CYSHIELD

Variational Auto Encoder Report

Introduction

This project implements a Variational Autoencoder (VAE) for facial image synthesis using a pre-trained model. The objective is to take a reference image as input, encode it into the latent space, and then generate new outputs through decoding. Two generation modes are provided:

- **Variation Mode**, where Gaussian noise is added to the latent representation to produce diverse variations of the original face.
- **Cycle Mode**, where the image is repeatedly encoded and decoded to study how reconstructions evolve over multiple iterations.

The implementation includes loading the trained encoder–decoder architecture, preprocessing input images, applying the reparameterization trick for latent sampling, and saving the generated outputs. Experimental results are analyzed to evaluate reconstruction quality, variation strength, and stability across cycles.

Methodology

Encoder–Decoder Architecture

The implemented VAE consists of two main components: an **encoder** that maps input images into a latent distribution, and a **decoder** that reconstructs images from latent samples.

- **Encoder**: a series of convolutional layers with increasing channel depth (3 → 512), followed by two linear layers that output the mean and log-variance of the latent distribution. The reparameterization trick is applied to sample a latent vector.
- **Decoder**: transposed convolutional layers progressively upsample the latent representation back into an image. The final output is passed through a sigmoid activation to constrain pixel values to [0,1][0,1][0,1].
- **Latent Dimension**: set to 512, balancing reconstruction capacity with manageable latent space size.

Implementation Details

1. **Model Loading**
    - The encoder and decoder weights are loaded from a pre-trained checkpoint (model.pt).
    - Both models are set to evaluation mode and run on GPU if available.
2. **Image Preprocessing**
    - Input images are resized to 178×218 pixels to match the VAE's training resolution.
    - Images are converted to tensors and normalized to [0,1][0,1][0,1].
3. **Encoding and Decoding**
    - Encoding produces a mean μ and log-variance σ.

- The latent vector z is obtained using the reparameterization trick.

4. **Modes of Operation**
    - **Variation Mode**: perturb the latent vector z with Gaussian noise of adjustable strength to create multiple variations of the same input.
    - **Cycle Mode**: repeatedly encode and decode the reconstructed output, observing how quality and content evolve across iterations.

Experiments and Results

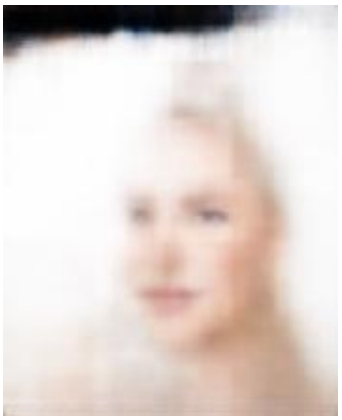
Experimental Setup

The experiments were conducted using a pre-trained VAE with a latent dimension of 512. The input image used for testing was *ronaldo.jpg*. Two modes of image generation were tested:

1. **Variation Mode**
    - Number of generated images: 10
    - Variation strengths: 0.2
    - Purpose: to study how noise in the latent space affects the diversity and quality of generated images.
2. **Cycle Mode**
    - Number of iterations: 10
    - Purpose: to observe how repeated encode–decode operations influence image stability and reconstruction fidelity.

