codebook collapse를 피하기 위한 stochastic quantization.

* Introduction.

VQ 방식에는 codebook collapse 문제가 대두되는데,

이를 해결하기 위해 EMA, codebook reset, tuning 이 사용된다.

본문에서는 알고리즘의 문제라 보고, stochastic quantization을 제시.

↳ 일반적인 VAE framework 내에서 훈련 가능

학습 중의 quantization process의 stochasticity → self-annealing

→ CE의 봉우 성능이 안좋아서 Gaussian 대신 vMF(von Mises-Fisher) 사용

Contribution.

1. SQ-VAE 제시

2. annealing으로 codebook usage ↑.

3. 그에의 SQ-VAE 설계. (Gaussian, vMF)

4. 성능 향상 및 평가.

* Background.

VAE와 VQVAE는 모두 reconstruction loss와 latent penalty로 구성된다고 말할 수 있다.

*Stochastically quantized VAE

SQ-VAE는 VAE와 VQVAE의 훈련을 연결합니다.

↳ hueristic tech와 hyperparam의 무게를 줄여준다.

훈련 중에 점점 quantization에 근접함.

-Overview of SQ-VAE

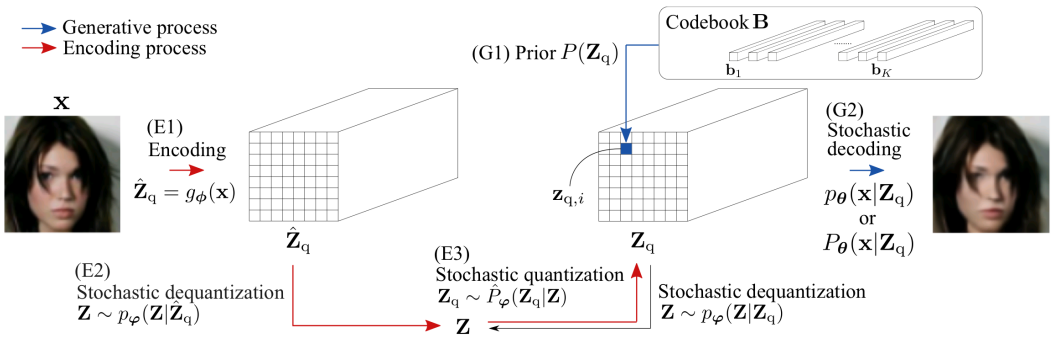


Figure 1. Encoding and generative processes of SQ-VAE. The encoding path from $\mathbf{x}$ to $\mathbf{Z}_q$ consists of (E1) deterministic encoding, (E2) stochastic dequantization, and (E3) quantization processes. For generation, in (G1) we first sample $\mathbf{Z}_q \in \mathbf{B}^{d_z}$ from the prior $p(\mathbf{Z}_q)$. Then, in (G2) we feed $\mathbf{Z}_q$ into the stochastic decoder to generate data samples.

Codebook을 B라고 할 때, $B := \{b_k\}_{k=1}^K$

θ에 대한 G는 $x \sim P(x|z_q)$, $x \sim P_{data}$, $z_q \sim P(z_q)$

$P(z_q)$는 uniform i.i.d. (확률이 다 1/k)

Z는 z_q로부터 dequantization 된 것이고,

↳ $P_\varphi(z|z_q)$, quantization : $\hat{P}_\varphi(z_q|z)$, $z|\hat{z}_q \sim P_\varphi(z|\hat{z}_q)$

↳ $\hat{P}_\varphi(z_q|z) \propto P_\varphi(z|z_q)P(z_q)$

$\hat{z}_q = g_\phi(x)$, $g_\phi : R^D \to R^{d_b \times d_z}$ (code의 개수(latent), codebook depth)

$\hat{z}_q \leftrightarrow z_q$ 사이유지.

$P_\varphi(z|z_q)$와 $\hat{P}_\varphi(z_q|z)$를 쌓아, $\hat{z}_q$와 z_q를 연결

$Q_w(z_q|x) := E_{q_w(z|x)}[\hat{P}_\varphi(z_q|z)]$; $w = \{\phi, \varphi\}$, $q_w(z|x) := P_\varphi(z|g_\phi(x))$ 로, x부터 z_q까지 stochastic encoding.

