VAE는 NLL base로 복잡하게 만들지만, autoregressive 보다 성능이 좋지는 않다.

그리고 flow-based는 high-dim으로 학습된다.

따라서 본문에서는 hierarchical 과 complex covariance structure model로 VAE의 성능 향상을 보인다.

↳ skip-connection과 bidirectional stochastic 를 가진 BIVA 소개.

↳ semantic latent.

* Introduction.

Implicit(GAN) 과 explicit(NLL) model 중 explicit 선택

Autoregressive는 성능은 좋지만, runtime이 좋지가 않다.

flow-based 는 시간소요적지만 학습 불포에서 autoregressive 보다 나쁨.

VAE는 posterior를 근사해야하고 low-dim manifold 에서 blured poor NLL로 보여짐.

flow로 covariance structure를 보완하려고, stacked 가 유망하지만, 여전히 성능이 낮음

논문에서는 좋은 representation의 latent는 autoregressive의 경쟁력이 있다고함.

latent의 inactive를 활성화시키고, 정보의 흐름을 향상시키기 위해, skip-connection을 사용하는 BIVA 소개.

flexible posterior approxim을 위해, bottom-up과 up-bottum의 양방향 inference 를 구성

Contribution

1. ablation

2. 성능개선

3. semi-supervise 가능

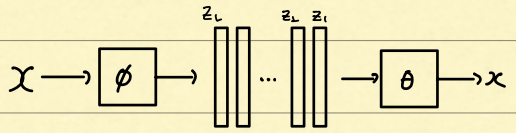
4. 이상 탐지 가능

* Variational Autoencoder.

Hierarchy 구조는 보통

$p_\theta(x,z) = p_\theta(x|z_1)p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})$

posterior approximation은 bottom-up으로 이루어짐. (1→L)



$q_\phi(z|x) = q_\phi(z_1|x)\prod_{i=1}^{L-1} q_\phi(z_{i+1}|z_i)$

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(\mathbf{z}|x)}\left[\log \frac{p_\theta(x,\mathbf{z})}{q_\phi(\mathbf{z}|x)}\right] \equiv \mathcal{L}(\theta,\phi)\ .$$

Hirarchical은 더 expressive 하지만, L이 커짐에 latent는 collapse가 일어나는 경향이 있다.

Ladder VAE는 G의 bottom-up과 inference top-down을 공유한다.

$q_{\phi,\theta}(z|x) = q_\phi(z_L|x)\prod_{i=1}^{L-1} q_{\phi,\theta}(z_i|z_{i+1},x)$

LVAE에서는 bottom-up(G) 모듈의 모든

z가 x의 영향을 받고, 이로인해 collapse가

덜 일어난다.

하지만 너무 깊으면 여전히 일어남.

∴ LVAE에서는 top-bottom path를 추가하고,

Bottom-up stochastic inference 를 구성하도록

각 z_i에서 latent의 factorization을 정의.

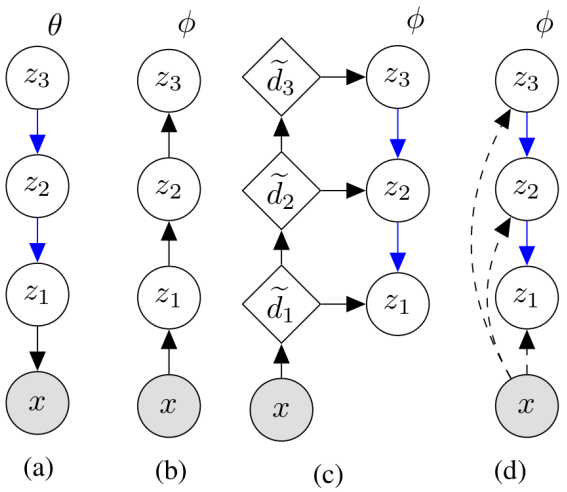


Figure 5: (a) Generative model of a VAE/LVAE with $L = 3$ stochastic variables, (b) VAE inference model, (c) LVAE inference model, and (d) skip connections among stochastic variables in the LVAE where dashed lines denote a skip-connection. Blue arrows indicate that there are shared parameters between the inference and generative model.

