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

Alireza Shafaei, Mark Schmidt, James Little

University of British Columbia

BMVC 2019

The Problem

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

DenseNet 161 (2017)	Balance Beam 52%	Chainlink Fence 31%	Chest 37%	Tench 36%
SqueezeNet (2016)	Balance Beam 18%	Poncho 32%	Jean 30%	Suit 21%
ResNet 152 (2015)	Pacifier 33%	Chain Mail 29%	Dust Cover 52%	Sweatshirt 25%
VGG 19 (2014)	Dust Cover 44%	Window Screen 5%	Chest 11%	Sweatshirt 46%
AlexNet (2012)	Dust Cover 22%	Cardigan 12%	Theater Curtain 3%	Coho 37%

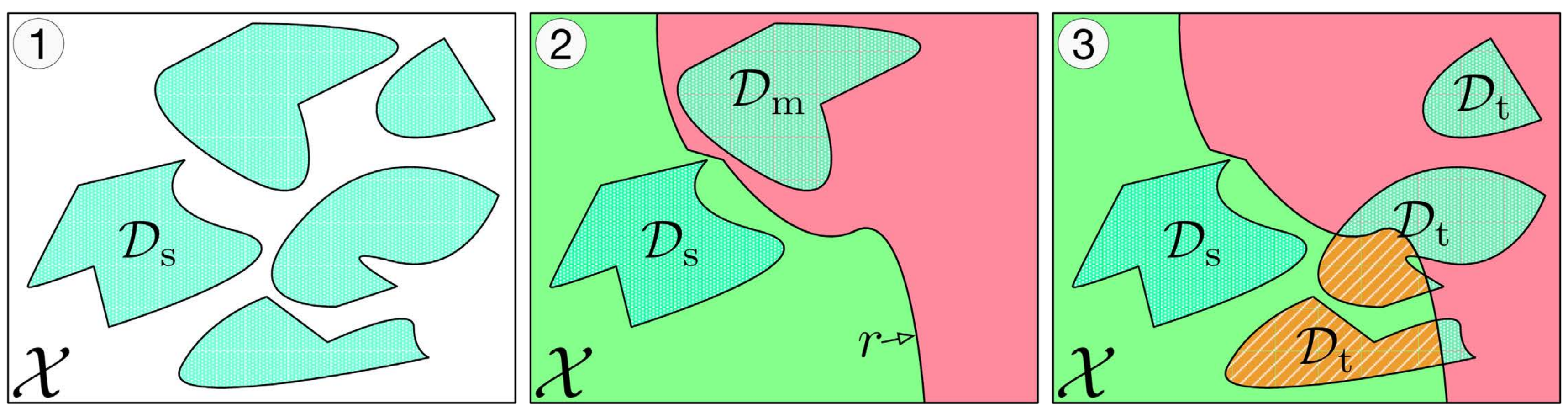
OOD Detectors detect the examples where the model cannot give reliable predictions.

- We show that current evaluation strategies over-estimate accuracy.
- We present **a more practical evaluation framework**.
- We show that the state-of-art methods are not reliable in practical scenarios.

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**): Given an inlier dataset D_s and outlier datasets D_m, and D_t.
  1. Observe a clean D_s.
  2. Learn a binary reject function $\mathbf{r}$ 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.

