# Hierarchical Variational Autoencoder and Denoise Diffusion Probabilistic Model

**Duan zhibin**

# Variational Autoencoder

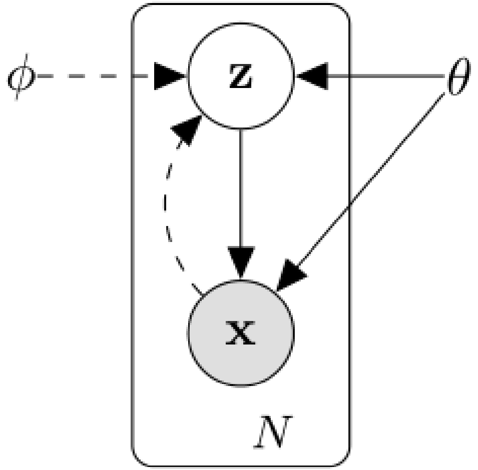



Figure 1: The type of directed graphical model under consideration.

- Generate model

  Given $N$ i.i.d sameples: $X = \left\{x^{(i)}\right\}_i^N$

  $(i).\quad z^{(i)} \sim p_{\theta^*}(z)$

  $(ii).\quad x^{(i)} \sim p_{\theta^*}(x \mid z)$

- Intractable posterior distributions

  $$p_\theta(z \mid x) = p_\theta(x \mid z)\, p_\theta(z) / p_\theta(x)$$

- Variational posterior

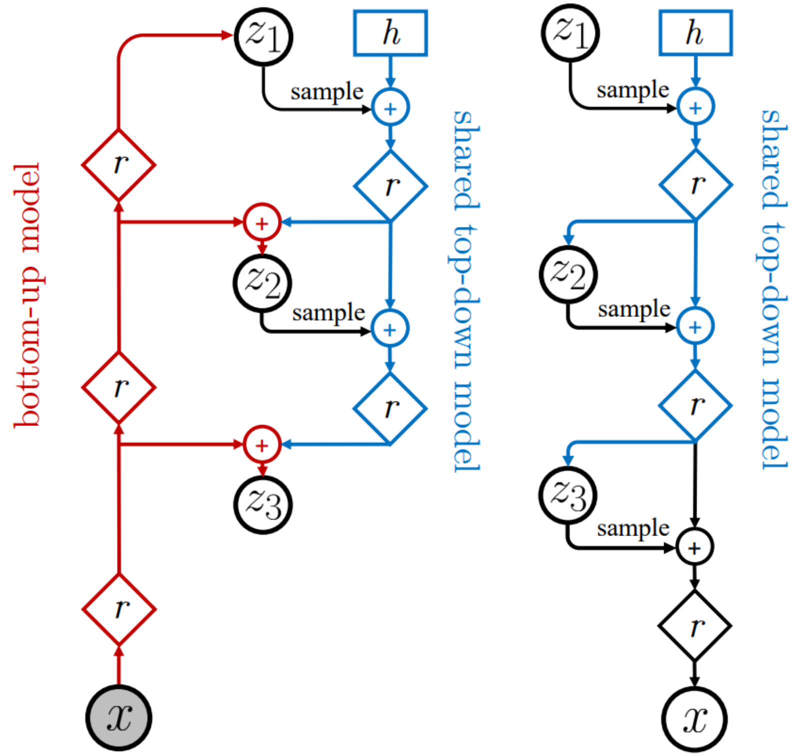
  $$q_\phi\left(z^{(i)} \mid x^{(i)}\right) \sim N\left(f_\mu\left(x^{(i)}\right), f_\sigma\left(x^{(i)}\right)\right)$$

- The variational bound

  $$\log p_\theta\left(x^{(i)}\right) = D_{KL}\left(q_\phi\left(z^{(i)} \mid x^{(i)}\right) \| \, p_\theta\left(z^{(i)} \mid x^{(i)}\right)\right) + L\left(\theta, \phi; x^{(i)}\right)$$


[1] Auto-Encoding Variational Bayes. Diederik P. Kingma, Max Welling. NIPS 2014

# Hierarchical variational autoencoder




(a) Bidirectional Encoder (b) Generative Model

- Generate model

Given $N$ i.i.d sameples: $\mathrm{X} = \left\{ x^{(i)} \right\}_i^N$

$$(i).\quad z_1^{(i)} \sim p_{\theta^*}(z_1), \cdots, z_l^{(i)} \sim p_{\theta^*}(z_l),$$

$$(ii).\quad x^{(i)} \sim p_{\theta^*}(x \mid z_L)$$

- Variational conditional posterior

$$q_\phi\left(z_l^{(i)} \mid x_t^{(i)}, z_{l-1}^{(i)}\right) \sim N\left(f_\mu\left(x^{(i)}, z_{l-1}^{(i)}\right), f_\sigma\left(x^{(i)}, z_{l-1}^{(i)}\right)\right)$$

