Week2_예습과제_김도희

High-Resolution Image Synthesis with Latent Diffusion Models

Abstract

Paper’s latent diffusion models(LDMs) achieve new state-of-the-arts scores for image inpaintings and class-conditonal image synthesis and highly competitive performance on various tasks, including text-to-image synthesis, unconditional image generation and super resoluion, while significantly reducing computational requriements compared to pixel-based DMs

⇒ 기존의 픽셀 기반 확산 모델의 계산 비용과 성능의 한계를 극복하기 위해 **잠재 공간**에서 확산 모델을 학습하는 새로운 접근 방식을 제안. 이를 통해 **효율성**을 높이면서도 **세부 묘사**를 유지하고, 다양한 조건부 생성 작업에서도 유연하게 적용할 수 있는 **고해상도 이미지 합성** 모델

1. Introduction

이전

scaling up likelihood-based-models (ex. autoregressive transformers) → billions of parameters

GAN → limited variability

최근

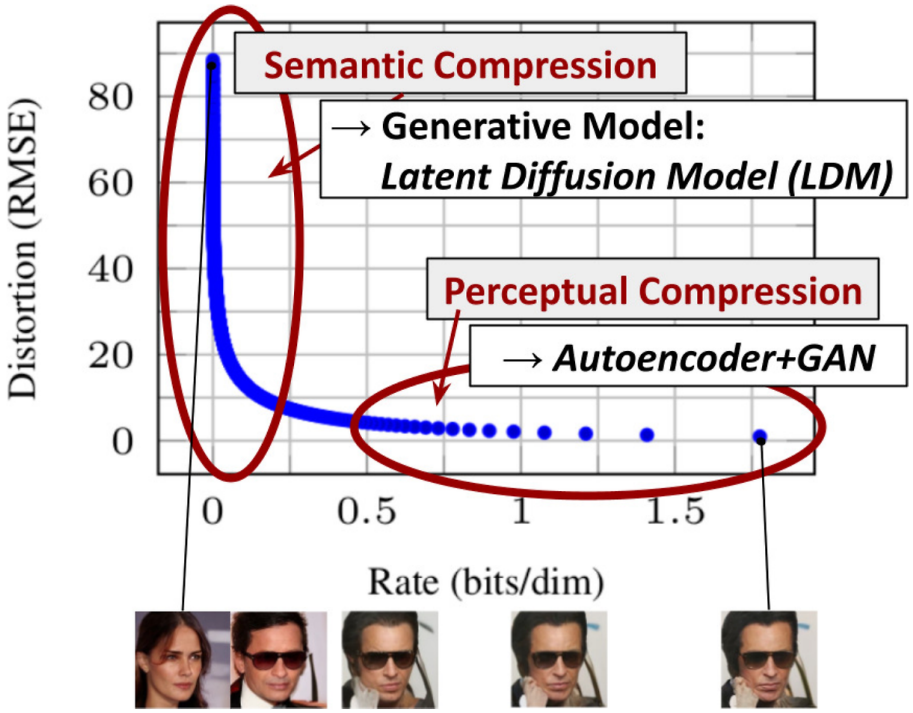
Denoising Autoencoder로 구성된 Diffusion models이 class-conditional image synthesis와 super-resolution에서 SOTA 달성. Likelihood-based 모델을 선택함에 따라

DM은 GAN의 비슷한 data만 계속 생성하는 문제(model-collapse)나 학습의 불안정성을 겪지 않았고 많은 파라미터가 없어도 parameter sharing을 통해 이미지의 복잡한 분포를 모델링할 수 있다.

Democratizing High-Resolution Image Synthesis

: DM의 성능을 유지하면서 computing demand를 줄여야한다.

Departure to Latent Space



⇒ Latent Diffusion model의 Rate(bits/dim)가 가장 낮은 것을 알 수 있는데, 이는 더 적은 비트 수로 더 많은 dim을 포함할 수 있다는 것이다. 위의 그림은 rate-distortion trade-off를 보여준다.

학습은 크게 아래의 두가지로 구분할 수 있다.

