A Performance of AAE and WAE on Synthetic Dataset

We illustrate the hole problem of Adversarial Auto-Encoder (AAE) [3] and Wasserstein Auto-Encoder (WAE) [4] through a toy experiment on a synthetic dataset, which contains 128 datapoints represented as 2-D Gaussian posterior distributions. We compare VAE, AE, AAE, and WAE with our proposed DG-VAE in this experiment through latent space visualization, including the visualization of the aggregated posterior distribution and the distribution of datapoints' posterior centers along the training process, as depicted in Figure 1. The models have an embedding layer as the encoder and a two-layer MLP classifier as the decoder. The batch size is set to 16 and all models are trained with Adam optimizer with an initial learning rate of 0.1.

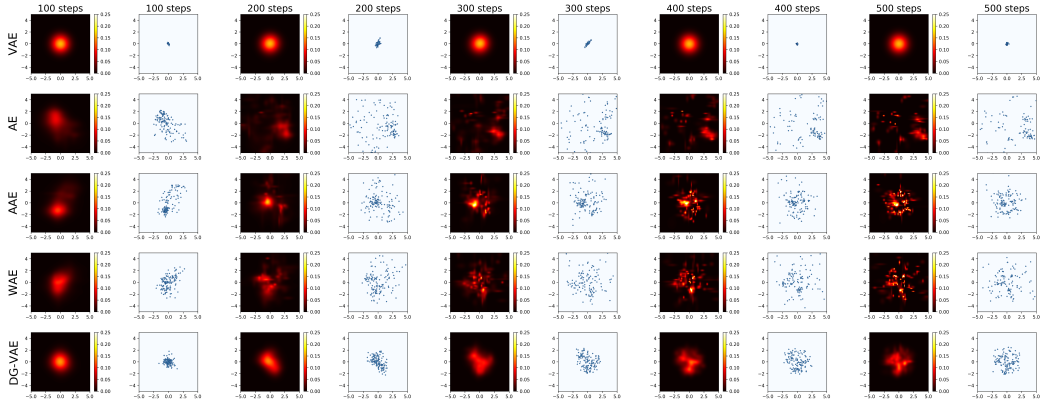

Figure 1: The visualization of aggregated posterior distribution (red-in-black) and datapoints' posterior centers distribution (blue-in-white) for VAE, AE, AAE, WAE and DG-VAE along 500 training steps.

It can be observed that the datapoints' posterior centers in VAE all collapse to the same point, i.e., posterior collapse, while the aggregated posterior fails to match the prior in WAE, AAE, and AE. Interestingly, a sampling set from the aggregated posterior distribution of WAE or AAE can already simulate that from the prior distribution to some degree; in this state, their sampling sets-based discrepancies between the aggregated posterior and the prior can be nearly the optimum.

In contrast, DG-VAE can solve posterior collapse and form a continuous latent space that matches the prior much better, as it optimizes the divergence between the aggregated posterior and the prior depicted by their density gap (instead of merely their sampling sets).

B Configurations

The configurations for the experiments are as follow: The dimension of word embeddings is 512 and the weights are randomly initialized by $U(-0.1, 0.1)$, while the other trainable parameters are initialized by $U(-0.01, 0.01)$. The encoder and the decoder are both implemented by a single layer LSTM [2] with 1024 hidden size. The sampled latent variable z is used to generate the initial hidden state of the decoder and concatenated with the word embedding for decoder input at each timestep.[1] The default batch size $|B|$ is set to 32, and each batch contains sentences of the same length. The latent dimension Dim is set to 32 for both Gaussian VAEs and vMF VAEs.

Each model is trained on one NVIDIA Tesla v100 by mini-batch SGD for at most 100 epochs except it performs overfit according to the valid set for 5 times. Training a model on a short dataset usually takes about 40 minutes while training on a long dataset usually takes about 8 hours. The sampling times M for Monte Carlo approximation in Eq. 9 and Eq. 11 is set to 32. The averaged training time of our model (over all experimental datasets) is only 11% higher than that of the vanilla VAE.

[1] Skip-VAE further feeds z into the vocabulary classifier.

C Language Modeling Metrics

We include the following metrics for the evaluation of VAEs on language modeling:

- $priorLL(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}}[\log \mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}(\mathbf{z})}[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]]$: the prior log likelihood, the log likelihood of sentences for the decoder given a latent variable from the prior distribution, which measures the unconditional generation ability of the decoder $\boldsymbol{\theta}$.
- $postLL(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}}[\log \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]]$: the posterior log likelihood, the log likelihood of sentences for the decoder given a latent variable from the posterior distribution of the corresponding sentences, which measures the conditional generation ability of the decoder $\boldsymbol{\theta}$ and the representation ability of the encoder $\boldsymbol{\phi}$;
- $KL(\boldsymbol{\phi}) = \mathbb{E}_{\mathbf{x}\sim\mathbf{X}}[KL(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(z))]$: the KL term in ELBo, a small $KL(\boldsymbol{\phi})$ indicates the phenomenon of posterior collapse;
- $MI(\boldsymbol{\phi}) = H(q_{\boldsymbol{\phi}}(\mathbf{z})) - \mathbb{E}_{\mathbf{x}\sim\mathbf{X}}[H(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}))]$: the mutual information of $\mathbf{z}$ and n in their joint distribution $q_{\boldsymbol{\phi}}(n, \mathbf{z})$, where $n = 1, 2, ..., |\mathbf{X}|$.
- $AU(\boldsymbol{\phi})$: the number of active units, a dimension of $\mathbf{z}$ is referred to as an active unit when the posterior centers of datapoints has an evident marginal variance, i.e., $Var_{\mathbf{x}\sim\mathbf{X}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{z}_i]] > 0.01$ means the i^{th} dimension is an active unit, which is proposed by Burda et al. [1] to measure the posterior collapse in a dimension-wise perspective. A lower value of $AU(\boldsymbol{\phi})$ indicates a severer posterior collapse issue;
- $CU(\boldsymbol{\phi})$: the number of consistent units, a dimension of $\mathbf{z}$ is referred to as a consistent unit when the aggregated posterior is close enough to the prior, i.e., $KL(p_{\boldsymbol{\theta}}(\mathbf{z}_i)\|q_{\boldsymbol{\phi}}(\mathbf{z}_i)) < 0.03$ means the i^{th} dimension is a consistent unit, which is proposed by us to quantify the severity of the hole problem in a dimension-wise perspective. A lower value of $CU(\boldsymbol{\phi})$ indicates a severer hole issue.

Among those metrics, $MI(\boldsymbol{\phi})$ has an upper bound of $\log N$, while $AU(\boldsymbol{\phi})$ and $CU(\boldsymbol{\phi})$ have an upper bound of Dim. Both the prior log likelihood $priorLL(\boldsymbol{\theta})$ and the posterior log likelihood $postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$ are the higher the better. A high value of $KL(\boldsymbol{\phi})$, $MI(\boldsymbol{\phi})$, $AU(\boldsymbol{\phi})$ or $CU(\boldsymbol{\phi})$ can not assure good performance, but too low a value of them can infer bad performance.

We approximate the inner expectation term of $priorLL(\boldsymbol{\theta})$ through importance weighted sampling [1], where S samples from the prior distribution $z_{s,prior} \overset{idd}{\sim} p_{\boldsymbol{\theta}}(\mathbf{z})$ and S samples from the posterior distribution $z_{s,post} \overset{idd}{\sim} q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ are used for Monte Carlo estimation:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{z}\sim p_{\boldsymbol{\theta}}(\mathbf{z})}[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] &= \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}) d\mathbf{z} \\
&= \int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \frac{p_{\boldsymbol{\theta}}(\mathbf{z})}{\frac{1}{2}(p_{\boldsymbol{\theta}}(\mathbf{z}) + q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}))} \frac{1}{2}(p_{\boldsymbol{\theta}}(\mathbf{z}) + q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})) d\mathbf{z} \\
&\approx \frac{1}{2S}\Big(\sum_{z_{s,prior}}^{S} p_{\boldsymbol{\theta}}(\mathbf{x}|z_{s,prior}) \frac{p_{\boldsymbol{\theta}}(z_{s,prior})}{\frac{1}{2}(p_{\boldsymbol{\theta}}(z_{s,prior}) + q_{\boldsymbol{\phi}}(z_{s,prior}|\mathbf{x}))} \\
&\quad + \sum_{z_{s,post}}^{S} p_{\boldsymbol{\theta}}(\mathbf{x}|z_{s,post}) \frac{p_{\boldsymbol{\theta}}(z_{s,post})}{\frac{1}{2}(p_{\boldsymbol{\theta}}(z_{s,post}) + q_{\boldsymbol{\phi}}(z_{s,post}|\mathbf{x}))} \Big)
\end{aligned}
\tag{1}
$$

Similarly, we also apply importance weighted sampling to approximate the inner expectation term of $postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$. We empirically set the sampling size $2S = 16$ and conducted evaluations for all models across 10 different random seeds under this setting and reported the mean values of $priorLL(\boldsymbol{\theta})$ and $postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$ at the precision of 0.1, where the variances are all less than 0.01.

