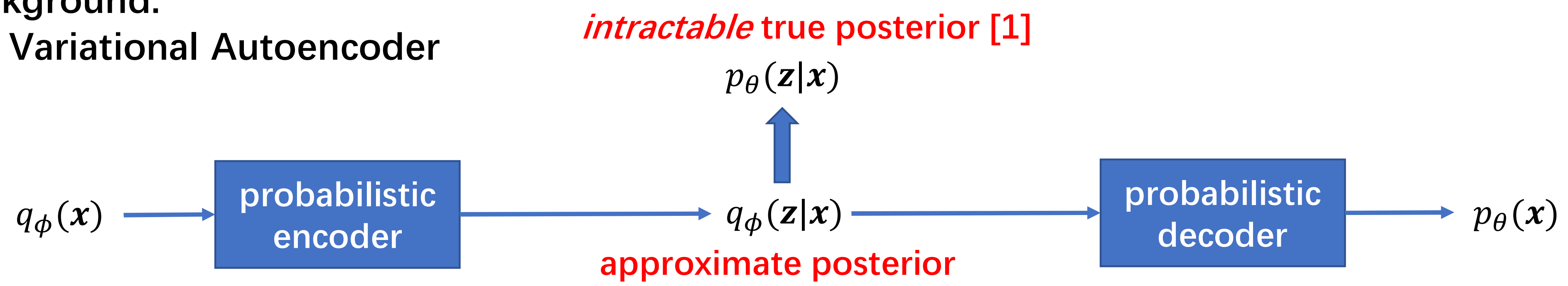
Improving Variational Autoencoders with Density Gap-based Regularization

Jianfei Zhang[1,2] **Jun Bai**[1,2] **Chenghua Lin**[3] **Yanmeng Wang**[4] **Wenge Rong**[1,2]

[1]State Key Laboratory of Software Development Environment, Beihang University, China
[2]School of Computer Science and Engineering, Beihang University, China
[3]Department of Computer Science, University of Sheffield, United Kingdom
[4]Ping An Technology, China
{zhangjf,ba1_jun,w.rong}@buaa.edu.cn
c.lin@sheffield.ac.uk, wangyanmeng219@pingan.com.cn

Catalogue

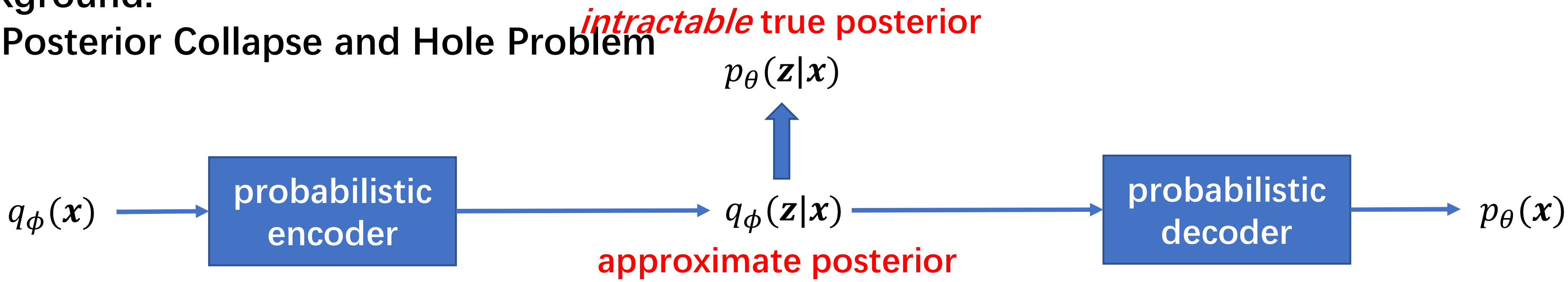
1. Background:
   a. Variational Autoencoder



$$
\begin{aligned}
\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) &= \underbrace{E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]}_{\text{conditional log-likelihood}} - \underbrace{D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))}_{\text{KL regularization}} \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) + \log p_\theta(\boldsymbol{z})] \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x})] \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{z}|\boldsymbol{x}) - \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})] \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) \\
&\leq E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] \\
&= \log p_\theta(\boldsymbol{x})
\end{aligned}
$$

where,
$q_\phi(\boldsymbol{x})$: the data distribution, described by the dataset and received by the encoder $\boldsymbol{\phi}$
$p_\theta(\boldsymbol{z})$: the prior distribution of latent variable $\boldsymbol{z}$ in decoder $\boldsymbol{\theta}$
$p_\theta(\boldsymbol{x})$: the generative data distribution by decoder $\boldsymbol{\theta}$ (or the generative likelihood)

1. Background:
   b. Posterior Collapse and Hole Problem



$$\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

Posterior Collapse:

$$\forall x\ D_{KL}(q_\phi(\boldsymbol{z}|x)||p_\theta(\boldsymbol{z})) \approx 0$$

➔ $\forall x\ p_\theta(\boldsymbol{z}|x) \approx q_\phi(\boldsymbol{z}|x) \approx p_\theta(\boldsymbol{z})$

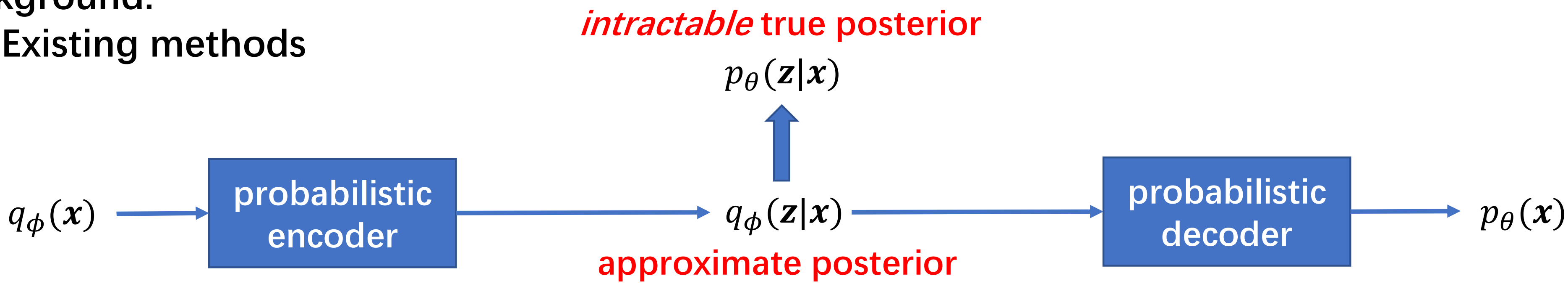
i.e., the latent variable $\boldsymbol{z}$ contains little information of $\boldsymbol{x}$

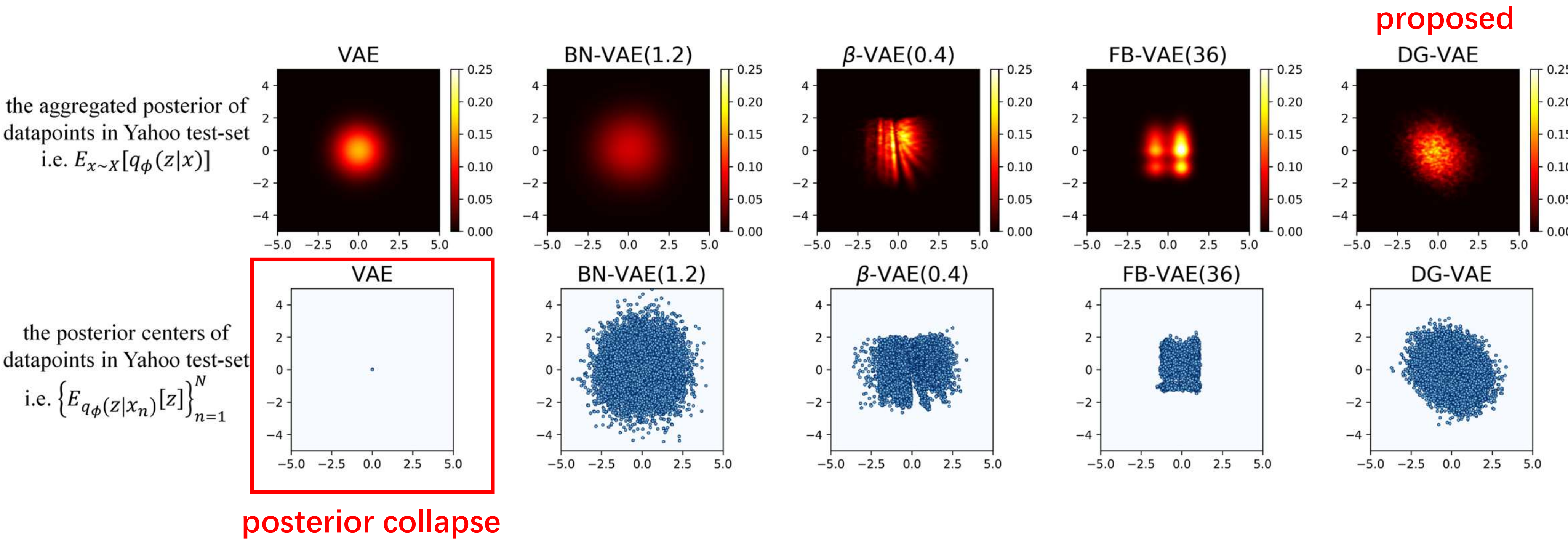
➔ $\forall x\ p_\theta(x|\boldsymbol{z}) = \frac{p_\theta(x,\boldsymbol{z})}{p_\theta(\boldsymbol{z})} \approx \frac{p_\theta(x,\boldsymbol{z})}{p_\theta(\boldsymbol{z}|x)} = p_\theta(x)$

i.e., the decoder $\boldsymbol{\theta}$ becomes insensitive to $\boldsymbol{z}$

i.e., the decoder degenerates to an unconditional language model (for NLG)

