Regularizing Variational Autoencoder with Diversity and Uncertainty Awareness

Dazhong Shen[1,2] , **Chuan Qin**[2] , **Chao Wang**[1,2] , **Hengshu Zhu**[2,*] , **Enhong Chen**[1] , **Hui Xiong**[3,*]
[1] School of Computer Science and Technology, University of Science and Technology of China
[2]Baidu Talent Intelligence Center
[3]Rutgers, The State University of New Jersey
sdz@mail.ustc.edu.cn, chuanqin0426@gmail.com, wdyx2012@mail.ustc.edu.cn,
zhuhengshu@baidu.com, cheneh@ustc.edu.cn, hxiong@rutgers.edu

Abstract

As one of the most popular generative models, Variational Autoencoder (VAE) approximates the posterior of latent variables based on amortized variational inference. However, when the decoder network is sufficiently expressive, VAE may lead to *posterior collapse*; that is, uninformative latent representations may be learned. To this end, in this paper, we propose an alternative model, DU-VAE, for learning a more ***D****iverse* and less ***U****ncertain* latent space, and thus the representation can be learned in a meaningful and compact manner. Specifically, we first theoretically demonstrate that it will result in better latent space with high diversity and low uncertainty awareness by controlling the distribution of posterior's parameters across the whole data accordingly. Then, without the introduction of new loss terms or modifying training strategies, we propose to exploit Dropout on the variances and Batch-Normalization on the means simultaneously to regularize their distributions implicitly. Furthermore, to evaluate the generalization effect, we also exploit DU-VAE for inverse autoregressive flow based-VAE (VAE-IAF) empirically. Finally, extensive experiments on three benchmark datasets clearly show that our approach can outperform state-of-the-art baselines on both likelihood estimation and underlying classification tasks.

1 Introduction

Recent years have witnessed the great success of Variational Autoencoder (VAE) [Kingma and Welling, 2013] as a generative model for representation learning, which has been widely exploited in various challenging domains, such as natural language modeling and image processing [Bowman *et al.*, 2015b; Pu *et al.*, 2016]. Indeed, VAE models the generative process of observed data by defining a joint distribution with latent space, and approximates the posterior of latent variables based on the amortized variational inference. While the use of VAE has been well-recognized, it may lead to uninformative latent representations, particularly when the expressive and powerful decoder networks are employed, such as LSTMs [Hochreiter and Schmidhuber, 1997] on text or PixelCNN [Van den Oord *et al.*, 2016] on images. This is widely known as the *posterior collapse* phenomenon [Zhao *et al.*, 2019]. In other words, the model may fail to diversify the posteriors of different data by simply using the single posterior distribution component to model all data instances. Also, the traditional VAE model usually produces the redundant information of representation due to the lack of guidance to characterize posterior space [Bowman *et al.*, 2015a; Chen *et al.*, 2017]. Therefore, the learned representation of VAE often results in an unsatisfied performance for downstream tasks, such as classification, even if it can approximate the marginal likelihood of observed data very well.

In the literature, tremendous efforts have been made for improving the representation learning of VAE and alleviating the problem of posterior collapse. One thread of these works is to attribute the posterior collapse to optimization challenges of VAEs and design various strategies, including KL annealing [Bowman *et al.*, 2015a; Fu *et al.*, 2019], Free-Bits(FB) [Kingma *et al.*, 2016], aggressive training [He *et al.*, 2018], encoder network pretraining and decoder network weakening [Yang *et al.*, 2017]. Among them, BN-VAE [Zhu *et al.*, 2020] applies the Batch-Normalization (BN) [Ioffe and Szegedy, 2015] to ensure one positive lower bound of the KL term. However, the theoretical basis of the effectiveness of BN on latent space learning is not yet understood, and more possible explanations based on the geometry analysis of latent space are needed. Other studies attempt to modify the objective carefully to direct the latent space learning [Makhzani *et al.*, 2016; Zheng *et al.*, 2019]. One feasible direction is to add additional Mutual Information (MI) based term to enhance the relation between data and latent space. However, due to the intractability, additional designs are always required for approximating MI-based objectives [Fang *et al.*, 2019; Zhao *et al.*, 2019]. Recently, Mutual Posterior-Divergence (MPD) [Ma *et al.*, 2018] is introduced to measure the diversity of the latent space, which is analytic and has one similar goal with MI. However, the scales of MPD and original objective are unbalanced, which requires deliberate normalization.

In this paper, to improve the representation learning performances of VAE, we propose a novel generative model, DU-

*This work was done when Dazhong Shen was an intern at Talent Intelligent Center, Baidu Inc. Hui Xiong and Hengshu Zhu are the corresponding authors.

