Stochastically Quantized

Variational Auto-Encoders
VMF-SQ-VAE

-Mayank Jain, Shikhar Agrawal
Prof. Suyash Awate

An Overview of VMF-SQ-VAE

- An intuitive way to adapting SQ-VAE for categorical data distribution is to model the decoder output as a categorical distribution
- Consider a typical classification scenario that the last layer of a decoder is a linear layer followed by a softmax.

max. The decoder can be represented as the combination of the linear layer $\mathbf{w}_{\text{last},c} \in \mathbb{R}^F$ and the rest $\tilde{f}^{\text{rest}}_{\boldsymbol{\theta}^-,d} : \mathbf{B}^{d_z} \to \mathbb{R}^F$. It becomes $f^c_{\boldsymbol{\theta},d}(\mathbf{Z}_\text{q}) = \mathbf{w}^\top_{\text{last},c} \tilde{f}^{\text{rest}}_{\boldsymbol{\theta}^-,d}(\mathbf{Z}_\text{q})$, where $\boldsymbol{\theta}^-$ denotes the trainable parameters excluding $\mathbf{w}_{\text{last},c}$. We may represent

Reminder of what an ELBO for SQ-VAE looks like!

$$
\begin{aligned}
&\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq -\mathcal{L}_{\mathrm{SQ}}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\omega}, \mathbf{B}) := \\
&\quad \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})} \left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}}) p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}}) P(\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x}) \hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})} \right] \\
&\quad = \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})} \left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}}) p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})} \right] \\
&\qquad + \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})} H(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})) + \mathrm{const.}, \qquad (5)
\end{aligned}
$$

- Where H(P) denotes the entropy of P. In (5), since P(Z_q) is assumed to follow a uniform distribution, it results into a constant term and is thus omitted.

ELBO of Naive-CE-SQ-VAE for Categorical Distributions

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

$$\left. + \mathcal{R}_{\boldsymbol{\varphi}}^{\mathcal{N}}(\mathbf{Z}, \mathbf{Z}_{\mathrm{q}}) \right] - \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})} H\left(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})\right) + \text{const.} \tag{9a}$$

with

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

Why not Naive SQ-VAE for Categorical Distributions?

- However, we found that the performance of this Naive categorical (NC) SQ-VAE is often unsatisfactory
- A possible cause can be found by observing the difference between the ELBO for Gaussian SQ-VAE and ELBO for NC SQ-VAE.
- In ELBO for NC SQ-VAE, owing to the replacement of Gaussian with categorical distribution, trainable parameters such as variance no longer exist in the objective function. This means that the model cannot be benefited from the self-annealing effect.
- To gain the advantage from self-annealing, we introduce the vMF distribution to refine the model, and we call it vMF SQ-VAE.

Mathematical Formulation of VMF-SQ-VAE

Consider a hypersphere $\mathcal{S}^{F-1}$ that is embedded in an F-dimensional space. Let $\mathbf{w}_c$ denote the projection vector[3] of the cth data category on the surface of

$\mathcal{S}^{F-1}$. Next, we represent the projection of data x_d on the hypersphere as $\mathbf{v}_d \in \{\mathbf{w}_c\}_{c=1}^{C_{\mathrm{all}}}$. If x_d belongs to a category c, that is, $x_d = c$, then $\mathbf{v}_d = \mathbf{w}_c$ and vice versa.

- For each element x_d of the input vector x(which belongs to one of the category in C_{all}), we have an associated vector v_d, which is same as the projection of the cth data category parameter onto the hypersphere.

How to Decode?

Decoding The first step is to decode $\mathbf{Z}_{\mathrm{q}}$ into $\mathbf{V} := \{\mathbf{v}_d\}_{d=1}^{D}$ with the decoder $\tilde{f}_{\boldsymbol{\theta},d} : \mathbf{B}^{d_z} \rightarrow \mathcal{S}^{F-1}$. Then, determine the probability of $\mathbf{v}_d = \mathbf{w}_c$ with a trainable scalar $\kappa \in \mathbb{R}_+$ by using

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

which resembles the categorical decoder in (9b) except for the normalization onto $\mathcal{S}^{F-1}$ and the scaling factor κ.

How to decode?

