Paper summary

AI VISION Lab

1. 공부한 논문의 제목, 게재된 학회 혹은 저널 등 논문 기본 정보를 적으세요.
   - **A. 이름: High-Resolution Image Synthesis with Latent Diffusion Models**
   - **B. 저널: CVPR**
   - **C. 도메인: Image Generation**
   - **D. 출판연도: 2022**
   - **E. 저자: Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser Bjorn Ommer**

2. 논문에서 제안한 알고리즘 및 프레임워크에 대해 본인이 이해한대로 다이어그램을 그려보세요. **논문 Figure를 그대로 따라 그리면 안됩니다.**
   A. 모델이 학습하고자 하는 공간은 Input image 자체가 아닌, Latent Space공간의 정보이므로 Downsampling 과정 없이 학습을 하다보니 엄청난 Computation 자원이 필요한 기존의 DM(Diffusion Model)의 구조대신 Latent Space의 Z를 사용하는 것이 주된 아이디어임.

   B. Encoding과정은 AutoEncoder 구조를 가져와서 사용함. 이때 과도한 Variance를 예방하고자 Regularization을 주입 (->일반화하지 못하는 혹은 불안정한 구조를 유발하는 것을 방지하려 함.)

   C. Auto Encoder를 통해 구한 Representation Z를 토대로, Forward – Reverse 과정을 진행. 이후 Decoder 과정을 통해 output을 생성하는 구조.

   D. 기존의 Auto Encoder과 가장 큰 차이점은 1D 이 아닌, 2D구조의 Z를 사용한다는 점인데, 논문에서 이야기하는 바로는 2D Z가 정보 유지 및 이후 생성에 더 많은 도움이 된다고 함.

   E. 또한, Z의 크기를 Down Sample 하는 Factor의 크기로 LDM의 종류를 구분하고 있는데 -> 가장 효과적인 LDM의 경우 4-8임을 실험적으로 보였음.

   F. 뿐만 아니라 기존의 Sampling과정에 Conditioning해주던 기존 모델들과 달리(Classifier, classifier-Free) Cross Attention block을 추가해, Condition을 직

접 주입해주는 구조를 선보임.

G. 다만, SR Task에서 SR3와 동일하게 Bicubic interpolation을 Condition으로 주입하는 경우 모델의 성능이 잘 나오지 않는 문제가 있어서, 해당 Degradation뿐만 아니라 추가적인 degradation을 주입해야 하는 한계도 있었음.

3. 본인이 생각하는 이 논문의 장점이 무엇이라고 생각하나요? **논문 Contribution bullet을 그대로 따라 적으면 안됩니다.**
   A. Input을 Noisy하게 만드는 것이 아닌, Z를 Noisy하게 만듦으로써 Computation 성능을 상당히 높였다는 점에서 가장 큰 의의가 있다고 생각됨.
   B. 또한 Condition을 Embedding -> Cross attention을 토대로 다루다 보니, 다양한 Task를 접목시킬 수 있는 Framework의 등장도 큰 기여 점

4. 이 논문을 읽으면서 느낀 점, 혹은 배운 점이 있으면 적어보세요.
   A. Cross Attention을 적용해 Conditioning하는 과정의 시작점임을 알 수 있었음.

   B. 해당과정에서 사용된 Encoder의 경우 -> 다른 논문에서 Feature extraction 용으로 사용하고 있으며 앞으로의 연구에서도 사용해보면 어떨까 생각됨

5. 이 논문의 한계점이 있다면 무엇이라고 생각하나요?
   A. Detail이 필요한 Generation Task(Pixel generation같은)에 대해서는 아직 성능이 미세하게 낮다는 한계가 있음 (Z를 사용하기 때문에 그렇다고 판단됨.)

6. 본인의 연구에 접목시켜볼 점이 있을지 생각하고 적어보세요.
   A. Z를 직접 Diffusion을 적용하는 방법은 현재 읽고 있는 논문인 Denoising Representative의 구조와 닮았다고 할 수 있음(해당 논문에서도 Encoding -> denoising 구조를 사용.)

   B. 본 논문과 읽고 있는 논문을 연결시킨다면 좋은 연구가 될 수 있지 않을까 생각됨.

7. 본 Summary를 작성하는 과정에서 생성형AI를 사용했나요?

A. 아니요

날짜: 2025-08-13

이름: 신준원

High-resolution Image synthesis with latent diffusion models

Introduction

→ 대규모의 diffusion models.

기존 Diffusion based (SR) ⇒ 연산량 + 시간 "많다"
(not-DM SR)

목표 ⇒ 성능은 유지하면서 Computational 자원 감소.

방식 → trained diffusion model의 pixel space 분석

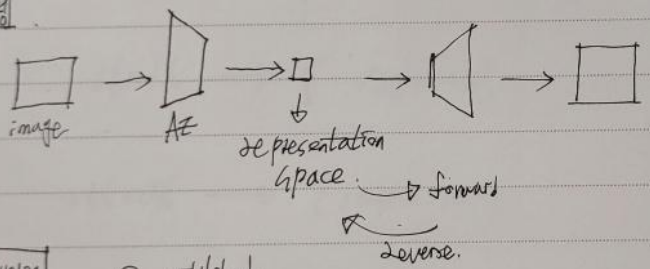
likelihood based 모델의 학습단계

과정 ① Detail → x, Semantic variation / (Latent space, Robust representation 추출)
(압축)

과정 ② 압축·추출된 데이터 기반 학습 진행

→ model에게 효율적인 Latent Space!

과정



최적화점

① 연산량 ↓

☆ ② 학습된 AE의 재활용

다양한 task를 위해 재사용 가능 → ☆ Scale 감소X

Perceptual Image Compression

AE(E) (trained by perceptual Loss & patch-based adversarial objective)

[$d \in H, W, 3$ (RGB)

$AE(x) = z \rightarrow D(z) = \tilde{x}$.

└ d dim (압축 ⇒ 1D)

⊕ variance가 큰 latent space 피하기 (General한 space 목표)

⇒ regularization 추가 (VAE 방식)

① KL-reg (KL-Penalty)

⇒ learned Latent to standard normal (추천)

② VQ-reg

⇒ Vector quantization layer 추가 (in decoder)

Diffusion part

[Loss → DDPM based. ⇒ $L_{LDM} = E_{\mathcal{E}(x), \epsilon \sim N(0,1), t}\left[\| \epsilon - \epsilon_\theta(z_t, t) \|_2^2\right]$

강의: ∿D

⊕ Condition 추가 ⇒ $\epsilon_\theta(z_t, t, y)$

(조건) ↓ └ Condition (text, Semantic map)

↓ 방법 활용의 Condition 추가 방법 ⇒ "Cross-attention mechanism"

Condition into key & value.

⇒ $L_{LDM} = E_{\mathcal{E}(x), y, \epsilon \sim N(0,1)}\left[\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)) \|_2^2\right]$

Experiments

Compression trade off

→ (Compression) down Sampling Factor.

- ⇒ H, W (original - LDM-1)
- ⇒ H/8, W/8 (LDM-8) ↓ → 성능 저하. (Best → 4 & 8)

Conditional Latent Diffusion

① text-to-Image

text ⇒ Bert-tokenizer → Cross-attention.
Image → encode

② Image-to-Image

→ Semantic Synthesis, Super-resolution, inpainting

※ SR3 방법 ⇒ 성능 (SR3 4)

LDM ← SR3 process

⇒ bicubic 줄여서 → Camera/blur 추가 degradation ⇒ Condition 적용 ⇒ 성능 ↑.