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*

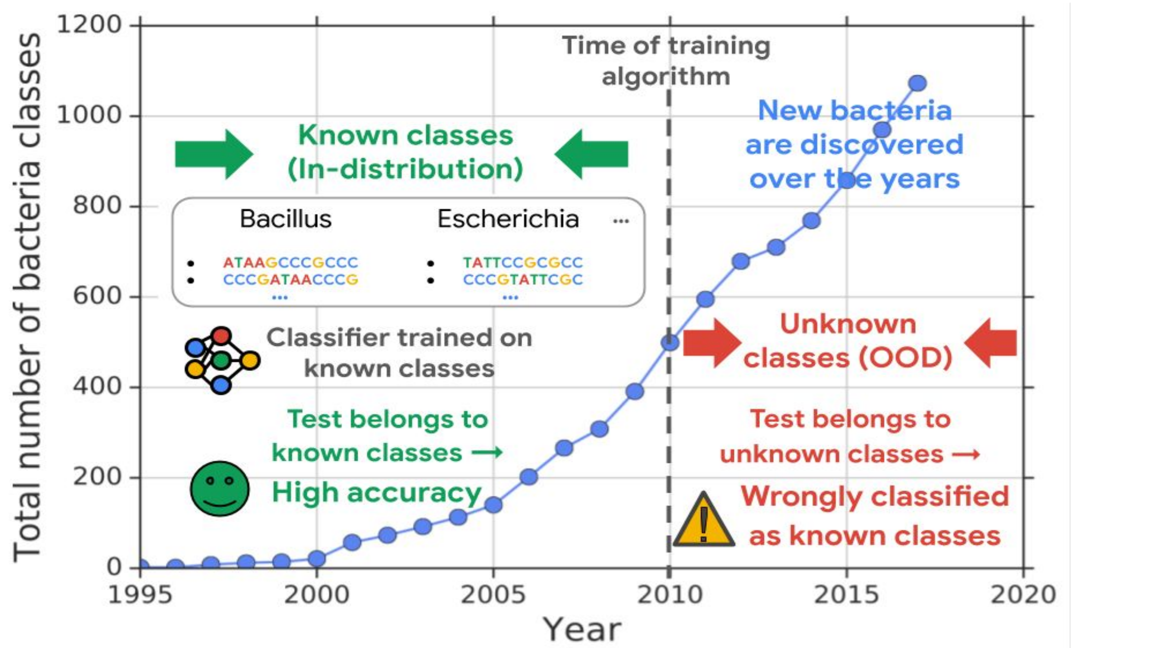
New benchmark dataset + code is available at
https://github.com/google-research/google-research/tree/master/genomics_ood

*Contact: {jjren, peterjliu, balajiln}@google.com

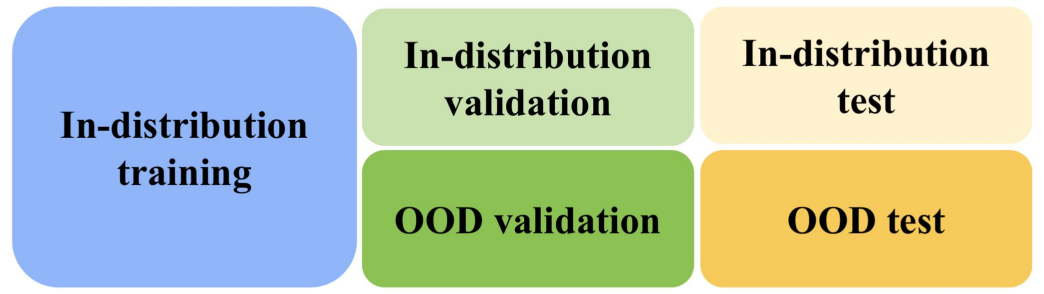
Google AI
DeepMind

1. Motivation

- **Bacteria identification based on genomic sequences**
  - ACGTTAACAACC...GGCTTC ⇒ label
  - Promising for early detection of disease
- Classifier can achieve high accuracy on known classes, but perform poorly in real world:
  - **60-80%** of real-world test inputs belong to as yet **unknown** bacteria
  - **Ideally, say "I don't know" on OOD inputs** than **assign high-confidence predictions**
- Need accurate OOD detection to ensure safe deployment of classifier



