OOD detection of Hierarchical VAE

Jaehyoung Hong

Unlike theoretical idea, deep generative model cannot detect out-of-distribution (OOD) inputs

- Out-of-distribution (OOD) detection success → Important when our task is filtering (Anomaly detection)
: Factory, Network (Security) and Healthcare (e.g. Dementia)

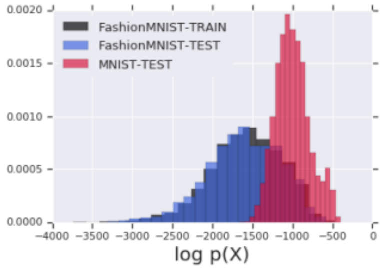
- Deep generative model: Training $\log p(X)$ → OOD detection success (Bishop,1994); Trained on simple data

- Likelihood: Higher is better

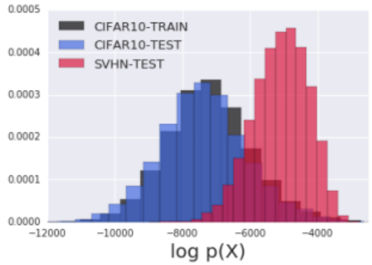
- Bits Per Dimension (BPD): Lower is better

Data Set	Avg. Bits Per Dimension
Glow Trained on FashionMNIST	
FashionMNIST-Train	2.902
FashionMNIST-Test	2.958
MNIST-Test	**1.833**
Glow Trained on MNIST	
MNIST-Test	1.262

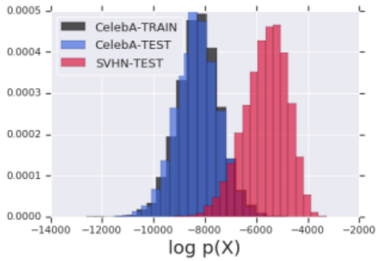
Data Set	Avg. Bits Per Dimension
Glow Trained on CIFAR-10	
CIFAR10-Train	3.386
CIFAR10-Test	3.464
SVHN-Test	**2.389**
Glow Trained on SVHN	
SVHN-Test	2.057



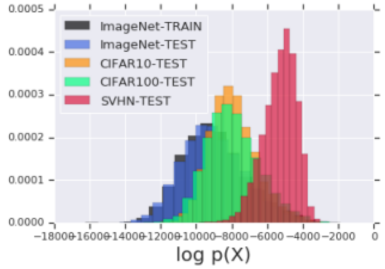
(a) Train on FashionMNIST, Test on MNIST



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



(c) Train on CelebA, Test on SVHN



(d) Train on ImageNet,
Test on CIFAR-10 / CIFAR-100 / SVHN

Why does OOD detection fail?

Latent variable $\boldsymbol{z}$: $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_L$

$\boldsymbol{x}$ $\boldsymbol{z}_1$ $\boldsymbol{z}_2$ … $\boldsymbol{z}_L$

$$\mathbb{R}^D \geq \mathbb{R}^{d_1} \geq \mathbb{R}^{d_2} \geq \mathbb{R}^{d_L}$$

Low-level feature   High-level feature

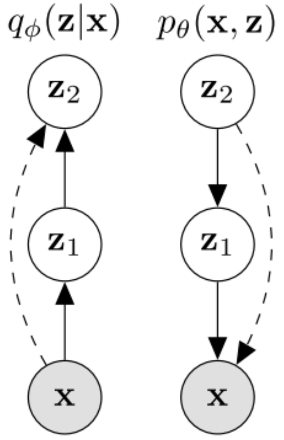
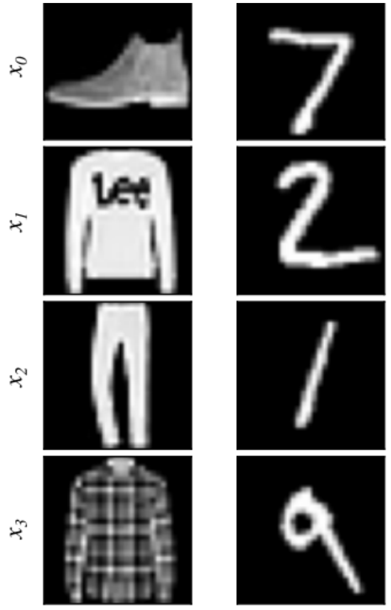


Figure 4. The inference and generative models, q_ϕ and p_θ, for an $L = 2$ layered bottom-up hierarchical VAE as the one used in our experiments. Dashed lines indicate deterministic skip connections which are employed in both networks. Skip connections are found to be useful for optimizing latent variable models (Dieng et al., 2019; Maaløe et al., 2019).

