Hierchical VAEs Know What They Don't Know

Jakob D. Havtorn[1,2], Søren Hauberg[1], Jes Frellsen[1], Lars Maaløe[1,2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark

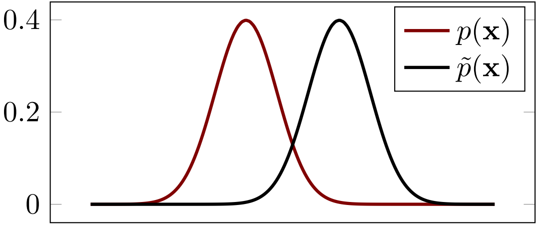
[2]Corti AI, Copenhagen

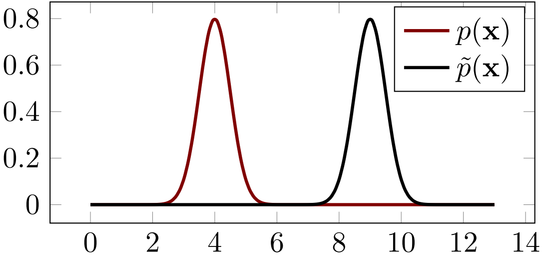
Defining OOD detection

Out-of-distribution (OOD) detection is about enabling models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$.

We are concerned with doing this on a per-datapoint basis, i.e. answering the question:

"Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"