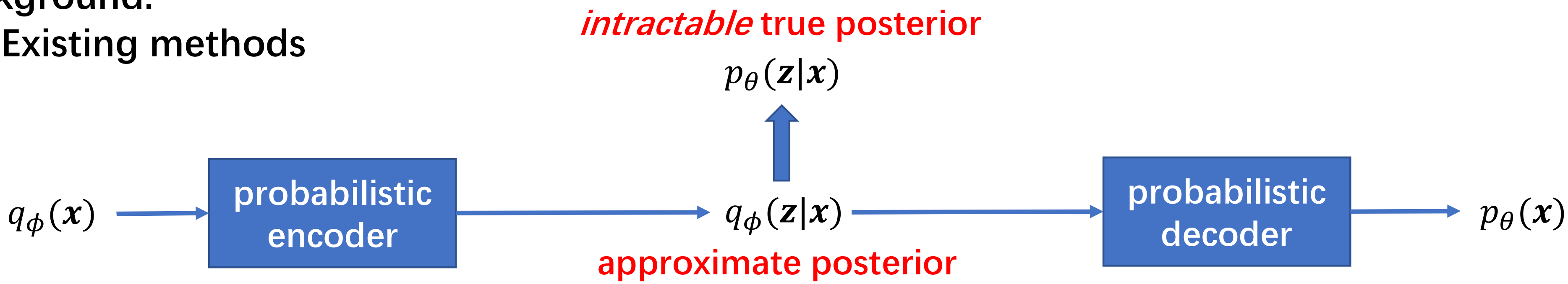
1. Background:
   c. Existing methods





Improving Variational Autoencoders with Density Gap-based Regularization

1. Background:
    c. Existing methods



$$\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

Posterior Collapse:

$$\forall x\ D_{KL}(q_\phi(\boldsymbol{z}|x)||p_\theta(\boldsymbol{z})) \approx 0$$

➔**training strategy:**

Cyclic-VAEs (cyclic annealing schedule); **AE pretraining;**

➔**semantic learning of $\boldsymbol{z}$:**

Skip-VAE (skip connection on z); **BOW-VAEs** (Bag-of-Word loss term on z);

➔**hard restriction on $q_\phi(\boldsymbol{z}|\boldsymbol{x})$:**

BN-VAEs (BN layer on $q_\phi(\boldsymbol{z}|\boldsymbol{x})$); **vMF-VAEs** (vMF distributions for $q_\phi(\boldsymbol{z}|\boldsymbol{x})$) and $p_\theta(\boldsymbol{z})$);

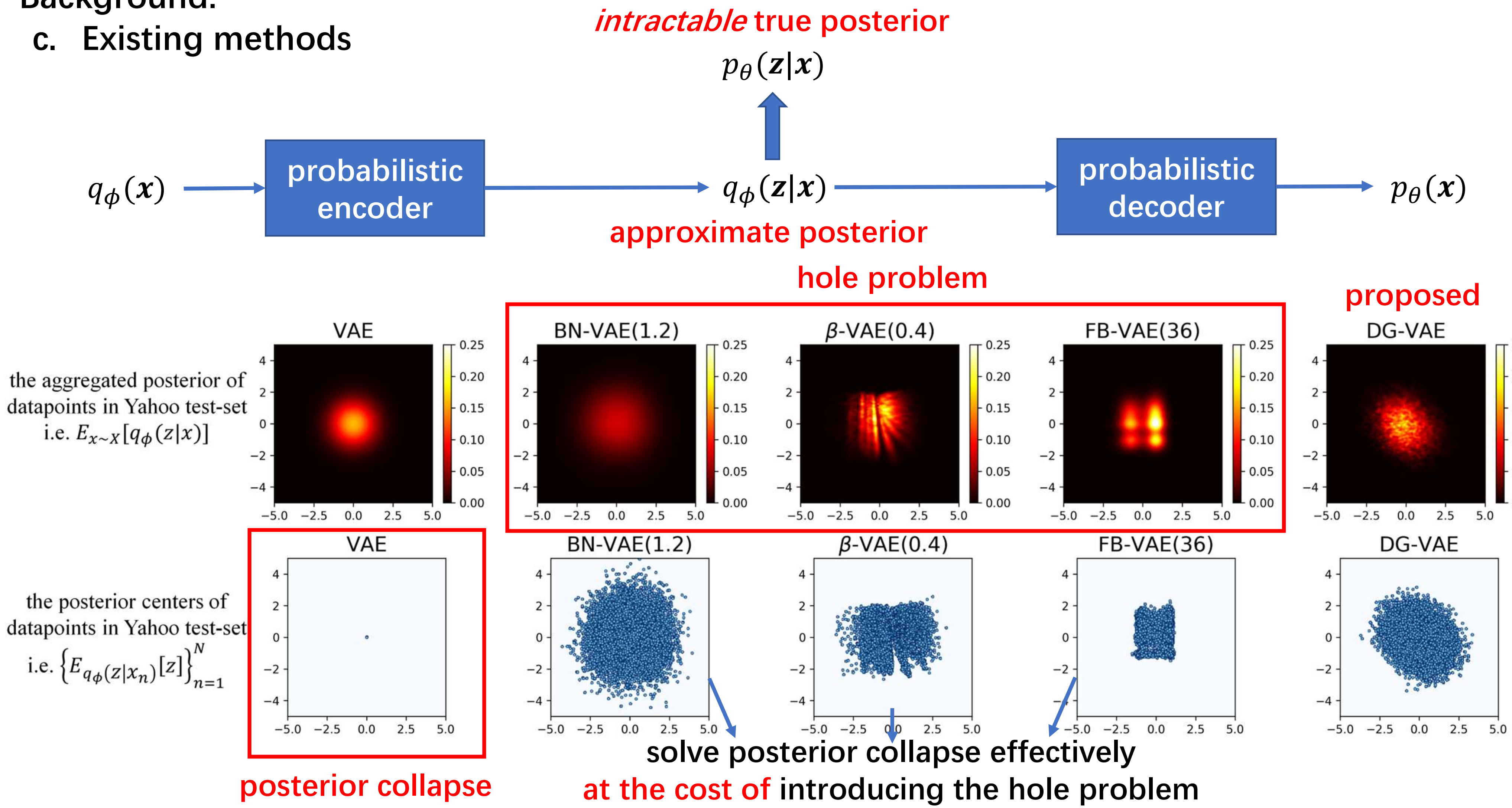
➔**weakening $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$ in $\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x})$:**

$\boldsymbol{\beta}$-VAEs (smaller weight of $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$ in $\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x})$);

FB-VAEs (hinge loss of $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$ in $\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x})$);

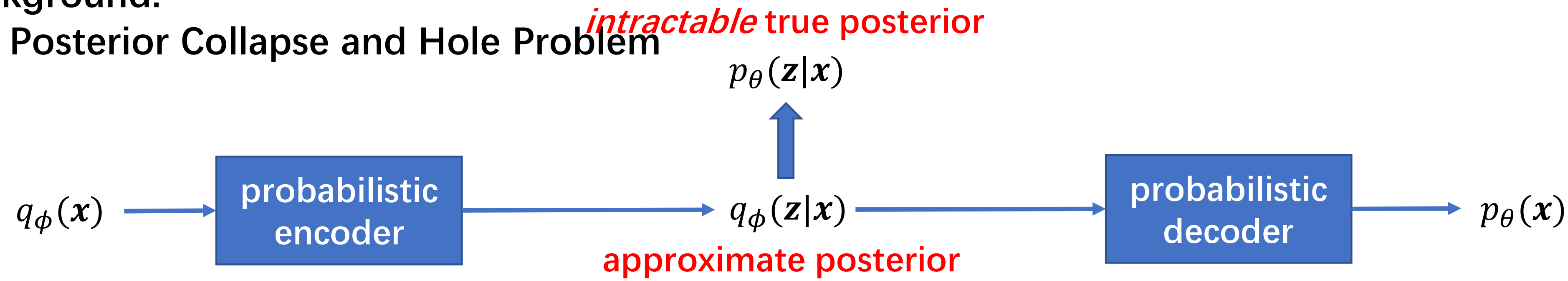
1. **Background:**

   c. **Existing methods**



Improving Variational Autoencoders with Density Gap-based Regularization

1. Background:
    b. Posterior Collapse and Hole Problem



$$\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

Hole Problem:

$$q_\phi(\boldsymbol{z}) \neq p_\theta(\boldsymbol{z})$$

where, $q_\phi(\boldsymbol{z}) = E_{q_\phi(\boldsymbol{x})}[q_\phi(\boldsymbol{z}|\boldsymbol{x})]$: the aggregated approximate posterior distribution

