2주차: High-Resolution Image Synthesis with Latent Diffusion Models

논문 링크: https://arxiv.org/pdf/2112.10752

주요 키워드: Latent Diffusion Model(LDM)

Abstract

denoising autoecoder를 사용해 이미지를 분해하면서 디퓨젼 모델은 이미지 합성에 좋은 결과를 달성. 재학습 없이 이미지 데이터 생성 가능.

→ 픽셀 단위로 학습을 진행해 GPU 매우 많이 소모.

⇒ 이러한 문제 해결하고 성능과 유연성은 보존하기 위해 잠재 공간에서 확산 과정을 수행하며 계산 효율성 증가. 메모리와 계산 자원 절감 및 고품질 이미지 생성 가능. 이미지 합성 등에 있어 픽셀 단위 기존 모델보다 좋은 성능을 보임.

1. **Introduction**

이미지 합성 발전 과정:

고차원 이미지 합성은 주로 likelihood 기반 모델을 사용하며 많은 파라미터를 활용

GAN 모델은 적대적 학습 절차가 복잡하고 다중 모드 분포를 모델링하는 것이 어려워 비교적 변동성이 제한된 데이터에서만 사용

최근의 Diffusion 모델은 denoising autoencoder를 기반으로 만들어졌고, 이미지 합성 등에 인상적인 결과를 보임.

Democratizng High-Resolution Image Synthesis:

Diffusion 모델은 likelihood 기반 모델로 데이터의 디테일을 모델링하기 위해서는 많은 용량이 필요.

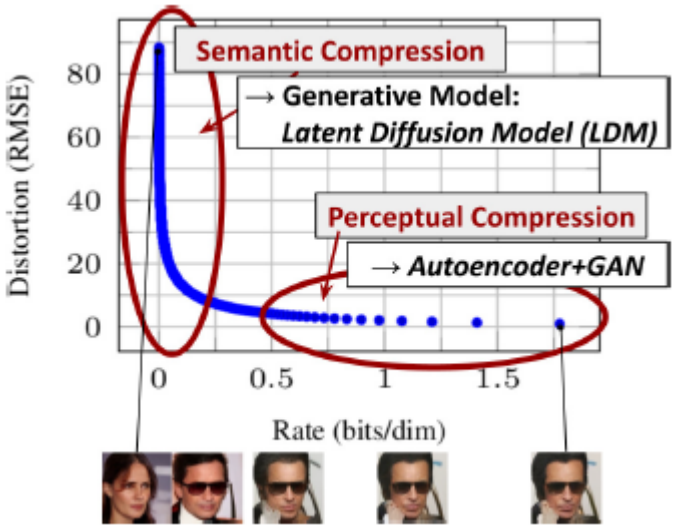
고차원의 RGB 공간에서 학습과 평가가 반복되어 계산량이 많음.

이러한 방식은 계산 자원이 많이 소모되고 메모리상으로 시간과 자원이 많이 들어간다는 단점이 있음.

⇒ Diffusion 모델 성능 저하 없이 훈련과 샘플링의 계산 복잡성을 줄여 자원 소비를 줄일 필요가 있다

Departure to Latent Space:

픽셀 공간의 pretrained diffusion 모델에서부터 접근 시작



pretrained 모델의 왜곡 트레이드 오프

보편적인 likelihood 기반 모델의 학습 과정:

1. 고차원 디테일은 제거하지만 semantic 다양성을 학습하는 perceptual compression 단계
2. 고차원 이미지 합성을 위해 diffusion 모델을 더 적합한 공간에서 수행

해당 논문의 모델 학습 과정:

1. 저차원의 respresentational 공간을 제공하는 오토 인코더 학습
2. Diffusion 모델은 잠재 공간에서 학습시켜 공간 압축에 신경쓸 필요 없음. 하나의 네트워크를 통과하는 잠재 공간에서 이미지 생성하는 것은 복잡성을 낮추고 효율적인 이미지 생성에 도움됨.

