Can contrastive learning help fight posterior collapse in VAEs?

Abhijith Asokan

RPTU Kaiserslautern-Landau
asokan@rptu.de

Abstract. Posterior collapse remains a significant challenge in VAE. Over the years, several approaches have been proposed to mitigate this issue. A recent advancement introduces a fresh perspective by incorporating contrastive learning. In this project, we thoroughly explore this innovative approach.

Keywords: VAEs · Posterior collapse · Contrastive learning

1 Introduction

VAEs are a widely successful class of generative models. However, a notable concern in VAEs is their vulnerability to posterior collapse. Over the years, there have been multiple approaches towards mitigating the posterior collapse.

A recent novel and promising approach that utilizes contrastive learning, was proposed in the paper titled "Forget-me-not! Contrastive Critics for Mitigating Posterior Collapse" [1]. The idea is to have a critic work against the posterior collapse.

In this project, we implement the three types of critic described in the paper and evaluate on three different datasets.

2 Fundamentals

2.1 VAE

Let us briefly review the VAE model. VAE describes a general model, where latent variables $\mathbf{z}$ gives rise to the data $\mathbf{x}$. It has 2 parts, the inference network (encoder) and the model network (decoder).

The generative model defines a marginal distribution as follows -

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(z)d\mathbf{z}$$

The generator $p_\theta(\mathbf{x}|\mathbf{z})$ is parameterized by a complex neural network. An auxiliary variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced to approximate the model posterior $p_\theta(\mathbf{z}|\mathbf{x})$.

Training the VAE involves optimizing the ELBO

$$\text{ELBO} = \underbrace{\mathbf{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction loss}} - \underbrace{\text{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{KL regularizer}}$$

2.2 Posterior collapse

Posterior collapse occurs when the training optimizes the objective without learning a meaningful latent space. This happens when the likelihood is flexible enough to learn to always output the data distribution, disregarding the latent entirely.

From ELBO, we see that this keeps the first term high, leaving no incentive to take penalty for the second term. This allows the approximate posterior to exactly match the prior.

2.3 Contrastive learning

Contrastive learning is an unsupervised representation learning technique, where the goal is the learn the representation of data such that similar instances are close together in the representation space (referred to as the contrastive space hereafter), while dissimilar instances are far apart.

3 Contrastive critics for mitigating posterior collapse

(Menon et al.)[1] propose a novel approach using contrastive learning to mitigate posterior collapse in VAEs. The key concept involves employing a *critic* to identify and discourage posterior collapse directly.

When posterior collapse occurs, there is a lack of shared information between the latent variables and the observed data. Essentially, it becomes impossible to match observations to their respective latent representations.

To facilitate this matching, there must exist shared information between each observation and its corresponding latent variable. Following this insight, a critic is designed to enforce this association and is integrated into the VAE objective. The role of the critic is to ensure the neural network preserves mutual information between observations and their corresponding latents. This is strategy is called *forget-me-not* regularization.

3.1 Critic objective

Let f be a function that scores the pairing, assigning high value for correct pairing ($x^+ \leftrightarrow z^+$) and low value otherwise.

Then, the critic objective would be to maximize

$$c(\mathbf{x}, \mathbf{z}) = \mathbf{E}\left[\log \frac{f(x^+, z^+)}{\sum_{x \in X} f(x, z^+)}\right]$$

And the final objective would be to optimize $\mathcal{L} = \text{ELBO} + \lambda c(\mathbf{x}, \mathbf{z})$

3.2 Types of critics

Three types of critics were discussed in the paper. They primarily differ in the way f is computed.

Neural network critic uses a neural network, separate from the VAE to implement the critic. In this project, we implement the neural network critic as 2 separate neural network that maps both x and z to the contrastive space. $f(x, z)$ would be the cosine distance between their representation.

Hybrid critic is similar to neural network critic, except that it shares some initial layers with the variational network.

Self critic doesn't use an additional network. We have $f(x, z) = \log q(z|x)$

4 Experiments

We implement the three types of critic on three datasets and evaluate their performance using standard metrics commonly utilized in prior research on posterior collapse.

4.1 Evaluation metrics

We evaluate all the models on the below metrics

Negative log likelihood (NLL) measures the modeling performance of VAE. A smaller value means better generalization.

Mutual information (MI) measures the shared information. Values close to zero indicates that latents have become independent of the data.

Active units (AU) is a measure of number of latent dimensions that are active.

Density and **coverage** [2] are improved version of precision and recall.

4.2 Common experimental setup

In all the experiments, for neural network critic, we use a two-layer Multi-Layer Perceptron (MLP) for z, and for x we use an encoder (of the same architecture as the encoder of VAE) followed by a two-layer MLP.

For hybrid critic, the setup is similar as above but few of the initial layers of VAE's encoder are shared with the critic's encoder for x.

4.3 Experiments on toy dataset

For toy experiment, we use a Gaussian Mixture Model with 10 classes with 20k samples for each class.

The VAE model has a two-layer MLP for encoder and decoder. The neural critic also uses two-layer MLP.