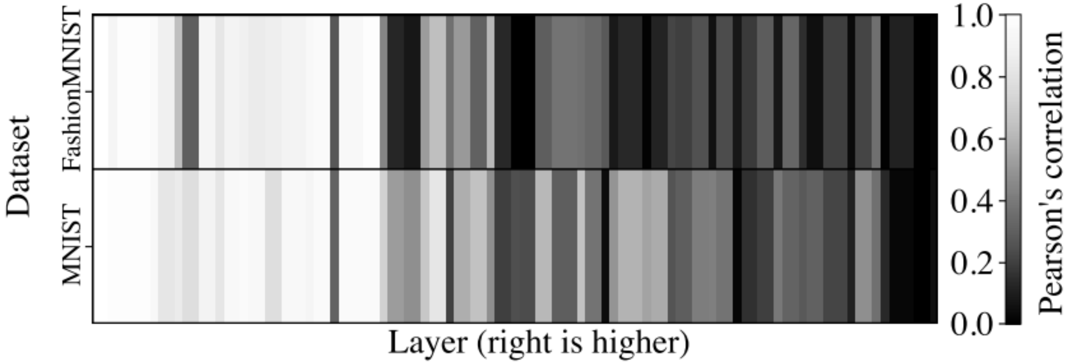
- ✓ Authors assume low-level feature is global structure
  → Important for transfer learning
  → Not suitable for OOD detection



- ✓ Low-level feature: Edge detector
- ✓ High-level feature: Cloth
- ✓ MNIST (simpler data) can sufficiently generated from edge (OOD detection Fails)

Why does OOD detection fail?

- **Low level features correlate strongly**



✓ Low-level feature is not important for OOD detection

✓ If Low-level feature is a large part of the reconstruction, then $p_\theta(\boldsymbol{x}|\boldsymbol{z}_1)$ will be high for both in- and out-of-distribution data

- **Reconstruction from each layer**

Example Reconstructions from latent hierarchy (right is higher)

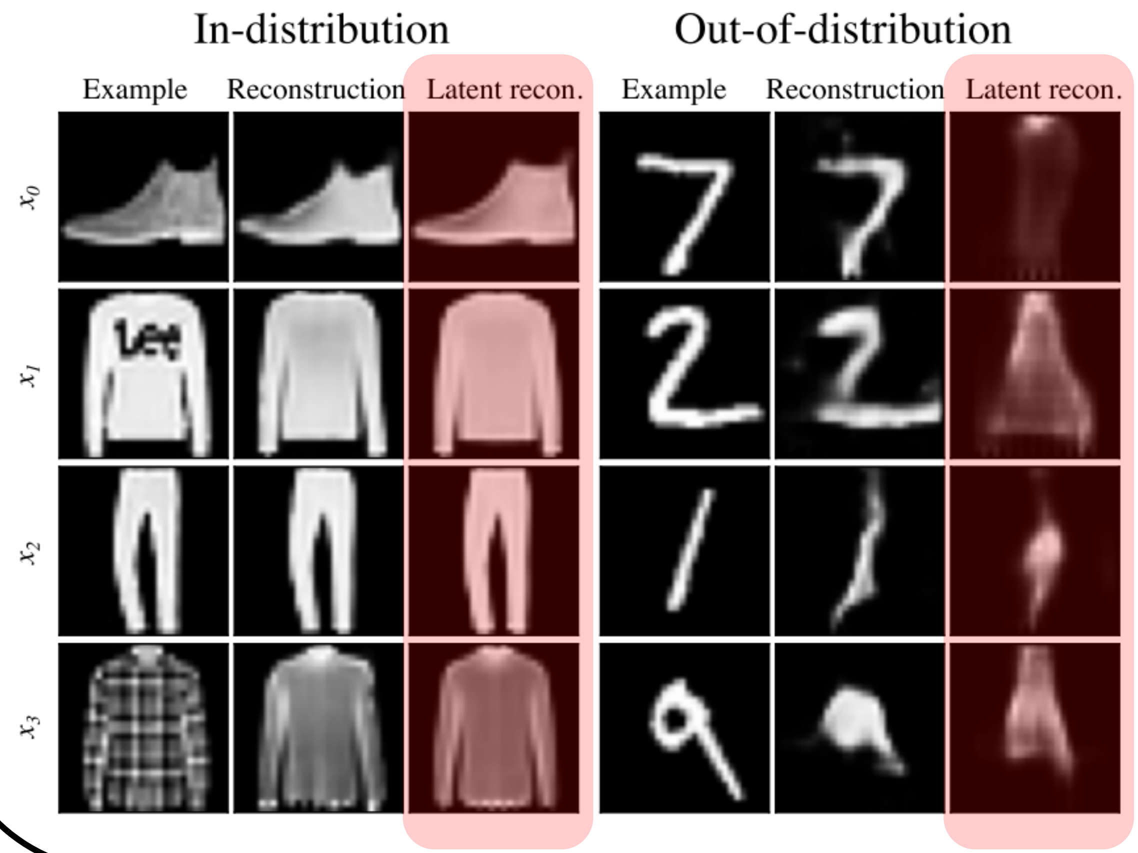
✓ Sampling from the $p(z_{\leq k}|z_{>k})$ instead of $q(z_{>k}|z_{\leq k})$

