Leveraging Conventional and Variational Autoencoder for Reconstruction & Generation

Kalpan Mukherjee[a], Ayush Naique[b]

[a]*km6276@nyu.edu,*
[b]*asn9772@nyu.edu,*

Abstract

This report provides an exhaustive examination of the application of Variational Autoencoders (VAEs) and conventional Autoencoders in the domain of image reconstruction. VAEs and autoencoders, representing unsupervised learning paradigms, adeptly capture meaningful data representations through the acquisition of efficient encoding and decoding schemes. The Variational Autoencoder introduces probabilistic latent variables, facilitating the generation of diverse reconstructions in response to a given input. This distinctive feature renders VAEs especially pertinent for applications requiring inherent uncertainty or variability in the reconstruction process. Conversely, Traditional Autoencoders hinge on deterministic latent representations, presenting a computationally efficient alternative. Our investigation scrutinizes the architectural nuances and training methodologies of autoencoders, elucidating their role in engendering faithful reconstructions. We assess the influence of varying latent space dimensions on reconstruction quality and navigate the delicate balance between computational efficiency and reconstruction fidelity. Comparative analyses discern the distinctive strengths and limitations of VAEs and autoencoders in the realm of image reconstruction tasks. The report delves into practical considerations encompassing hyperparameter tuning and network architectures that exert a substantial impact on the performance of these models.
Code Repository

1. Introduction

Image reconstruction stands as a foundational process within the realms of computer vision and image processing, encompassing the restoration or generation of high-fidelity images from data that may be incomplete, noisy, or encoded. This undertaking is pivotal across diverse applications, spanning medical imaging, surveillance, and the enhancement of photographic content. The recent integration of advanced deep learning methodologies, exemplified by Variational Autoencoders (VAEs) and conventional Autoencoders, has markedly propelled the efficacy of image reconstruction. VAEs incorporate probabilistic elements into their framework, facilitating the generation of diverse reconstructions, whereas Autoencoders optimize computational efficiency through deterministic latent representations.
In the context of this project, we embark on the training of Variational Autoencoders and traditional Autoencoders using an extensive array of training images. Subsequently, a meticulous comparative analysis is conducted to elucidate and contrast their respective capacities for image reconstruction.

2. Case Studies

In the domain of image reconstruction, various models and techniques have been explored, each offering unique strengths and addressing different aspects of this challenging task. The advancements in Variational Autoencoders (VAEs) and traditional Autoencoders, as highlighted in our study, draw from several key developments in the field.

The concept of Autoencoders, fundamental to our research, revolves around unsupervised learning paradigms that capture efficient data representations through encoding and decoding processes. Traditional Autoencoders are known for their deterministic latent representations, providing a computationally efficient approach for image reconstruction. Their simplicity and effectiveness in learning representations have made them a staple in various applications, particularly in scenarios where computational resources are a constraint.

Variational Autoencoders, on the other hand, introduce a probabilistic twist to the autoencoding process. The introduction of probabilistic latent variables by Kingma and Welling (1) marked a significant advancement in the field. VAEs generate diverse reconstructions in response to a given input, handling inherent uncertainty and variability in the reconstruction process. This feature is particularly beneficial in applications requiring a degree of randomness or creativity in the output.

The development of VAEs and traditional Autoencoders has been influenced by foundational work in related fields. For instance, the Transformer model, introduced by Vaswani et al.(2), and its adaptation in Vision Transformers (ViTs) by Dosovitskiy et al.(3) have revolutionized the approach to handling sequences, whether in text or image form. These models' ability to capture long-range dependencies has informed approaches in image reconstruction, particularly in understanding and manipulating complex image structures.

