BayesOD: Bayesian Inference for Fusing Epistemic and Aleatoric Uncertainty in Deep Object Detectors

Ali Harakeh, and Steven L. Waslander

Toronto Robotics and Artificial Intelligence Laboratory
University of Toronto Institute for Aerospace Studies
Goals:

• Provide uncertainty estimates for the **category** and **bounding box** states associated with detected object instances.

• Uncertainty estimates should be **meaningful**, and well correlated to the correctness of a detection.

• Allow the treatment of the neural network as a sensor within the robot sensor suite.
Related Work

**Sample-Free** [Feng, Rosenbaum, Dietmeyer 2018], [Le, Diehl, Brunner, Knol 2018]

- A single instance detector with box covariance regression

**Black Box** [Miller, Nicholson, Dayoub, Sünderhauf 2018], [Miller, Dayoub, Milford, Sünderhauf 2018]

- MC-dropout with T instances
- Cluster after NMS for each instance and estimate sample mean and covariance
Related Work

**Redundancy** [Le, Diehl, Brunner, Knoll 2018]

- Replace NMS with anchor output clustering and sample mean and covariance estimation

**BayesOD** [Harakeh, Waslander 2019]

- Combine box covariance regression, MC dropout and clustering for joint aleatoric and epistemic uncertainty estimation
- Incorporate Dirichlet and Gaussian priors on classification and regression of anchors and objects
- Perform Bayesian inference over anchor and MC dropout clusters, replacing NMS
Bayes OD – Box Covariance Regression

- Feature Extractor
- Anchor Level Priors
- Box Mean Regression
- Box Covariance Regression
- Classification
- Clustering
- Bayesian Inference
- Object Level Prior
- MC-Dropout
Estimating Per-Anchor Aleatoric Uncertainty:

\[ f(x_i, \theta) = \mathbb{E}[p(\hat{B}_i | x_i, D, \theta)] \]
\[ p(\hat{S}_i | x_i, D, \theta) = \text{Cat}(\hat{p}_1, \ldots, \hat{p}_K) \]

\[ p(\hat{B}_i | x_i, D, \theta) = \mathcal{N}(\mu(x_i, \theta), \Sigma_a(x_i, \theta)) \]
\[ p(\hat{S}_i | x_i, D, \theta) = \text{Cat}(\hat{p}_1, \ldots, \hat{p}_K) \]

- Aleatoric uncertainty already captured by the parameters of the categorical distribution.
- The only requirement is to estimate the covariance matrix of the Gaussian distribution describing objects' bounding boxes.
Learning to Estimate a Multivariate Covariance Matrix:

• Extension to stable multi-variate covariance training through LDL matrix decomposition:

\[ \Sigma_a(x_i, \theta) = L(x_i, \theta)D(x_i, \theta)L(x_i, \theta)^\top \]

\[
\begin{array}{ccc}
1 & 0 & 0 \\
a & 1 & 0 \\
b & c & 1 \\
\end{array}
\]

\[
\begin{array}{ccc}
e & 0 & 0 \\
f & 0 & 0 \\
g & 0 & 0 \\
\end{array}
\]

• Learning the components through cross-(differential) entropy loss function:

\[
\min L(x_i, \theta) = \frac{1}{2} \left[ ||D(x_i, \theta)^{-1/2}\text{adj}(L(x_i, \theta))\mu(x_i, \theta) - \mu_{gt}||^2_2 + \text{Tr}(\log(D(x_i, \theta))) \right]
\]
Bayes OD – Epistemic Uncertainty

Anchor Level Priors

Feature Extractor

Box Mean Regression

Box Covariance Regression

Classification

Clustering

Bayesian Inference

Object Level Prior

MC-Dropout
Estimating Per-Anchor Epistemic Uncertainty

Given
\[
p(\hat{B}_i | x_i, \mathcal{D}, \theta) = \mathcal{N}(\mu(x_i, \theta), \Sigma(a(x_i, \theta)))
\]
\[
p(\hat{S}_i | x_i, \mathcal{D}, \theta) = \text{Cat}(\hat{p}_1, \ldots, \hat{p}_K)
\]

Required
\[
p(\hat{B}_i | x_i, \mathcal{D}) = \mathcal{N}(\mu(x_i), \Sigma(x_i))
\]
\[
p(\hat{S}_i | x_i, \mathcal{D}) = \text{Cat}(\hat{p}_1, \ldots, \hat{p}_K)
\]