VAE, for learning a more ***Diverse*** and less ***Uncertain*** latent space, and thus ensures the representation can be learned in a meaningful and compact manner. To be specific, we first analyze the expected latent space theoretically from two geometry properties, diversity and uncertainty, based on the MPD and Conditional Entropy (CE) metrics, respectively. We demonstrate that it will lead to a better latent space with high diversity and low uncertainty by controlling the distribution of posterior's parameters across the whole data. Then, instead of introducing new loss terms or modifying training strategies, we propose to apply Dropout [Srivastava *et al.*, 2014] on the variances and Batch-Normalization on the means simultaneously to regularize their distributions implicitly. In particular, we also discuss and prove the effectiveness of two regularizations in a rigorous way. Furthermore, to verify the generalization of our approaches, we also demonstrate that DU-VAE can be extended empirically into VAE-IAF [Kingma *et al.*, 2016], a well-known normalizing flow-based VAE. Finally, extensive experiments have been conducted on three benchmark datasets, and the results clearly show that our approach can outperform state-of-the-art baselines on both likelihood estimation and classification tasks.

2 Background of VAE

Given the input space $x \in \mathcal{X}$, VAE aims to construct a smooth latent space $z \in \mathcal{Z}$ by learning a generative model $p(x, z)$. Starting from a prior distribution $p(z)$, such as standard multivariate Gaussian $\mathcal{N}(0, I)$, VAE generates data with a complex conditional distribution $p_\theta(x|z)$ parameterized by one neural network $f_\theta(\cdot)$. The goal of the model training is to maximize the marginal likelihood $E_{p_\mathcal{D}(x)}[\log p_\theta(x)]$, where the $p_\mathcal{D}(x)$ is the true underlying distribution. To calculate this intractable marginal likelihood, an amortized inference distribution $q_\phi(z|x)$ parameterized by one neural network $f_\phi(\cdot)$ has been utilized to approximate the true posterior. Then, it turns out to optimize the following lower bound:

$$\mathcal{L} = E_{p_\mathcal{D}(x)}[E_{q_\phi(z|x)}[\log p_\theta(x|z)] - [D_{KL}[q_\phi(z|x)||p(z)]], \quad (1)$$

where the first term is the reconstruction loss and the second one is the Kullback-Leibler (KL) divergence between the approximated posterior and prior.

Unfortunately, in practice, VAE may fail to capture meaningful representation. In particular, when applying auto-regressive models as the decoder network, such as LSTMs or PixelCNN, it is likely to model the data marginal distribution $p_\mathcal{D}(x)$ very well even without latent variable z, i.e., $p(x|z) = \Pi_i p(x_i|x_{<i})$. In this case, VAE degenerates to auto-regressive, the latent variable z tends to be independent with the data x. Meanwhile, with the goal to minimize $D_{KL}[q(z|x)||p(z)]$ in ELBO objective, $q(z|x)$ vanishes to $p(z)$, i.e., $q(z|x_i) = q(z|x_j) = p(z)$, $\forall x_i, x_j \in \mathcal{X}$. To solve this problem, we will direct the latent space learning carefully and purposefully for high diversity and low uncertainty in the following.

3 The Proposed Method

Here, we start with theoretical analysis on the latent space of VAE from two geometric properties: diversity and uncertainty, respectively. Then, we design Dropout on the variance parameters and Batch-Normalization on the mean parameters to encourage the latent space with high diversity and low uncertainty. In particular, the effectiveness of our approach will be discussed and proved rigorously. Finally, we extend DU-VAE into VAE-IAF [Kingma *et al.*, 2016] empirically.

3.1 Geometric Properties of Latent Space

For enabling meaningful and compact representation learning in VAE model, we have two intuitions: 1) for different data samples x_1, x_2, the posteriors $q(z_1|x_1)$ and $q(z_2|x_2)$ should mutually diversify from each other, which encourages posteriors to capture the characteristic or discriminative information from data; 2) given data sample x, the degree of uncertainty of the latent variable z should be minimized, which encourages removing redundant information from z. Guided by those intuitions, we first analyze the diversity and uncertainty of latent space under quantitative metric, respectively.

Diversity of Latent Space

Here, we attempt to measure the divergence among the posterior distribution family. One intuitive and reasonable metric is the expectation of the mutual divergence between a pair of posteriors. Following this idea, [Ma *et al.*, 2018] proposed the mutual posterior diversity (MPD) to measure the diversity of posteriors, which can be computed by:

$$MPD_{p_\mathcal{D}(x)}[z] = E_{p_\mathcal{D}(x)}[D_{SKL}[q_\phi(z_1|x_1)|q_\phi(z_2||x_2)]], \quad (2)$$

where $x_1, x_2 \sim p_\mathcal{D}(x)$ are $i.i.d.$ and $D_{SKL}[q_1||q_2]$ is symmetric KL divergence defined as the mean of $D_{KL}[q_2||q_1]$ and $D_{KL}[q_2||q_1]$, which is analytical under Gaussian distributions. Specifically, we have:

$$\begin{aligned} 2MPD_{p_\mathcal{D}(x)}[z] = &\sum_{d=1}^{n} E_{p_\mathcal{D}(x)}[\frac{(\mu_{x_1,d} - \mu_{x_2,d})^2}{\delta^2_{x_1,d}}] \\ &+ \sum_{d=1}^{n} E_{p_\mathcal{D}(x)}[\delta^2_{x,d}] E_{p_\mathcal{D}(x)}[\frac{1}{\delta^2_{x,d}}] - 1. \end{aligned} \quad (3)$$

Interestingly, if the value of $\delta^2_{x,d}$ is upper bounded, like less than 1 in most practical case for VAEs. then, we can find that MPD has one lower and strict bound proportional to $\sum_{d=1}^{n} Var_{P_\mathcal{D}(x)}[\mu_{x,d}]$ (see Supplementary).

