DL - Assignment - 5

Members - Ritu Singh, Bikram Majhi, Prabhat Ranjan

Colab Link - M23CSA017_M23CSA007_M23CSE017.ipynb

Generative Modeling of Skin Lesions: VQ-VAE and Autoregressive Models

Objective:

Utilizes VQ-VAE to encode and decode skin lesion data, capturing meaningful latent representations. Additionally, Auto Regressive Models generate realistic images from learned latent space representations.

Dataset - ISIC Skin Lesion Images.

The dataset includes:

Train Data: 9015 training images

Test Data: 1000 test images

Data Preprocessing

A custom data loader is designed for the above dataset for data preprocessing.

Resizing: Images are resized to 256*256 pixels. This is performed for training datasets.

Data augmentation:

Data augmentation includes random horizontal and vertical flips, as well as color jittering. The images are resized to (256, 256) and normalized with a mean of (0.5, 0.5, 0.5) and a standard deviation of (0.5, 0.5, 0.5).

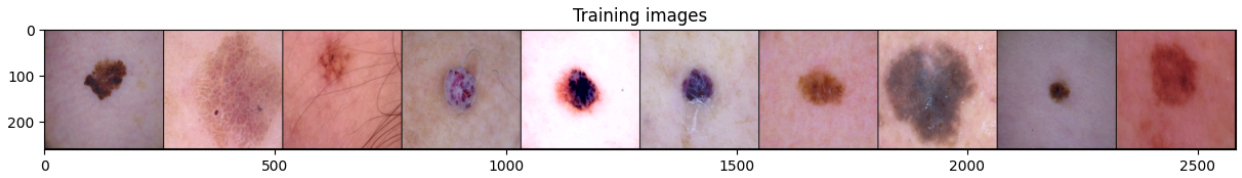


Fig 1: Visualization of training images (after data augmentations)

Normalization: Scaling the pixel values so that they are centered around 0 with a standard deviation of 1.

Class Imbalance in the Dataset:

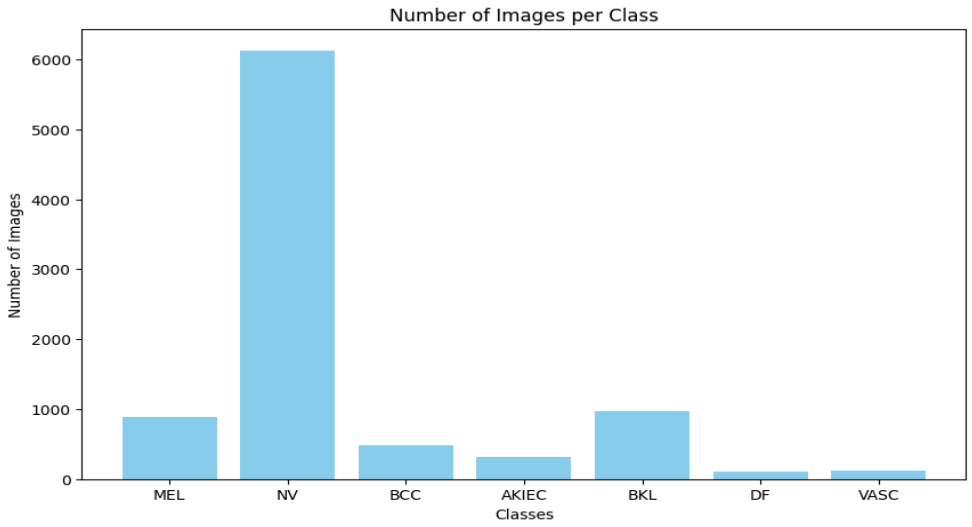


Fig 2: Distribution of training images across classes

NV class has the highest number of images, with a count of 6130. This indicates that the dataset is heavily skewed towards this class.

VQ-VAE Model

Encoder: A convolutional neural network (CNN) downsamples the input images into a compact latent representation. Residual blocks are used for efficient feature extraction.

Vector Quantization (VQ) Layer:

The encoder output is discretized using a learned embedding codebook.

