Addressing Posterior Collapse in Variational Autoencoders with β-VAE

Ziyang Wang*

School of Marine Science and Technology, Northwestern Polytechnical University,

Xi'an, 450000, China

*Corresponding author: wzyqjyy@mail.nwpu.edu.cn

Abstract. Posterior collapse is a pervasive issue in Variational Autoencoders (VAEs) that leads to the learned latent representations becoming trivial and devoid of meaningful information. To address this problem, this paper presents a novel β-VAE approach, which incorporates a hyperparameter β to strike an optimal balance between the reconstruction loss and the KL divergence loss. By conducting a comprehensive series of experiments and drawing comparisons with existing methods, robust evidence is provided that the proposed β-VAE method effectively mitigates posterior collapse and yields more expressive and informative latent representations.The experimental setup involves various architectures and datasets to demonstrate the versatility and efficacy of the β-VAE approach in diverse settings. Additionally, ablation studies are performed to investigate the impact of different β values on the model's performance, elucidating the role of this hyperparameter in controlling the trade-off between reconstruction quality and latent representation expressiveness. Furthermore, the disentanglement properties of the learned latent space are analyzed, which is a crucial aspect of VAEs, especially when applied to complex, real-world data.In-depth analysis of the results offers valuable insights into the underlying mechanisms of β-VAE, thereby contributing to a more profound understanding of VAEs and their inherent limitations. The findings not only establish the effectiveness of the β-VAE method in preventing posterior collapse but also pave the way for future research on improving VAEs' performance in various applications. Potential future work could explore alternative techniques for balancing the competing objectives of reconstruction and latent representation learning or delve into the theoretical properties of β-VAE, providing a more rigorous foundation for this approach.

Keywords: Variational autoencoder, Posterior collapse, β-VAE.

1. Introduction

The Variational Autoencoder (VAE) represents a deep generative model that seamlessly integrates the principles of autoencoders and probabilistic graphical models [1]. VAEs have found widespread applications in various domains, such as unsupervised learning, data compression, denoising, and generative modeling. The primary objective during training is to maximize the evidence lower bound (ELBO), which consists of a reconstruction error term and a KL divergence term. However, VAEs may experience posterior collapse in practice, where the posterior distribution closely approximates the prior distribution, resulting in suboptimal sample quality. To tackle this issue, the β-VAE method has been introduced, which employs a hyperparameter β to modulate the balance between the reconstruction loss and the KL divergence loss [2, 3].

This paper presents an in-depth investigation of the β-VAE method, encompassing its theoretical underpinnings, model architecture, experimental outcomes, and comprehensive analysis. The efficacy of β-VAE in alleviating the posterior collapse issue is assessed through an extensive set of experiments and juxtapositions with the conventional VAE [4].

Additionally, the paper delves into the impact of varying β values on the model's performance, elucidating the significance of this hyperparameter in balancing the trade-off between reconstruction quality and latent representation expressiveness. Furthermore, an exploration of the disentanglement properties of the learned latent space is conducted, which is an essential aspect of VAEs, particularly when applied to intricate, real-world data.

The study's findings not only substantiate the effectiveness of the β-VAE method in mitigating posterior collapse but also serve as a foundation for further research in refining VAEs' performance across a variety of applications. Future work may include the investigation of alternative strategies for reconciling the competing objectives of reconstruction and latent representation learning or an examination of the theoretical properties of β-VAE, thus providing a more rigorous basis for this approach..

2. Related Theory

This section offers a comprehensive and theoretical overview of Variational Autoencoders (VAEs) and derives the Evidence Lower Bound (ELBO) [5]. Additionally, the posterior collapse phenomenon and its underlying causes are explored.

2.1. Theory of VAE

Composed of an encoder and decoder framework, variational autoencoders are designed to reduce the reconstruction disparity between the original data and the data following encoding and decoding. To enforce particular regularization in the latent space, adjustments are implemented in the encoding-decoding mechanism: the input is encoded as a probability distribution in the latent space, rather than as an individual point.

