Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

November 6, 2024

.

1 Implementation Details

The pipeline given in the paper integrates three major components:

- Pre-trained Autoencoder: Encodes input images into latent space without up-sampling layers.
- Latent Diffusion Model (LDM): Operates in a latent space and learns via denoising and alignment losses. The diffusion process in the latent space enables efficient high-quality image generation.
- Local Implicit Image Function (LIIF): Decodes images from the latent space and generates arbitrary-scale outputs by leveraging MLP-based architecture that decodes latent vectors to continuous pixel coordinates.

The paper is implemented using the publicly available code. Some suitable and required changes were made in the code for it's successful execution on the local machine. To train the implicit neural decoder and diffusion model, I used Adam optimizer with learning rates of 5e-5 and 1e 6, respectively. The network is implemented in two parts the first stage model and the latent diffusion model. First the first stage model is trained using the given dataset and the obtained checkpoints are stored which are then used for training the latent diffusion model.

2 Dataset Description

For the training purpose a dataset containing diverse set of high quality images with high resolution is used. The dataset comprises of 1000 images which are further splitted into training and validation and testing sets. But due to limited computing power available I used a subset of these images.

3 Evaluation Metrics

Frechet Inception Distance (FID)

FID measures the distance between the feature distributions of real and generated images. A lower FID indicates better perceptual quality.

Precision and Recall

Precision evaluates the quality of generated samples, while Recall assesses their diversity. Higher values indicate better-quality and more diverse images.

Self-Similarity Structural Similarity Index (SelfSSIM)

SelfSSIM measures consistency across scales by comparing structural similarity between images generated at different resolutions.

Peak Signal-to-Noise Ratio (PSNR)

PSNR measures pixel-level similarity between generated and ground-truth images. Higher PSNR indicates better reconstruction accuracy but may not correlate well with perceptual quality.

Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS evaluates perceptual similarity between generated and ground-truth images. Lower LPIPS indicates better perceptual fidelity.

Frames Per Second (FPS)

FPS measures the inference speed, with higher values indicating faster processing, important for real-time applications.

4 Novelity

For improving performance we can add a lightweight transformer within the latent diffusion model. This would help in compressing high dimensional latent space representation much effectively.