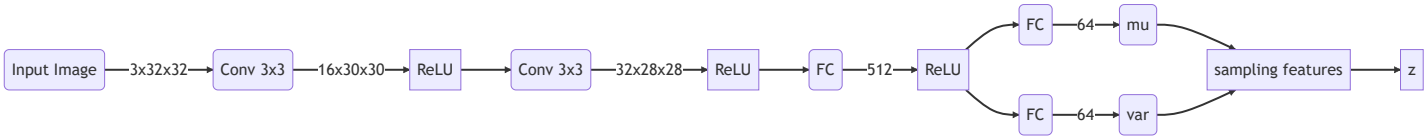
Image reconstruction

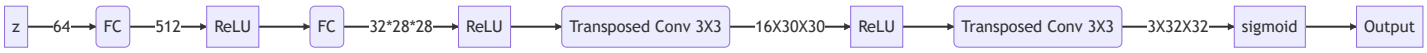
Wu Yuhang 520030910187

VAE Model Architecture

The network structure of the encoder is as follows:



The network structure of the decoder is as follows:



On the whole, the VAE model is divided into two parts: encoder and decoder. In the encoder part, I choose to use two layers of 3X3 convolutional layers for feature extraction, and use ReLU as the activation function.

After the convolutional layer is a fully connected layer, the input dimension is a flattened vector of 32X28X28, and the output is a 512-dimensional vector. Then two fully connected layers are used to map to the 64-dimensional potential feature space to obtain the mean and variance of the distribution. After sampling, a 64-dimensional vector is output as a feature output.

The decoder is similar in structure to the encoder. The deconvolution layer uses a 3X3 convolution kernel. It first maps to a high-dimensional space through a fully connected layer, and then uses deconvolution to generate an image. The activation function uses the ReLU function.

Implementation

The pytorch platform is used to build the model network, and the Adam optimizer is used to train and optimize the network parameters.

At first, only fully connected layers are used to map high-dimensional images to low-dimensional extracted features. Due to the lack of feature extraction process, the face generation effect is extremely poor.

So I chose to add two convolutional layers to extract the local features of the image. On this basis, the purpose of using a fully connected layer to map to a 512-dimensional space is to reduce feature loss. Finally, the features are mapped to a 64-dimensional latent space and sampled from a normal distribution.

At the beginning, I chose to map to the 20-dimensional latent space, but due to the small dimension, it is difficult for feature extraction to support the generation of complex faces.

The implementation of the decoder is to reverse the operation of the encoder. A model is trained to generate images from features.

Interpolation

Pass two pictures of different categories into the encoder of the model to get feature outputs z1 and z2. Select the parameter α to interpolate the two features.

$$z = \alpha z_1 + (1-\alpha)z_2, \alpha \in (0,1)$$

Input the processed feature z into the decoder of the model to generate an image.

