Likelihood Ratios for Out-of-Distribution Detection

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

Presented By:
Deepanshu

Department of Engineering Design
Indian Institute of Technology Madras, Chennai

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Outline

- Problem Statement
- Introduction
- Background
- High level idea
- Likelihood Ratio for OOD detection
    - Algorithm:Training the Background Model
    - Algorithm:LLR for OOD Detection
- Results
    - OOD detection for images
    - OOD detection for genomic sequences
- Comparison with baseline methods
- Summary
- Review
- References

Problem Statement 1/2

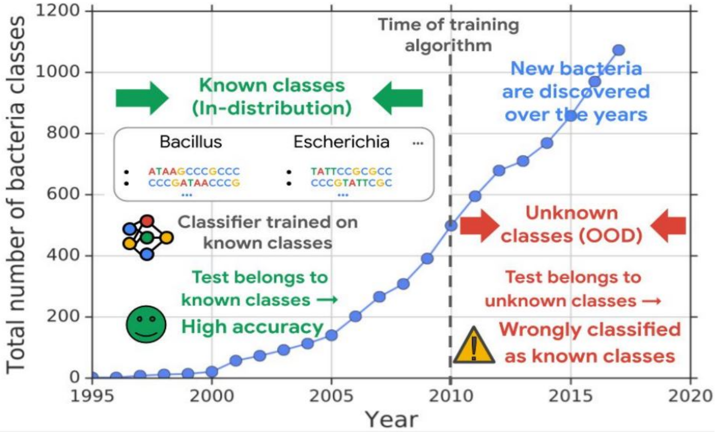

Figure 1: Bacteria identification based on genomic sequences

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Problem Statement 2/2

- Bacteria identification based on genomic sequences:
  - **ACGTTAACAACC...GGCTTC : label**
  - Promising for early detection of disease

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Problem Statement 2/2

- 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 percent 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**

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Problem Statement 2/2

- 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 percent 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**
- **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

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Problem Statement 2/2

- 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 percent 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**
- **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

"Need accurate OOD detection to ensure safe deployment of classifier"

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Introduction

In-distribution dataset D of (x,y) pairs sampled from the distribution p*(x,y):

- $x_d \in [A,C,G,T]$ for genomic sequences and $x_d \in [0,...,255]$ for images
- $y \in Y := [1,...,k,...,K]$ is the label

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Introduction

In-distribution dataset D of (x,y) pairs sampled from the distribution p*(x,y):

- $x_d \in$ [A,C,G,T] for genomic sequences and $x_d \in$ [0,...,255] for images
- $y \in Y :=$ [1,...,k,...,K] is the label

OOD inputs are samples (x,y) generated from distribution other than p*(x,y):

- An input (x,y) is OOD if $y \notin Y$
- Goal is to accurately detect if an input x is OOD or not

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Introduction

In-distribution dataset D of (x,y) pairs sampled from the distribution p*(x,y):

- $x_d \in$ [A,C,G,T] for genomic sequences and $x_d \in [0,...,255]$ for images
- $y \in Y := [1,...,k,...,K]$ is the label

OOD inputs are samples (x,y) generated from distribution other than p*(x,y):

- An input (x,y) is OOD if $y \notin Y$
- Goal is to accurately detect if an input x is OOD or not

Existing methods:

- **Classifier-based:** taking the confidence or entropy of the predictive distribution p(y|x)
- **Density-based:** fit a generative model p(x) to the input data, and then evaluate the likelihood of new inputs under that model

Background

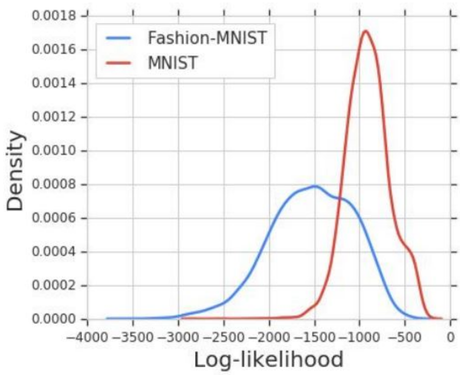

Figure 2: MNIST (OOD) vs Fashion-MNIST (in-dist.)
Prior work [Nalisnick et al., 2018, Choi et al. 2019]

