Image Generation using VQ-VAE and Pixel CNN

Divyaansh Mertia (M23CSE013)
Obed Jamir (D23CSE004)
Ritobina Ghosh (M23CSA021)

May 1, 2024

Indian Institute of Technology, Jodhpur

Contents

List of Tables

List of Figures

Chapter 1

Model Architecture

1.1 Overview

Our project employs a sophisticated architecture combining a Vector Quantized Variational Autoencoder (VQ-VAE) with a PixelCNN to perform both analysis and generation of dermatological images effectively. This architecture is designed to capture the complex patterns typical of medical imagery, specifically within the ISIC dataset, and consists of four main components:

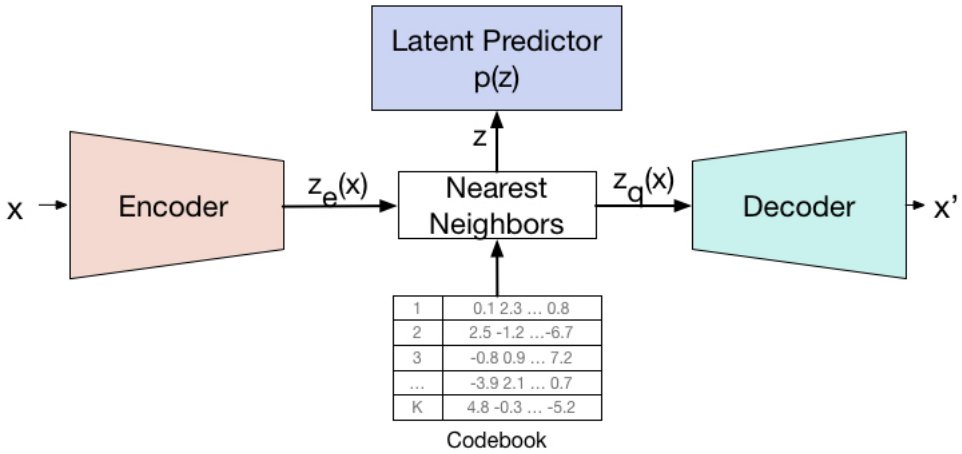


Fig. 1.1: VQ-VAE Architecture

1. **Encoder:** Compresses high-dimensional image data into a lower-dimensional latent space.
2. **Vector Quantizer:** Transforms the continuous latent representation into a discrete format, enabling effective reconstruction and generation.
3. **Decoder:** Reconstructs images from the quantized latent representation.
4. **Auto-Regressive Model (PixelCNN):** Generates new images by modeling the distribution of the latent representations.

1.2 Encoder

The Encoder is a crucial component designed to reduce the dimensionality of the input images while capturing essential information in a dense latent representation.