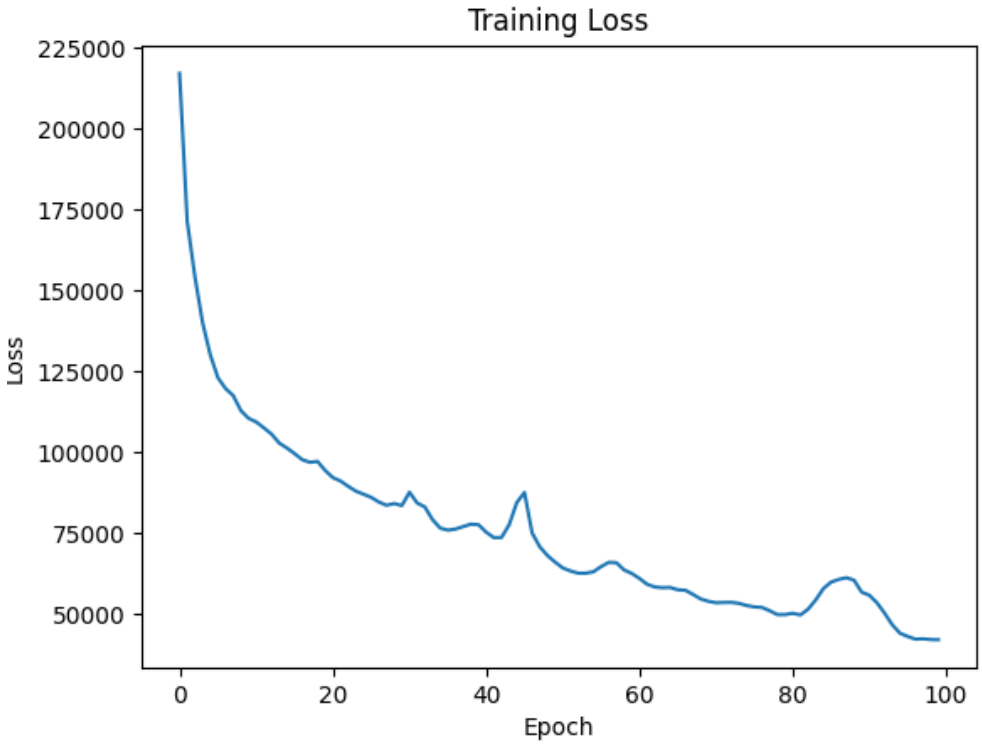
Model Training

Use the training set to train the model, and use MSE Loss as the reconstruction loss and calculate the KL divergence generated by the model as the loss function of model training.

$$Loss = MSE_{loss} + 0.5 \times KL$$

During the training of the model, I halved the KL divergence. This is because the KL divergence is too large during training and the loss drop is concentrated on too much KL divergence. This will lead to slower reduction of reconstruction loss and poorer model performance. So I think that in this model, reconstruction loss is the key direction of optimization.

The model loss curve looks like this:

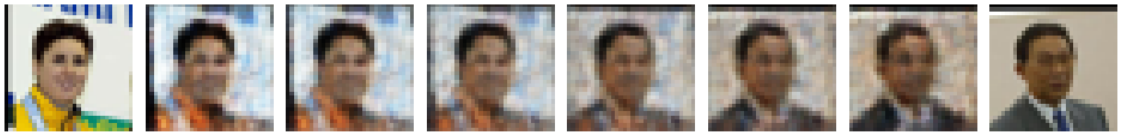


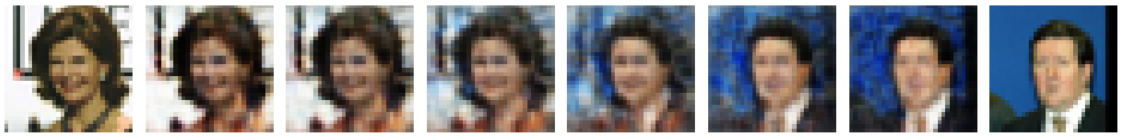
The model was trained for 80 epochs to get a better fit. The reconstruction loss of the model keeps decreasing throughout the process, and the KL divergence increases slightly.

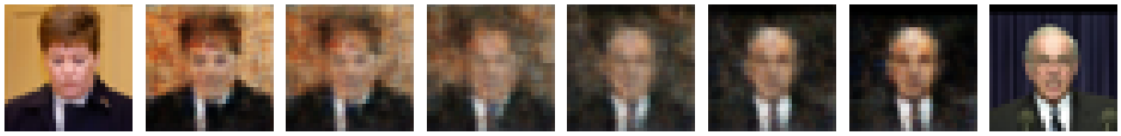
Reconstructed Images

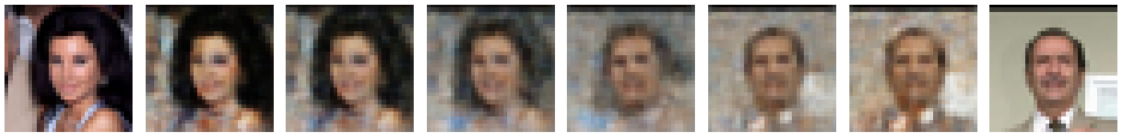
I choose men and women as two categories of pictures. The leftmost images are visualizations of the original training samples. The pictures from left to right are the images generated when the alpha is 0, 0.2, 0.4, 0.6, 0.8, 1. The leftmost image is the generated image from the input woman portrait.

I choose 7 groups of samples of men and women as the result display.









Discussion

1. From the results, it can be seen that under different conditions of α, the generated images reflect the characteristics of men and women to different degrees.
2. Since the input picture is 32X32, the picture output by the model is blurry. The features extracted by the model are not very obvious.
3. Due to factors such as race, skin color, hair color, background, etc., it will have a certain impact on the characteristics of the male and female categories.
4. It can be seen from the generated images that the fine feature extraction of the model needs to be further strengthened. It may be necessary to further enhance the complexity of the model, or increase the convolutional layer to extract more features.
5. Optimize the reconstruction loss to obtain similar results to the original image. The KL divergence is sub-optimized to obtain normal distribution differences to generate different faces.