Realistic Human Face Generation Using Diffusion

Daniel Lobo & Moinak Dey

Overview

- **Objective**: Adapt a pretrained Stable Diffusion model to produce high-quality face images
- **Dataset**: CelebA Dataset (celebrity faces dataset)
- **Approach**: Only refine the UNet component while keeping VAE and text encoder fixed.

Stable Diffusion Architecture (High-Level)

- **VAE (AutoencoderKL):** Encodes and decodes images to/from a latent space.
- **CLIP Text Encoder:** Transforms text prompts into latent text embeddings.
- **UNet Diffusion Model:** Iteratively denoises latent representations to generate images.

Fine-Tuning Process

What we did:

- Loaded pretrained Stable Diffusion (VAE, text encoder, UNet).
- **Froze VAE & Text Encoder:** No updates to these weights.
- **Trained Only the UNet:** Adjusted its weights to better represent faces from CelebA.

Why we did this:

- Speeds up training
- Leverages strong prior knowledge from the pretrained model.
- Focuses learning on domain adaptation (faces).

Training Details

Configuration:

- **Dataset:** CelebA
- **Prompt:** "A high-resolution photo of a face"
- **Batch Size & Epochs:** Tuned for available GPU (NVIDIA A6000) memory (16 batch size, 5 epochs)
- **Learning Rate:** 5e-6 for stable fine-tuning
- **Checkpointing:** Saved model every 500 steps

Outcome:

- Final model specialized in generating face images aligned with the target domain.

The Key Lines of Fine Tuning

- *optimizer.step()*
  - Updates the UNet's parameters using the gradients just computed. This is where the model's weights are actually adjusted, fine-tuning the UNet to better represent the training data.
- *torch.nn.utils.clip_grad_norm_(unet.parameters(), config.max_grad_norm)*
  - Limits the magnitude of the gradients to prevent overly large updates, helping maintain stable and controlled training. If gradients exceed **config.max_grad_norm**, they are scaled down.
- *loss.backward()*
  - Computes gradients of the loss with respect to the UNet's parameters. This tells us how the model should adjust each weight to reduce the loss.

Results & FID Score

Generated Images:

- Model produces realistic faces matching the prompt.
- Some images appear slightly tilted or cropped.

Quantitative Metric (FID):

- FID (Fréchet Inception Distance) ≈ 59.6
- Indicates improvements possible in alignment and framing.

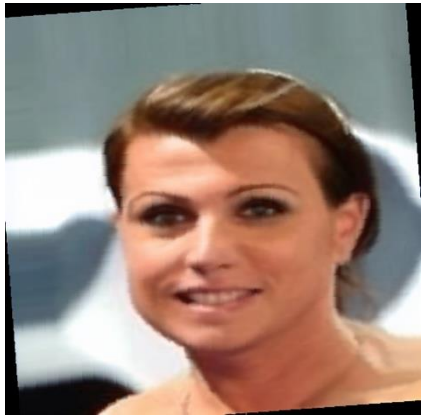
Interpretation:

- Model captures facial features but not perfectly aligned distribution of CelebA.
- Further refinement needed for lower FID.

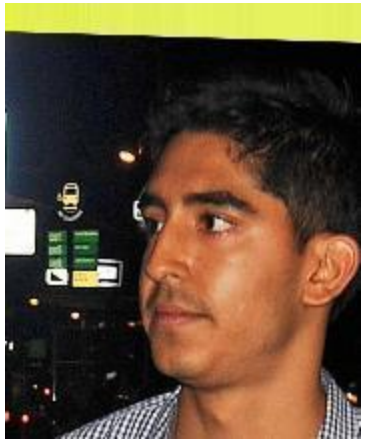
Examples of Data

Generated



Real

Conclusion & Future Work

Conclusions:

- Fine-tuning a pretrained diffusion model can adapt it to a new image domain with minimal overhead.
- Achieved a working face-generation model, though quality and alignment can improve.

Future Work:

- Refine prompts for better image framing.
- Use face alignment or data augmentation during training.
- Experiment with different guidance scales, hyperparameters, or extended training.