Stochastically Quantized
Variational Auto-Encoders

•••

-Mayank Jain, Shikhar Agrawal
Prof. Suyash Awate

An Overview

- One noted issue of vector-quantized variational autoencoder (VQ-VAE) is that the learned discrete representation uses only a fraction of the full capacity of the codebook, also known as codebook collapse.
- We propose a new training scheme that extends the standard VAE via novel stochastic dequantization and quantization.
- In SQ-VAE, we observe a trend that the quantization is stochastic at the initial stage of the training but gradually converges toward a deterministic quantization.
- Our experiments show that SQ-VAE improves codebook utilization without using common heuristics.

Introduction

- In VQ-VAE, the encoded latent variables are quantized to their nearest neighbors in a learnable codebook, and the data samples are decoded from the quantized latent variables.
- Although VQ-VAE shares some similarities with VAE, its training does not follow the standard variational Bayes framework. Instead, it relies on carefully designed heuristics such as the use of a stop-gradient operator and the straight-through estimation of gradients.
- We propose a framework that combines stochastic quantization and VAE, called stochastically quantized VAE (SQ-VAE).
- It can address the low codebook utilization issue of VQ-VAE and can be explained within the scope of the usual variational Bayes framework.

Summarizing our approach

- SQ-VAE introduces a pair of stochastic dequantization and quantization processes in the latent space. These processes are characterized by probability distributions with trainable parameters.
- Optimizing the ELBO gradually reduces the stochasticity of the quantization process during the training, which we call self-annealing.
- In general, SQ-VAE does not impose any assumption on the data distribution; hence, we can model the stochastic quantization and dequantization processes via Gaussian distributions. (Yaay).

Background (VAE)

VAE Consider an observation $\mathbf{x} \in \mathbb{R}^D$ and a target data distribution $p_{\text{data}}(\mathbf{x})$, which models finite samples. The standard VAE consists of a stochastic encoder–decoder pair: a decoder $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and an approximated posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are trainable parameters. The latent variables $\mathbf{z} \in \mathbb{R}^{d_z}$ are assumed to follow a prior distribution $p(\mathbf{z})$. Data are generated by first sampling $\mathbf{z}$ from the prior $p(\mathbf{z})$ then obtaining $\mathbf{x}$ by feeding $\mathbf{z}$ into the stochastic decoder, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$. The negative ELBO per sample $\mathbf{x}$ is expressed as $\mathcal{L}_{\text{VAE}} =$

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right] + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \quad (1)$$

If the target data distribution is continuous, the stochastic decoder can be modeled by a Gaussian distribution with a mapping $f_{\boldsymbol{\theta}} : \mathbb{R}^{d_z} \to \mathbb{R}^D$ as

$$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{z}), \sigma^2\mathbf{I}), \quad (2)$$

Background (VQ-VAE)

VQ-VAE In contrast to VAE, VQ-VAE consists of a deterministic encoder–decoder path and a trainable *codebook*. The codebook is a set $\mathbf{B}$, which contains K d_b-dimensional vectors $\{\mathbf{b}_k\}_{k=1}^{K}$. A d_z-dimensional discrete latent space related to the codebook can be interpreted as the d_z-ary Cartesian power of $\mathbf{B}$, $\mathbf{B}^{d_z} \subset \mathbb{R}^{d_b \times d_z}$. We denote a latent variable in $\mathbf{B}^{d_z}$ and its ith column vector as $\mathbf{Z}_\mathrm{q} \in \mathbf{B}^{d_z}$ and $\mathbf{z}_{\mathrm{q},i} \in \mathbf{B}$, respectively. The deterministic encoding process from $\mathbf{x}$ to $\mathbf{Z}_\mathrm{q}$ includes a mapping $\hat{\mathbf{Z}}_\mathrm{q} = g_{\boldsymbol{\phi}}(\mathbf{x})$ with $g_{\boldsymbol{\phi}} : \mathbb{R}^D \to \mathbb{R}^{d_b \times d_z}$ and the quantization process of $\hat{\mathbf{Z}}_\mathrm{q}$ onto $\mathbf{B}^{d_z}$. The quan-

The objective function of VQ-VAE is

$$\mathcal{L}_{\mathrm{VQ}} = -\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_\mathrm{q}) + \|\mathrm{sg}[g_{\boldsymbol{\phi}}(\mathbf{x})] - \mathbf{Z}_\mathrm{q}\|_F^2 + \beta\|g_{\boldsymbol{\phi}}(\mathbf{x}) - \mathrm{sg}[\mathbf{Z}_\mathrm{q}]\|_F^2, \quad (4)$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator and β is set between 0.1 and 2.0 (van den Oord et al., 2017). To improve