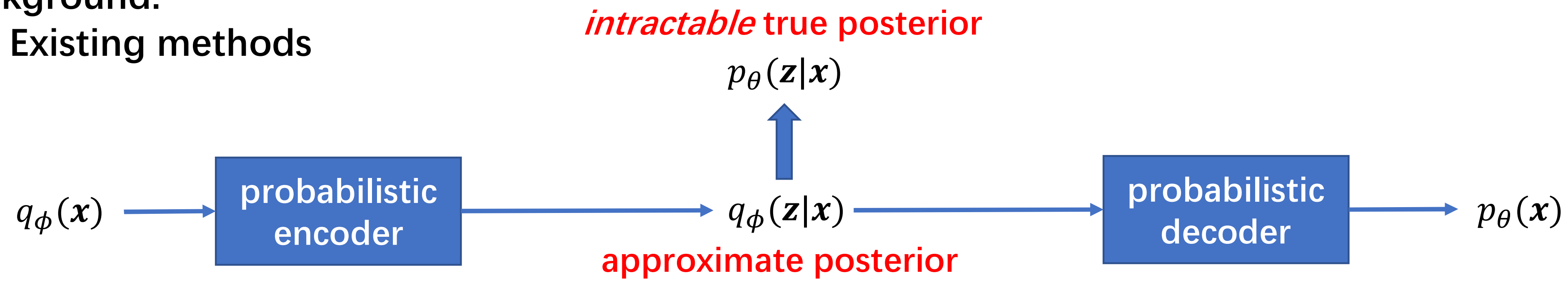
➔ $\exists z \; q_\phi(z) \neq p_\theta(z)$

i.e. there exist areas (named as holes) with mismatch between density in $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$

Empirically, inferences located in such areas are observed to perform low-quality generation, e.g., obscure and corrupted images, or sentences against commonsense.

1. Background:

   c. Existing methods



$$\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

Hole Problem:

$$q_\phi(\boldsymbol{z}) \neq p_\theta(\boldsymbol{z})$$

For image generation:

➔ascribed to the limited expressivity of $p_\theta(\boldsymbol{z})$ ($p_\theta(\boldsymbol{z}) = N(\mathbf{0},\mathbf{I})$ by default)

➔ tackled by increasing the flexibility of $p_\theta(\boldsymbol{z})$ through:

hierarchical priors, energy-based models, a mixture of encoders, etc.

For text generation:

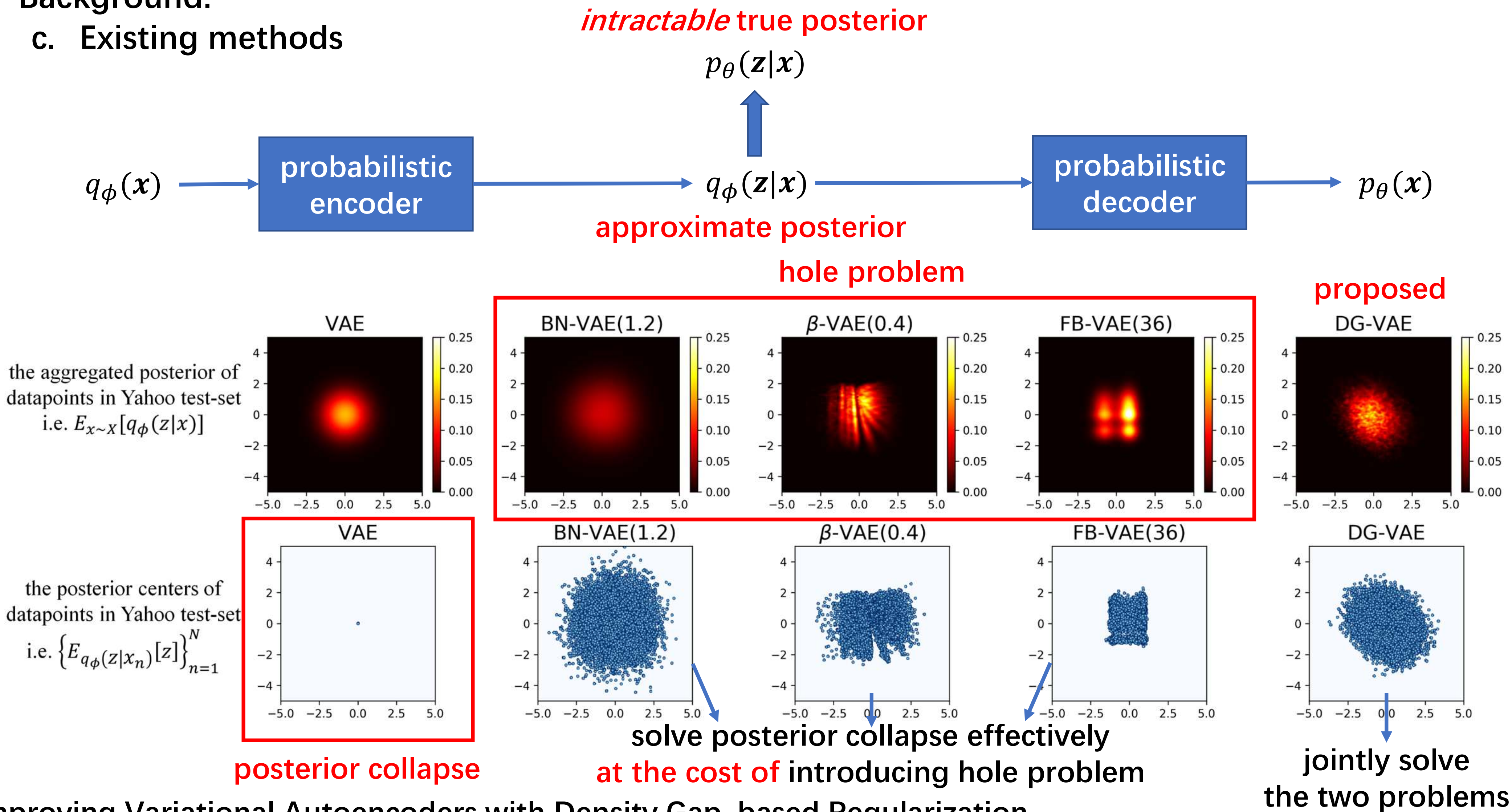
➔there's still little work on this, and we found that:

1. the vanilla VAEs (with $p_\theta(\boldsymbol{z}) = N(\mathbf{0},\mathbf{I})$) for text generation has no hole problem;

2. existing methods can solve posterior collapse effectively at the cost of introducing hole problem;

1. Background:

c. Existing methods



2. Methodology:
   a. Regularization on the aggregated posterior distribution

rethink of $\mathcal{L}_{ELBo}(\theta, \phi, \boldsymbol{x})$:

$$\mathcal{L}_{ELBo}(\theta, \phi, \boldsymbol{x}) = E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \underline{D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))}$$

Q1: Since $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ should not be too close to $p_\theta(\boldsymbol{z})$ (otherwise it will lead to posterior collapse), what should be close to $p_\theta(\boldsymbol{z}) = E_{p_\theta(\boldsymbol{x})}[p_\theta(\boldsymbol{z}|\boldsymbol{x})]$?
A1: The aggregated posterior distribution $q_\phi(\boldsymbol{z}) = E_{q_\phi(\boldsymbol{x})}[q_\phi(\boldsymbol{z}|\boldsymbol{x})]$.

Q2: So, how about regularizing $q_\phi(\boldsymbol{z})$ towards $p_\theta(\boldsymbol{z})$ instead in VAEs?
A2: It turns out to maximize $E_{q_\phi(\boldsymbol{x})}\mathcal{L}_{ELBo}(\theta, \phi, \boldsymbol{x}) + \mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n}, \boldsymbol{z}]$ (Hoffman et al. 2016):

$$E_{q_\phi(\boldsymbol{x})}\mathcal{L}_{ELBo}(\theta, \phi, \boldsymbol{x}) + \mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n}, \boldsymbol{z}] = E_{q_\phi(\boldsymbol{x})}E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \underline{D_{KL}(q_\phi(\boldsymbol{z})||p_\theta(\boldsymbol{z}))}$$

$$\mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n}, \boldsymbol{z}] = E_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\log \frac{q_\phi(\boldsymbol{n}, \boldsymbol{z})}{q_\phi(\boldsymbol{n})q_\phi(\boldsymbol{z})}]$$

where $\boldsymbol{n}$ is the identity of datapoints in $\boldsymbol{x}$, i.e., $q_\phi(\boldsymbol{n} = n) = \frac{1}{N}, (n = 1,2, \dots, N)$

effect: 1. weaken the regularization on $q_\phi(\boldsymbol{z}|\boldsymbol{x})$; 2. ensure $q_\phi(\boldsymbol{z}) \approx p_\theta(\boldsymbol{z})$.