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Background

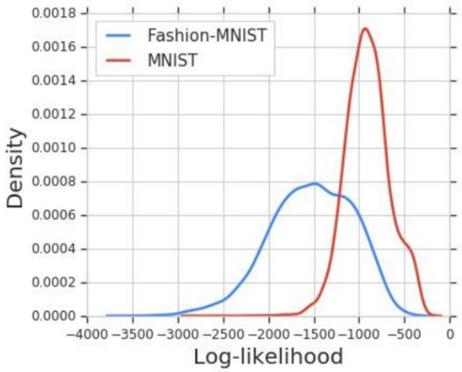

Figure 2: MNIST (OOD) vs Fashion-MNIST (in-dist.)
Prior work [Nalisnick et al., 2018, Choi et al. 2019]

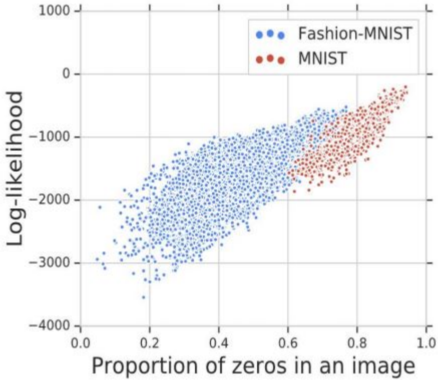

Figure 3: Likelihood is highly correlated with the background

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

High level idea 1/2

Background vs Semantic Component:

Assume that an input **x** is composed of two components:

$$x = x_B + x_S$$

- Background component (x_B) characterized by population level background statistics
- Semantic component (x_S) characterized by patterns specific to the in-distribution data

High level idea 1/2

Background vs Semantic Component:

Assume that an input **x** is composed of two components:

$$x = x_B + x_S$$

- Background component (x_B) characterized by population level background statistics
- Semantic component (x_S) characterized by patterns specific to the in-distribution data

Background vs. Semantics Examples:

- **Images: background + object**
- **Text: stop words + key words**
- **Genomics: GC content + motifs**
- **Speech: background noise + speaker**

Probabilistic Techniques laboratory

High level idea 2/2

For simplicity, assume that the background and semantic components are generated independently. The likelihood can be then decomposed as follows:

$$\boxed{p(x) = p(x_B) * p(x_S)} \tag{1}$$

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

High level idea 2/2

For simplicity, assume that the background and semantic components are generated independently. The likelihood can be then decomposed as follows:

$$\boxed{p(x) = p(x_B) * p(x_S)} \tag{1}$$

Assume that p_θ is a model trained using in-distribution data, and p_{θ_0} is a background model that captures general background statistics. A likelihood ratio statistic can be defined as:

$$\boxed{LLR(x) = log\frac{p_\theta(x)}{p_{\theta_0}(x)} = log\frac{p_\theta(x_B)p_\theta(x_S)}{p_{\theta_0}(x_B)p_{\theta_0}(x_S)}} \tag{2}$$

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

High level idea 2/2

For simplicity, assume that the background and semantic components are generated independently. The likelihood can be then decomposed as follows:

$$\boxed{p(x) = p(x_B) * p(x_S)} \tag{1}$$

Assume that p_θ is a model trained using in-distribution data, and p_{θ_0} is a background model that captures general background statistics. A likelihood ratio statistic can be defined as:

$$\boxed{LLR(x) = log \frac{p_\theta(x)}{p_{\theta_0}(x)} = log \frac{p_\theta(x_B)p_\theta(x_S)}{p_{\theta_0}(x_B)p_{\theta_0}(x_S)}} \tag{2}$$

Assume that both models capture the background information equally well:

$$\boxed{LLR(x) = log(p_\theta(x_S)) - log(p_{\theta_0}(x_S))} \tag{3}$$

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Likelihood Ratio for OOD detection 1/2

Algorithm 1:Training the Background Model

- **Inputs:** D-dimensional input $x = x_1 ... x_D$, $x_d \in F$, where F= [A,C,G,T] or [0,...,255]
- **Output:** perturbed input $\bar{x}$

