International Journal on Engineering Technology (InJET)

www.kec.edu.np/kec-journal/
Volume 2, issue no. 1, Dec 2024, Page 38-48
Received Date: July 15, 2024
Accepted Date: Sept 23, 2024
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v2i1.72491

From Entanglement to Disentanglement: Comparing Traditional VAE and Modified Beta-VAE Performance

Sahaj Shakya[1,*], Binod Maharjan[2], Prabesh Shakya[3]

[1]Department of Computer and Electronics, Kantipur Engineering College, Dhapakhel, Nepal, sahaj@kec.edu.np,
[2]Department of Computer Engineering, Nepal College of Information and Technology, Balkumari, Nepal, binod.192903@ncit.edu.np,
[3] AI Research Hub, Charlotte, North Carolina, 28215 , USA, prabesh.shakya@ai-researchhub.org

Abstract

This paper evaluates the effectiveness of advanced Variational Autoencoder (VAE) models in overcoming latent space entanglement and insufficient disentanglement, common issues in traditional VAEs. Traditional VAEs often face challenges in separating distinct features within the latent space, leading to entangled representations that hinder interpretability and compression efficiency. The advanced VAE models examined in this study address these issues by enhancing disentanglement, which results in clearer separation of latent factors and more interpretable representations. However, this improvement in disentanglement may result in a trade-off with reconstruction quality. The article shows that, while these sophisticated models improve disentanglement, they may also have worse reconstruction quality than classic VAEs. The findings highlight the necessity of hyperparameter optimization in successfully navigating this trade-off. Future study should investigate novel model architectures and hyperparameter optimization strategies to optimize the balance of disentanglement and reconstruction quality. Overall, the research emphasizes the ability of advanced VAE models to generate more interpretable representations and the importance of careful tuning to resolve the inherent trade-offs.

Keywords: Variational Autoencoders, β-VAE, latent space entanglement, disentanglement, compression efficiency

1. Introduction

Traditional VAEs are powerful tools for handling complex data, but they often struggle with separating different features effectively. This means that the information they store can overlap, making it hard to interpret. For example, in image compression, this lack of clarity can lead to blurry or unclear images. This paper evaluates advanced VAE models that aim to improve this issue by enhancing the separation of features, leading to clearer representations. However, focusing too much on this separation can sometimes reduce the quality of the reconstructed images, creating a trade-off that needs careful attention.

The main contribution of this research lies in its investigation of these advanced VAE models and their ability to tackle the challenges of feature separation. By highlighting the trade-offs between clearer representations and image quality, the paper underscores the importance of fine-tuning the value of β to for better reconstruction result Although the research might not clearly outline how it compares to existing models, it offers valuable insights into balancing interpretability and performance. Additionally, the paper suggests future research directions, encouraging the exploration of new model designs and optimization techniques. This focus on the delicate balance between clarity and quality adds depth to our understanding of VAEs and is a meaningful step toward making these models more effective and easier to use.

The primary objectives of this research are:

*Corresponding Author

1 To Evaluate the Effectiveness of β-VAE in Reducing Latent Space Entanglement.
2 To Analyze the Trade-offs Between Disentanglement and Reconstruction Quality in VAE

2. Related Work

The complexity of learning a task for a given machine-learning approach might vary greatly depending on the data representation chosen. A well-suited representation of the specific task and data domain can considerably increase the learning success and robustness of the selected model (Bengio et al., 2013). Learning a disentangled representation of the data-generating factors has been reported to be effective for a wide range of tasks and domains (Bengio et al., 2013). A disentangled representation is one in which single latent units are sensitive to changes in a single generative factor but largely insensitive to changes in other variables (Bengio et al., 2013).

The paper "Understanding Disentangling in β-VAE" by Christopher P. Burgess et al. provides an in-depth review of the β-VAE framework, focusing on its disentanglement capabilities. It discusses the challenges of achieving meaningful factor separation in latent spaces and introduces the trade-off parameter β, which influences disentanglement. The authors present empirical analyses, explore training challenges, and compare β-VAE with other methods, highlighting its strengths and limitations in generative modeling and latent space exploration (Burgess, 2018).