Initially, the premise is that all data samples are independent and identically distributed, indicating no mutual influence between observations. The objective entails estimating the parameters for the generative model $p_\theta(z|x)$ through the application of the maximum likelihood approach, requiring the maximization of the subsequent log-likelihood function:

$$\log p_\theta\left(x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right) = \sum_{i=1}^{N} \log p_\theta(x^{(i)}) \tag{1}$$

Variational autoencoders employ the recognition model $q_\emptyset(\mathrm{z}|x)$ to approximate the true posterior probability $p_\theta(z|x)$. To gauge the resemblance between these two distributions, typically, the Kullback-Leibler (KL) divergence is utilized, as follows:

$$\begin{aligned} \mathrm{KL}(q_\emptyset(\mathrm{z}|x) \| p_\theta(z|x)) &= \mathrm{E}_{q_\emptyset(\mathrm{z}|x)} \log \frac{q_\emptyset(\mathrm{z}|x)}{p_\theta(z|x)} \\ &= \mathrm{E}_{q_\emptyset(\mathrm{z}|x)} \log \frac{q_\emptyset(\mathrm{z}|x) p_\theta(x)}{p_\theta(z|x) p_\theta(x)} \\ &= \mathrm{E}_{q_\emptyset(\mathrm{z}|x)} \log \frac{q_\emptyset(\mathrm{z}|x)}{p_\theta(z|x)} + \mathrm{E}_{q_\emptyset(\mathrm{z}|x)} \log p_\theta(x) \\ &= \mathrm{E}_{q_\emptyset(\mathrm{z}|x)} \log \frac{q_\emptyset(\mathrm{z}|x)}{p_\theta(z|x)} + \log p_\theta(x) \end{aligned} \tag{2}$$

Thus

$$\log p_\theta(x) = \mathrm{KL}\Big(q_\emptyset(\mathrm{z}|x), p_\theta(z|x)\Big) + \Gamma(\theta, \emptyset, x) \tag{3}$$

where

$$\Gamma(\theta, \emptyset, x) = E_{q(Z|x)}[\log p(x|z)] - \mathrm{KL}\Big(q_\emptyset(\mathrm{z}|x), p_\theta(z|x)\Big) \tag{4}$$

$\Gamma(\theta, \emptyset, x)$ is called Evidence Lower Bound (ELBO) [5], function a lower bound for the log-likelihood of $p_{\theta(x)}$.

Notably, ELBO exhibits a structure akin to an autoencoder, which is represented in figure 1.

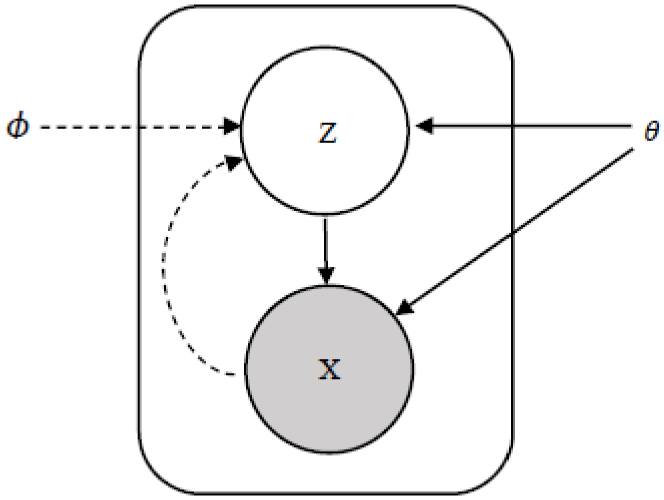

Figure. 1 Basic VAE structure

2.2. Posterior Collapse

During training, the KL term D_{KL} can easily disappear and approach 0. As a result, the VAE network's decoder will ignore the posterior distribution $q_{\emptyset}(z|x)$ produced by the encoder (which has already collapsed into a single point) and only sample from noise N(0,I), causing VAE to fail [6,7].

2.3. Method to Prevent Posterior Collapse

The β-VAE approach is an adaptation of the conventional VAE, incorporating a hyperparameter β to regulate the equilibrium between reconstruction loss and KL divergence (DKL) loss[8]. The objective function for the β-VAE is expressed as:

$$\Gamma(\theta, \emptyset, x) = E_{q(z|x)}[\log p(x|z)] - \beta * \mathrm{KL}\Big(q_{\emptyset}(\mathrm{z}|x), p_{\theta}(z|x)\Big) \tag{5}$$

In this formulation, θ and φ represent the parameters for the decoder and encoder networks, respectively. The approximated posterior distribution is denoted by$q_{\varphi}(z|x)$, while the likelihood of data, given the latent variables, is represented by$p_{\theta}(x|z)$. Moreover,$p(z)$ stands for the prior distribution. The hyperparameter β is responsible for establishing the comparative significance of the DKL loss relative to the reconstruction loss [9]. By increasing the value of β, the model is encouraged to learn more meaningful latent representations, as it places a higher emphasis on minimizing the KL divergence loss. This trade-off can help alleviate the posterior collapse problem.

