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

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, Balaji Lakshminarayanan

DeepMind

1. Introduction

- Discriminative models are susceptible to overconfidence on **out-of-distribution (OOD) inputs**. Generative models are widely believed to be more robust to such inputs as they also model *p(x)* [Bishop, 1994].
- We challenge this assumption, showing that **deep generative models can assign higher density estimates to an OOD dataset than to the training data!**
- This phenomenon has implications not just for anomaly detection but also for detecting covariate shift, open-set classification, active learning, semi-supervised learning, etc

2. Motivating Observation: CIFAR-10 vs SVHN

CIFAR-10

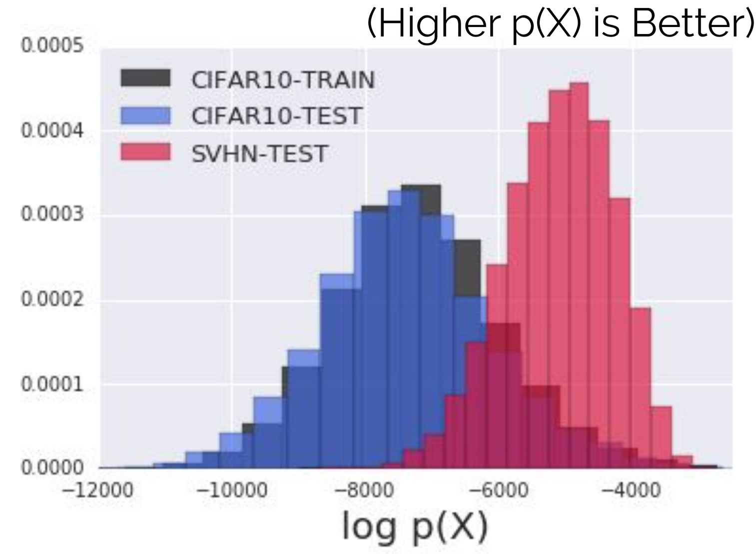
SVHN

We trained **Glow** [Kingma & Dhariwal, 2018] on CIFAR-10 and evaluated on the model on SVHN. We find that **Glow assigns a higher likelihood to SVHN than to CIFAR-10** (both train/test splits).

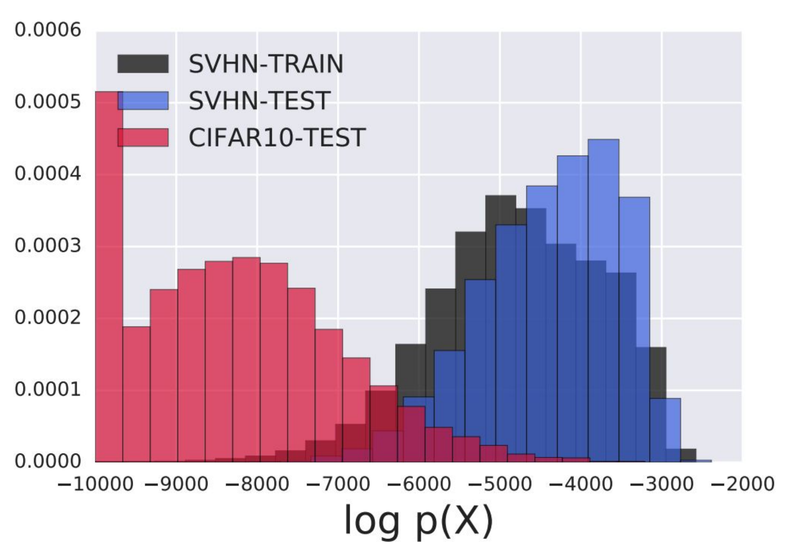
Data Set	Avg. Bits Per Dimension
Glow Trained on CIFAR-10	
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	**2.389**
Glow Trained on SVHN	
SVHN-Test	2.057

(Lower BPD is Better)

(Higher p(X) is Better)



Histogram of Glow Log-Likelihoods



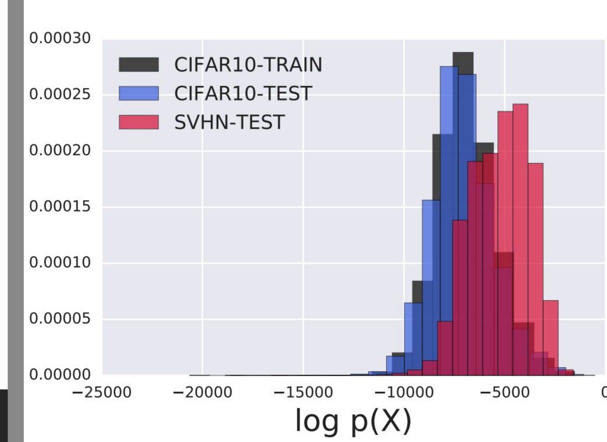
Glow Log-Likelihoods: SVHN train, CIFAR-10 test

The phenomenon is asymmetric w.r.t. datasets: Training on SVHN and evaluating on CIFAR-10 results in the expected ordering (SVHN is assigned higher likelihood).