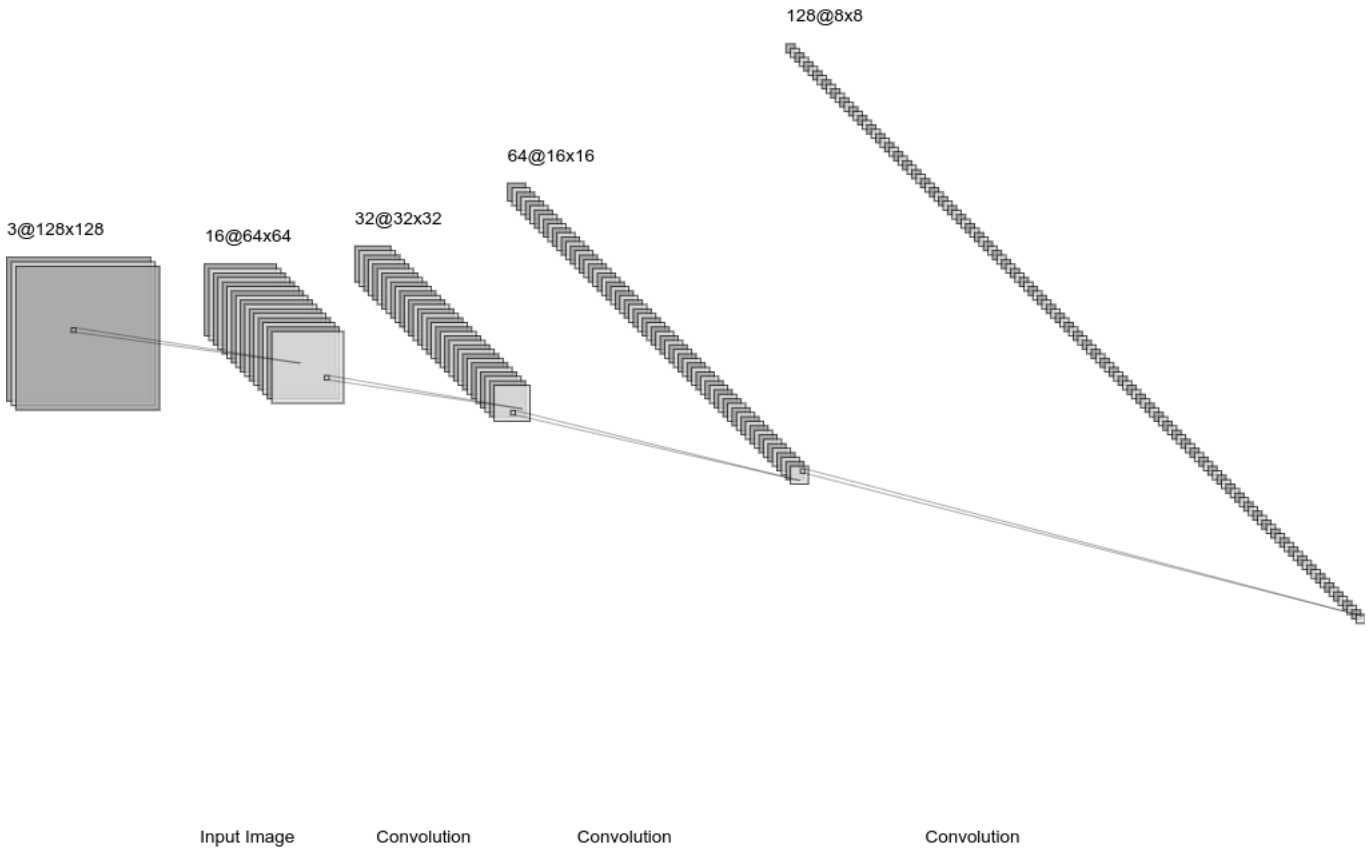


Fig. 1.2: Encoder Architecture for VQ-VAE

1.2.1 Functional Specifications

- **Input:** The encoder accepts 128x128 pixel images with 3 color channels.
- **Architecture:** The network architecture features several layers of convolutions, each followed by batch normalization and ReLU activation. These layers progressively decrease spatial dimensions and increase the depth of feature maps, culminating in a compact feature representation of the input image.

1.3 Vector Quantizer

Situated between the encoder and decoder, the Vector Quantizer is pivotal in converting the dense, continuous latent features into a discrete format, facilitating both effective reconstruction and generative processes.

1.3.1 Operational Details

- **Quantization Process:** The encoder's output is flattened and then mapped to a discrete set of embeddings (codebook vectors). This mapping is determined by calculating and minimizing the distance between each feature and the embeddings, ensuring an efficient representation of the input features.
- **Commitment Loss:** Included to ensure the quantization does not deviate significantly from the original latent features, thereby preserving the fidelity of reconstruction.

1.4 Decoder

The Decoder mirrors the Encoder's architecture in reverse, using transposed convolutions to gradually upscale the quantized representations back to the original image dimensions.