Uncertainty of Latent Space

Here, we aim at quantifying the uncertainty about the outcome of latent variable z given data sample x and the learned encoder distribution $q_\phi(z|x)$. In information theory, conditional entropy is utilized to measure the average level of the uncertainty inherent in the variable's possible outcomes when giving another variable. Due to the same goal, we follow this idea and use the Conditional Entropy (CE) $H_{q_\phi}(z|x)$ for z conditioned on x to measure the uncertainty of latent space:

$$H_{q_\phi}(z|x) = E_{p_\mathcal{D}(x)}[H(q_\phi(z|x))], \quad (4)$$

where $H(q_\phi(z|x))$ denotes the differential entropy of posterior $q_\phi(z|x)$. Actually, $H(q_\phi(z|x))$ can be computed analytically as $\sum_{d=1}^{n} \frac{1}{2}\log(2\pi e \delta^2_{x,d})$, then we have:

$$H_{q_\phi}(z|x) = \frac{n}{2}\log 2\pi e + \frac{1}{2}\sum_{d=1}^{n} E_{p_\mathcal{D}(x)}[\log \delta^2_{x,d}]. \quad (5)$$

Intuitively, in order to reduce the uncertainty in the latent space, we need to minimize the conditional entropy $H_{q_\phi}(z|x)$. However, the differential entropies $H(q_\phi(z|x))$ defined on continuous spaces are not bounded from below. That is, the variance $\delta^2_{x,d}$ can be scaled to be arbitrarily small achieving arbitrarily high-magnitude negative entropy. As a result, the optimization trajectories will invariably end with garbage networks as activations approach zero or infinite. To solve this problem, we enforce the differential entropy non-negativity by adding noise to the latent variable. For one latent variable z, we replace it with $z + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \alpha)$ is one zero-entropy noise, where we set constant $\alpha = \frac{1}{2\pi e}$ for convenience. Then based on properties of Gaussian distribution, we have $H(q_\phi(z|x)) > H(\epsilon) = 0$ and $\delta^2_{x,d} > \alpha$.

In sum, in order to encourage the diversity and decrease the uncertainty of latent space, we need to constrain both MPD in Equation 3 and CE in Equation 4 . One feasible solution is to regard them as additional objectives explicitly and approximate them by using Monte Carlo in each mini-batch. However, the scales among different objective terms are unbalanced, which require deliberately designed normalization or careful weight parameters tuning [Ma *et al.*, 2018].

Instead, we propose control implicitly the MPD and CE without modifying the objective function. Based on Equation 3 and Equation 4, we note that both MPD and CE are only dependent on approximated posterior' s parameters, i.e., $\mu_{x,d}$ or $\delta_{x,d}$. This inspires us to select proper regularization on the distribution of posterior' s parameters to encourage higher MPD and lower CE. Specifically, in the following two sub-sections. we will introduce the application of the Dropout on variance parameters and Batch-Normalization on mean parameters respectively, and provide theoretical analysis about the effectiveness of our approach.

3.2 Dropout on Variance Parameters

In order to encourage high diversity and low uncertainty of latent space, we need to increase the MPD in Equation 3 and decrease the CE in Equation 5, simultaneously. Meanwhile, we also need to avoid $E_{p_\mathcal{D}(x)}[\delta^2_{x,d}]$ to be too small for ensuring the smoothing of the latent space. One extreme case is that when $E_{p_\mathcal{D}(x)}[\delta^2_{x,d}]$ convergence to 0, i.e., $\delta^2_{x,d} \approx 0, \forall x, d$, each data point is associated with a delta distribution in latent space and the VAEs degenerate into Autoencoders in this dimension. To accomplish these requirements together, we propose to apply Dropout [Srivastava *et al.*, 2014] to regularize posterior's variance parameters in training as following,

$$\hat{\delta^2}_{x,d} = g_{x,d}(\delta^2_{x,d} - \alpha) + \alpha, \tag{6}$$

where $g_{x,d}$ denotes the independent random variable generated from the normalized Bernoulli distribution $1/pB(1,p),\ p \in (0,1)$, where $E_B[g_{x,d}] = 1$. Then, we have the following proposition (see Supplementary for the proof.):

Proposition 1. *Given the Dropout strategy defined in Equation 6, we have:*

$$\begin{aligned} &E_{p_{\mathcal{D}(x)\cdot B}}[\hat{\delta}^2_{x,d}] = E_{p_\mathcal{D}(x)}[\delta^2_{x,d}], \\ &MPD_{p_{\mathcal{D}(x)\cdot B}}[z] > MPD_{p_{\mathcal{D}(x)}}[z], \\ &H_{q_\phi \cdot B}(z|x) < H_{q_\phi}(z|x), \end{aligned} \tag{7}$$