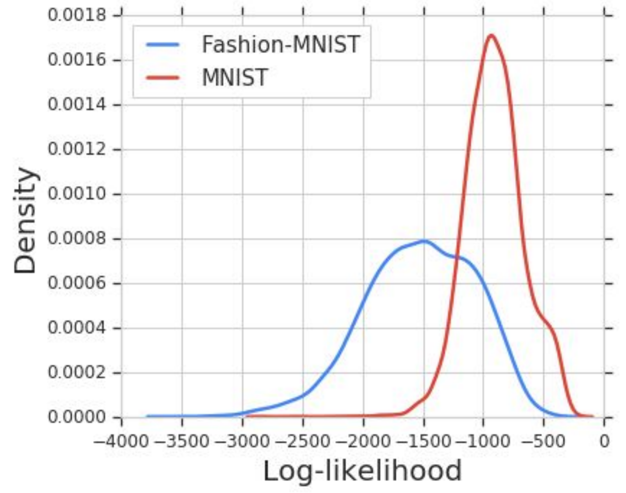
- We create **a realistic benchmark for OOD detection on genomics data**.
  - 10 in-distribution, 60 OOD validation, 60 OOD test classes.
  - Classes split by year to reflect challenges faced when classifier trained only on known classes

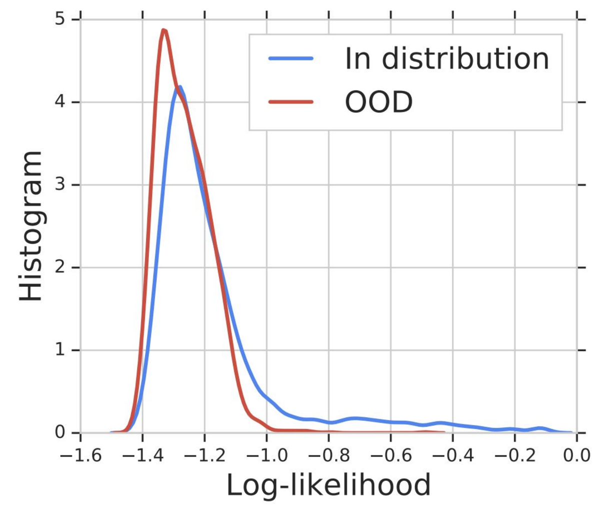


- **Challenge**: Detect if a test input is OOD (i.e. it does not belong to any of the training classes)
  - Unsupervised: Density-based approaches
  - Supervised: Classifier-based approaches

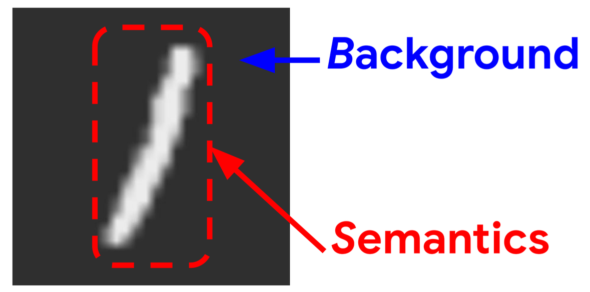
2. Generative Models Can Assign Higher Likelihood to OOD Inputs

- **Generative models for OOD detection:**
  - do not require labeled data
  - model the input distribution $p_{TRAIN}(\boldsymbol{x})$ and evaluate the likelihood of new inputs.
- Prior work [Nalisnick et al., 2018, Choi et al. 2019] observed failure modes of generative models: **Higher likelihoods for OOD than in-dist.** e.g. Fashion-MNIST (in-dist.) vs. MNIST (OOD)
- We observe a **similar failure mode on generative models trained on genomic sequences.**





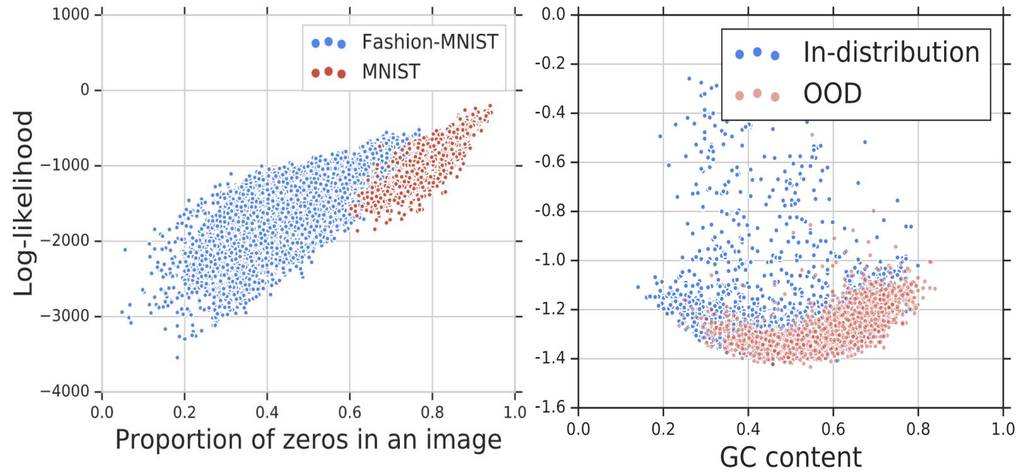
3. Explaining Why Density Models Fail At OOD Detection