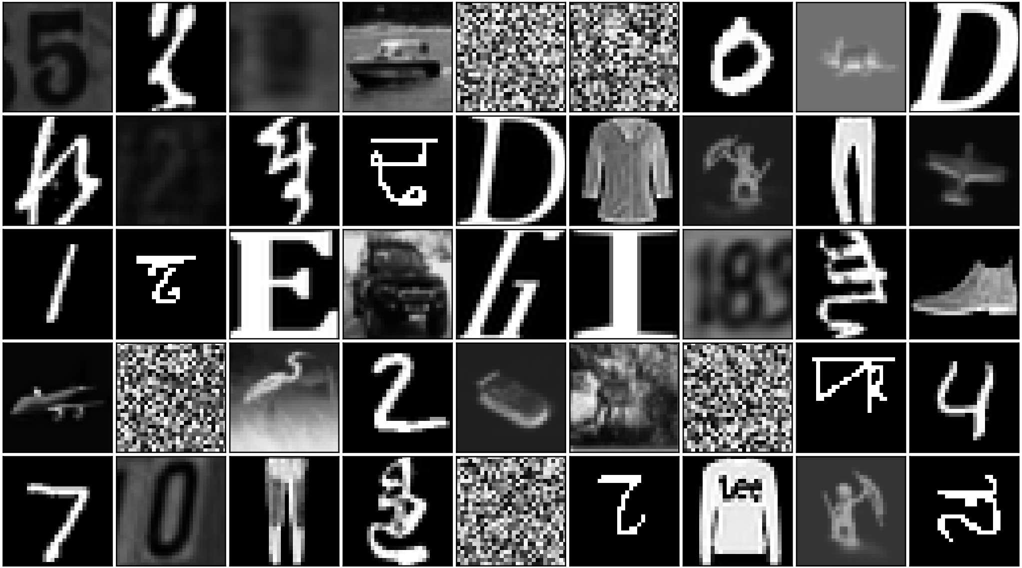




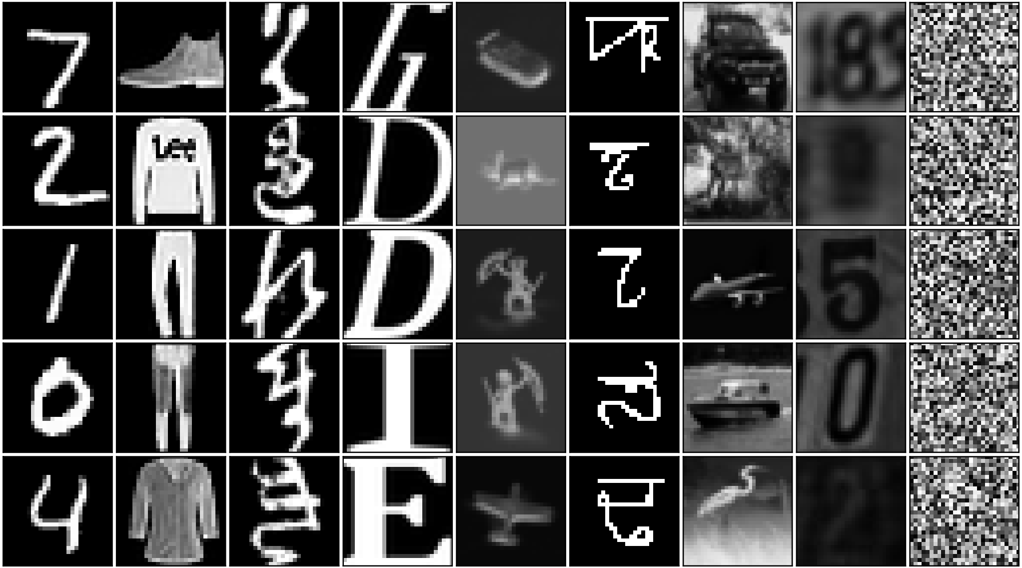
Problem and Contributions

- Deep generative models often fail at OOD detection task when using their likelihood estimate as the score function [6] by, perhaps surprisingly, assigning **higher likelihoods** to the OOD data.
- Contributions:
  - We present a fast and fully unsupervised method for OOD detection competitive with the state-of-the-art
  - We provide evidence that out-of-distribution detection fails due to learned low-level features that generalize across datasets.

In distribution?



Out of distribution?



Hierarchical VAE

We choose the hierarchical VAE as our model [2, 3].

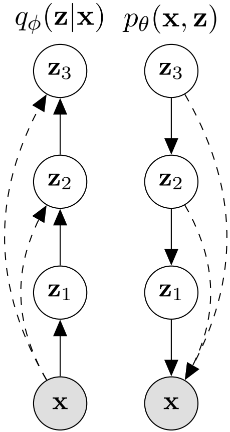
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z})\mathrm{d}\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})\mathrm{d}\mathbf{z}$$

Specifically we use

1. a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.

$$\text{Generative model:} \quad p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1)p_\theta(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_3),$$
$$\text{Inference model:} \quad q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1)q_\phi(\mathbf{z}_3|\mathbf{z}_2).$$

2. a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [5].



What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction likelihood}} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{regularization penalty}} \ . \quad (1)$$

The first term is high if the data is well-explained by $\mathbf{z}$.

The second term we can rewrite as,

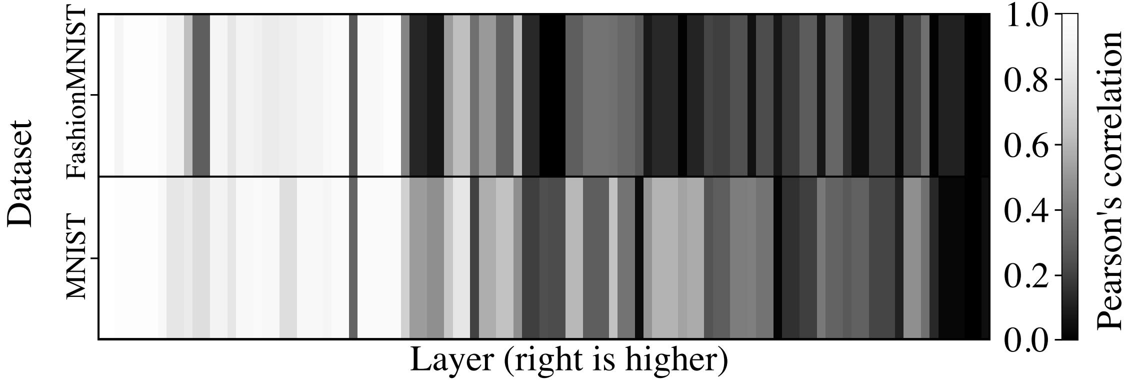
$$D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] \ . \quad (2)$$

The absolute log-ratios grow with $\dim(\mathbf{z}_i)$ since the log probability terms are computed by summing over the dimensionality of $\mathbf{z}_i$.

What do the lowest latent variables code for?

Absolute Pearson correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST.

Correlation computed between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).



An alternative likelihood bound, $\mathcal{L}^{>k}$

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [5]

$$\mathcal{L}^{>k}(\mathbf{x};\theta,\phi) = \mathbb{E}_{p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \tag{3}$$

Here, we have replaced the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [1].

$$LLR^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) \ . \tag{4}$$

We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right), \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leq}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \ . \end{aligned} \tag{5}$$

In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} LLR^{>k}(\mathbf{x}) = &-D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leq}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \ . \end{aligned} \tag{6}$$

Importance sampling the ELBO

The well-known importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples, $S \to \infty$ [4],