The model learns a finite set of embeddings, and each input representation is mapped to its closest embedding.

Commitment cost is used to encourage the model to effectively utilize the codebook.

Decoder: A CNN with transposed convolutions up samples the discretized latent representation to reconstruct the input image.

Loss Function

The model is trained using a combined loss:

Reconstruction loss: Mean squared error (MSE) between the input and reconstructed image.

Quantization loss: Measures the difference between the encoder output and its quantized representation.

Experimental Setup

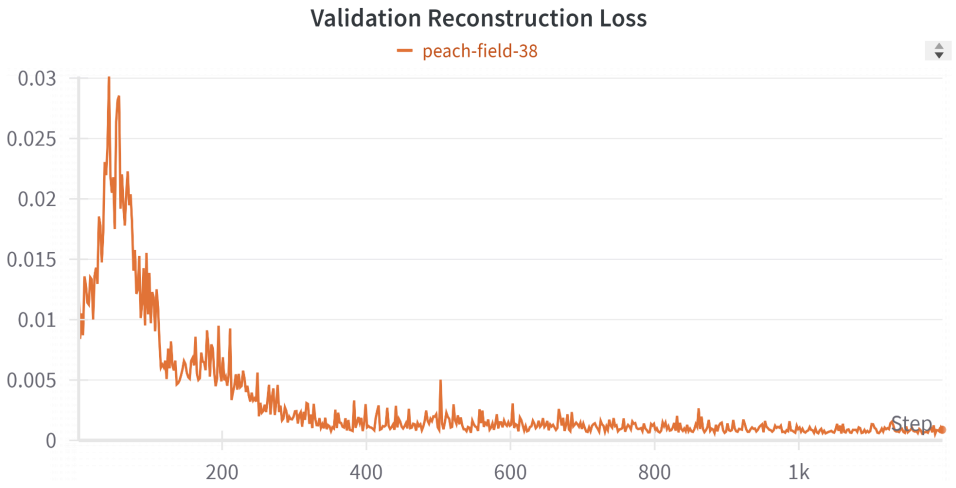
Hyperparameters:

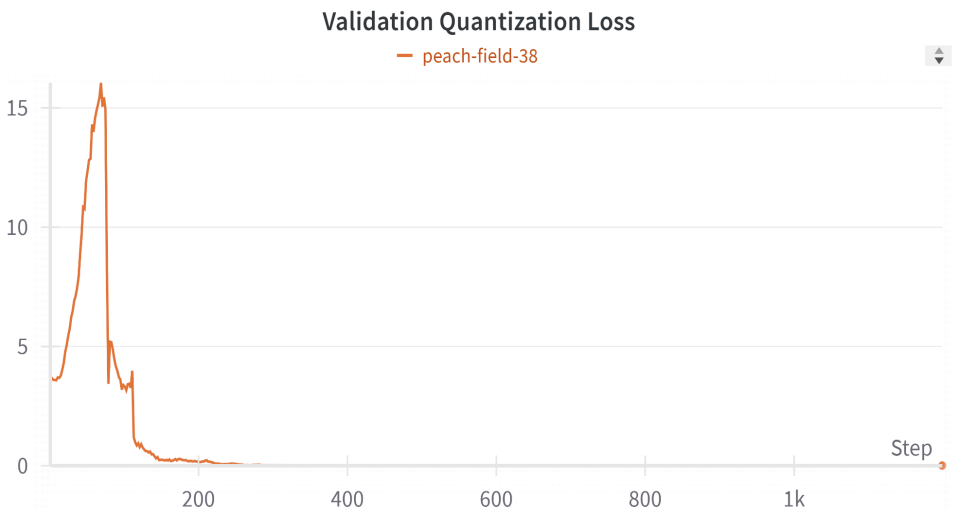
1. Batch size: 32
2. Epochs: 300
3. Learning rate: 0.0002
4. Weight decay: 0.0001
5. Number of embeddings: 128
6. Embedding size: 512

Results

Quantitative:

Reconstruction loss and quantization loss on training and validation sets over epochs (plotted using wandb).