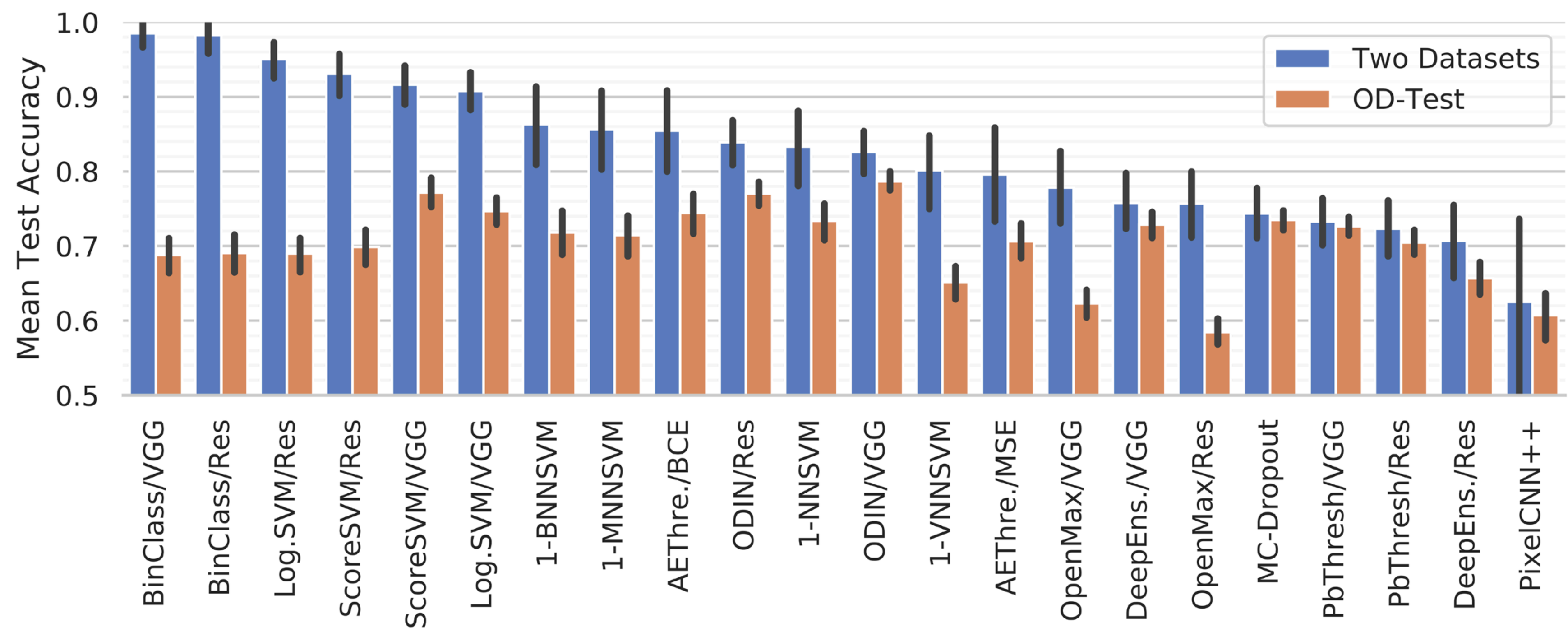
Models.

- VGG-16
- Resnet-50

Datasets.

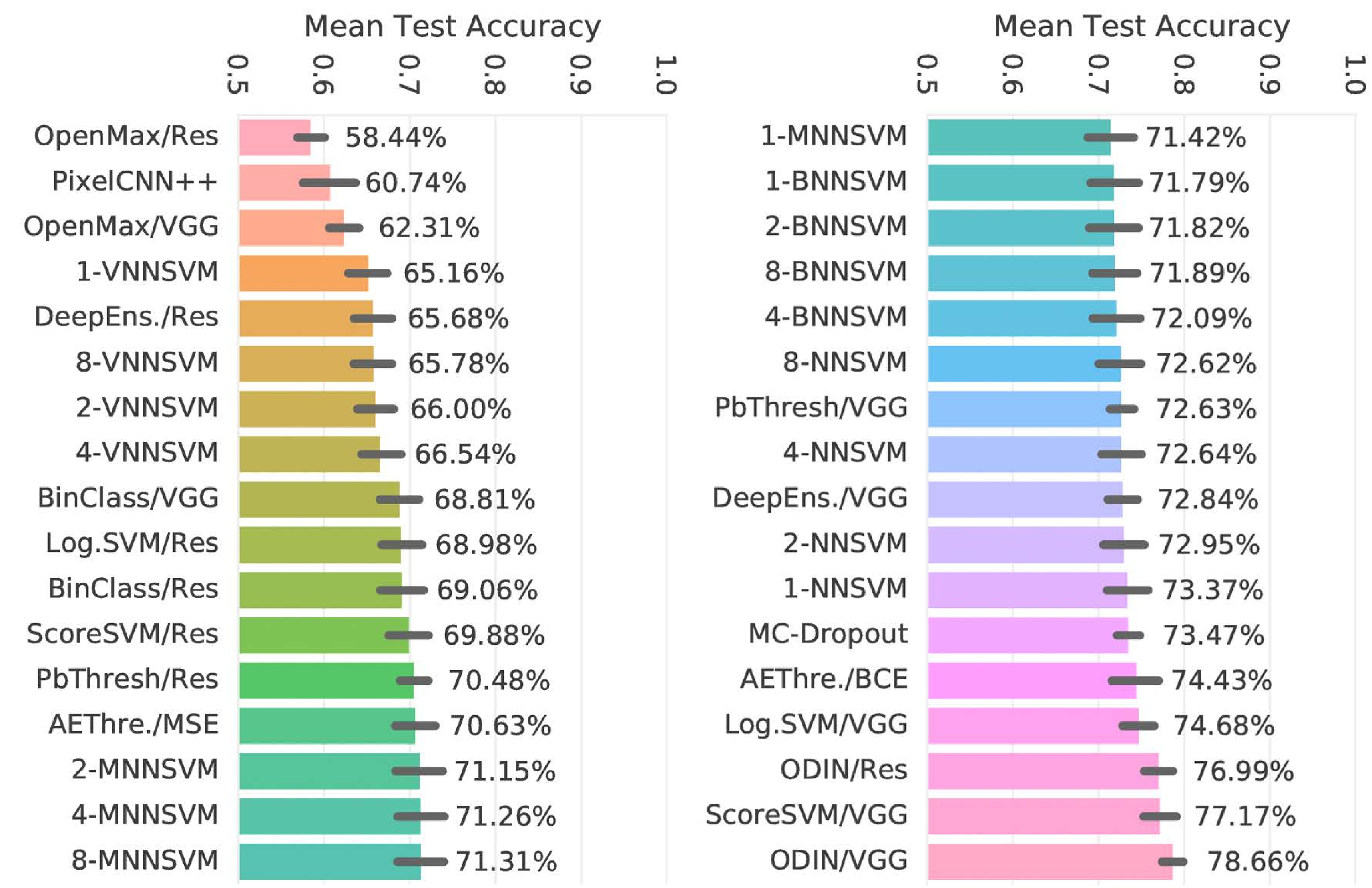
- MNIST
- FashionMNIST
- NotMNIST
- CIFAR10
- CIFAR100
- STL10
- TinyImagenet
- Uniform Noise
- Gaussian Noise