• Use Bayesian Neural Networks to eliminate the dependence on a point estimate of the parameters theta:
\[
p(\hat{B}_i | x_i, \mathcal{D}) = \int_{\theta} p(\hat{B}_i | x_i, \mathcal{D}, \theta)p(\theta | \mathcal{D})d\theta
\]
\[
p(\hat{S}_i | x_i, \mathcal{D}) = \int_{\theta} p(\hat{S}_i | x_i, \mathcal{D}, \theta)p(\theta | \mathcal{D})d\theta
\]
Estimating Per-Anchor Epistemic Uncertainty

- Monte-Carlo Dropout as variational Bayesian approximation of the integral:

**Regression**

\[
p(\hat{Y}_i | x_i, D) = N(\mu(x_i), \Sigma(x_i))
\]

\[
\mu(x_i) = \frac{1}{T} \sum_{t=1}^{T} f(x_i, \theta_t)
\]

\[
\Sigma(x_i) = \frac{1}{T} \left( \sum_{t=1}^{T} f(x_i, \theta_t) f(x_i, \theta_t)^\top \right)
\]

\[- \mu(x_i)\mu(x_i)^\top + \frac{1}{T} \sum_{t=1}^{T} \Sigma_\alpha(x_i, \theta_t)\]

**Classification**

\[
p(\hat{S}_i | x_i, D) = Cat([\hat{p}_1, \ldots, \hat{p}_K])
\]

\[
\hat{p}_k = \frac{1}{T} \sum_{t=1}^{T} g(x_i, \theta_t)
\]
Bayes OD - Per-Anchor Probability Distributions

\[
p(\hat{B}_i|x_i, D) = \mathcal{N}(\mu(x_i), \Sigma(x_i))
\]
\[
p(\hat{S}_i|x_i, D) = \text{Cat}(\hat{p}_1, \ldots, \hat{p}_K)
\]
Bayes OD – Incorporating Priors

- Feature Extractor
- Box Mean Regression
- Classification
- Box Covariance Regression
- Bayesian Inference
- Anchor Level Priors
- Object Level Prior
- MC-Dropout

Clustering
Bounding Box Per-Anchor Posterior

- The bounding box per anchor posterior distribution can be written as:

\[ p(B|x_i, D, \hat{B}_i) \propto p(\hat{B}_i|x_i, D, B)p(B|x_i) \]

\[ N(\mu(x_i), \Sigma(x_i)) \quad N(\mu_0, \Sigma_0) \]

- Can be computed in closed form as:

\[ p(B|x_i, D, \hat{B}_i) = N(\mu'(x_i), \Sigma'(x_i)) \]

\[ \Sigma'(x_i) = (\Sigma_0^{-1} + \Sigma(x_i)^{-1})^{-1} \]

\[ \mu'(x_i) = \Sigma'(x_i)(\Sigma_0^{-1}\mu_0 + \Sigma(x_i)\mu(x_i)). \]
Bayes OD – Incorporating Priors

Feature Extractor

Box Mean Regression
Classification
Box Covariance Regression
Bayesian Inference
Anchor Level Priors
Object Level Prior
Clustering
MC-Dropout

Anchor Level Priors
Box Covariance Regression
Box Mean Regression
Classification
Clustering
Bayesian Inference
Object Level Prior

MC-Dropout
Bayesian Inference Over Object Clusters

• **Step 1:** Cluster detections using your favorite clustering algorithm [Miller Dayoub, Milford, Sünderhauf, 2019].

• **Step 2:** Use cluster members to update the states of the center of the cluster. **Assume independent measurements.**

\[
p(B|\mathcal{X}, D, [\hat{B}_1, \ldots, \hat{B}_M]) \propto p(B|x_1, D, \hat{B}_1) \prod_{i=2}^{M} p(\hat{B}_i|x_i, D, B)
\]

\[
= \mathcal{N}(\mu'', (\mathcal{X}), \Sigma''(\mathcal{X}))
\]

\[
\Sigma''(\mathcal{X}) = \left( \sum_{i=1}^{M} \Sigma'(x_i)^{-1} \right)^{-1}
\]

\[
\mu''(\mathcal{X}) = \Sigma''(\mathcal{X}) \left( \sum_{i=1}^{M} \Sigma'(x_i)^{-1} \mu'(x_i) \right)
\]

