VQVAE

Problem Statement:

The goal of this assignment is to develop a deep learning solution for generating high-quality images of skin lesions. This involves the use of advanced generative models to learn from a dataset of skin lesion images and then produce new, realistic images that could potentially expand the dataset or be used for further research in medical diagnostics.

Objective and our Approach

1. Train a Vector-Quantized Variational Autoencoder (VQ-VAE): To efficiently encode and decode high-dimensional image data while capturing meaningful latent representations of skin lesions.
2. Develop an Auto-Regressive Model (Gated PixelCNN): To generate new, realistic images based on the learned latent space representations from the VQ-VAE.

Dataset

- The ISIC dataset includes thousands of high-quality, dermoscopic images of various skin lesions, categorized by diagnosis, making it a valuable resource for research and educational purposes in medical imaging.

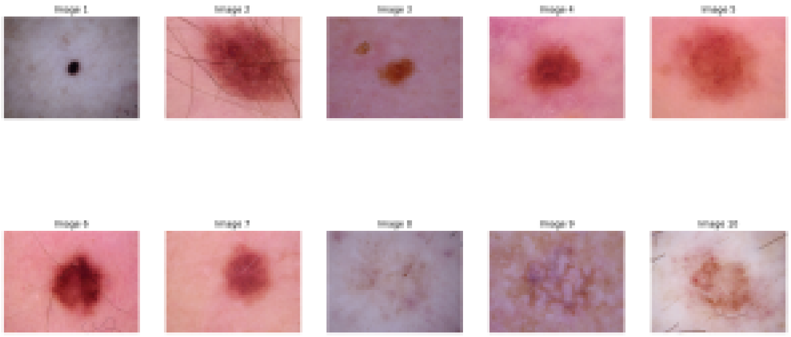


Figure: ISIC dataset samples

Preprocessing

1) **Normalization :** The normalization implemented here scales the pixel values of each image to a range between 0 and 1 based on the minimum and maximum values in that image. This specific type of normalization is sometimes referred to as **min-max scaling** or feature scaling.

2) **Resizing:** Standardizing all images to a fixed size (256x256 pixels) to maintain consistency in input dimensions and save computational resources. 3) **Transformations :** Applying transformations such as horizontal and vertical flips, rotations, and color jitter. These transformations help the model become invariant to orientation and lighting differences and enhance the generalizability of the learned features. One of each was randomly passed for the input to avoid extra computational resources.

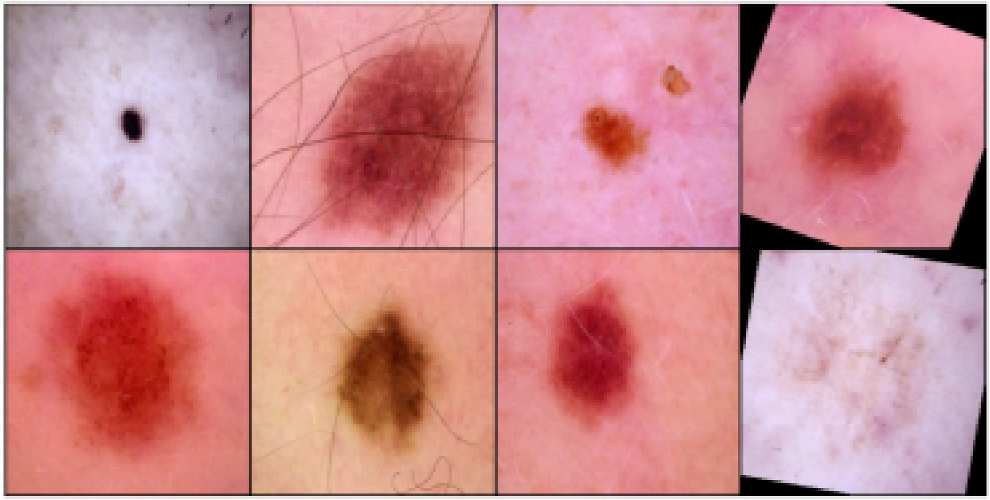


Figure: Final images in the dataloader after applying transformations

Implementation of VQVAE :

The VQ-VAE model is designed to efficiently encode images into a latent space,discretize the latent representations into a set of embeddings (codebook), and then reconstruct the images from these embeddings. This process is split across three core components: the Encoder, the Vector Quantizer, and the Decoder.