- **B**ackground vs. **S**emantics Examples::
  - *Images*: background + objects
  - *Text*: stop words + key words
  - *Genomics*: GC background + motifs
  - *Speech*: background noise + speaker

$$p(\mathbf{x}) = \underbrace{p(\mathbf{x}_B)}_{\text{can be dominant}}\ \underbrace{p(\mathbf{x}_S)}_{\text{the focus}}$$

- $p(\boldsymbol{x})$ has to explain both semantic & background components
- Humans ignore background and focus primarily on semantics for OOD



Likelihood is highly correlated with the background

- proportion of zeros in an image
- GC-content in genomic sequence

4. Proposed Solution: Likelihood Ratios For OOD Detection

- How do we automatically extract the semantic component of p(x)?
- We propose **training a background model** on perturbed inputs and computing the **likelihood ratio**:

$$\mathrm{LLR}(\mathbf{x}) = \log\frac{p_\theta(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} = \log\frac{\cancel{p_\theta(\mathbf{x}_B)}\ p_\theta(\mathbf{x}_S)}{\cancel{p_{\theta_0}(\mathbf{x}_B)}\ p_{\theta_0}(\mathbf{x}_S)} \approx \log\frac{p_\theta(\mathbf{x}_S)}{p_{\theta_0}(\mathbf{x}_S)}$$

assuming both models capture background equally well.

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

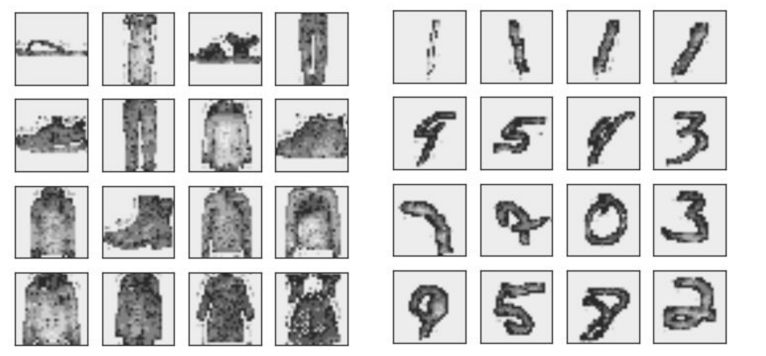
Algorithm

- Fit $p_\theta(\mathbf{x})$ using in-distribution data
- Fit $p_{\theta_0}(\mathbf{x})$ using perturbed input data and optionally model regularization*.
- Compute the likelihood ratio.
- Predict OOD if likelihood ratio is small.

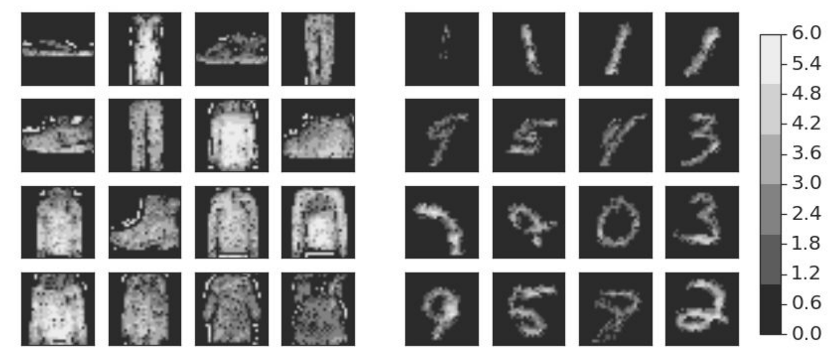
*Hyperparameters (mutation rate and L2 coefficient) are tuned using an independent OOD dataset different from test OOD.

5. OOD Detection For Images

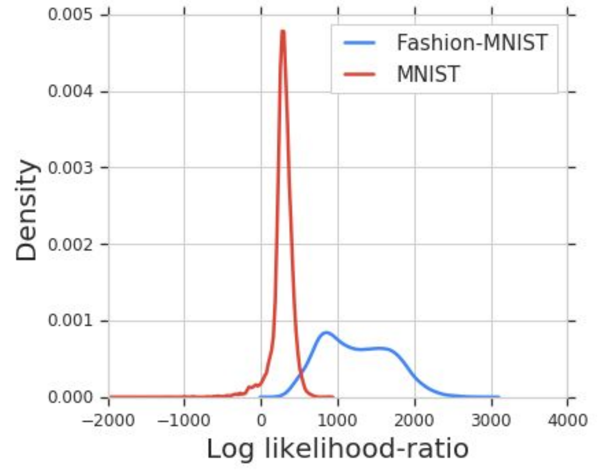
- Investigate auto-regressive models: ***which pixels contribute the most to the likelihood (ratio)?***
- **Fashion-MNIST (in-dist.) vs. MNIST (OOD).** PixelCNN++ model is trained on Fashion-MNIST.
- **Likelihood** is **dominated** by the **background pixels** ⇒ p(Fashion-MNIST) < p(MNIST)
- **Likelihood ratio** focuses on the **semantic pixels** ⇒ LLR(Fashion-MNIST) > LLR(MNIST)

