A Comparison of The Methods of Alleviating The Posterior Collapse Problem for VAE and ControlVAE

David Canagasabey 1004301588
Runtian Wang 1003454102
Guanjie Wang 1004835422
Jun Wei Wong 1004317731

Abstract

Variational autoencoders (VAE) can model data with a latent-variable model and then fit the model by maximizing a lower bound of log marginal likelihood. However, they often suffer from a problem known as the "posterior collapse". The posterior collapses when it is equal to the prior and the posterior is independent of the input data. We investigate how architectures influence the posterior collapse, especially the newly proposed ControlVAE [1] model. Previous studies have pointed out adding skip connection and introducing aggressive training are ways to address the posterior collapse problem. We proposed two methods of the skip connections added to the model and experiment with basic variational autoencoders implementations. We study the "posterior collapse" effect on images (MEDMNIST) and text (Yelp) dataset and compare them with the corresponding architecture with a controller. We conclude that ControlVAE improves upon the standard VAE models, but fails to substantially improve upon aggressive encoder training and skip connections. We also observed the alternative skip connection model which adds skip connections between the encoder network and decoder network improves reconstruction but does not provide benefits in alleviating posterior collapse.

1 Introduction

A variational autoencoder (VAE) [2] is a composition of a inference model $q_\phi(\mathbf{z} \mid \mathbf{x})$ (encoder) and a generative model $p_\theta(\mathbf{x} \mid \mathbf{z})$ (decoder) under prior $p(\mathbf{z})$ and observations x. The encoder approximates the marginal distribution of the latent variables $\mathbf{z}$ over observations $\mathbf{x}$, and the decoder model the distribution over latent variables $\mathbf{z}$. The training objective of VAEs is to maximize the Evidence Lower Bound (ELBO), which is a sum of the reconstruction loss and a KL divergence term which is the difference between the learned approximate posterior of the latent variables and their true distribution:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))$$

However, VAEs often suffer from a problem known as the "posterior collapse". Due to the expressive power of neural networks, $\mathbf{z}$ and $\mathbf{x}$ could be independent under $q_\phi(\mathbf{z} \mid \mathbf{x})$ when the trained parameters achieve a local optimum where learned approximate posterior closely approximates the prior: $q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p(\mathbf{z})$. This is not a desired behaviour as an important goal of VAEs is to learn meaningful latent features from the observations. Without meaningful representation of the observations in the latent dimensions, the capacity of the generative model quickly deteriorates.

In this paper,we will investigate ways to address the posterior collapse problem. We will explore its effect on a newly proposed ControlVAE model, and investigate if adding skip connection layers to the generative model of VAE helps to alleviate the problem.

2 Related Work

To understand the motivation behind ControlVAE, we need to first look at the β-VAE [3] framework. β-VAE is a modification to the basic VAE model with emphasis on learning a disentangled representation of the latent variables $\mathbf{z}$. A latent variable is said to have a disentangled representation when the latent variable is only sensitive to a single generative factor (the underlying factor that generates the observations $\mathbf{x}$) and relatively invariant to other generative factors. The benefit of this disentangled representation includes good interpretability and easy generalization to tasks of different nature. To achieve that, β-VAE introduced an extra hyperparameter β, a weight for the KL-divergence term in the VAE objective function:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))$$

However, β-VAE tends to apply large β to the objective function, causing the optimization algorithm to prioritize the second term and sacrifices the reconstruction accuracy. ControlVAE addresses this issue by employing a PID controller [4], an algorithm borrowed from control theory, to control the weight β. At each training step, ControlVAE samples from the output of the KL-divergence term and feeds it to the PID Controller that tunes the hyperparameter β accordingly, in order to stabilize the KL-divergence term at a desired value, therefore preventing KL vanishing.

Some more recent methods have moved away from weighing the KL-divergence term. Many proposed solutions to posterior collapse focus on weakening the decoder or modifying the training objective. Dieng et al. [5] have proposed adding residual connections (or "skip connections") to the decoder network to force a stronger connection between the latent representation and the output.

