High Resolution Image Synthesis with Latent Diffusion Models

Abstract

(기존 모델 operate in pixel space → 자원 소모가 큼)

! DM training 과정을 the latent space 에서 수행하는 것으로 계산에 필요한 비용을 줄인다. + pretrained autoencoder 사용

! cross-attention layer의 사용으로 텍스트, 바운딩 박스 등을 인풋으로 받는 generater 제작 가능

Introduce

Img synthesis - greatest computational demands

natural scene, AR transformer(likelihood model 기반) 주도

GANs, limited variability

Diffusion model

- img synthesis, class-conditional image synthesis, super resolution, inpainting, colorization, stroke-based synthesis
- (a hierarchy of denoising autoencoders) → mode-collapse & training instabilities 없음
- (heavily exploiting parameter sharing) → 수많은 파라미터 없이 complex distribution으로 모델링 가능

Democratizing High-Resolution Image Synthesis

: DM 이 소모하는 용량/자원이 과도하다.

학습에서도, RGB space에서의 repeated function evaluation 도.

→ 이런 대량의 자원을 소모하는 모델 학습에는 소수만 접근이 가능하고, 환경에도 영향을 준다(carbon footpring)

→ 평가 과정에서도 대량의 자원이 소모된다

이런 powerful model 에 대한 접근성을 높이기 위해서, 소모 자원을 줄이기 위해서 : computational complexity 줄이기

(DM의 computational complexity 줄이기)

Departure to Latent Space

pre-trained diffusion model in pixel space 분석

likelihood based model의 learning (2 steps)

1. perceptual compression : high-frequency 한 부분 삭제
2. semantic compression : 모델이 데이터의 개념적 구성을 학습

⇒ We thus aim to first find a perceptually equivalent, but computationally more suitable space. 인식적으로는 동등하지만, 계산에 있어서 더 적합한 공간을 찾을 것

<이 연구에서의(LDMs) 학습>

적합한 저차원 공간을 생성할 autoencoder 학습 : excessive spatial compression X, reduced complecity, 다른 DMs 학습에서 재사용 가능한 autoencoder

<contributions 요약>

1. transformer-based와 달리 고차원에 잘 적용됨, 세심한 복원이 가능하며 고해상도의 megapixel 이미지 합성에도 적용 가능
2. 다양한 문제에서 높은 성능 (px based와 비교)
3. 재구성 및 생성 능력 동시 측정할 필요 X, regularization of the latent space가 매우 적다
4. super-resolution, inpainting, semantic synthesis task 에서 큰 이미지를 보여줄 수 있다
5. general-purpose conditioning → multi-modal training가능
6. pretrained latent diffusion & autoencoder 제공

Related Work

- Generative models for image synthesis(GANs) : 효율적으로 고해상도의 이미지 생성이 가능 / 어려운 최적화, full data distribution을 잡기 어려움
- Likelihood-based models, VAE, flow-based, ARM : 해상도, 용량, 계산 비용 등에서 문제 또한 존재

⇒ 문제를 해결하기 위해 raw px 대신 a compressed latent image space 사용

Diffusion Probabilistic models(DM)

- density estimation, sample quality에서 좋은 성능
- UNet 사용 : fit to inductive biases of image-like data
- reweighted objective가 학습에 사용되었을 때 , 가장 좋은 합성 품질을 보임

하지만, 높은 계산 비용과 느린 추론 속도 ⇒ LDMs 으로 결점 보완

Two-Spate Image Synthesis

개별 모델의 단점 보완을 위해 combing

- VQ-VAE : autoregressive model, discretized latent space에서 사용
- VQGANs : adversarial & perceptual objective 사용으로 큰 이미지 처리

수많은 파라미터 - 성능 제한

⇒ LDM으로 문제 해결. 고차원으로 더 자연스럽게 확장이 가능하고, the level of compression 선택이 가능

Method

: To lower the computational demands of training diffusion models towards high-resolution image synthesis