- Training object

$$L_{\mathrm{VAE}}(x) := \mathrm{E}_{q(z|x)}\left[\log p(x \mid z) - \mathrm{KL}\left(q(z_1 \mid x) \,\|\, p(z_1)\right)\right]$$
$$- \sum_{l=2}^{L} \mathrm{E}_{q(z<l|x)}\left[\mathrm{KL}\left(q(z_l \mid x, z_{<l}) \,\|\, p(z_l \mid z_{<l})\right)\right]$$


[1] Vahdat, Arash, and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder." NeurIPS 2020

# Hierarchical variational autoencoder

Fig 2: unconditional CIFAR10 generative
From NVAE

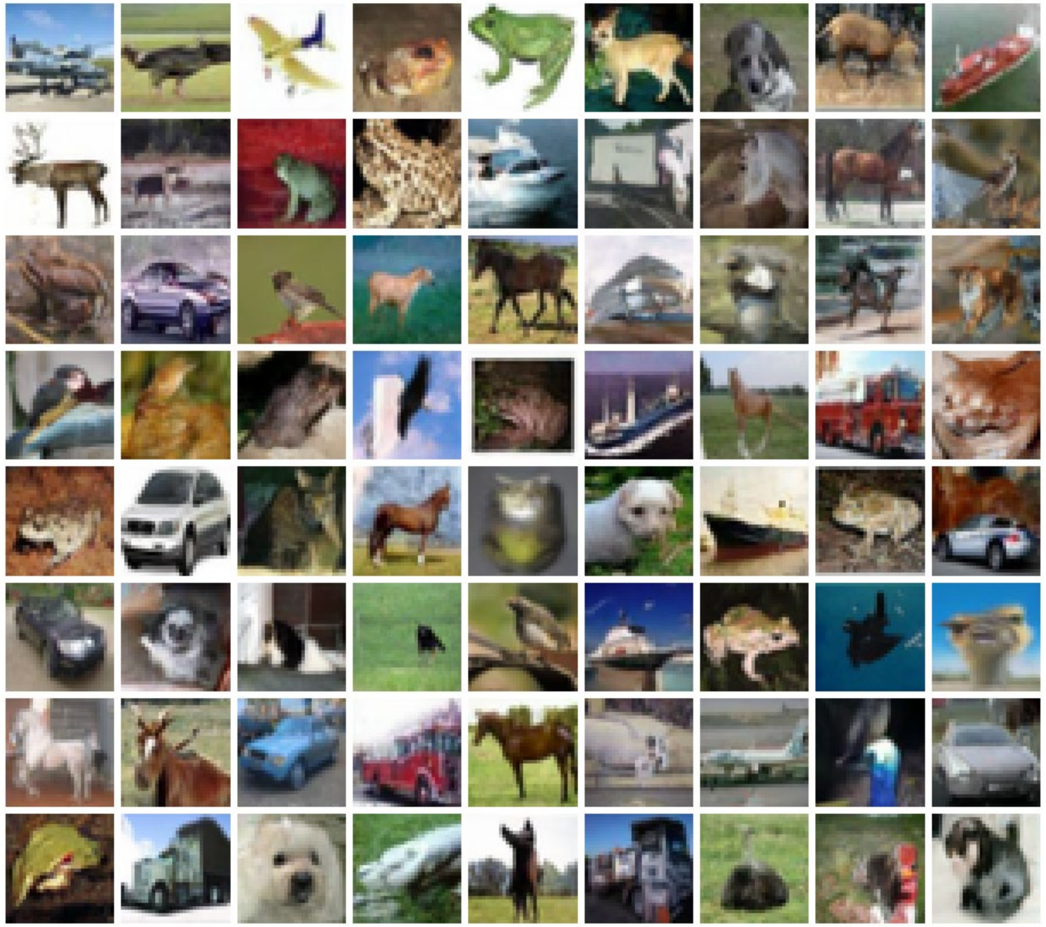


Fig 2: unconditional CIFAR10 generative
from denoise diffusion model

[1] Vahdat, Arash, and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder." NeurIPS 2020

# Denoise diffusion probabilistic model

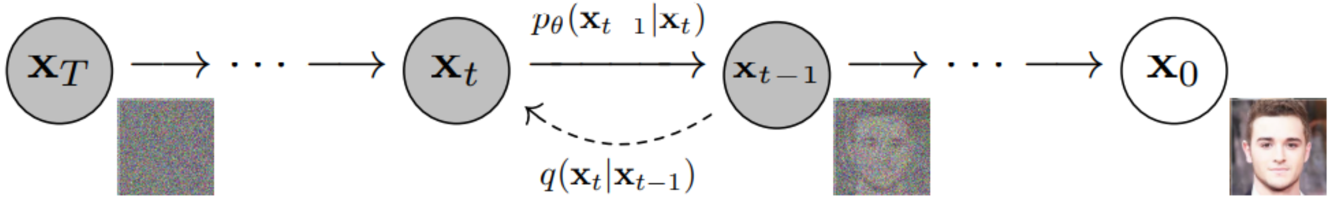



Figure 2: The directed graphical model in this work.