He et al. [6] investigated the posterior collapse problem from the perspective of training dynamics. They suggested that the posterior approximation $q_\phi(\mathbf{z} \mid \mathbf{x})$ often lags far behind the true model posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ in the initial stages of training. This lagging behaviour, drives the generative model towards a collapsed local optimum (posterior collpase). He et al. hypothesize that this is due to the optimization of the inference network and the generative network is imbalanced during training, and proposed to separate the optimization of the two:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{X}; \theta, \phi^*), \text{ where } \phi^* = \underset{\phi}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{X}; \theta, \phi)$$

Where the optimization of the inference network $q_\phi(\mathbf{z} \mid \mathbf{x})$ is performed separately within the the training process. This method aggressively optimizes the inference network with more updates to reduce lag.

Our goal is to expand on the existing work by providing a level comparison of these methods (as well as some new methods we propose) using our chosen metrics. We will also explore how a combination of these methods may improve their impact in alleviating posterior collapse.

3 Methods

3.1 Skip-VAE

We add the skip connections into VAE architectures in the following two ways. SKIP-VAE-I architecture is shown in Figure 1. It adds the skip connection layers from latent variables into the decoder layers. Notice this matches the proposal skip connection from Dieng [5]. SKIP-VAE-II architecture is shown in the right of Figure 1. It adds the skip connection layers from the encoder layers to the matched decoder layers without adding skip connection from latent space.

3.2 Inference Network Optimization

As previously mentioned, it has been demonstrated that the initial stages of training the inference network often fail to approximate the model's true posterior [6], which causes posterior collapse as the model is encouraged to ignore the latent encoding. Therefore, optimizing the inference network before performing model updates can reduce inference lag and help avoid the posterior collapse problem [6]. To do this, we trained the encoder more aggressively than the decoder in the initial stages.

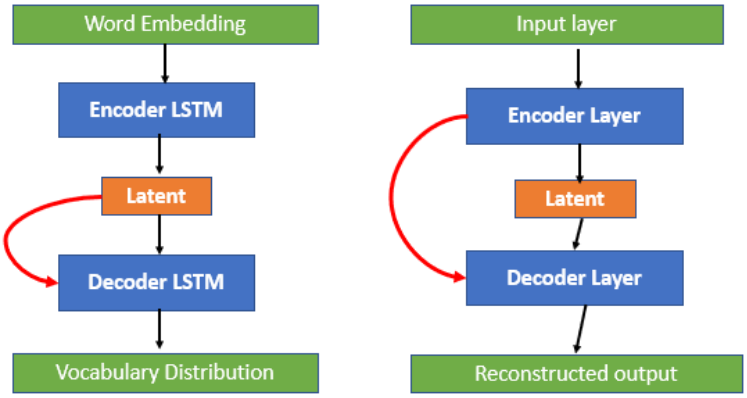

Figure 1: The two architectures for the skip-VAE. (Left: SKIP-VAE-I. Right: SKIP-VAE-II)

3.3 ControlVAE

The ControlVAE model uses a PID controller [4] to dynamically adjust the beta parameter in the loss function during training. After every epoch, the parameter is updated according to the KL-divergence observed in that epoch and a summation of KL-divergences in previous epoch (which acts as an estimated integral).

4 Experiments

We explored the posterior collapse problem in variational autoencoders and investigated different methods of circumventing this problem for VAEs and the recently proposed ControlVAE [1].

4.1 Datasets

There are two datasets used for our experiments:

- Image generation: The MedMNIST dataset [7] was used. This dataset consists of a collection of 10 pre-processed medical datasets with images of size 28 x 28. For our experiments, we used the PathMNIST dataset for pathology. PathMNIST contains 89,996 training images, and 7,180 test images from 9 classes. The images are resized to 128 x 128 images before they are used in our models.
- Text generation: A subset of the Yelp review dataset [8] was used, which consist of user-created reviews (with metadata) from the Yelp website. From this dataset, we used (text, rating) pairs. Our subset consists of 6126 training examples, 1312 validation examples, and 1312 test examples.