D Full Results of Language Modeling

We illustrate the language modeling performance of all the Gaussian distribution-based VAEs we consider in table 1 (on Short-Yelp), table 2 (on SNLI), table 3 (on Yahoo) and table 4 (on Yelp).

D.1 On Short-Yelp dataset

Table 1: Full results of Language Modeling on Short-Yelp dataset. Here we bold up $MI(\boldsymbol{\phi}) \geq 9.0$, $AU(\boldsymbol{\phi}) \geq 30$ and $CU(\boldsymbol{\phi}) \geq 30$.

Models	$priorLL(\boldsymbol{\theta})$	$postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$	$KL(\boldsymbol{\phi})$	$MI(\boldsymbol{\phi})$	$AU(\boldsymbol{\phi})$	$CU(\boldsymbol{\phi})$
VAE (default)	-34.1	-33.1	0.9	0.8	3	**32**
cyclic-VAE	-34.0	-31.6	2.3	2.3	4	**32**
bow-VAE	-33.9	-31.4	2.6	2.6	3	**32**
skip-VAE	-33.9	-29.3	4.2	4.0	14	**32**
δ-VAE(0.15)	-35.0	-32.8	4.8	1.7	23	2
BN-VAE(0.6)	-33.9	-27.4	6.2	5.5	**32**	**32**
BN-VAE(0.7)	-34.0	-25.9	8.5	7.0	**32**	**32**
BN-VAE(0.9)	-34.8	-23.4	13.7	8.6	**32**	14
BN-VAE(1.2)	-37.3	-20.3	23.3	**9.0**	**32**	0
BN-VAE(1.5)	-42.6	-19.6	35.9	**9.1**	**32**	0
BN-VAE(1.8)	-47.9	-18.7	49.7	**9.1**	**32**	0
FB-VAE(4)	-33.9	-31.0	4.5	3.1	**32**	**31**
FB-VAE(9)	-31.5	-27.9	9.3	6.3	**32**	**31**
FB-VAE(16)	-30.1	-23.6	16.4	8.8	**32**	11
FB-VAE(25)	-32.2	-19.1	24.6	**9.1**	**32**	0
FB-VAE(36)	-38.7	-15.7	34.8	**9.1**	**32**	0
FB-VAE(49)	-46.6	-13.7	45.0	**9.1**	**32**	0
β-VAE(0.8)	-34.1	-30.3	4.0	4.0	3	**32**
β-VAE(0.4)	-36.1	-22.9	14.6	**9.0**	8	**30**
β-VAE(0.2)	-44.6	-13.7	34.8	**9.1**	21	**31**
β-VAE(0.1)	-54.7	-9.2	52.5	**9.1**	**32**	29
β-VAE(0.0)	-71.1	-10.3	147.5	**9.1**	**32**	0
DG-VAE ($\|b\| = 1$)	-34.1	-32.8	1.2	1.2	2	**32**
DG-VAE ($\|b\| = 2$)	-33.6	-26.8	8.0	7.3	8	**32**
DG-VAE ($\|b\| = 4$)	-35.0	-20.3	18.6	**9.1**	23	**32**
DG-VAE ($\|b\| = 8$)	-38.7	-14.2	34.8	**9.1**	**32**	**32**
DG-VAE ($\|b\| = 16$)	-41.2	-14.1	41.0	**9.1**	**32**	**32**
DG-VAE ($\|b\| = 32$)	-47.5	-11.2	53.1	**9.1**	**32**	**32**
DG-VAE (default)	-47.5	-11.2	53.1	**9.1**	**32**	**32**

D.2 On SNLI dataset

Table 2: Full results of Language Modeling on SNLI dataset. Here we bold up $MI(\boldsymbol{\phi}) \geq 9.0$, $AU(\boldsymbol{\phi}) \geq 30$ and $CU(\boldsymbol{\phi}) \geq 30$.

Models	$priorLL(\boldsymbol{\theta})$	$postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$	$KL(\boldsymbol{\phi})$	$MI(\boldsymbol{\phi})$	$AU(\boldsymbol{\phi})$	$CU(\boldsymbol{\phi})$
VAE (default)	-32.8	-31.2	1.3	1.3	2	**32**
cyclic-VAE	-32.7	-29.9	2.5	2.5	5	**32**
bow-VAE	-32.8	-30.8	2.0	2.0	2	**32**
skip-VAE	-32.7	-28.6	3.8	3.7	17	**32**
δ-VAE(0.15)	-33.6	-31.6	4.8	1.4	28	0
BN-VAE(0.6)	-32.6	-25.7	6.3	5.6	**32**	**32**
BN-VAE(0.7)	-32.6	-23.7	8.8	7.3	**32**	**32**
BN-VAE(0.9)	-32.8	-20.5	13.9	8.8	**32**	24
BN-VAE(1.2)	-36.5	-18.8	24.0	**9.2**	**32**	0
BN-VAE(1.5)	-41.2	-17.4	36.5	**9.2**	**32**	0
BN-VAE(1.8)	-47.1	-16.5	52.1	**9.2**	**32**	0
FB-VAE(4)	-32.6	-30.2	4.0	2.2	**32**	**32**
FB-VAE(9)	-30.4	-27.2	9.0	5.4	**32**	28
FB-VAE(16)	-28.3	-23.8	15.9	8.4	**32**	25
FB-VAE(25)	-28.6	-17.1	24.8	**9.2**	**32**	1
FB-VAE(36)	-35.0	-13.9	34.7	**9.2**	**32**	0
FB-VAE(49)	-43.0	-11.5	46.1	**9.2**	**32**	0
β-VAE(0.8)	-32.5	-27.1	5.8	5.6	5	**32**
β-VAE(0.4)	-35.2	-19.6	17.1	**9.2**	15	**31**
β-VAE(0.2)	-40.4	-13.7	30.8	**9.2**	23	**31**
β-VAE(0.1)	-46.7	-10.6	45.9	**9.2**	**30**	**30**
β-VAE(0.0)	-61.5	-9.1	138.3	**9.2**	**32**	0
DG-VAE ($\|b\| = 1$)	-32.8	-31.1	1.3	1.3	3	**32**
DG-VAE ($\|b\| = 2$)	-32.0	-25.4	8.0	7.3	8	**32**
DG-VAE ($\|b\| = 4$)	-33.3	-19.5	17.0	**9.2**	19	**32**
DG-VAE ($\|b\| = 8$)	-34.9	-15.0	28.6	**9.2**	**31**	**32**
DG-VAE ($\|b\| = 16$)	-38.9	-12.1	40.8	**9.2**	**31**	**32**
DG-VAE ($\|b\| = 32$)	-42.7	-11.1	48.6	**9.2**	**32**	**32**
DG-VAE (default)	-42.7	-11.1	48.6	**9.2**	**32**	**32**

D.3 On Yahoo dataset

Table 3: Full results of Language Modeling on Yahoo dataset. Here we bold up $MI(\boldsymbol{\phi}) \geq 9.0$, $AU(\boldsymbol{\phi}) \geq 30$ and $CU(\boldsymbol{\phi}) \geq 30$.