✓ Reconstruction from higher latent variable losses details (e.g. Sunglass)

➢ ELBO $\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$

Why does OOD detection fail?

- **Reconstruction from higher latent variable for OOD**

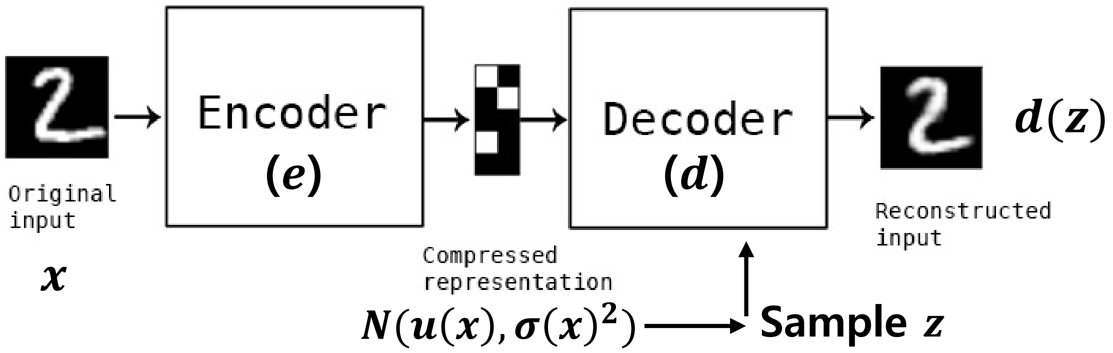


✓ Latent reconstruction: Sampling from the $p(z_{\leq k}|z_{>k})$ instead of $q(z_{>k}|z_{\leq k})$

✓ Reconstruction form higher latent variables cannot reconstruct OOD data (Stick to trained data)

Hierarchical VAE is used to make VAE more flexible

- VAE (Variational autoencoder)



✓ e, d : Deep network & $e(x) = (u(x), \sigma(x))$

✓ Cost function is depending on latent distribution

✓ Single latent variable limits the ability to learn a high likelihood representation

Hierarchical VAE
(Deeper hierarchy of latent variables)

- Hierarchical VAE

✓ Introduce hierarchy of latent variables $\boldsymbol{z} = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_L$

✓ Two types of inference model

➢ **Bottom-up**: $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_1|\boldsymbol{x}) \prod_{i=2}^{L} q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i-1})$

➢ Top-down: $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = q_\phi(\boldsymbol{z}_L|\boldsymbol{x}) \prod_{i=L-1}^{1} q_\phi(\boldsymbol{z}_i|\boldsymbol{z}_{i+1})$

✓ Until NVAE, hierarchical VAE
< SOTA autoregressive and flow-based model

Regular ELBO is not suitable for OOD detection because of bottleneck structure

- **Lowest level latent variable contribute the most to the approximate likelihood**

$$\boldsymbol{x} \quad \boldsymbol{z_1} \quad \boldsymbol{z_2} \quad \cdots \quad \boldsymbol{z_L}$$

$$\mathbb{R}^D \geq \mathbb{R}^{d_1} \geq \mathbb{R}^{d_2} \geq \mathbb{R}^{d_L}$$

➢ Generative mapping $f: \mathbb{R}^d \to \mathbb{R}^D$ s.t $\boldsymbol{x} = f(\mathbf{z}_L)$

$$f(\mathbf{z}_L) = f_1(..f_{L-1}(f_L(\mathbf{z})))$$

where $f_i: \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i-1}}$

➢ Likelihood $p(\boldsymbol{x})$

$$p(\boldsymbol{x}) = p(\mathbf{z}) \prod_{i=1}^{L} \left(\sqrt{\det \mathbf{J}_i^T \mathbf{J}_i} \right)^{-1}$$

where $\mathbf{J}_i$ is the Jacobian of f_i i.e. $J_i = \frac{\partial f_i}{\partial \mathbf{z}_i} \in \mathbb{R}^{d_i \times d_{i-1}}$

