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*

Google AI DeepMind

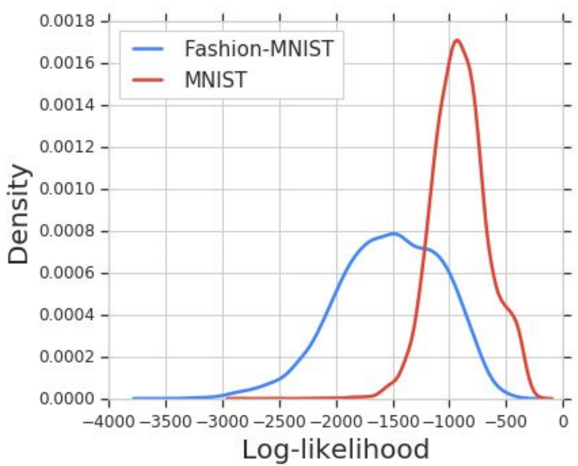
Motivation: Why is OOD detection important?

- **Bacteria identification based on genomic sequences**
  - **ACGTTAACAACC...GGCTTC ⇒ label**
  - Holds the promise of early detection of disease
- Classifier can achieve high accuracy on cross-validation
- But, the classifier can perform poorly in real world:
  - **60-80%** data belonging to as yet **unknown** bacteria
  - Assign high-confidence predictions to OOD inputs, than say “I don’t know”

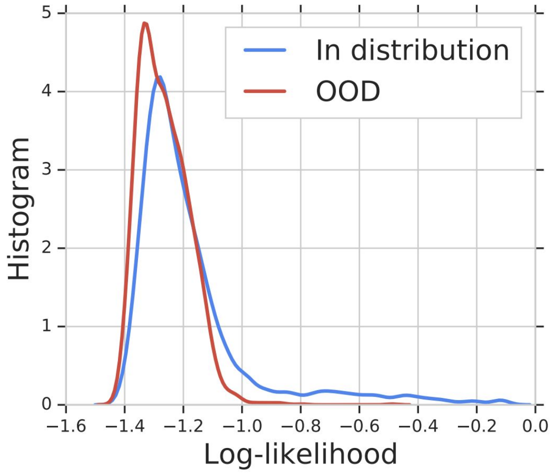
- **Need accurate OOD detection to ensure safe deployment of classifier**

Generative models for OOD detection?

- Pros: do not require labeled data; model the input distribution $p(\boldsymbol{x})$ and then evaluate the likelihood of new inputs
- Cons: can **assign higher likelihood to OOD** inputs!
  - Nalisnick et al., 2018, Choi et al. 2019.

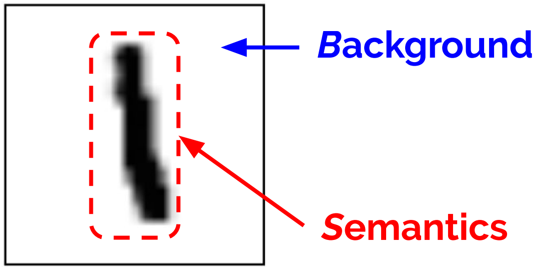


Fashion-MNIST (in-dist.) vs. MNIST (OOD)



Genomics

What does *p*(***x***) represent?



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

- Likelihood *p*(***x***) has to explain both semantic and background components

$$p(\mathbf{x}) = \boxed{p(\mathbf{x}_B)}\,\boxed{p(\mathbf{x}_S)}$$

can be dominant

the focus

- Humans ignore background and focus primarily on semantics for OOD
- **Question**: how do we automatically extract semantic component of p(x)?

Likelihood Ratio for OOD Detection

To focus on $\mathbf{x}_S$ we propose:

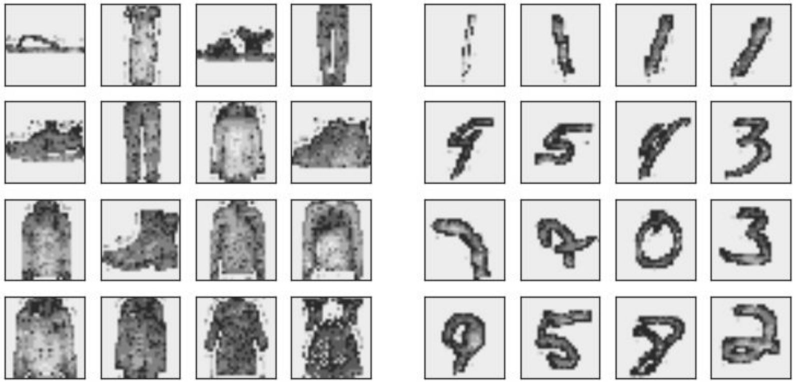
1. Training a **background model** on perturbed inputs
2. Computing the likelihood ratio