SQVAE 로부터의 ELBO는

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

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

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

$$+ \mathbb{E}_{q_\omega(\mathbf{Z}|\mathbf{x})} H(\hat{P}_\varphi(\mathbf{Z}_\mathrm{q}|\mathbf{Z})) + \text{const.}, \qquad (5)$$

→ uniform 이므로 constant.

→ $H(P)$는 P의 entropy

↳ E, G, B가 동시에 optimizing

- Gaussian SQ-VAE

Gaussian SQ-VAE는 dequantization을 gaussian으로 가정.

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

↳ trainable

Inverse 이 해당하는 quantization은 ($\mathbf{z}_q$ 복구)

$$\hat{P}_\varphi(\mathbf{z}_{q,i}=\mathbf{b}_k|\mathbf{z}) = \mathrm{softmax}_k\left(\left\{-\frac{(\mathbf{b}_j-\mathbf{z}_i)^T\boldsymbol{\Sigma}_\varphi^{-1}(\mathbf{b}_j-\mathbf{z}_i)}{2}\right\}_{j=1}^K\right)$$

↳ variance Σ_φ 와 $\mathbf{z}_i$ 로부터의 $\mathbf{b}_j$의 Mahalanobis 거리의 unnormalized log-probability

Table 1. Different parameterizations of the variance $\boldsymbol{\Sigma}_\varphi$ in Gaussian SQ-VAE.

	Variance $\boldsymbol{\Sigma}_\varphi$	Unnormalized log-probability	Regularization objective $\mathcal{R}_\varphi^{\mathcal{N}}(\mathbf{Z},\mathbf{Z}_\mathrm{q})$
(I)	$\sigma_\varphi^2\mathbf{I}$	$\|\mathbf{b}_k-\mathbf{z}_i(\mathbf{x})\|_2^2/2\sigma_\varphi^2$	$\|\mathbf{Z}-\mathbf{Z}_\mathrm{q}\|_F^2/2\sigma_\varphi^2$
(II)	$\sigma_\varphi^2(\mathbf{x})\mathbf{I}$	$\|\mathbf{b}_k-\mathbf{z}_i(\mathbf{x})\|_2^2/2\sigma_\varphi^2(\mathbf{x})$	$\|\mathbf{Z}-\mathbf{Z}_\mathrm{q}\|_F^2/2\sigma_\varphi^2(\mathbf{x})$
(III)	$\sigma_{\varphi,i}^2(\mathbf{x})\mathbf{I}$	$\|\mathbf{b}_k-\mathbf{z}_i(\mathbf{x})\|_2^2/2\sigma_{\varphi,i}^2(\mathbf{x})$	$\sum_{i=1}^{d_z}\|\mathbf{z}_i(\mathbf{x})-\mathbf{z}_{\mathrm{q},i}\|_2^2/2\sigma_{\varphi,i}^2(\mathbf{x})$
(IV)	$\mathrm{diag}(\boldsymbol{\sigma}_{\varphi,i}^2(\mathbf{x}))$	$\sum_{j=1}^{d_b}(b_{k,j}-z_{i,j}(\mathbf{x}))^2/2\sigma_{\varphi,i,j}^2(\mathbf{x})$	$\sum_{i=1}^{d_z}\sum_{j=1}^{d_b}(z_{i,j}(\mathbf{x})-z_{\mathrm{q},i,j})^2/2\sigma_{\varphi,i,j}^2(\mathbf{x})$

Decoding : $P_\theta(\mathbf{x}|\mathbf{z}_q) = \mathcal{N}(f_\theta(\mathbf{z}_q), \sigma^2\mathbf{I})$, $\sigma^2 \in \mathbb{R}_+$ 와 θ는 trainable.

Encoding : $p_\varphi(\mathbf{z}_i|\hat{\mathbf{Z}}_\mathrm{q}) = \mathcal{N}(\hat{\mathbf{z}}_{\mathrm{q},i}, \boldsymbol{\Sigma}_\varphi)$.

Objective function.

eq.5를 변형해보면.

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

→ table 1의 regularization objective.
↳ Appendix B.1

- Self-annealed Qauntization.