Components of VQ-VAE

1. Encoder:
   - The purpose of the encoder is to compress the input image into a

smaller, dense representation, capturing the essential features of the image in a latent space.

- Residual blocks within the encoder help in building a deeper network by enabling training through the addition of shortcut connections that skip one or more layers.

2. Vector Quantization :

- It takes the continuous latent space vectors produced by the encoder and quantizes them into a finite set of discrete embeddings.
- This quantization process involves mapping each vector to the nearest vector in a predefined set (codebook).
- The codebook is learned during training and enables the model to generate more stable and coherent outputs during the decoding phase. 3. Decoder :
- The Decoder's role is to reconstruct an image from the quantized latent representations.

Loss Calculation :

1. Reconstruction Loss : The reconstruction loss measures how well the reconstructed image matches the original input image.

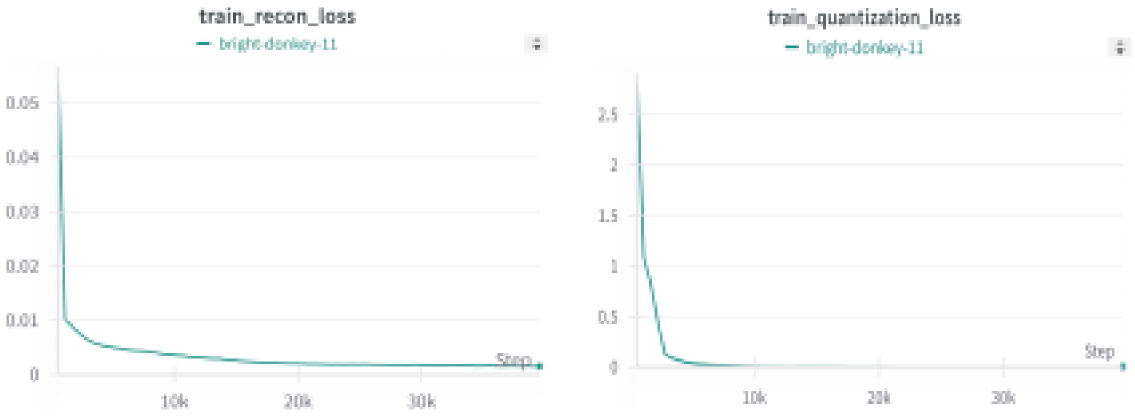
$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

2. Quantization Loss :

- The quantization loss consists of two components: the codebook loss and the commitment loss.
- Codebook loss encourages the selected vectors (quantized vectors) in the codebook to move closer to the encoder outputs (latent vectors), minimizing the distance between them.
- The commitment loss penalizes large distances between the encoder outputs and their corresponding quantized vectors.Here we have taken 0.25 as β.

Total Loss=Reconstruction Loss+Codebook Loss+β×Commitment Loss • The reconstruction loss started with an initial spike but quickly stabilized and decreased to below 0.01, demonstrating the model's increasing accuracy in reproducing the input images.

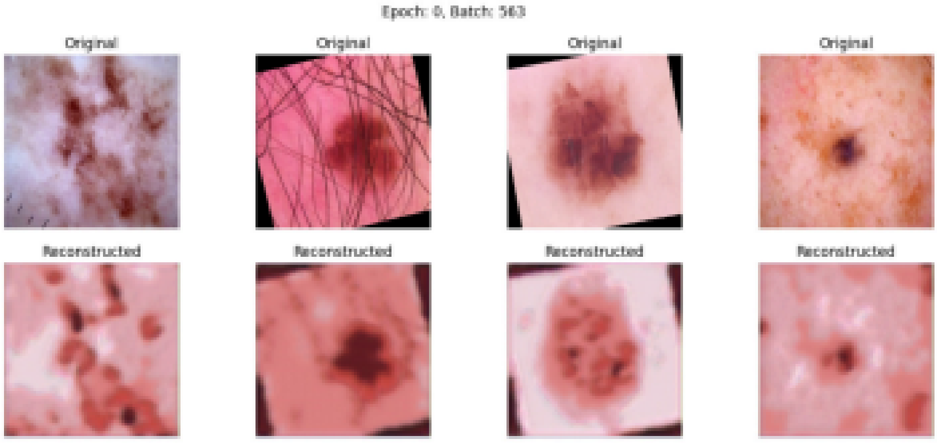
- Similarly, the quantization loss showed a sharp decline in the initial epochs and continued to diminish, reflecting the effective learning and stabilization of the quantization process within the VQ-VAE's encoder-decoder structure.