Validation Reconstruction Loss
peach-field-38
0.03
0.025
0.02
0.015
0.01
0.005
0
200
400
600
800
1k
Step

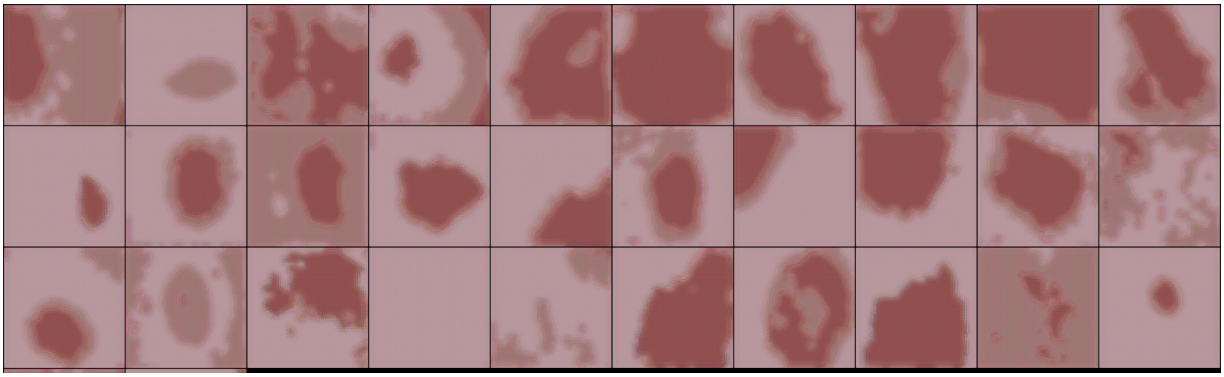

Validation Quantization Loss
peach-field-38
15
10
5
0
200
400
600
800
1k
Step

Qualitative:

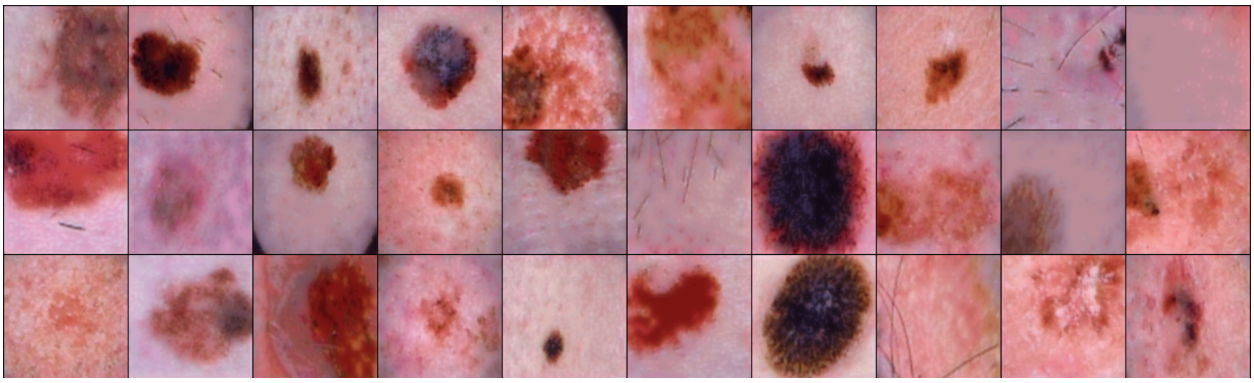
Visualizations of reconstructed images over epochs, saved periodically.

Comparison of original and reconstructed images side-by-side in grids.