➢ Log-likelihood $p(\boldsymbol{x})$

$$\log p(\boldsymbol{x}) = \log p(\mathbf{z}) - \frac{1}{2} \sum_{i=1}^{L} \log \det \mathbf{J}_i^T \mathbf{J}_i$$

➢ We can expect that

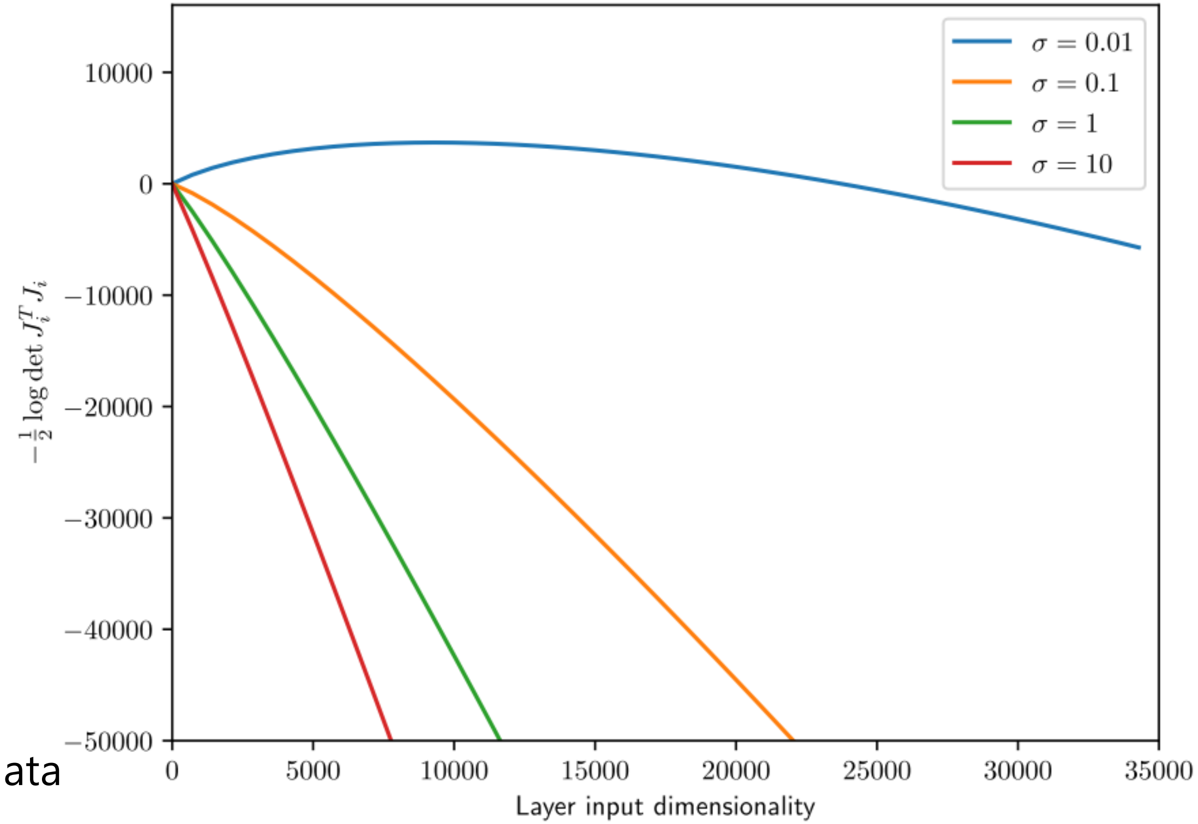
$$\det \mathbf{J}_{i+1}^T \mathbf{J}_{i+1} > \det \mathbf{J}_i^T \mathbf{J}_i$$

since determinant of $d \times d$ matrix $\approx O(\lambda^d)$

✓ Dominant of lowest level latent variable

Regular ELBO is not suitable for OOD detection because of bottleneck structure

- **Lowest level latent variable contribute the most to the approximate likelihood**



Simple data

Complex data

✓ $\mathbf{J}_i \sim N(0, \sigma^2)$ & $d_i - d_{i-1} = C$

New bound $\mathcal{L}^{>k}$ for semantic OOD detection

- **New bound $\mathcal{L}^{>k}$**
  - ELBO $\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$
  - $\mathcal{L}^{>k} = \mathbb{E}_{p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)}\left[\log \frac{p_\theta(x|Z)p_\theta(z_{>k})}{q_\phi(z_{>k}|x)}\right]$
  - $q(z_{>k}|x) = q(z_{>k}|d_k(x))$ with $d_k(x) = \mathbb{E}[q(z_k|d_{k-1}(x))]$, $d_0(x) = x$