- Encoder

$$\begin{cases} q(x_t \mid x_{t-1}) = N\left(x_t; \alpha_t x_{t-1}, \beta_t^2 I\right) \\ q(x_{1:T} \mid x_0) := \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \end{cases}$$

$$\Rightarrow q(x_{1:T} \mid x_0) := N\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right)$$

- Analytic conditional posterior

$$p(x_{t-1} \mid x_t, x_0) = N\left(x_{t-1}; \frac{\alpha_t \bar{\beta}_{t-1}}{\bar{\beta}_t^2} x_t + \frac{\bar{\alpha}_{t-1} \bar{\beta}_t^{2}}{\bar{\beta}_t^2} x_0, \frac{\bar{\beta}_{t-1}^{2} \beta_t^2}{\bar{\beta}_t^{2}} I\right),$$


[1] Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics.“ ICML2015

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models.“ NeuiIPS 2020.

# Denoise diffusion probabilistic model

- Generative model

$$p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$$

$$p_\theta(x_{t-1} \mid x_t) := N\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

- Object function

$$L_{\mathrm{vb}} = \mathbb{E}_{q(\boldsymbol{x}_0)}\Bigg[\underbrace{D_{\mathrm{KL}}[q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)]}_{L_T} + \sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\big[D_{\mathrm{KL}}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)]\big]}_{L_{t-1}}$$

$$\underbrace{-\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]}_{L_0}\Bigg]. \quad (1)$$

[1] Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics.“ ICML2015

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models.“ NeuiIPS 2020.

# Lossy compressor

$$x \xrightarrow[q]{\textit{lossily}\ \text{mappling}} z \xrightarrow[p]{\textit{decoding}} \hat{x}$$

- Rate

$$p(z) = N(0, I)$$

$$R = E_{p_{data}(x)} D_{KL}\left(q(z \mid x) \,\|\, p(z)\right)$$

- Distortion

$$D = E_{p_{data}(x)} E_{q(z|x)} f\left(x, d(z)\right)$$

- A rate-distortion trade-off

$$L_{vae}\left(x^{(i)}\right) = D_{KL}\left(q_\phi\left(z^{(i)} \mid x^{(i)}\right) \| \, p_\theta\left(z^{(i)} \mid x^{(i)}\right)\right) + L\left(\theta, \phi; x^{(i)}\right)$$

$$= R\left(x^{(i)}\right) + D\left(x^{(i)}\right)$$

- $\gamma -$ optimal

$$\lambda^* = \min_{d,p,q} E_{p_{data}(x)} E_{q(z|x)}\left(x, d(z)\right)$$

$$s.t. \quad \min_{d,p,q} E_{p_{data}(x)} D_{KL}\left(q(z \mid x) \,\|\, p(z)\right) \le \gamma$$

- $\lambda -$ optimal

$$\gamma^* = \min_{d,p,q} E_{p_{data}(x)} D_{KL}\left(q(z \mid x) \,\|\, p(z)\right)$$

$$s.t. \quad E_{p_{data}(x)} E_{q(z|x)}\left(x, d(z)\right) \le \lambda$$

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# Progressively lossier compressors

- Rate

$$Given: p\left(z_{1:k}\right)$$

$$R_k = E_{p_{data}(x)} D_{KL}\left(q\left(z_{1:k} \mid x\right) \| p\left(z_{1:k}\right)\right)$$

- Distortion

$$D_k = E_{p_{data}(x)} E_{q\left(z_{1:k} \mid x\right)} f\left(x, d_k\left(z_{1:k}\right)\right)$$

$$A \ sequence \ \ of \left(R_k, D_k\right), and \ \ d_k$$

- Pareto Optimality

帕累托最优（Pareto Optimality），是指资源分配的一种理想状态，假定固有的一群人和可分配的资源，从一种分配状态到另一种状态的变化中，在没有使任何人境况变坏的前提下，使得至少一个人变得更好。

- VAE are pareto optimality

$$L_{vae}\left(x^{(i)}\right) = D_{KL}\left(q_\phi\left(z^{(i)} \mid x^{(i)}\right) \| p_\theta\left(z^{(i)} \mid x^{(i)}\right)\right) + L\left(\theta, \phi; x^{(i)}\right)$$