- DM은 인식에 중요하지 않은 정보 무시 가능 - 그래도 많은 계산 비용 요구

→ compressive와 generative를 명확히 구분 : autoencoder 사용으로 perceptually equivalent & reduced complexity 달성

<이러한(구분) 접근의 장점>

1. 저차원으로의 샘플링 → 계산 효율
2. Unet 을 계승한 DM : 공간 구조의 데이터에 효과적, 퀄리티 저하 조절하여 압축 레벨 선정?
3. general-purpose compression model → other downstream applications

(아래 GPT의 설명 - 다시 읽어보기)

Perceptual Image Compression

: autoencoder - compression, reconstruction

- autoencoder : perceptual loss & patch_based adversarial objective로 학습 → L1, L2 기반의 모델에서 생기는 bluriness 완화
- input → encoder 압축(the latent space로) → decoder reconstruction : latent space가 인풋의 크기를 줄이는 동시에 중요 정보 유지
- Downsampling 비율 실험(factor m)
- 정규화 - avoid arbitrary high-variance latent space, 두 가지 방식의 정규화 사용
  - KL-reg : VAE와 비슷, standard normal on the learned latent
  - VQ-reg : 디코더의 내부 vector quantization layer 사용, VQGAN과 유사
- 모델이 2차원 구조의 latent space 유지하여 압축 → 고유 주고를 최대한 활용한 재구성(더 세밀한 복원 과정)

Latent Diffusion Models

Diffusion Models

- probabilistic model : denoising 을 수행해나가며 데이터 분산 학습 - Markov Chain의 resverse process. 가장 최적 모델은 변분 하한을 사용해 노이즈 제거를 최적화

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

Generative Modeling of Latent Representations

- attention-based transformer model, discrete latent space
- Unet 기반 구조 : 2d conv layer 사용 - 이미지 공간적 구조 활용
- 변형된 변분 하

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

Conditioning Mechanisms

- cross-attention → Unet으로 연결
- 도메인별 인코더
- Transformer 기반의 인코더 → text-to-image

Experiments

유연 & 계산적 효율 - LDM, image synthesis of various image modalities

- pixel-based DM과 비교
- VQ-reg 사용했을 때 더 나은 sample quality 가지는 경우 - first stage regularization 비교
- 실험의 구조, 구현, 학습, 평가

On Perceptual Compression Tradeoffs

- downsampling factor(f)에 따라 어떤 성능을 보이는지 (f=1이면 픽셀 기반, f 높아질수록 해상도는 낮아짐)

→ f 작으면 속도는 느려지고 품질 높음, f 크면 정보 손실(퀄리티 저하)

→ f=4, 8

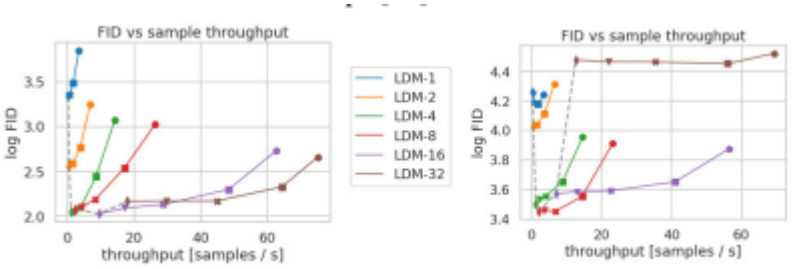

Image Generation with Latent Diffusion

- unconditional model을 다양한 데이터셋에 적용
- precision & recall 로 성능 측정
- CelebA-HQ에서 비교적 뛰어난 성능
- GAN(+ 기존 DM)과 비교해서 data-mode 잘 커버, 더 적은 자원을 요구

Figure 4. Samples from *LDMs* trained on CelebAHQ [39], FFHQ [41], LSUN-Churches [102], LSUN-Bedrooms [102] and class-conditional ImageNet [12], each with a resolution of 256 × 256. Best viewed when zoomed in. For more samples *cf*. the supplement.