2. Methodology:
   a. Regularization on the aggregated posterior distribution

Q3: Has anyone tried “regularizing $q_\phi(\boldsymbol{z})$ towards $p_\theta(\boldsymbol{z})$ instead in VAEs”?

A3: Yes, as below:

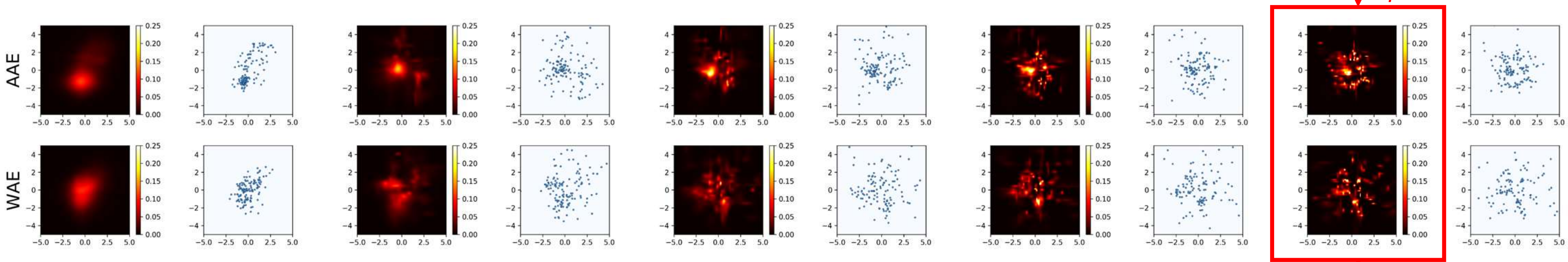
AAE (Adversarial Auto-Encoder): minimize their JS divergence in the framework of GAN

WAE (Wasserstein Auto-Encoder): minimize the Maximum Mean Discrepancy between them

iVAE$_{\text{MI}}$ (implicit VAE + MI regularization): minimize a dual form of KL divergence between them

But all their implementations of regularization are based on merely sampling sets from $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$, and lead to a kind of local optimums.

$q_\phi(\boldsymbol{z})$



1. a sampling set from such a $q_\phi(\boldsymbol{z})$ can already stimulate that from $p_\theta(\boldsymbol{z})$ to some degree;

2. but such a $q_\phi(\boldsymbol{z})$ still have evident difference from $p_\theta(\boldsymbol{z})$

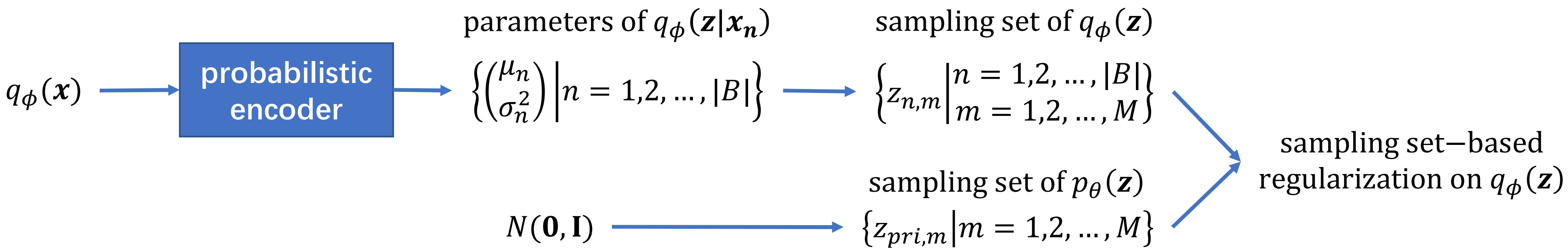
Intuitively, a sampling set from $q_\phi(\boldsymbol{z})$ can hardly be the same as that from $p_\theta(\boldsymbol{z})$, even when $q_\phi(\boldsymbol{z}) = p_\theta(\boldsymbol{z})$

2. Methodology:

b. Density Gap-based regularization

For example,

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}_n) = N(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n^2), p_\theta(\boldsymbol{z}) = N(\mathbf{0}, \mathbf{I})$$



2. Methodology:

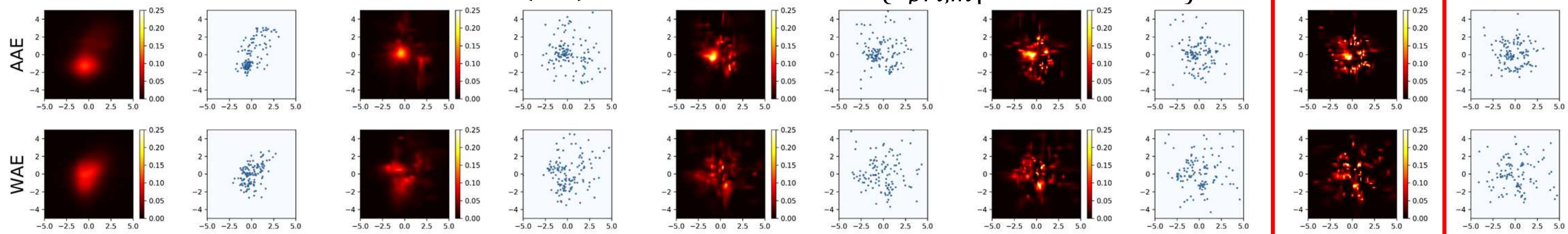
b. Density Gap-based regularization

For example,

$$q_\phi(\boldsymbol{z}|\boldsymbol{x_n}) = N\left(\boldsymbol{\mu_n}, \boldsymbol{\sigma_n^2}\right), p_\theta(\boldsymbol{z}) = N(\mathbf{0}, \mathbf{I})$$

$q_\phi(\boldsymbol{x})$ → probabilistic encoder → parameters of $q_\phi(\boldsymbol{z}|\boldsymbol{x_n})$: $\left\{\begin{pmatrix}\mu_n \\ \sigma_n^2\end{pmatrix} \middle| n = 1,2,\dots,|B|\right\}$ → sampling set of $q_\phi(\boldsymbol{z})$: $\left\{z_{n,m} \middle| \begin{matrix} n = 1,2,\dots,|B| \\ m = 1,2,\dots,M \end{matrix}\right\}$ → sampling set−based regularization on $q_\phi(\boldsymbol{z})$

$N(\mathbf{0}, \mathbf{I})$ → sampling set of $p_\theta(\boldsymbol{z})$: $\left\{z_{pri,m} \middle| m = 1,2,\dots,M\right\}$ → sampling set−based regularization on $q_\phi(\boldsymbol{z})$



1. a sampling set from such a $q_\phi(\boldsymbol{z})$ can already stimulate that from $p_\theta(\boldsymbol{z})$ to some degree;

2. but such a $q_\phi(\boldsymbol{z})$ still have evident difference from $p_\theta(\boldsymbol{z})$

Intuitively, a sampling set from $q_\phi(\boldsymbol{z})$ can hardly be the same as that from $p_\theta(\boldsymbol{z})$, even when $q_\phi(\boldsymbol{z}) = p_\theta(\boldsymbol{z})$

2. Methodology:

b. Density Gap-based regularization

Intuitively, a sampling set from $q_\phi(\boldsymbol{z})$ can hardly be the same as that from $p_\theta(\boldsymbol{z})$, even when $q_\phi(\boldsymbol{z}) = p_\theta(\boldsymbol{z})$

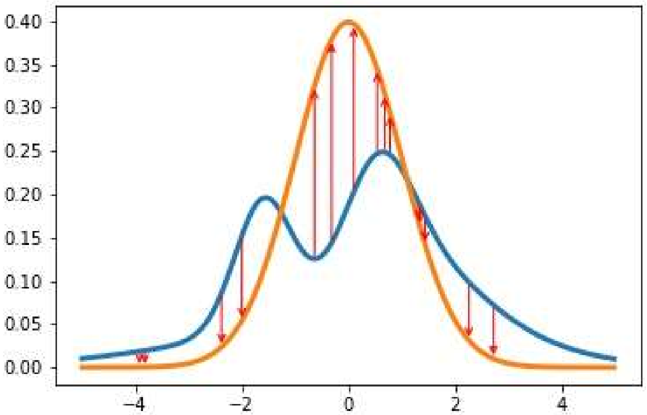
➔The probability density of $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$ are the same everywhere when $q_\phi(\boldsymbol{z}) = p_\theta(\boldsymbol{z})$