The proposed process

As a generative model, the goal of SQ-VAE is to learn a generative process $\mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})$ with $\mathbf{Z}_{\mathrm{q}} \sim P(\mathbf{Z}_{\mathrm{q}})$ to generate samples that belong to the data distribution $p_{\mathrm{data}}(\mathbf{x})$, where $P(\mathbf{Z}_{\mathrm{q}})$ denotes the prior distribution of the discrete latent space $\mathbf{B}^{d_z}$. The prior $P(\mathbf{Z}_{\mathrm{q}})$ is assumed to be an i.i.d. uniform distribution in the main training stage as in VQ-VAE, i.e., $P(\mathbf{z}_{\mathrm{q},i} = \mathbf{b}_k) = 1/K$ for $k \in [K]$. A second training will take place to learn $P(\mathbf{Z}_{\mathrm{q}})$ after the main training stage. Since the exact evaluation of $p_{\boldsymbol{\theta}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{x})$ is intractable, the approximated posterior $q_{\boldsymbol{\phi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{x})$ is used instead.

In this setup, although we can establish the generative process following that in VQ-VAE, the construction of the encoding process from x to Zq is not straightforward owing to the discrete property of Zq. Therefore, we introduce two auxiliary variables to ease the explanation: $\mathbf{Z}$ and $\hat{\mathbf{Z}}_{\mathrm{q}}$

The proposed process (II)

variables to ease the explanation: $\mathbf{Z}$ and $\hat{\mathbf{Z}}_{\mathrm{q}}$. $\mathbf{Z}$ is the continuous variable converted from $\mathbf{Z}_{\mathrm{q}}$ via the dequantization process $p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})$, where $\boldsymbol{\varphi}$ indicates its parameters. Furthermore, we may derive the inverse process of $p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})$, i.e., the stochastic quantization process $\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})$, from Bayes' theorem $\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z}) \propto p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})P(\mathbf{Z}_{\mathrm{q}})$. On the other hand, $\hat{\mathbf{Z}}_{\mathrm{q}}$ is defined as $\hat{\mathbf{Z}}_{\mathrm{q}} = g_{\boldsymbol{\phi}}(\mathbf{x})$, which is the output of the deterministic encoder $g_{\boldsymbol{\phi}} : \mathbb{R}^D \to \mathbb{R}^{d_b \times d_z}$ given a sample $\mathbf{x}$. Ideally, $\hat{\mathbf{Z}}_{\mathrm{q}}$ should be close to $\mathbf{Z}_{\mathrm{q}}$. Similarly, the dequantization process of $\hat{\mathbf{Z}}_{\mathrm{q}}$ can be written as $\mathbf{Z}|\hat{\mathbf{Z}}_{\mathrm{q}} \sim p_{\boldsymbol{\varphi}}(\mathbf{Z}|\hat{\mathbf{Z}}_{\mathrm{q}})$. As in Figure 1, stacking the processes $p_{\boldsymbol{\varphi}}(\mathbf{Z}|\hat{\mathbf{Z}}_{\mathrm{q}})$ and $\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})$ connects $\hat{\mathbf{Z}}_{\mathrm{q}}$ and $\mathbf{Z}_{\mathrm{q}}$, and thus establishes the stochastic encoding process from $\mathbf{x}$ to $\mathbf{Z}_{\mathrm{q}}$ as $Q_{\boldsymbol{\omega}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{x}) := \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}[\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})]$, where $\boldsymbol{\omega} := \{\boldsymbol{\phi}, \boldsymbol{\varphi}\}$ and $q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x}) := p_{\boldsymbol{\varphi}}(\mathbf{Z}|g_{\boldsymbol{\phi}}(\mathbf{x}))$.