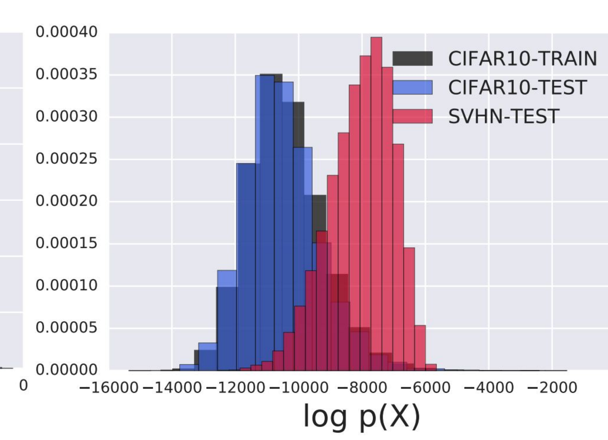
3. Testing Other Deep Generative Model Classes

This **phenomenon is also observed in two other classes of deep generative models**: auto-regressive (PixelCNN) and latent variable models (Variational Auto-Encoders).

CIFAR-10 Train vs SVHN Test

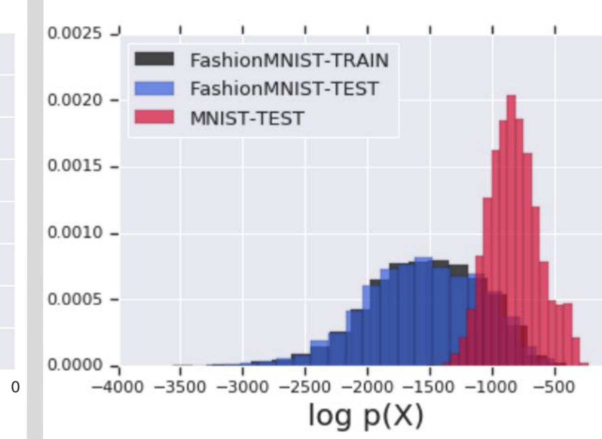


PixelCNN Log-Likelihoods

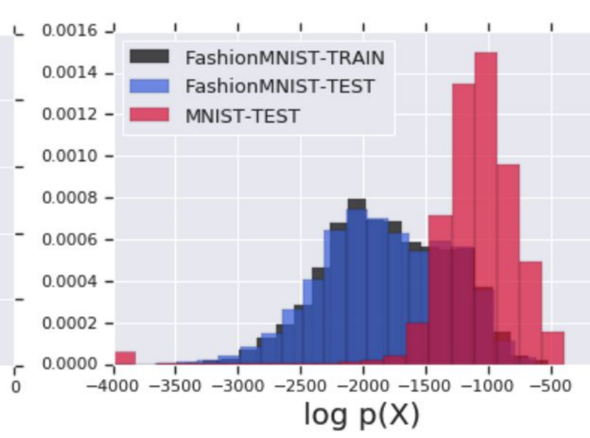


VAE Log-Likelihoods

FashionMNIST Train vs MNIST Test



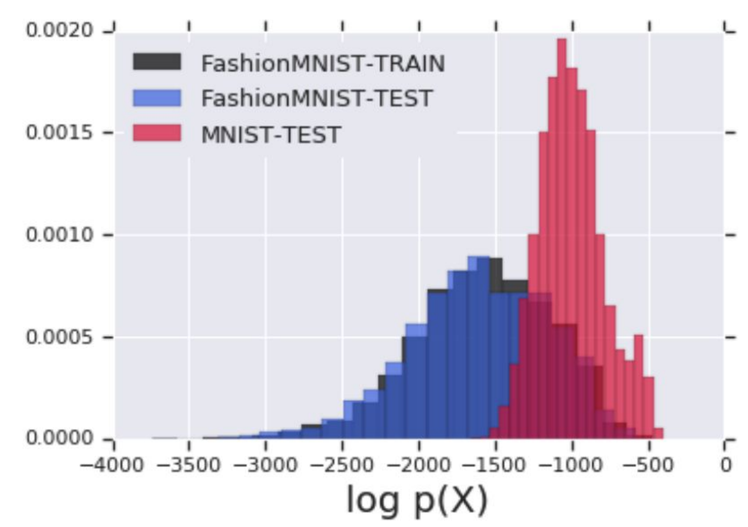
PixelCNN Log-Likelihoods



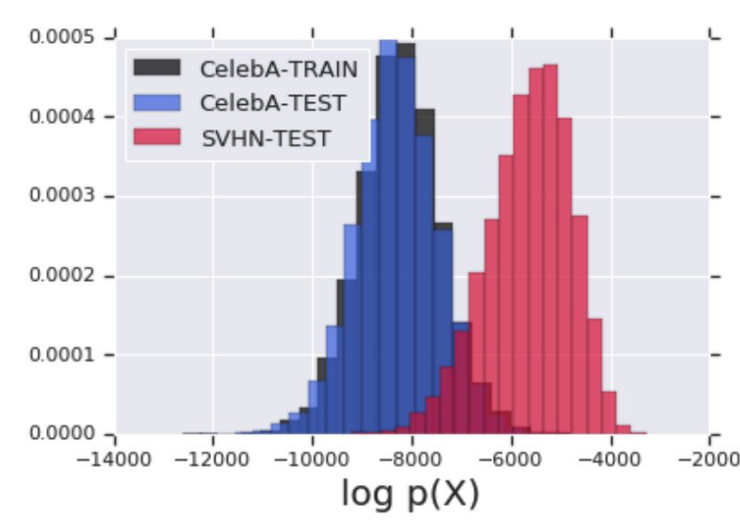
