* Introduction.

대부분의 VAE는 prior, tighter bound, posterior estimate의 초점을 맞춘다.

Arch는 classification과 대부분 비슷한데, 기본적으로 VAE의 arch는 요구 사항이 조금 다르다.

VAE는 input과 latent 간의 상호정보 maximize.

↳ input의 복원에는 관심없는 classification과 다름

Decoder는 generator이기 때문에, overfit되면 안되지만, encoder는 괜찮다.

encoder-based가 아닌 marginal LL이나 encoder overfit에 민감하지 않다.

그리고 큰 receptive field가 필요. ELBO 때문에 불안정 (연산 VAE는 그래서 BN사용)

따라서 NVAE 제안. → non-autoregressive likelihood. 중 SOTA

↳ depthwise conv 사용.

↳ BN은 좋고, hierarchical 이 많아질수록 unstable.

↳ KL minimize를 위해, residual parameterization과 spectral Norm이 stable에 중요.

즉, 새로운 계층 VAE 제시. (NVAE)

새로운 posterior approximate residual parameterization.

SN으로 stable.

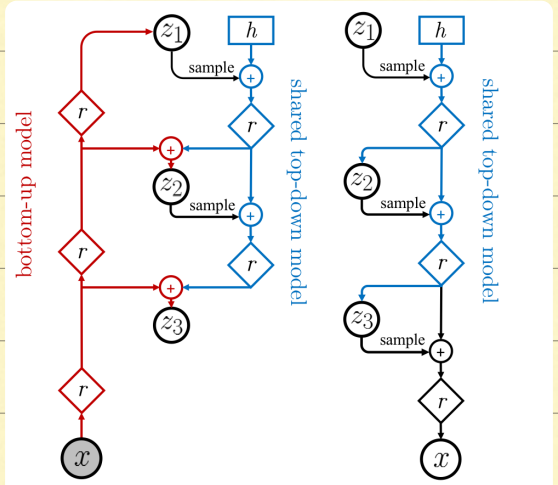
메모리 부담을 줄이는 solution

기본 VAE에서 초차도 성능↑.

IAF-VAE과 관련 있지만, 성능↑.

BIVA보다 간단하지만 성능↑.

* Related Work.




(a) Bidirectional Encoder (b) Generative Model

Figure 2: The neural networks implementing an encoder $q(\boldsymbol{z}|\boldsymbol{x})$ and generative model $p(\boldsymbol{x}, \boldsymbol{z})$ for a 3-group hierarchical VAE. (r) denotes residual neural networks, (+) denotes feature combination (e.g., concatenation), and [h] is a trainable parameter.

* Method.

NVAE의 challenge.

1) VAE에 맞는 network

2) scale up, stability를 유지한 hierarchical.

* Residual cell for VAE

VAE에서 문제는 long-term dependency.

NVAE 같은 Unconditional 에서는 hierarchical로 global 부터 fine 까지.

* Residual Cell for Generative Model.

kernel size를 늘리는 것은 receptive ↔ param trade off.

Depthwise는 즉 각 channel 별 개별에 대응해 표현력 ↓

∴ MobileNet V2와 같이, 1x1 channel conv 축 적용.

- BN.

최근 BN의 성능저하로 대부분 WN을 쓴다.

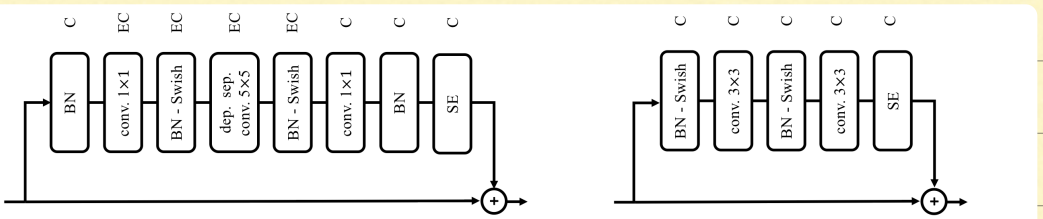
하지만 Testing시에만 일어난다. train과 test의 동작이 약간 다르기 때문.