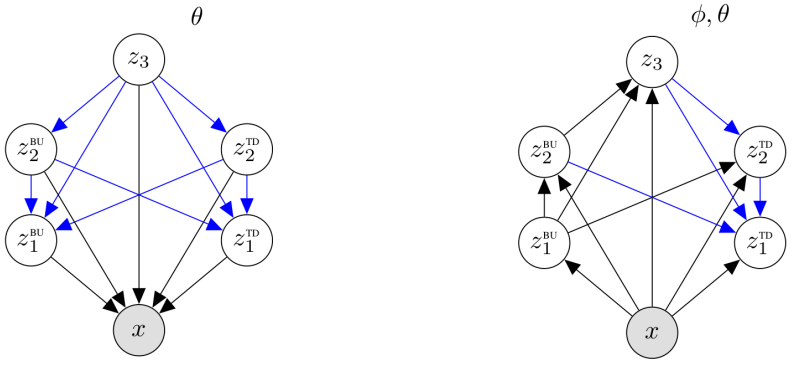
★ Bidirectional-Inference VAE.

* Model Architecture.

- Generative model

BU = Bottom-up. TD = Top-down.

latent z_i를 $z_i = [z_i^{BU}, z_i^{TD}]$로 분해



(a) Generative model (b) Inference model

Figure 1: A $L = 3$ layered BIVA with (a) the generative model and (b) inference model. Blue arrows indicate that the deterministic parameters are shared between the inference and generative models. See Appendix B for a detailed explanation and a graphical model that includes the deterministic variables.

z_{i+1}를 input으로 들어가는 NN으로 parameterization 된. deterministic TD d_i가 있다.

↳ $z_{>i}$를 포함하는 latent

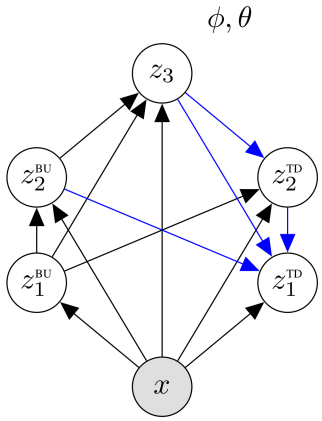
$z_i = [z_i^{BU}, z_i^{TD}]$는 concate,

$$p_\theta(x, \mathbf{z}) = p_\theta(x|\mathbf{z}) p_\theta(z_L) \prod_{i=1}^{L-1} p_\theta(z_i^{\text{BU}}|z_{>i}) p_\theta(z_i^{\text{TD}}|z_{>i}) \,,$$

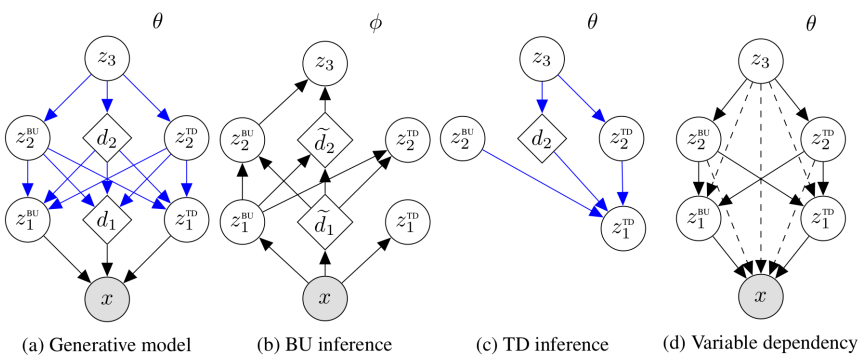
$P_\theta(x, z)$는 z_i와 직접적으로 dependent 하고. d_i로부터 $z_{>i}$와 dependent하다.

즉 z_i ($i \sim L$)는 μ와 σ의 대한 각각의 선형합의 대한 mean, var parameterization.

- Bidirectional inference network.



(b) Inference model



(a) Generative model (b) BU inference (c) TD inference (d) Variable dependency

Figure 6: A $L = 3$ layered BIVA with (a) the generative model, (b) bottom-up (BU) inference path, (c) top-down (TD) inference path, and (d) variable dependency of the generative models where dashed lines denote a skip-connection. Blue arrows indicate that the deterministic parameters are shared within the generative model or between the generative and inference model.

BU는 모든 bottom 에 대해 영향을 받는다.

z_i^{BU} 는 z_{i-1}^{BU} 에는 직접적으로 dependency. $z_{<i-1}^{\text{BU}}$ 에는 d_i 에 의해 dependent 하다.

z_i^{TD}는 z_{i+1}^{BU} 와 z_{i+1}^{TD}. 그리고 data 에 의존적.