A hybrid critic was not used for this setup, as the encoder network was too shallow to actually have a shared layer with the critic.

Table 1. Results on the toy dataset

Results are average over 3 runs, with standard deviation in parenthesis.
30 epochs, with $\lambda = 20$

Model	NLL	MI	AU	Density	Coverage
VAE	**144.09** (12.59)	5.38 (0.15)	4.0 (0.0)	0.69 (0.09)	**0.20** (0.05)
Self critic	159.54 (5.11)	**6.79** (0.01)	4.0 (0.0)	0.63 (0.03)	0.14 (0.03)
Neural critic	158.69 (18.64)	5.78 (0.08)	4.0 (0.0)	**0.73** (0.04)	0.15 (0.02)

4.4 Experiments on image dataset

We use preprocessed Omniglot dataset from (Kim et al.)[3]'s experiment.

The VAE uses a CNN encoder with 5 convolutional layers and a final fully connected layer. The decoder has 5 transposed convolutional layers.

For the hybrid critic, we share the initial 4 convolutional layers with the VAE's encoder.

Table 2. Results on the Omniglot dataset

Results are average over 3 runs, with standard deviation in parenthesis.
30 epochs, with $\lambda = 20$

Model	NLL	MI	AU	Density	Coverage
VAE	**64.29** (0.0)	0.03 (0.03)	32.0 (0.0)	0.99 (0.03)	**0.97** (0.0)
Self critic	67.54 (6.48)	**3.86** (0.03)	32.0 (0.0)	0.97 (0.11)	0.87 (0.05)
Hybrid critic	71.77 (0.29)	3.74 (0.01)	32.0 (0.0)	0.91 (0.28)	0.48 (0.12)
Neural critic	71.29 (0.26)	3.77 (0.04)	32.0 (0.0)	**1.02** (0.01)	0.63 (0.01)

4.5 Experiments on text dataset

We use Yahoo dataset with the train/val/test split provided in the experiments from [4]. The vocabulary size was approximately 20k.

A LSTM encoder-decoder architecture, similar to the one used in the experiments from [4], is used for VAE. For the purpose of computing density/coverage, a greedy-decode method is used to generate samples using the decoder.

In the hybrid critic setup, the embedding layer of the VAE's encoder is shared with the critic encoder for x .

Table 3. Results on the Yahoo dataset

Results are average over 3 runs, with standard deviation in parenthesis.
15 epochs, with $\lambda = 40$

Model	NLL	MI	AU	Density	Coverage
VAE	**373.45** (1.72)	0.04 (0.05)	32.0 (0.0)	0.07 (0.08)	0.23 (0.2)
Self critic	433.62 (12.4)	**4.09** (0.02)	32.0 (0.0)	0.13 (0.03)	0.24 (0.06)
Hybrid critic	380.66 (1.37)	4.07 (0.11)	32.0 (0.0)	0.07 (0.05)	0.21 (0.11)
Neural critic	379.28 (1.62)	3.95 (0.25)	32.0 (0.0)	**0.22** (0.12)	**0.38** (0.06)

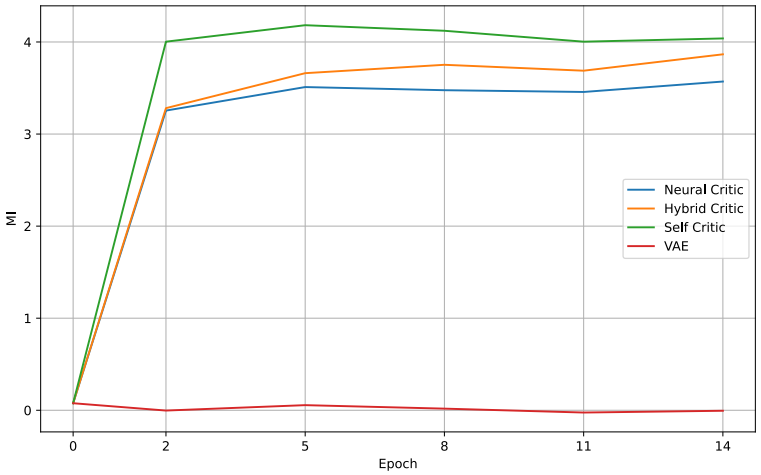

Fig. 1. Mutual information over epochs for Yahoo dataset

5 Discussion on the results

The self critic achieves the highest Mutual Information (MI) values, while the Neural critic excels in the density metric. Standard VAE has the best NLL scores.

6 Conclusion

In this project, we ventured into the novel concept of bridging VAEs and contrastive learning. Through a series of experiments, we observe a notable trend - the utilization of contrastive critics consistently enhances the mutual information between observations and latent variables. This compelling observation serves as a strong indicator of their efficacy in combating the issue of posterior collapse.

References

1. Sachit Menon, David Blei, and Carl Vondrick. Forget-me-not! contrastive critics for mitigating posterior collapse, 2022.
2. Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020.
3. Yoon Kim, Sam Wiseman, Andrew C. Miller, David Sontag, and Alexander M. Rush. Semi-amortized variational autoencoders, 2018.
4. Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders, 2019.