Models	$priorLL(\boldsymbol{\theta})$	$postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$	$KL(\boldsymbol{\phi})$	$MI(\boldsymbol{\phi})$	$AU(\boldsymbol{\phi})$	$CU(\boldsymbol{\phi})$
VAE (default)	-330.7	-330.7	0.0	0.0	0	**32**
cyclic-VAE	-329.9	-329.0	1.1	1.1	2	**31**
bow-VAE	-330.5	-330.5	0.0	0.0	0	**32**
skip-VAE	-330.2	-325.2	5.1	4.3	8	**31**
δ-VAE(0.15)	-330.5	-330.7	4.8	0.0	0	0
BN-VAE(0.6)	-327.6	-321.1	6.6	6.0	**32**	**32**
BN-VAE(0.7)	-326.8	-318.5	9.1	7.5	**32**	**32**
BN-VAE(0.9)	-327.1	-313.8	15.6	**9.0**	**32**	**32**
BN-VAE(1.2)	-330.9	-310.1	26.3	**9.2**	**32**	0
BN-VAE(1.5)	-337.8	-310.3	37.6	**9.2**	**32**	0
BN-VAE(1.8)	-343.6	-308.6	51.4	**9.2**	**32**	0
FB-VAE(4)	-329.8	-328.5	4.0	1.8	**32**	**32**
FB-VAE(9)	-327.9	-326.3	8.9	4.2	**32**	12
FB-VAE(16)	-325.8	-320.8	16.2	8.5	**32**	8
FB-VAE(25)	-333.5	-316.3	25.8	**9.2**	**32**	0
FB-VAE(36)	-341.3	-307.1	36.9	**9.2**	**32**	0
FB-VAE(49)	-344.7	-296.1	50.1	**9.2**	**32**	0
β-VAE(0.8)	-330.2	-328.5	2.0	1.9	2	**30**
β-VAE(0.4)	-330.9	-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)	-370.0	-289.6	83.7	**9.2**	**32**	0
β-VAE(0.0)	-445.3	-280.4	178.8	**9.2**	**32**	0
DG-VAE ($\|b\| = 1$)	-330.7	-330.8	0.0	-0.0	0	**32**
DG-VAE ($\|b\| = 2$)	-330.1	-326.5	4.1	4.1	4	**32**
DG-VAE ($\|b\| = 4$)	-330.4	-318.3	14.4	**9.1**	11	**32**
DG-VAE ($\|b\| = 8$)	-338.3	-308.3	32.1	**9.2**	**30**	**32**
DG-VAE ($\|b\| = 16$)	-349.5	-295.1	57.7	**9.2**	**32**	**32**
DG-VAE ($\|b\| = 32$)	-355.4	-294.1	65.2	**9.2**	**32**	**32**
DG-VAE (default)	-358.0	-290.9	70.8	**9.2**	**32**	**32**

D.4 On Yelp dataset

As depicted in table 4, it can be observed that BN-VAEs perform abnormally on Yelp dataset when $\gamma \geq 1.2$. We investigate this phenomenon and find that their batch normalization layers have already crashed in training.

Normally, the batch normalization layer performs the following operation for input x, where ϵ is a small value to avoid division by zero:

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta \tag{2}$$

However, in those BN-VAEs that perform abnormally:

$$\mathrm{Var}[x] \ll \epsilon \tag{3}$$

So, their batch normalization layers can no longer fix the variance of output y to γ, and thus they can no longer ensure a lower bound of the KL term in ELBo.

Intuitively, the KL term in ELBo tends to minimize $\mathrm{Var}[y]$ while the batch normalization layer persists in fixing $\mathrm{Var}[y]$ to γ. So they finally minimize $\mathrm{Var}[x]$ and lead to this phenomenon on Yelp dataset when $\gamma \geq 1.2$.

Table 4: Full results of Language Modeling on Yelp dataset. Here we bold up $MI(\boldsymbol{\phi}) \geq 9.0$, $AU(\boldsymbol{\phi}) \geq 30$ and $CU(\boldsymbol{\phi}) \geq 30$.

Models	$priorLL(\boldsymbol{\theta})$	$postLL(\boldsymbol{\theta}, \boldsymbol{\phi})$	$KL(\boldsymbol{\phi})$	$MI(\boldsymbol{\phi})$	$AU(\boldsymbol{\phi})$	$CU(\boldsymbol{\phi})$
VAE (default)	-360.2	-360.2	0.1	0.0	0	**32**
cyclic-VAE	-358.9	-358.2	0.5	0.5	2	**32**
bow-VAE	-359.2	-358.8	0.3	0.3	1	**32**
skip-VAE	-359.8	-356.6	3.2	2.5	4	**30**
δ-VAE(0.15)	-359.4	-359.6	4.8	0.0	0	0
BN-VAE(0.6)	-356.5	-349.6	7.4	6.1	**32**	**32**
BN-VAE(0.7)	-356.6	-347.8	10.0	7.7	**32**	**31**
BN-VAE(0.9)	-356.5	-343.8	15.8	**9.0**	**32**	25
BN-VAE(1.2)	-362.0	-357.7	7.2	4.0	28	28
BN-VAE(1.5)	-359.8	-357.5	4.0	1.7	22	**30**
BN-VAE(1.8)	-365.5	-360.4	11.3	4.5	**30**	17
FB-VAE(4)	-358.6	-357.3	4.0	1.8	**32**	**32**
FB-VAE(9)	-358.4	-357.4	8.8	2.8	**32**	0
FB-VAE(16)	-355.3	-351.3	16.1	7.9	**32**	13
FB-VAE(25)	-367.9	-355.2	24.3	**9.1**	**32**	0
FB-VAE(36)	-368.8	-338.0	36.6	**9.2**	**32**	0
FB-VAE(49)	-375.1	-329.1	48.9	**9.2**	**32**	0
β-VAE(0.8)	-358.8	-357.4	1.7	1.7	2	**31**
β-VAE(0.4)	-359.5	-353.9	6.6	6.2	3	**31**
β-VAE(0.2)	-366.9	-344.1	24.8	**9.2**	17	0
β-VAE(0.1)	-376.1	-336.8	42.0	**9.2**	24	0
β-VAE(0.0)	-483.3	-309.4	190.6	**9.2**	**32**	0
DG-VAE ($\|b\| = 1$)	-359.3	-358.9	0.3	0.3	1	**32**
DG-VAE ($\|b\| = 2$)	-361.3	-359.0	2.9	2.7	4	**31**
DG-VAE ($\|b\| = 4$)	-359.3	-351.0	9.9	8.4	7	**32**
DG-VAE ($\|b\| = 8$)	-362.7	-344.1	20.9	**9.1**	**30**	**32**
DG-VAE ($\|b\| = 16$)	-368.1	-337.8	33.4	**9.1**	**31**	**32**
DG-VAE ($\|b\| = 32$)	-378.2	-331.0	51.2	**9.1**	**31**	**32**
DG-VAE (default)	-381.8	-324.6	62.4	**9.1**	**32**	**31**

E Interpolation Rouge-L F1-score curves

E.1 On Short-Yelp dataset

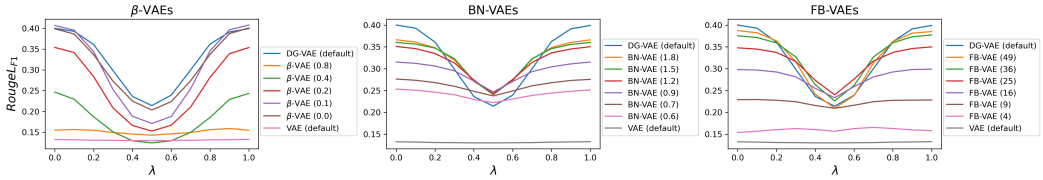

Figure 2: The curves of Rouge-L F1-score and λ for models' interpolation performance on Short-Yelp.

E.2 On SNLI dataset

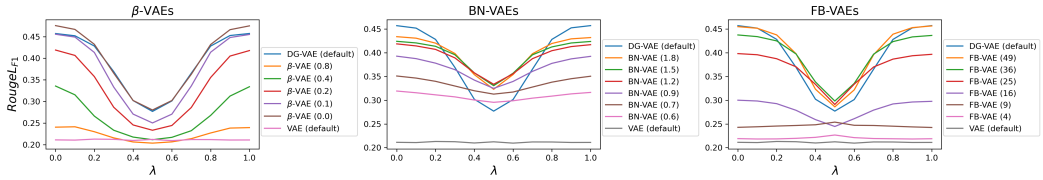

Figure 3: The curves of Rouge-L F1-score and λ for models' interpolation performance on SNLI.

It can be observed in Figure 2 and Figure 3 that although DG-VAE outperforms β-VAEs on short datasets under nearly all conditions, BN-VAEs and FB-VAEs with proper parameter settings outperform DG-VAE when $\lambda \approx 0.5$. We think this is due to the capacity of DG-VAE may be too big for short sentences, as it maximizes the mutual information between the input sentences and the latent variables on 32 dimensions respectively while those short sentences contain only about 10 tokens on average.

So, for such short datasets, BN-VAEs and FB-VAEs with proper parameter settings may be better choices for latent-guided generation.

E.3 On Yahoo dataset

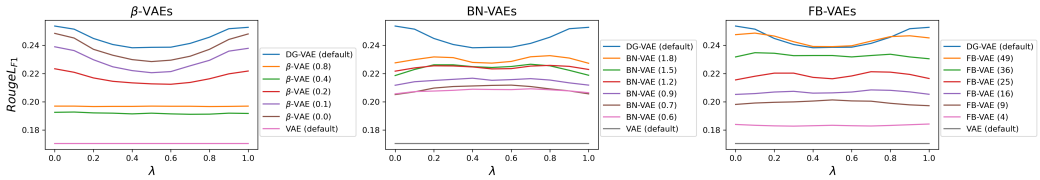

Figure 4: The curves of Rouge-L F1-score and λ for models' interpolation performance on Yahoo.