4.2 Experiments on Image Generation

Firstly, we modified the proposed VAE and ControlVAE models [1] and trained them on the PathMNIST dataset 900 iterations. We then experimented with the following methods of diminishing the posterior collapse problem on both models:

- Include skip connections in the VAE decoder (ie. at hidden layers of the decoder neural network, the input will include the output of the previous hidden layer as well as the outputs of intermediate encoder layers).
- Train the encoder more aggressively than the decoder in first 400 iterations as described in the Methods section in an attempt the improve the inference network. For every time the decoder weights were updated, we updated the encoder weights 5 times.

4.3 Experiments on Text Generation

We use recurrent versions of the VAE and ControlVAE models to experiment on our text dataset. We also experimented with adding skip connections to each model, as proposed by Dieng et al. [5].

4.4 Evaluation Metrics

Despite the large volume of work done on posterior collapse, there is no consistent ways of measuring this issue. We evaluated the posterior collapse for each model and each method in 3 ways:

- Visualize the learned latent representations as t-SNE embeddings
- Define the (ϵ, δ)-collapse of latent dimension i to be [9]:

  $$\mathbb{P}[KL(q(z_i|x)||p(z_i)) < \epsilon] \geq 1 - \delta$$

  Delta was fixed as 0.01 (1%), while epsilon was varied from 0 to 1. This provides the posterior collapse percentage as a function of epsilon. This percentage is the percentage of latent dimensions that are within ϵ KL divergence of the prior for at least $(1 - \delta)\%$ over all test data [9].
- The mutual information can defined as the following [5]:

  $$\mathcal{I}_q(\mathbf{x}, \mathbf{z}) = KL(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - KL(q_\phi(\mathbf{z})||p(\mathbf{z}))$$

  This will be estimated by using Monte Carlo estimates of the KL terms [5]. Higher mutual information indicates less posterior collapse.

5 Results and Discussion

Table 1: Posterior Collapse Measurement Metrics (Negative ELBO, Construction Loss, KL-divergence) on MEDMNIST dataset using Controller VAE (PID-VAE), VAE, Controller VAE with Skip Connection I (PID-SKIP-I), Skip Connection I (SKIP-I) , Controller VAE with Aggressive Training VAE (PID-Agg), Aggressive Training VAE (Agg) at convergence after training

	PID-VAE	VAE	PID-SKIP-I	SKIP-I	PID-Agg	Agg
Negative ELBO	**283.888**	458.188	**32.753**	65.195	**309.961**	460.499
Construction-Loss	**256.066**	399.072	**32.753**	64.710	**264.023**	407.711
Total KL-divergence	**198.438**	59.116	262.725	**0.485**	247.829	**52.788**
Mean KL-divergence	**0.397**	0.118	0.525	**0.001**	0.496	**0.106**
MI	**14.9150**	9.1210	**18.9913**	5.4777	**15.6894**	9.9914

5.1 Controllers Reduce Posterior collapse

5.1.1 Image Generation

Table 1 summarizes the results of the posterior collapse metrics on VAE, Skip-I-VAE, Agg-VAE with or without controller on MEDMNIST dataset. In general, models with controllers have better performance than those without the controller, showing lower training loss and higher reconstructed image quality. This can also be observed from the train loss and reconstruction loss curves in Figure 5.2.1, as well as the reconstructed images in Figure 9 in the Appendix. This is expected as the ControlVAE automatically adjusts the β hyperparameter, which helps stabilize the KL-divergence.

Based on the evaluation metrics we used for posterior collapse, we observe that the ControlVAE does not suffer as much from the posterior collapse problem. Looking at Figure 6 of the graphs produced by measuring posterior collapse percentage as a function of ϵ and δ in the Appendix, we see that the VAE begins to suffer from posterior collapse for very small epsilons, while the ControlVAE is less prone to this. Table 1 also shows higher mutual information for ControlVAE, which indicates less posterior collapse.