➔Density Gap-based regularization:

$$KL\left(q_\phi(\boldsymbol{z})||p_\theta(\boldsymbol{z})\right) = E_{q_\phi(\boldsymbol{z})}[\log\frac{q_\phi(\boldsymbol{z})}{p_\theta(\boldsymbol{z})}]$$

$$\approx \frac{1}{S}\sum_{s=1}^{S}\underbrace{[\log q_\phi(z_s) - \log p_\theta(z_s)]}_{\text{we refer to this as}}$$

the density gap between $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$ at position z_s



where,

z_s is the s^{th} sample from $q_\phi(\boldsymbol{z})$ ➔ stratified sampling & reparameterization trick

$q_\phi(z_s)$ and $p_\theta(z_s)$ are the values of corresponding PDFs ➔ parametric differentiable PDFs

2. Methodology:

b. Density Gap-based regularization

For example,

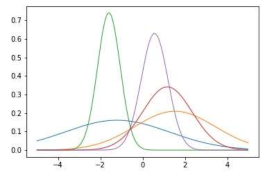
$$q_\phi(\boldsymbol{z}|\boldsymbol{x_n}) = N\left(\boldsymbol{\mu_n}, \boldsymbol{\sigma_n^2}\right), p_\theta(\boldsymbol{z}) = N(\mathbf{0}, \mathbf{I})$$

$q_\phi(\boldsymbol{x})$ → probabilistic encoder → parameters of $q_\phi(\boldsymbol{z}|\boldsymbol{x_n})$: $\left\{\begin{pmatrix}\mu_n \\ \sigma_n^2\end{pmatrix} \middle| n = 1,2,\dots,|B|\right\}$ → sampling set of $q_\phi(\boldsymbol{z})$: $\left\{z_{n,m} \middle| \begin{matrix} n = 1,2,\dots,|B| \\ m = 1,2,\dots,M \end{matrix}\right\}$

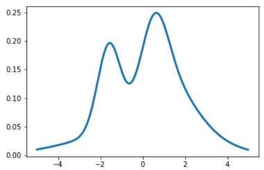
$N(\mathbf{0}, \mathbf{I})$ → sampling set of $p_\theta(\boldsymbol{z})$: $\{z_{pri,m} | m = 1,2,\dots,M\}$

sampling set−based regularization on $q_\phi(\boldsymbol{z})$

$\{q_\phi(\boldsymbol{z}|\boldsymbol{x_n})|n\}$



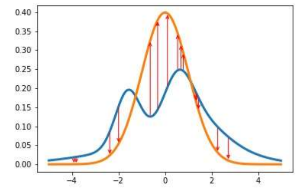
$q_\phi(\boldsymbol{z})$



$$q_\phi(z) = \frac{1}{|B|}\sum_{n=1}^{|B|} \exp\left(-\frac{(z-\mu_n)^2}{2\sigma_n^2}\right) / \sqrt{2\pi}\sigma_n$$

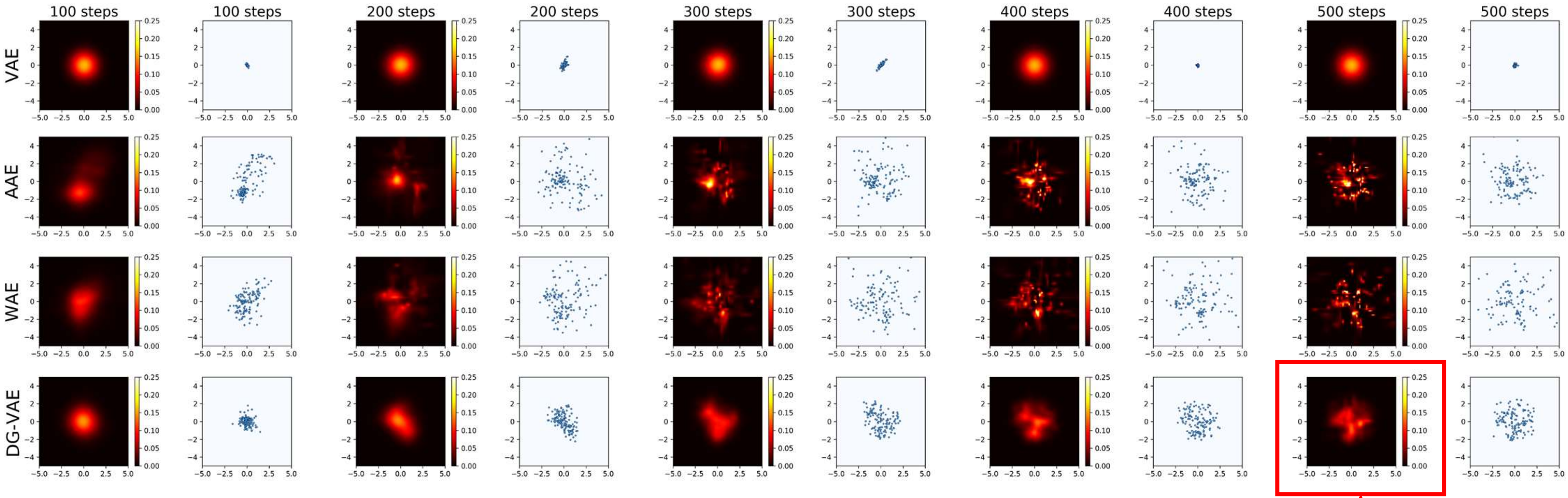
$$p_\theta(z) = \exp(-z^2/2) / \sqrt{2\pi}$$

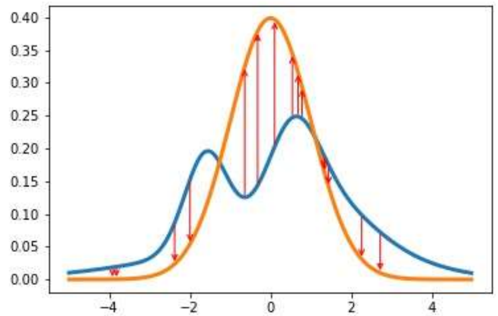
density gap−based regularization on $q_\phi(\boldsymbol{z})$



2. Methodology:

b. Density Gap-based regularization





regularize $q_\phi(z)$ towards $p_\theta(z)$ in the perspective of their mismatch in PDFs

Improving Variational Autoencoders with Density Gap-based Regularization

2. Methodology:

c. Marginal regularization for more Mutual Information

We can apply the proposed regularization in training with mini-batch gradient descent:

$$E_{q_\phi(\boldsymbol{x})}\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) + \mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n},\boldsymbol{z}] = E_{q_\phi(\boldsymbol{x})}E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z})||p_\theta(\boldsymbol{z}))$$

where the data distribution $q_\phi(\boldsymbol{x})$ is described by the current mini-batch B:

$$B = \{x_1, x_2, \dots, x_{|B|}\}$$

$$q_\phi(\boldsymbol{x} = x_{\boldsymbol{n}}) = q_\phi(\boldsymbol{n}) = \frac{1}{|B|}$$

➔the mutual information term to maximize has a limited upper bound:

$$\mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n},\boldsymbol{z}] = H_{q_\phi(\boldsymbol{n})}(\boldsymbol{n}) - H_{q_\phi(\boldsymbol{n},\boldsymbol{z})}(\boldsymbol{n}|\boldsymbol{z}) \le H_{q_\phi(\boldsymbol{n})}(\boldsymbol{n}) = \log|B| < \log N$$

➔for a high dimensional prior distribution, it still have limited effect on solving posterior collapse (it is already enough for $\mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z})}[\boldsymbol{n},\boldsymbol{z}]$ to reach $\log|B|$ with limited dimensions of $\boldsymbol{z}$ being activated)

➔in order to activate all dimensions of $\boldsymbol{z}$, we propose marginal regularization:

$$E_{q_\phi(\boldsymbol{x})}\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x}) + \sum_{i=1}^{Dim} \mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z}_i)}[\boldsymbol{n},\boldsymbol{z}_i] = E_{q_\phi(\boldsymbol{x})}E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \sum_{i=1}^{Dim} D_{KL}(q_\phi(\boldsymbol{z}_i)||p_\theta(\boldsymbol{z}_i))$$

where $i = 1,2,\dots,Dim$ denotes the index of dimension, $\boldsymbol{z}_i$ denotes the i^{th} component of $\boldsymbol{z}$, $q_\phi(\boldsymbol{z}_i)$ and $p_\theta(\boldsymbol{z}_i)$ denote the marginal distribution of $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$ on the i^{th} dimension respectively.

