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.

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).

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).

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.

Density estimates from (current) deep generative models are not always able to detect out-of-distribution inputs. We find further evidence of the phenomenon in five other data set pairs:

- FashionMNIST vs MNIST
- CelebA vs SVHN
- ImageNet vs SVHN / CIFAR

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 likelihood.
3. Early stopping / ensembling would not help: expression holds true for all values of CV-Glow’s parameters.

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.

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

This phenomenon is also observed in two other classes of deep generative models: auto-regressive (PixelCNN) and latent variable models (Variational Auto-Encoders). We find that Glow assigns a higher likelihood to SVHN than to CIFAR-10 (both train/test splits).

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.

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.

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.