∴ BN의 momentum을 조절하고, 또한 불편하고 BN에 의해 종속되지 않도록 scaling parameterization의 reg

- Swish.

- Squeeze and Excitation

- Final cell.




(a) Residual Cell for NVAE Generative Model

(b) Residual Cell for NVAE Encoder

Figure 3: The NVAE residual cells for generative and encoder models are shown in (a) and (b). The number of output channels is shown above. The residual cell in (a) expands the number of channels $E$ times before applying the depthwise separable convolution, and then maps it back to $C$ channels. The cell in (b) applies two series of BN-Swish-Conv without changing the number of channels.

* Residual cell for Encoder.

Depthwise는 bottom-up Encoder에는 효과x

∴ 일반 conv

그래서 논문과 BN-conv-activate 가 다른 듯.

* Reducing Memory require.

Depthwise는 memory가 많이 필요하다.

따라서 그래서 tricks

1). APEX의 mixed precision.

2) 그림 3에서 각 feature map의 분산값의 backward를 통해 저장함.

따라서 메모리를 줄이기 위해 ReLU, swish를 활성화하고, backward에 대해 하나의 feature map만 저장.

↳ = gradient checking points.

* Taming the Unbounded KL Term

Hierarchical에서 $p(z_{<l}|x)$는 모델링하다.

Encoder dec과 훈련 중 너무 먼 distribution을 생성하면 KL이 급격하게 weight를 update하게되고, unstable하다.

- Residual Normal Distribution.

$p(z_l^i|z_{<l}) = \mathcal{N}(\mu_i(z_{<l}), \sigma_i(z_{<l}))$

$q(z_l^i|z_{<l}, x) = \mathcal{N}(\mu_i(z_{<l}) + \Delta\mu_i(z_{<l}, x), \sigma_i(z_{<l}) \cdot \Delta\sigma_i(z_{<l}, x))$

↳ prior의 대한 posterior approximate의 상대적 location, scale.

↳ prior가 움직이면 posterior도 움직임.

$$\mathrm{KL}\big(q(z^i|\boldsymbol{x})||p(z^i)\big) = \frac{1}{2}\left(\frac{\Delta\mu_i^2}{\sigma_i^2} + \Delta\sigma_i^2 - \log\Delta\sigma_i^2 - 1\right),$$

↳ Decoder의 의한 $\sigma_i$가 below bound 될 경우, encoder의 의존한다.

↳ weight averaging prior, posterior도 같은 맥락.

- Spectral Regularization (SR)

KL을 bound 하기 위해, encoder의 출력이 input의 변화에도 급격히 바뀌지 않아야한다.

↳ Lipschiz regularization.

∴ SR 적용

$\mathcal{L}_{SR} = \lambda\Sigma_i s^{(i)}$ , $s^{(1)}$ = largest singular value.

- More expressive Approximate Posteriors with Normalizing flow.

NVAE는 $p(z)$와 $q(z|x)$의 그룹간 autoregressive 분포와 각 그룹의 독립적 분포로 modeling

따라서 병렬 sampling 이지만 representation↓.

∴ 추가적인 normalization flow 수행.

encoder 에만 적용되기 때문에 explicit inversion이 필요없고, IAF 사용가능. Sampling 시간이 증가하지 않음.

