Diffusion Bridges Vector Quantized Variational Autoencoders

Max Cohen[1,2] Guillaume Quispe[3] Sylvain Le Corff[1] Charles Ollion[3] Éric Moulines[3]

[1]Samovar, Télécom SudParis, Département CITI, Institut Polytechnique de Paris, Palaiseau, France.

[2]Accenta, Boulogne-Billancourt, France.

[3]Centre de Mathématiques Appliquées, École polytechnique, Institut Polytechnique de Paris, Palaiseau, France

Thirty-ninth International Conference on Machine Learning

Discrete latent models

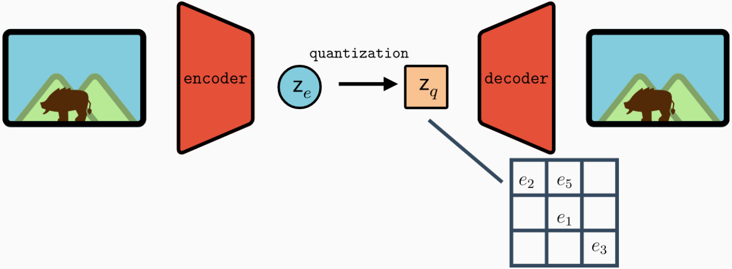


Figure: A discrete latent model.

Assume the distribution of the input $x \in \mathbb{R}^m$ depends on a hidden discrete state z_q, derived from a continuous state $z_e = f_\varphi(x)$, where in practice:

$$p_\theta(z_q = e_k | z_e) \propto \mathtt{softmax}(-\|z_e - e_k\|), \; z_q \in \mathcal{E} = \{e_1, \ldots, e_K\}$$

Discrete latent models

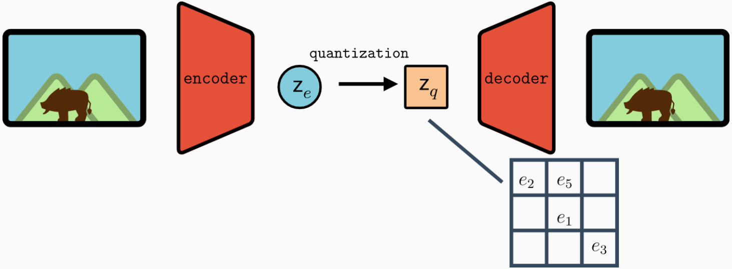


Figure: A discrete latent model.

Assume the distribution of the input $x \in \mathbb{R}^m$ depends on a hidden discrete state z_q, derived from a continuous state $z_e = f_\varphi(x)$, where in practice:

$$p_\theta(z_q = e_k | z_e) \propto \texttt{softmax}(-\|z_e - e_k\|), \; z_q \in \mathcal{E} = \{e_1, \ldots, e_K\}$$

Diffusion Bridge

We model z_q as the final sample of a chain $z_q^{0:T}$, associated with a Markov chain of continuous samples $z_e^{0:T}$.

- The initial distribution $p_\theta(z_e^T)$ is an uninformative prior;
- Each transition $p_\theta(z_e^t | z_e^{t+1})$ aims at producing a consistent sample through a Deep Neural Network.

$$p_\theta(z_q^{0:T}, z_e^{0:T}) = p_\theta(z_e^T) p_\theta(z_q^T | z_e^T) \prod_{t=0}^{T-1} p_\theta(z_e^t | z_e^{t+1}) p_\theta(z_q^t | z_e^t)$$

Architecture

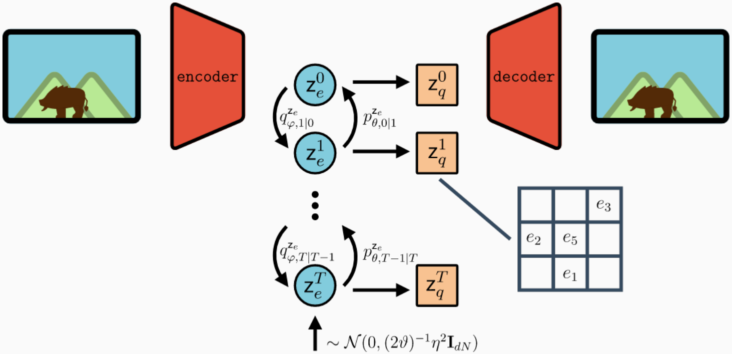


Figure: Our proposed architecture, for a prior based on a Ornstein-Uhlenbeck bridge. The top pathway from *input image* to z_e^0, to z_q^0, to *reconstructed image* resembles the original VQ-VAE model. The vertical pathway from (z_e^0, z_q^0) to (z_e^T, z_q^T) and backwards is based on a denoising diffusion process.