데이터셋셋

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.

Conditional Latent Diffusion

- **Transformer Encoders for LDMs**
  - cross attention 기반
  - various conditioning modalities(test-to-image)
  - test prompt : BERT, transformer encoder → UNet

  → AR, GAN 기반보다 나은 성능, 적은 params

Figure 8. Layout-to-image synthesis with an *LDM* on COCO [4], see Sec. 4.3.1. Quantitative evaluation in the supplement D.3.

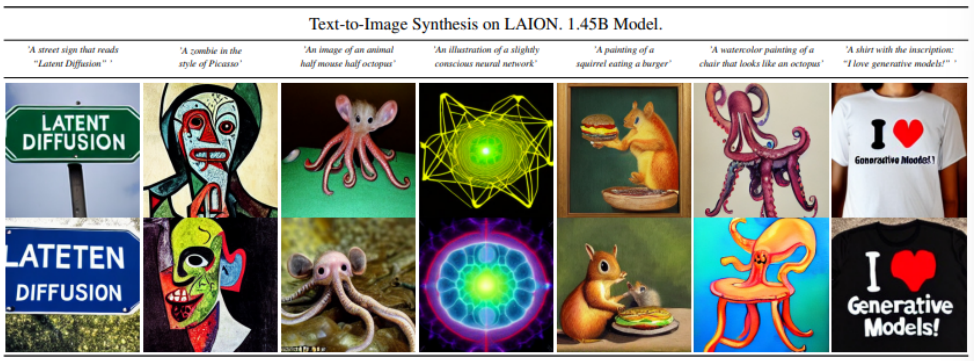

Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

- **Convolutional Sampling Beyond 256×256** : 고해상도에서 좋은 성능 (image-to-image)
- **Super-Resolution with Latent Diffusion**
  - low → high-resolution
  - 저해상도 이미지를 조건부로 입력, SR3 방식 → x4 downsampling
  - ImageNet
- **Inpainting**(빈 부분 채우기) **with Latent Diffusion**
  - LDM-1 (픽셀 기반)과 LDM-4 (latent)의 성능을 비교

  → LDM-4 가 빠르고 FID 점수 낮음
  - cross attention 사용 → quality, 사용자 선호 결과 제공 / 고해상도에서 성능 변화 (512x512에서 개선)

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

Method	40-50% masked		All samples	
	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
LDM-4 (ours, big, w/ ft)	**9.39**	0.246± 0.042	**1.50**	0.137± 0.080
LDM-4 (ours, big, w/o ft)	12.89	0.257± 0.047	2.40	0.142± 0.085
LDM-4 (ours, w/ attn)	11.87	0.257± 0.042	2.15	0.144± 0.084
LDM-4 (ours, w/o attn)	12.60	0.259± 0.041	2.37	0.145± 0.084
LaMa [88]†	12.31	**0.243**± 0.038	2.23	**0.134**± 0.080
LaMa [88]	12.0	**0.24**	2.21	0.14
CoModGAN [107]	10.4	0.26	1.82	0.15
RegionWise [52]	21.3	0.27	4.75	0.15
DeepFill v2 [104]	22.1	0.28	5.20	0.16
EdgeConnect [58]	30.5	0.28	8.37	0.16

Table 7. Comparison of inpainting performance on 30k crops of size 512 × 512 from test images of Places [108]. The column *40-50%* reports metrics computed over hard examples where 40-50% of the image region have to be inpainted. †recomputed on our test set, since the original test set used in [88] was not available.

Limitations & Societal Impact

(limitation)

- Speed : sequential sampling process에서 GANs보다 느린 속도
- Precision : px 기반 모델보다 낮음. 특히 super-resolution에서

(Impact)

- generative tech에 대한 접근성 향상(비용 감소)
- 미디어 복제, misinformation 등

Conclusion

- LDMs
- 퀄리티 저하 없이 훈련 및 샘플링 효율 개선
- cross-attention conditioning → specific architecture에서 벗어남