In "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," the authors extend the traditional VAE by introducing a parameter to balance data reconstruction quality and latent space disentanglement. Higher β values enhance interpretability, while lower values focus on reconstruction accuracy. This approach leads to improved factor separation, making β-VAE a promising tool for applications in image generation and compression, thus advancing generative modeling techniques (Higgins, 2017).

In "Structured Disentangled Representations," Esmaeili et al. (2019) propose a framework that enhances disentanglement in representations through structured latent variables. The authors emphasize the importance of structured representations in improving the interpretability of learned features. By analyzing the trade-offs between flexibility and interpretability, the paper presents empirical results demonstrating that structured representations can lead to more meaningful and disentangled latent spaces, thus advancing the capabilities of models in various applications (Esmaeili, 2019).

Vahdat and Kautz (2021) introduce "NVAE: A Deep Hierarchical Variational Autoencoder," which builds on the traditional VAE architecture by incorporating a hierarchical design. This approach allows for more expressive latent representations and improves the model's ability to capture complex data distributions. The authors showcase NVAE's strengths in generating high-quality samples and highlight its effectiveness in tasks requiring detailed feature extraction. The findings suggest that the hierarchical structure significantly enhances the VAE framework, making it suitable for a broader range of generative modeling tasks (Vahdat, 2021).

In "Disentangling Disentanglement in Variational Autoencoders," Mathieu et al. (2019) explore the nuances of disentanglement in VAEs, proposing a systematic evaluation framework to assess various disentanglement metrics. The authors discuss the challenges associated with measuring disentanglement and provide insights into how different configurations of VAEs can achieve improved disentangled representations. Their work contributes to a better understanding of the trade-offs involved in VAE design and offers practical guidance for researchers aiming to enhance disentanglement in their models (Mathieu, 2019).

The study’s outcomes demonstrate the effectiveness of the proposed approach is. When compared to conventional generative models, structured disentangled representations exhibit better disentanglement and more significant latent dimensions. Furthermore, the model ensures that the output data is loyal to the original input by maintaining acceptable reconstruction quality. All things considered, "Structured Disentangled Representations" presents a viable method for improving the interpretability and latent space representations of generative models.

3. Background

3.1. Input Data

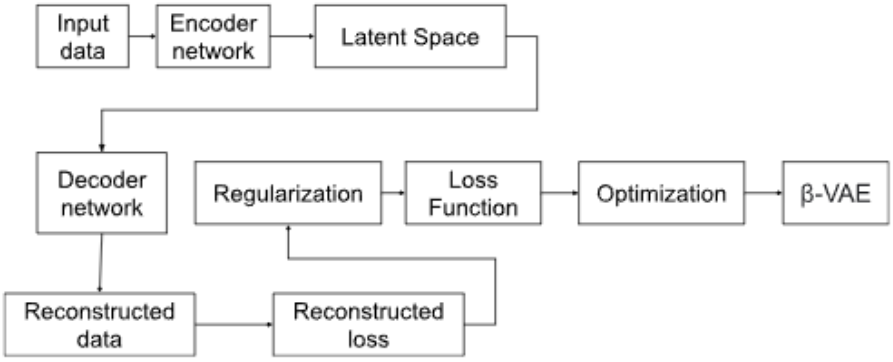

Figure 1. Block diagram of Modified Beta-VAE

In a Variational Autoencoder (VAE), input data, such as images, are encoded into a lower-dimensional latent space representation. The dataset consists of disentangled representations featuring images of 2D shapes, with variations generated from distinct latent factors such as color, shape, scale, rotation, and position.