Table 1의 type I을 가정함

Σ_φ는 훈련 과정에서 qauntization의 정도를 조절한다.

Appendix D.2에서 $\sigma^2 \to \infty$ 와 $\sigma^2 \to 0$의 그때의 extreme case를 설명.

Proposition 1

$P_{data}(x)$가 finite하지만, g_θ와 $\{b_k\}_{k=1}^K$는 bounded라고 가정.

$w^* = \{\theta^*, \varphi^*\}$가 fixed $\Theta, \sigma^2, \{b_k\}_{k=1}^K$에서, $E_{P_{data}(x)} D_{KL}(Q_w(z_q|x) \| P_\theta(z_q|x))$의 minimizer라고 하면.

$\sigma^2 \to 0$ 이면, $\sigma_{\varphi^*}^2 \to 0$

$\sigma^2 \to \infty$ 일 때, eq.8의 첫번째 term은 사라짐. → $P_\varphi(\mathbf{z}_{\mathrm{q},i} = \mathbf{b}_k|\mathbf{Z})$가 uniform이 되기 때문.

$\sigma^2 \to 0$ 이면, kronecker delta δ로 수렴 $\delta_{k,\hat{k}}$. $\hat{k} = \arg\min_k \|z_i - b_k\|_2$

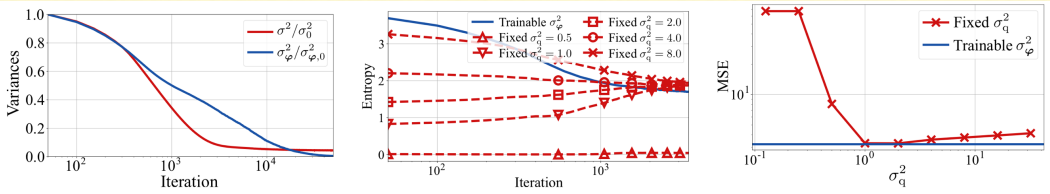
↳ quantization이 VQVAE의 posterior와 같다.

→ σ^2가 감소하면, quantization process도 확률을 감소시키고, 엔트로피가 줄어듦 → self-annealing.

Dynamics of the variance parameter

self-annealing을 verify하기 위해, $\Sigma_p = \sigma_p^2 I$의 gaussian SQ-VAE라고 하면.

σ_q^2를 설계하기 위해 σ_p^2를 고정. → Appendix D.3.



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

- vMF SQ-VAE for Categorical Distributions.

Categorical classification으로 생각. (softmax 사용)

CE에 대해서 ELBO는.

$$\mathcal{L}_{\text{CE-SQ}}^{\text{naïve}} = \mathbb{E}_{q_{\omega}(\mathbf{Z}|\mathbf{x})\hat{P}_{\varphi}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[-\sum_{d=1}^{D}\log(P_{\boldsymbol{\theta}}(x_d = c|\mathbf{Z}_{\mathrm{q}})) + \mathcal{R}_{\varphi}^{\mathcal{N}}(\mathbf{Z}, \mathbf{Z}_{\mathrm{q}})\right] - \mathbb{E}_{q_{\omega}(\mathbf{Z}|\mathbf{x})}H\left(\hat{P}_{\varphi}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})\right) + \text{const.}$$

with (9a)

$$P_{\boldsymbol{\theta}}(x_d = c|\mathbf{Z}_{\mathrm{q}}) = \text{softmax}_c\left(\{\mathbf{w}_{\text{last},c'}^{\top}\tilde{f}_{\boldsymbol{\theta}^-,d}^{\text{rest}}(\mathbf{Z}_{\mathrm{q}})\}_{c'=1}^{C_{\text{all}}}\right). \quad (9b)$$

$\tilde{f}_{\theta^-,d}^{\text{rest}}(Z_q) = \mathbb{B}^{d_z} \to \mathbb{R}^F$

위의 방식, eq.8과 다르게 σ^2에 대한 것이 없이 때문에, self-annealing의 효과를 볼 수 없다.