그리고 TD path는 GCR parameter를 sharing.

$$q_\phi(\mathbf{z}|x) = q_\phi(z_L|x, z_{<L}^{\text{BU}}) \prod_{i=1}^{L-1} q_\phi(z_i^{\text{BU}}|x, z_{<i}^{\text{BU}}) q_{\phi,\theta}(z_i^{\text{TD}}|x, z_{<i}^{\text{BU}}, z_{>i}^{\text{BU}}, z_{>i}^{\text{TD}})$$

↳ ELBO 에 의해 Training.

* Motivation

- Deterministic top-down path.

Skip connection은 간단하지만, deep NN의 효과적이다.

따라서 LVAE와 deep hierarchy의 이점이 없기에 deterministic TD path를 추가하여,

Gen latent 간의 skip connection 추가. latent가 모든 hierarchy의 dependent 하도록함.

↳ 더 나은 information flow, collapse 방지

- Bidirectional Inference

Auxiliary VAE (AVAE)에서 영감을 얻음.

↳ 보조변수 a를 추가하여, covariance structure 추가.

BIVA에서는 latent를 BU, TD로 분리하여 동일한 효과

BU가 모든 bottom에 대해 factorize 되기 때문에

TDown 복잡한 cov를 학습할 수 있도록 하여, high-level semantic feature에 focus.

* Anomaly detection with BIVA

BIVA의 high-hierarchy는 high-level semantic을 capture 할 수 있다고 가정.

ELBO를 alternative log likelihood lower bound로 씀.

↳ 식을 명시적으로 나눔.

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

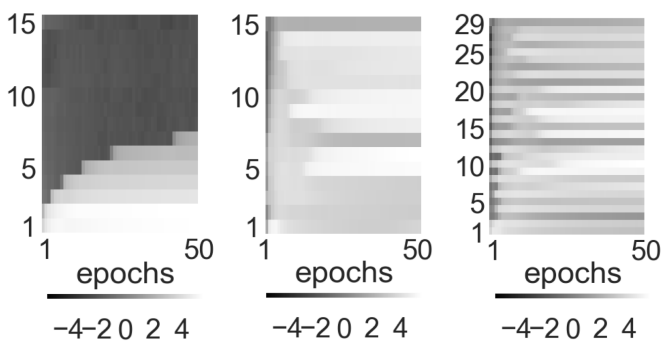
* Experiments

* Ablation.

Table 1: A comparison of the LVAE with no skip-connections and no bottom-up inference, the LVAE+ with skip-connections and no bottom-up inference, and BIVA. All models are trained on the CIFAR-10 dataset.

	PARAM.	BITS/DIM
LVAE L=15, $\mathcal{L}_1$	72.36M	$\leq$ 3.60
LVAE+ L=15, $\mathcal{L}_1$	73.35M	$\leq$ 3.41
LVAE+ L=29, $\mathcal{L}_1$	119.71M	$\leq$ 3.45
BIVA L=15, $\mathcal{L}_1$	102.95M	$\leq$ 3.12

LVAE+는 BIVA에서 BU만 없는것.



(a) LVAE $L = 15$ (b) LVAE+ $L = 15$ (c) BIVA $L = 15$

Figure 2: The $\log KL(q||p)$ for each stochastic latent variable as a function of the training epochs on CIFAR-10. (a) is a $L = N = 15$ stochastic latent layer LVAE with no skip-connections and no bottom-up inference. (b) is a $L = N = 15$ LVAE+ with skip-connections and no bottom-up inference. (c) is a $L = 15$ stochastic latent layer ($N = 29$ latent variables) BIVA for which $1, 2, ..., N$ denotes the stochastic latent variables following the order $z_1^{\text{BU}}, z_1^{\text{TD}}, z_2^{\text{BU}}, z_2^{\text{TD}}, ..., z_L$.

↳ LVAE는 기껏해야 활성화.

∴ BU가 latent 변수에 더 정보 전달을 하나보이 기여

* Binary Image.

Table 2: Test log-likelihood on statically binarized MNIST for different number of importance weighted samples. The finetuned models are trained for an additional number of epochs with no *free bits*, $\lambda = 0$. For testing resiliency we trained 4 models and evaluated the standard deviations to be ± 0.031 for $\mathcal{L}_1$.