Two-dataset evaluation vs. OD-test (n = 46/bar, 308/bar)

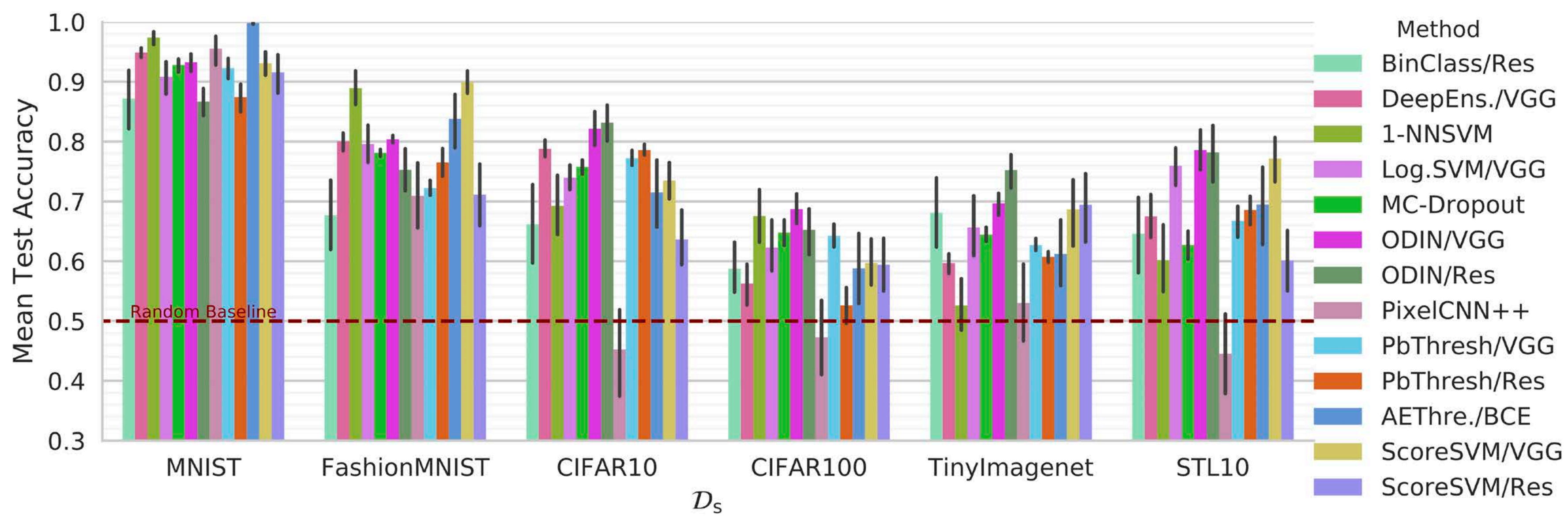


A two-dataset evaluation scheme can be too optimistic in identifying the best available method.

Mean test accuracy, averaging over D_s, D_m, D_t (n = 308/bar)



Mean test accuracy per source dataset D_s (n = 54/bar)



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

[1] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *ICML*, 2016.
[2] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *NIPS*, 2017.
[3] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *ICLR*, 2017.
[4] A. Bendale and T. E. Boult, "Towards Open Set Deep Networks," in *CVPR*, 2016.
[5] S. Liang, Y. Li, and R. Srikant, "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks," *ICLR*, 2018.

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