Likelihood Ratio for OOD detection 1/2

Algorithm 1:Training the Background Model

- **Inputs:** D-dimensional input $x = x_1 ... x_D$, $x_d \in F$, where F= [A,C,G,T] or [0,...,255]
- **Output:** perturbed input $\bar{x}$
- Generate a D-dimensional vector $v = v_1 ..., v_D$, where $v_d \in [0, 1]$ are independent and identically distributed according to a Bernoulli distribution with rate μ

Algorithm 1:Training the Background Model

- **Inputs:** D-dimensional input $x = x_1 ... x_D$, $x_d \in F$, where F= [A,C,G,T] or [0,...,255]
- **Output:** perturbed input $\bar{x}$
- Generate a D-dimensional vector $v = v_1 ..., v_D$, where $v_d \in [0, 1]$ are independent and identically distributed according to a Bernoulli distribution with rate μ
- for index d $\in [1, ..., D]$

  if $v_d = 1$

  Sample $\bar{x_d}$ from the set F with equal probability

  else

  $\bar{x_d} = x_d$

  end

  end

Likelihood Ratio for OOD detection 2/2

Algorithm 2: OOD detection using Likelihood Ratio

- **Inputs:** D-dimensional test input $x = x_1 ... x_D$
- **Output:** Predict OOD

Likelihood Ratio for OOD detection 2/2

Algorithm 2: OOD detection using Likelihood Ratio

- **Inputs:** D-dimensional test input $x = x_1 ... x_D$
- **Output:** Predict OOD
- Fit a model $p_\theta(x)$ using in-distribution data-set D_{in}
- Fit a background model $p_{\theta_0}(x)$ using perturbed input data $\bar{D_{in}}$ (generated using Algorithm 1) and (optionally) model regularization techniques
- Compute the likelihood ratio statistic:

$$\boxed{LLR(x) = log(p_\theta(x_S)) - log(p_{\theta_0}(x_S))} \tag{4}$$

Likelihood Ratio for OOD detection 2/2

Algorithm 2: OOD detection using Likelihood Ratio

- **Inputs:** D-dimensional test input $x = x_1 ... x_D$
- **Output:** Predict OOD
- Fit a model $p_\theta(x)$ using in-distribution data-set D_{in}
- Fit a background model $p_{\theta_0}(x)$ using perturbed input data $\bar{D_{in}}$ (generated using Algorithm 1) and (optionally) model regularization techniques
- Compute the likelihood ratio statistic:

$$\boxed{LLR(x) = log(p_\theta(x_S)) - log(p_{\theta_0}(x_S))} \tag{4}$$

- **Predict OOD if LLR(x) is small**

OOD detection for Images 1/2

- **Fashion-MNIST (in-dist.) vs. MNIST (OOD):** PixelCNN++ model is trained on Fashion-MNIST
- **Likelihood** is dominated by the **background pixels** $\implies$ **p(Fashion-MNIST)** $<$ **p(MNIST)**
- **Likelihood ratio** focuses on the **semantic pixels** $\implies$ **LLR(Fashion-MNIST)** $>$ **LLR(MNIST)**

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Images 1/2

- **Fashion-MNIST (in-dist.) vs. MNIST (OOD):** PixelCNN++ model is trained on Fashion-MNIST
- **Likelihood** is dominated by the **background pixels** $\implies$ **p(Fashion-MNIST)** $<$ **p(MNIST)**
- **Likelihood ratio** focuses on the **semantic pixels** $\implies$ **LLR(Fashion-MNIST)** $>$ **LLR(MNIST)**

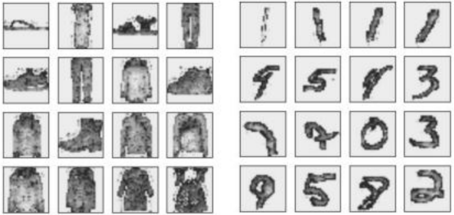

Figure 4: likelihood of pixels

OOD detection for Images 1/2

- **Fashion-MNIST (in-dist.) vs. MNIST (OOD):** PixelCNN++ model is trained on Fashion-MNIST
- **Likelihood** is dominated by the **background pixels** $\Longrightarrow$ **p(Fashion-MNIST)** $<$ **p(MNIST)**
- **Likelihood ratio** focuses on the **semantic pixels** $\Longrightarrow$ **LLR(Fashion-MNIST)** $>$ **LLR(MNIST)**