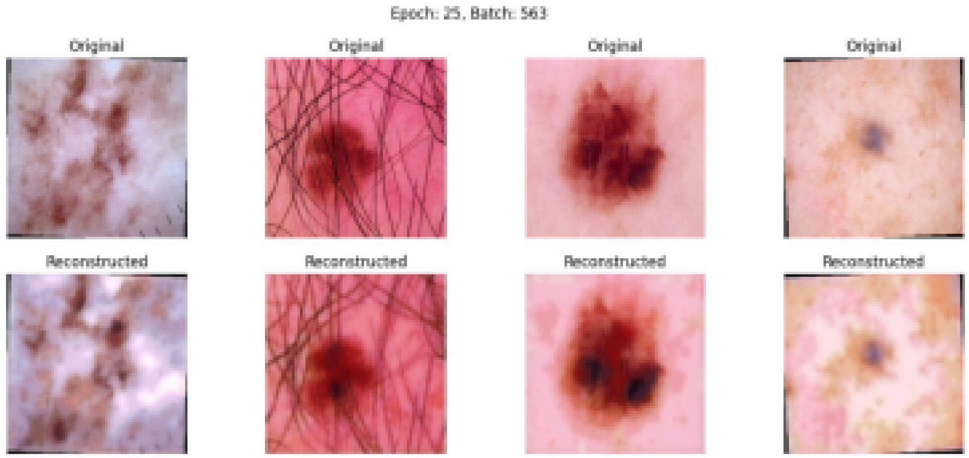
Reconstruction Loss Quantization loss

Results

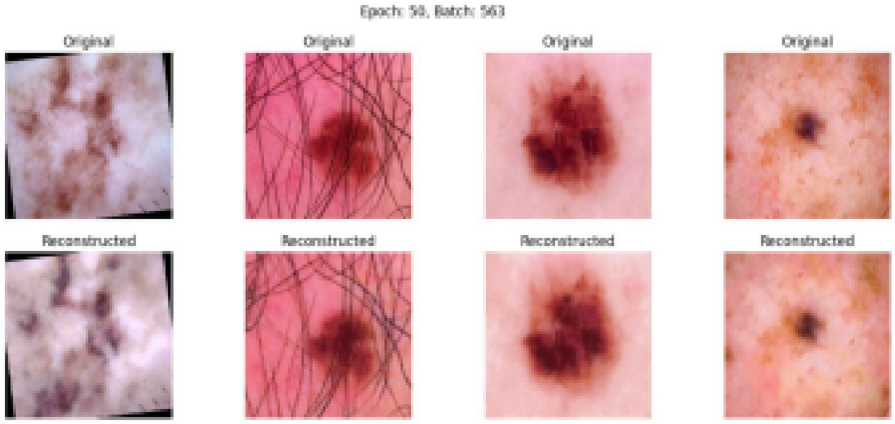
- As illustrated in the visual results, the VQ-VAE model demonstrates substantial improvement in the reconstruction quality of skin lesion images across training epochs.