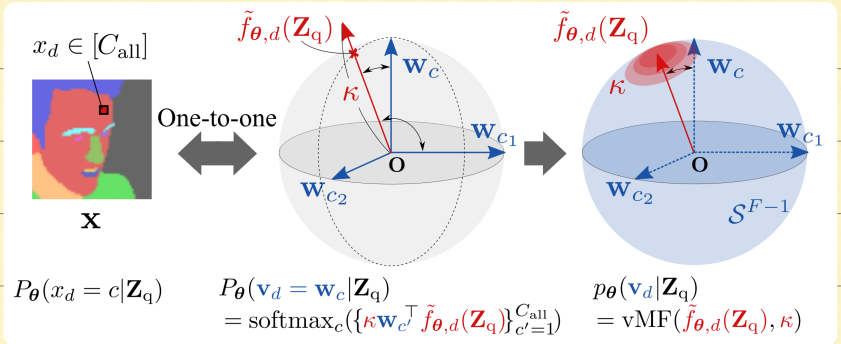
Self-annealing을 위해, vMF 분포와 model을 refine.

F-dim에서 embedded된 hyper-sphere S^{F-1}을 가정.

W_c는 S^{F-1}의 surface에서, C번째 data categorical의 projected vector로 명시한다.

projected data x_d의 hypersphere에서의 표현은 $V_d \in \{W_c\}_{c=1}^{C_{all}}$

$x_d = c$라면, $V_d = W_c$, $V_d = W_c$면 $x_d = c$



Decoding

$$P_{\boldsymbol{\theta}}(\mathbf{v}_d = \mathbf{w}_c|\mathbf{Z}_{\text{q}}) = \text{softmax}_c\left(\left\{\kappa \mathbf{w}_{c'}^\top \tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\text{q}})\right\}_{c'=1}^{C_{\text{all}}}\right), \tag{10}$$

$$p_{\boldsymbol{\theta}}(\mathbf{v}_d|\mathbf{Z}_{\text{q}}) \propto \exp\left(\kappa \mathbf{v}_d^\top \tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\text{q}})\right). \tag{11}$$

$$p_{\boldsymbol{\theta}}(\mathbf{v}_d|\mathbf{Z}_{\text{q}}) = \text{vMF}(\tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\text{q}}), \kappa)$$

Encoding

VMF를 사용하여, dequantization을 modeling.

$$p_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{Z}_{\text{q}}) = \text{vMF}(\mathbf{z}_{\text{q},i}, \kappa_{\boldsymbol{\varphi}}), \tag{12}$$

$\kappa_{\boldsymbol{\varphi}}$ → trainable concentration parameter.

$$\hat{P}_{\boldsymbol{\varphi}}(\mathbf{z}_{\text{q},i} = \mathbf{b}_k|\mathbf{Z}) = \text{softmax}_k\left(\{\kappa_{\boldsymbol{\varphi}} \mathbf{b}_j^\top \mathbf{z}_i\}_{j=1}^K\right), \tag{13}$$

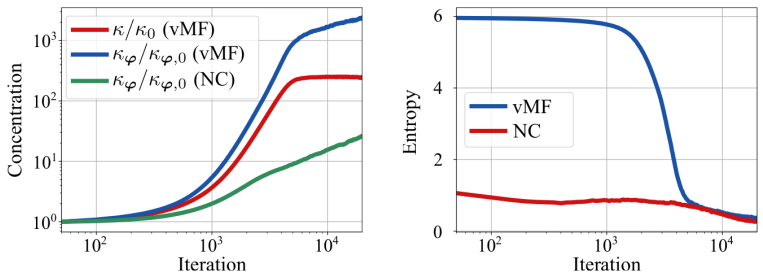
Objective function.

$$\mathcal{L}_{\text{VMF-SQ}} = \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\text{q}}|\mathbf{Z})}\left[-\kappa \sum_{d=1}^{D} \mathbf{v}_d^\top \tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\text{q}}) + \mathcal{R}_{\boldsymbol{\varphi}}^{\text{vMF}}(\mathbf{Z}, \mathbf{Z}_{\text{q}})\right] - \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{V})} H\left(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\text{q}}|\mathbf{Z})\right) - \log C_F(\kappa) + \text{const.}, \tag{14}$$

$$\mathcal{R}_{\boldsymbol{\varphi}}^{\text{vMF}}(\mathbf{Z}, \mathbf{Z}_{\text{q}}) = \sum_{i=1}^{d_z} \kappa_{\boldsymbol{\varphi},i}(1 - \mathbf{z}_{\text{q},i}^\top \mathbf{z}_i)$$