Although there is a noticeable improvement in image reconstruction quality for models using controllers, our experiments show that the model architecture is still the most crucial factor. The reconstruction loss and negative ELBO in both VAE and ControlVAE decreased significantly after

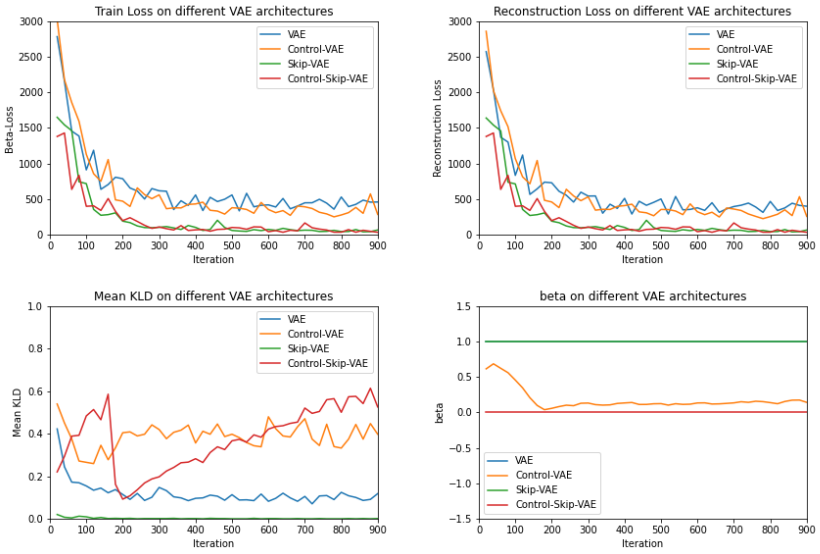

Figure 2: Training loss and KL divergence curves. Note that beta is 1 for the VAE without the Controller

adding the skip connection layers. However, one interesting observation is that the VAE and ControlVAE with skip connection both suffer more from the posterior collapse problem. This may be due to the decoder receiving extra information from the skip connection layers, which may encourage the model to ignore the latent encoding and focus instead on the additional features provided to it.

The experiment of training the encoder more aggressively in the initial stages did not yield the results we expected. The performance of the aggressively trained models is very similar to the models without the aggressive training. We can see from Table 1 that this did not improve the losses of the VAE and ControlVAE. Observing from the posterior collapse measurements as well, the posterior collapse was not decreased by much. The similarity in performance shows that the problem might not be related to the beginning stages of the encoder training in our setup.

5.1.2 Text Generation

Table 2: Posterior Collapse Measurement Metrics (Negative ELBO, KL-divergence, Mutual Information) on our **test** subset of the Yelp dataset using VAE, ControlVAE, VAE with latent variable skip connections (SkipVAE-I), and ControlVAE with latent variable skip connections (ControlSkipVAE) after training

	VAE	ControlVAE	SkipVAE-I	ControlSkipVAE
Negative ELBO	673.4669	**667.3101**	671.7986	**665.5868**
KL-divergence	0.2313	**0.0966**	**7.1003**	7.1476
MI	0.1900	**0.4375**	**13.6456**	13.6152

We observe similar results on Yelp dataset as shown in Table 2. It summarizes the results of the posterior collapse metrics on VAE and SkipVAE-I with or without a PID controller when training on our subset of the Yelp dataset. In general, models with controllers have better performance than those without the controller, showing lower training loss. However, PID with skip connection has worse KL divergence compared to VAE with skip connection. Though there is a noticeable improvement in reconstruction for models using controllers, the architecture structure is still the most crucial factor in addressing the posterior collapse problem. This is further supported by Figure 3, which shows that ϵ-δ collapse is negligibly impacted by the addition of a PID controller during training. Due to the significant quality difference between each architecture, adding a controller can only decrease the performance gap.

Figure 5 shows the tSNE visualizations of the latent spaces learned by each model. The

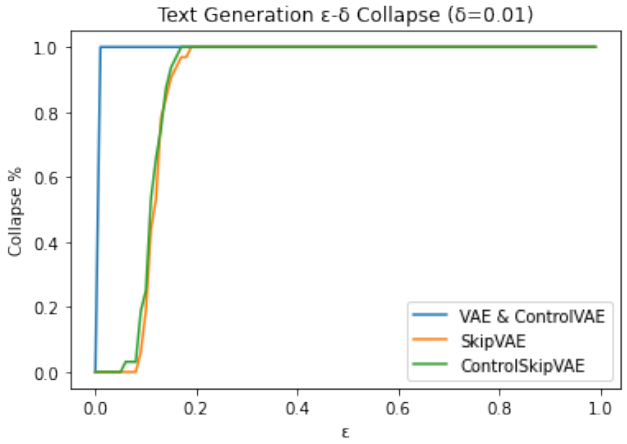

Figure 3: ϵ-δ collapse on **test** subset of Yelp dataset ($\delta = 0.01$). Results for VAE and ControlVAE are identical