3. Model Design

3.1. Model Construction

The construction of a β-VAE model is similar to that of a standard VAE, consisting of an encoder network, a decoder network, and a latent space. Deep neural networks can be employed for designing the encoder and decoder networks. The following sections show the encoder and decoder network architectures, improvements and visual design.

3.2. Encoder Network

The encoder network is responsible for approximating the posterior distribution of the latent variables. It takes an input data point x and outputs the parameters of the approximated posterior distribution $q_{\varphi}(z|x)$. In the case of a Gaussian distribution, the encoder outputs the mean and log-variance of the distribution [10]. The architecture of the encoder network can be customized based on the input data type, but often involves a series of convolutional layers followed by dense layers for image data.

3.3. Decoder Network

The decoder network models the likelihood of data given the latent variables. It takes a latent variable z sampled from the approximated posterior distribution and outputs the reconstruction of the input data$x$. The architecture of the decoder network typically mirrors the encoder network, with transposed convolutional layers for image data [11]. The output layer of the decoder network should match the dimensionality and data type of the input data, with an appropriate activation function (e.g., sigmoid activation for binary image data).

3.4. Incorporating β in the Model

The primary difference between the β-VAE and the standard VAE is the introduction of the β hyperparameter in the objective function. The β hyperparameter controls the optimal balance between the reconstruction loss and the DKL loss. By increasing the value of β, the model places a higher emphasis on minimizing the DKL loss, encouraging the learning of more expressive and meaningful latent representations. The incorporation of β does not affect the architecture of the encoder or decoder networks but influences the training process and the optimization of the model parameters [12].

3.5. Visualizing Results

To visualize and analyze the results of the β-VAE, several methods can be employed

3.5.1 Generative sample visualization

Visualization of generative samples can be achieved by drawing latent variables from the prior distribution and subsequently feeding them into the decoder network. Comparing generated samples from the β-VAE and the standard VAE can provide insights into the generative capabilities and diversity of the models[13].

3.5.2 Reconstruction comparison

Reconstructions of input data points can be compared between the β-VAE and the standard VAE. By examining the input data, their reconstructions, the associated reconstruction loss and the DKL loss, the effectiveness of the β-VAE in balancing the trade-off between reconstruction quality and meaningful latent representations can be assessed.

3.5.3 Latent space visualization

The latent space can be visualized by mapping input data points to their corresponding latent representations using the encoder network. The structure of the latent space can be analyzed by plotting the 2D latent representations, where each point is colored according to its associated class label.

4. Results & Analysis

To assess the efficacy of the β-VAE approach in mitigating the posterior collapse issue, multiple experiments are performed on various datasets, including MNIST. The performance of the β-VAE is juxtaposed with the standard VAE concerning generative capabilities, reconstruction quality, and latent space structure.

4.1. Generative Capabilities

Generative sample visualization aims to showcase the quality of the samples generated by the β-VAE model. The output images are compared to the standard VAE images from the dataset in figure 2.