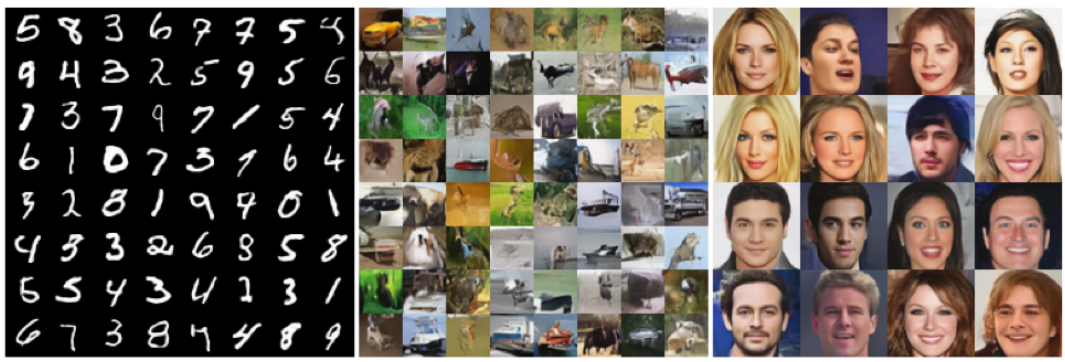
* Experiments.

* Main quantitative results.

Table 1: Comparison against the state-of-the-art likelihood-based generative models. The performance is measured in bits/dimension (bpd) for all the datasets but MNIST in which negative log-likelihood in nats is reported (lower is better in all cases). NVAE outperforms previous non-autoregressive models on most datasets and reduces the gap with autoregressive models.

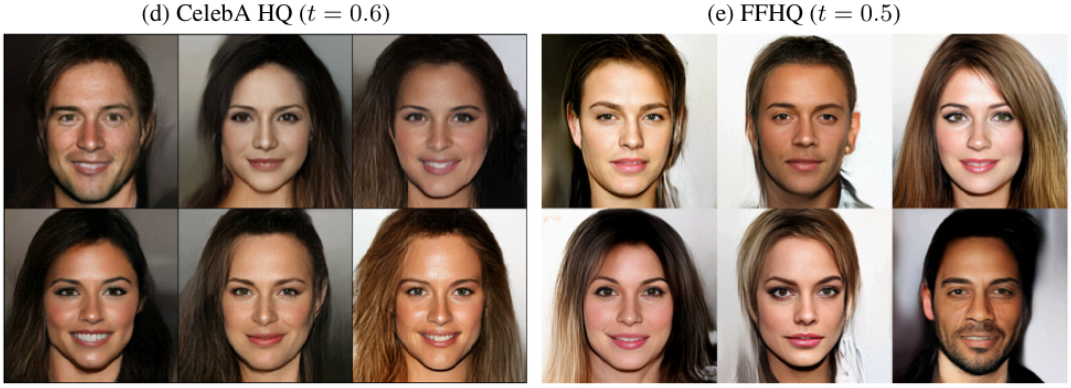
| Method | MNIST 28×28 | CIFAR-10 32×32 | ImageNet 32×32 | CelebA 64×64 | CelebA HQ 256×256 | FFHQ 256×256 |
|---|---|---|---|---|---|---|
| NVAE w/o flow | **78.01** | 2.93 | - | 2.04 | - | 0.71 |
| NVAE w/ flow | 78.19 | **2.91** | 3.92 | **2.03** | **0.70** | **0.69** |
| **VAE Models with an Unconditional Decoder** | | | | | | |
| BIVA [36] | 78.41 | 3.08 | 3.96 | 2.48 | - | - |
| IAF-VAE [4] | 79.10 | 3.11 | - | - | - | - |
| DVAE++ [20] | 78.49 | 3.38 | - | - | - | - |
| Conv Draw [42] | - | 3.58 | 4.40 | - | - | - |
| **Flow Models without any Autoregressive Components in the Generative Model** | | | | | | |
| VFlow [59] | - | 2.98 | - | - | - | - |
| ANF [60] | - | 3.05 | 3.92 | - | 0.72 | - |
| Flow++ [61] | - | 3.08 | **3.86** | - | - | - |
| Residual flow [50] | - | 3.28 | 4.01 | - | 0.99 | - |
| GLOW [62] | - | 3.35 | 4.09 | - | 1.03 | - |
| Real NVP [63] | - | 3.49 | 4.28 | 3.02 | - | - |
| **VAE and Flow Models with Autoregressive Components in the Generative Model** | | | | | | |
| $\delta$-VAE [25] | - | 2.83 | 3.77 | - | - | - |
| PixelVAE++ [35] | 78.00 | 2.90 | - | - | - | - |
| VampPrior [64] | 78.45 | - | - | - | - | - |
| MAE [65] | 77.98 | 2.95 | - | - | - | - |
| Lossy VAE [66] | 78.53 | 2.95 | - | - | - | - |
| MaCow [67] | - | 3.16 | - | - | 0.67 | - |
| **Autoregressive Models** | | | | | | |
| SPN [68] | - | - | 3.85 | - | 0.61 | - |
| PixelSNAIL [34] | - | 2.85 | 3.80 | - | - | - |
| Image Transformer [69] | - | 2.90 | 3.77 | - | - | - |
| PixelCNN++ [70] | - | 2.92 | - | - | - | - |
| PixelRNN [41] | - | 3.00 | 3.86 | - | - | - |
| Gated PixelCNN [71] | - | 3.03 | 3.83 | - | - | - |