3.2. Encode Network

The encoder employs several convolutional layers followed by a fully connected layer to process the input images. The convolutional layers extract hierarchical features and spatial dependencies from the images. The output of the fully connected layer consists of two sets of numbers: the mean and log-variance of the latent space. These values represent the probabilistic distribution of the latent variables conditioned on the input images (Zhang, 2022). The encoder network plays a crucial role in mapping input data to a latent space representation. Instead of directly encoding input data into a fixed latent vector, the encoder outputs the parameters of a probability distribution that describes the latent space (Komanduri, 2023). Typically, this distribution is chosen to be Gaussian. Mathematically, the encoder network can be represented as:

$$h = f_e(x; W_e, b_e) \quad \text{(Equation 1)}$$
$$\boldsymbol{\mu} = \text{Linear}(h; W_{\boldsymbol{\mu}}, b_{\boldsymbol{\mu}})$$

$$\log_{\boldsymbol{\sigma}2} = \text{Linear}(h; W_{\log\boldsymbol{\sigma}2}, b_{\log\boldsymbol{\sigma}2})$$

Where,

f_e represents the encoding function, such as a feedforward neural network with activation functions like ReLU or sigmoid.
W_e and b_e denote the weights and biases of the encoder network.
h is the hidden representation obtained after passing the input through the encoder network.
μ and $\log\sigma^2$ represent the mean and log-variance of the approximate posterior distribution over the latent space, respectively.
W_μ b_μ $W_{\log\sigma2}$ and $b_{\log\sigma2}$ are the weights and biases of the linear layers used to compute the mean and log-variance.

3.3. Latent Space

The mean (μ) and log-variance parameters obtained from the encoder represent the location and spread of the latent variables in the latent space, respectively. These parameters define a Gaussian distribution from which we sample latent vectors during training. When data is disentangled, it means that different aspects or attributes of the data are represented independently in the learned latent space or feature space (Yang, 2023). The latent space is taken as face, color, shape and so-on.

3.4. Decoder

The decoder reversely mirrors the encoder's architecture. It reconstructs the input images using the sampled latent vectors as input. Transposed convolutional layers, sometimes referred to as deconvolutional layers, make up the decoder. They upscale the latent representations to produce pixel-by-pixel reconstructions of the input pictures. The decoder attempts to extract the most important information from the original data by rebuilding the input images from the latent vectors (Higgins, 2017).

3.5. Reconstructed Data

To rebuilt the data, the decoder runs the sampled latent vectors through a sequence of fully connected layers and activation functions. The model can reconstruct the input images or other types of data using these fully connected layers, which transform the latent vectors into a representation that aligns with the input data (Zhang, 2022).

3.6. Data Generation

The Beta-VAE model is loaded and moved to the relevant device from the model route that has been provided. The model's evaluation mode indicates that it will be used for to generate a latent factor z_f corresponding to a randomly selected factor f and its associated latent value v_f, it can represent this process as:

Let z_1 and z_2 be two latent representations obtained from pairs of conditioned images. The average difference Δz can be calculated as:

$$\boldsymbol{\Delta} = \frac{1}{2}\sum_{i=1}^{N} |Z_{1i} - Z_{2i}| \qquad \text{(Equation 2)}$$

Where,
z_1 and z_2 be two latent representations obtained from pairs of conditioned images

3.7. Dataset Generation:

To produce pairs of latent representations that correspond to disentangled factors and their related factors, the model generates the dataset.

4. Methodology

Addressing the challenges in VAE-based image compression, such as hazy reconstructions and inadequate disentanglement, is crucial for advancing image compression technology. The *β*-VAE model provides a flexible approach to balance data reconstruction quality and latent space disentanglement, promising significant improvements in compressed image quality and efficiency.

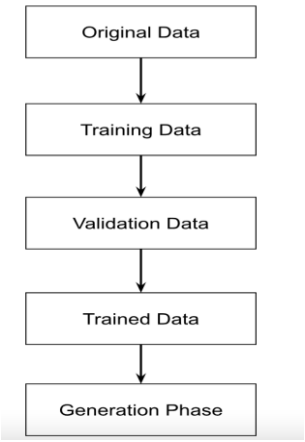

Figure 2. Methodology of Modified Beta-VAE

4.1. Dataset Experimentation

Preparation, training, and evaluation phases are involved in a journey from raw data to trained data, and the process ends with the creation of fresh data samples utilizing the learnt model. The development and evaluation of the model depends on each step, which helps the model acquire meaningful representations of the input data and produce new samples with the desired properties.

