A Less Biased Evaluation of Out-of-distribution Sample Detectors

Alireza Shafaei, Mark Schmidt, James Little
University of British Columbia

The Problem
In a typical supervised learning scenario, we assume the samples are drawn from a fixed distribution. What can go wrong in practice?

OD-Test: A less biased evaluation strategy

- A binary classifier: in-distribution vs. out-of-distribution (OOD).
- We do not have access to OOD samples in practice.
- Supervised outlier detection: train a binary classifier on a fixed mixture of outlier and inlier datasets (two-dataset evaluation).
- Complex models can easily overfit to two-dataset classifications. Previous work uses a fixed mixture of two low-dimensional datasets. We show that it yields unreliably optimistic results (see top right).
- A more realistic setup with three datasets (OD-Test):
  1. Observe a clean $D_s$.
  2. Learn a binary reject function $r$ on the mixture of $D_s$ and $D_m$.
  3. Test the reject function on the mixture of $D_s$ and $D_t$.
Repeat over different outlier datasets to obtain a reliable estimate of performance on $D_s$.

Experimental Setup

- Methods:
  - Uncertainty: MC-Dropout [1], DeepEnsemble [2].
  - Density estimation: PixelCNN++ [3].
  - Open-set recognition: OpenMax [4].
  - Deep learning literature: ODIN [5], Probability Threshold.
  - Outlier/Anomaly detection: K-NN, Reconstruction-based.
  - Other: K-NN on Autoencoder and VAE latent representations, SVM on logits, K-way logistic regression loss, direct binary classification.

- Models:
  - VGG-16
  - Resnet-50

- Datasets:
  - MNIST
  - FashionMNIST
  - NotMNIST
  - CIFAR10
  - CIFAR100
  - STL10
  - Tinyimagenet
  - Uniform Noise
  - Gaussian Noise

A Short Summary of Results

- A two-dataset evaluation can make us too optimistic.
- Simpler/cheaper data mining approaches work as well as the recently proposed methods in low-dimensional settings.
- None of the methods work well on high-dimensional data.
- VGG-16 is better than Resnet-50 for this task, even though the Resnet model has a higher image classification accuracy.
- For a more reliable assessment, future work should use OD-test instead of two-dataset evaluations.

Selected References


Replicate the results on GitHub
https://github.com/ashafaei/OD-test