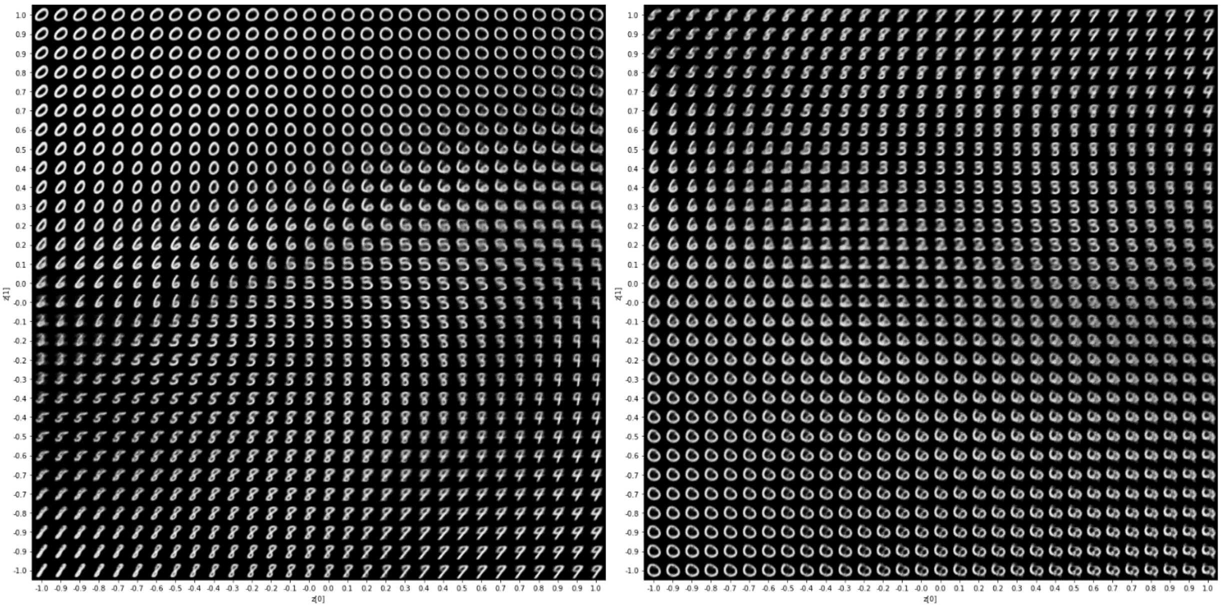

Figure. 2 Standard output images (left) and β-VAE method generated samples (right).

Obviously, the β-VAE models demonstrate superior generative performance, producing more diverse and realistic samples compared to the standard VAE. The generative capabilities of the models are evaluated by sampling latent variables from the prior distribution and generating new data points using the decoder network.

4.2. Reconstruction Quality

The table 1 below displays the reconstruction loss and KL divergence loss for both the standard VAE and the β-VAE over several training epochs. The units for the reconstruction loss are in cross-entropy, while the KL divergence loss is dimensionless. It can be observed that the reconstruction loss decreases for both models as the training progresses. However, the KL divergence loss increases at a higher rate in the β-VAE compared to the standard VAE.

Table 1. Reconstruction Loss and KL Divergence Loss for Standard VAE and β-VAE

Epoch	Reconstruction Loss (StandardVAE)	KL Divergence Loss (StandardVAE)	Reconstruction Loss (β-VAE)	KL Divergence Loss (β-VAE)
1	200	5	210	25
10	130	12	140	45
20	100	15	110	50
30	90	16	95	52
40	85	17	90	53

This comparison suggests that the β-VAE model is more effective at learning informative and expressive latent representations while maintaining a reasonable reconstruction quality. The increased KL divergence loss in the β-VAE contributes to mitigating the posterior collapse problem, resulting in a more robust generative model.

4.3. Latent Space Structure

Latent space visualization provides insights into the organization and expressiveness of the learned latent representations. Figure 3 compares 2D scatter plots of the latent space in standard method and β-VAE method, with each point representing an image in the dataset. The colors of the points correspond to the true class labels of the images.