Similarly, the concept of self-supervised learning, which has seen extensive use in models like Masked AutoEncoders (4), plays a role in our research. This approach, where models learn from auxiliary tasks like predicting masked parts of an image,

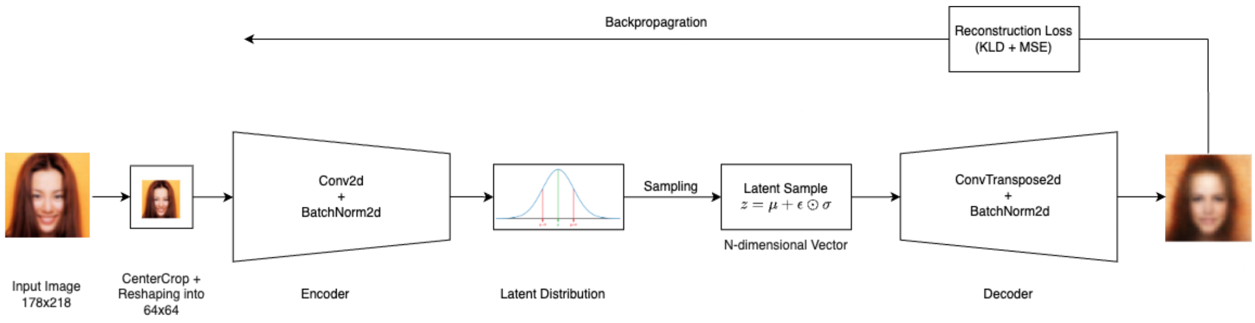
Backpropagration
Reconstruction Loss (KLD + MSE)
Conv2d + BatchNorm2d
Sampling
Latent Sample $z = \mu + \epsilon \odot \sigma$
N-dimensional Vector
ConvTranspose2d + BatchNorm2d
Input Image 178x218
CenterCrop + Reshaping into 64x64
Encoder
Latent Distribution
Decoder

