Diffusion Model 요약

1. Diffusion Model의 기초
2. Diffusion Model이 할 수 있는 것

Diffusion Model: 이미지 생성모델의 일종(ex. GAN)

vs. GAN -> 둘은 다르지만 Diffusion 모델의 이미지 생성 퀄리티가 전혀 떨어지지 않음

Text를 입력 받아 이에 해당하는 이미지를 생성- 생성된 사진이 DALLE2가 학습한 데이터셋에는 없다는 점이 놀라움

위치를 제한해서 요구조건에 맞는 고화질 이미지를 생성 가능

1. Diffusion Model 작동 원리

Gaussian Distribution을 따르는 분자의 움직임을 매시간마다 계산 -> 역으로 Diffusion process의 과정을 거꾸로 되돌려서 분자의 움직임의 과정을 반대로 볼 수 있음

먼저 분자나 픽셀에 노이즈를 추가하다 보면 점점 완전한 노이즈의 이미지를 얻을 수 있다.

스크린샷, 텍스트, 빛, 디자인이(가) 표시된 사진

자동 생성된 설명

역으로 과정을 되돌려서 분자의 움직임을 반대로 계산할 수 있으면

스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명

이와 같이 첨가된 노이즈의 값을 계산하여 원래의 이미지로 복원이 가능하고 이는 이미지를 생성하는 것과 같다.

1. Architecture

Diffusion모델은 Input과 Output의 Resolution이 동일하게 나오기 위해서 UNet구조를 가지고 있다. 과정을 처리하기 위해서 받는 값으로는 process의 과정을 알게 해주는 t값과 클래스의 정보나 text 정보인 condition이 있다.

1. Loss Function

Diffusion 모델이 각 픽셀의 첨가된 noise를 예측할 때 실제 값과의 차이가 거의 나지 않도록 Diffusion Model을 학습한다. 아래는 이와 관련된 식이다.

텍스트, 폰트, 친필, 서예이(가) 표시된 사진

자동 생성된 설명

1. Diffusion Model의 응용
2. Class Guided Image Generation

: ImageNet의 여러 클래스 중 Diffusion Model에게 내가 원하는 class의 이미지만을 생성하게 함

1. Classifier Guidance

Classifier이 noisy image를 먼저 학습하여 gradient를 구하고 이를 Diffusion model이 이미지를 생성 하려할 때 원하던 이미지가 아니면 큰 gradient값을 보내어 Guidance의 역할을 하여 원하는 이미지만 생성함.

1. Classifier free Guidance

Guidance없이 class에 해당하는 condition정보를 Diffusion Model에서 t를 Positional Encoding 방식을 통해 Condition으로 입력하는 방식처럼 Class 정보 또한 Condition으로 Diffusion Model에 입력함. 이때 condition을 받을 때와 안 받을 때로 Noise를 예측하고 최종 noise를 interpolation으로 내고 실제와 같도록 하면 guidance없이 원하는 이미지를 생성하는 게 가능하다.

1. Super-Resolution

: 저화질 이미지를 고화질 이미지로 변환

입력: 기존의 High Resolution 이미지와 Low Resolution 이미지를 Concat

(High Resolution 이미지를 먼저 작은 사이즈로 Resize -> 작은 이미지를 다시 원래 사이즈로 Resize -> 이렇게 만들어진 이미지는 작은 사이즈 이미지로부터 확대한 이미지 이므로 Low Resolution 이미지가 됨)

-> Diffusion 모델이 이를 학습

학습 후: Low resolution 이미지만을 입력으로 받아도 High Resolution의 이미지 생성가능

1. Inpainting

: 이미지의 가려진 부분을 그럴듯하게 채워 넣는 기술

입력: Mask 이미지

Condition: Mask 정보 -> Diffusion 모델이 Mask 처리된 부분의 원래 모습을 예측하도록 학습

출력: 학습이 완료된 모델에게 Mask 처리된 다른 이미지와 Mask 정보를 주면 그럴듯하게 Mask 부분을 채워서 출력

1. Text Guided Image Generation
2. Glide

- Text Embedding을 학습하는 Transformer-> 이미지 생성을 위한 Diffusion Model-> Super-Resolution을 위한 Diffusion Model로 구성

1. Imagen

-Caption의 Text Embedding을 추출-> Diffusion Model이 이미지생성 ->Super Resolution까지 직렬로 연결

-Glide와의 차이: Pretrained text encoder, 더 고화질의 super-resolution을 수행

1. DALLE2

Clip 활용

맨 위: pretrained Clip로 Text와 Image의 Embedding을 추출

아래: Prior와 Decoder로 표현된 Diffusion모델에서 각각 Image Embedding과 이미지 생성

특징: 이미지 생성 뿐만 아니라 이미지 조작방법도 제안

1. Multi Modal Guided Image Generation

: 위의 한계는 condition으로 text만을 입력 받을 수 있음, 저화질의 이미지만을 생성하여 SR모델을 따로 붙여줘야 함

Stable Diffusion은 Perceptual 한 부분에 초점을 맞춰 학습하는 Diffusion Model을 제안

Non-perceptual한 부분에 집중하지 않기 위해 Auto encoder을 사용:  Auto Encoder로부터 압축된 Latent Embedding을 예측하는 방법

Image - Auto encoder -> Diffusion Model -> Decoder- Noise

\*Auto Encoder를 학습과정 = Perceptual Image Compression: Auto Encoder로 Diffusion Model은 Auto Encoder로부터 출력되는 Latent Embedding을 생성하도록 학습함.

\*Latent Diffusion Model: Auto Encoder로 정보를 압축-> Non-perceptual한 부분인 Noise성 정보들은 제거 -> Semantic한 부분에 집중 + 계산 복잡도는 줄어듦

\*Condition: Stable은 다양한 Condition을 입력 가능 but 이중 한 가지만 입력 Condition은 각자에 맞는 적절한 Encoder 𝜏가 필요, 이후 입력된 condition은 연산중인 image와 cross attention(이미지와 Condition의 상관 관계를 고려하여 Condition 정보에 가중치를 반영)으로 정보에 반영됨.

1. Multi Modal Input Image Generation

: Stable Diffusion에도 한계가 존재 다양한 형태의 Condition을 입력으로 받을 수 있지만, 하나만 선택해야 함 + 엄청난 computing resource 필요

GLIGEN

: bounding box class, caption등 여러 입력 받는 거 가능

Stable Diffusion이 그대로(image generation기능 그대로 활용), 이 외에 Grounding 정보를 연산하기 위한 Gated Self Attention 부분만 추가

Grounding:

Class 정보는 Pretrained Text Encoder를 사용하여 Text Embedding으로 변환

Bounding Box의 위치 정보는 Fourier Transform을 사용해 변환  
두 정보를 결합해서 MLP를 통해 GLIGEN이 연산할 수 있는 형태로 변환

Gated Self Attention:

Grounding 정보를 기존 정보들과 결합하여 연산

Grounding 과 Visual 정보 간의 Self Attention을 수행 후 Visual 정보만 추출

이렇게 나온 Visual 정보에는 Scale을 조절하기 위해 𝛾를 곱해주고요. 다시 기존 Visual 정보에 더해줌.