Algorithm 1 Training Procedure of DU-VAE

1: Initialize ϕ, θ, $\gamma_\mu = \gamma$, and $\beta_\mu = 0$
2: **while** not convergence **do**
3: Sample a mini-batch x
4: $\mu_x, \delta^2_x = f_\phi(x)$.
5: $\hat{\mu}_x = BN_{\gamma_\mu,\beta_\mu}(\mu_x)$, $\hat{\delta}^2_x = Dropout_p(\delta^2_x)$.
6: Sample $z \sim \mathcal{N}(\hat{\mu}_x, \hat{\delta}^2_x)$ and generate x from $f_\theta(z)$.
7: Compute gradients $g_{\phi,\theta} \leftarrow \nabla_{\phi,\theta}\mathcal{L}_{ELBO}(x;\phi,\theta)$.
8: Update ϕ, θ, γ_μ, β_μ according to $g_{\phi,\theta}$.
9: $\gamma_\mu = \frac{\gamma}{\sqrt{E[\gamma^2_\mu]}} \odot \gamma_\mu$
10: **end while**

where two inequalities are both strict, the gaps between two sides are greater as p decreases to 0. Then, we also have:

$$MPD_{p_{\mathcal{D}(x)}\cdot B}[z] > \frac{1-p}{\alpha}\sum_{d=1}^{n} Var_{p_\mathcal{D}(x)}[\mu_{x,d}]. \tag{8}$$

Proposition 1 tells us that: 1) Dropout regularization encourages the increase of $MPD_{p_\mathcal{D}(x)}[z]$ and the decrease of the conditional entropy $H_{q_\phi}(z|x)$ of the latent space while preserving the expectation of variance parameters, which is actually a simple but useful strategy what we need. 2) Dropout regularization also provides one lower bound of $MPD_{p_\mathcal{D}(x)}[z]$ independent on the variance parameters, which makes it possible to ensure positive MPD with further controls on the variance $\sum_{d=1}^{n} Var_{p_\mathcal{D}(x)}[\mu_{x,d}]$.

3.3 Batch-Normalization on Mean Parameters

Inspired by Batch-Normalization (BN) [Ioffe and Szegedy, 2015], which is an effective approach to control the distribution of the output of neural network layer. We apply BN on the mean parameters $\mu_{x,d}$ to constrain $\sum_{d=1}^{n} Var_{p_\mathcal{D}(x)}[\mu_{x,d}]$. Mathematically, our BN is defined as:

$$\hat{\mu}_{x,d} = \gamma_{\mu_d}\frac{\mu_{x,d} - \mu_{\mathcal{B}_d}}{\delta_{\mathcal{B}_d}} + \beta_{\mu_d}, \tag{9}$$

where $\hat{\mu}_{x,d}$ represents the output of BN layer, and $\mu_{\mathcal{B}_d}$ and $\delta_{\mathcal{B}_d}$ denote the mean and standard deviation of $\mu_{x,d}$ estimated within each mini-batch. γ_{μ_d} and β_{μ_d} are the scale and shift parameters, which lead that the distribution of $\hat{\mu}_{x,d}$ has the variance $\gamma^2_{\mu_d}$ and mean β_{μ_d}. Therefore, we can control the $\sum_{d=1}^{n} Var_{p_\mathcal{D}(x)}[\mu_{x,d}]$ by fixing the mean $E_d[\gamma^2_{\mu_d}] = \gamma^2$ with respect to each dimension d. Specifically, we regard each γ_d as learnable parameters with initialization γ. Then after each training iteration, we re-scale each parameter γ_{μ_d} with coefficient $\gamma/\sqrt{E_d[\gamma^2_{\mu_d}]}$. In addition, all β_{μ_d} is learnable with initialization 0 and no constraint.

Overall, based on the analysis above, we propose our approach, namely DU-VAE, to encourage high diversity and low uncertainty of the latent space by applying Dropout regularizations on variance parameters and Batch-Normalization on mean parameters of approximated posteriors, simultaneously. Specifically, we train DU-VAE following Algorithm 1.

Connections with BN-VAE. In the literature, BN-VAE [Zhu *et al.*, 2020] also applies BN on mean parameters. Zhu et

al. claim that keeping one positive lower bound of the KL term, i.e., the expectation of the square of mean parameters $\sum_{d=1}^{n} E_{q_\phi}[\mu_{x,d}^2]$, is *sufficient* for preventing posterior collapse. In practice, they ensure $E_{q_\phi}[\mu_{x,d}^2] > 0$ by fixing scale parameter γ_{μ_d} of BN for each dimension d. However, here, we will demonstrate that keeping one positive lower bound of MPD is one more powerful strategy for preventing collapse posterior. As the discussion in Section 2, when posterior collapse occurs, we have $q(z|x_i) = q(z|x_j) = p(z)$, $\forall x_i, x_j \in \mathcal{X}$. Therefore, to avoid this phenomenon, we actually need to control posterior distributions carefully so that:

$$\begin{aligned} &q(z|x) \neq p(z), \ \exists\, x \in \mathcal{X}, \\ &q(z|x_i) \neq q(z|x_j), \ \exists\, x_i, x_j \in \mathcal{X}. \end{aligned} \tag{10}$$