2. Methodology:

c. Marginal regularization for more Mutual Information

➔in order to activate all dimensions of z, we propose marginal regularization:

$$E_{q_\phi(\boldsymbol{x})}\mathcal{L}_{ELBo}(\theta,\phi,\boldsymbol{x})+\sum_{i=1}^{Dim}\mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z}_i)}[\boldsymbol{n},\boldsymbol{z}_i]=E_{q_\phi(\boldsymbol{x})}E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]-\sum_{i=1}^{Dim}D_{KL}(q_\phi(\boldsymbol{z}_i)||p_\theta(\boldsymbol{z}_i))$$

where $i=1,2,\dots,Dim$ denotes the index of dimension, $\boldsymbol{z}_i$ denotes the i^{th} component of $\boldsymbol{z}$, $q_\phi(\boldsymbol{z}_i)$ and $p_\theta(\boldsymbol{z}_i)$ denote the marginal distribution of $q_\phi(\boldsymbol{z})$ and $p_\theta(\boldsymbol{z})$ on the i^{th} dimension respectively.

➔in such way, the mutual information term to maximize has an upper bound linear with Dim:

$$\sum_{i=1}^{Dim}\mathbb{I}_{q_\phi(\boldsymbol{n},\boldsymbol{z}_i)}[\boldsymbol{n},\boldsymbol{z}_i]\le\sum_{i=1}^{Dim}H_{q_\phi(\boldsymbol{n})}(\boldsymbol{n})=Dim*\log|B|$$

➔we implement this for VAEs with $p_\theta(\boldsymbol{z})=N(\mathbf{0},\mathbf{I})$, as its marginal distributions are independent:

$$p_\theta(\boldsymbol{z})=\prod_{i=1}^{Dim}p_\theta(\boldsymbol{z}_i)$$

➔it should be noted that, this independency-based decomposition of $p_\theta(\boldsymbol{z})$ is not established for von Mises-Fisher distributions, e.g., $p_\theta(\boldsymbol{z})=vMF(\boldsymbol{\mu},\kappa)$, so we only implement the joint regularization for von Mises-Fisher distribution-based VAEs.

2. Methodology:

d. Aggregation size for ablation

➔to further investigate the effect of maximizing mutual information, we split the mini-batch B into non-overlapping subsets:

$$B = \bigcup_{i=1}^{C} b_i \text{, s.t. } b_i \cap b_j = \emptyset \text{ iff } i \neq j$$

those subsets have the same size $|b| = |b_i| = |b_j| = \frac{|B|}{C}$ which we refer to as the aggregation size, as we only calculate the aggregated posterior distributions inside each subsets, and regularize them to the prior distribution respectively:

$$q_{\phi,j}(\boldsymbol{z}) = E_{\boldsymbol{x}\sim b_j}[q_\phi(\boldsymbol{z}|\boldsymbol{x})]$$

$$\sum_{j=1}^{C} \sum_{i=1}^{Dim} D_{KL}(q_{\phi,j}(\boldsymbol{z}) || p_\theta(\boldsymbol{z}_i))$$

➔in such way, the maximized mutual information term has an upper bound linear with $\log|b|$:

$$\sum_{j=1}^{C} \sum_{i=1}^{Dim} \mathbb{I}_{q_{\phi,j}(\boldsymbol{n},\boldsymbol{z}_i)}[\boldsymbol{n}, \boldsymbol{z}_i] \leq \sum_{j=1}^{C} \sum_{i=1}^{Dim} H_{q_{\phi,j}(\boldsymbol{n})}(\boldsymbol{n}) = C * Dim * \log|b|$$

when $|b| = 1$, the proposed method is equivalent to the vanilla VAE.

3. Experiment

a. Language modeling

Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg $\pm$ std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

Table 2: Results of Language Modeling on Yahoo dataset. We bold up $MI(\phi) \geq 9.0$, $AU(\phi) \geq 30$, $CU(\phi) \geq 30$, the highest $priorLL(\boldsymbol{\theta})$ and $postLL(\boldsymbol{\theta}, \phi)$ for the same methods.

Models	$priorLL(\boldsymbol{\theta})$	$postLL(\boldsymbol{\theta}, \phi)$	$KL(\phi)$	$MI(\phi)$	$AU(\phi)$	$CU(\phi)$
VAE (default)	-330.7	-330.7	0.0	0.0	0	**32**
cyclic-VAE	-329.8	-328.9	1.1	1.0	2	**31**
bow-VAE	-330.5	-330.5	0.0	0.0	0	**32**
skip-VAE	-330.1	-325.2	5.0	4.3	8	**31**
δ-VAE(0.15)	-330.5	-330.6	4.8	0.0	0	0
BN-VAE(0.6)	**-327.6**	-321.1	6.6	5.9	**32**	**32**
BN-VAE(1.2)	-330.9	-310.1	26.2	**9.2**	**32**	0
BN-VAE(1.8)	-343.5	**-308.6**	51.3	**9.2**	**32**	0
FB-VAE(4)	-329.8	-328.4	3.9	1.8	**32**	**32**
FB-VAE(16)	**-325.7**	-320.8	16.1	8.5	**32**	8
FB-VAE(49)	-344.6	**-296.1**	50.0	**9.2**	**32**	0
β-VAE(0.4)	**-330.8**	-324.8	7.0	6.7	3	**31**
β-VAE(0.2)	-338.6	-310.3	30.1	**9.2**	22	25
β-VAE(0.1)	-369.9	**-289.6**	83.7	**9.2**	**32**	0
DG-VAE ($\lvert b \rvert = 1$)	-330.7	-330.7	0.0	0.0	0	**32**
DG-VAE ($\lvert b \rvert = 4$)	**-330.4**	-318.3	14.3	**9.1**	11	**32**
DG-VAE ($\lvert b \rvert = 32$)	-355.4	-294.1	65.2	**9.1**	**32**	**32**
DG-VAE (default)	-358.0	**-290.8**	70.8	**9.1**	**32**	**32**

$$priorLL(\theta) = E_x \log E_{p_\theta(z)}[p_\theta(x|z)]$$

$$postLL(\theta, \phi) = E_x \log E_{q_\phi(z|x)}[p_\theta(x|z)]$$

$$KL(\phi) = E_x KL(q_\phi(z|x)||p_\theta(z))$$

$$MI(\phi) = H(q_\phi(z)) - E_x H(q_\phi(z|x))$$

$$AU(\phi) = |\{i|Var_x E_{q_\phi(z|x)}[z_i] > 0.01\}|$$

Small values indicate posterior collapse

$$CU(\phi) = |\{i|KL(q_\phi(z_i)||p_\theta(z_i)) < 0.03\}|$$

Small values indicate the hole problem

Improving Variational Autoencoders with Density Gap-based Regularization

3. Experiment

a. Language modeling

Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg $\pm$ std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

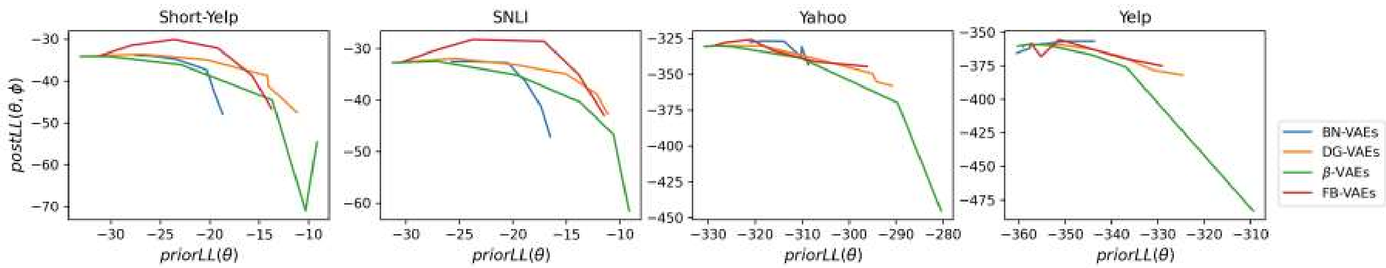


Figure 2: The curves of $priorLL(\boldsymbol{\theta})$ and $postLL(\boldsymbol{\theta}, \phi)$ in Gaussian distribution-based VAEs.

$$priorLL(\theta) = E_x \log E_{p_\theta(z)}[p_\theta(x|z)]$$

$$postLL(\theta, \phi) = E_x \log E_{q_\phi(z|x)}[p_\theta(x|z)]$$

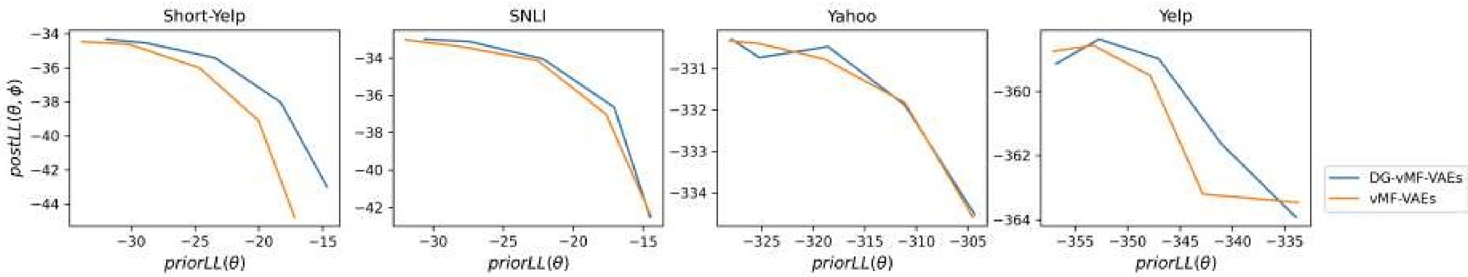


Figure 3: The curves of $priorLL(\boldsymbol{\theta})$ and $postLL(\boldsymbol{\theta}, \phi)$ in vMF distribution-based VAEs.