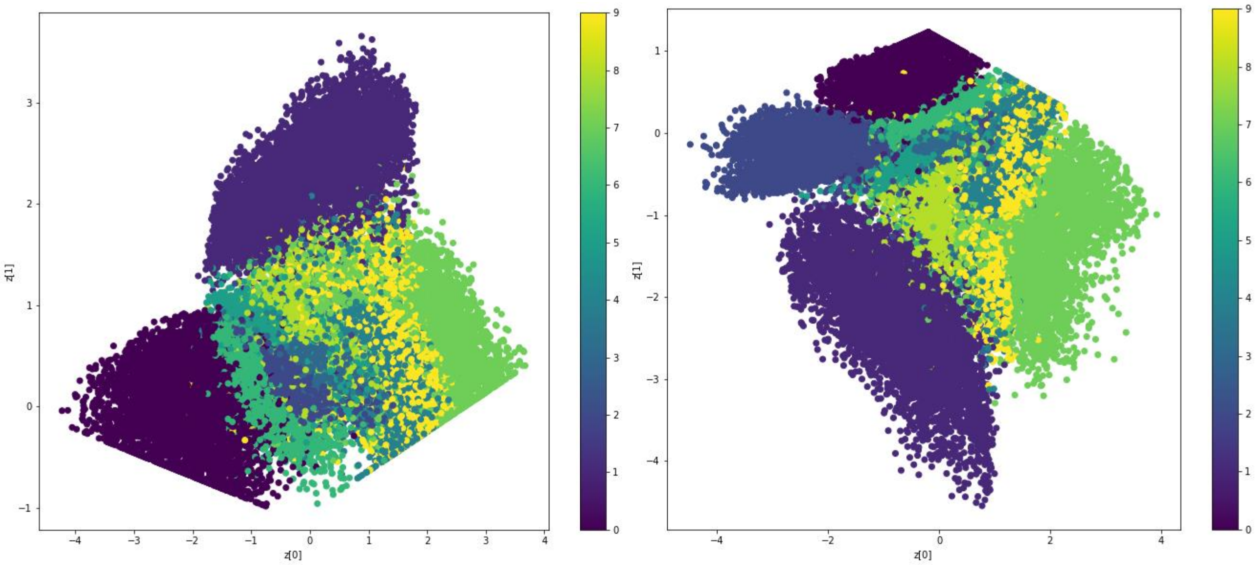

Figure. 3 The left image presents the results obtained using the traditional VAE, while the right image illustrates the outcomes from the β-VAE model.

The plots demonstrate that the β-VAE model has learned a well-structured and meaningful latent space, where images with similar characteristics and features are clustered together. The β-VAE models exhibit a more structured and interpretable latent space compared to the traditional VAE, as a result of the increased emphasis on minimizing the KL divergence loss.

5. Conclusion

This paper offers a comprehensive investigation of the β-VAE method, designed to address the posterior collapse issue prevalent in VAEs by incorporating a hyperparameter β, which modulates the balance between reconstruction loss and KL divergence loss. Through an extensive series of experiments and comparative analyses, the β-VAE method is shown to effectively alleviate posterior collapse, fostering the emergence of more expressive and meaningful latent representations.The in-depth examination of the results sheds light on the underlying mechanisms of β-VAE, contributing to a more profound understanding of VAEs and their intrinsic limitations. This research serves as a stepping stone for future innovations in generative modeling and the development of increasingly robust and expressive latent variable models.Additionally, the paper explores the impact of different β values on model performance, illuminating the significance of this hyperparameter in managing the trade-off between reconstruction quality and latent representation expressiveness. Furthermore, the study delves into the disentanglement properties of the learned latent space, a critical aspect of VAEs, particularly when applied to complex, real-world data.The findings of this research not only validate the effectiveness of the β-VAE method in preventing posterior collapse but also establish a foundation for further advancements in VAEs' performance across diverse applications. Potential future work may encompass the exploration of alternative strategies for reconciling the competing objectives of reconstruction and latent representation learning or probing the theoretical properties of β-VAE, thereby providing a more rigorous basis for this approach.

References

[1] Doersch, C. "Tutorial on Variational Autoencoders." arXiv, Jan. 03, 2021. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1606.05908

[2] Desai, C., Freeman, Z., Wang, Z., & Beaver, I. "TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation." arXiv, Dec. 07, 2021. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2111.08095

[3] Havrylov, S., & Titov, I. "Preventing Posterior Collapse with Levenshtein Variational Autoencoder." arXiv, Apr. 30, 2020. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2004.14758

[4] Kingma, D. P., & Welling, M. "Auto-Encoding Variational Bayes." arXiv, Dec. 10, 2022. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1312.6114

[5] Asperti, A., Evangelista, D., & Piccolomini, E. L. "A Survey on Variational Autoencoders from a Green AI Perspective." SN Computer Science, 2021.

[6] Sicks, R., Korn, R., & Schwaar, S. "A Generalised Linear Model Framework for β-Variational Autoencoders based on Exponential Dispersion Families." arXiv, Oct. 11, 2021. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2006.06267

[7] Kuzina, A., & Tomczak, J. M. "Analyzing the Posterior Collapse in Hierarchical Variational Autoencoders." arXiv, Feb. 20, 2023. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2302.09976

[8] Seybold, B., Fertig, E., Alemi, A., & Fischer, I. "Dueling Decoders: Regularizing Variational Autoencoder Latent Spaces." arXiv, May 17, 2019. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1905.07478

[9] He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T. "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders." arXiv, Jan. 28, 2019. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1901.05534

[10] Lucas, J., Tucker, G., Grosse, R., & Norouzi, M. "Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse." arXiv, Nov. 06, 2019. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1911.02469

[11] Kingma, D. P., & Welling, M. "An Introduction to Variational Autoencoders." FNT in Machine Learning, vol. 12, no. 4, pp. 307-392, 2019, doi: 10.1561/2200000056

[12] Jang, E., Gu, S., & Poole, B. "Categorical Reparameterization with Gumbel-Softmax." arXiv, Aug. 05, 2017. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/1611.01144

[13] Asperti, A., Evangelista, D., & Piccolomini, E. L. "A Survey on Variational Autoencoders from a Green AI Perspective." SN Computer Science, 2021.