# 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

We will test how the posterior collapse problem is affecting the newly proposed ControlVAE model. To resolve the posterior collapse problem in VAE, we are proposing adversarial training method and pretrained VAE method (details described in the technical aspect), and compare with the previously proposed methods.

## 1 Background

### 1.1 Variational Autoencoders

The variational autoencoder (VAE) was introduced by Kingma & Welling (2014) as a method for variational inference using neural networks. The method consists of two models: an inference model $q_\phi(\mathbf{z} \mid \mathbf{x})$ (often called the "encoder") and a generative model $p_\theta(\mathbf{x} \mid \mathbf{z})$ (often called the "decoder"). We also assume a prior $p(\mathbf{z})$. The goal for the decoder is to model the marginal distribution of the observation over the latent vector:

$$p(\mathbf{x}) \approx \int p_\theta(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

The goal of the encoder is to approximate the marginal distribution of the latent vector over the observation.

### 1.2 Evidence Lower Bound

The models are trained simultaneously to maximize the Evidence Lower Bound (ELBO) (Chen et al., 2017). ELBO is typically characterized in two forms. The first form is a sum of a "reconstruction loss" (which grows as the decoder fails to reconstruct the input given to the encoder) and a regularizer (which pushes the approximate posterior closer to the prior):

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

The second form is as a sum of the log-likelihood and a KL term which measures the difference between the variational posterior and the model posterior.

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z} \mid \mathbf{x}))$$

The former is tractable and used for calculating gradients, while the latter is useful for analysis.

### 1.3 Posterior Collapse

Previous work has demonstrated that VAEs often suffer from a problem referred to as "posterior collapse". When the model parameters $\phi$ and $\theta$ are initialized, $\mathbf{x}$ and $\mathbf{z}$ are nearly independent under $q_\phi(\mathbf{z} \mid \mathbf{x})$ and $p_\theta(\mathbf{x} \mid \mathbf{z})$ (He et al, 2019). Due to the expressive power of neural networks, it is possible for trained parameters to achieve a local optimum where the encoder approximates the prior ($q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p(\mathbf{z})$) and the decoder approximates the posterior ($p_\theta(\mathbf{x} \mid \mathbf{z}) \approx p(\mathbf{x})$) while still maintaining this independence, thereby ignoring the latent variable $\mathbf{z}$ (Bowman et al., 2016).

## 2 Related Work

The most common approaches to this problem involve regulating the KL term ($D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))$) in the loss function. KL cost annealing (Bowman et al, 2016) consists of applying a coefficient to the KL term that starts at 0 and gradually increases to 1 throughout training. A later work by Higgins et al. (2017) proposed applying a constant weight $\beta$ to the KL term throughout training, although posterior collapse was not the primary motivation for this approach. ControlVAE (Shao et al, 2020) is a more recent model which uses a PID controller (Åström et al, 2006), an algorithm borrowed from control theory, to automatically tune the $\beta$ parameter.
Some more recent methods have moved away from weighing the KL regularizer. Many proposed solutions to posterior collapse focus on weakening the decoder (Yang et al., 2017; Semeniuta et al., 2017) or modifying the training objective (Zhao et al., 2017; Tolstikhin et al., 2018). He et al. (2019) have proposed that posterior collapse is caused by the encoder lagging behind the decoder during training, and implemented a change to the training algorithm (where the encoder is trained to convergence before each training step of the decoder) to balance the optimization of the encoder and decoder networks. Dieng et al. (2020) have proposed adding residual connections (or "skip connections") to the decoder network to force a stronger connection between the latent representation and the output.
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 Extended Abstract

Here we will describe the intuition behind our proposed method resolving the posterior collapse problem. As introduced in the background, one possible approach is to improve the encoder to optimize the ELBO objective. We can pretrain the encoder network on a task relevant to the dataset (eg. classification task) to force a dependence between the input image and the encoding before we train the decoder.Add noise to the model input and perform adversarial training. Similar to the previous method, we believe this may help alleviate the posterior collapse problem by improving VAE encoding.