Improving Variational Autoencoders with Density Gap-based Regularization

3. Experiment

b. Visualization of the posterior

Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg $\pm$ std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

ours

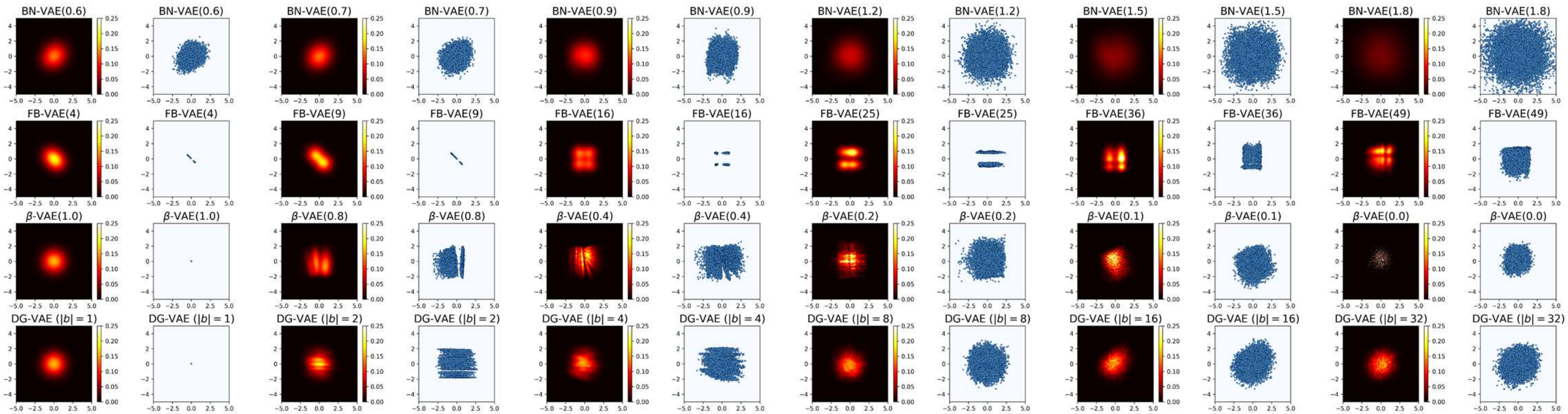


Figure 4: The visualization of the aggregated posterior distributions (red-in-black) and the posterior centers distributions (blue-in-white) for BN-VAEs, FB-VAEs, β-VAEs, and DG-VAEs on the Yahoo test-set. Illustrations for more datasets, more models, and more dimensions, are shown in Appendix G.

3. Experiment

c. Interpolation study

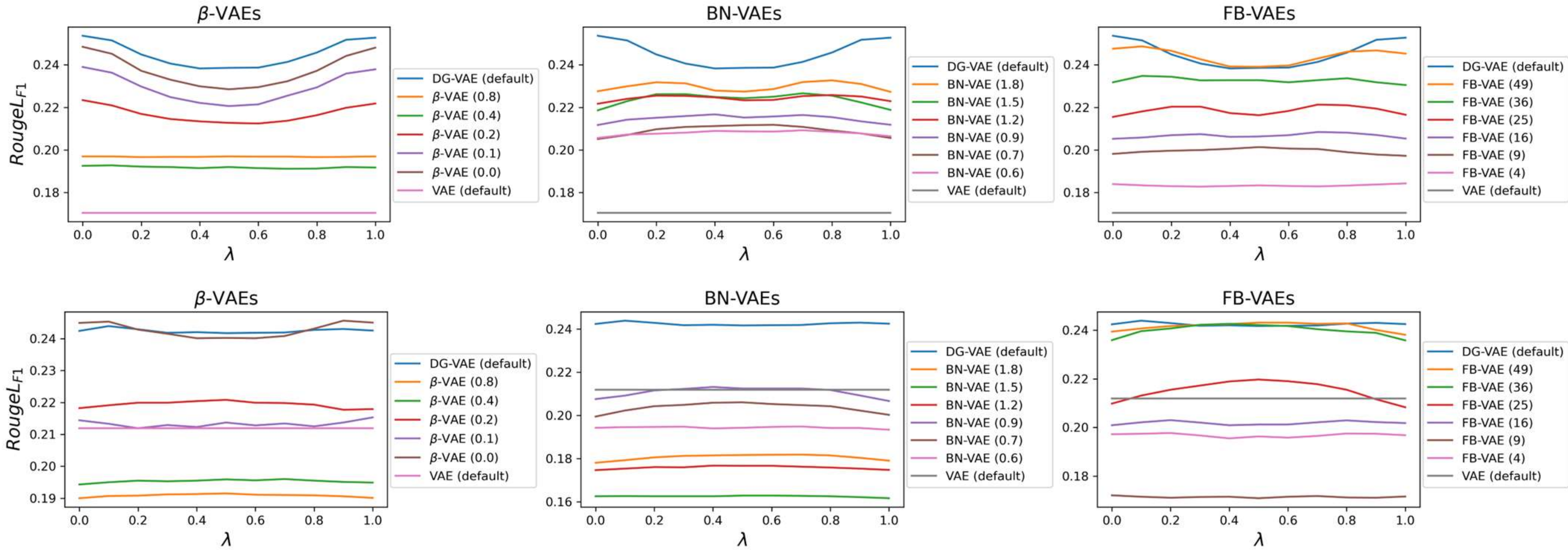
Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg ± std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

$$z_a, z_b \sim q_\phi(\boldsymbol{z}|x_a), q_\phi(\boldsymbol{z}|x_b)$$
$$z_\lambda = \lambda * z_a + (1-\lambda) * z_b$$
$$x_\lambda \sim p_\theta(\boldsymbol{x}|z_\lambda)$$
$$RougeL_{F1} = \frac{1}{2}\left(F_{lcs}(x_a, x_\lambda) + F_{lcs}(x_b, x_\lambda)\right)$$

Yahoo

Yelp



Improving Variational Autoencoders with Density Gap-based Regularization

3. Experiment

c. Interpolation study

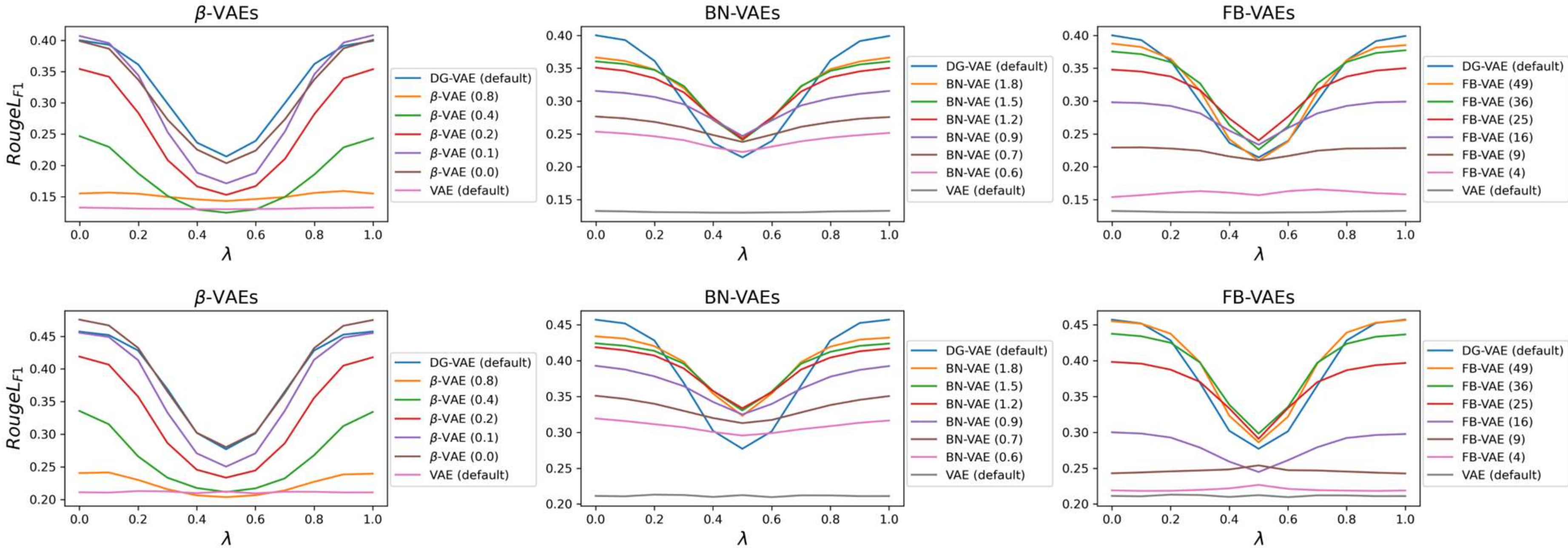
Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg $\pm$ std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

$$z_a, z_b \sim q_\phi(\boldsymbol{z}|x_a), q_\phi(\boldsymbol{z}|x_b)$$
$$z_\lambda = \lambda * z_a + (1-\lambda) * z_b$$
$$x_\lambda \sim p_\theta(\boldsymbol{x}|z_\lambda)$$
$$RougeL_{F1} = \frac{1}{2}(F_{lcs}(x_a, x_\lambda) + F_{lcs}(x_b, x_\lambda))$$