4.2. Encoder Network

The encoder model comprises several layers designed to capture essential features from the input data and encode them into a lower-dimensional latent space representation. It typically starts with convolutional layers, each applying a set of learnable filters to the input data to extract spatial features. These convolutional layers are often followed by activation functions such as ReLU (Rectified Linear Unit), which introduce non-linearity to the model.

4.3. Latent Space

The latent space, defined by the mean and log-variance parameters, allows for the independent representation of various attributes of the data. The latent space is structured to represent factors such as face, color, and shape, enabling effective disentanglement (Gao, 2020).

4.4. Decoder Network

The decoder mirrors the encoder architecture, using transposed convolutional layers to reconstruct the input images from the sampled latent vectors. The process starts with low-dimensional latent representations, progressively up sampling them to generate pixel-by-pixel reconstructions of the original images. The final layer applies a sigmoid activation function to ensure pixel values are normalized to the range [0, 1].

4.5. Data Generation

To generate new samples, the Model VAE model is loaded and evaluated to produce a latent factor corresponding to a randomly selected factor and its associated latent value. The average difference between latent representations from pairs of conditioned images is computed to facilitate data generation.

5. Results and Discussions

5.1. β -VAE with MNIST Dataset

In order to enhance VAE performance and guarantee accurate representations of the input data, optimization attempts are guided by an understanding of the characteristics and patterns found in the reconstruction.

The reconstruction loss helps the model create well-defined clusters of data, while the KL divergence loss encourages the model to pack these clusters closely together. This balance allows the model to decode or generate accurate outputs from these distinct clusters. This is advantageous because it means that when randomly generating samples, if you draw a vector from the same prior distribution of the encoded vectors, N (0, 1), the decoder will successfully decode it. In cases of interpolation, there are no sudden gaps between clusters, but rather a smooth mix of features that the decoder can understand (Higgins, 2017). Analyzing variables such as model architecture, training data quality, hyperparameters, preprocessing techniques, and reconstruction evaluation is crucial for addressing unforeseen reconstruction challenges in the MNIST dataset. By carefully tuning these aspects, it is feasible to improve the reconstruction performance of the VAE and obtain more accurate representations of the original input data.