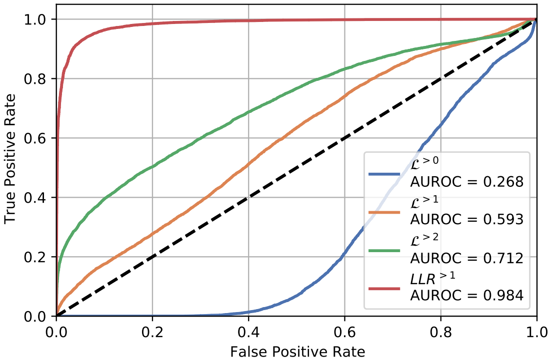
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N} \sum_{s=1}^{S} \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] \leq \log p_\theta(\mathbf{x}) \ , \tag{7}$$

Consequently, by importance sampling the ELBO, the associated KL-divergence associated vanishes and our likelihood ratio reduces to the KL-divergence associated with $\mathcal{L}^{>k}$.

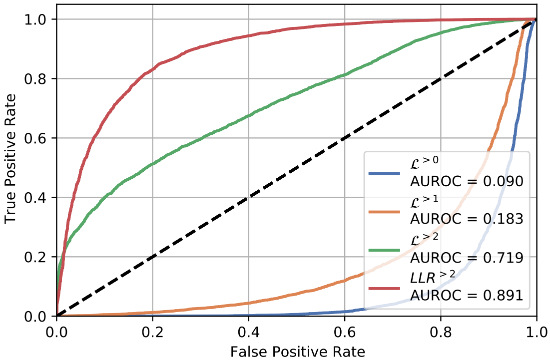
$$LLR_S^{>k}(\mathbf{x}) \to D_{\mathrm{KL}}(p(\mathbf{z}_{\leq}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \ . \tag{8}$$

We can now see that $LLR_S^{>k}(\mathbf{x})$ performs OOD detection based on the top-most latent variables.

Results with $LLR^{>k}$



(a) FashionMNIST HVAE evaluated on MNIST



(b) CIFAR10 BIVA evaluated on SVHN

Results with $LLR^{>k}$

OOD dataset	Metric	AUROC↑
Trained on CIFAR10		
SVHN	$LLR^{>2}$	0.811
CIFAR10	$LLR^{>1}$	0.469
Trained on SVHN		
CIFAR10	$LLR^{>1}$	0.939
SVHN	$LLR^{>1}$	0.489

OOD dataset	Metric	AUROC↑
Trained on FashionMNIST		
MNIST	$LLR^{>1}$	0.986
notMNIST	$LLR^{>1}$	0.998
KMNIST	$LLR^{>1}$	0.974
Omniglot28x28	$LLR^{>2}$	1.000
Omniglot28x28Inverted	$LLR^{>1}$	0.954
SmallNORB28x28	$LLR^{>2}$	0.999
SmallNORB28x28Inverted	$LLR^{>2}$	0.941
FashionMNIST	$LLR^{>1}$	0.488
Trained on MNIST		
FashionMNIST	$LLR^{>1}$	0.999
notMNIST	$LLR^{>1}$	1.000
KMNIST	$LLR^{>1}$	0.999
Omniglot28x28	$LLR^{>1}$	1.000
Omniglot28x28Inverted	$LLR^{>1}$	0.944
SmallNORB28x28	$LLR^{>1}$	1.000
SmallNORB28x28Inverted	$LLR^{>1}$	0.985
MNIST	$LLR^{>2}$	0.515

Thank you for your attention

Bibliography I

[1] Adolf Buse. "The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note". In: *The American Statistician* 36.3 (1982), pp. 153–157.

[2] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Banff, AB, Canada, 2014. arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114 (visited on 06/07/2018).

[3] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models". In: *Proceedings of Machine Learning Research*. International Conference on Machine Learning. Vol. 32. Beijing, China: PMLR, Jan. 16, 2014, pp. 1278–1286. URL: http://proceedings.mlr.press/v32/rezende14.pdf (visited on 08/12/2018).

Bibliography II

[4] Yuri Burda, Roger Grosse, and Ruslan R. Salakhutdinov. “Importance Weighted Autoencoders”. In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. San Juan, Puerto Rico, 2016, p. 8. URL: `https://arxiv.org/abs/1509.00519` (visited on 10/04/2017).

[5] Lars Maaløe et al. “BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling”. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Vancouver, Canada, Feb. 6, 2019, pp. 6548–6558. URL: `http://arxiv.org/abs/1902.02102` (visited on 03/19/2019).

Bibliography III

[6] Eric Nalisnick et al. "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019. arXiv: 1810.09136. URL: http://arxiv.org/abs/1810.09136 (visited on 10/02/2019).