where the first term is actually implied in the second term as the *necessary* condition and $D_{KL}[q(z|x)||p(x)] > 0$ is *equivalent* (both *sufficient* and *necessary*) to the first term, we can claim that keeping one positive lower bound of the KL term is not *sufficient* for the second term along with several certain abnormal cases (detailed analysis can be found in Supplementary.). By contrast, keeping one positive MPD in the latent space is actually one *equivalent* condition for the second term, which implies the first term. Actually, from the perspective of the diversity of latent space, we can provide one more possible explanation for the effectiveness of BN-VAE. That is, the application of BN on μ_x ensures one positive value of $Var_{p_\mathcal{D}(x)}[\mu_x, d]$ for each d, which is also one lower bound of MPD defined in Equation 2 when the variance parameters has one constant upper bound, like 1 in practice.

3.4 Extension to VAE-IAF

Here, to further examine the generalization of DU-VAE, we aim to extend our approach for other VAE variants, such as, VAE-IAF [Kingma *et al.*, 2016], one well-known normalizing flow-based VAE. Different from classic VAEs which assume the posterior distributions are diagonal Gaussian distributions, VAE-IAF can construct more flexible posterior distributions through applying one chain of invertible transformations, named the IAF chain, on an initial random variable drown from one diagonal Gaussian distribution. Specifically, the initial random variable z^0 is sampled from the diagonal Gaussian with parameters μ^0 and δ^0 outputted from the encoder network. Then, T invertible transformations, are applied to transform z^0 into the final random variable z^T. More details can be found in [Kingma *et al.*, 2016].

Indeed, noting that the MPD and CE of the initial random variable z^0 have the same form as these for classic VAEs in Equation 2 and Equation 4, one intuitive idea is to apply Dropout on δ^0 and Batch Normalization on μ^0 with the guidance in Algorithm 1 to control the MPD and CE of z^0. It is surprising to find that this simple extension of DU-VAE, called DU-IAF, demonstrated comparative performance in our experiments. This may be attributed to the close connection between z^0 and z^T. In particular, we find that the CE of z^0 is the upper bound of CE of z^T. Meanwhile, $MPD_{p_\mathcal{D}(x)}[z^0]$ is closely related with $MPD_{p_\mathcal{D}(x)}[z^T]$, even they are equal to each other when each invertible transformation in IAF chain is independent on the input data. Further discussion and proof can be found in Supplementary.

4 Experiments

In this section, our method would be evaluated on three benchmark datasets in terms of various metrics and tasks. Complete experimental setup can be found in Supplementary.

4.1 Experimental Setup

Setting. Following the same configuration as [He *et al.*, 2018], we evaluated our method on two text benchmark datasets, i.e., Yahoo and Yelp corpora [Yang *et al.*, 2017] and one image benchmark dataset, i.e., OMNIGLOT [Lake *et al.*, 2015]. For text datasets, we utilized a single layer LSTM as both encoder and decoder networks, where the initial state of the decoder is projected by the latent variable z. For images, a 3-layer ResNet [He *et al.*, 2016] encoder and a 13-layer Gated PixelCNN [Van den Oord *et al.*, 2016] decoder are applied. We set the dimension of z as 32. and utilized SGD to optimize the ELBO objective for text and Adam [Kingma and Ba, 2015] for images. Following [Burda *et al.*, 2016], we utilized dynamically binarized images for training and the fixed binarization as test data. Meanwhile, following [Bowman *et al.*, 2015a], we applied a linear annealing strategy to increasing the KL weight from 0 to 1 in the first 10 epochs if possible.

Evaluation Metrics. Following [Burda *et al.*, 2016], we computed the approximate negative log-likelihood (NLL) by estimating 500 importance weighted samples. In addition, we also considered the value of KL term, mutual information (MI) $I(x, z)$ [Alemi *et al.*, 2016] under the joint distribution $q(x, z)$ and the number of activate units (AU) [He *et al.*, 2018] as additional metrics. In particular, the activity of each dimension z_d is measured as $A_{z,d} = Cov(E_{z_d \sim q(z_d|x)}[z_d])$. One dimension is regarded to be active when $A_{z,d} > 0.01$.

Baselines. We compare our method with various VAE-based models, which can be grouped into two categories: 1) Classic VAEs: **VAE** with annealing [Bowman *et al.*, 2015a]; Semi-Amortized VAE (**SA-VAE**) [Kim *et al.*, 2018]; **Agg-VAE** [He *et al.*, 2018]; β**-VAE** [Higgins *et al.*, 2017] with parameter β re-weighting the KL term; **FB** [Kingma *et al.*, 2016] with parameter λ constraining the minimum of KL term in each dimension; δ**-VAE** [Razavi *et al.*, 2018] with parameter δ constraining the range of KL term; **BN-VAE** [Zhu *et al.*, 2020] with parameter γ keeping one positive KL term; **MAE** [Ma *et al.*, 2018] with parameters γ and η controlling the diversity and smoothness of the latent space. Note that we implemented MAE with the standard Gaussian prior, instead of the AF prior in [Ma *et al.*, 2018] for one fair comparison. 2) IAF-based models: **IAF+FB** [Kingma *et al.*, 2016], which utilized the FB strategy with the parameter λ to avoid the posterior collapse in VAE-IAF; **IAF+BN**, where we applied BN regularization on the mean parameters of the distributions of z^0 with the fixed scale parameters γ in each dimension.