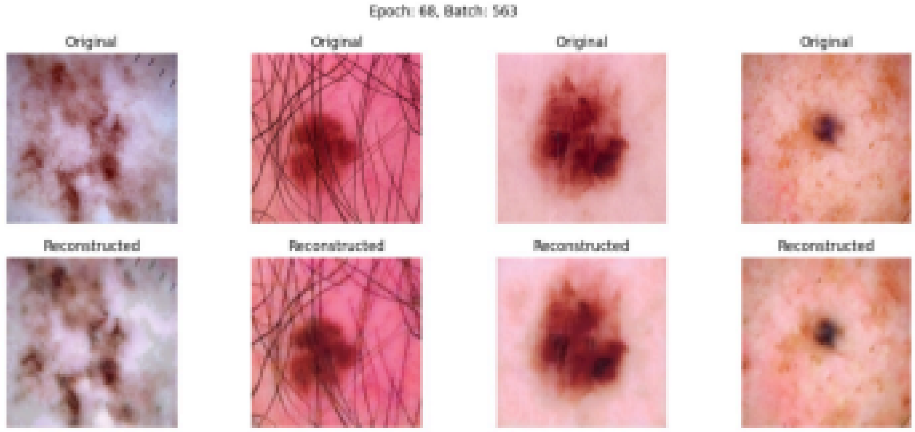
- By epoch 25, there is noticeable progress in how the reconstructed images retain more of the original details and colors, though some blurring remains evident.



- By epoch 50, the reconstructed images exhibit even greater clarity and color fidelity, closely mirroring the original images.



- This progression validates the model's ability to learn effective representations of complex medical images over time, enhancing its potential utility in diagnostic applications where precise image details are crucial.



Conversion of Dataset Using Trained VQ-VAE Model for Gated PixelCNN Training

- The images are encoded into a latent space representation using the encoder part of the VQ-VAE model.
- These continuous latent vectors are then quantized.
- The quantization process involves mapping each vector to the closest vector in the predefined codebook, resulting in discrete indices.
- These indices effectively compress the image data into a compact, discrete form that retains the essential information in a format suitable for the next stage of generative modeling.

Gated PixelCNN :

1. Gated Activation Function:
    - It splits the convolution output into two halves and applies a **tanh** activation to one half and a **sigmoid** activation to the other, then multiplies the outputs.
    - This operation helps in managing gradients during training, preventing vanishing or exploding gradients, and enabling deeper models.
2. **Gated Masked Convolutional Layers:**
    - These layers are crucial for maintaining the autoregressive property of the model, ensuring that each output pixel can depend only on the previous pixels in the input.
    - This is achieved by masking the convolutions so that no information from future pixels is used in predicting current pixels.
    - The layers are specifically designed with separate pathways for vertical and horizontal stacks, allowing the model to effectively capture dependencies in both dimensions.
3. **Residual Connections:** Incorporated within the gated masked convolutions, these connections help in stabilizing the training of deep networks by allowing gradients to flow through multiple layers directly.

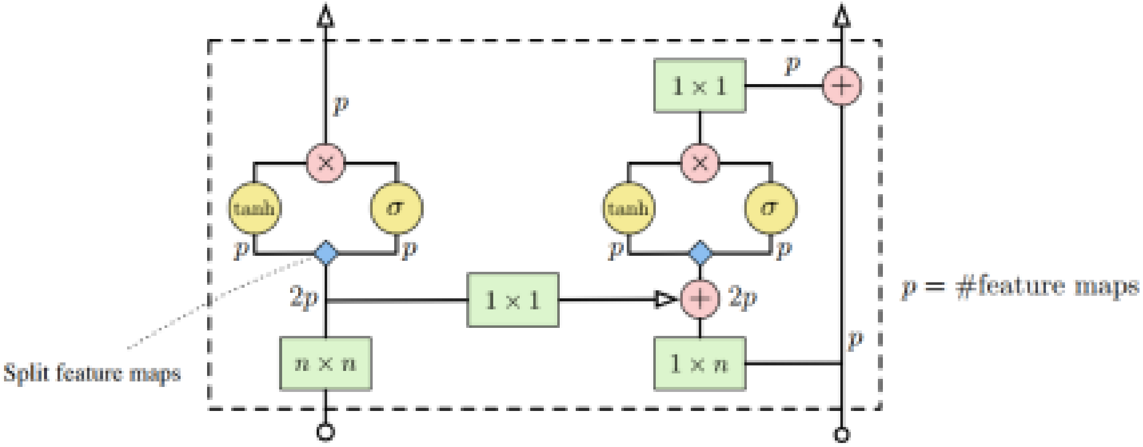


Figure: Architecture of Gated PixelCNN

- Cross-entropy loss is employed to train the Gated PixelCNN, which measures the discrepancy between the predicted probabilities of the next index and the actual index from the dataset.
- An Adam optimizer is used for minimizing this loss, we had experimented with SGDMomentum loss as well to solve the plateau issue but Adam performed significantly well.