	$-\log p(x)$
With autoregressive components	
PIXELCNN [57]	$= 81.30$
DRAW [13]	< 80.97
IAFVAE [23]	≤ 79.88
PIXELVAE [14]	≤ 79.66
PIXELRNN [57]	$= 79.20$
VLAE [5]	≤ 79.03
Without autoregressive components	
DISCRETE VAE [42]	≤ 81.01
BIVA, $\mathcal{L}_1$	≤ 81.20
BIVA, $\mathcal{L}_{1e3}$	≤ 78.67
BIVA FINETUNED, $\mathcal{L}_1$	≤ 80.47
BIVA FINETUNED, $\mathcal{L}_{1e3}$	≤ 78.59

Table 3: Semi-supervised test error for BIVA on MNIST for 100 randomly chosen and evenly distributed labelled samples.

	ERROR %
M1+M2 [22]	3.33% ($\pm$0.14)
VAT [32]	2.12%
CATGAN [51]	1.91% ($\pm$0.10)
SDGM [31]	1.32% ($\pm$0.07)
LADDERNET [38]	1.06% ($\pm$0.37)
ADGM [31]	0.96% ($\pm$0.02)
IMPGAN [44]	0.93% ($\pm$0.07)
TRIPLEGAN [29]	0.91% ($\pm$0.58)
SSLGAN [6]	0.80% ($\pm$0.10)
BIVA	0.83% ($\pm$0.02)

* Natural Image.

Table 4: Test log-likelihood on CIFAR-10 for different number of importance weighted samples. We evaluated two different BIVA with various number of layers (L). For testing resiliency we trained 3 models and evaluated the standard deviations to be ± 0.013 for $\mathcal{L}_1$ and $L = 15$.

	BITS/DIM
With autoregressive components	
CONVDRAW [12]	< 3.58
IAFVAE $\mathcal{L}_1$ [23]	≤ 3.15
IAFVAE $\mathcal{L}_{1e3}$ [23]	≤ 3.12
GATEDPIXELCNN [56]	$= 3.03$
PIXELRNN [57]	$= 3.00$
VLAE [5]	≤ 2.95
PIXELCNN++ [45]	$= 2.92$
Without autoregressive components	
NICE [8]	$= 4.48$
DEEPGMMS [58]	$= 4.00$
REALNVP [9]	$= 3.49$
DISCRETEVAE++ [54]	≤ 3.38
GLOW [21]	$= 3.35$
FLOW++ [16]	$= 3.08$
BIVA L=10, $\mathcal{L}_1$	≤ 3.17
BIVA L=15, $\mathcal{L}_1$	≤ 3.12
BIVA L=15, $\mathcal{L}_{1e3}$	≤ 3.08

ImageNet은 훨씬 복잡하니 데이터

Auto regressive 보다는 생성

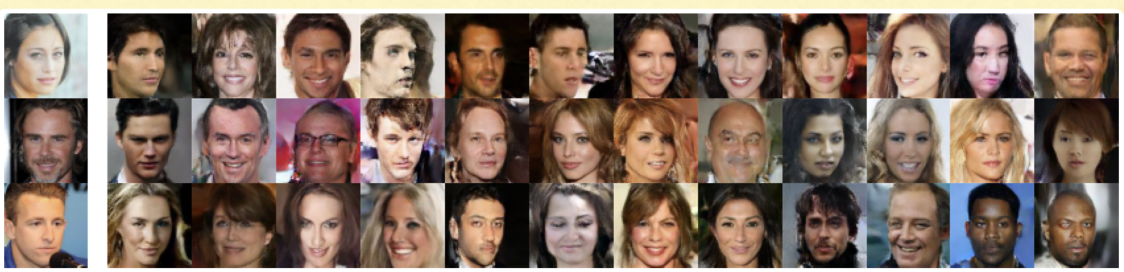


Figure 3: (left) images from the CelebA dataset preprocessed to 64x64 following [27]. (right) $\mathcal{N}(0, I)$ generations of BIVA with $L = 20$ layers that achieves a $\mathcal{L}_1 = 2.48$ bits/dim on the test set.

너 흐릿하지 않은 이미지생성

* Conclusion.

BIVA를 제시하고, semi-supervised에서 우수한 성능