- HVAE are not pareto optimality

$$L_{\mathrm{VAE}}(x) := \mathrm{E}_{q(z|x)}\left[\log p(x \mid z) - \mathrm{KL}\left(q\left(z_1 \mid x\right) \| p\left(z_1\right)\right)\right] - \sum_{l=2}^{L} \mathrm{E}_{q(z<l|x)}\left[\mathrm{KL}\left(q\left(z_l \mid x, z_{<l}\right) \| p\left(z_l \mid z_{<l}\right)\right)\right]$$


[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# DDPM are good progressively lossier compressors

- Progressive coding objective

$$\min_{d_{1:T},p,q} \sum_{k=1}^{T} \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q(z|x)} f(x, d_k(z_{1:k}))$$

$$\text{s.t. } \mathbb{E}_{p_{\text{data}}(x)} D_{\text{KL}}(q(z_{1:k} \mid x) \parallel p(z_{1:k})) \leq \gamma_k \quad \forall k \in \{1, \ldots, T\}.$$

- Optimizing

$$d_k^* = \arg\min_{d_k} \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q(z_{1:k}|x)} f(x, d_k(z_{1:k}))$$

$$p^* = \arg\min_{p} \mathbb{E}_{p_{\text{data}}(x)} D_{\text{KL}}(q(z_{1:T} \mid x) \parallel p(z_{1:T}))$$

- Rate $\{\gamma_1, \ldots, \gamma_T\}$

$$\gamma_k = \mathbb{E}_{p_{\text{data}}(x)} D_{\text{KL}}(q(z_{1:k} \mid x) \parallel p^*(z_{1:k})).$$

- Parameter-sharing setting

$$p(z_k \mid z_{k-1}) = q(z_k \mid x = d_{k-1}(z_{k-1})).$$

- The progressive coding objective

$$\min_{d_{1:T}} \sum_{k=1}^{T} \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q(z_k|x)} f(x, d_k(z_k))$$

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# HAVE are not good progressively lossier compressors

- Rate

$$R_K = \mathbb{E}_{p_{\mathrm{data}}(x)} D_{\mathrm{KL}}(q(z_{1:K} \mid x) \parallel p(z_{1:K}))$$

- Distortion

$$D_K = \min_d \mathbb{E}_{p_{\mathrm{data}}(x)} \mathbb{E}_{q(z_{1:K}|x)} \|x - d(z_{1:K})\|^2$$

- $\gamma -$ optimal using a Lagrange multiplier

$$\lambda^* = \min_{d,p,q} \mathbb{E}_{p_{\mathrm{data}}(x)} \mathbb{E}_{q(z_{1:K}|x)} \|x - d(z_{1:K})\|^2$$
$$\mathrm{s.\,t.}\ \mathbb{E}_{p_{\mathrm{data}}(x)} D_{\mathrm{KL}}(q(z_{1:K} \mid x) \parallel p(z_{1:K})) \le R_K$$

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# HAVE are not good progressively lossier compressors

- Rate

| Dataset | Conventional Training | | $\gamma$-Optimal Training | |
|---|---|---|---|---|
| | $R_K$ | $D_K$ | $R'_K$ | $\geq \lambda^*$ |
| **SVHN** | 0.0168 | 0.0200 | 0.0162 | **0.0063** |
| **CelebA** | 0.0142 | 0.0274 | 0.0135 | **0.0128** |
| **LSUN** | 0.0187 | 0.0611 | 0.0181 | **0.0339** |

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# Two-stage vae

To enforce pareto-optimality $(R_K, D_K)$

- First stage

$$\min_{d,p,q} \mathbb{E}_{p_{\text{data}}(x)} D_{\text{KL}}(q(z_{1:K} \mid x) \parallel p(z_{1:K}))$$
$$\text{s.t. } \mathbb{E}_{p_{\text{data}}(x)} \mathbb{E}_{q(z_{1:K}|x)} f(x, d(z_{1:K})) \leq \lambda$$

- Second stage

$$\min_{p',q'} \mathbb{E}_{p_{\text{data}}(x)} \Big[ \mathbb{E}_{q(z_{1:T})} \ln p(x \mid z_{1:T}) + D_{\text{KL}}(q(z_{1:T} \mid x) \parallel p(z_{1:T})) \Big],$$

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding."." ICML2022

# Experiments

*Table 2.* Quantitative evaluation of conventionally-trained versus our progressively-coded HVAEs. ELBOs are per-dimension (i.e., normalized by the data dimensionality).

| | Conventional Training | | Progressive Coding | |
|---|---|---|---|---|
| Dataset | ELBO | FID | ELBO | FID |
| SVHN | **−1.31** | 18.92 | −1.46 | **10.07** |
| CelebA | **−1.44** | 13.33 | −1.58 | **8.54** |
| LSUN | **−1.72** | 40.71 | −1.78 | **36.87** |

Conventional training

Progressive coding

[1] Shu, Rui, and Stefano Ermon. "Bit Prioritization in Variational Autoencoders via Progressive Coding.".“ ICML2022