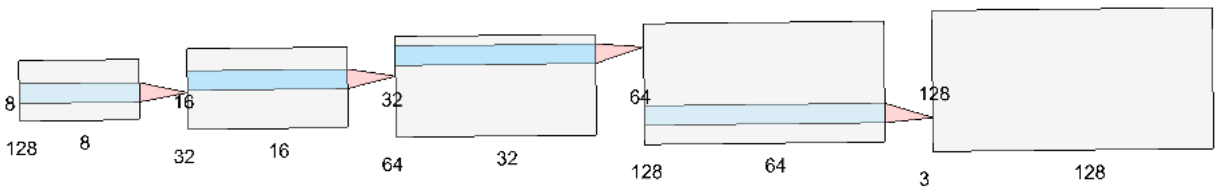


Fig. 1.3: Decoder Architecture for VQ-VAE

1.4.1 Architectural Configuration

- **Upscaling:** Begins with dense, quantized features and employs successive layers of transposed convolutions, batch normalization, and ReLU activations to increase spatial resolution while decreasing feature depth.
- **Final Layer:** The output layer reconstructs the original image dimensions and color channels, using a sigmoid activation to normalize the reconstructed pixel values.

1.5 Auto-Regressive Model (PixelCNN)

The PixelCNN plays a critical role as an auto-regressive model in generating new images from the discrete latent representations provided by the VQ-VAE's Vector Quantizer. This section details the functionality and integration of the PixelCNN in our model architecture.

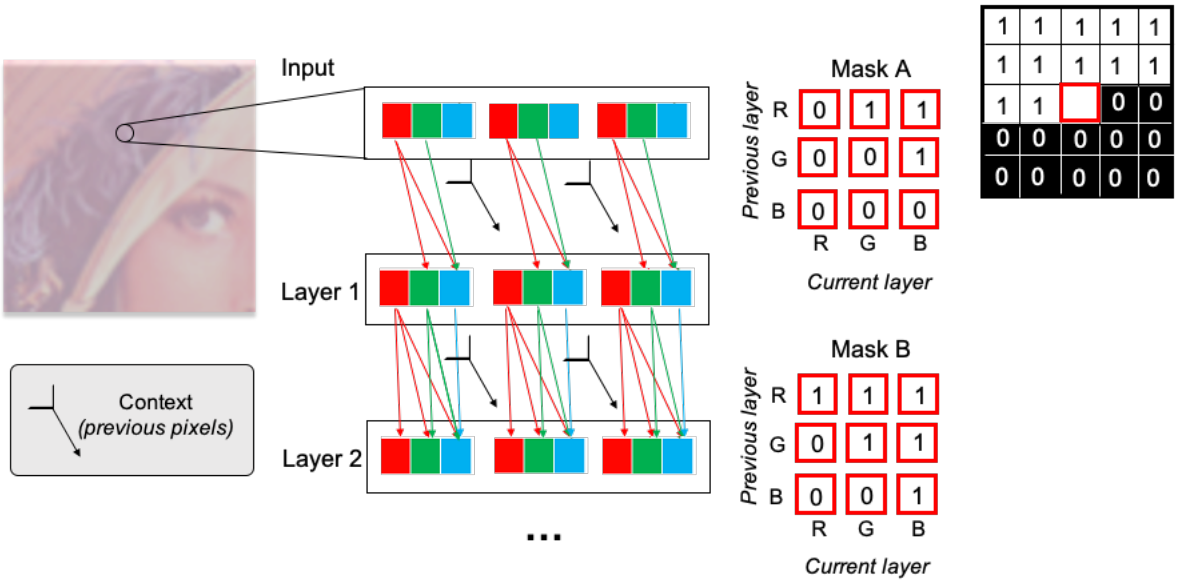


Fig. 1.4: Our Auto Regressive Model Pixel CNN

1.5.1 Functional Overview

PixelCNN utilizes the auto-regressive property to generate images pixel by pixel. Each pixel's value is predicted based on the previously generated pixels, adhering to a raster-scan order. This ensures that the generation of each pixel can consider all previously generated context, which is crucial for maintaining the structural integrity of generated images.