4.2 Overall Performance

Log-Likelihood Estimation. Table 1 shows the results in terms of log-likelihood estimation. We can note that DU-VAE and DU-IAF achieve the best NLL among classic VAEs and IAF-based VAEs in all datasets, respectively. Besides, we also have some interesting findings. First, MAE does not perform well in all datasets, which may be caused by the

	Yahoo				Yelp				OMNIGLOT			
Model	NLL	KL	MI	AU	NLL	KL	MI	AU	NLL	KL	MI	AU
VAE	328.5	0.0	0.0	5.0	357.5	0.0	0.0	0.0	89.21	2.20	2.16	5.0
β-VAE* (0.4/ 0.4/ 0.8)	328.7	6.4	6.0	13.0	357.4	5.8	5.6	4.0	89.15	9.98	3.84	13.0
SA-VAE*	327.2	5.2	2.7	8.6	355.9	2.8	1.7	8.4	89.07	3.32	2.63	8.6
Agg-VAE	**326.7**	5.7	2.9	6.0	355.9	3.8	2.4	11.3	89.04	2.48	2.50	6.0
FB (0.1)	328.1	3.4	2.5	32.0	357.1	4.8	2.5	32.0	89.17	7.98	6.87	32.0
δ-VAE (0.1)	329.0	3.2	0.0	2.0	357.6	3.2	0.0	0.0	89.62	3.20	2.36	2.0
BN-VAE (0.6/ 0.6/ 0.5)	326.9	8.3	7.0	32.0	355.7	6.0	5.2	32.0	89.26	4.34	4.03	32.0
MAE (1/ 2/ 0.5, 0.2/ 0.2/ 0.2)	332.1	5.8	3.5	28.0	362.8	8.0	4.6	32.0	89.62	15.61	8.90	32.0
DU-VAE (0.5, 0.9)	327.0	5.2	4.3	18.0	355.6	5.3	4.9	18.0	**89.00**	6.63	5.97	19.0
DU-VAE (0.5, 0.8)	327.0	6.7	6.0	19.0	**355.5**	6.8	5.9	18.0	89.04	7.46	6.31	32.0
DU-VAE (0.6, 0.8)	**326.7**	8.7	7.2	28.0	355.8	9.6	7.7	23.0	89.18	10.99	8.22	32.0
IAF+FB (0.15/0.25/0.15)	328.4	5.2	-	-	357.1	7.7	-	-	88.98	6.77	-	-
IAF+BN (0.6/0.7/0.5)	328.1	0.2	-	-	356.6	0.6	-	-	89.32	1.30	-	-
DU-IAF (0.7/0.6/0.5, 0.70/0.70/0.85)	**327.4**	5.4	-	-	**356.1**	5.1	-	-	**88.97**	6.77	-	-

Table 1: The performance on likelihood estimation. Due to the intractability of MI and AU metrics for IAF-based models, we just report NLL and KL same as [Kingma *et al.*, 2016]. * indictes the results are referred from [He *et al.*, 2018]. Hyper-parameters are reported in brackets and split by slashes if different on different datasets.

#label	100	500	1k	2k	10k
AE	84.05	86.82	87.93	88.19	88.75
VAE	71.10	71.43	71.58	72.96	77.11
δ-VAE (0.1)	60.11	60.52	61.46	63.79	64.38
Agg-VAE	75.05	77.16	78.50	79.29	80.07
FB (0.1)	75.19	80.78	81.63	82.28	82.39
BN-VAE(0.6)	84.53	88.22	89.45	89.63	89.72
MAE (2, 0.2)	61.50	61.70	62.42	63.58	63.68
DU-VAE (0.5, 0.8)	**88.91**	**89.63**	**90.36**	**90.51**	**90.77**
IAF+FB(0.25)	89.73	90.60	90.94	90.91	91.01
IAF+BN(0.7)	87.98	89.03	89.18	89.35	90.29
DU-IAF (0.6, 0.7)	**91.25**	**91.10**	**91.52**	**91.97**	**92.31**

Table 2: The accuracy of the classification on Yelp.