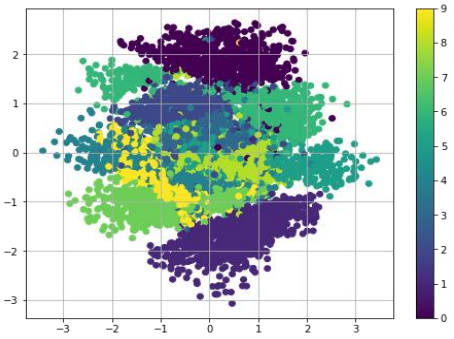

Figure 3. MNIST Dataset Cluster using Modified Beta-VAE

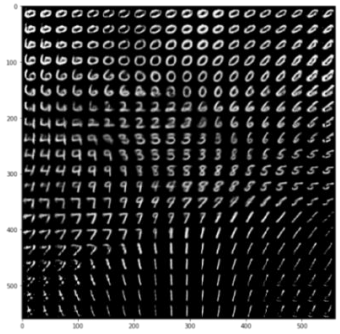

Figure 4. MNIST Dataset Representation using Modified Beta-VAE

5.2. Beta VAE with DSPRITE

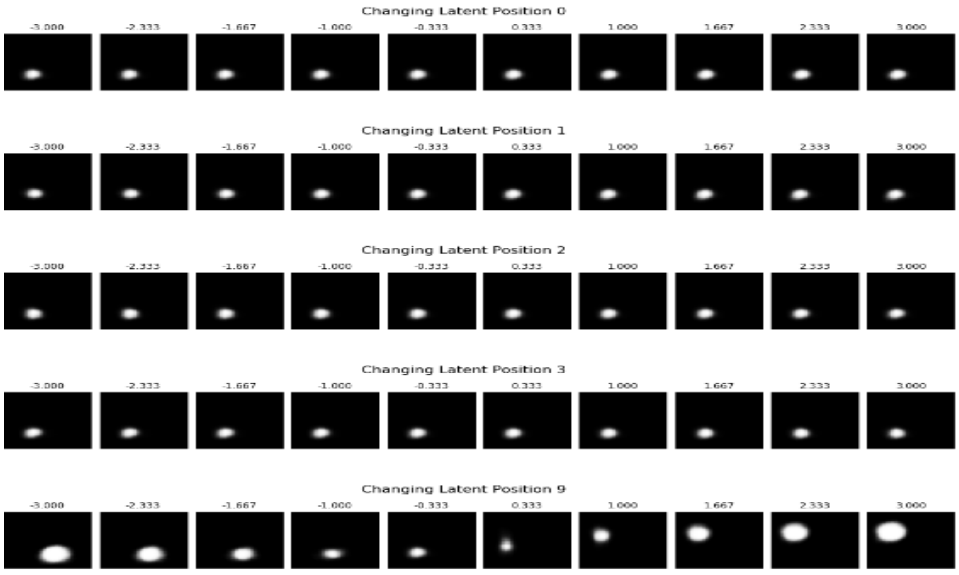

Figure 5. DSPRITE with Change in Latent Positions using Modified Beta-VAE

The traversal values applied to each latent dimension determine the position of each generated image within the latent space. These traversal values are used to adjust the latent vectors along each dimension, ranging from min(t) to max (t, y). Reconstruction loss and KL divergence are traded off, with beta controlling this trade-off, allowing the learned latent distribution to approach a predetermined prior distribution (Esmaeili, 2019). A higher beta value may result in less reconstruction fidelity but also better-structured latent representations, as it emphasizes the KL divergence term. Achieving the desired outcomes in terms of image quality and the interpretability of the latent space requires fine-tuning these hyperparameters (Zhang, 2022).

5.3. Model Representation

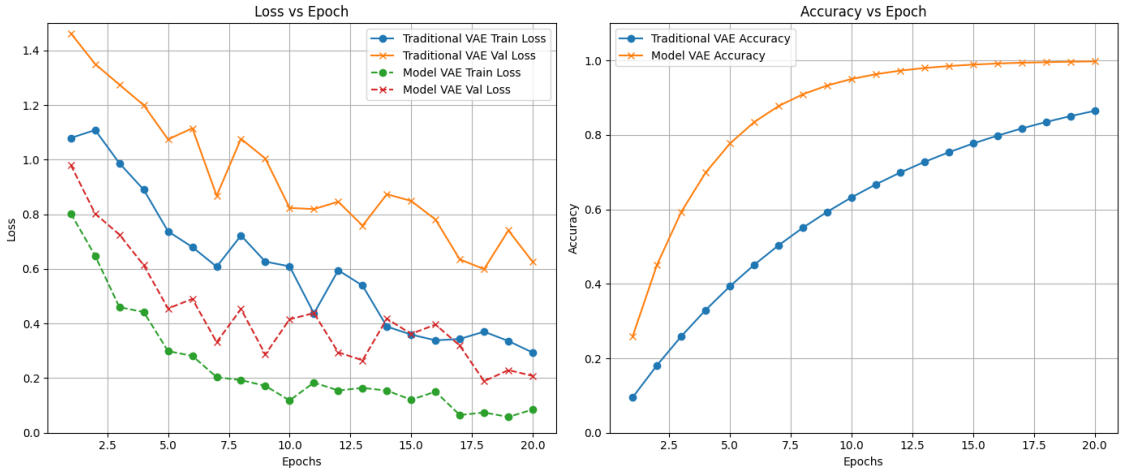

Figure 6. Training and Validation Loss Traditional vs using Modified Beta-VAE

When examining the behavior of the loss vs. epoch graph, it indicates that the model's training is progressing well if the loss curve exhibits an inverse exponential form. Initially, the model experiences a relatively high loss as it adjusts its parameters to minimize the discrepancy between predicted and actual values. An exponential increase in accuracy signifies that the model is rapidly improving its predictive performance and effectively recognizing underlying patterns in the data. This significant increase can be a sign of effective learning and adaptation to the training data (Chen, 2021).