1.5.2 Generation of Indices

- **Input**: The model receives quantized latent codes, which are discrete indices representing compressed image features.
- **Operational Mechanism**: Starting with an empty grid, PixelCNN fills in this grid one pixel at a time. At each position, it predicts a distribution over possible pixel values (indices) based on the context provided by the already filled-in pixels.
- **Training**: During training, the model learns the conditional distributions of pixel values, optimizing its parameters to minimize prediction errors across the training dataset using cross-entropy loss.

1.5.3 Conversion from Indices to Vectors

Once PixelCNN generates a full set of new indices for an image, these indices are mapped back to their corresponding vectors using the codebook established during the Vector Quantizer's training. This step converts the discrete indices back into continuous vector representations, preparing them for the final image reconstruction phase.

1.5.4 Image Reconstruction

- **Decoder Integration**: The continuous vectors obtained from the indices are input into the Decoder. The Decoder architecture, designed to mirror the Encoder, uses transposed convolutions to progressively upscale these vectors back to the original image dimensions.
- **Final Output**: The Decoder reconstructs the image detail by detail, ultimately producing a high-resolution output that reflects both the learned data distribution and the nuances captured by the VQ-VAE's encoding process.

Chapter 2

Visualizations

2.1 Reconstructed Images

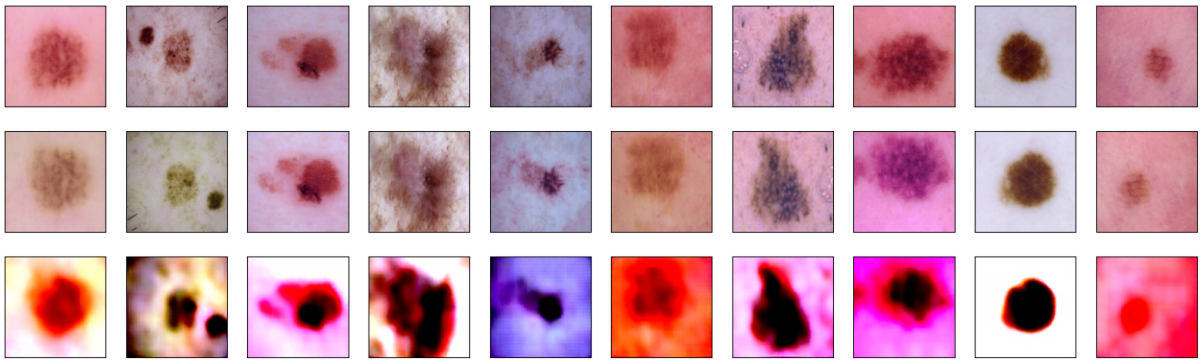


Fig. 2.1: Reconstructed Images:

2.2 Histogram of Codebook Vector Usage

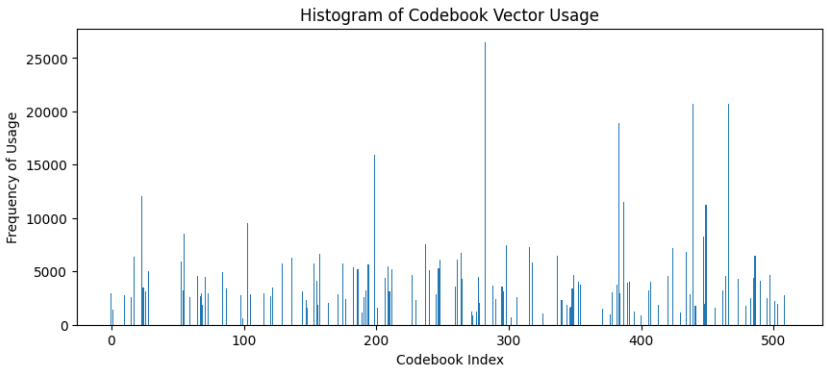
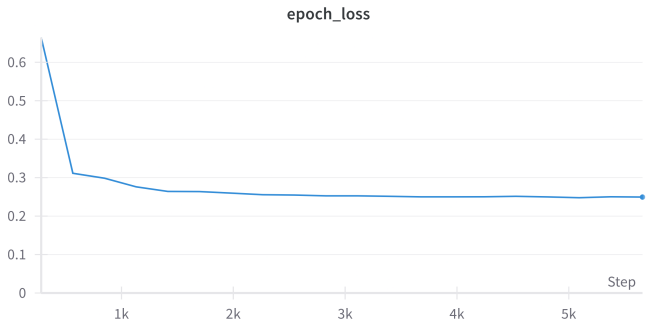