E.4 On Yelp dataset

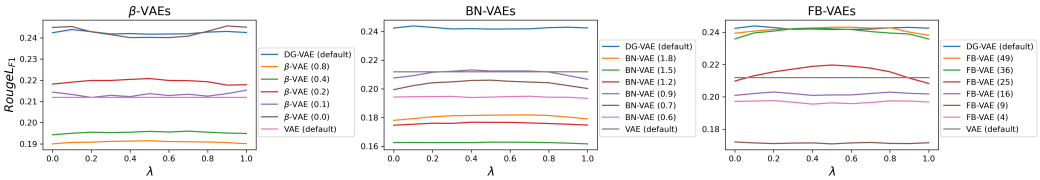

Figure 5: The curves of Rouge-L F1-score and λ for models' interpolation performance on Yelp.

It can be observed in Figure 4 and Figure 5 that DG-VAE outperforms all the other models under nearly all conditions on long datasets, which suggests that DG-VAE can be an excellent choice for latent-guided text generation for long datasets.

F Interpolation Case Study

As β-VAE(0.1) performs the most competitively according to the previous evaluation, here we do case study for interpolation of β-VAE(0.1) and our proposed DG-VAE. For each sentence pair, $\mathbf{x}_a$ and $\mathbf{x}_b$, we report the generation results $\mathbf{x}_\lambda$ and the corresponding values of density gap $DG(z_\lambda)$ (short for $DG(\boldsymbol{\theta}, \boldsymbol{\phi}; z_\lambda)$) along with λ, as shown in Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11. We highlight the tokens in the longest common subsequences in yellow (with $\mathbf{x}_a$), blue (with $\mathbf{x}_b$) or green (with both).

It can be observed that the values of $DG(z_\lambda)$ tend to be more negative in the middle, i.e., $\lambda \approx 0.5$. This indicates that the aggregated posterior distribution (where z is sampling from during fitting) has much lower probability density than the prior distribution (where z is sampling from during unconditional inference) does in the middle.

Compared with β-VAE(0.1), our proposed DG-VAE has less negative $DG(z_\lambda)$, and thus provides smoother interpolation results.

	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 .

Figure 6: The (short) interpolation case of β-VAE(0.1) and DG-VAE on Short-Yelp dataset.

	x_a:	two girls walking in a park .	x_b:	the two kids are playing in water
		β-VAE(0.1)		DG-VAE
λ	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
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

Figure 7: The (short) interpolation case of β-VAE(0.1) and DG-VAE on SNLI dataset.

	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 !

Figure 8: The (long) interpolation case of β-VAE(0.1) and DG-VAE on Short-Yelp dataset.

	x_a:	a man in a white shirt and black pants poses in front of a large banner .	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	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
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 .

Figure 9: The (long) interpolation case of β-VAE(0.1) and DG-VAE on SNLI dataset.

	x_a:	do you like marilyn monroe and why ? i think marilyn monroe is so beautiful a great actress and she was so _UNK she spoke her _UNK just want to know if you like her and why ? ? ! ! yes i love her . she was beautiful , sexy and a little mysterious too .	x_b:	where can i find a book full information about religions ? _UNK _UNK [_UNK] wrote a series of three books called " a history of religion " that covers just about every aspect of all of the major religions and treats upon all sorts of minor ones all the way back to prehistoric religions .
	β-VAE(0.1)		DG-VAE	
λ	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
0.0	91.2	do you like ashley parker or what ? ? i love her and she has a great voice and _UNK , but she 's not really good at her _UNK she 's not really good for her but she 's not really good . she is a great singer and she 's hot too .	75.9	do you think girls are sexy in paris ? i mean , they are just a bunch of idiots who are _UNK and _UNK and they do n't want to wear hair ! ! ! ? i think it 's a fad . you are better than wearing thongs and sexy .
0.2	-302.3	do you think she 's hot ? she is a cutie and she is hot and she is hot and i am _UNK she is hot and i 'm not sure what she is doing in her room ? she is a hot _UNK . i have been friends with her since she was 18 , but she 's hot and she sucks .	-49.6	do you think angelina jolie is hot ? i 'm a huge fan of her and i do n't know what her _UNK is _UNK is her name ? she is very talented and she is hot ! ! ! she is not a real actress . she is a very beautiful actress , but she has a lot of talent ... .
0.4	-1453.8	why do i say goodbye my girlfriend ? a friend of mine is mad at me and she is _UNK to me and she 's mad at me ? she 's a _UNK , but she has been playing for over 30 years . she is the only one who knows how to spell and play for her own time .	-404.3	do you think jack bauer is a good actor ? i 'm a fan of _UNK and _UNK and he 's a great actor ... but i 'm not sure what _UNK is ? tom cruise is a great actor ... but he is a great actor and he is very good . he is the best actor and he is very intelligent and clever , and is very talented .
0.6	-1016.9	can you help me with my boyfriend ? he is an _UNK girl and he is in a wheelchair and he is _UNK and he is working on his feet ! ! what is your opinion on this one and the other person has to be with his feet with his hands .	-318.1	where can i find a biography on this topic ? i 'm looking for a _UNK _UNK ' story of _UNK and a man who is a christian ... .. what is a good book to read about ? the bible is not a religious belief , but there are many books that are written by books and books .
0.8	-187.8	how do i become a successful friend ? 1) become _UNK and b) _UNK the person with his friends 4) a friend who is in his class and he is a good friend 2) _UNK 2) work with your friends and talk with the person , talk to your friends and talk about your family .	-23.4	where can i find a biography of this man ? this is a _UNK _UNK of _UNK 's syndrome and his son 's son is a doctor who has been diagnosed with this disease and the symptoms are very helpful to him in his quest to help .
1.0	90.8	where can i find a list of a good christian youth in _UNK ? the national association of _UNK 's church has a lot of great friends and family members and family members and _UNK groups are very close to the public and also includes a variety of activities and services (for example) and other places .	79.2	where can i find a list of a good scientist ? _UNK _UNK [_UNK] is a fictional character of a fictional character who has been used to describe the author of a fictional character that has been used to describe the character of a person who has been unable to achieve his own greatness .

Figure 10: The interpolation case of β-VAE(0.1) and DG-VAE on Yahoo dataset.

	x_a:	i love this place ! i usually get an iced soy white mocha , which is love ! i also had a soy chai blended ! to _UNK for ! love it ! if you are in downtown glendale make this your go to place for coffee !	x_b:	i 'm not a big fan of '' coffee '' but i do enjoy smoothies and iced drinks , which are n't bad here , but starbucks still has my heart . i do enjoy their free wi-fi , and it 's a very calm atmosphere . -rrb-
	β-VAE(0.1)		DG-VAE	
λ	$DG(z_\lambda)$	x_λ	$DG(z_\lambda)$	x_λ
0.0	35.3	i love this place ! it 's a great place to go for a quick bite ! i love the _UNK , the _UNK and the _UNK ! they have a great selection of beer ! great place to hang out with friends ! i love the _UNK and _UNK !	63.4	i love this place ! i love the fact that they have a variety of flavors ! i love the fact that they have a variety of toppings ! also , they have a bakery ! i 'll definitely be coming back here for sure !
0.2	-131.2	i love this place ! it 's a little pricey but it 's worth it ! i love the fact that they have a lot of different flavors , including the _UNK ! i love the _UNK and the chocolate chip !	6.4	i love this place ! i have been here a few times , and the coffee is always good ! i 've had the _UNK _UNK & it 's always good . if you 're looking for a quick bite to eat at starbucks i would recommend stopping by here for a quick bite to eat !
0.4	-457.4	i have been to this location a few times ... it 's always clean and the staff is friendly ! i 've never had a bad experience here , and it 's always great . i 've been here a few times and it 's always a great experience and i love it !	-152.9	i love this place ! i have n't tried the coffee yet , but i have n't tried the coffee yet but i love it . a little pricey but hey ... it 's worth it ! if you are looking for a coffee shop this is the place to go !
0.6	-517.9	i am a big fan of the _UNK ! i 've been here a few times and it 's always been a good time to go . the staff is friendly , and the staff is friendly . i 've been here a few times and it 's always a great time ! i 'll be back soon ...	-281.5	i 'm a big fan of the coffee bean and i love this place ! i like the fact that they have a lot of options . but , i do n't like the fact that they do n't accept credit cards . i 'll stick to the starbucks on the weekends , but they have a starbucks for $ 5 !
0.8	-354.1	i 'm not a big fan of the _UNK -lrb- _UNK -rrb- and the _UNK _UNK -lrb- _UNK -rrb- and _UNK . i love the _UNK and _UNK . the _UNK is a great place to go , and enjoy a movie or two ...	-21.5	i 'm not a big fan of starbucks but i love their coffee and i 've never had a bad experience . coffee is good , but i prefer starbucks . it 's a little pricey but you can get a good cup of coffee here ! . d
1.0	42.3	i 'm not a big fan of the _UNK -lrb- _UNK -rrb- and the _UNK _UNK _UNK . i 've had better luck . the _UNK is a great place . i love the fact that they have a _UNK . _UNK and _UNK _UNK	68.8	i 'm not a big fan of starbucks but i love the coffee and coffee here . but i do n't think it 's worth it . but , it 's a starbucks . it 's a starbucks and a starbucks in the middle of the strip . -rrb-

Figure 11: The interpolation case of β-VAE(0.1) and DG-VAE on Yelp dataset.