Return of the ELBO

- At this point, we can derive the ELBO for SQ-VAE:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq -\mathcal{L}_{\mathrm{SQ}}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\omega},\mathbf{B}) :=$$

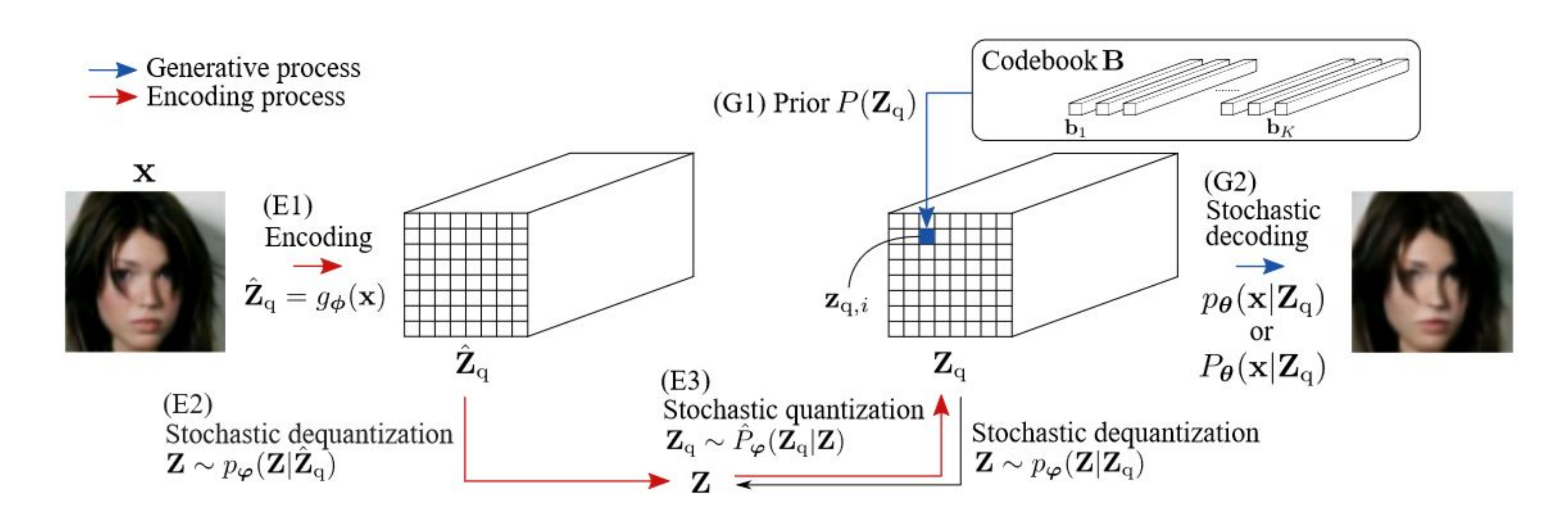
$$\mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})P(\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\right]$$

$$= \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}\right]$$

$$+ \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})} H(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})) + \mathrm{const.},$$

et al., 2020). The expectation in the first term of (5) involves the categorical distribution $\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})$, which can be approximated by the Gumbel–softmax relaxation (Jang et al., 2017; Maddison et al., 2017) to use the reparameterization trick in the backward pass of conventional VAE.

A thousand words...



Gaussian SQ-VAE (well... obviously)

We design Gaussian SQ-VAE by assuming that the dequantization process follows a Gaussian distribution. On the basis of the assumption, the dequantization process is modeled as

$$p_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{Z}_{\mathrm{q}}) = \mathcal{N}(\mathbf{z}_{\mathrm{q},i}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}}), \tag{6}$$

Decoding The usual Gaussian setup is adopted in the decoding such that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}}) = \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{Z}_{\mathrm{q}}), \sigma^2\mathbf{I})$, where $\sigma^2 \in \mathbb{R}_+$ and $\boldsymbol{\theta}$ are trainable parameters.

Encoding The encoding follows the process depicted in Figure 1, and the dequantization process applied to $\hat{\mathbf{Z}}_{\mathrm{q}}$ is $p_{\boldsymbol{\varphi}}(\mathbf{z}_i|\hat{\mathbf{Z}}_{\mathrm{q}}) = \mathcal{N}(\hat{\mathbf{z}}_{\mathrm{q},i}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}})$.

Another Objective Function

Objective Function The substitution of the encoding and decoding processes above into (5) gives $\mathcal{L}_{\mathcal{N}\text{-SQ}} =$

$$\mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\frac{1}{2\sigma^2}\|\mathbf{x}-f_{\boldsymbol{\theta}}(\mathbf{Z})\|_2^2+\mathcal{R}_{\boldsymbol{\varphi}}^{\mathcal{N}}(\mathbf{Z},\mathbf{Z}_{\mathrm{q}})\right]$$
$$-\mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}H\left(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})\right)+\frac{D}{2}\log\sigma^2+\mathrm{const.}, \quad (8)$$

where $\mathcal{R}_{\boldsymbol{\varphi}}^{\mathcal{N}}(\mathbf{Z},\mathbf{Z}_{\mathrm{q}})$ denotes the regularization objective in Table 1, depending on the parameterization of $\boldsymbol{\Sigma}_{\boldsymbol{\varphi}}$. The derivation detail can be found in Appendix B.1.