Multiple datasets will be used to perform our experiments. In particular, both an image based dataset and a text based dataset will be included. Text based datasets have been shown to be sensitive to posterior collapse with VAEs [1]. Some potential datasets include MNIST, Omniglot, CelebA, MedNIST and Yelp corpus. These are frequently used in related research.

## 4 Methods

We will be exploring the posterior collapse problem in variational autoencoders and investigating different methods of circumventing this problem. We will be working with VAE and the recently proposed ControlVAE [3].

We will implement both VAE and ControlVAE and test the following methods of diminishing the posterior collapse problem on both models. The first two methods are methods proposed in prior research papers on this topic, and the latter two methods are our ideas for tackling this problem. The results from the methods in prior research papers will be used as a benchmark that we will compare our methods' results against:

- 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 latent variable). As demonstrated by Dieng et al., skip connections help strengthen the links between latent variables and the likelihood function [1].
- He et al. [2] showed that the initial stages of training the inference network often fail to approximate the model's true posterior, which causes posterior collapse as the model is encouraged to ignore the latent encoding. They demonstrated that optimizing the inference network before performing model updates can reduce inference lag and avoid the posterior collapse problem [2].

- Reserve a portion of the training data and pre-train the VAE encoder on it, then train the VAE model on the remaining training data.

- Add noise to the model input and perform adversarial training.

Our experiments will be evaluated on multiple datasets. We will evaluate the posterior collapse for each model and each method in 3 ways. These methods of evaluation were used in related research:

- Visualize the learned latent representations as t-SNE embeddings

- Define the $(\epsilon, \delta)$-collapse of latent dimension i to be [4]:

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

  Delta will be fixed, while epsilon will be varied. This provides the posterior collapse percentage as a function of epsilon. This percentage is the percentage of latent dimensions that are within $\epsilon$ KL divergence of the prior for at least $(1 - \delta)\%$ of the training data [4].

- The mutual information can defined as the following [1]:

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

  This will be estimated by using Monte Carlo estimates of the KL terms as done in Avoiding Latent Variable Collapse with Generative Skip Models [1]. Higher mutual information indicates less posterior collapse.

Comparing the results of the VAE and ControlVAE with the each of the proposed methods, we will draw our conclusions on the effectiveness of each method, as well as ControlVAE's behaviour in regards to the posterior collapse problem.

## 5 Experiments

There are two major milestones we have planned to achieve within the timeframe of the project. The first milestone is to implement and test the recently proposed ControlVAE to find out how susceptible it is to the problem of KL divergence compared to VAE. The second milestone is to implement two existing methods and two proposed methods on ControlVAE and VAE to measure the effectiveness of these methods on alleviating the problem of posterior collapse.

We will evaluate ControlVAE and VAE on different datasets on the task of image generation based on three distinct metics . The metrics, which we have defined in the previous section, are t-SNE for visualization, measuring posterior collapse percentage as a function of $\epsilon$-threshold and measuring the mutual information term as an indicator of posterior collapse. The results of the evaluation will be presented in the report. We will experiment with our proposed methods on alleviating posterior collapse and upon observing notable improvements of our proposed methods on each of the evaluation metrics, that will mark the completion of the project.

## 6 Results and Discussion

## 7 Summary

## References


[1] Dieng, A. B., Kim, Y., Rush, A. M., Blei, D. M. (2019). Avoiding latent variable collapse with generative skip models. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 2397-2405). PMLR.

[2] He, J., Spokoyny, D., Neubig, G., Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. arXiv preprint arXiv:1901.05534.

[3] Shao, Huajie, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. "ControlVAE: Controllable Variational Autoencoder." 37th Proceedings of ICML, 2020.

[4] Lucas, James, George Tucker, Roger Grosse, and Mohammad Norouzi. "Don't Blame the ELBO!A Linear VAE Perspective on Posterior Collapse." NeurIPS 2019.

[5] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR.

[6] Åström, K. J., Hägglund, T. (2006). Advanced PID Control. ISA - The Instrumentation, Systems and Automation Society.

[7] Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., Bengio, S. (2016). Generating Sentences from a Continuous Space. arXiv:1511.06349.

[8] Kingma, D. P., Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv:1312.6114.