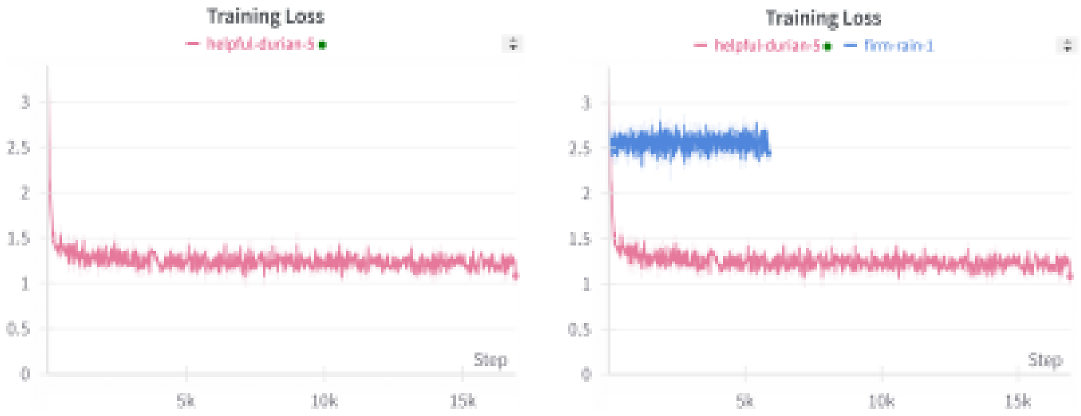


Figure: training loss during training of gated PixelCNN

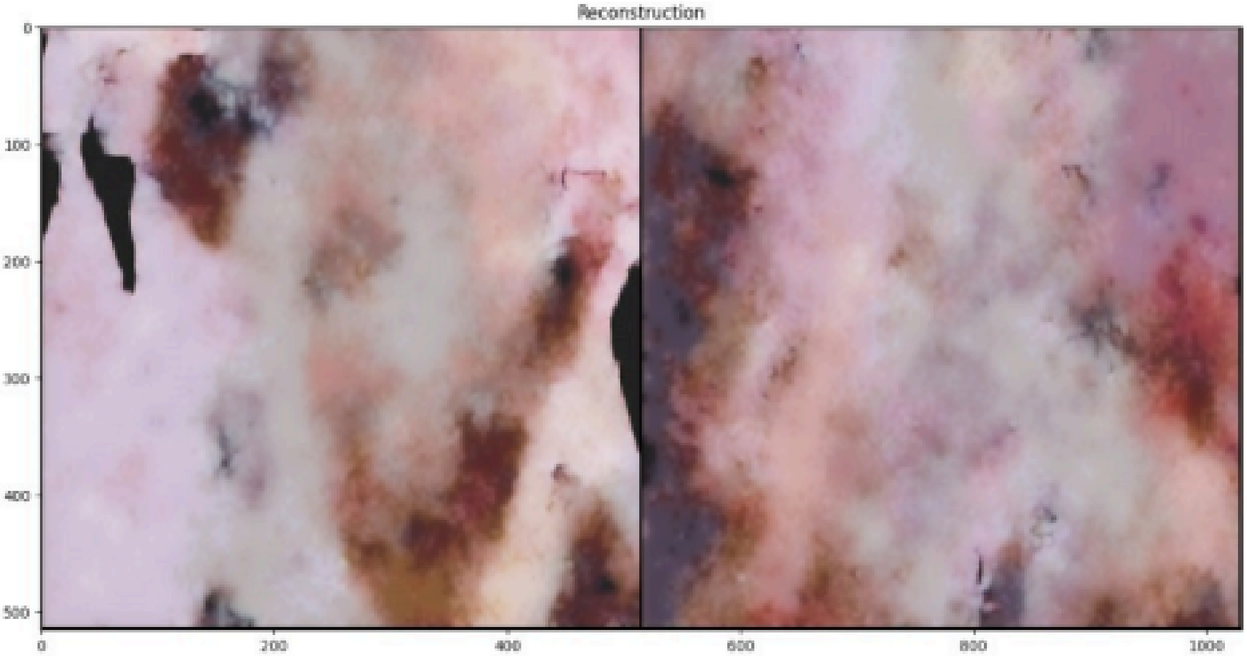
Generating Images Using Trained VQ-VAE and PixelCNN Models:

