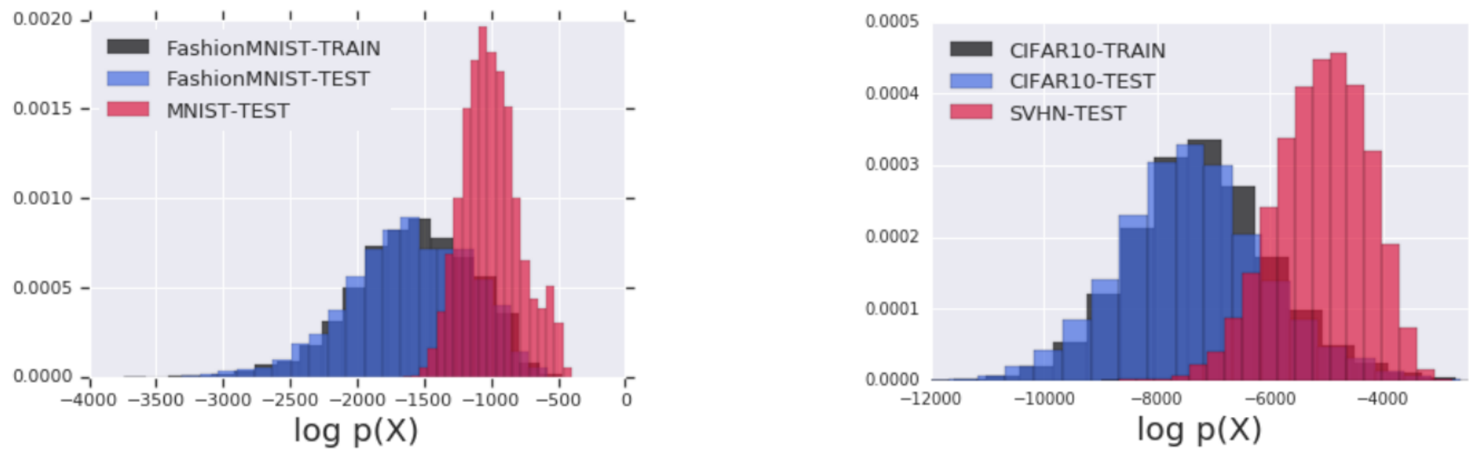
Further Analysis of Outlier Detection with Deep Generative Models

Ziyu Wang (Tsinghua), Bin Dai (Samsung Research China), David Wipf (Amazon Research), Jun Zhu (Tsinghua)

arXiv code

- The recent discovery that DGMs may assign higher likelihood to outliers with clearly different semantics led to the question of whether they are calibrated.
- We propose a new outlier test using DGMs, by generalizing the idea of typicality.
  The empirical success of our test, along with an experiment showing the difficulty of pdf estimation in OOD regions, suggest that previous observations do not necessarily imply the corresponding DGMs are uncalibrated.
- Additional experiments suggest that we need new benchmarks which disentangle the effect of texture vs semantics.



(a) Train on FashionMNIST, Test on MNIST

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

(Taken from Nalisnick et al (2019). See also Hendrycks et al (2019).)

Background: Typicality

- "Longitudinal view of data": p(high-dim rv) ⇒ p(random sequence)
- Certain random sequences fall into a **typical set** with high prob, which does not necessarily have high density
  E.g. x iid sequence ⇒ $\|x\|_2^2/d \to^p \mathrm{E}\, x_1^2$

Understanding the previous findings through typicality?

- Similar concentration guarantees don't exist for general p
- Transforming p_{inlier} back to a simple distribution (e.g. iid, using flow), and use the iid test?
  Estimating the transform may require far more samples than to generate plausible samples

From Typicality to an Outlier Test

- **Idea**: transform p_{inlier} to random sequence with a **weaker** property than IID, and test for that property
- IID ⊃ MD ⊃ (weak) WN[1]
- Generate MD seq, approximated with an autoregressive model

$$\tilde{R}_t(x) := x_t - \mathbf{E}_{\mathrm{inlier}}(x_t|x_{<t}) \approx x_t - \mathbf{E}_\theta(x_t|x_{<t})$$

- Test for WN: Box-Pierce

$Q_{\mathrm{BP}} = d\sum_{l=1}^{L}\hat{\rho}_l^2/L$, where ρ_l is the l-lag autocorrelation of R

Intuitively

- Under H_0,[2] $L\, Q_{BP}$ approximately $\sim\chi_L^2$, or $Q_{BP} \sim 1$
- For outlier, the residual sequence R contains unexplained semantic information, which likely has autocorrelation. Non-zero ACF leads to $Q_{BP} \sim d/L \gg 1$.

