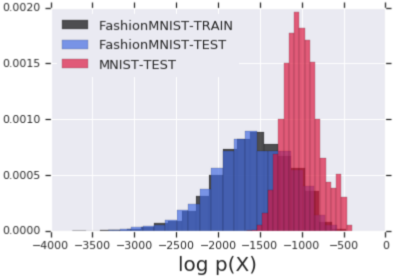
Further Analysis of Outlier Detection with DGMs

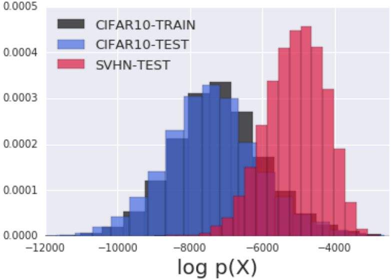
Ziyu Wang[1], Bin Dai[2], David Wipf[3], Jun Zhu[1]

[1]Tsinghua University, [2]Samsung Research China, [3]AWS AI Lab

Background



(a) Train on Fashion, test on MNIST



(b) Train on CIFAR-10, test on SVHN

"Do Deep Generative Models Know What They Don't Know?"

Figure taken from Nalisnick et al (2019). See also Hendrycks et al (2019).

The Typicality Argument

A "longitudinal view" of data: high-d rv $\Leftrightarrow$ random sequence

- $\mathcal{N}(0, I_d) \Leftrightarrow$ a sequence of d scalar rvs

Certain random sequences fall into a **typical set** with high probability, which does not necessarily coincide with region of high density

Ex. an IID random sequence of length d will have ℓ_2 norm of $O(\sqrt{d})$ with high probability

- "Gaussian distributions are like soap bubbles"
- Test for outlier using $\|x\|$

The Typicality Argument

So far, the typicality argument has not been successfully applied to explain the peculiarity of single-sample outlier detection[1]

Check $\log p_{inlier}(x_{test})$*?*

- $\log p$ doesn't always concentrate, unlike the IID case

Transform $x \sim p_{inlier}$ *to an **IID** sequence (e.g. latents of flows) and test in that space?*

- Doesn't work in practice, estimating that transformation is probably too hard

[1] See paper for discussion about previous work, alternative explanation, etc

An Outlier Test Generalizing the Idea of Typicality

Proposal: transform x into a sequence with a **weaker** property than IID, and test for that property

IID $\subset$ **M**artingale **D**ifference $\subset$ (weak) **W**hite **N**oise

$\tilde{R}_i(x) := x_i - \mathsf{E}_p(x_i|x_{<i}) \approx x_i - \mathsf{E}_\theta(x_i|x_{<i})$ is MD for $x \sim p_{inlier}$

- Still using autoregressive GMs
- But estimating $\mathsf{E}(x_i|x_{<i})$ is easier than estimating $p(x_i|x_{<i})$

Test for outlier by applying WN tests to R

Results

Table 1: AUROC and average ranks. $\boxed{\text{Worse than random}}$.

Inlier Dist.		CIFAR-10		CelebA		TinyImageNet		Avg.
Outlier Dist.		CelebA	SVHN	CIFAR-10	SVHN	CIFAR-10	SVHN	Rank↓
	LH	0.88	$\boxed{0.16}$	0.82	$\boxed{0.15}$	$\boxed{0.28}$	$\boxed{0.05}$	3.67
AR-	LH-2S	0.77	0.69	0.84	0.78	0.55	0.93	2.50
DGM	LR	0.86	0.86	0.99	1.00	$\boxed{0.39}$	0.56	2.00
	Ours	0.97	0.83	0.85	0.93	0.85	0.62	**1.67**

- Our test works well under the previous setup, supporting a (generalized) typicality argument
- DGMs probably know what they don't know?

Results

Inlier Dist.		CIFAR-10		CelebA		TinyImageNet		Avg.
Outlier Dist.		CelebA	SVHN	CIFAR-10	SVHN	CIFAR-10	SVHN	Rank↓
Linear	LH	0.77	$\boxed{0.02}$	0.72	$\boxed{0.03}$	$\boxed{0.11}$	$\boxed{0.00}$	2.50
	LH-2S	0.69	0.76	0.70	0.80	0.64	0.81	2.17
	Ours	0.67	0.95	0.90	0.99	0.92	0.99	**1.33**

- A **linear generative model** also seems to know … about semantics?

Further Analysis of Generative Outlier Detection

- New benchmarks to disentangle the influence of low-level textual information vs image semantics:
    - CIFAR-10 vs subset-of-CIFAR-100, and BigGAN-synthesized images
- On the intrinsic difficulty of high-dimensional density estimation in OOD regions
    - SoTA DGMs generate visually plausible images, yet may deviate significantly from a <u>known</u> ground truth in density estimation
    - Model's inductive bias has more influence on density estimation in OOD regions $\Rightarrow$ likelihood-based tests should be used with care

See paper for details