- Perceptual Compression: 의미는 거의 학습하지 않는 단계

- Sementic Compression: 의미와 개념적인 구성을 학습하는 단계

Goal : perceptual하게 동등하지만 계산적으로 더 적합한 space를 찾는 것

Phase 1

Date space와 perceptual하게 동일한 저차원 representational space로 보내는 autoencoder를 학습한다.

Phase2

이를 통해 complexity가 감소되고 네트워크를 한번만 통과하고 latent space에서 효율적인 이미지 생성이 가능하다. (LDMs)

⇒ Pros : universal autoencoding 단계를 한 번만 학습하면 되므로 여러 diffusion model 학습에 재사용하거나 완전히 다른 task를 탐색할 수 있다. 이를 이용하여, image-to-image 및 text-to-image task를 위한 여러 diffusion model을 효율적으로 탐색할 수 있다

Contributions

- 고차원 데이터에 대한 효율적인 확장 및 재구성
- 성능과 계산 비용
- 인코더/디코더 아키텍처와 사전 학습 과정의 단순화
- Convolution method를 이용한 밀집 조건부(super-resolution, inpainting)에 대한 적용성
- 멀티모달 훈련을 위한 범용 조건부 메커니즘
- 사전 학습된 모델의 공개

2. Related Work

Generative Models for Image Synthesis

- GANs
- VAE와 Flow-based Model
- Autoregressive Model

Diffusion Probabilistic Models(DM)

최고의 밀도 추정과 샘플 품질을 보여주고 있지만, 픽셀 기반 모델의 특성상 추론 속도가 느리고 훈련 비용이 높다. LDM은 latent space에서 학습하고, 이를 통해 계산 비용을 줄이면서도 품질을 유지할 수 있는 모델이다. U-Net 구조를 기반으로 하여 고해상도 이미지의 세밀한 정보를 보존하면서도, 압축된 잠재 공간에서 효율적인 학습이 가능하게 설계되었다.

Two-stage Image Synthesis

이미지 생성 성능을 향상시키기 위해 서로 다른 방법을 혼합하는 방법

3. Method

> We propose to cicumvent this drawback **by introducing an explicit separation of the compressive** from the generative learning phase. To achieve this, we utilize **an autoencoding mode**l which learns a space that is perceptually equivalent to the image space, but offers significantly reduced computational complexity.

Advantages

- 저차원 공간에서 샘플링이 수행되어 계산이 효율적
- UNet 아키텍처에서 상속된 DM의 inductive bias를 활용하여 공간 구조가 있는 데이터에 유리
- Latent space가 여러 생성 모델을 학습시키는 데 사용 가능

3.1. Perceptual Image Compression

perceptual loss+ patch-based adversarial objective으로 학습되는 Autoencoder로 구성

→ local realism이 적용되어 재구성이 image manifold에 국한되고 L2 또는 L1 objective와 같은 pixel space loss에만 의존하여 발생하는 흐릿함을 방지

RGB Image : $x \in \mathbb{R}^{H \times W \times 3}$

Encoder : $\mathcal{E}$

Latent Representation : $z = \mathcal{E}(x), \quad z \in \mathbb{R}^{h \times w \times c}$

Decoder: $\mathcal{D}$

Data Reconstruction : $\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$

- KL-reg

  : A small KL penalty towards a standard normal distribution over the learned latent, similar to VAE

- VQ-reg

  : Uses a vector quantization layer within the decoder, like VQVAE but the quantization layer is absorbed by the decoder.

3.2. Latent Diffusion Models

Diffusion Models

: 정규 분포 변수의 noise를 점진적으로 제거하여 data distribution $p(x)$를 학습하기 위해 디자인 되었다. 길이가 T인 고정된 Markov chain의 reverse process를 학습하는 것이다. 아래는 단순화된 Objective이다.

