we use both the original and the modified PDQ to evaluate real probabilistic detections from applying BayesOD [5] and Black Box [6] to the validation set of the COCO dataset. Table I shows that using the modified PDQ, the label quality of both detectors drop by around $13-15\%$, while their spatial quality drop by $16\% - 22\%$. This observation demonstrates that the original PDQ formulation misses a substantial effect of false positives on the quality of a probabilistic object detector.

Another observation from Table I is that the modified PDQ allows detectors to compensate for false positives by assigning them a lower category or bounding box probability mass. Being able to utilize such mechanism in a better manner than Black Box, BayesOD scored 4.72% higher when using the modified PDQ formulation, but only 0.77% higher when using the original PDQ formulation as it does not take into account false positives.

V. Conclusion

In this note, we provide a brief comparative study of common metrics used to evaluate probabilistic object detectors. We also provide a modification for PDQ to take into account the probability mass assigned by a probabilistic detector to false positive detections. As a result of this study, we recommend the research community rely on AP, MUE, and PDQ jointly for fair evaluation of probabilistic object detectors.

REFERENCES

[1] David Hall, Feras Dayoub, John Skinner, Peter Corke, Gustavo Carneiro, and Niko Sünderhauf. Probability-based detection quality (PDQ): A probabilistic approach to detection evaluation. *CoRR*, abs/1811.10800, 2018.

[2] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. *arXiv preprint arXiv:1809.06006*, 2018.

[3] 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.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[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.

[6] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.