Therefore, we may represent the categorical probabilities for the decoded $\mathbf{Z}_{\mathrm{q}}$ as

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

By normalizing (11) w.r.t. $\mathbf{v}_d$ over $\mathcal{S}^{F-1}$, we obtain $p_{\boldsymbol{\theta}}(\mathbf{v}_d|\mathbf{Z}_{\mathrm{q}}) = \mathrm{vMF}(\tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\mathrm{q}}), \kappa)$, where $\tilde{f}_{\boldsymbol{\theta},d}(\mathbf{Z}_{\mathrm{q}})$ and κ correspond to the mean direction and the concentration parameter of the vMF distribution, respectively.

Von Mises Fisher Distribution says hii!

Again 3000 words

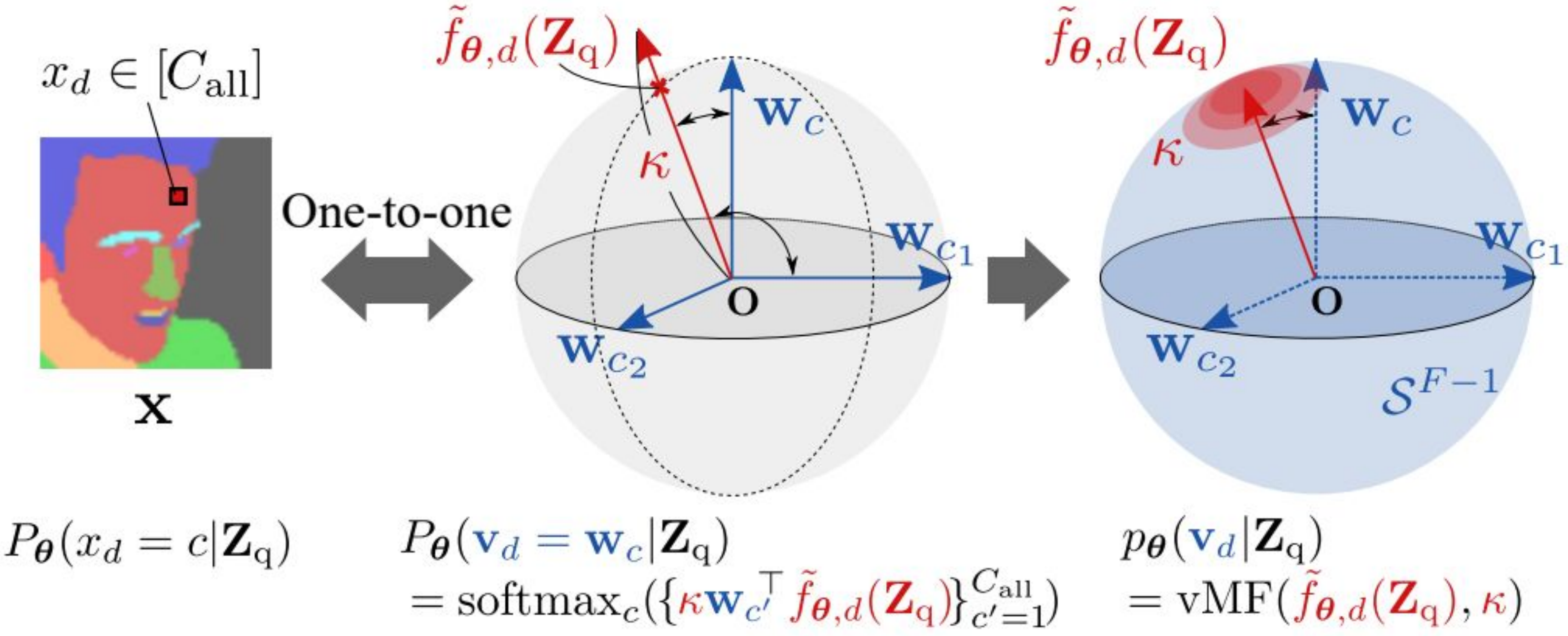

Figure 3. vMF decoder.

Encoding in VMF-SQ-VAE

Encoding Accordingly, we model the stochastic dequantization process of the encoder with the vMF distribution:

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

where $\kappa_{\boldsymbol{\varphi}}$ is the trainable concentration parameter[4]. Similarly to Gaussian SQ-VAE in Section 3.2, the discrete $\mathbf{Z}_{\mathrm{q}}$ is recovered using Bayes' theorem as

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

where the unnormalized log-probabilities of $\mathbf{b}_k$ in (13) correspond to the $\kappa_{\boldsymbol{\varphi}}$-scaled cosine similarity between $\mathbf{b}_k$ and $\mathbf{z}_i$.