Figure 1: Steps followed

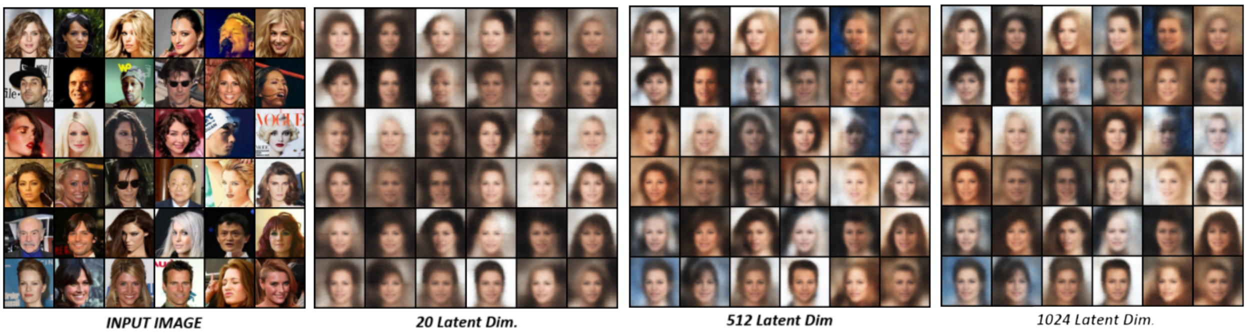

Figure 2: Image Reconstruction: Variational Autoencoders

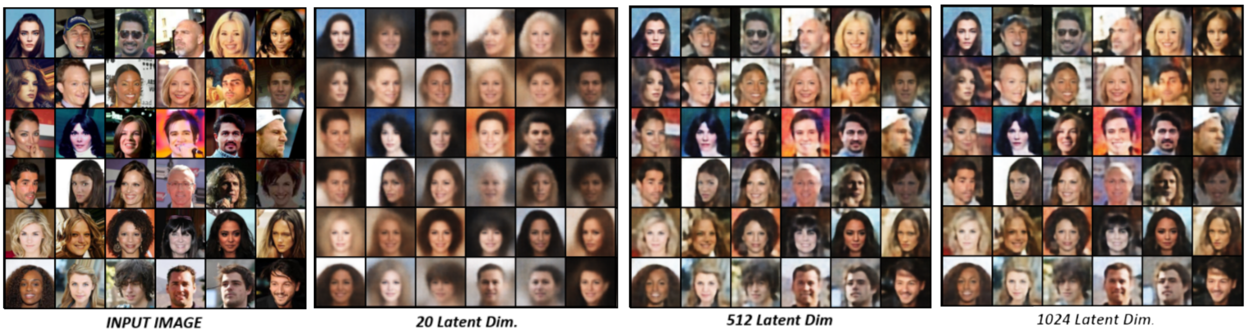

Figure 3: Image Reconstruction: Autoencoders

has proven effective in learning robust and generalizable representations. Such techniques are crucial in improving the quality of image reconstruction by enabling models to understand and predict image content more accurately.

In our work, we examine the architectural nuances and training methodologies of both VAEs and traditional Autoencoders, considering their role in image reconstruction tasks. We explore the influence of varying latent space dimensions on reconstruction quality, a key factor in both VAE and Autoencoder performance. Additionally, we assess how hyperparameter tuning and network architecture choices impact these models, drawing on the rich history of advancements in the broader field of image processing and machine learning. Our comparative analyses aim to discern the distinctive strengths and limitations of VAEs and Autoencoders, providing insights into their optimal applications and potential developments in the field of image reconstruction.

3. Dataset

In this undertaking, the Large-scale CelebFaces (5) Attributes (CelebA) dataset, as presented by Ziwei Liu et al., served as the foundational corpus for model training and validation. Comprising in excess of 200,000 images capturing 10,077 distinct celebrity identities, this dataset was chosen due to its extensive image repository and inherent variability. The copious image data, exhibiting a spectrum of facial variations, renders the CelebA dataset well-suited for the comprehensive training of the Autoencoder models under consideration.

4. Methodology

Figure 1 delineates the procedural steps undertaken in this research project for image reconstruction. Initial preprocessing of images sourced from the CelebA dataset involves resizing and reformatting into 64x64 dimensions, thereby facilitating a

Figure 4: CelebA dataset Images

streamlined model training process.
Subsequently, the reshaped images undergo processing through the Encoder segment of the Autoencoder architecture, comprising four Convolutional layers. The Encoder operates to encode the input images, originally in the format of 64x64x3, into a latent vector of n dimensions. This dimensionality, denoted by 'n,' is predetermined during runtime. The resultant latent vector serves as the essential input for the Decoder system, tasked with reconstructing the images by transforming the n-dimensional vector back into an RGB image format.

4.1. Components

In the employed methodology, a modified Autoencoder and Variational Autoencoder architecture is utilized, as previously outlined. The design principle maintains symmetry between the Encoder and Decoder Objects, fostering a one-to-one reconstruction of data. This symmetrical configuration is instrumental in preserving the integrity of the encoded information during the reconstruction process. Subsequent to the convolutional operations at each layer within the network, the resulting outputs traverse through Rectified Linear Unit (ReLU) layers. The incorporation of ReLU layers introduces non-linearity to the network, promoting enhanced model flexibility and the ability to capture intricate data patterns. Furthermore, the computational efficiency of ReLU, stemming from its straightforward mathematical formulation, contributes to expedited training durations, rendering it a judicious choice within the presented architectural framework.

4.2. Latent Representation

The latent space in VAEs is where the model encodes the data into a compressed, abstract, lower dimensional form. This allows the model to understand the nuances of the input data instead of attaching importance to secondary features (for example, understanding that the foreground is more important than the background in facial image reconstruction). Where variational autoencoders diverge from traditional autoencoders is by introducing a probabilistic layer to the encoding process. Instead of encoding an input into a fixed latent representation, VAEs map inputs into a distribution over the latent space. This allows the model to accurately imbibe the features of the input data and create brand new examples instead of just mimicking it (as in the case of traditional autoencoders). The latent representation is characterised by two key parameters: the mean (μ) and variance (σ^2) which define a Gaussian distribution for each latent dimension.