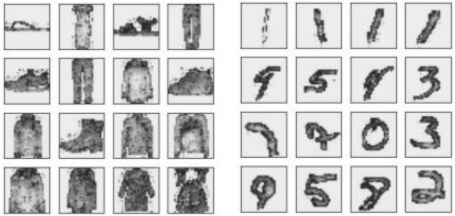

Figure 4: likelihood of pixels

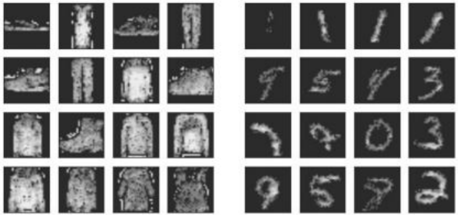

Figure 5: Likelihood ratio of pixels

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Images 2/2

Error Metric

- **AUROC↑:** Area under the ROC curve
- **AUPRC↑:** Area under the precision-recall curve
- **FPR80↓:** False positive rate at 80 percent true positive rate

OOD detection for Images 2/2

Error Metric

- **AUROC↑:** Area under the ROC curve
- **AUPRC↑:** Area under the precision-recall curve
- **FPR80↓:** False positive rate at 80 percent true positive rate

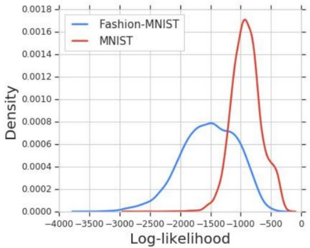

Figure 6: Log-likelihood is lower for Fashion-MNIST (in-dist) than MNIST (OOD)

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Error Metric

- **AUROC↑:** Area under the ROC curve
- **AUPRC↑:** Area under the precision-recall curve
- **FPR80↓:** False positive rate at 80 percent true positive rate

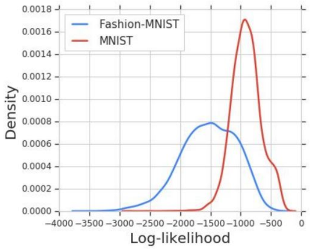

Figure 6: Log-likelihood is lower for Fashion-MNIST (in-dist) than MNIST (OOD)

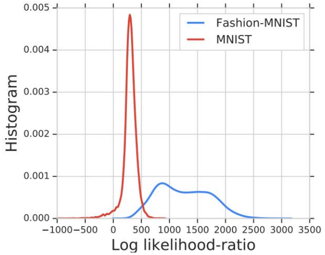

Figure 7: Log-likelihood ratio is higher for Fashion-MNIST (in-dist) than MNIST (OOD)

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Images 2/2

Error Metric

- **AUROC↑:** Area under the ROC curve
- **AUPRC↑:** Area under the precision-recall curve
- **FPR80↓:** False positive rate at 80 percent true positive rate

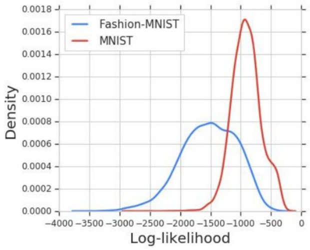

Figure 6: Log-likelihood is lower for Fashion-MNIST (in-dist) than MNIST (OOD)

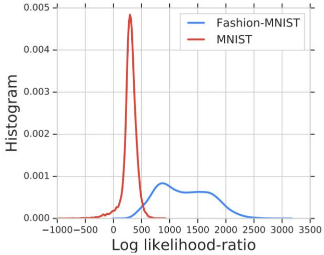

Figure 7: Log-likelihood ratio is higher for Fashion-MNIST (in-dist) than MNIST (OOD)

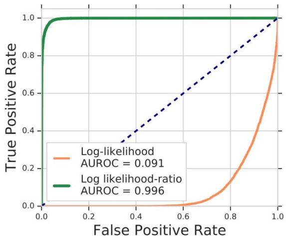

Figure 8: Likelihood ratio significantly improves the AUROC of OOD detection from 0.091 to 0.996

OOD detection for Genomic Sequences 1/2

- 10 in-distribution, 60 OOD validation, 60 OOD test
- Classes split by year to reflect challenges faced when classifier trained only on known classes