#label for each character	5	10	15
AE	37.28	43.38	46.94
VAE	29.48	37.79	42.24
δ-VAE (0.1)	37.28	43.38	46.94
Agg-VAE	33.72	41.31	46.27
FB (0.1)	33.93	41.05	45.21
BN-VAE (0.5)	31.17	39.15	43.24
MAE (05, 0.2)	35.05	41.72	44.95
DU-VAE (0.5, 0.1)	**40.54**	**48.09**	**52.47**
IAF+FB(0.15)	38.33	45.85	49.90
IAF+BN(0.5)	16.58	19.49	21.11
DU-IAF (0.5, 0.15)	**41.84**	**49.86**	**52.97**

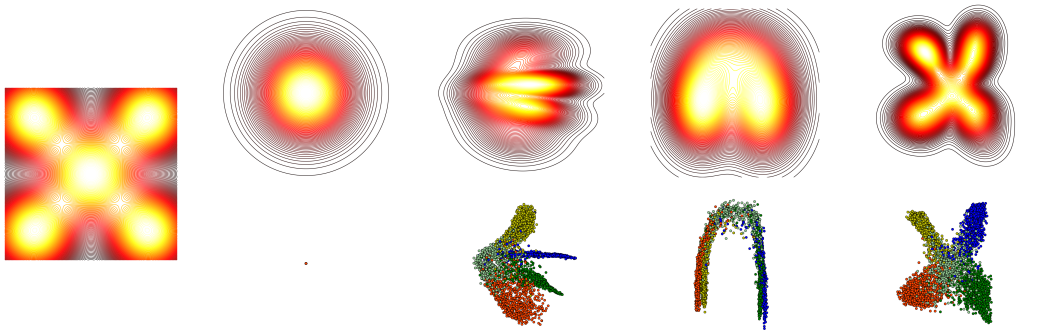
Table 3: The average accuracy of classifications on OMNIGLOT.

difficulty to balance the additional training objective terms and ELBO. Second, although, Agg-VAE and SA-VAE also reach the great NLL in both datasets, they require the additional training procedure on the inference network, leading to the high training time cost [Zhu *et al.*, 2020]. Third, BN-VAE also achieves completive performance on text datasets. However, for images, where the posterior collapse may be less of an issue [Kim *et al.*, 2018], BN-VAE fails to catch up with other models, even worse than basic VAE on NLL. Fourth, DU-VAE prefers to capture higher KL and MI compared with BN-VAE with the same scale parameter γ. In other words, DU-VAE can convert more information from the observed data into the latent variable. Fifth, based on the results of IAF+BN, we can find that the BN strategy used in BN-VAE can not prevent the collapse posterior in VAE-IAF with small KL. By contrast, our approach can be easily extended for VAE-IAF with the best performance. Finally, we also note that IAF based models may be more suitable for image dataset without sound performance on text, while DU-IAF nevertheless achieves competitive performance.

Classification. To evaluate the quality of learned representation, we train a one-layer linear with the output from the trained model as the input for classification tasks on both text and image datasets. For classic VAEs, the mean parameter μ of each latent variable has been used as the representation vector. For IAF-based models, we first selected the initial sample z^0 in latent space as its mean parameter μ^0. Then, the combination of z^0 and z^T is used as the representation vector.

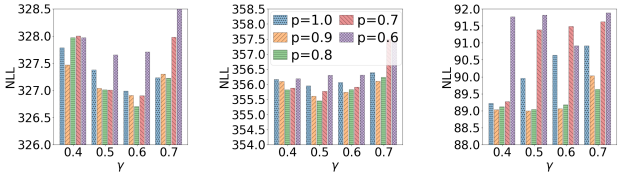
Specifically, for text datasets, following [Shen *et al.*, 2017], we work with one downsampled version of Yelp sentiment dataset for binary classification. Table 2 shows the performance under varying number of labeled data. For the image dataset, noting that OMNIGLOT contains 1623 different handwritten characters from 50 different alphabets, where each character has 15 images in our training data and 5 images in our testing data, we conducted classifications on each alphabet with varying number of training samples for each character. Table 3 reports the average accuracy.

We can find that DU-VAE and DU-IAF achieve the best accuracy under all settings for classic VAEs and IAF-based models respectively. Interestingly, we also find that most baselines show inconsistent results on text and image classification. For example, Agg-VAE and BN-VAE may be better at text classification without sound accuracy in Table 3. On the contrary, δ-VAE and MAE adapt to image classification better with uncompetitive performance in Table 2. Meanwhile,



(a) True Latent Space (b) VAE (c) Agg-VAE (d) BN-VAE (1.0) (e) DU-VAE (1.0, 0.5)

Figure 1: The visualization of the latent space learned by DU-VAE and other baselines. Figure (a) is the counter plot of the true latent space for generating the synthetic dataset. In the rest, the first line shows the counter plot of the aggregated posterior $q_\phi(z)$. The brighter the color, the higher the probability. Meanwhile, the location of mean parameters are displayed in the second line with colors to distinguish different categories generated from different Gaussian components, where the blue ones correspond to the component in the center in Figure (a), and the others denote the other four components. All figures are located in the same region, i.e., $z \in [-3, 3] \times [-3, 3]$, with the same scale.



(a) Yahoo (b) Yelp (c) OMNIGLOT

Figure 2: Parameter Analysis.

we note IAF chain trends to improve the classification accuracy for FB and our approach to both text and image datasets. However, IAF+BN fails to achieve competitive performance on image classification, which indicates that the applications of BN in BN-VAE may not be suitable for image again.

Parameter Analysis. Here, we train DU-VAE by varying γ from 0.4 to 0.7 and p from 1 to 0.6. As the Figure 2 shows, we find that, DU-VAE would achieve the best NLL with parameters (γ, p) as when (0.6, $p = 0.8$) for Yahoo, (0.5, $p = 0.8$) for Yelp, and (0.5, $p = 0.9$) for OMNIGLOT, respectively.