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{1}$$

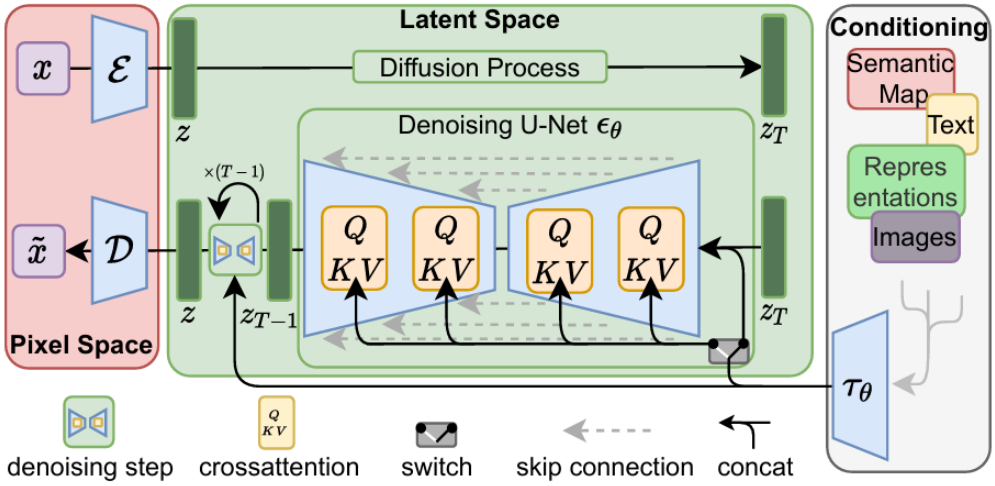
Generative Modeling of Latent Representations

$\mathcal{E}$와 $\mathcal{D}$로 구성된 학습된 perceptual compression 모델을 사용하여 high-frequency의 감지할 수 없는 detail이 추상화되는 효율적이고 낮은 차원의 latent space에 접근한다. 이 공간은 likelihood 기반 생성 모델에 더 적합하다. 또한 이미지별 inductive bias를 활용할 수 있다.

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] \tag{2}$$

backbone $\epsilon_\theta(\cdot, t)$는 time-conditional UNet으로 구현된다. Forward process가 고정되어 있으므로 학습 중에 z_t를 $\mathcal{E}$에서 효율적으로 얻을 수 있고 $p(z)$의 샘플을 $\mathcal{D}$에 한 번 통과시켜 이미지 space로 쉽게 디코딩할 수 있다.

3.3. Conditioning Mechanisms



$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V \tag{3}$$

$$Q = W_Q^{(i)} \cdot \phi_i(z_t), \quad K = W_K^{(i)} \cdot \tau_\theta(y), \quad V = W_V^{(i)} \cdot \tau_\epsilon(y)$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2 \right] \tag{4}$$

⇒ cross attention mechanism

: Query와 Key와의 비교를 통해, Query가 가져야하는 값을 찾는 과정

4. Experiments

4.1. On Perceptual Compression Tradeoffs

f	$\lvert\mathcal{Z}\rvert$	c	R-FID ↓	R-IS ↑	PSNR ↑	PSIM ↓	SSIM ↑
16 *VQGAN* [23]	16384	256	4.98	–	19.9 ±3.4	1.83 ±0.42	0.51 ±0.18
16 *VQGAN* [23]	1024	256	7.94	–	19.4 ±3.3	1.98 ±0.43	0.50 ±0.18
8 *DALL-E* [66]	8192	-	32.01	–	22.8 ±2.1	1.95 ±0.51	0.73 ±0.13
32	16384	16	31.83	40.40 ±1.07	17.45 ±2.90	2.58 ±0.48	0.41 ±0.18
16	16384	8	5.15	144.55 ±3.74	20.83 ±3.61	1.73 ±0.43	0.54 ±0.18
8	16384	4	1.14	201.92 ±3.97	23.07 ±3.99	1.17 ±0.36	0.65 ±0.16
8	256	4	1.49	194.20 ±3.87	22.35 ±3.81	1.26 ±0.37	0.62 ±0.16
4	8192	3	0.58	224.78 ±5.35	27.43 ±4.26	0.53 ±0.21	0.82 ±0.10
4†	8192	3	1.06	221.94 ±4.58	25.21 ±4.17	0.72 ±0.26	0.76 ±0.12
4	256	3	0.47	223.81 ±4.58	26.43 ±4.22	0.62 ±0.24	0.80 ±0.11
2	2048	2	0.16	232.75 ±5.09	30.85 ±4.12	0.27 ±0.12	0.91 ±0.05
2	64	2	0.40	226.62 ±4.83	29.13 ±3.46	0.38 ±0.13	0.90 ±0.05
32	KL	64	2.04	189.53 ±3.68	22.27 ±3.93	1.41 ±0.40	0.61 ±0.17
32	KL	16	7.3	132.75 ±2.71	20.38 ±3.56	1.88 ±0.45	0.53 ±0.18
16	KL	16	0.87	210.31 ±3.97	24.08 ±4.22	1.07 ±0.36	0.68 ±0.15
16	KL	8	2.63	178.68 ±4.08	21.94 ±3.92	1.49 ±0.42	0.59 ±0.17
8	KL	4	0.90	209.90 ±4.92	24.19 ±4.19	1.02 ±0.35	0.69 ±0.15
4	KL	3	0.27	227.57 ±4.89	27.53 ±4.54	0.55 ±0.24	0.82 ±0.11
2	KL	2	0.086	232.66 ±5.16	32.47 ±4.19	0.20 ±0.09	0.93 ±0.04