vectors are coloured by the star rating associated with the review that it represents, which acts as an indicator for the sentiment of the text. We see that the ControlVAE has a latent distribution which is closer to a standard normal distribution than that of the naive VAE model. We also see that for both the naive VAE and SkipVAE-I, adding a PID controller improves the clustering of the latent vectors by star rating. However, this improvement is negligible for the SkipVAE-I.

5.2 Discussion

5.2.1 Skip Connection

We notice that with both Skip-VAE-I and Skip-VAE-II architectures, there are significant improvements in the decoded images shown in Appendix Figure . Skip-VAE-I can help alleviate the posterior collapse problem but Skip-VAE-II makes it worse based on Table 1 and Table 2. The Skip-VAE-I model addresses the posterior collapse problem directly by creating a functional link between the latent vector and the output layer of the decoder, which makes the training more robust. However, the Skip-VAE-II model creates a skip connection between encoders and decoders. Feeding encoder inputs to the decoder significantly help with the output reconstruction, but it weakens the training to latent space results in the model suffers more from posterior collapse.

5.2.2 Controller

In both image and text domains, we observe that using a PID controller for the beta parameter is effective for a standard VAE model, but minimally important and possibly a detriment for a SkipVAE model. The PID controller addresses posterior collapse by balancing the importance of the KL-divergence during training, which indirectly emphasizes the link between the latent space and reconstructed data. In this sense, the PID controller *indirectly* addresses an issue that the SkipVAE model addresses *directly*. This would result in the impact of adding the PID controller on the SkipVAE being negligible, while still being effective on a naive VAE model, which is what we observe.

6 Limitations and Future Work

Due to limitations in our available resources, our text generation models were trained using a relatively small dataset (in comparison to other natural language models) and using a relatively outdated language model. We suggest that further research be done into how these methods behave on state-of-the-art language models (ex. transformers). Our work also uses a fixed set of hyperparameters. It is worth investigating how the interaction of the studied methods varies across multiple hyperparameter settings (ex. latent space dimension, number of hidden layers).

7 Conclusion

We have evaluated the ControlVAE model and its' ability to alleviate the posterior collapse problem. We compared the model with other proposed methods and experimented with how these methods interact with each other. We observed that, in relation to this problem, ControlVAE improves upon the standard VAE model, but fails to substantially improve upon aggressive encoder training and skip connections. We also proposed an alternative skip connection model which adds skip connections between the encoder network and decoder network. We observed that this model improved reconstruction, but does not provide benefits in alleviating posterior collapse.

8 Code

The code used for this report is available on a Github repository at `https://github.com/guanjiew/csc412_vae`.

References

[1] H. Shao, S. Yao, D. Sun, A. Zhang, S. Liu, D. Liu, J. Wang, and T. Abdelzaher, "Controlvae: Controllable variational autoencoder," 2020.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[3] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[4] K. Åström and T. Hägglund, *Advanced PID Control*. ISA-The Instrumentation, Systems, and Automation Society, 2006.

[5] A. B. Dieng, Y. Kim, A. M. Rush, and D. M. Blei, "Avoiding latent variable collapse with generative skip models," 2019.

[6] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick, "Lagging inference networks and posterior collapse in variational autoencoders," 2019.

[7] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," *arXiv preprint arXiv:2010.14925*, 2020.

[8] "Yelp open dataset."

[9] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Don't blame the elbo! a linear vae perspective on posterior collapse," 2019.