- ✓ $\mathcal{L}^{>0}$: Regular ELBO
- ✓ $\mathcal{L} \geq \mathcal{L}^{>k}\ \forall k$ (Empirically)
- ✓ $p(z_{>k}) = p(z_L)p(z_{L-1}|z_L)\cdots p(z_{k+1}|z_{k+2})$ evaluated with samples from $q(z_{>k}|x)$
- ✓ $p(z_{\leq k}|z_{>k}) = p(z_k|z_{k+1})p(z_{k-1}|z_k)\cdots p(z_1|z_2)$ evaluated with samples from $p_\theta(z_{\leq k}|z_{>k})$

$p_\theta(x|z)$ depend on z_{k+1}

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int\int p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_2)d\mathbf{z}_1 d\mathbf{z}_2 \qquad (18)\\
&= \log \int\int \frac{q(\mathbf{z}_2|\mathbf{x})}{q(\mathbf{z}_2|\mathbf{x})} p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_1|\mathbf{z}_2)p(\mathbf{z}_2)d\mathbf{z}_1 d\mathbf{z}_2\\
&= \log \int\int q(\mathbf{z}_2|\mathbf{x})p(\mathbf{z}_1|\mathbf{z}_2)\frac{p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_2)}{q(\mathbf{z}_2|\mathbf{x})}d\mathbf{z}_1 d\mathbf{z}_2\\
&\geq \mathbb{E}_{p(\mathbf{z}_1|\mathbf{z}_2)q(\mathbf{z}_2|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z}_1)p(\mathbf{z}_2)}{q(\mathbf{z}_2|\mathbf{x})}\right] \equiv \mathcal{L}^{>1}.
\end{aligned}$$

- $q(z_2|x) = q(z_2|\mathbb{E}[q(z_1|x)])$

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \qquad (19)\\
&= \log \int \frac{q(\mathbf{z}_{>k}|\mathbf{x})}{q(\mathbf{z}_{>k}|\mathbf{x})} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}\\
&= \log \int q(\mathbf{z}_{>k}|\mathbf{x})p(\mathbf{z})\frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}_{>k}|\mathbf{x})}d\mathbf{z}\\
&= \log \int q(\mathbf{z}_{>k}|\mathbf{x})p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})p(\mathbf{z}_{>k})\frac{p(\mathbf{x}|\mathbf{z})}{q(\mathbf{z}_{>k}|\mathbf{x})}d\mathbf{z}\\
&= \log \int q(\mathbf{z}_{>k}|\mathbf{x})p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}_{>k})}{q(\mathbf{z}_{>k}|\mathbf{x})}d\mathbf{z}\\
&\geq \mathbb{E}_{p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})}\left[\log q(\mathbf{z}_{>k}|\mathbf{x})\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}_{>k})}{q(\mathbf{z}_{>k}|\mathbf{x})}\right]\\
&\geq \mathbb{E}_{p(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z}_{>k})}{q(\mathbf{z}_{>k}|\mathbf{x})}\right] \equiv \mathcal{L}^{>k}.
\end{aligned}$$

Likelihood-ratio score for OOD detection

- **Likelihood-score $LLR^{>k}$**
  - $LLR^{>k}(x) = (\mathcal{L} - L^{>k})(x)$

$$\mathcal{L} = \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right), \qquad (7)$$

$$\mathcal{L}^{>k} = \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leq k}|\mathbf{z}_{>k})q_\phi(\mathbf{z}_{>k}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right)$$

  - $LLR^{>k}(x) = -D_{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right) + D_{KL}\left(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)||p_\theta(z|x)\right)$

✓ $\mathcal{L} \geq \mathcal{L}^{>k}\ \forall k$ (Empirically) → $LLR^{>k} \geq 0$

✓ High $LLR^{>k}$ indicates $L^{>k}$ is looser on the data than the ELBO: **The data may be OOD**

✓ $LLR^{>k}$ does not include $\log p_\theta(x)$: Only depend on latent space → Suitable for OOD detection

✓ Note that $LLR^{>k}$ is only possible concept for HVAE

- **Consider stricter bound**
  - $LLR_S^{>k}(x) = (\mathcal{L}_S - L^{>k})(x)$
  - $\mathcal{L}_S = \mathbb{E}_{q(z|x)}\left[\log\left(\frac{1}{N}\sum_{s=1}^{S}\frac{p(x,z^{(s)})}{q(z^{(s)}|x)}\right)\right]$: Strictly tighter importance weighted bound (Burda et al., 2016)
  - $\mathcal{L}_S \to \log p_\theta(x)$ when $S \to \infty$ so that $LLR_S^{>k}(x) \to D_{KL}\left(p_\theta(z_{\leq k}|z_{>k})q_\phi(z_{>k}|x)||p_\theta(z|x)\right)$