4.3 Case Study–Latent Space Visualization

Here, we aim to provide one intuitive comparison of latent spaces learned by different models based on one simple synthetic dataset. Specifically, following [Kim *et al.*, 2018], we first sample 2-dimensional latent variable z from one mixture of Gaussian distributions that have 5 mixture components. Then one text-like dataset can be generated from one LSTM layer conditioned on those latent variables. Based on this synthetic dataset, we trained different VAEs with 2-dimensional standard Gaussian prior and diagonal Gaussian posterior. Then, we visualize the learned latent spaces by displaying the counter plot of the aggregated approximated posteriors $q(z) = E_{p_\mathcal{D}(x)}[q_\phi(z|x)]$ and the location of approximated posterior's mean parameters for different samples x.

According to the results in Figure 1, we have some interesting observations. First, due to the posterior collapse, VAE learns an almost meaningless latent space where the posterior $q(z|x)$ for all data are squeezed in the center. Actually, it is not surprising that the aggregated posterior matches the prior excessively in this case, because we almost have $q_\phi(z|x) = p(z)$, $\forall x$. Second, Agg-VAE, BN-VAE, and DU-VAE all tend to diverse samples in different categories, but in different manners and degrees. Intuitively, all three models force to embedding the blue category in the center around by the other four categories. However, only the average posterior learned by DU-VAE have five centers same as the true latent space. Meanwhile, DU-VAE with Dropout strategy encourages the aggregated posteriors to be more compact, while that of BN-VAE is more broad, compared with the prior. Those observations demonstrate that DU-VAE tends to guide the latent space to be more diverse and less uncertain.

5 Conclusion

In this paper, we developed a novel generative model, DU-VAE, for learning a more diverse and less uncertain latent space. The goal of DU-VAE is to ensure that more meaningful and compact representations can be learned. Specifically, we first demonstrated theoretically that it led to better latent space with high diversity and low uncertainty awareness by controlling the distribution of posterior's parameters across the whole dataset respectively. Then, instead of introducing new loss terms or modifying training strategies, we proposed to apply Dropout on the variances and Batch-Normalization on the means simultaneously to regularize their distributions implicitly. Furthermore, we extended DU-VAE into VAE-IAF empirically. The experimental results on three benchmark datasets clearly showed that DU-VAE outperformed state-of-the-art baselines on both likelihood estimation and underlying classification tasks.

Acknowledgements

This work was partially supported by grants from the National Natural Science Foundation of China (Grant No. 91746301, 61836013).

References

[Alemi *et al.*, 2016] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *ICLR*, 2016.

[Bowman *et al.*, 2015a] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[Bowman *et al.*, 2015b] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *CCNLP*, 2015.

[Burda *et al.*, 2016] Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR*, 2016.

[Chen *et al.*, 2017] Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, P. Dhariwal, John Schulman, Ilya Sutskever, and P. Abbeel. Variational lossy autoencoder. *ICLR*, 2017.

[Fang *et al.*, 2019] Le Fang, Chunyuan Li, Jianfeng Gao, Wen Dong, and Changyou Chen. Implicit deep latent variable models for text generation. In *EMNLP*, 2019.

[Fu *et al.*, 2019] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *ACL*, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[He *et al.*, 2018] Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR*, 2018.

[Higgins *et al.*, 2017] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

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

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

[Kim *et al.*, 2018] Yoon Kim, Sam Wiseman, Andrew Miller, David Sontag, and Alexander Rush. Semi-amortized variational autoencoders. In *ICML*, 2018.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013.

[Kingma *et al.*, 2016] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NeurIPS*, 2016.

[Lake *et al.*, 2015] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[Ma *et al.*, 2018] Xuezhe Ma, Chunting Zhou, and Eduard Hovy. Mae: Mutual posterior-divergence regularization for variational autoencoders. In *ICLR*, 2018.

[Makhzani *et al.*, 2016] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *ICLR*, 2016.

[Pu *et al.*, 2016] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *NeurIPS*, 2016.

[Razavi *et al.*, 2018] Ali Razavi, Aaron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. In *ICLR*, 2018.

[Shen *et al.*, 2017] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*, 2017.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JLMR*, 15(1):1929–1958, 2014.

[Van den Oord *et al.*, 2016] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, pages 4790–4798, 2016.

[Yang *et al.*, 2017] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *ICML*, 2017.

[Zhao *et al.*, 2019] Shengjia Zhao, Jiaming Song, and S. Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *AAAI*, 2019.

[Zheng *et al.*, 2019] Huangjie Zheng, Jiangchao Yao, Ya Zhang, Ivor W Tsang, and Jia Wang. Understanding vaes in fisher-shannon plane. In *AAAI*, 2019.

[Zhu *et al.*, 2020] Qile Zhu, Wei Bi, Xiaojiang Liu, Xiyao Ma, Xiaolin Li, and D. Wu. A batch normalized inference network keeps the kl vanishing away. In *ACL*, 2020.