AIML 425 | Assignment 4
Variational Autoencoders

Irshad Ul Ala | Student ID 300397080
Victoria University of Wellington

September 8, 2021

1 Introduction

In this study, we examine variational autoencoders, question 2 of the given assignment. The additional term, "variational", represents the relationship between the regularization and variational inference in statistics. Its aptitude for generation of basic shapes (rectangles, circles or triangles), and the information rate of the latent layer are subjects of interest in this paper. Additionally, we compare this standard variational autoencoder, to one that has its latent layer controlled to be Gaussian instead of a sampling layer, using the Maximum Mean Discrepancy(MMD) objective function.

2 Theory

2.1 Variational Autoencoders

Autoencoders are neural networks that attempt to reconstruct the input data given, a type of feature extraction algorithm. This is first done by first attempting to encode the data into a latent vector of lower dimensions then the given input data, typically via convolutional layers. A visual example would be attempting to decompose a 3 dimensional spiral into 1 dimension, θ, the phase of the spiral. The latent vector is then decoded to form the original input as closely as possible, the difference with which being quantified as the **reconstruction loss**. This network is trained to minimize said reconstruction error. It is important to note that the autoencoder is only trained to encode and decode with as low a reconstruction loss as possible, regardless of how the latent space is organised.

A **variational** autoencoder(VAE) is one which has its encodings distributions regularized to ensure that its latent vector space is continuous and complete, enabling generative process. Instead of encoding each data point as a single point in latent space, we encode it as a distribution. In the decoder segment, a point is sampled from that distribution, to generate the new datapoint or set. This is done by introducing a second loss or regularization term, that tends to regularise the organisation of the latent space by making the distributions returned by the encoder close to a standard normal distribution - the Kullback-Leibler divergence. The Kullback-Leibler divergence between two Gaussian distributions has a closed form that can be directly expressed in terms of the means and the covariance matrices of the the two distributions.

2.2 Probabilistic Interpretation

Conceptualizing VAEs in terms of probability space would be useful in explaining the choice of Kulback-Leibler divergence as the regularization term. Let x denote our data, and that is in reality perfectly represented by a lower dimensional variable z, not directly observed. This means that: a latent distribution z is sampled from the prior distribution p(z), and that x is sampled from the conditional likelihood distribution p(x|z).
This way, the encoder is an analog of p(z|x), describing the distribution of the latent variable given the unencoded datapoint x, and the decoder is an analog of p(x|z), describing the distribution of the decoded variable given the encoded input.
Now assume p(z) to be of a standard Gaussian distribution, and hence, p(x|z) to be

$$z \sim N(0, \mathbf{I})$$

$$x|z \sim N(f(z), c\mathbf{I})$$

where f is unspecified for now.
The best Gaussian distribution, $q_x(z) \cong N(g(x), h(x))$, that describes the p(z|x) is the one which minimizes our given error measure, which would be Kulback-Leibler divergence-found via gradient descent over the parameters that describe the family of distribution. In other words, we need to find the optimal functions (g^*, h^*)=

$$\arg\min_{(g,h)\in GxH} KL(q_x(z), p(z|x))$$

$$= \arg\min_{(g,h)\in GxH} (\mathbb{E}_{z\sim q_x}(lnq_x(z)) - \mathbb{E}_{z\sim q_x}(ln\frac{p(x|z)p(z)}{p(x)}))$$

$$= \arg\max_{(g,h)\in GxH} (\mathbb{E}_{z\sim q_x}(lnp(x|z)) - KL(q_x(z), p(z)))$$

$$= \arg\max_{(g,h)\in GxH} (\mathbb{E}_{z\sim q_x}(-\frac{||x - f(z)||^2}{2c}) - KL(q_x(z), p(z)))$$