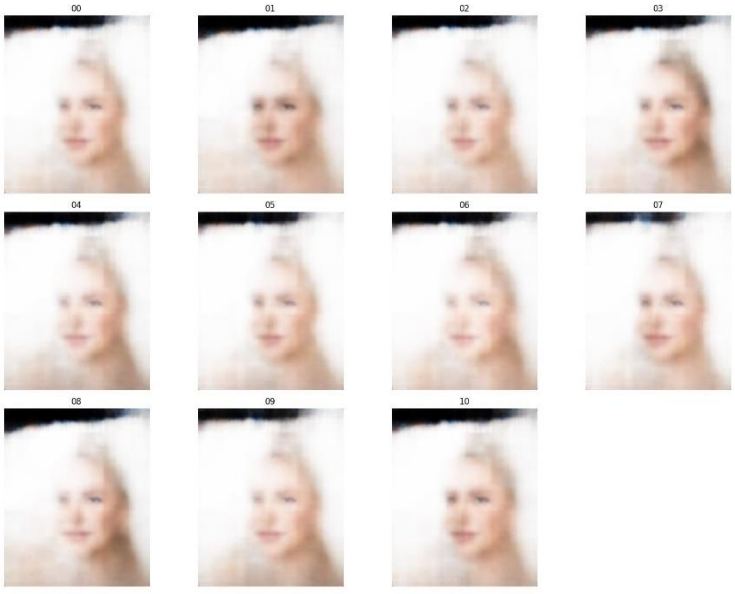
Results in Variation Mode

- **Reconstruction**: The VAE captures the overall facial structure but introduces significant **blurriness**, especially around the hairline and facial edges. This blur is expected because VAEs optimize for pixel-wise reconstruction, often producing smoother outputs rather than sharp details.



- **Variation (Image 7)**: The generated output diverges noticeably from the original male input, instead producing a more **feminine appearance**. This highlights how adding noise in the latent space can shift the model's interpretation of high-level attributes such as gender, hairstyle, or facial expression. At the same time, the blur becomes more pronounced, further reducing realism.

All Images:

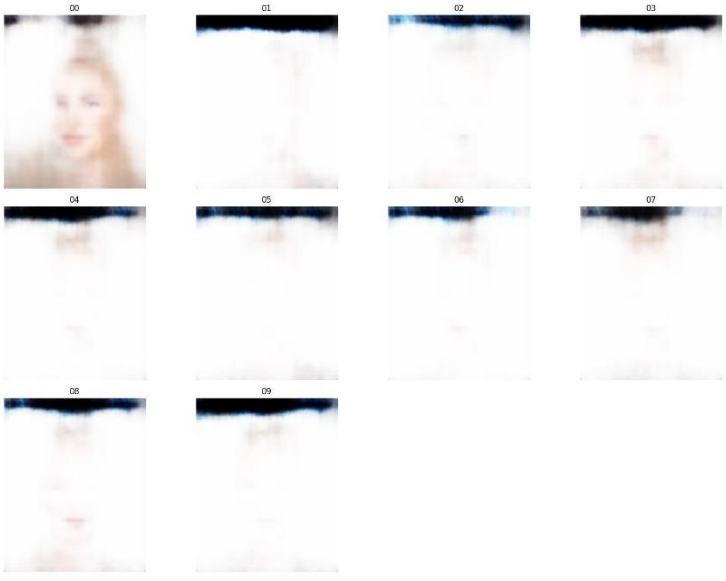


Results in Cycle Mode

- **Initial Reconstruction**: Even the first reconstructed image is heavily blurred compared to the original input. Although the rough structure of a face is still present, key details such as contours, eyes, and textures are missing.

- **Subsequent Cycles**: With each additional cycle, the degradation intensifies. After one cycle, the outputs no longer resemble a human face and instead collapse into indistinct blurred patterns.

All images:



Conclusion

This project implemented and tested a pre-trained Variational Autoencoder (VAE) for facial image synthesis using two approaches: **variation mode** and **cycle mode**. The experiments revealed several important insights:

- The VAE successfully reconstructed inputs and generated novel variations by perturbing the latent space, demonstrating its capacity for creative synthesis.
- However, the generated outputs were consistently **blurred**, reflecting the tendency of VAEs to produce smooth, low-frequency images due to their pixel-wise reconstruction objective.
- In **variation mode**, adding Gaussian noise produced diverse outputs but also led to unintended changes in high-level attributes such as gender, indicating that the latent space was not fully disentangled.
- In **cycle mode**, repeated encoding–decoding caused errors to accumulate, ultimately degrading the outputs into indistinct blurred patterns, showing the model's limitations in preserving information across iterations.

Overall, the study confirms both the **strengths** of VAEs in generating variations from latent space and their **weaknesses** in producing sharp, realistic images and maintaining long-term reconstruction fidelity. Future improvements could include using more advanced generative models such as **β-VAE, VQ-VAE, or diffusion models**, which address issues of disentanglement and sharpness more effectively.