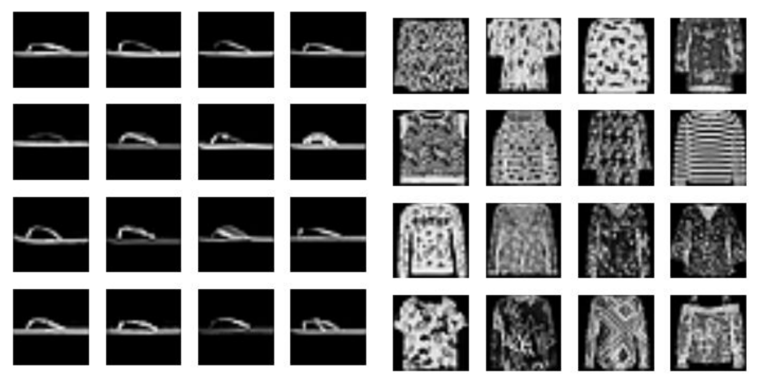


$\log p_\theta(x_d|x_{<d})$



$\log p_\theta(x_d|x_{<d}) - \log p_{\theta_0}(x_d|x_{<d})$





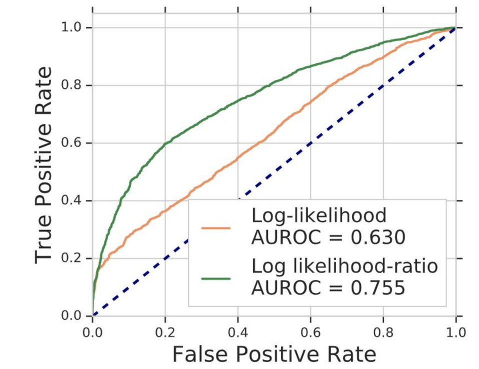
Images with highest (high portion of background) **and lowest likelihood**

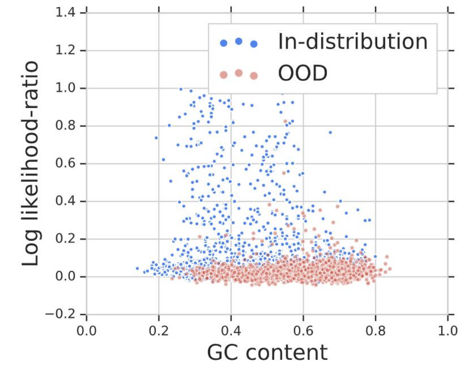
Images with highest (prototypical) **& lowest likelihood ratio** (rare patterns)

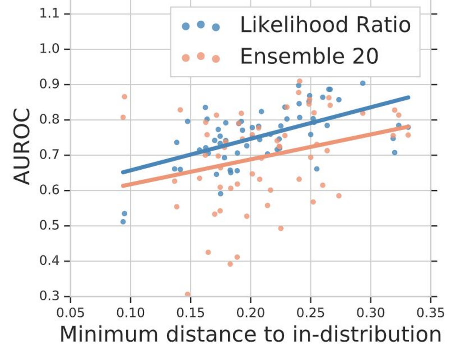
Method	AUROC
Likelihood	0.089
Likelihood Ratio	**0.994**
Classifier-based p(y|x)	0.734
Classifier-based Entropy	0.746
Classifier-based ODIN	0.752
Classifier Ensemble 5	0.839
Classifier-based Mahalanobis Distance	0.942

6. OOD Detection For Genomic Sequences

- 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







Method	AUROC
Likelihood	0.626
Likelihood Ratio	**0.755**
Classifier-based p(y|x)	0.634
Classifier-based Entropy	0.634
Classifier-based ODIN	0.697
Classifier Ensemble 5	0.682
Classifier-based Mahalanobis Distance	0.525

Summary

- Create a **realistic benchmark dataset** for OOD detection (and open-set classification) in genomics
- Show that the likelihood from deep generative models can be **confounded by background statistics**
- Propose a **likelihood ratio method for unsupervised OOD detection**, outperforming the raw likelihood
- Our method **performs well on images and achieves SOTA performance on genomic dataset.**