Given the **best** approximation of p(z|x), denoted by $q_x^*(z)$, we need to choose an f^* that maximizes the likelihood of x given z sampled from $q_x^*(z)$

$$f^* = \arg\max_{f \in F} \mathbb{E}_{z \sim q_x^*}(lnp(x|z))$$

$$= \arg\max_{f \in F} \mathbb{E}_{z \sim q_x^*}(-\frac{||x - f(z)||^2}{2c})$$

Hence, in a VAE, we are looking for the construction such that $(f^*, g^*, h^*) =$

$$\arg\max_{(f,g,h) \in FxGxH} (\mathbb{E}_{z \sim q_x}(-\frac{||x - f(z)||^2}{2c}) - KL(q_x(z), p(z))) \tag{1}$$

where the higher the value of c (set by user), the greater the weight of the regularization term over the reconstruction term and vice versa.

3 Results & Conclusion

Both models were trained on 10000 image arrays, for 20 epochs. Adjusted model was trained with Gaussian noise of variance 5 for best results.

3.1 Image Generation

3.1.1 VAE Model

Figure 1: 16 test dataset Images generated with VAE (20 epochs)

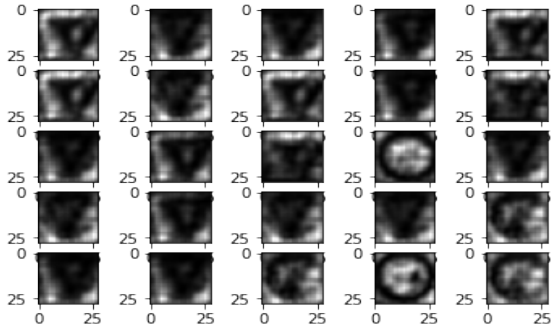

Figure 2: 25 Randomly generated images with VAE

3.1.2 Similar Model with modified ELBO

VAE model with latent layer controlled to be Gaussian with MMD.

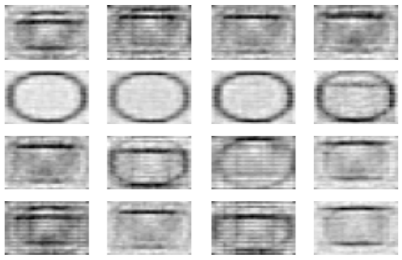

Figure 3: 16 test dataset Images generated with adjusted model (20 epochs)

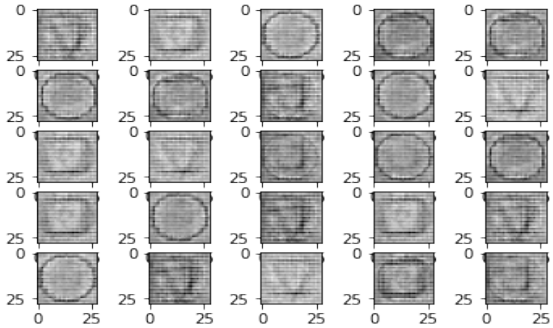

Figure 4: 25 Randomly generated images with adjusted model

3.2 Performance Comparison

Performance Metrics				
Model	Info loss (MSE Metric)	Test Image Generation Quality	Random Generation Quality	Runtime
VAE Model	12.022	93.75%	72.00%	243.09s
Adjusted VAE Model	0.048484	68.75%	80.00%	244.00s

Quality is determined by number of distinguishable shapes in generation set.

3.3 Conclusion

A large dataset of 10000 images was required to produce discernable shapes. Alternatives are training with more epochs, or much higher latent dimensional space (20 or above).

The adjusted model's generated images have visibly stronger, detailed edges and less smearing than the alternative model's. In contrast, the test image datasets are better generated by the standard VAE model than the modified one. This worsens even more for smaller variances of Gaussian noise.

The modified model is arguably better at generating images than the standard VAE model, in spite of performing relatively poorly with reconstructing test image datasets.

4 Code

Workspace on Colab `https://colab.research.google.com/drive/1iZ1BNwyySu08mg8Xjiaj0FLv2m5qmTdY?usp=sharing`

5 References

Diederik P. Kingma and Max Welling (2019), "An Introduction toVariational Autoencoders", Foundations and TrendsR©in Machine Learning: Vol. xx, No.xx, pp 1–18. DOI: 10.1561/XXXXXXXXX.