Fig. 2.2: Codebook Vector Usage

2.3 Loss Plots Using WandB

2.3.1 VQ-VAE

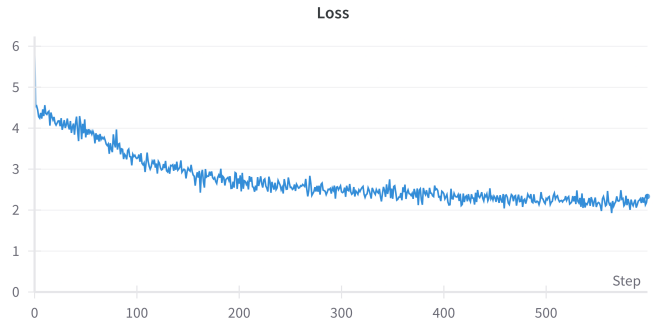


(a) Epoch Loss

(b) Batch Loss

Fig. 2.3: VQ-VAE Loss

2.3.2 Pixel-CNN



(a) Epoch Loss

Fig. 2.4: Pixel CNN Loss

2.4 Generated Images

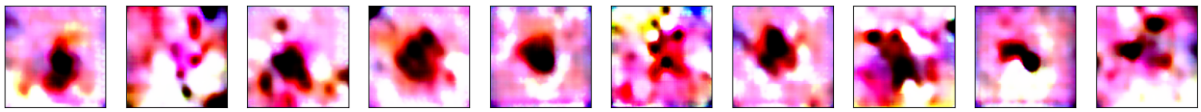


Fig. 2.5: Generated Images:

Chapter 3

Observations and Conclusion

3.0.1 Model Efficacy

- The **Encoder-Decoder structure**, coupled with the **Vector Quantizer**, effectively compresses and reconstructs high-dimensional dermatological images. This process retains crucial diagnostic features, essential for applications in medical imaging.
- **PixelCNN** demonstrates remarkable capabilities in generating new images from learned distributions, showcasing a profound understanding of complex skin lesion characteristics.

3.0.2 Quantization Benefits

- Introducing a **Vector Quantizer** within the VQ-VAE framework significantly enhances the model's ability to perform lossless reconstruction. The discrete latent space ensures a compact yet detailed capture of image features, which supports efficient data storage and processing.
- The quantization process maintains high fidelity in the reconstructed image quality.

3.0.3 Challenges Encountered

- A primary challenge was managing the **granularity of quantization** relative to model performance. Overly coarse quantization obscured critical fine details, while overly fine quantization raised the model complexity and computational demands.

3.0.4 Generalization and Overfitting

- Although the model performs exceptionally on the ISIC training dataset, achieving robust **generalization to unseen data** from other sources or variations within skin lesions remained challenging. Techniques like data augmentation in mitigating overfitting.

Chapter 4

References

1. "Visual AutoRegressive Modeling:Scalable Image Generation via Next-Scale Prediction," YouTube, [Online]. Available: https://youtu.be/yJ396Ksiv2s?si=2ziRmlKwpPh0D8xy.

2. "Visual AutoRegressive Modeling:Scalable Image Generation via Next-Scale Prediction," Paper, [Online]. Available: https://arxiv.org/abs/2404.02905.

Appendix A

Code

https://colab.research.google.com/drive/175-0rKQqkSdmwXil2Nf34NYTPk4lv0qP?usp=sharing