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Genomic Sequences 1/2

- 10 in-distribution, 60 OOD validation, 60 OOD test
- Classes split by year to reflect challenges faced when classifier trained only on known classes

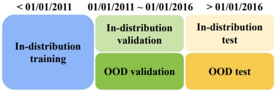

Figure 9: Genomic sequence data-set

OOD detection for Genomic Sequences 1/2

- 10 in-distribution, 60 OOD validation, 60 OOD test
- Classes split by year to reflect challenges faced when classifier trained only on known classes

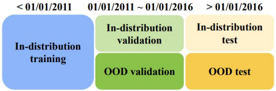

Figure 9: Genomic sequence data-set

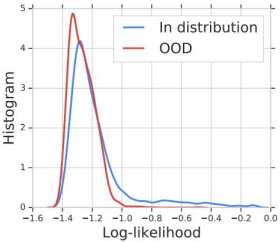

Figure 10: Log-likelihood hardly separates in-distribution and OOD input

OOD detection for Genomic Sequences 1/2

- 10 in-distribution, 60 OOD validation, 60 OOD test
- Classes split by year to reflect challenges faced when classifier trained only on known classes

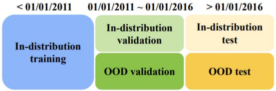

Figure 9: Genomic sequence data-set

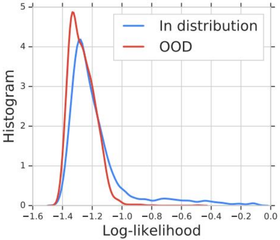

Figure 10: Log-likelihood hardly separates in-distribution and OOD input

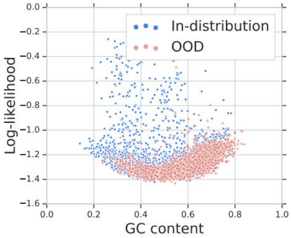

Figure 11: The log-likelihood is heavily affected by the GC-content of a sequence

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Genomic Sequences 1/2

- 10 in-distribution, 60 OOD validation, 60 OOD test
- Classes split by year to reflect challenges faced when classifier trained only on known classes

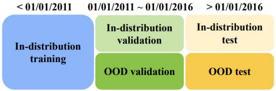

Figure 9: Genomic sequence data-set

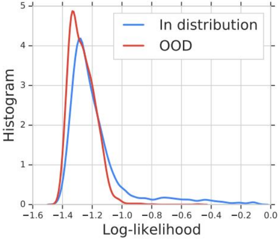

Figure 10: Log-likelihood hardly separates in-distribution and OOD input

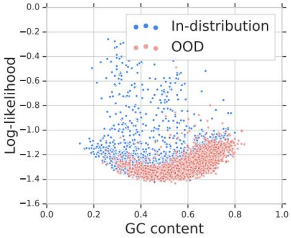

Figure 11: The log-likelihood is heavily affected by the GC-content of a sequence

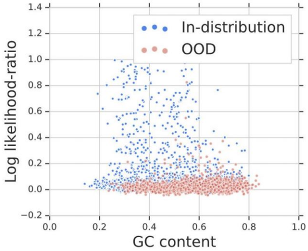

Figure 12: Corrected GC-content of a sequence

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Genomic Sequences 2/2

- 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

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Genomic Sequences 2/2

- 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

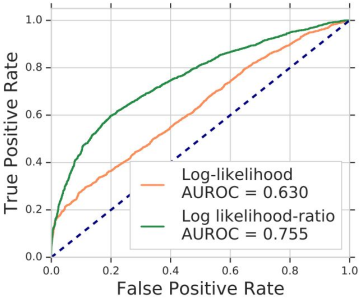

Figure 13: AUROC for likelihood and LLR

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

OOD detection for Genomic Sequences 2/2

- 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

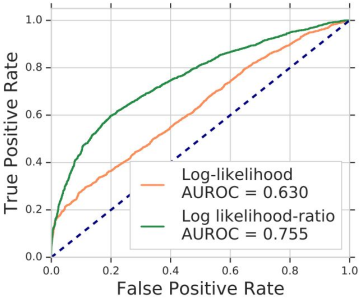

Figure 13: AUROC for likelihood and LLR

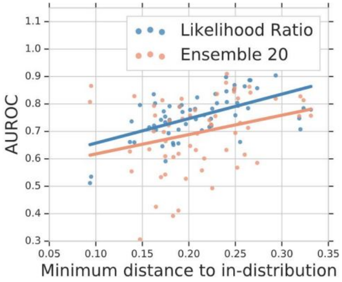

Figure 14: Correlation between the AUROC of OOD detection and distance to in-distribution classes