$\Sigma\varphi$ controls the degree of stochasticity of the quantization during the training. We first consider two extreme cases, $\sigma \to \infty$ and $\sigma \to 0$.

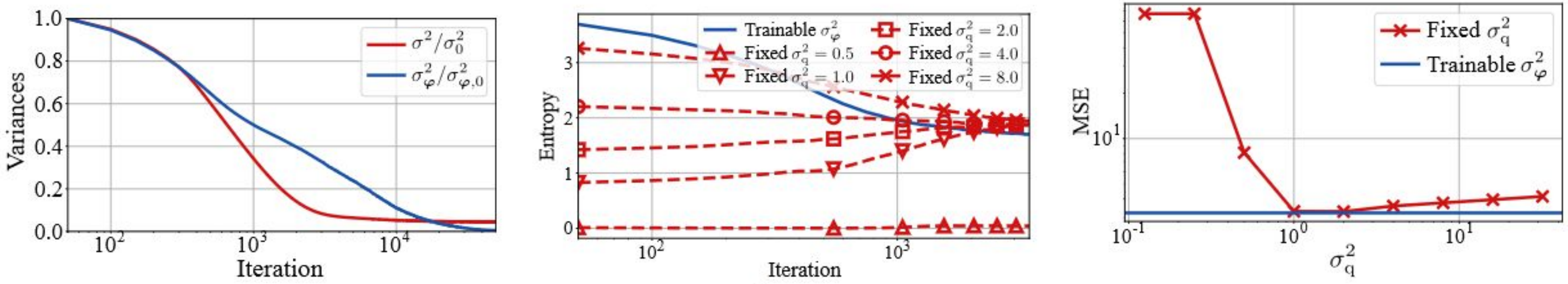
Self-annealing (the proposition)

Proposition 1. *Assume that $p_{\text{data}}(\mathbf{x})$ has finite support, whereas $g_{\boldsymbol{\phi}}$ and $\{\mathbf{b}_k\}_{k=1}^{K}$ are bounded. Let $\boldsymbol{\omega}^* = \{\boldsymbol{\phi}^*, \boldsymbol{\varphi}^*\}$ be a minimizer of $\mathbb{E}_{p_{\text{data}}(\mathbf{x})} D_{\text{KL}}(Q_{\boldsymbol{\omega}}(\mathbf{Z}_{\text{q}}|\mathbf{x}) \parallel P_{\boldsymbol{\theta}}(\mathbf{Z}_{\text{q}}|\mathbf{x}))$ with fixed $\boldsymbol{\theta}$, σ^2 and $\{\mathbf{b}_k\}_{k=1}^{K}$. If $\sigma^2 \to 0$, then $\sigma^2_{\boldsymbol{\varphi}^*} \to 0$.*

When σ → ∞, the first term in (8) diminishes. It is minimized when σφ → ∞.

means that $P_{\boldsymbol{\varphi}}(\mathbf{z}_{\text{q},i} = \mathbf{b}_k|\mathbf{Z})$ converges to the Kronecker delta function $\delta_{k,\hat{k}}$, where $\hat{k} = \arg\min_k \|\mathbf{z}_i - \mathbf{b}_k\|_2$. This deterministic quantization is exactly the posterior categorical distribution of VQ-VAE. According to the two cases above, if σ^2 decreases gradually during the training, the quantization process will also gradually decrease its stochasticity and

Three thousand words...



(a) Variance parameters (b) $H(\hat{P}_\varphi(\mathbf{z}_{q,i}|\mathbf{Z}))$ (c) MSE

Figure 2. Empirical study on the dynamics related to σ_φ^2 in Section 3.3. (a) The variance parameter σ_φ^2 (blue) decreased with σ^2 (red), where σ_0^2 and $\sigma_{\varphi,0}^2$ are their initial values. (b) Average entropy of the quantization process w.r.t. the iteration, which is obtained by Monte Carlo estimation. (c) MSE for trainable σ_φ^2 and various values of σ_q^2 on the test set.