We observe the phenomenon for autoregressive (PixelCNN) and latent variable (VAE) models as well.

The log-likelihood of flow-based models can be decomposed into prior and volume terms:
\[
\log p_{\theta}(x) = \log p_{\theta}(f_{\phi}(x); \psi) + \log \left| \frac{\partial f_{\phi}}{\partial x} \right|
\]

Discriminative models are susceptible to overconfident mistakes on out-of-distribution (OOD) inputs. Generative models are widely believed to be more robust as they model not only \( p(y|x) \) but also \( p(x) \). If an incoming sample has a low probability under \( p(x) \), then it can be flagged as OOD and an “I don’t know” message can be returned to the user (Bishop, 1994).

We challenge this assumption, showing that deep generative models can at times 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 open-set classification, active learning, semi-supervised learning, and others.

We trained Glow [Kingma & Dhariwal, 2018], a flow-based generative model, on CIFAR-10 and evaluated on the CIFAR-10 test split and on SVHN, a set of house numbers.

Constant (128) images are assigned a higher log-likelihood than to CIFAR-10 but also SVHN.

We find that CV-Glow assigns a higher likelihood to SVHN than to CIFAR-10. We see that SVHN’s higher likelihood is persistent across SVHN images and is not due to a few outliers.

CV-Glow’s log-likelihoods when trained on CIFAR-10 and tested on SVHN is shown to the right. We see that SVHN’s higher likelihood persists. Bits-per-dimension (BPD) over the course of training is also shown to the right. There is no cross-over during optimization.

CV-Glow in particular, the volume term drops out (since it does not depend on \( x \)), yielding:
\[
\frac{\partial^2}{\partial z^2} \log p_{\theta}(z; \psi) = \sum_{c=1}^{C} \left( \sum_{k=1}^{K} \sum_{x=1}^{w_{h,c}} \left( \sigma^2_{h,x,w,c} + \sigma^2_{w,h,x,c} \right) \right)
\]

where \( u \) denotes the parameters of the 1x1 conv. kernel, \( \sigma^2 \) is the variance of each distribution, \( C \) is the number of channels, and \( K \) is the number of flow steps. Plugging in the empirical variances (see right, top) results in the quantity being greater than zero, meaning SVHN will always have a higher likelihood under CV-Glow. This agrees with our plot of BDP during training shown above. Furthermore, this means that we can increase the likelihoods by lowering the input variance, which is equivalent to graying natural images. See figure to right.

Summary
- Density estimates from (current) deep generative models are not always able to detect out-of-distribution inputs.
- We should temper the enthusiasm with which we preach the benefits of generative models until their sensitivity to out-of-distribution inputs is better understood.

Contact: e.nalisnick@eng.cam.ac.uk, lamatsukawa, ywteh, dilang, balajiln@gmail.com