**Likelihood Ratios for Out-of-Distribution Detection**

Jie Ren*, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, Balaji Lakshminarayanan*

### 1. Introduction

- Discriminative models offer little performance guarantees on out-of-distribution (OOD) inputs.
- Limiting the AI safety in real-world applications.
- Bacteria identification based on genomic sequences holds the promise of early detection of disease.
- ML classifiers perform poorly in real-world, because real data contains 60-80% genomic sequences from unknown bacteria and other contaminants.
- We create a realistic benchmark for OOD detection on genomics data.
- We propose a Likelihood Ratio method for OOD detection, achieving SOTA on genomics data.

<01/01/2011 01/01/2011–01/01/2016 >01/01/2016

In-distribution training In-distribution validation In-distribution test

In-distribution, 60 OOD validation, 60 OOD test classes.

Bacterial classes are discovered gradually over the years (not saturated yet).

In-distribution and OOD classes are interlaced in phylogeny.

### 2. Generative Models Can Assign Higher Likelihood to OOD Inputs

- Generative models:
  - do not require labeled data
  - model the input distribution $p(x_{true})$ and then evaluate the likelihood of new inputs.

- Higher likelihood for OOD than in-dist. in Fashion-MNIST (in-dist.) vs. MNIST (OOD).

- We observe a similar phenomenon on genomic sequences.

- The likelihood is heavily affected by the sequence’s GC-content (background statistics).

### 3. Likelihood Ratios For OOD Detection

- Assumption: An input $x$ is composed of two components
  - Background $x_b$: population level background statistics
  - Semantic $x_s$: in-dist. specific features. See examples.

  $p(x) = p(x_b) p(x_s) \Rightarrow \text{can be dominant}$

  - To focus on $x_s$ we propose (1) training a background model on perturbed inputs and (2) computing the likelihood ratio

  $\text{LLR}(x) = \log \frac{p(x)}{p_b(x)} = \log \frac{p(x_b) p(x_s)}{p_b(x_b)} \approx \log \frac{p(x_b)}{p_b(x_b)}$ assuming both models capture background equally well.

  - LLR is a background contrastive score: the significance of the semantics compared with the background.

Examples of Background vs Semantics:

- Images: background → objects
- Text: stop words → key words
- Genomics: background → motifs
- Speech: background noise → speaker

Algorithm:
- Fit $p_b(x)$ using in-distribution data
- Fit $p_b(x)$ using perturbed input data and optionally model regularization.
- Compute the likelihood ratio.
- Predict OOD if likelihood ratio is small.
*mutation rate and L2 coefficient are tuned using an independent OOD dataset different from test OOD.

### 4. OOD Detection For Images

- Investigate auto-regressive models: which pixels contribute the most to the likelihood (ratio)?
- Fashion-MNIST (in-dist.) vs. MNIST (OOD). PixelCNN+ model is trained on Fashion-MNIST.
- Likelihood is dominated by the background pixels $p(x_b)$ (or MNIST).
- Likelihood ratio focuses on the semantic pixels $\Rightarrow \text{LLR(Fashion-MNIST)} > \text{LLR(MNIST)}$

- Table:

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>0.115</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>0.977</td>
</tr>
<tr>
<td>Classifier-based plylx</td>
<td>0.679</td>
</tr>
<tr>
<td>Classifier-based Entropy</td>
<td>0.688</td>
</tr>
<tr>
<td>Classifier-based ODIN</td>
<td>0.620</td>
</tr>
<tr>
<td>Classifier Ensemble 5</td>
<td>0.832</td>
</tr>
<tr>
<td>Classifier-based Mahalanobis Distance</td>
<td>0.986</td>
</tr>
</tbody>
</table>

### 5. OOD Detection For Genomics

- LSTM model is trained using sequences from in-distribution classes.
- Likelihood Ratio significantly improves OOD Detection
- Effect of background GC-content is corrected
- OOD detection correlates with its distance to in-distribution

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>0.630</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>0.755</td>
</tr>
<tr>
<td>Classifier-based plylx</td>
<td>0.622</td>
</tr>
<tr>
<td>Classifier-based Entropy</td>
<td>0.622</td>
</tr>
<tr>
<td>Classifier-based ODIN</td>
<td>0.645</td>
</tr>
<tr>
<td>Classifier Ensemble 5</td>
<td>0.673</td>
</tr>
<tr>
<td>Classifier-based Mahalanobis Distance</td>
<td>0.496</td>
</tr>
</tbody>
</table>

*Pearson Correlation: 0.509 (LLR), 0.272 (ensemble)

Summary:
- Create a realistic benchmark dataset for OOD detection in genomics.
- Show that the likelihood from deep generative models can be confounded by background statistics.
- Propose a likelihood ratio method for OOD detection, outperforming the raw likelihood.
- Our method achieves state-of-the-art performance on genomic dataset.

Check the ArXiv Version for details

Contact: jiren@google.com, balajiln@google.com