Objective Function of VMF-SQ-VAE

Objective Function Substituting the encoding and decoding processes into (5) leads to $\mathcal{L}_{\text{vMF-SQ}} =$

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

where $\mathcal{R}_{\boldsymbol{\varphi}}^{\mathrm{vMF}}(\mathbf{x},\mathbf{Z}_{\mathrm{q}})$ is a regularization objective defined by $\mathcal{R}_{\boldsymbol{\varphi}}^{\mathrm{vMF}}(\mathbf{Z},\mathbf{Z}_{\mathrm{q}})=\sum_{i=1}^{d_z}\kappa_{\boldsymbol{\varphi},i}(1-\mathbf{z}_{\mathrm{q},i}^{\top}\mathbf{z}_i)$ (see Appendix B.2 for details). Here, $C_F(\kappa)$ denotes the normalizing constant of the vMF distribution (see Appendix A).

Comparing NC-VQ-VAE and VMF-VQ_VAE

- In (14), the first two terms are scaled with κ and $\kappa_{\boldsymbol{\varphi}}$. Furthermore, vMF SQ-VAE has a property that, if $\kappa \to \infty$, then $\kappa_{\boldsymbol{\varphi}^*} \to \infty$. Its proof can be done similarly to Proposition 1 via setting $\kappa = 1/\sigma^2$ and $\kappa_{\boldsymbol{\varphi}^*} = 1/\sigma^2_{\boldsymbol{\varphi}^*}$.
- As a result, vMF SQ-VAE can also achieve self-annealing as described in Section 3.3 if $\kappa \to \infty$.
- On the other hand, self-annealing is impossible for NC SQ-VAE owing to the lack of scaling parameters.

Finally, we are done! Or are we?

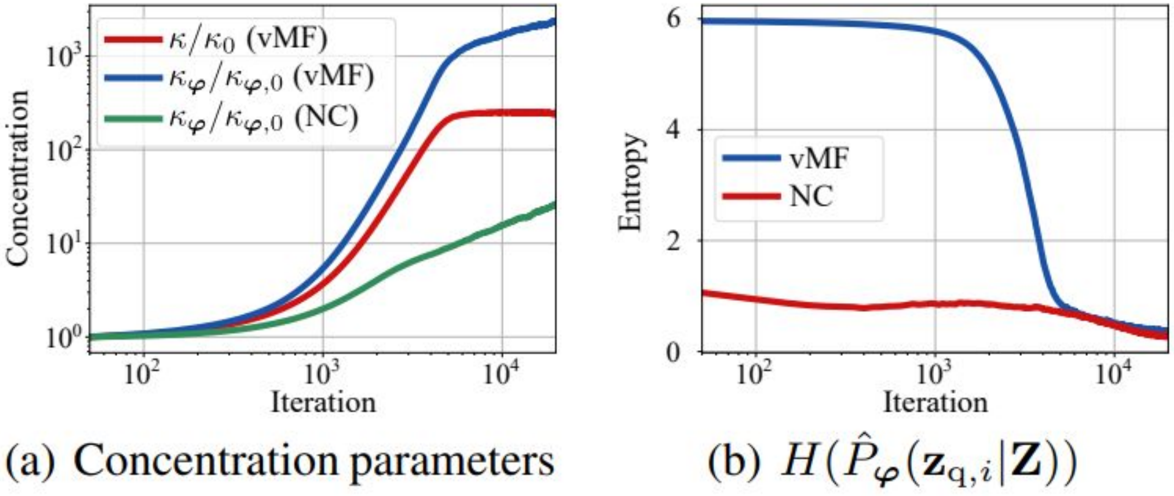

Figure 4. Comparison between vMF and NC decoders: (a) The concentration parameter of vMF decoder κ_φ increases with κ, whereas the growth of κ_φ of the NC decoder is relatively small. Here, κ_0 and $\kappa_{\varphi,0}$ indicate initial values. (b) Average entropy of probabilities of quantization processes.