Short-Yelp

SNLI



Improving Variational Autoencoders with Density Gap-based Regularization

3. Experiment

c. Interpolation study

$$z_a, z_b \sim q_\phi(\boldsymbol{z}|x_a), q_\phi(\boldsymbol{z}|x_b)$$
$$z_\lambda = \lambda * z_a + (1-\lambda) * z_b$$
$$x_\lambda \sim p_\theta(\boldsymbol{x}|z_\lambda)$$
$$RougeL_{F1} = \frac{1}{2}\left(F_{lcs}(x_a, x_\lambda) + F_{lcs}(x_b, x_\lambda)\right)$$

	$\boldsymbol{x}_a$:	two girls walking in a park .	$\boldsymbol{x}_b$:	the two kids are playing in water
		β-VAE(0.1)		DG-VAE
λ	$DG(z_\lambda)$	$\boldsymbol{x}_\lambda$	$DG(z_\lambda)$	$\boldsymbol{x}_\lambda$
0.0	32.9	two girls walking in a park .	38.9	two girls walking in a park .
0.1	26.1	two girls walking in a park .	35.5	two girls walking in a park .
0.2	11.4	two women walking in a park .	29.4	two girls walking in a park .
0.3	-11.3	two women walking in a pool .	20.6	two girls walking in a park .
0.4	-42.2	two cats playing in a pool .	9.2	two girls sit in a beach .
0.5	-55.4	an african man walks in the pool .	-4.8	two girls sit in beach
0.6	-48.0	an elderly man walks in water .	1.0	two girls are playing in a boat .
0.7	-15.1	the two children are playing in water .	18.7	four girls are playing in water
0.8	12.2	the two children are playing in water	31.7	the two kids are playing in water
0.9	30.0	the two kids are playing in water	40.1	the two kids are playing in water
1.0	38.2	the two kids are playing in water	43.9	the two kids are playing in water

Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg ± std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

	$\boldsymbol{x}_a$:	a man in a white shirt and black pants poses in front of a large banner .	$\boldsymbol{x}_b$:	two people hug each other to warm up while they are locked out of the house .
		β-VAE(0.1)		DG-VAE
λ	$DG(z_\lambda)$	$\boldsymbol{x}_\lambda$	$DG(z_\lambda)$	$\boldsymbol{x}_\lambda$
0.0	60.2	a man in a black shirt and black pants sits in front of a large gathering .	56.5	a man in a white shirt and black pants jumps in front of a large screen .
0.1	38.6	a man in a black shirt and blue pants walking in front of a large gathering .	44.2	a man in a white shirt and black pants jumps in front of a large screen .
0.2	-21.9	a man in jeans and a white shirt walking down in an orange kayak in the forest .	12.5	a man in a white shirt and black pants jumps in front of a large audience .
0.3	-121.4	a man in shorts and a black shirt walking through snow , on the street .	-38.6	a man in a black shirt and black pants kneels out in front of the <unk> .
0.4	-157.4	a toddler , wearing shorts and black pants walking a green scooter while looking in the water .	-47.2	a man wearing a black shirt shovels snow while standing in front of the <unk> .
0.5	-159.9	a toddler girl wearing black shorts and sandals walking through her house while on the sunny sidewalk .	-53.7	a man wearing a black shirt is shoveling snow , while standing in front of the <unk> .
0.6	-161.7	a toddler girl wearing pink pants and boots walks across the street in front of cars .	-63.5	two people wearing black shirts are waiting in front of two cars they made from a field .
0.7	-164.6	a girl , who looks over her head while she sits alone on the edge .	-76.7	two people chat while one is standing next to her friends on the deck .
0.8	-46.4	two people decide whether , as they walk up in the water while looking up .	-26.2	two people chat as they are standing next to two <unk> on the roof .
0.9	37.5	two people decide whether to each other . and one is out out of the window .	40.8	two people chat as they are trying to figure out how to the house .
1.0	66.8	two people hug as they walk out and sun to get out of the sun .	64.8	two people chat as they are trying to figure out how to get the best .

3. Experiment

c. Interpolation study

$$z_a, z_b \sim q_\phi(\boldsymbol{z}|x_a), q_\phi(\boldsymbol{z}|x_b)$$
$$z_\lambda = \lambda * z_a + (1-\lambda) * z_b$$
$$x_\lambda \sim p_\theta(\boldsymbol{x}|z_\lambda)$$
$$RougeL_{F1} = \frac{1}{2}(F_{lcs}(x_a, x_\lambda) + F_{lcs}(x_b, x_\lambda))$$

	x_a:	great place for a romantic <unk> .	x_b:	the asian cucumber salad was bland .
		β-VAE(0.1)		DG-VAE
λ	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
0.0	46.9	great place for a romantic <unk> .	52.1	great place for a romantic dinner .
0.1	32.2	great place for a romantic <unk> .	38.3	great place for a romantic dinner .
0.2	-5.6	great place for a chilly <unk> .	3.5	great place for lunch .
0.3	-55.6	oh you 're perfect and special .	-16.7	great place for lunch .
0.4	-60.4	oh you 'll enjoy the special .	-9.2	great for lunch .
0.5	-75.0	the guys keep it clean though .	-9.4	the best lunch was an tasteless .
0.6	-86.0	the apartments make it was comfortable .	-17.2	the usual street salad was boring .
0.7	-72.0	the specialty pie are good out .	-32.7	the usual cookie salad was bland .
0.8	-6.3	the wood martinis are very cheap .	11.6	the usual cookie salad was bland .
0.9	34.3	the wood martinis taste was bland .	43.6	the usual scallops which was bland .
1.0	49.7	the english muffins were good bland .	56.8	the usual scallops which was bland .

Table 1: Statistics of sentences in the datasets

Dataset	Train	Valid	Test	Vocab size	Length (avg $\pm$ std)
Yelp	100,000	10,000	10,000	19997	98.01 ± 48.86
Yahoo	100,000	10,000	10,000	20001	80.76 ± 46.21
Short-Yelp	100,000	10,000	10,000	8411	10.96 ± 3.60
SNLI	100,000	10,000	10,000	9990	11.73 ± 4.33

	x_a:	our server was not even <unk> familiar with the food or food preparation .	x_b:	i have had just about everything on the menu and everything is delicious !
		β-VAE(0.1)		DG-VAE
λ	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
0.0	82.1	our server was not even warm about the food and the quality service .	69.8	our server was not even <unk> with the food or service on food .
0.1	58.0	our server was not even busy on the menu and the food network .	51.9	our server was not even <unk> with the food or food poisoning .
0.2	-11.4	our waitress was not even busy on the menu and the food sucked .	1.0	our server was n't a few times but the food seems absolutely delicious .
0.3	-126.1	still they was not even warm by the food and taste was awesome .	-33.7	our dishes were n't a bit of everything but the food has been impeccable .
0.4	-179.0	still unfortunately the cashier just kept asking the menu and was seriously awesome .	-32.8	i actually was still having a bit on the menu that was terrible and beyond .
0.5	-177.4	even we was nothing as much of the menu and food was decent !	-32.4	i feel so just of a plus on the menu and food was beyond delicious .
0.6	-169.5	then i still had all about eating at the food and everything was good !	-33.2	i feel all just for the amount of food and everything was beyond delicious .
0.7	-63.0	did i say before they use no food and there are very quick !	-13.2	i have also had one of the items on food and everything was delicious !
0.8	13.6	did i say before they use from the menu and menu was decent !	29.4	i have also had no lunch on the menu and everything is outstanding !
0.9	60.0	i have had just just twice there and their food is very yummy !	56.5	i have also had no reservations on the menu and everything is delicious !
1.0	76.3	i have had just just because of the menu and everything is awesome !	68.1	i have also just ordered each on the menu and everything is delicious !