DENOISING LATENT DIFFUSION PROBABILISTIC MODELS

Bhoumik Patidar Guntas Singh Saran Vamsi Srivathsa

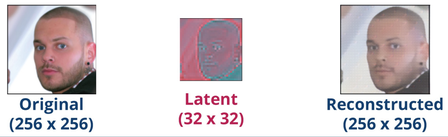
Prof. Shanmuganathan Raman

Indian Institute of Technology Gandhinagar

Computer Vision, Imaging, and Graphics (CVIG) Lab

Problem Statement

Training Diffusion Models over the pixel space is a computationally expensive task requiring days worth of compute and train time along with many GPUs involved. The idea is to train the DMs on a lower dimensional latent space and then obtaining the original image space using a trained autoencoder capable of reconstructing high quality images.



Diffusion Process

Forward Process

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbb{I})$$

Reverse Process

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \left(\frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\mathbf{x}_0\right), \frac{(1-\alpha_t)\cdot(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)}\mathbb{I})$$

Loss Function

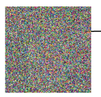
$$L_t^{\text{Simple}} = \mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_\theta((\sqrt{\bar{\alpha}_t})\mathbf{x}_0 + (\sqrt{1-\bar{\alpha}_t})\boldsymbol{\epsilon}_t, t) - \boldsymbol{\epsilon}_t\|^2\right]$$

By simply learning to **predict the noise** in the image, the model learns the parameters of the reverse distribution

$\boldsymbol{\epsilon}_t$

Denoising Diffusion

Probabilistic Model



$\boldsymbol{\epsilon}_\theta$

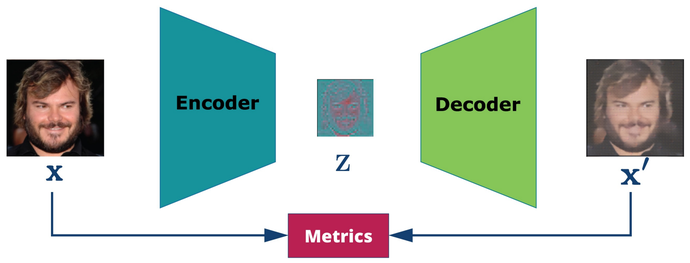
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$

Sample from the reverse distribution

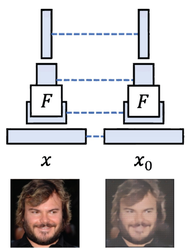
$\mathbf{x}_{t-1}$

$\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta + \sigma_t\mathbf{z} \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$

Training VQVAE as AutoEncoder

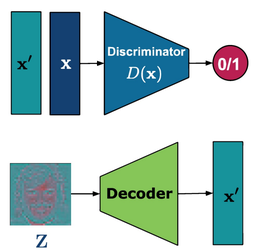


Perceptual Loss



Compare the feature maps of the original and the reconstructed from a learnt VGG16 Network

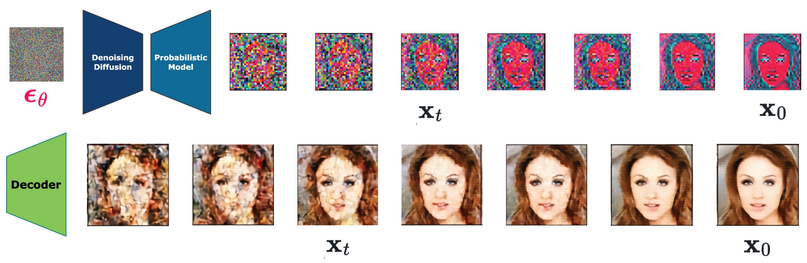
Adversarial Loss



In addition to the LPIPS Loss, also train the decoder to fool the Discriminator

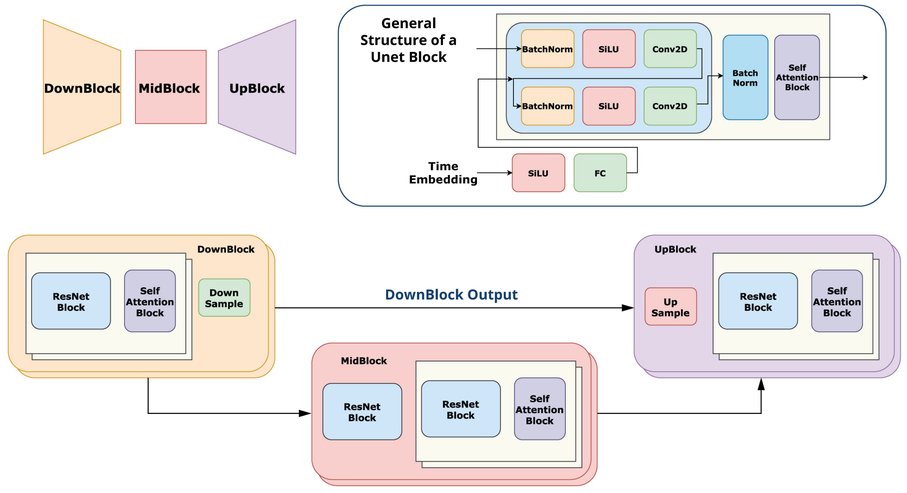
The overall loss function involves the contributions from the L2 Loss, LPIPS Loss, Adversarial Loss, VQVAE's codebook and commitment losses. For a batch size of 8 i.e. 3750 images in an epoch, the training took almost 9 hours for 20 epochs with each epoch taking 25 mins on an average

Sampling from Reverse Distribution



For a batch size of 16 i.e. 1875 images now all in the latent dimension (32 x 32) in an epoch, the training took almost 8:30 hours for 100 epochs with each epoch taking 4:30 mins on an average

UNet Architecture



Results

REFERENCES

[1] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. https://arxiv.org/abs/2006.11239

[2] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. https://arxiv.org/abs/1801.03924

[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. https://arxiv.org/abs/2112.10752

[4] van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2018). Neural Discrete Representation Learning. https://arxiv.org/abs/1711.00937

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. https://arxiv.org/abs/1505.04597