G Latent Space Visualization

We visualize the latent space for a model through the following two steps:

(1) Rank the 32 dimensions by the marginal variance of posterior centers, i.e. $Var_{\mathbf{x}\sim\mathbf{X}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{z}_i]]$ for the i^{th} dimension, from low to high, which in essence ranks the dimensions from inactive to active.

(2) Visualize the aggregated posterior distribution (red-in-black) and the posterior centers (blue-in-white) on a group of two adjacent dimensions. Here we illustrate the results on dimensions ranked the 0th paired with the 1st, the 6th paired with the 7th, the 12th paired with the 13th, the 18th paired with the 19th, the 24th paired with the 25th, and the 30th paired with the 31st.

G.1 On Yahoo dataset

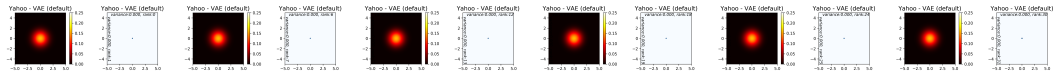

Figure 12: The latent space visualization of VAE (default) on Yahoo dataset.

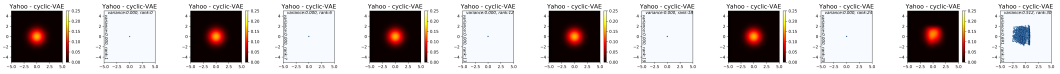

Figure 13: The latent space visualization of cyclic-VAE on Yahoo dataset.

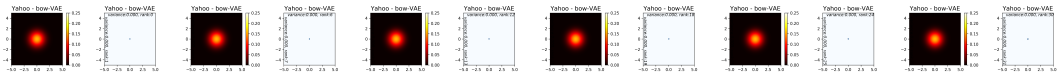

Figure 14: The latent space visualization of bow-VAE on Yahoo dataset.

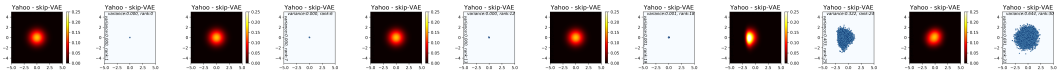

Figure 15: The latent space visualization of skip-VAE on Yahoo dataset.

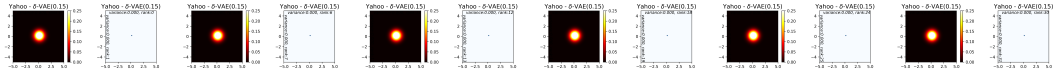

Figure 16: The latent space visualization of delta-VAE on Yahoo dataset.

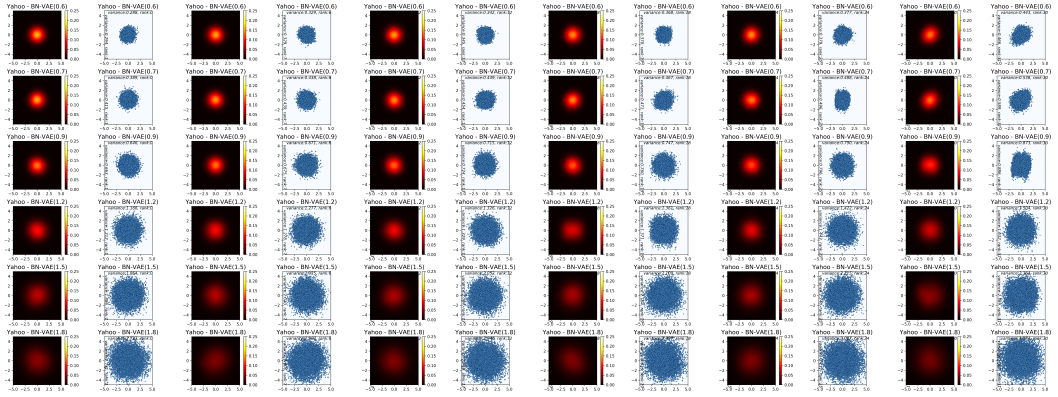

Figure 17: The latent space visualization of BN-VAEs on Yahoo dataset.

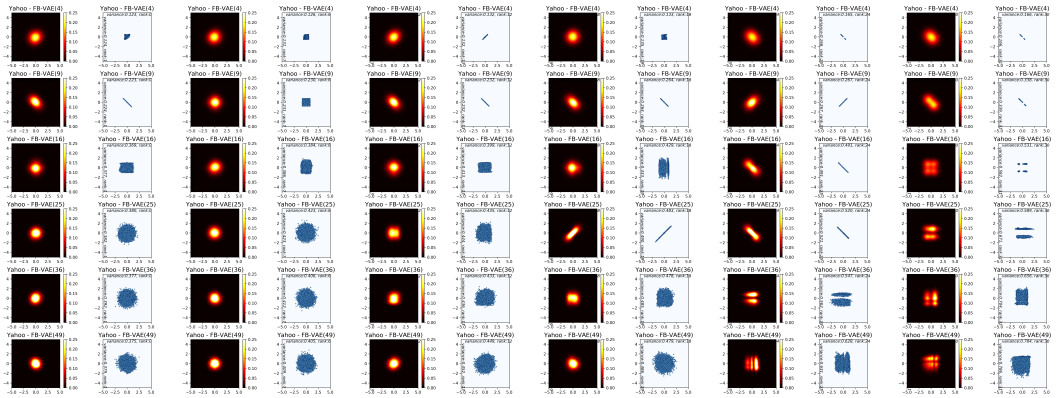

Figure 18: The latent space visualization of FB-VAEs on Yahoo dataset.

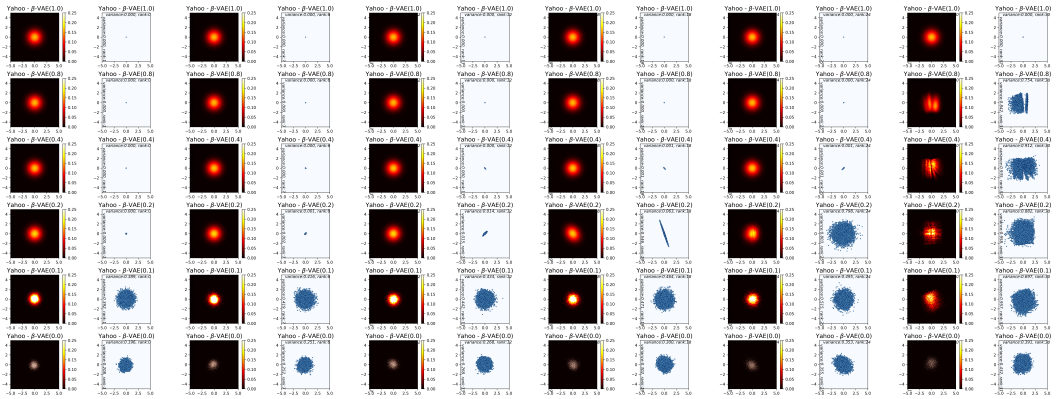

Figure 19: The latent space visualization of Beta-VAEs on Yahoo dataset.