Table 1: AUROC values for the single-sample test, and average ranks within each group. **Boldface** indicates best results; underline indicates notable failures (AUC < 0.5).

Inlier Dist.		CIFAR-10		CelebA		TinyImageNet		Avg.
Outlier Dist.		CelebA	SVHN	CIFAR-10	SVHN	CIFAR-10	SVHN	Rank
AR-DGM	LH	0.88	0.16	0.82	0.15	0.28	0.05	3.67
	LH-2S	0.77	0.69	0.84	0.78	0.55	0.93	2.50
	LR	0.86	0.86	0.99	1.00	0.39	0.56	2.00
	WN	0.97	0.83	0.85	0.93	0.85	0.62	**1.67**
VAE+Linear $n_z = 64$	LH	0.64	0.09	0.88	0.26	0.28	0.04	3.33
	LH-2S	0.47	0.81	0.85	0.69	0.51	0.87	3.00
	LR	0.39	0.90	0.98	0.99	0.64	0.91	1.83
	WN	0.64	0.67	0.93	0.99	0.92	0.99	**1.50**
VAE+Linear $n_z = 512$	LH	0.76	0.04	0.81	0.09	0.19	0.01	3.33
	LH-2S	0.61	0.85	0.76	0.81	0.59	0.90	2.67
	LR	0.56	0.86	0.97	0.99	0.55	0.90	2.50
	WN	0.61	0.88	0.88	1.00	0.94	0.99	**1.33**
Linear	LH	0.77	0.02	0.72	0.03	0.11	0.00	2.50
	LH-2S	0.69	0.76	0.70	0.80	0.64	0.81	2.17
	WN	0.67	0.95	0.90	0.99	0.92	0.99	**1.33**

Evaluating the WN Test

Baselines: single-side likelihood (**LH**); two-side likelihood (**LH-2S**, "weakly typical set"); likelihood latio (**LR**; Serra et al, 2020) [3]

Models: **a**uto**r**egressive **DGM** (PixelCNN++, PixelSNAIL); MVN = **linear** AR model; MVN applied to **VAE** residuals

Under previous settings: across all choices of model,

- The LH test fails, including when using the linear/MVN model
- The proposed test avoids such pathologies

⇒ Probably failure to assign lower likelihood to such outliers ↛ model miscalibration (more on this later)

Linear WN test also detects most outliers?

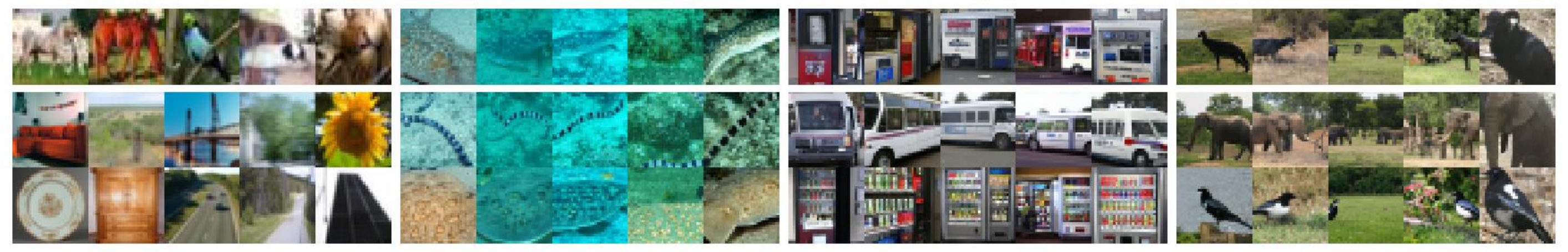
- The linear WN test could be useful. It's probably more useful than MVN likelihood
- New benchmarks are needed if we really care about test/model's ability to distinguish semantics