✓ $Var\left(\widehat{LLR}^{>k}\right) = Var\left(\hat{\mathcal{L}}\right) + Var\left(\hat{\mathcal{L}}^{>k}\right) - 2Cov\left(\hat{\mathcal{L}}, \hat{\mathcal{L}}^{>k}\right)$

✓ $Var\left(\hat{\mathcal{L}}\right) \leq Var\left(\widehat{LLR}^{>k}\right) \leq Var\left(\hat{\mathcal{L}}^{>k}\right)$ (Empirically)

Experimental Setup

- **Model architecture**

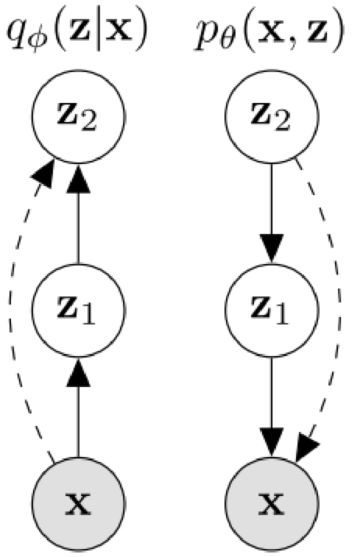


Figure 4. The inference and generative models, q_ϕ and p_θ, for an $L = 2$ layered bottom-up hierarchical VAE as the one used in our experiments. Dashed lines indicate deterministic skip connections which are employed in both networks. Skip connections are found to be useful for optimizing latent variable models (Dieng et al., 2019; Maaløe et al., 2019).

Hyperparameter	Setting/Range
All	
Optimization	Adam (Kingma & Ba, 2015)
Learning rate	$3e-4$
Batch size	128
Epochs	2000
Free bits	2 nats shared among all $\mathbf{z}_i$
Free bits constant	200 epochs
Free bits annealed	200 epochs
Activation	ReLU
Initialization	Data-dependent (Salimans & Kingma, 2016)
HVAE	
Latent dimensionality	8-16-32 (natural) / 8-16-8 (grey)
Convolution kernel	5-3-3
Stride	2-2-2 (natural) / 2-2-1 (grey)
Warmup anneal period	200 epochs
BIVA	
Latent dimensionality	10-8-6 (spatial) 42-40-38-36-34-32-30 (dense)
Convolution kernel	5-3-3-3-3-3-3-3-3-3
Stride	2-1-1-2-1-2-1-1-1-1

The baseline result of trained model

Method	Dataset	Avg. bits/dim $\log p(x)$	$\mathcal{L}^{>1}$	$\mathcal{L}^{>2}$	$\mathcal{L}^{>3}$
Trained on FashionMNIST					
Glow	FashionMNIST	2.96	-	-	
	MNIST	1.83	-	-	
HVAE (Ours)	FashionMNIST	0.420	0.476	0.579	-
	MNIST	0.317	0.601	0.881	-
Trained on CIFAR10					
Glow	CIFAR10	3.46	-	-	
	SVHN	2.39	-	-	
HVAE (Ours)	CIFAR10	3.74	17.8	54.3	75.7
	SVHN	2.62	10.2	64.0	93.9
BIVA (Ours)	CIFAR10	3.46	8.74	19.7	37.3
	SVHN	2.35	6.62	25.1	59.0