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

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

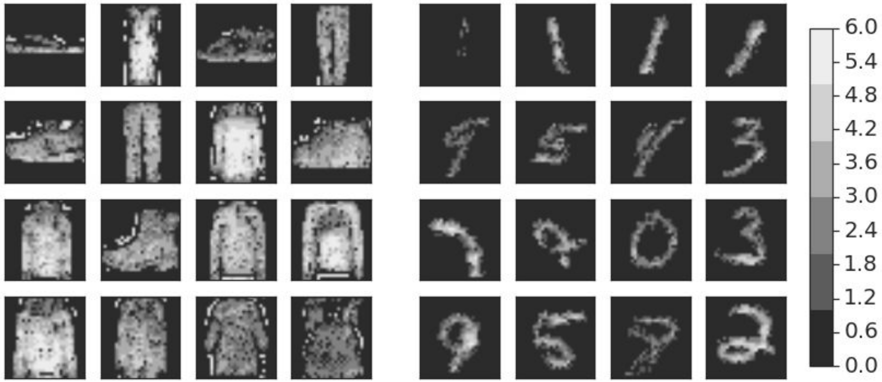
Which pixels contribute the most to likelihood (ratio)?

- PixelCNN++ model trained on FashionMNIST
- Heatmap showing per-pixel contributions on Fashion-MNIST (in-dist) and MNIST (OOD)



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

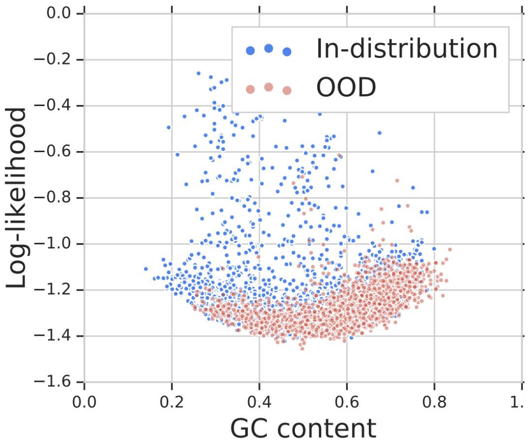
Likelihood is dominated by background pixels, which explains why MNIST (OOD) is assigned higher p(x)



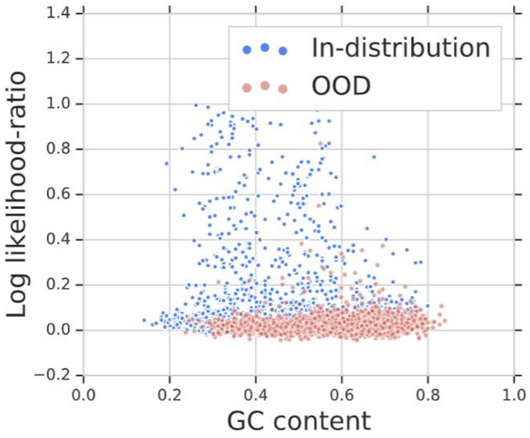
$$\log p_\theta(x_d|x_{<d}) - \log p_{\theta_0}(x_d|x_{<d})$$

Likelihood ratio focuses more on the semantic pixels and significantly outperforms likelihood on OOD detection

OOD detection for genomic sequences



Likelihood is heavily affected by GC bias



Likelihood Ratio corrects for GC bias

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

- Likelihood from deep generative models can be affected by background
- The proposed Likelihood Ratio method effectively corrects for background, and outperforms the raw likelihood on OOD detection

- Release a realistic benchmark dataset for OOD detection in genomics
- Our method achieves SOTA performance on genomic dataset

New benchmark dataset + code is available at

https://github.com/google-research/google-research/tree/master/genomics_ood