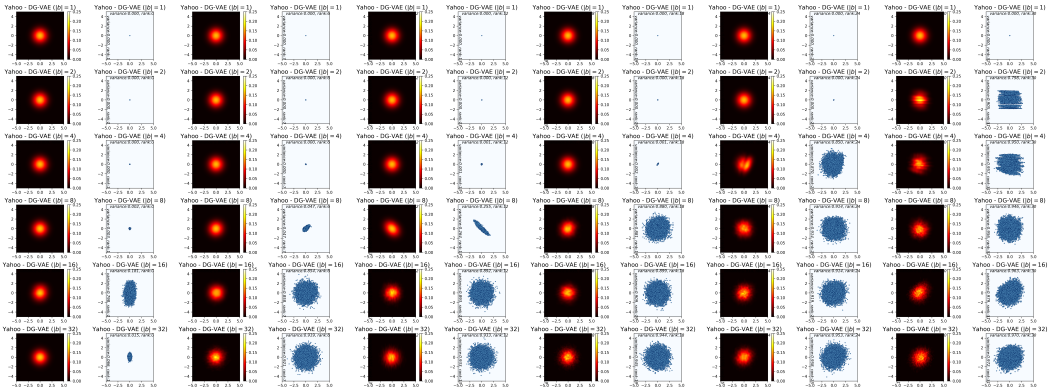

Figure 20: The latent space visualization of DG-VAEs on Yahoo dataset.

G.2 On Yelp dataset

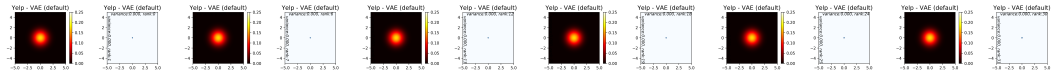

Figure 21: The latent space visualization of VAE (default) on Yelp dataset.

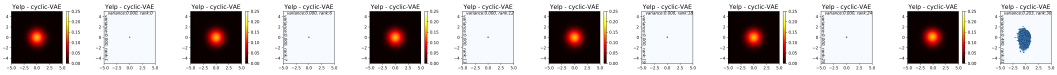

Figure 22: The latent space visualization of cyclic-VAE on Yelp dataset.

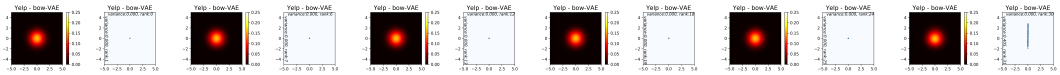

Figure 23: The latent space visualization of bow-VAE on Yelp dataset.

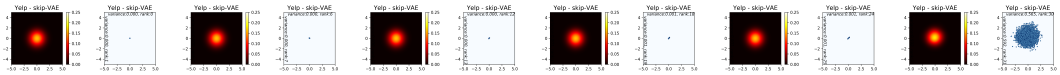

Figure 24: The latent space visualization of skip-VAE on Yelp dataset.

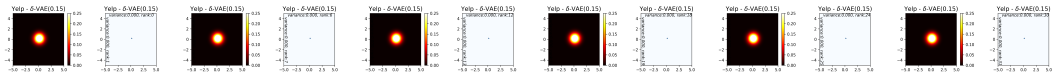

Figure 25: The latent space visualization of delta-VAE on Yelp dataset.

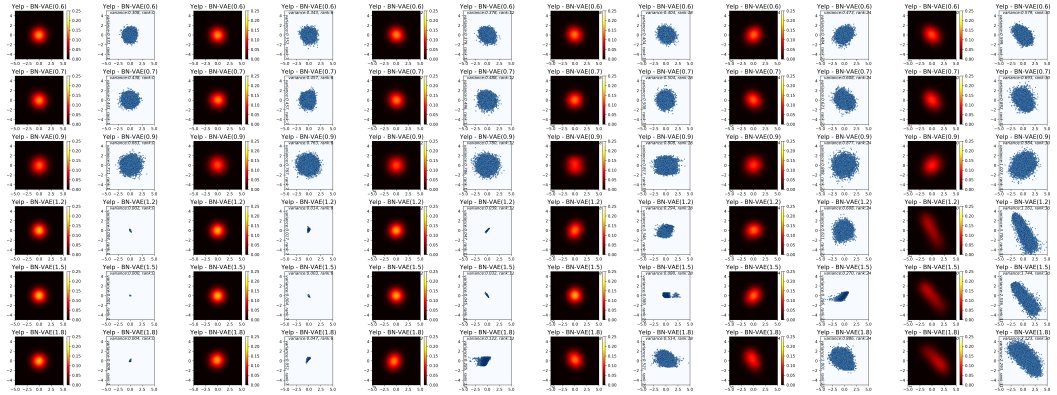

Figure 26: The latent space visualization of BN-VAEs on Yelp dataset.

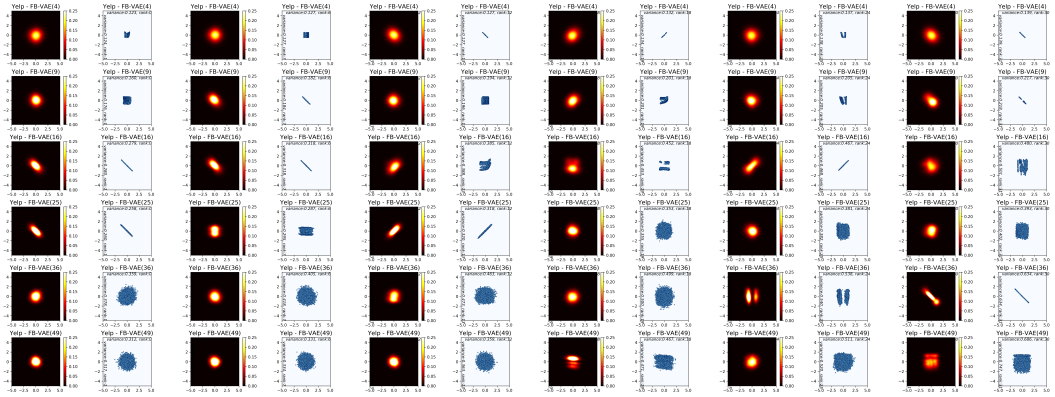

Figure 27: The latent space visualization of FB-VAEs on Yelp dataset.

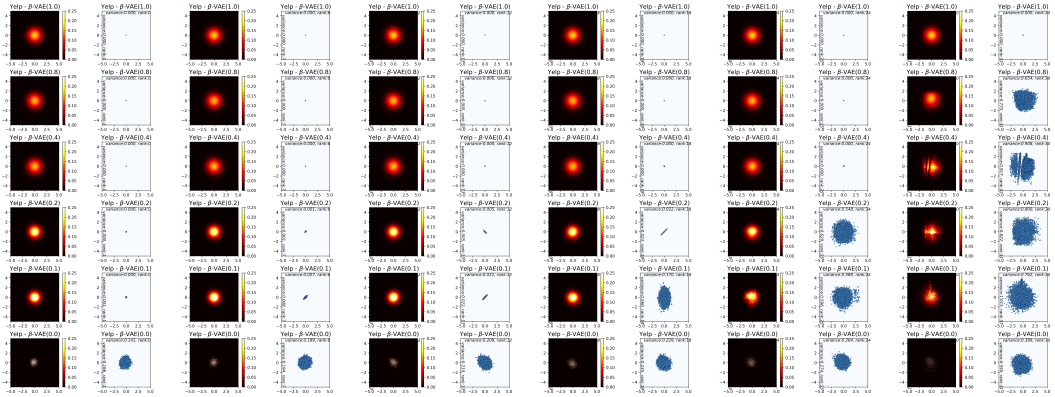

Figure 28: The latent space visualization of Beta-VAEs on Yelp dataset.

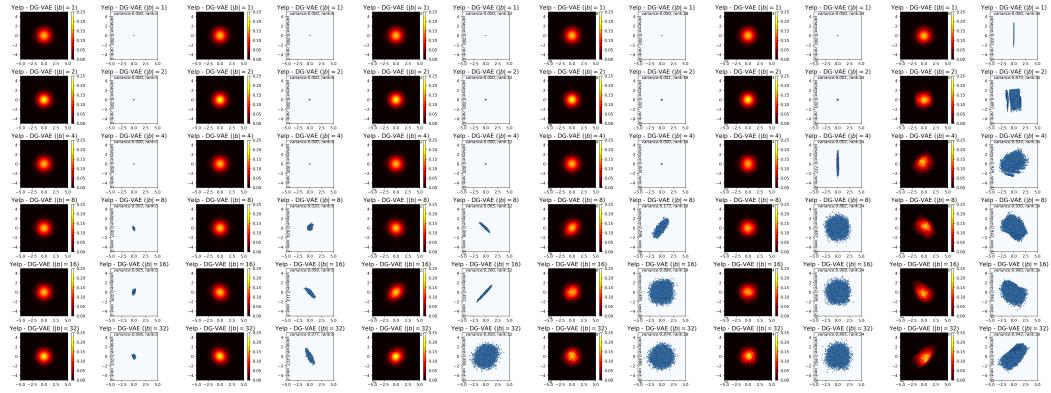

Figure 29: The latent space visualization of DG-VAEs on Yelp dataset.