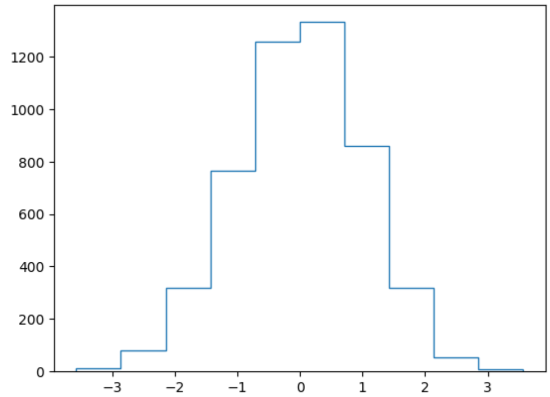

Figure 5: First dimension of the Latent vector representing a Gaussian Curve

This probabilistic encoding, however, presents a challenge for training the model using back propagation, as it requires the gradient of a sampled value with respect to the parameters of the distribution. Directly sampling from the distribution makes the process non-differentiable. This is where something called a **reparameterisation trick** is used (1). This 'trick' involves sampling from a standard distribution (like a standard normal distribution) and then transforming this sample using the parameters (mean and variance) of the desired distribution. Specifically, in a VAE, a sample z from the latent space is generated by :

$$z = \mu + \sigma \odot \epsilon \tag{1}$$

where ϵ is a random sample drawn from a standard normal distribution. This approach makes the sampling process differentiable as μ and σ are deterministic functions of the input and the model parameters, and ϵ is independent of these parameters (Figure 5).

5. Experiments

5.1. Training

Both the conventional and Variational Autoencoders underwent training on a dataset comprising 200,000 images extracted from the CelebA dataset. The training process encompassed the development of three distinct model variations for each architecture, wherein the dimensionality of the latent vector ranged from 20 to 1024 dimensions. Post-training, the models were employed for image reconstruction tasks using the validation

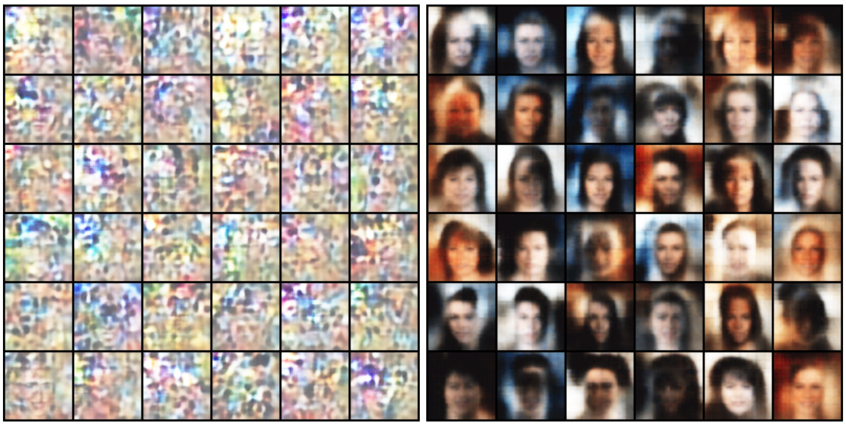

Figure 6: Image Generation: Autoencoder (left), Variational Autoencoder (right)