ELBO

We approximate the posterior using Variational Inference, by optimizing the following Evidence LOwer Bound:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}, z_e^{0:T}, x)}{q_\varphi(z_q^{0:T}, z_e^{0:T} | x)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\varphi}\left[\log p_\theta(x|z_q^0)\right]}_{\mathcal{L}^{rec}\text{ Reconstruction cost for the VQ-VAE architecture.}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}|z_e^{0:T})}{q_\varphi(z_q^{0:T}|z_e^{0:T})}\right]}_{\mathcal{L}^{reg}\text{ Generalization of the commitment cost, proportional to } -\|z_e - e_*\|.} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^{0:T})}{q_\varphi(z_e^{0:T})}\right]}_{\mathcal{L}^{prior}\text{ Distance between the corrupting and denoising models.}}$$

The choice of diffusion bridge appears in the last term, over all time steps: $\mathcal{L}^{prior} = \sum_{t=1}^{T} \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^t|z_e^{t+1})}{q_\varphi(z_e^t|z_e^0, z_e^{t+1})}\right]$

ELBO

We approximate the posterior using Variational Inference, by optimizing the following Evidence LOwer Bound:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}, z_e^{0:T}, x)}{q_\varphi(z_q^{0:T}, z_e^{0:T}|x)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\varphi}\left[\log p_\theta(x|z_q^0)\right]}_{\mathcal{L}^{rec}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}|z_e^{0:T})}{q_\varphi(z_q^{0:T}|z_e^{0:T})}\right]}_{\mathcal{L}^{reg}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^{0:T})}{q_\varphi(z_e^{0:T})}\right]}_{\mathcal{L}^{prior}}$$

$\mathcal{L}^{rec}$ Reconstruction cost for the VQ-VAE architecture.

$\mathcal{L}^{reg}$ Generalization of the commitment cost, proportional to $-\|z_e - e_*\|$.

$\mathcal{L}^{prior}$ Distance between the corrupting and denoising models.

The choice of diffusion bridge appears in the last term, over all time steps: $\mathcal{L}^{prior} = \sum_{t=1}^{T} \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^t|z_e^{t+1})}{q_\varphi(z_e^t|z_e^0, z_e^{t+1})}\right]$

ELBO

We approximate the posterior using Variational Inference, by optimizing the following Evidence LOwer Bound:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}, z_e^{0:T}, x)}{q_\varphi(z_q^{0:T}, z_e^{0:T}|x)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\varphi}\left[\log p_\theta(x|z_q^0)\right]}_{} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}|z_e^{0:T})}{q_\varphi(z_q^{0:T}|z_e^{0:T})}\right]}_{} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^{0:T})}{q_\varphi(z_e^{0:T})}\right]}_{}$$

$\mathcal{L}^{rec}$ Reconstruction cost for the VQ-VAE architecture.

$\mathcal{L}^{reg}$ Generalization of the commitment cost, proportional to $-\|z_e - e_*\|$.

$\mathcal{L}^{prior}$ Distance between the corrupting and denoising models.

The choice of diffusion bridge appears in the last term, over all time steps: $\mathcal{L}^{prior} = \sum_{t=1}^{T} \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^t|z_e^{t+1})}{q_\varphi(z_e^t|z_e^0, z_e^{t+1})}\right]$

ELBO

We approximate the posterior using Variational Inference, by optimizing the following Evidence LOwer Bound:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}, z_e^{0:T}, x)}{q_\varphi(z_q^{0:T}, z_e^{0:T}|x)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\varphi}\left[\log p_\theta(x|z_q^0)\right]}_{\mathcal{L}^{rec}\text{ Reconstruction cost for the VQ-VAE architecture.}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}|z_e^{0:T})}{q_\varphi(z_q^{0:T}|z_e^{0:T})}\right]}_{\mathcal{L}^{reg}\text{ Generalization of the commitment cost, proportional to } -\|z_e - e_*\|.} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^{0:T})}{q_\varphi(z_e^{0:T})}\right]}_{\mathcal{L}^{prior}\text{ Distance between the corrupting and denoising models.}}$$