Comparing vMF SQ-VAE with Naive Categorical SQ-VAE

SQ-VAE는 self-annealing 가능



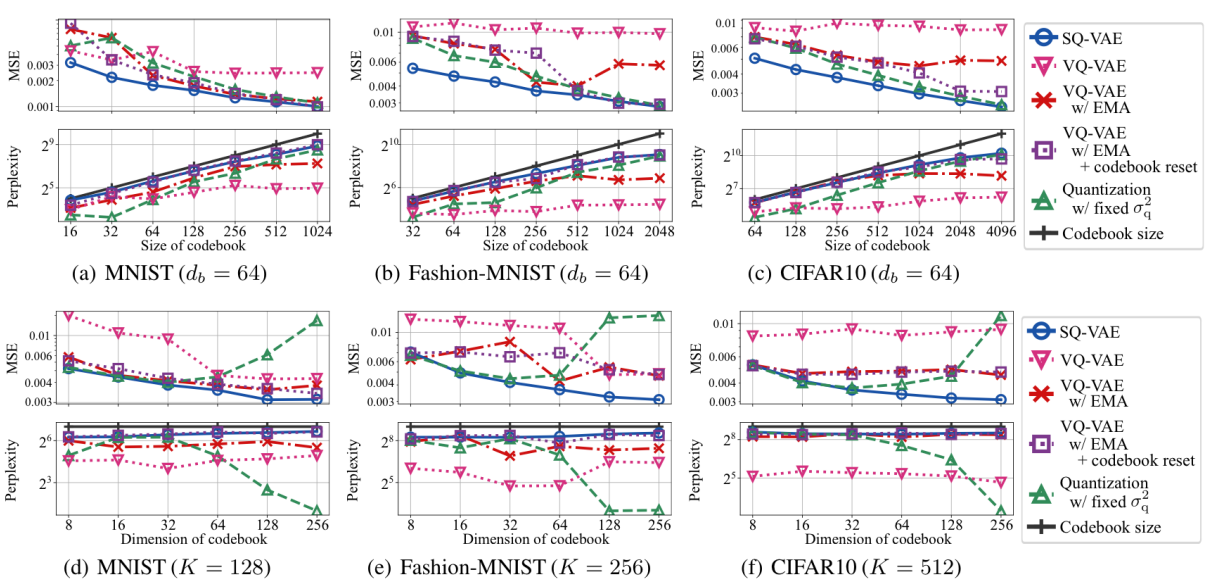
(a) Concentration parameters (b) $H(\hat{P}_{\varphi}(\mathbf{z}_{\mathrm{q},i}|\mathbf{Z}))$

Figure 4. Comparison between vMF and NC decoders: (a) The concentration parameter of vMF decoder κ_φ increases with κ, whereas the growth of κ_φ of the NC decoder is relatively small. Here, κ_0 and $\kappa_{\varphi,0}$ indicate initial values. (b) Average entropy of probabilities of quantization processes.

* Experiments

Continuous data Distribution.

Vision



(a) MNIST ($d_b = 64$) (b) Fashion-MNIST ($d_b = 64$) (c) CIFAR10 ($d_b = 64$)

(d) MNIST ($K = 128$) (e) Fashion-MNIST ($K = 256$) (f) CIFAR10 ($K = 512$)

Figure 5. Empirical studies on the impact of codebook capacity examined on MNIST Fashion-MNIST and CIFAR10. (a)–(c) The size K is swept with the dimension d_b fixed to 64. (d)–(f) Various d_b values are tested with the size K fixed as 128, 256, and 512, respectively. The black lines with "+" marks indicate the upper bounds of the perplexities, i.e., K. All the y-axes are in log-scale.

Table 2. Evaluation on CelebA. The MSE ($\times 10^3$) and reconstructed FID (rFID) are evaluated using the test set. The codebook capacity for the discrete latent space is set to $(n_b, k) = (64, 512)$. The Roman numerals for Gaussian SQ-VAEs correspond to those in Table 1. We also show the FID of samples generated with a prior learned with PixelCNN.