Keydana, S. (2021). RStudio AI Blog: Representation learning with MMD-VAE. RStudio AI Blog. Retrieved 8 September 2021, from `https://blogs.rstudio.com/ai/posts/2018-10-22-mmd-vae/`.

Lamberta, B., Daoust, M., Katariya, Y., Chen, W., Babu, S., Gardener, T. (2021). Convolutional Variational Autoencoder [Ebook]. Tensorflow. Retrieved 8 September 2021, from `https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/generative/cvae.ipynb`

Rocca, J. (2019). Understanding Variational Autoencoders (VAEs). Medium. Retrieved 8 September 2021, from `https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73` .

6 Appendix

6.1 Random Image Generation for standard VAE model trained to 10 epochs

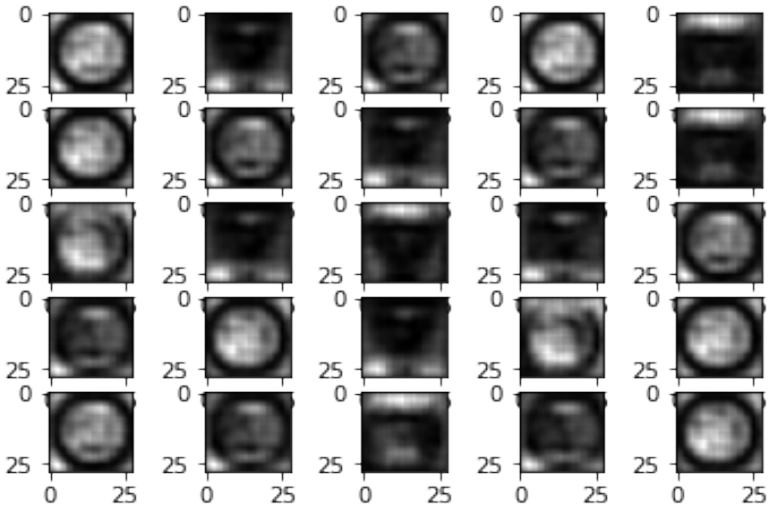

Figure 5: 25 Randomly generated images with VAE at 10 epochs only

6.2 Random Image Generation for modified model with Gaussian noise variance of 15

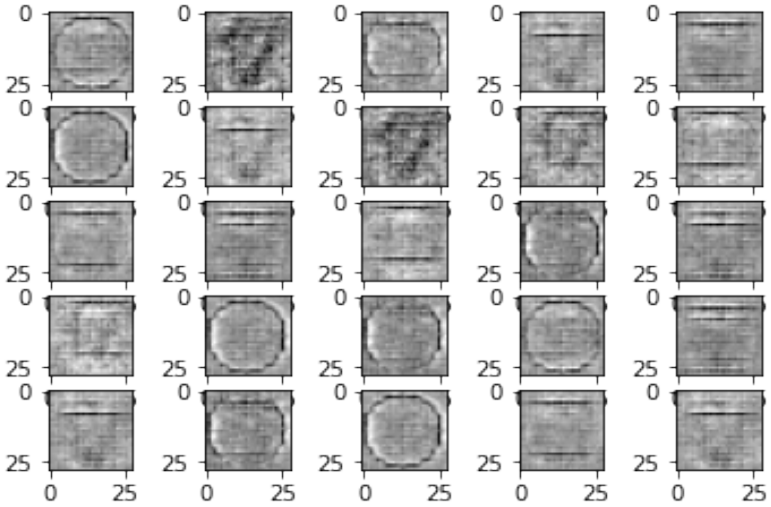

Figure 6: 25 Randomly generated images with VAE_ MMD

6.3 Test Image Generation for boths models

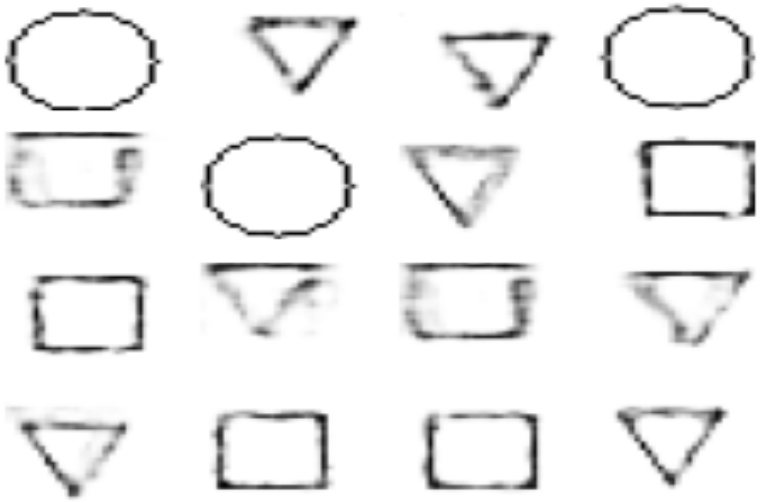

Figure 7: Test image generation at 50 epochs for standard VAE model

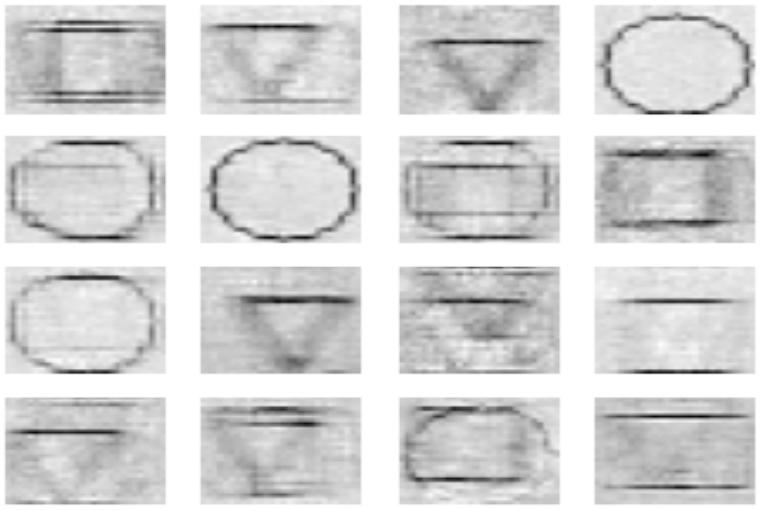

Figure 8: Test image generation at 50 epochs for modified model, with variance 2 for noise

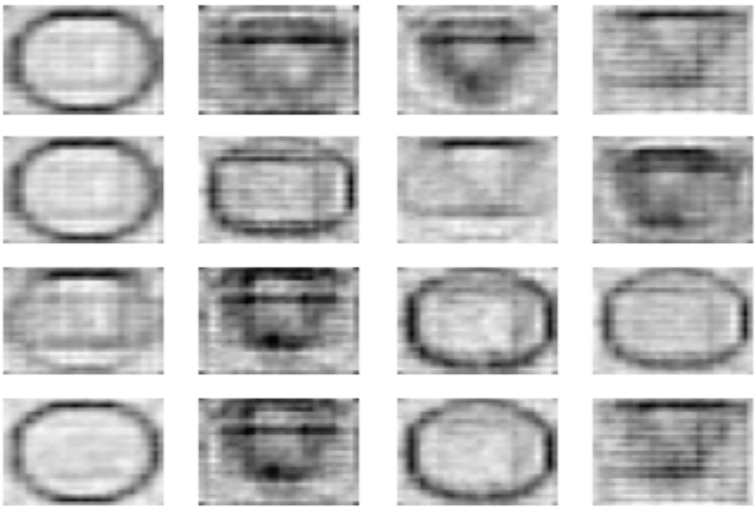

Figure 9: Different set of test image generation at 20 epochs for modified model, with variance 2 for noise