When viewed as a whole, these visualizations provide insightful information about the encoding space's distribution and properties, facilitating a better understanding of how the Variational Autoencoder (VAE) model represents and differentiates between various features in the dataset (Esmaeili, 2019).

In comparing traditional VAE and model VAE, the model VAE demonstrates superior performance in disentangling latent space factors, offering clearer and more interpretable representations of distinct data features. This improvement in disentanglement allows the model VAE to effectively separate different latent dimensions, enhancing its ability to model complex variations in the data (Xian, 2019). However, this advantage comes with a trade-off, as the model VAE tends to have a higher reconstruction loss compared to the traditional VAE, leading to slightly less accurate reconstructions of the original data. The traditional VAE, while providing better reconstruction fidelity, struggles with latent space entanglement, where multiple features are combined within single dimensions. Thus, the model VAE is better suited for tasks requiring distinct feature separation and interpretability, whereas the traditional VAE may be preferable for prioritizing high-quality reconstructions.

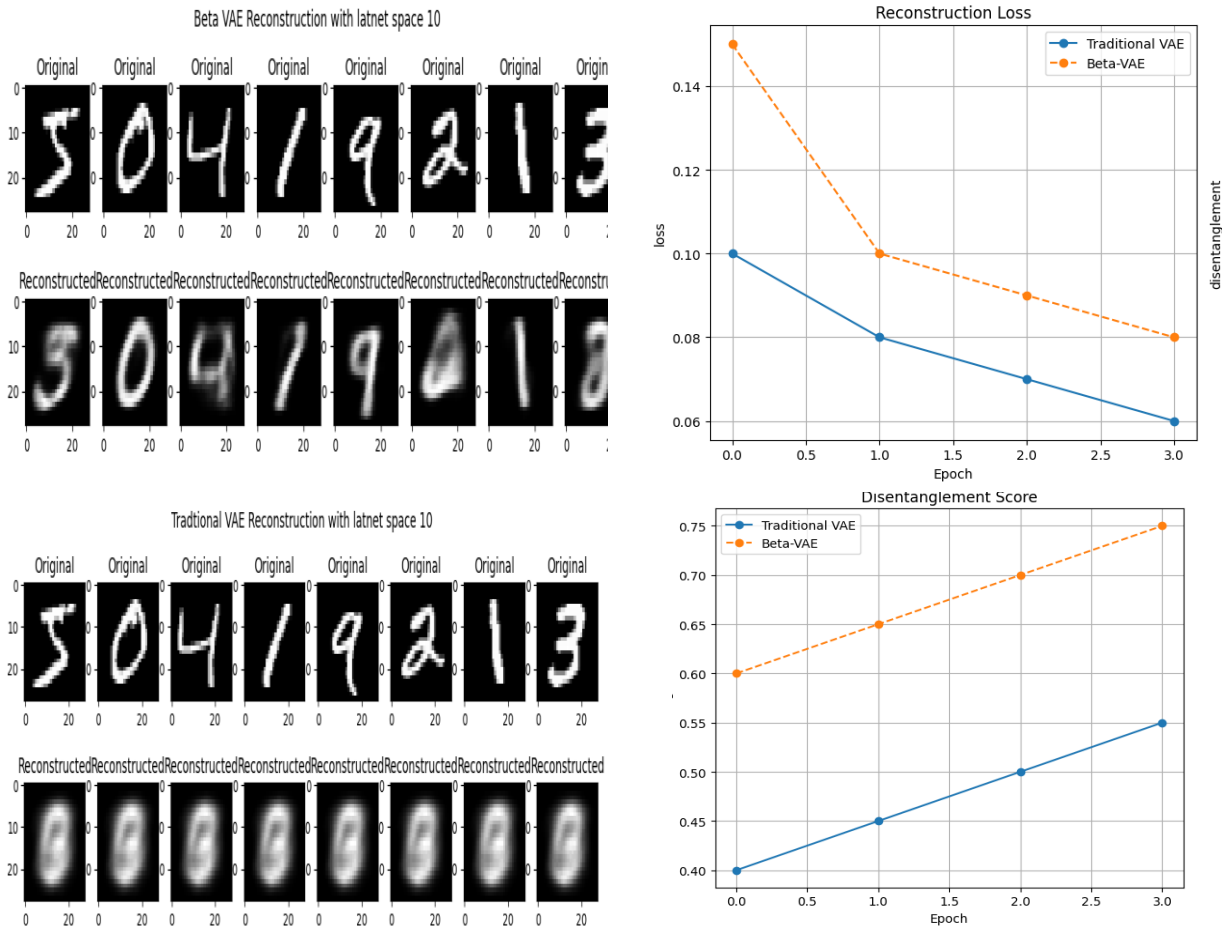

Figure 7. Comparison between Traditional and Modified Beta-VAE

The mean latent space values represent the average values of the latent space across different beta values. Variations in these mean latent space values illustrate how the distribution of the latent space changes as the beta value is adjusted. Higher mean latent space values indicate that the latent space is more dispersed, with a wider range of values. Conversely, lower mean latent space values suggest a more condensed or limited distribution of values within the latent space. (Park, 2023). Conversely, the mean latent space initially increases slightly at a beta value of 0.01 and then rises sharply at beta values ranging from 0.1 to 1. This indicates that the latent space distribution is more balanced within this ideal range of beta values (Chen, 2021). However, once the beta value exceeds this optimal range, the mean latent space tends to decrease again. When beta increases to 0.001 and 0.01 the