The choice of diffusion bridge appears in the last term, over all time steps: $\mathcal{L}^{prior} = \sum_{t=1}^{T} \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^t|z_e^{t+1})}{q_\varphi(z_e^t|z_e^0, z_e^{t+1})}\right]$

ELBO

We approximate the posterior using Variational Inference, by optimizing the following Evidence LOwer Bound:

$$\mathcal{L}(\theta, \varphi) = \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}, z_e^{0:T}, x)}{q_\varphi(z_q^{0:T}, z_e^{0:T} | x)}\right]$$

$$= \underbrace{\mathbb{E}_{q_\varphi}\left[\log p_\theta(x|z_q^0)\right]}_{\mathcal{L}^{rec}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_q^{0:T}|z_e^{0:T})}{q_\varphi(z_q^{0:T}|z_e^{0:T})}\right]}_{\mathcal{L}^{reg}} + \underbrace{\mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^{0:T})}{q_\varphi(z_e^{0:T})}\right]}_{\mathcal{L}^{prior}}$$

$\mathcal{L}^{rec}$ Reconstruction cost for the VQ-VAE architecture.

$\mathcal{L}^{reg}$ Generalization of the commitment cost, proportional to $-\|z_e - e_*\|$.

$\mathcal{L}^{prior}$ Distance between the corrupting and denoising models.

The choice of diffusion bridge appears in the last term, over all time steps: $\mathcal{L}^{prior} = \sum_{t=1}^{T} \mathbb{E}_{q_\varphi}\left[\log \frac{p_\theta(z_e^t|z_e^{t+1})}{q_\varphi(z_e^t|z_e^0, z_e^{t+1})}\right]$

Sampling

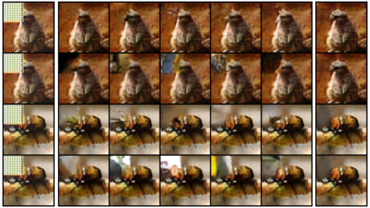
Algorithm Sampling procedure

Sample $\mathrm{z}_e^T \sim \mathcal{N}(0, (2\vartheta)^{-1}\eta^2 \mathbf{I}_{dN})$
for $t = T - 1$ **to** 0 **do**
    Sample $\mathrm{z}_e^t \sim p_\theta(\mathrm{z}_e^t | \mathrm{z}_e^{t+1})$ ▷ *diffusion bridge*
end for
Sample $\mathrm{z}_q^0 \sim p_\theta(\mathrm{z}_q^0 | \mathrm{z}_e^0)$ ▷ *quantization*
Sample $x \sim p_\theta(x | \mathrm{z}_q^0)$ ▷ *decoding*

Figure: Sampling denoising chain from $t = 500$ up to $t = 0$.

Inpainting

(a) Centered mask



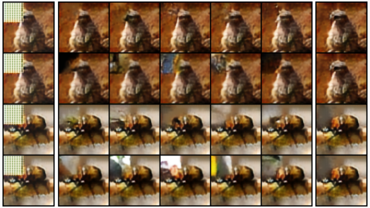
(b) Top-left mask

Figure: Conditional sampling for 2 different images, where samples from our diffusion are on top and from PixelCNN on the bottom. Each row contains independent conditional samples, with the original reconstruction on the right.

- We propose a new mathematical framework for quantized latent models.
- Our methodology focuses on VQVAE but allows sampling from any discrete law.
- To our best knowledge, this is the first probabilistic generative model to use denoising diffusion in discrete latent space.

Inpainting

(a) Centered mask



(b) Top-left mask

Figure: Conditional sampling for 2 different images, where samples from our diffusion are on top and from PixelCNN on the bottom. Each row contains independent conditional samples, with the original reconstruction on the right.

- We propose a new mathematical framework for quantized latent models.
- Our methodology focuses on VQVAE but allows sampling from any discrete law.
- To our best knowledge, this is the first probabilistic generative model to use denoising diffusion in discrete latent space.