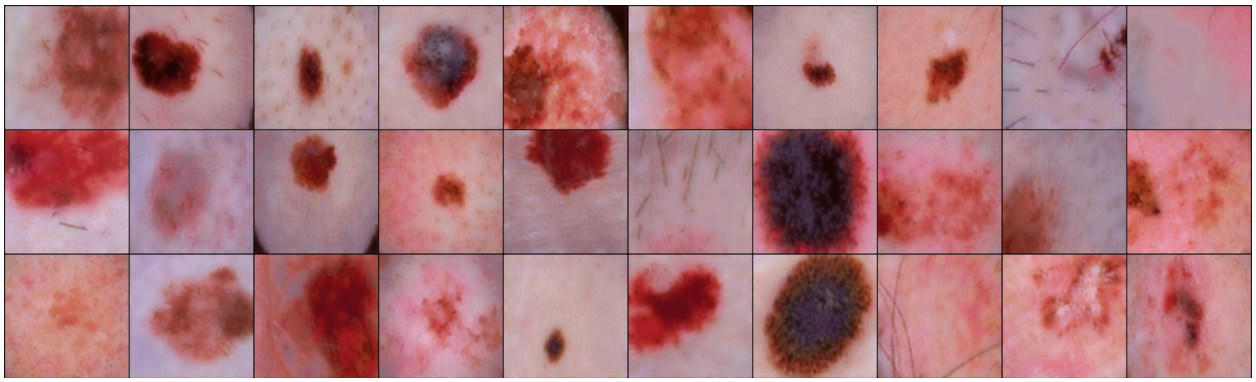
Epoch 1



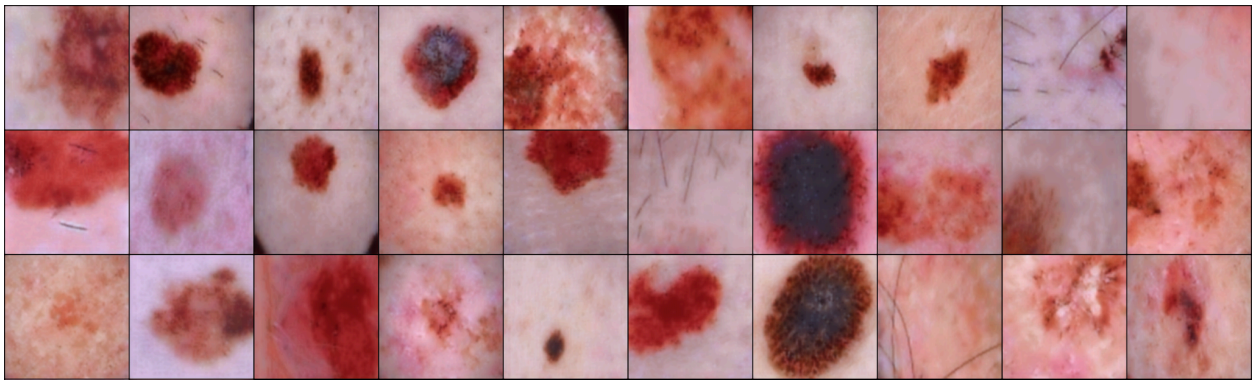
Epoch - 100



Epoch - 200



Epoch - 300



Reconstruction Quality:

Improved reconstruction quality observed with increasing epochs of training.