• **Alternative Step 2:** Inverse Covariance Intersection to estimate correlation between anchor measurements.
Results - MUE and AP

- We used RetinaNet as 2D object detector
- Trained on BDD, tested on BDD, KITTI
- Minimum Uncertainty Error Evaluation:
  - Categorical Minimum Uncertainty Error (CMUE): Lowest possible uncertainty error using a threshold on the categorical entropy.
  - Gaussian Minimum Uncertainty Error (GMUE): Lowest possible uncertainty error using a threshold on the gaussian entropy.

\[
UE(\delta) = 0.5 \frac{|TP > \delta|}{|TP|} + 0.5 \frac{|FP \leq \delta|}{|FP|}
\]
Results - MUE and AP

- Bayesian inference treats anchors as separate measurements, and fuses them
- Leads to better estimates and higher AP
- Significant improvement in both Gaussian and Categorical Minimum Uncertainty Error

<table>
<thead>
<tr>
<th>Test Dataset</th>
<th>Method</th>
<th>AP(%) ↑</th>
<th>GMUE(%) ↓</th>
<th>CMUE(%) ↓</th>
<th>AP(%) ↑</th>
<th>GMUE(%) ↓</th>
<th>CMUE(%) ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black Box [13, 14]</td>
<td>57.34</td>
<td>49.75</td>
<td>21.71</td>
<td>41.54</td>
<td>49.86</td>
<td>29.43</td>
</tr>
<tr>
<td></td>
<td>Redundancy [16]</td>
<td>56.43</td>
<td>49.71</td>
<td>24.80</td>
<td>40.43</td>
<td>49.96</td>
<td>38.56</td>
</tr>
<tr>
<td></td>
<td>Ours (Diagonal)</td>
<td>60.98</td>
<td>25.76</td>
<td>16.67</td>
<td>42.97</td>
<td>26.68</td>
<td>22.72</td>
</tr>
<tr>
<td></td>
<td>Ours (Full Covar)</td>
<td>60.79</td>
<td>25.64</td>
<td>16.50</td>
<td>42.05</td>
<td>27.25</td>
<td>23.02</td>
</tr>
</tbody>
</table>
Results - Inverse Covariance Intersection

- Bayesian inference treats anchors as separate measurements, and fuses them
- Leads to better estimates and higher AP
- Significant improvement in both Gaussian and Categorical Minimum Uncertainty Error
- Inverse Covariance Intersection leads to reduced AP, increased GMUE

<table>
<thead>
<tr>
<th>Test Dataset</th>
<th>Method</th>
<th>AP(%)</th>
<th>GMUE(%)</th>
<th>CMUE(%)</th>
<th>AP(%)</th>
<th>GMUE(%)</th>
<th>CMUE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Black Box [13, 14]</td>
<td>57.34</td>
<td>49.75</td>
<td>21.71</td>
<td>41.54</td>
<td>49.86</td>
<td>29.43</td>
</tr>
<tr>
<td></td>
<td>Redundancy [16]</td>
<td>56.43</td>
<td>49.71</td>
<td>24.80</td>
<td>40.43</td>
<td>49.96</td>
<td>38.56</td>
</tr>
<tr>
<td></td>
<td>Ours (Diagonal)</td>
<td>60.98</td>
<td>25.76</td>
<td>16.67</td>
<td><strong>42.97</strong></td>
<td><strong>26.68</strong></td>
<td><strong>22.72</strong></td>
</tr>
<tr>
<td></td>
<td>Ours (Full Covar)</td>
<td>60.79</td>
<td>25.64</td>
<td><strong>16.50</strong></td>
<td>42.05</td>
<td>27.25</td>
<td>23.02</td>
</tr>
<tr>
<td></td>
<td>Ours (Full Covar + ICI)</td>
<td>60.63</td>
<td>36.04</td>
<td>16.83</td>
<td>41.65</td>
<td>40.00</td>
<td>23.76</td>
</tr>
</tbody>
</table>
PDQ Score

- PDQ results show some complex behaviours
  - PDQ spatial quality does not seem to be positively correlated to minimum uncertainty error or AP performance in these comparisons.
  - ICI has a strong positive influence on spatial quality