G.3 On SNLI dataset

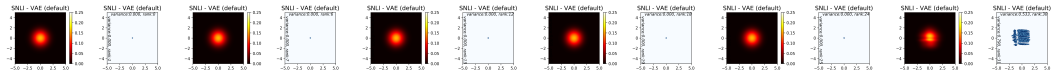

Figure 30: The latent space visualization of VAE (default) on SNLI dataset.

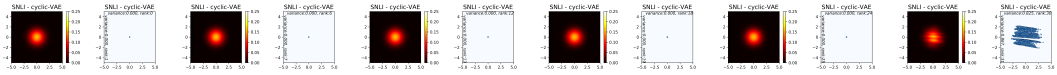

Figure 31: The latent space visualization of cyclic-VAE on SNLI dataset.

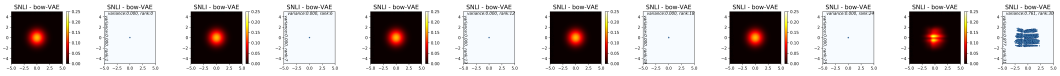

Figure 32: The latent space visualization of bow-VAE on SNLI dataset.

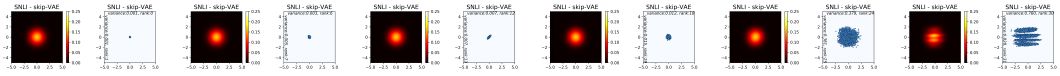

Figure 33: The latent space visualization of skip-VAE on SNLI dataset.

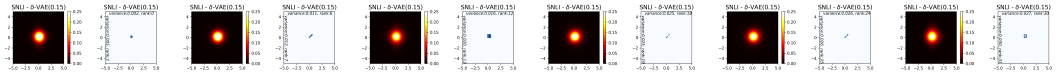

Figure 34: The latent space visualization of delta-VAE on SNLI dataset.

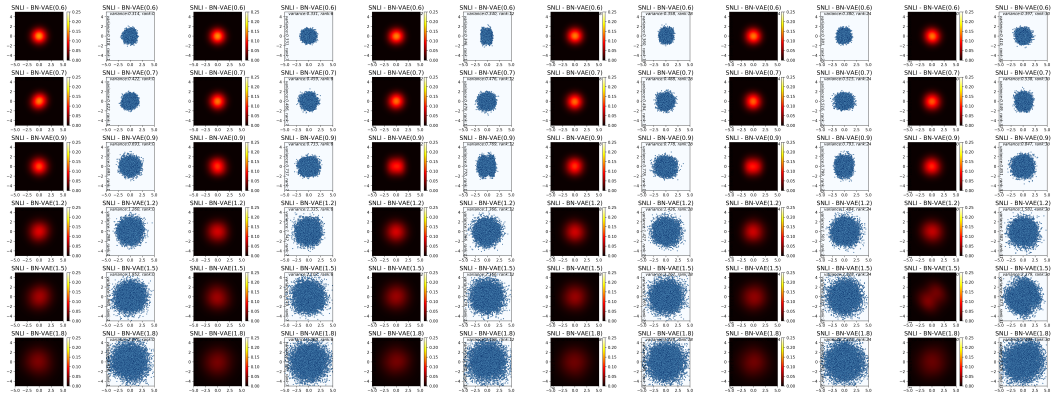

Figure 35: The latent space visualization of BN-VAEs on SNLI dataset.

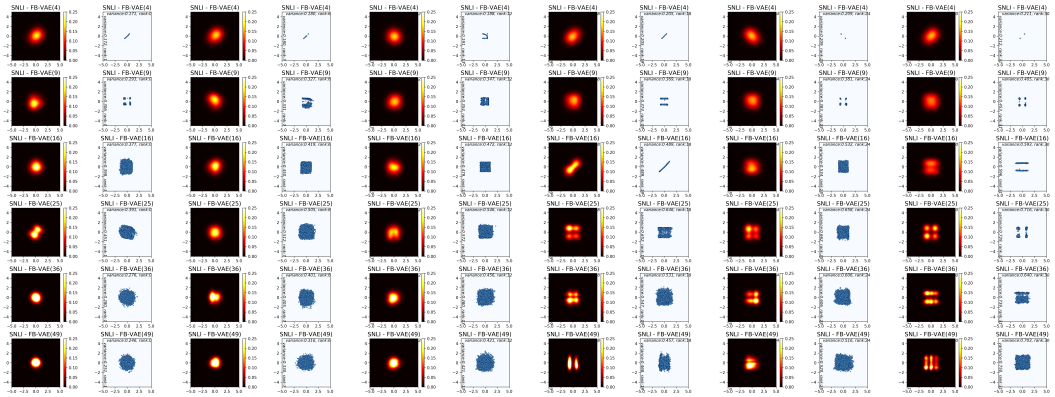

Figure 36: The latent space visualization of FB-VAEs on SNLI dataset.

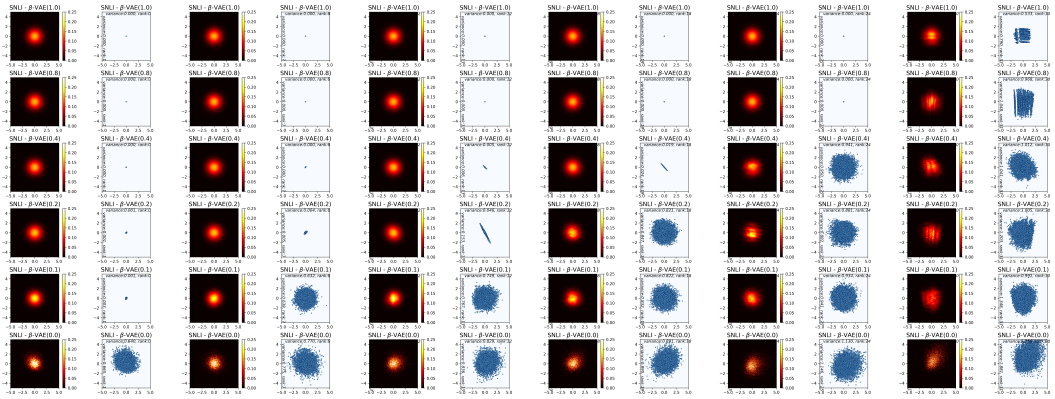

Figure 37: The latent space visualization of Beta-VAEs on SNLI dataset.

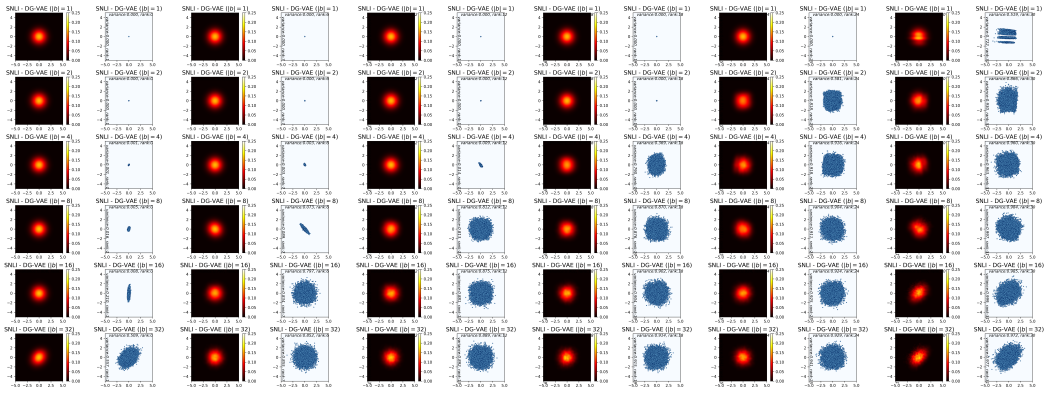

Figure 38: The latent space visualization of DG-VAEs on SNLI dataset.

G.4 On Short-Yelp dataset

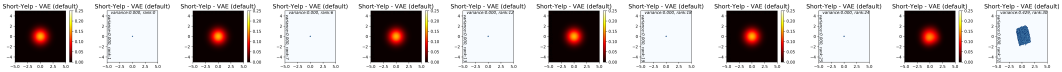

Figure 39: The latent space visualization of VAE (default) on Short-Yelp dataset.

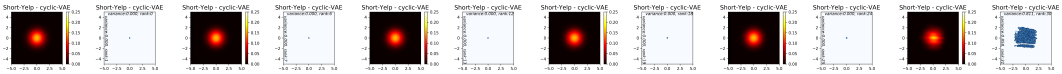

Figure 40: The latent space visualization of cyclic-VAE on Short-Yelp dataset.