- In the final phase of our project, we combined the trained VQ-VAE and PixelCNN models to generate new skin lesion images.
- The process starts by using PixelCNN to sample discrete latent space indices, which are then converted into quantized vectors using VQ-VAE's embedding.

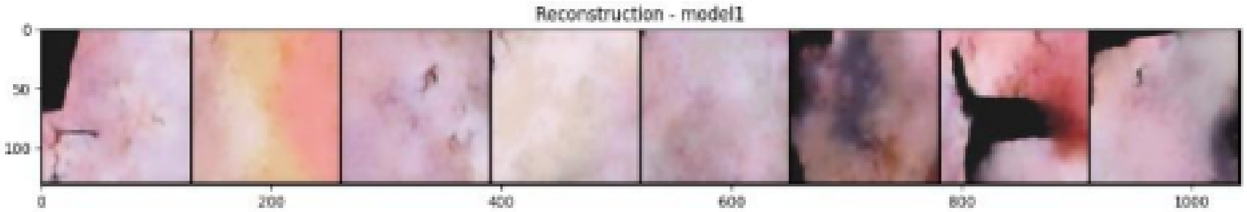
These vectors are decoded back into images using the VQ-VAE's decoder.

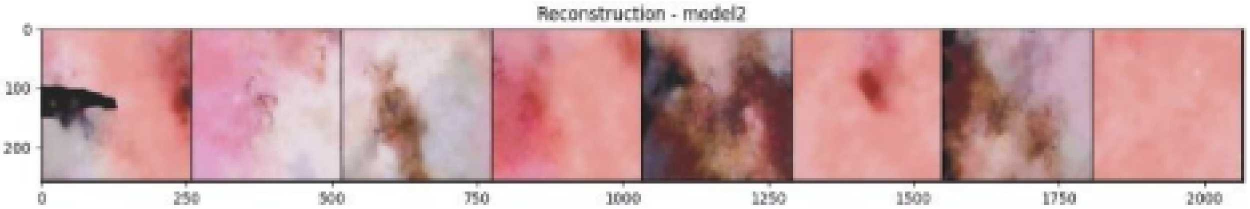
Experimentations and Results:

- We experimented with three different sizes for the latent space grid: 32x32, 64x64, and 128x128.
- **The 128x128 configuration yielded the best results in terms of image clarity and detail**. This improvement can be attributed to a **larger latent space providing a more granular and detailed representation of the images**, allowing the model to capture and generate finer details and textures more effectively. Larger grids in the latent space offer more capacity to encode distinct features of the images, leading to higher-quality reconstructions.



- Conversely, smaller configurations like 32x32 and 64x64, **while faster in processing**, tended to produce images that were **less detailed and sometimes blurrier**, indicating a less effective capture of the complexities inherent in skin lesion images.





Conclusion :

- The use of VQ-VAE for encoding and decoding images, combined with PixelCNN for generating new image data, proved to be a potent combination. VQ-VAE effectively compressed images into a latent space and maintained high-quality reconstructions, while PixelCNN generated new, diverse images based on the learned distribution of the latent codes.
- The project confirmed that the quality of generated images is highly dependent on the size of the latent space.
- The ability to generate realistic skin lesion images holds significant potential for medical research, particularly in areas where data is scarce or privacy concerns limit the availability of real images.