✓ $\mathcal{L}^{>k}$ is higher for OOD data as k increased

The Likelihood-based OOD detection result of trained model

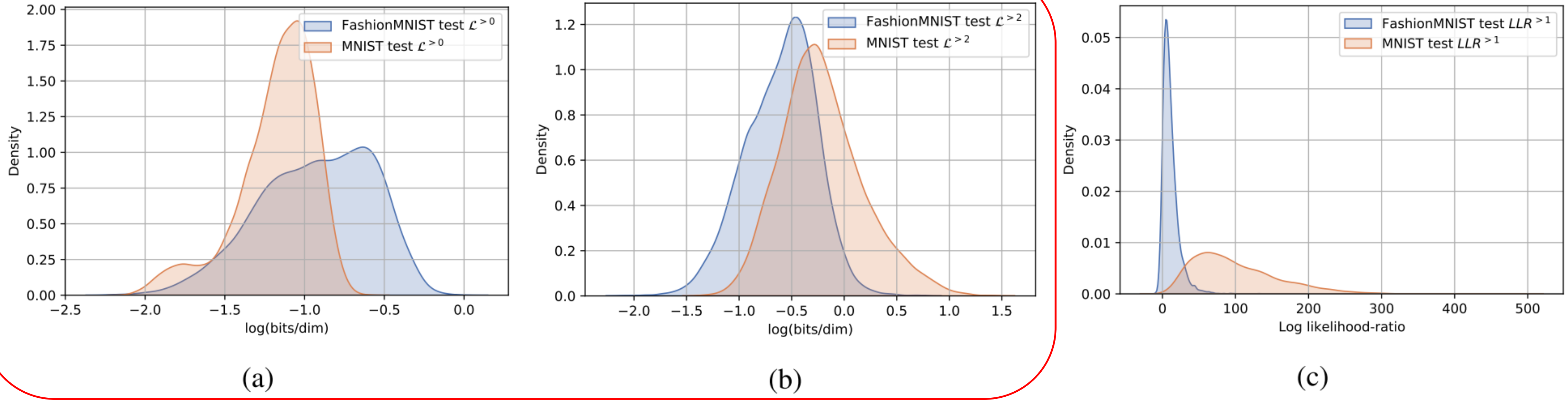
Method	AUROC↑	AUPRC↑	FPR80↓
FashionMNIST (in) / MNIST (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [1]	0.994	0.993	0.001
Backgr. contrast. LR (VAE) [7]	0.924	-	-
Binary classifier [1]	0.455	0.505	0.886
$p(\hat{y}\|\mathbf{x})$ with OOD as noise class [1]	0.877	0.871	0.195
$p(\hat{y}\|\mathbf{x})$ with calibration on OOD [1]	0.904	0.895	0.139
Input complexity (S, Glow) [9]	0.998	-	-
Input complexity (S, PixelCNN++) [9]	0.967	-	-
Use in-distribution data labels y			
$p(\hat{y}\|\mathbf{x})$ [1], [2]	0.734	0.702	0.506
Entropy of $p(y\|\mathbf{x})$ [1]	0.746	0.726	0.448
ODIN [1, 3]	0.752	0.763	0.432
VIB [4, 7]	0.941	-	-
Mahalanobis distance, CNN [1]	0.942	0.928	0.088
Mahalanobis distance, DenseNet [5]	0.986	-	-
Ensemble, 20 classifiers [1, 6]	0.857	0.849	0.240
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, VAE [7]	0.766	-	-
WAIC, 5 models, PixelCNN [1]	0.221	0.401	0.911
- Not ensembles			
Likelihood regret [8]	**0.988**	-	-
$\mathcal{L}^{>0}$ + HVAE (ours)	0.268	0.363	0.882
$\mathcal{L}^{>1}$ + HVAE (ours)	0.593	0.591	0.658
$\mathcal{L}^{>2}$ + HVAE (ours)	0.712	0.750	0.548
$LLR^{>1}$ + HVAE (ours)	0.964	0.961	0.036
$LLR^{>1}_{250}$ + HVAE (ours)	0.984	**0.984**	**0.013**

Method	AUROC↑	AUPRC↑	FPR80↓
CIFAR10 (in) / SVHN (out)			
Use prior knowledge of OOD			
Backgr. contrast. LR (PixelCNN) [1]	0.930	0.881	0.066
Backgr. contrast. LR (VAE) [8]	0.265	-	-
Outlier exposure [9]	0.984	-	-
Input complexity (S, Glow) [10]	0.950	-	-
Input complexity (S, PixelCNN++) [10]	0.929	-	-
Input complexity (S, HVAE) (Ours) [10][3]	0.833	0.855	0.344
Use in-distribution data labels y			
Mahalanobis distance [5]	0.991	-	-
No OOD-specific assumptions			
- Ensembles			
WAIC, 5 models, Glow [7]	1.000	-	-
WAIC, 5 models, PixelCNN [1]	0.628	0.616	0.657
- Not ensembles			
Likelihood regret [8]	0.875	-	-
$LLR^{>2}$ + HVAE (ours)	0.811	0.837	0.394
$LLR^{>2}$ + BIVA (ours)	**0.891**	**0.875**	**0.172**

✓ $\mathcal{L}^{>0}$ is higher for OOD data as expected, so that have inferior AUROC↑ to random

ELBO and $\mathcal{L}^{>k}$ are not sufficient for OOD detection while $LLR^{>k}$ shows potential