Comparison with baseline methods

	AUROC↑	AUPRC↑	FPR80↓
Likelihood	0.626 (0.001)	0.613 (0.001)	0.661 (0.002)
Likelihood Ratio (ours, μ)	0.732 (0.015)	0.685 (0.017)	0.534 (0.031)
Likelihood Ratio (ours, μ, λ)	**0.755 (0.005)**	**0.719 (0.006)**	**0.474 (0.011)**
$p(\hat{y}\|\boldsymbol{x})$	0.634 (0.003)	0.599 (0.003)	0.669 (0.007)
Entropy of $p(y\|\boldsymbol{x})$	0.634 (0.003)	0.599 (0.003)	0.617 (0.007)
Adjusted ODIN	0.697 (0.010)	0.671 (0.012)	0.550 (0.021)
Mahalanobis distance	0.525 (0.010)	0.503 (0.007)	0.747 (0.014)
Ensemble, 5 classifiers	0.682 (0.002)	0.647 (0.002)	0.589 (0.004)
Ensemble, 10 classifiers	0.690 (0.001)	0.655 (0.002)	0.574 (0.004)
Ensemble, 20 classifiers	0.695 (0.001)	0.659 (0.001)	0.570 (0.004)
Binary classifier	0.635 (0.016)	0.634 (0.015)	0.619 (0.025)
$p(\hat{y}\|\boldsymbol{x})$ with noise class	0.652 (0.004)	0.627 (0.005)	0.643 (0.008)
$p(\hat{y}\|\boldsymbol{x})$ with calibration	0.669 (0.005)	0.635 (0.004)	0.627 (0.006)
WAIC, 5 models	0.628 (0.001)	0.616 (0.001)	0.657 (0.002)

Figure 15: Error metric for genomic sequence using different methods

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

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
- Proposed method performs well on **images and achieves state of the art (SOTA) performance on genomic dataset**

ADOPT | Advanced Design, Optimization and Probabilistic Techniques laboratory

Review/Comments

- Author assumes that background and semantic component of input are independent, which may not be true in many practical application
- GC content of a sequence is similarly a function of the semantic component when classifying bacterial sequences
- The AUROC being significantly worse than random on the Fashion MNIST dataset isn't explained
- Given the experimental evidence and the novelty of the method, it is important contribution for OOD detection
- **Given the genomics sequence, this method can be used for finding out new strain of COVID'19**
- **Proposed method can be used for early detection of disease, which is significant contribution in respective area**

From the News

THE HINDU

MENU

HOME NEWS OPINION BUSINESS SPORT ENTERTAINMENT Free Games CROSSWORD+ SCIENCE

NEWS NATIONAL INTERNATIONAL STATES CITIES

NEWS › NATIONAL

NATIONAL

Coronavirus | 5% of COVID-19 cases to be tested for genome sequence

SPECIAL CORRESPONDENT

NEW DELHI, DECEMBER 26, 2020 23:00 IST
UPDATED: DECEMBER 26, 2020 23:00 IST

SHARE ARTICLE PRINT A | A | A

Figure 16: New variant of COVID'19 identification based on genomic sequencing

References

- Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, Balaji Lakshminarayanan. Likelihood Ratios for Out-of-Distribution Detection. *arXiv:1906.02845v2,2019*
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565, 2016.*
- Bishop, C. M. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7 (1):108–116, 1995b.
- Busia, A., Dahl, G. E., Fannjiang, C., Alexander, D. H., Dorfman, E., Poplin, R., McLean, C. Y., Chang, P.-C., and DePristo, M. A deep learning approach to pattern recognition for short DNA sequences. *bioRxiv*, pp. 353474, 2018.
- Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136, 2016*.
- Hendrycks, D., Mazeika, M., and Dietterich, T. G. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606, 2018*.
- Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325, 2017*.

ADOPT Advanced Design, Optimization and Probabilistic Techniques laboratory

Thank You