9 Appendix

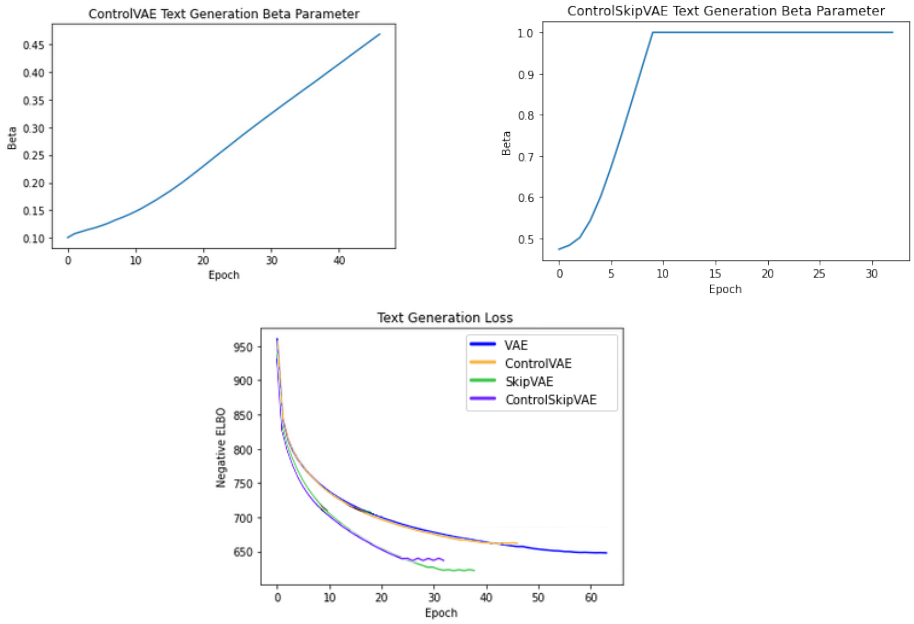

Figure 4: Beta parameter (coefficient of KL-divergence in loss function) and negative ELBO on **training** subset of Yelp dataset. For models without a PID controller, the beta parameter is always 1

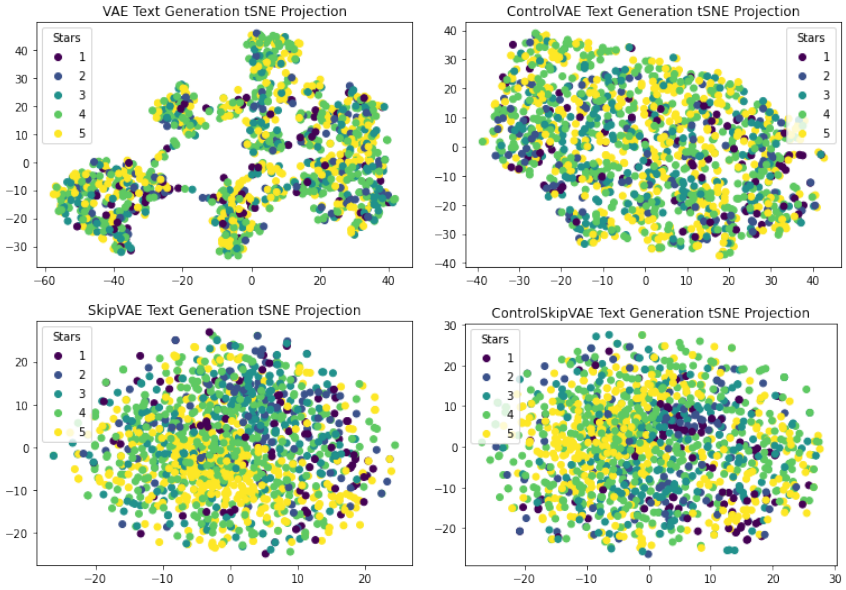

Figure 5: tSNE projections of **test** subset of Yelp dataset into latent space of each model, coloured by the star rating of the review