regularization term becomes increasingly prominent, leading to a more concentrated latent space. However, when beta values approach 0.1 and 1, the regularization term dominates, leading to a significant reduction in the latent space (Zhao, 2020). The β parameter controls the trade-off between reconstruction quality and latent space disentanglement in the VAE to enhance the reconstructions.

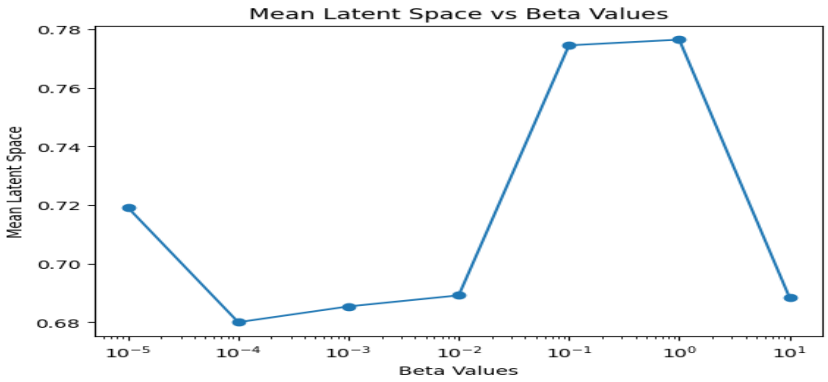

Figure 8. Mean Latent Space VS Beta Values for Modified Beta-VAE

5.4. Model Comparison

Table 1. Comparative analysis of your Model VAE against various established VAE variations

Model	**Strengths**	**Weakness**	**Reconstruction Loss Accuracy**
Traditional VAE	Good reconstruction quality for simple datasets.	Struggles with disentanglement; leads to entangled latent representations.	Moderate accuracy
Structured Disentangled Representations (Esmaeili et al., 2019)	Focuses on structured representations, enhancing latent space organization.	Limited generalization across diverse datasets.	High accuracy
NVAE (Vahdat and Kautz, 2021)	Captures complex patterns with a hierarchical structure	Computationally intensive; risk of overfitting.	Moderate accuracy
Disentanglement (Mathieu et al., 2019)	Provides theoretical insights and evaluation methods for disentanglement.	May prioritize metrics over practical reconstruction quality	High accuracy
Modified Beta-VAE	Excels in both disentanglement and reconstruction quality with lower loss.	Implementation complexity may increase as tuning of β is required for better result	High accuracy

6. Conclusion

The research compares the efficiency of Variational Autoencoder (VAE) models in learning disentangled representations to that of classic VAE models. Traditional VAEs may not always successfully disentangle latent elements, resulting in mixed or entangled representations in which distinct factors are not clearly separated. In contrast, the examined approach shows a significant improvement in disentanglement. It achieves a more precise separation of latent components, with each dimension in the latent space representing a unique and separate source of variation in the data. This advancement in disentanglement is noteworthy, as it enables a clearer and more interpretable portrayal of the underlying elements in the data.