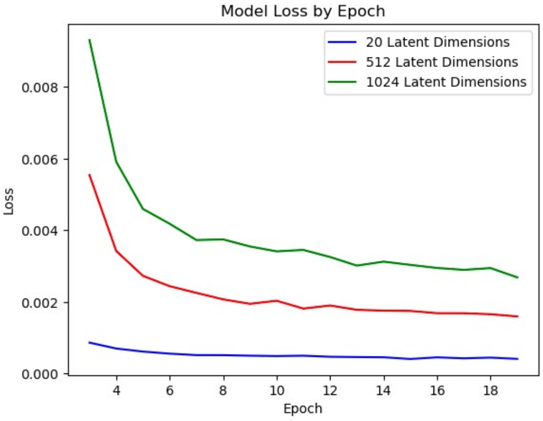

Figure 7: Training loss for VAE over 20 epochs

set. Additionally, the models' image generation capabilities were evaluated by subjecting them to random inputs, serving as a comprehensive test of their capacity to generate meaningful visual content.

The Variational Autoencoder (VAE) computes its total loss using two integral components: the reconstruction loss and the regularization term, expressed as the Kullback-Leibler (KL) divergence. The reconstruction loss serves to gauge the VAE's proficiency in reconstructing input data and is rooted in the likelihood of the input data given its reconstructed counterpart. By minimizing the reconstruction loss, the VAE is incentivized to generate outputs closely mirroring the input data. Concurrently, the regularization term, computed through KL divergence, imposes constraints upon the latent space, directing it to conform to a specified prior distribution, commonly a standard normal distribution. This regularization mechanism mitigates overfitting tendencies towards the training data, facilitating the disentanglement of learned latent representations within the VAE framework.

$$L_{recon} = -E_{q(z|x)}[logp(x|z)] \tag{2}$$

$$L_{KL} = 0.5 * (\sum (1 + log(\sigma_j{}^2) - \mu_j{}^2 - \sigma_j{}^2) \tag{3}$$

5.2. Results

In Figure 3, we demonstrate that traditional autoencoders exhibit superior performance in image reconstruction tasks compared to their variational counterparts. This enhanced capability is attributed to the fact that standard autoencoders primarily focus on replicating the input images. Conversely, variational autoencoders (VAEs) are designed to learn the underlying data distribution, enabling them to generate images from scratch—a task where traditional autoencoders typically face challenges.

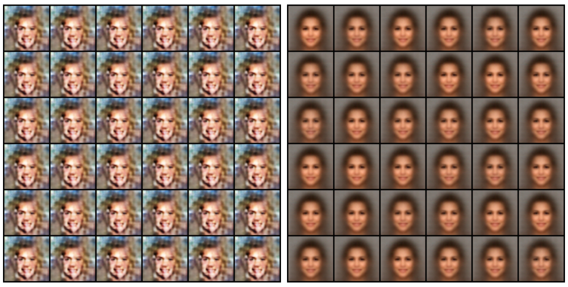

Figure 8: Representation learnt for autoencoder and VAE

Further exploration of this concept is presented in Figure 6. Here, we input a randomly generated latent variable exclusively into the decoder segment of a variational autoencoder. The model successfully generates a facial image of an individual who was not part of either the training or test dataset. This phenomenon underscores the model's stochastic nature, which empowers it to create novel examples that are similar yet distinct from the training data. Such capability aligns with the foundational concept of websites like 'thispersondoesnotexist.com' and others, showcasing the potential of VAEs in generating diverse, unseen data instances. Fig 8 shows the latent representation learned by the trained models.

6. Future Work

The efficacy of the reconstructed images can be improved by training on larger images (eg. 128x128, 256x256 etc.). This will gives the model scope to understand and imbibe minute details like eye color, skin texture, hair etc. Another direction that could be taken would be to use masked autoencoders for reconstruction allowing the model to attached weighted importance to different parts of the image.

References

[1] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In International Conference on Learning Representations (ICLR).

[2] Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS).

[3] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[4] He, K., et al. (2021). Masked Autoencoders are Scalable Vision Learners: In Computer Vision and Pattern Recognition (CVPR).

[5] Liu, Ziwei et al. (2015). Deep Learning Face Attributes in the Wild: Proceedings of International Conference on Computer Vision (ICCV).