Reconstruction on test dataset

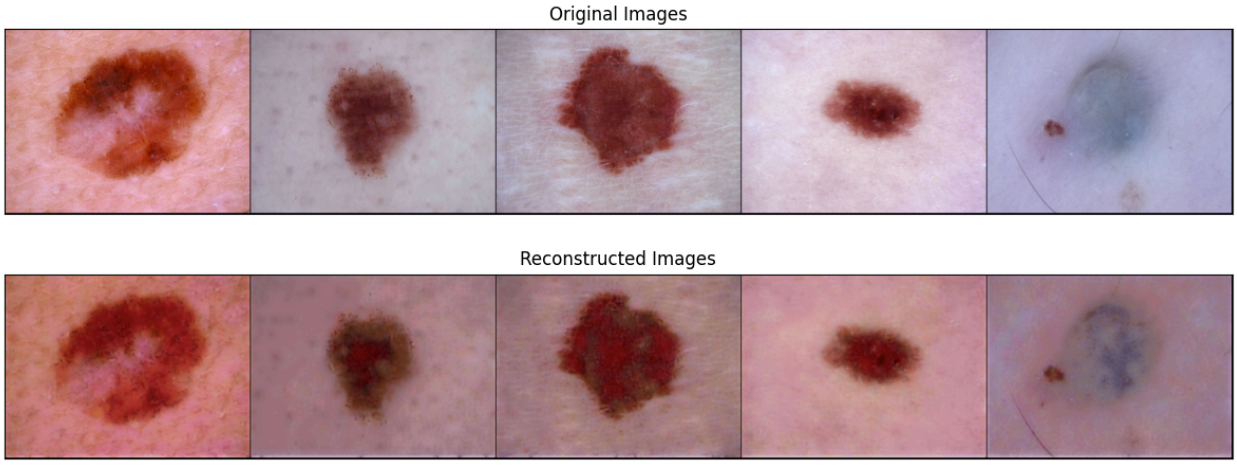


Fig 1: Visualization of reconstructed images on test dataset

Autoregressive Model - PixelCNN

We use a modified dataset to train the autoregressive model to generate diverse realistic images using the codebook and the decoder trained previously. In the modified dataset each image is replaced by its corresponding quantized latent vector from the VQ-VAE model. Gated PixelCNN is trained to model the distribution of the quantized vectors produced by the VQ-VAE. This allows it to generate new quantized vectors, which can then be decoded by the VQ-VAE into new images.

Experimental Setup

Hyperparameters:

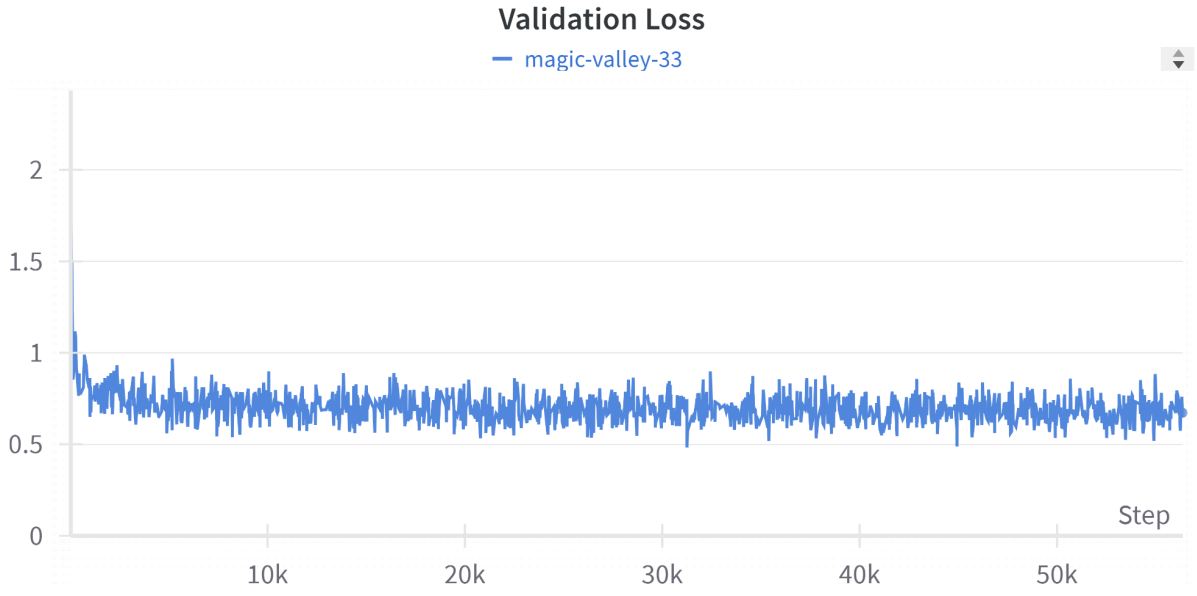
- Batch size: 32
- Epochs: 500
- Learning rate: 0.0002
- Weight decay: 0.0001
- Number of embeddings: 128
- Embedding size: 512

Results

Quantitative:

Training loss and validation loss on training and validation sets over epochs (plotted using wandb).