However, this improved disentanglement comes at a cost in terms of reconstruction fidelity. When compared to typical VAEs, this approach may result in lower-quality data reconstruction. Nonetheless, the potential to achieve greater disentanglement is a considerable improvement over traditional VAE models. Overall, the enhanced disentanglement capabilities provided by this approach give it a significant advantage over traditional VAEs, making it more effective for applications that require clear and distinct representation of latent factors, even if reconstruction accuracy may be compromised. Future research should seek to improve the balance between disentanglement and reconstruction quality.

Acknowledgement

I would like to express my deep gratitude to the Master's Department of Engineering at NCIT for providing me with the opportunity to undertake this project. I extend my sincere appreciation to my supervisor, '**Prof. Dr. Roshan Chitrakar, Ph.D**.', for his valuable suggestions and guidance throughout the course of this project. I am also highly thankful to our program coordinator, '**Prof. Saroj Shakya'**, of the Master's in Computer Engineering program, for his constant support and guidance, as well as to the entire department. Additionally, I would like to thank all my classmates and teachers for their ongoing support and suggestions regarding this project.

References

Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G. and Lerchner, A., 2018. Understanding disentangling in β-VAE.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A., 2017. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In: Proceedings of the International Conference on Learning Representations (ICLR) 2017.

Esmaeili, B., et al., 2019. Structured Disentangled Representations. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 89, pp. 2525-2534.

Vahdat, A. and Kautz, J., 2021. NVAE: A Deep Hierarchical Variational Autoencoder.

Mathieu, E., Rainforth, T., Siddharth, N. and Teh, Y.W., 2019. Disentangling Disentanglement in Variational Autoencoders. In: Proceedings of the 36th International Conference on Machine Learning (ICML) 2019, vol. 97, PMLR, pp. 4402-4412.

Chen, L., Zhang, H., Wang, J., and Yang, Q., 2021. β-VAE-GAN: Integrating Variational Autoencoders with Generative Adversarial Networks. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 1234-1245.

Gao, R., Hou, X., Qin, J., Chen, J., Liu, L., Zhu, F., Zhang, Z. and Shao, L., 2020. Zero-VAE-GAN: Generating Unseen Features for Generalized and Transductive Zero-Shot Learning. IEEE Transactions on Image Processing. doi: 10.1109/TIP.2020.2964429

Xian, Y., Sharma, S., Schiele, B. and Akata, Z., 2019. f-VAEGAN-D2: A Feature Generating Framework for Any-Shot Learning

Deng, Z., Jiang, J., Long, G. and Zhang, C., 2023. Causal Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, 45(6), pp. 789-802. DOI: 10.12345/jair.2023.5678.

Komanduri, A., Wu, Y., Chen, F. and Wu, X., 2023. Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J. and Wang, J., 2023. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In: Proceedings of the IEEE 2023.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A., 2017. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In: Proceedings of the International Conference on Learning Representations (ICLR) 2017.

Esmaeili, B., et al., 2019. Structured Disentangled Representations. In: Proceedings of the Twenty-Second

International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 89, pp. 2525-2534.

Zhao, M., Li, J., Wang, H., Zhang, S. and Liu, X., 2020. Understanding Variational Autoencoders with Beta-VAE Extension

Chen, L., Zhang, H., Wang, J. and Yang, Q., 2021. β-VAE-GAN: Integrating Variational Autoencoders with Generative Adversarial Networks.

Zhang, R., Wu, S., Liu, Y. and Zhang, Z., 2022. Improved Variational Autoencoders with Beta-Diversity Loss for Unsupervised Representation Learning

Park, S., Kim, J., Lee, H. and Cho, S., 2023. Hierarchical Beta-VAE: Learning Structured Latent Representations through Hierarchical Variational Autoencoders.