[^1] assuming sequence has zero mean and unit variance

[^2] technically, the more restricted hypothesis that R be IID. Note this is different from requiring x being IID

[^3] comparison to more baselines in appendix

[^4] technically, its lower bounds



(a) CIFAR (b) Synthetic-1 (c) Synthetic-2 (d) Synthetic-3

Figure 1: Overview of inlier (top) and outlier (bottom) distributions used in Section 3.2.

New Benchmarks

CIFAR: clearly different semantics

- inlier CIFAR-10, outlier subset of CIFAR-100

Synthetic: texture difference minimized, semantics still different

- inlier $\mathrm{GAN}(z, c_1)$ or $\mathrm{GAN}(z, c_2)$, outlier $\mathrm{GAN}(z, (c_1 + c_2)/2)$

Comparison less clear, developing universally effective OOD tests might be difficult

Table 2: Results for the semantics-oriented experiments. Boldface indicates the best result.

	CIFAR, AUROC↑				Synthetic, Avg. Rank↓			
	LH	LH-2S	LR	WN	LH	LH-2S	LR	WN
AR-DGM	0.49	0.57	**0.61**	0.58	**2**	3.5	2.5	**2**
Linear	0.56	0.59	-	**0.60**	2.33	**1.67**	-	2
VAE+Linear, 64	0.51	0.55	0.64	**0.84**	**1.67**	3.33	2.67	2.33
VAE+Linear, 512	0.59	0.58	0.73	**0.80**	**2**	3.67	**2**	2.33

More on OoD Density Estimation

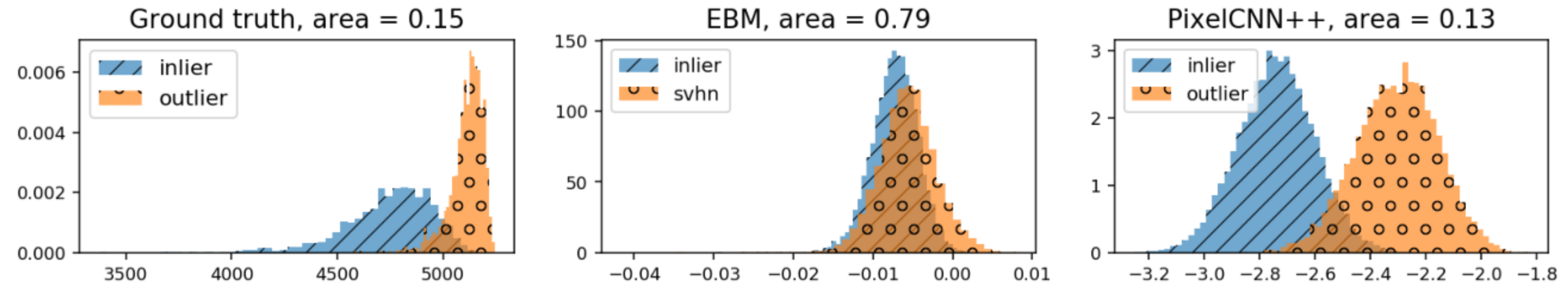
"Still it is possible a 'truly calibrated model' should assign lower likelihood to these 'natural' outliers?"

Density estimation in in-distribution locations is (relatively) easy, but in OOD regions, the inductive bias / "prior" of the GM may still overwhelm the evidence

Experiment:

- train SoTA EBM and AR-DGM on VAE-synthesized images
- The VAE is trained on CIFAR-10, so its output resembles natural image, but different from CIFAR we know the ground truth pdf[4]
- Generate outliers with high ground-truth density by masking VAE latents. Compare the learnt models' likelihood with ground truth

Result: learned pdf (below) indeed different in OoD regions


Ground truth, area = 0.15
inlier
outlier
EBM, area = 0.79
inlier
svhn
PixelCNN++, area = 0.13
inlier
outlier