<table>
<thead>
<tr>
<th>Method</th>
<th>Score ↑</th>
<th>TP ↑</th>
<th>FP ↓</th>
<th>FN ↓</th>
<th>Spatial Quality(%) ↑</th>
<th>Label Quality(%) ↑</th>
<th>Overall Quality(%) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Free</td>
<td>32.98</td>
<td>71597</td>
<td>12608</td>
<td>36897</td>
<td>45.47</td>
<td>81.78</td>
<td>55.79</td>
</tr>
<tr>
<td>Black Box</td>
<td>27.80</td>
<td>69274</td>
<td>13709</td>
<td>39220</td>
<td>39.85</td>
<td>78.47</td>
<td>49.04</td>
</tr>
<tr>
<td>Redundancy</td>
<td>28.24</td>
<td>70323</td>
<td>13916</td>
<td>38171</td>
<td>46.57</td>
<td>63.44</td>
<td>49.17</td>
</tr>
<tr>
<td>Ours (Diag )</td>
<td>17.92</td>
<td>73757</td>
<td>31881</td>
<td>35059</td>
<td>24.16</td>
<td>77.82</td>
<td>34.20</td>
</tr>
<tr>
<td>Ours (Full Covar)</td>
<td>17.84</td>
<td>73453</td>
<td>31977</td>
<td>35342</td>
<td>24.16</td>
<td>77.78</td>
<td>34.20</td>
</tr>
<tr>
<td>Ours (Full Covar + ICI)</td>
<td>31.20</td>
<td>76207</td>
<td>25464</td>
<td>32588</td>
<td>46.93</td>
<td>77.52</td>
<td>54.97</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Dataset</th>
<th>Method</th>
<th>Car</th>
<th>Pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDD [21]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sampling Free [16, 15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black Box [13, 14]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Redundancy [16]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (Diagonal)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (Diagonal + ICI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours (Full Covar + ICI)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AP(%) ↑</td>
<td>GMUE(%) ↓</td>
<td>CMUE(%) ↓</td>
</tr>
<tr>
<td></td>
<td>55.16</td>
<td>38.99</td>
<td>21.96</td>
</tr>
<tr>
<td></td>
<td>57.34</td>
<td>49.75</td>
<td>21.71</td>
</tr>
<tr>
<td></td>
<td>56.43</td>
<td>49.71</td>
<td>24.80</td>
</tr>
<tr>
<td></td>
<td>61.35</td>
<td>25.53</td>
<td>16.96</td>
</tr>
<tr>
<td></td>
<td>60.59</td>
<td>36.94</td>
<td>16.87</td>
</tr>
<tr>
<td></td>
<td>60.63</td>
<td>36.04</td>
<td>16.83</td>
</tr>
</tbody>
</table>
• Can achieve 100% label quality by replacing predicted category probability vector with a one hot vector.

• We have a few ideas to tackle this issue, mainly by incorporating the predicted probability of false positives into the label quality.

<table>
<thead>
<tr>
<th>Method</th>
<th>Score ↑</th>
<th>TP ↑</th>
<th>FP ↓</th>
<th>FN ↓</th>
<th>Spatial Quality(%) ↑</th>
<th>Label Quality(%) ↑</th>
<th>Overall Quality(%) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling Free</td>
<td>35.79</td>
<td>71197</td>
<td>12503</td>
<td>37297</td>
<td>45.62</td>
<td>100.00</td>
<td>60.83</td>
</tr>
<tr>
<td>Black Box</td>
<td>30.45</td>
<td>68999</td>
<td>13524</td>
<td>39495</td>
<td>39.99</td>
<td>100.00</td>
<td>53.86</td>
</tr>
<tr>
<td>Redundancy</td>
<td>34.84</td>
<td>69987</td>
<td>13730</td>
<td>38507</td>
<td>46.73</td>
<td>100.00</td>
<td>60.85</td>
</tr>
<tr>
<td>Ours (Full Covar + ICI)</td>
<td>34.60</td>
<td>75897</td>
<td>25784</td>
<td>32898</td>
<td>47.11</td>
<td>100.00</td>
<td>61.36</td>
</tr>
</tbody>
</table>
Next Steps

• Further investigate the effect of correlated measurements.

• Uninformative priors only the first step, dataset based anchor priors and tracking based object priors to be exploited.

• Alternatives to MC Dropout for epistemic uncertainty estimation
Extra Slides
Quantitative Representations
Quantitative Representations
Quantitative Representations (Kitti Dataset)