Out-of-Distribution Detection Using Deep Neural Network Latent Space Uncertainty

SafeAI 2023 @ AAAI-23

Fabio Arnez, Ansgar Radermacher, François Terrier
Semantic segmentation predictive uncertainty estimation comparison for In-Distribution (InD) and Out-of-Distribution (OoD) data.

**Contribution**

*Use the Deep Neural Network Uncertainty (Entropy) from intermediate latent representations for the Out-of-Distribution Detection Task*
Semantic Segmentation Architectures: Capturing Uncertainty from Intermediate Latent Representations

1. We apply Monte-Carlo DropBlock2D to capture epistemic uncertainty from the Encoder, Prior Net in the probabilistic U-Net Architecture.

\[ q_{\phi}(z \mid x, D_p) = \int q(z \mid x, \phi)p(\phi \mid D_p)d\phi \]

\[ \phi_m \sim p(\phi \mid D_p) \]

\[ \Phi = \{\phi_m\}_m^M \]

2. Obtain the average surprise (Entropy) of observing latent representation \( z \) at the output of the Encoder (Prior Net), for an input image \( x \)

\[ H(z \mid x) = \int_{z} p(z \mid x) \log \frac{1}{p(z \mid x)}dz \]
For each type of sample, obtain the Entropy empirical densities $T(z)$ using a Kernel Density Estimation (KDE).

$$Y = \{\text{normal, anomaly}\}$$

For an unknown latent representation $z$, we can compute the posterior probability of each class (normal, anomaly) using Bayes’ Rule:

$$p(y \mid T(z)) = \frac{p(T(z) \mid y)p(y)}{\sum_{y \in Y} p(T(z) \mid y)p(y)}$$

Class Label Prior
- counting frequencies - samples in each class

Figure 5: Empirical latent representation uncertainty density (KDE) comparison for InD and OoD samples
Experiments and Results 1

- Early Experiments and results with **Probabilistic U-Net** architecture:

Figure 6: Woodscape dataset (InD) vs. Woodscape-soiling (OoD) samples in custom dataset

Figure 7: Empirical densities with KDE

Table 1: Evaluation of OoD detection methods using DNN latent representations

<table>
<thead>
<tr>
<th>Method</th>
<th>MCC</th>
<th>F1</th>
<th>AUROC</th>
<th>FPR90</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_M$</td>
<td>0.473</td>
<td>0.763</td>
<td>0.769</td>
<td>0.5</td>
</tr>
<tr>
<td>$\hat{H}_\phi(z_i \mid x)$</td>
<td>0.572</td>
<td>0.797</td>
<td>0.855</td>
<td>0.4</td>
</tr>
<tr>
<td>$\hat{H}_\phi(z_i \mid x)$</td>
<td><strong>0.685</strong></td>
<td><strong>0.849</strong></td>
<td><strong>0.946</strong></td>
<td><strong>0.16</strong></td>
</tr>
</tbody>
</table>

Figure 8: OoD Detection methods ROC Curves
Experiments and Results 2

- Experiments and results with DeeplabV3+ architecture:

Figure 9: InD datasets – Woodscape dataset and Cityscapes datasets

Figure 10: OoD Datasets - Woodscape-anomalies, Cityscapes-anomalies, and Woodscape-soiling

Table 2: OoD Detection evaluation using DNN latent representations Entropy

| InD Dataset (training) | OoD Dataset           | OoD Method: $\hat{H}_θ(z_i | x)$ |
|------------------------|-----------------------|-----------------------------------|
|                        | MCC (↑)   | F1 (↑)   | AUROC (↑) | FPR95 (↓) |
| Woodscape             | Woodscape-anomalies  | 0.9698   | 0.9847   | 0.9983   | 0.0049   |
| Cityscapes            | Cityscapes-anomalies | 0.9981   | 0.9993   | 1.0      | 0.0      |
| Woodscape            | Woodscape-soiling    | 0.9770   | 0.9894   | 0.9997   | 0.0      |
| Cityscapes            | Cityscapes           | 0.9850   | 0.9924   | 0.9999   | 0.0      |
| Woodscape            | Woodscape            | 0.9991   | 0.994    | 1.0      | 0.0      |
| Woodscape            | Woodscape-soiling    | 0.9983   | 0.9990   | 1.0      | 0.0      |
Conclusion

- We use the uncertainty from intermediate latent representations for Out-of-distribution detection in a semantic segmentation task.

- Our results show that using the entropy from latent representation can be useful to build classifiers that act as data-driven monitoring functions.

- Our uncertainty estimation method relies on samples which can be prohibitive in systems with tight time constraints.

- In future work, we aim to explore the impact of the structure in the latent space and its effect on the statistics and metrics used for the OoD detection task.