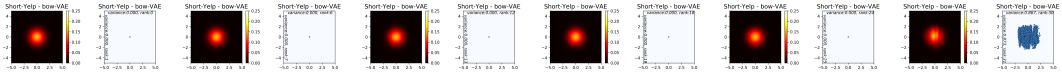

Figure 41: The latent space visualization of bow-VAE on Short-Yelp dataset.

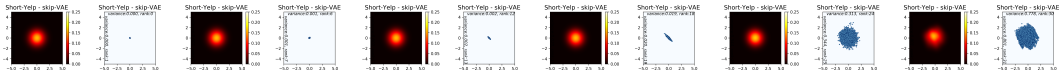

Figure 42: The latent space visualization of skip-VAE on Short-Yelp dataset.

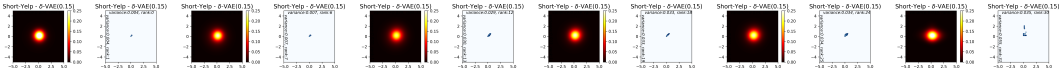

Figure 43: The latent space visualization of delta-VAE on Short-Yelp dataset.

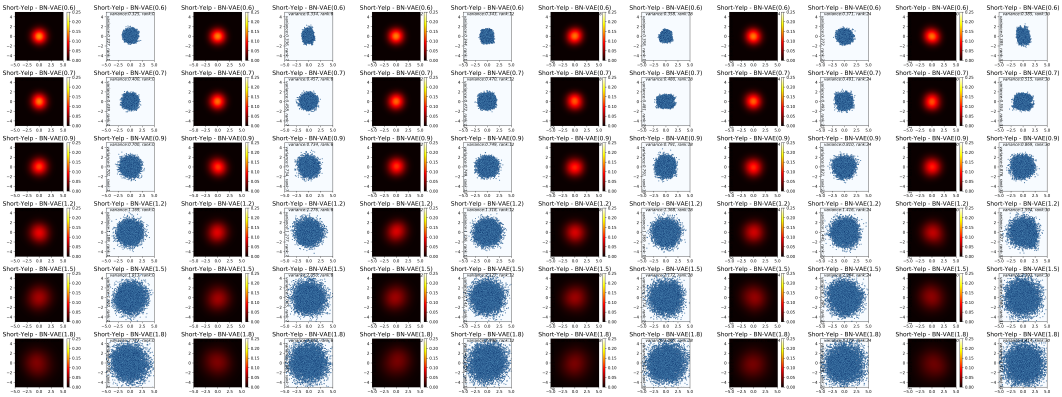

Figure 44: The latent space visualization of BN-VAEs on Short-Yelp dataset.

In conclusion, cyclic-VAEs (depicted in Figures 13, 22, 31 and 40), bow-VAEs (depicted in Figures 14, 23, 32 and 41), skip-VAEs (depicted in Figures 15, 24, 33 and 42) and δ-VAEs (depicted in Figures 16, 25, 34 and 43) have limited effect on solving posterior collapse as most of their dimensions are still inactive (according to the posterior centers distributions).

Meanwhile, FB-VAEs (depicted in Figures 18, 27, 36 and 45) and β-VAEs (depicted in Figures 19, 28, 37 and 46) can solve posterior collapse effectively through weakening the KL term in ELBo by a large margin, e.g., FB-VAE(49) or β-VAE(0.1), but they also introduce mismatch between the aggregated posterior and the prior through doing so.

According to the visualization, BN-VAEs (depicted in Figures 17, 26, 35 and 44) can form a latent space without posterior collapse or significant hole problem with a proper γ, e.g., $\gamma = 0.6$, but they indeed perform poorly on latent-guided generation in such circumstances according to experiments

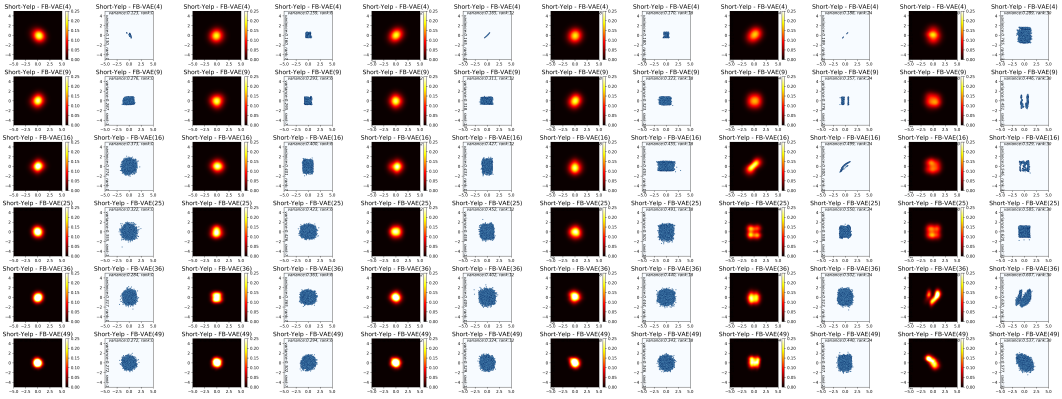

Figure 45: The latent space visualization of FB-VAEs on Short-Yelp dataset.

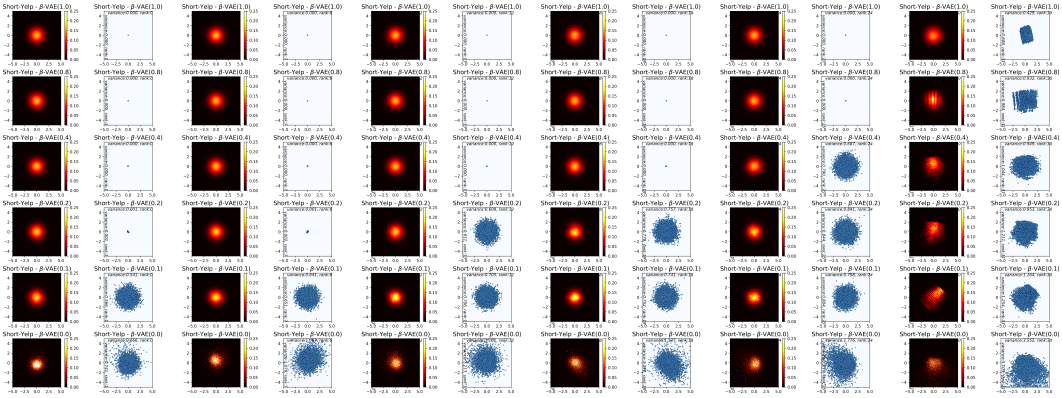

Figure 46: The latent space visualization of Beta-VAEs on Short-Yelp dataset.

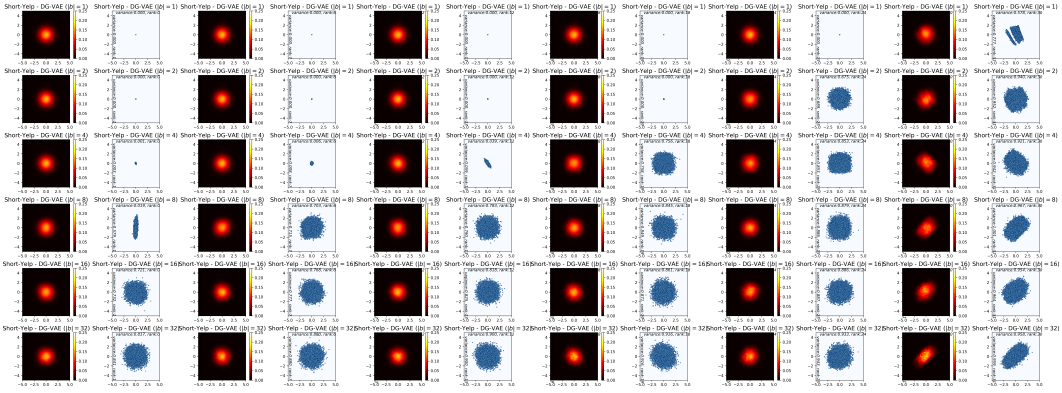

Figure 47: The latent space visualization of DG-VAEs on Short-Yelp dataset.

on language modeling and interpolation. With the increase of γ, BN-VAEs also introduce mismatch between the aggregated posterior and the prior. Moreover, it can be observed in Figure 26 that BN-VAEs with high values of γ perform unsteadily on Yelp dataset, as we discuss and explain in Appendix D.4.

In contrast, DG-VAEs (depicted in Figures 20, 29, 38 and 47) can gradually solve posterior collapse with the increase of $|b|$, and avoid the mismatch between the aggregated posterior and the prior throughout the process.

References

[1] Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[3] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015.

[4] Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein autoencoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.