- **BPD densities of in-distribution and OOD data**



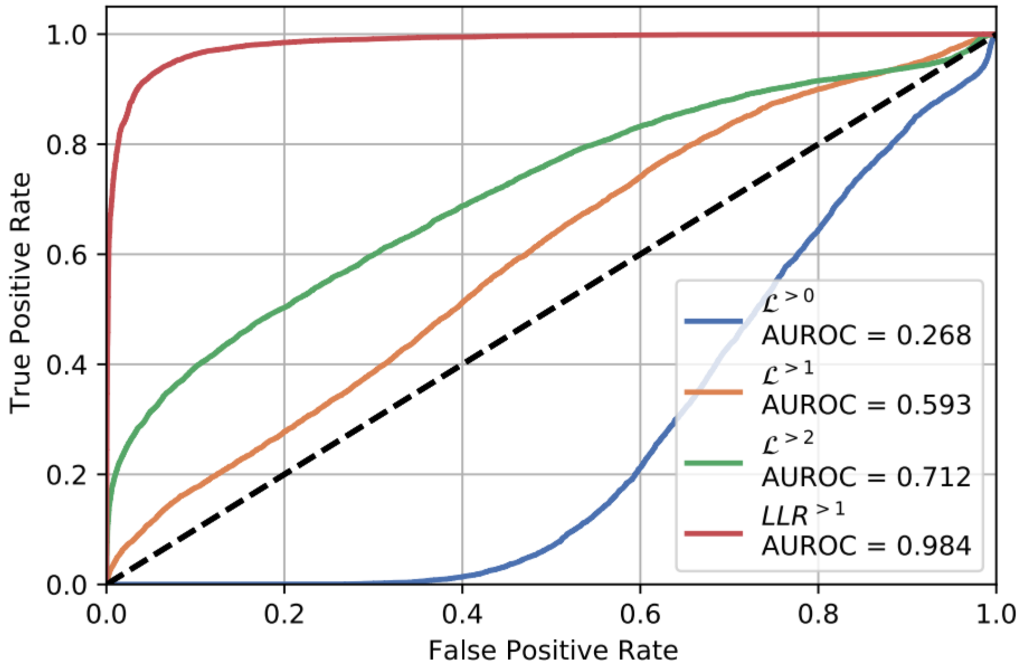
✓ Too much overlap

✓ Much less overlap

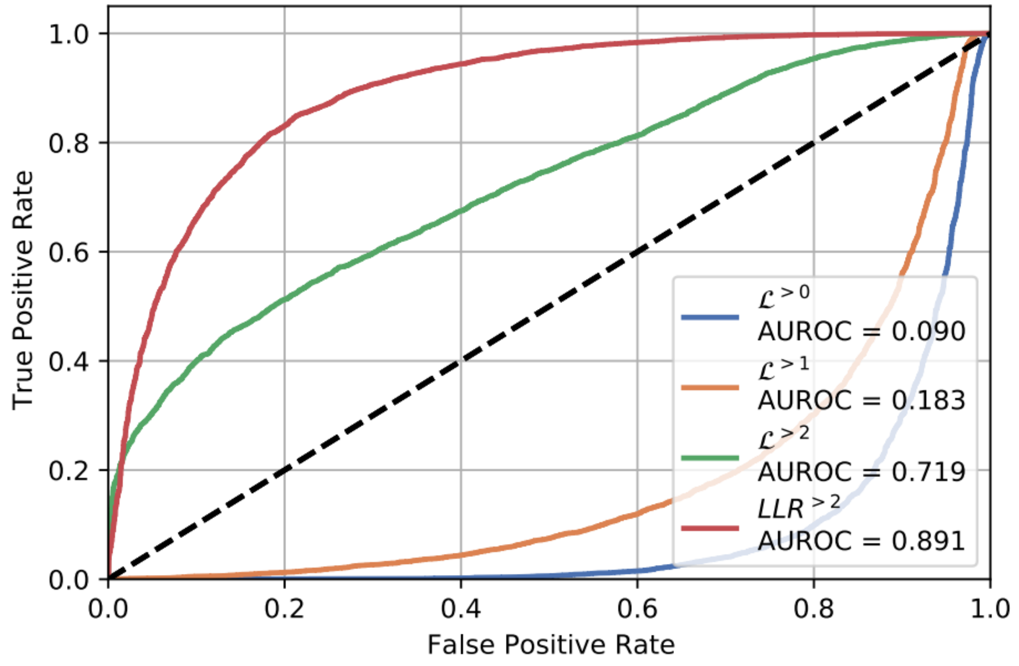
✓ Trained on FashionMNIST, test on MNIST

$LLR^{>k}$ has better performance for OOD detection

- **ROC curves for detecting MNIST**



- **ROC curves for detecting SVHN**



Two conflict intuition

- Generative model is strong for transfer learning (Complex to simple)
  - Complex data have enough 'latent information' to generate simple data

- HVAE is strong for OOD detection
  - $LLR^{>k}$ only suitable for HVAE

NVAE