VAE Log-Likelihoods

4. Testing Glow on Other Data Sets

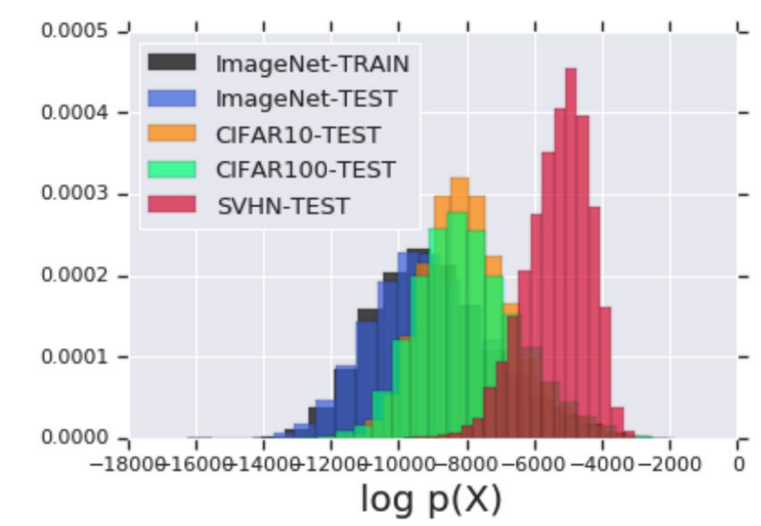
We find further evidence of the phenomenon in five other data set pairs:



FashionMNIST vs MNIST



CelebA vs SVHN

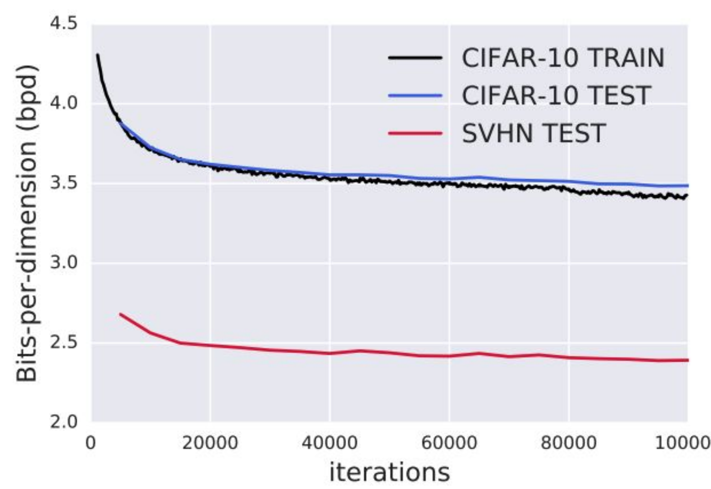


ImageNet vs SVHN / CIFAR

- We also observe that **constant inputs have the highest log-likelihood of any (tested) input**.
- Furthermore, we find that **SVHN has higher likelihood over the entire duration of training**.
- Ensembling generative models does not help.

