Stable Diffusion

The new state of art in Text2Img and Img2img

UNIVERSITAT DE BARCELONA

Ramon Mateo Navarro
Vislang @ HuPBA
13/09/2022

Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development but have a greatest computational demands.

Diffusion models which are build from hierarchy of denoising autoencoders have shown to achieve impressive results in image synthesis.

Training the most powerful DMs often takes hundreds of GPU days (e.g 150-1000 V100 days). Example. Producing 50k samples takes approximately 5 days on a single A100 GPU.

Introduction

This has two consequences:

1. Requires a massive computational cost.
2. Evaluating an already trained model is expensive in time and memory.

Introduction

So what they do for Stable diffusion?

- Separate training in two distinct phases.
    - Training first an autoencoder that provides a lower dimensional representational space which is perpetually equivalent to the data space.
    - Training DM in the learned space.

Doing this have advantage that's only training an universal autoencoding stage only once can be reused for a multiple DM training to explore possible completely different task.

Method

They observe that although diffusion models allow to ignore perceptually irrelevant details by undersampling the corresponding loss terms, they still require costly function evaluations in pixel space, which causes huge demands in computation time and energy resources.

They propose to circumvent this drawback by introducing an explicit separation of the compressive from the generative learning phase

To achieve this, they utilize an autoencoding model which learns a space that is perceptually equivalent to the image space, but offers significantly reduced computational complexity.

Method

With this they obtain DMs which are computationally much more efficient because sampling is performed on a low-dimensional space.

They exploit the inductive bias of DMs inherited from their UNet architecture, which makes them particularly effective for data with spatial structure and therefore alleviates the need for aggressive, quality-reducing compression levels as required by previous approaches.

Finally, they obtain general-purpose compression models whose latent space can be used to train multiple generative models

Perceptual image compression

Perceptual compression consists of an autoencoder trained by combination of perceptual loss and a patch-based adversarial objective.

The autoencoder used in Stable Diffusion has a reduction factor of 8. This means that an image of shape (3, 512, 512) becomes (3, 64, 64) in latent space, which requires $8 \times 8 = 64$ times less memory.

Perceptual loss functions are used when comparing two different images that look similar, like the same photo but shifted by one pixel. The function is used to compare high level differences, like content and style discrepancies, between images.

Conditioning Mechanisms

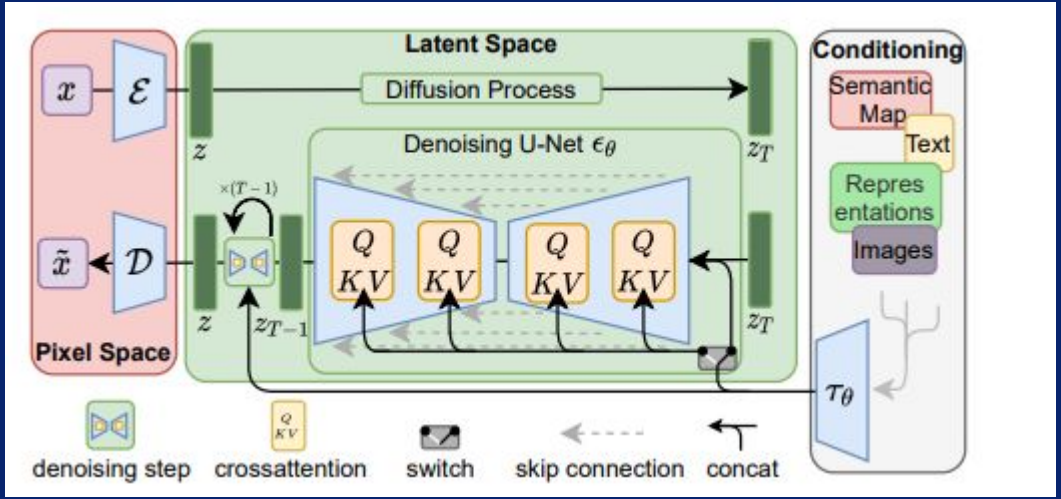
Diffusion models are capable of modeling conditional distributions of the form p(z|y). This is implemented using a conditional denoising autoencoder.

They turn DM into more flexible conditional image generators by augmenting their underlying UNet backbone with cross-attention mechanism.

For preprocess y from different domains they introduce a specific encoder τ_θ that projects y to an intermediate representation , which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing Attention

$$(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$$

Architecture



Demos

Text2Img : Colab Text2Img

Img2Img: Colab Img2Img