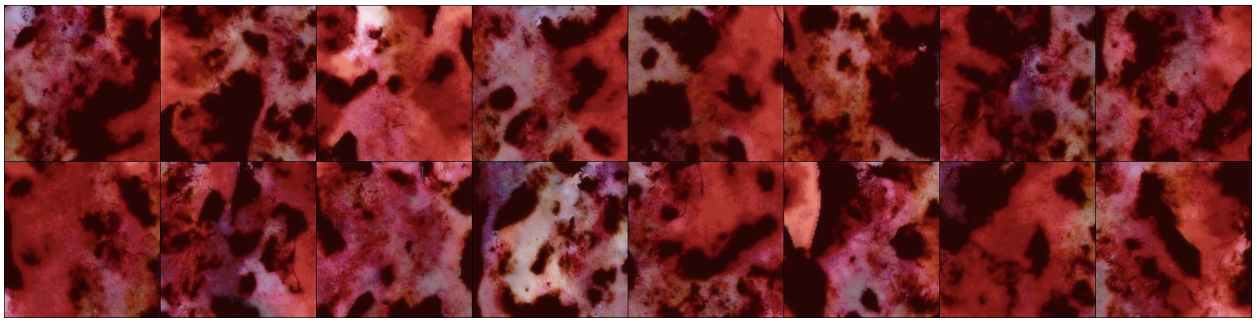

Training Loss
magic-valley-33
4
3
2
1
0
Step
10k
20k
30k
40k
50k


Validation Loss
magic-valley-33
2
1.5
1
0.5
0
Step
10k
20k
30k
40k
50k

Qualitative:

Visualizations of reconstructed images after the completion of training.

Sampling images from the model.



End-to-end process overview:

1. VQ-VAE Training:
   A Vector Quantized Variational AutoEncoder (VQ-VAE) was trained on a dataset of images. The VQ-VAE consists of an encoder that maps input images to a latent space, a codebook of latent vectors, and a decoder that maps quantized latent vectors back to the input space. The VQ-VAE was trained to minimize the reconstruction loss between the original and reconstructed images, as well as a loss term that encourages the encoder's outputs to match the codebook vectors.

2. Gated PixelCNN Training:
   A Gated PixelCNN, an autoregressive model, was then trained to model the distribution of the quantized vectors produced by the VQ-VAE. The PixelCNN was trained to predict each pixel in the image given all previously generated pixels, ensuring that the model captures the dependencies between pixels in the image.

3. Combining VQ-VAE and PixelCNN:
   A final model, PixelCNNVQVAE, was created by combining the trained VQ-VAE and PixelCNN models. This model uses the PixelCNN to generate new indices in the latent space, converts these indices into quantized vectors using the VQ-VAE's codebook, and then decodes these quantized vectors into new images using the VQ-VAE's decoder

4. Image Generation:

The PixelCNNVQVAE model was then used to generate new, realistic images. This was done by first sampling a batch of indices from the PixelCNN, converting these indices into quantized vectors using the VQ-VAE's codebook, and then decoding these quantized vectors into images using the VQ-VAE's decoder.

Analysis

We can observe that the images produced by the trained VQ-VAE model are high-quality reconstructions of the original dataset. The images, which are close-ups of various skin conditions, display a variety of patterns and colors, demonstrating the model's ability to capture the complex features of skin lesions effectively.

However, when it comes to generating new images using the Gated PixelCNN trained on the latent discrete representations from the VQ-VAE, the results are not as expected. The generated images do not closely resemble the skin lesion images, indicating that the PixelCNN model may not have fully captured the distribution of the latent space representations.

Conclusion

The VQ-VAE model has demonstrated excellent performance in reconstructing skin lesion images, capturing the intricate details and variations in the data. This success underscores the potential of VQ-VAE models in tasks that involve complex and diverse visual data.

On the other hand, the Gated PixelCNN, while a powerful model for generating new data in many contexts, may require further tuning or a different approach when applied to the specific task of generating skin lesion images based on learned latent representations. This could involve adjustments to the model architecture, training process, or even exploring other types of autoregressive models.

Despite the current results, the combination of VQ-VAE and Gated PixelCNN holds promise. With further refinement and experimentation, it's plausible that the generation of realistic and diverse skin lesion images could be achieved.