- Qualitative Results.



(a) MNIST ($t = 1.0$) (b) CIFAR-10 ($t = 0.7$) (c) CelebA 64 ($t = 0.6$)

(d) CelebA HQ ($t = 0.6$) (e) FFHQ ($t = 0.5$)



(f) MaCow [67] trained on CelebA HQ ($t = 0.7$) (g) Glow [62] trained on CelebA HQ ($t = 0.7$)

Figure 4: (a)-(e) Sampled images from NVAE with the temperature in prior ($t$). (f)-(g) A few images generated by MaCow [67] and Glow [62] are shown for comparison (images are from the original publications). NVAE generates diverse high quality samples even with a small temperature, and it exhibits remarkably better hair details and diversity (best seen when zoomed in).

* Ablation studies.

- Norm and Activation flow.

Table 2: Normalization & activation

| Functions | $L = 10$ | $L = 20$ | $L = 40$ |
|---|---|---|---|
| WN + ELU | 3.36 | 3.27 | 3.31 |
| BN + ELU | 3.36 | 3.26 | 3.22 |
| BN + Swish | **3.34** | **3.23** | **3.16** |

- Residual Cell.

Table 3: Residual cells in NVAE

| Bottom-up model | Top-down model | Test (bpd) | Train time (h) | Mem. (GB) |
|---|---|---|---|---|
| Regular | Regular | 3.11 | 43.3 | 6.3 |
| Separable | Regular | 3.12 | 49.0 | 10.6 |
| Regular | Separable | **3.07** | 48.0 | 10.7 |
| Separable | Separable | **3.07** | 50.4 | 14.9 |

- Residual Normal Distribution.

Residual도 역시 posterior collapse에 나쁜 영향을 끼치지 않음.

- The effect of SR and SE

Table 5: SR & SE

| Model | Test NLL |
|---|---|
| NVAE | **3.16** |
| NVAE w/o SR | 3.18 |
| NVAE w/o SE | 3.22 |

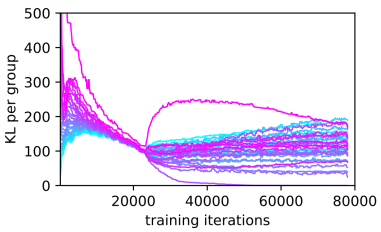
- Sampling Speed.

빠름.

- Reconstruction

  좋음.

(a) Reconstruction results (best seen when zoomed in).




(b) Average KL per group.

Figure 5: (a) Input input on the left and reconstructed images on the right for CelebA HQ. (b) KL per group on CIFAR-10.

* Conclusion.

  NVAE 에서

  ㄴ Generator 에는 depthwise 를 사용하고, Encoder 에는 일반 residual 사용.

  ㄴ Enc 에서 normal distribution 의 residual parameterization 과, stable 을 위한 SN 적용

  ㄴ 또한 memory 줄임.

  ㄴ 고화질의 큰 image,

  ㄴ 신뢰한 architecture 소개.