Data Set	Avg. Bits Per Dimension
Glow Trained on FashionMNIST	
Random	8.686
Constant (0)	**0.339**

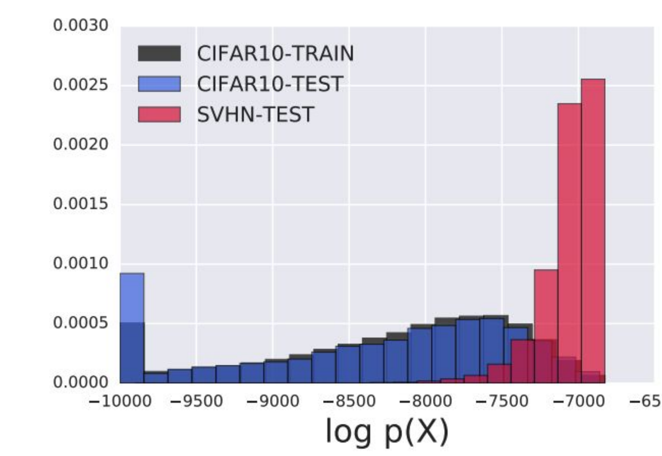
Data Set	Avg. Bits Per Dimension
Glow Trained on CIFAR-10	
Random	15.773
Constant (128)	**0.589**



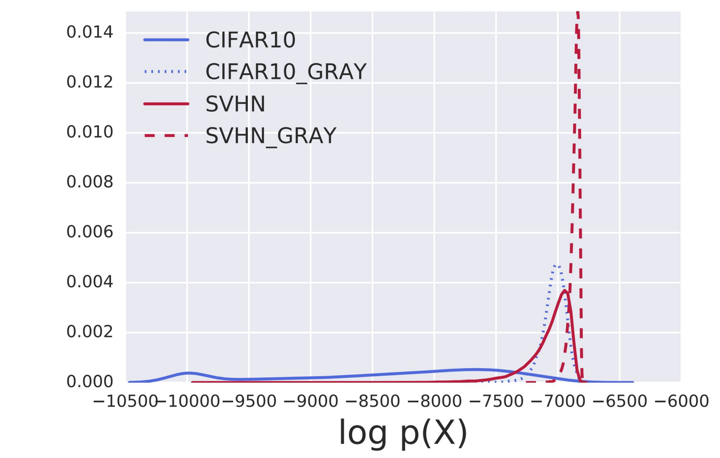
BPD vs Training Iteration

5. Digging Deeper Into Glow

To make theoretical analysis more tractable, **we restrict Glow to have constant volume (CV) transformations (w.r.t. input)**. We see similar CIFAR-vs-SVHN results for this model.



CV-Glow: CIFAR-10 vs SVHN



Graying Images Increases Likelihoods

For CV-Glow, we can approximate the difference in likelihoods between the training and OOD data as follows:

$$0 < \mathbb{E}_{q}[\log p(\boldsymbol{x};\boldsymbol{\theta})] - \mathbb{E}_{p^*}[\log p(\boldsymbol{x};\boldsymbol{\theta})]$$

Non-Training Distribution

Training Distribution

$$\approx \frac{\partial^2}{\partial z^2}\log p(z;\psi) \sum_{c=1}^{C}\left(\prod_{k=1}^{K}\sum_{j=1}^{C} u_{k,c,j}\right)^2 \sum_{h,w}(\sigma^2_{q,h,w,c} - \sigma^2_{p^*,h,w,c})$$

< 0 for all log-concave densities (e.g. Gaussian)

Non-negative due to square

Second Moment of Non-Training Distribution

Second Moment of Training Distribution

This expression helps explain several observations:

1. **Asymmetry**: difference between 2nd moments does not commute.
2. **Constant / grayscale inputs**: equivalent to non-training moment being zero. **Graying images increases likelihoods.**
3. **Early stopping / ensembling would not help**: expression holds true for all values of CV-Glow's parameters.

6. Summary

Density estimates from (current) deep generative models are **not always able to detect out-of-distribution inputs**.

Paper: https://arxiv.org/abs/1810.09136