Model	Reconstruction		Generation	Latent manipulation (FID)			
	MSE	rFID	(FID)	Neighbor-3	Neighbor-5	Neighbor-10	Interp.
VAE	4.79 ± 0.01	40.3 ± 0.3	–	–	–	–	–
VQ-VAE (EMA)	1.33 ± 0.41	18.5 ± 5.1	42.0 ± 11.5	31.9 ± 14.8	42.8 ± 20.7	70.7 ± 35.4	28.2 ± 6.4
VQ-VAE (EMA+code reset)	1.62 ± 0.36	22.0 ± 5.9	51.8 ± 10.8	39.7 ± 12.0	52.7 ± 14.7	83.2 ± 20.4	32.6 ± 7.1
Quantization w/ fixed σ_q^2	1.09 ± 0.01	15.9 ± 0.1	38.2 ± 0.9	20.0 ± 0.4	26.4 ± 0.8	41.5 ± 2.1	18.6 ± 0.3
Gaussian SQ-VAE (I)	$\mathbf{0.96} \pm 0.01$	14.8 ± 0.3	28.2 ± 0.9	17.8 ± 0.1	21.9 ± 0.1	$\mathbf{33.1} \pm 0.3$	$\mathbf{17.6} \pm 0.6$
Gaussian SQ-VAE (II)	0.98 ± 0.01	14.3 ± 0.2	$\mathbf{27.7} \pm 1.1$	17.8 ± 0.2	22.2 ± 0.4	34.0 ± 0.9	$\mathbf{17.6} \pm 0.1$
Gaussian SQ-VAE (III)	$\mathbf{0.96} \pm 0.00$	$\mathbf{13.9} \pm 0.1$	28.1 ± 0.3	$\mathbf{17.3} \pm 0.2$	$\mathbf{21.6} \pm 0.3$	33.5 ± 0.6	18.5 ± 0.4

Speech .

Table 3. Evaluation on VCTK and ZeroSpeech 2019. The MSE (dB^2) of sample reconstruction is evaluated using the test set. We do not apply SQ-VAE (II) in this evaluation because of the variable length property of speech data and the different manipulations of speech signals between training and inference (see Appendix E.2).

Model	MSE (dB^2)	
	VCTK	ZeroSpeech 2019
VQ-VAE w/ EMA	29.59 ± 0.25	34.33 ± 1.57
Gaussian SQ-VAE (I)	25.52 ± 0.08	33.17 ± 1.11
Gaussian SQ-VAE (III)	25.94 ± 0.22	34.35 ± 1.07
Gaussian SQ-VAE (IV)	$\mathbf{24.68} \pm 0.21$	$\mathbf{32.32} \pm 0.88$

Categorical Distribution.

Table 4. Evaluation on CelebA-Mask. The pixel error (%), mIoU, and perplexity are evaluated using the test set.

Model	Pixel error	mIoU	Perplexity
VAE	8.79 ± 0.01	55.8 ± 0.3	–
VQ-VAE w/ EMA	6.95 ± 0.14	59.7 ± 0.7	46.2 ± 2.0
NC SQ-VAE	6.63 ± 1.38	64.1 ± 5.4	12.6 ± 5.2
vMF SQ-VAE	$\mathbf{3.51} \pm 0.17$	$\mathbf{74.6} \pm 0.0$	$\mathbf{52.4} \pm 0.8$

Table 5. Evaluation on MNIST and gray-scaled CelebA. The MSE ($\times 10^3$) is evaluated using the test set.

Model	MNIST	Gray-CelebA
VAE	22.80 ± 0.32	17.73 ± 0.23
VQ-VAE w/ EMA	6.24 ± 0.18	5.19 ± 0.06
NC SQ-VAE	10.89 ± 0.47	3.88 ± 0.02
vMF SQ-VAE	$\mathbf{1.63} \pm 0.21$	$\mathbf{2.37} \pm 0.01$

Conclusion.

VAE와 VQVAE를 합쳐 dequantized와 quantized를 쌍으로 학습한다.