Table 8. Complete autoencoder zoo trained on OpenImages, evaluated on ImageNet-Val. † denotes an attention-free autoencoder.

pixel-based DM이다. *LDM-{4-16}*이 효율과 품질 간의 좋은 balance를 보였고, *LDM-4*, *LDM-8*이 high-quality 결과에 최적의 조건이었다.

4.2. Image Generation with Latent Diffusion

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49	*LDM-4* (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.59**	0.61	0.44	ProjectedGAN [76]	**1.52**	0.61	0.34
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**	*LDM-4* (ours, 200-s)	2.95	**0.66**	0.48

Table 1. Evaluation metrics for unconditional image synthesis. CelebA-HQ results reproduced from [43, 63, 100], FFHQ from [42, 43]. †: N-s refers to N sampling steps with the DDIM [84] sampler. *: trained in KL-regularized latent space. Additional results can be found in the supplementary.

On CelebA-HQ, we report a new state-of-the-art FID GAN과 같은 likelihood based model

4.3. Conditional Latent Diffusion

Figure 9. A *LDM* trained on 256^2 resolution can generalize to larger resolution (here: 512×1024) for spatially conditioned tasks such as semantic synthesis of landscape images. See Sec. 4.3.2.

4.4. Super-Resolution with Latent Diffusion

Figure 10. ImageNet 64→256 super-resolution on ImageNet-Val. *LDM-SR* has advantages at rendering realistic textures but SR3 can synthesize more coherent fine structures. See appendix for additional samples and cropouts. SR3 results from [72].

4.5. Inpainting with Latent Diffusion

Figure 11. Qualitative results on object removal with our *big, w/ ft* inpainting model. For more results, see Fig. 22.

5. Limitations & Societal Impact

Limitation

pixel에 기반한 접근법보다 계산량이 줄었지만, 순차적인 샘플링 과정은 GAN보다 느리다.

높은 정확도가 필요한 경우 LDM의 사용이 의문스럽고 세밀한 정확도가 필요한 경우 재구성 기능이 병목을 일으킬 수 있다.

Societal Impact - A double edged sword

기술에 대한 접근이 용이하게 되어 조작된 데이터를 생성 및 유포하거나 잘못된 정보 및 스팸을 유포하기가 더 쉬워진다. 데이터의 수집에 대한 우려가 발생할 수 있다.

6. Conclusion

LDM은 확산 모델의 효율성과 성능을 동시에 향상시킬 수 있는 간단하고 효과적인 방법!

- 기존의 denoising DM의 성능을 유지하면서 계산 효율 개선
- 다양한 조건부 이미지 합성 작업에서 좋은 성능을 보임

[참고]

https://velog.io/@hewas1230/StableDiffusion

https://kimjy99.github.io/논문리뷰/ldm/

https://jang-inspiration.com/latent-diffusion-model

https://pitas.tistory.com/9