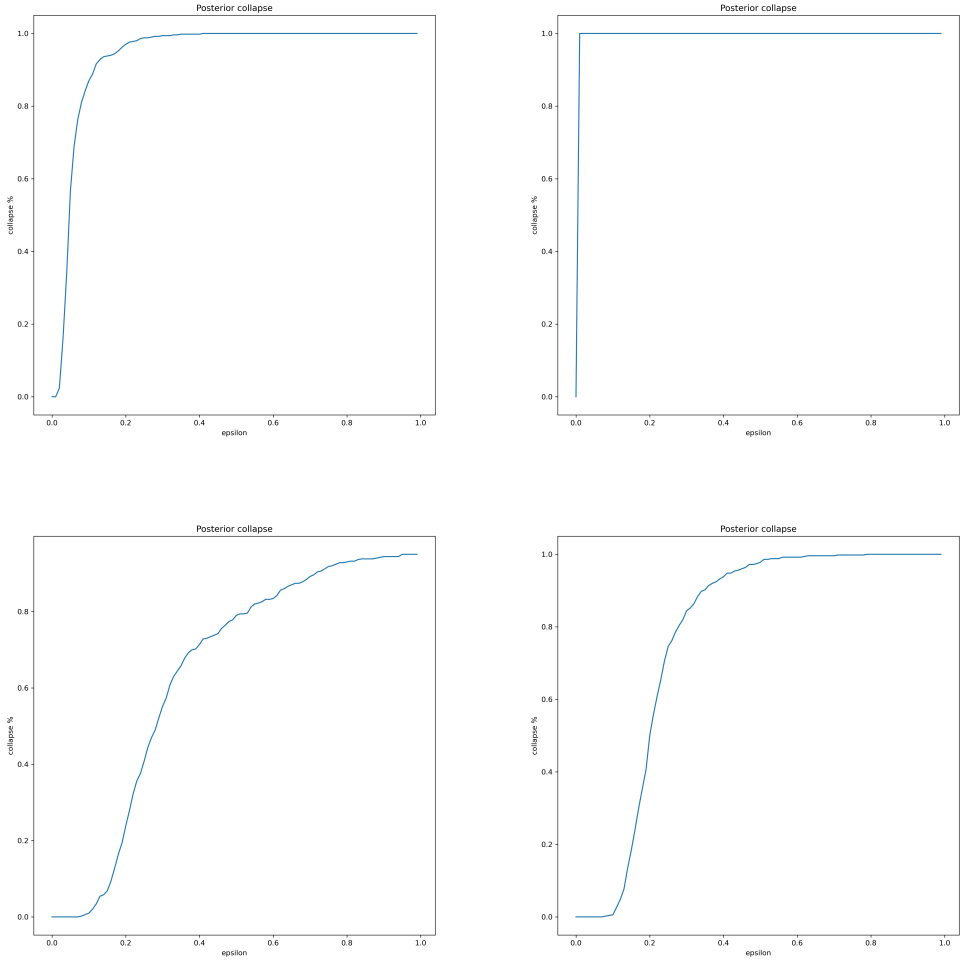

Figure 6: ϵ-δ collapse on **test** subset of MedMNIST dataset ($\delta = 0.01$) using VAE (top-left), SkipVAE-I (top-right), ControlSkipVAE (bottom-left), and ControlVAE (bottom-right) models

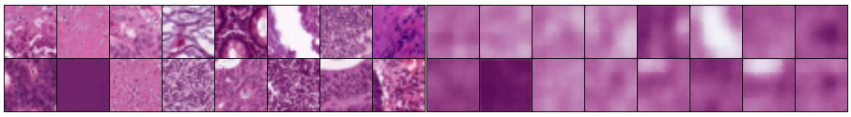

Figure 7: Reconstructed images using the VAE model

Figure 8: Reconstructed images using the ControlVAE model

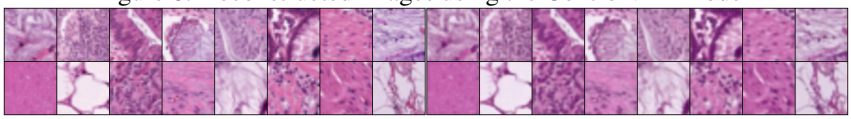

Figure 9: Reconstructed images using the VAE model with skip connections

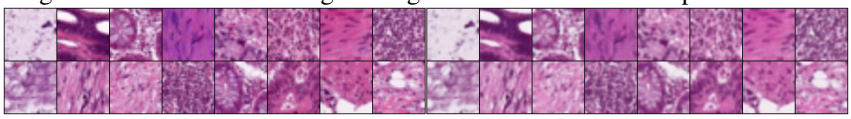

Figure 10: Reconstructed images using the ControlVAE model with skip connections