⇒ 장점: 오토 인코딩 단계를 한 번만 진행. 여러 Diffusion 모델 학습에 재사용하거나 완전 다른 태스크에 적용 가능해 효율적

논문 요점:

1. 트랜스포머 기반 방식보다 더 고차원 데이터까지 확장 가능
2. 여러 태스크에 대해 경쟁력 있는 성능 달성. 픽셀 기반 방식보다 들어가는 비용 감소

3. 정교한 reconstruction 과 생성 능력이 필요 없음. 잠재 공간의 규제 조금만 필요
4. densely conditioned 태스크를 컨볼루션 방식으로 처리
5. Diffusion 모델보다 더 다양한 태스크에 재사용할 수 있는 pretrained latent diffusion & autoencoding 모델

2. **Related work**

Generative Models for Image Synthesis:

GAN은 고해상도 이미지를 고성능으로 이해 가능하지만 전체 데이터 분포를 파악하고 최적화하기 어려움이 있음.

likelihood 기반 모델은 최적화 성능이 좋아 고해상도 이미지를 효율적으로 합성할 수 있지만 품질이 나쁨.

Auto Regressive 모델은 밀도 추정의 성능이 좋지만 계산 비용이 많이 들고 저해상도 이미지만 처리 가능

Diffusion Probabilistic Models:

밀도 추정에 좋은 성능을 보임. reweight 과정을 통해 픽셀 공간에서 최적화 결과 도출해 느린 속도와 비용이 많이 듦.

Two-Stage Image Synthesis:

각각의 접근의 단점을 완화하기 위해 두 방법 결합.

VQ-VAE 는 autoregressive 모델로 먼저 잠재 공간을 학습하고 결합 분포로 이미지와 텍스트 representation 학습

3. **Method**

generative 학습 과정에서 compression 과정을 분리해서 계산량 줄임. auto encoding 방법 활용

장점:

1. 고차원이 아닌 공간에서 Diffusion 모델이 더 효율적으로 진행. 샘플링이 저차원에서 수행되어서.
2. UNet 모델로부터 상속된 inductive bias 를 활용해 공간 구조의 데이터에 특히 효율적이고 압축이 필요 없음
3. 잠재 공간이 다양한 생성 모델을 학습할 수 있도록 할 수 있는 general 한 압축 모델을 얻는다.

3.1 Perceptual Image Compression

perceptual loss 와 patch-based adversarial objective의 조합으로 학습된 오토인코더를 기반으로 함.

지역 realism을 적용해 reconstruction이 이미지 manifold에 제한되고 L2,L1 규제 같이 픽셀 공간 로스에만 의존해 흐려짐 방지

잠재 공간에서 높은 변동성을 피하기 위해 두가지 규제로 실험

KL-reg: 약간의 KL 페널티를 기준에 적용. 학습된 잠재 공간으로 이차원 구조를 통해 학습하므로 상대적으로 약합 압축을 하게 되어 reconstruction에 용이함.

3.2 Latent Diffusion Models

diffusion 모델은 확률 모델로 기본 분포를 점차적으로 denoising 하며 학습한다.

Denoising 오토 인코더에 의해 일정한 가중치를 가지고 있다고 해석될 수 있다.

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

Latent Representation. 효율적으로 저차원의 잠재 공간에 반영해 더 중요한 데이터에 집중하고 저차원에서 학습하며 효율적으로 계산 가능. 특히 이미지 데이터 학습에 이점이 있다.

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

reweighted bound를 사용함. zt는 학습 과정중에 얻어지는 값임.

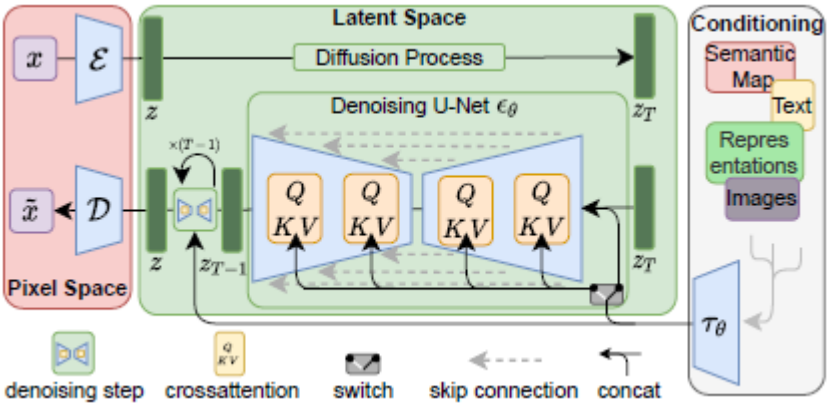
3.3 Conditioning Mechanisms

조건부 denoising 오토인코더로 인풋을 합성 과정에 활용 가능.

Diffusion 모델을 더 유연한 조건무 이미지 생성기로 바꿈. cross-attention 메커니즘을 통해. → 다양한 인풋을 학습할 때 효율적

cross attention layer는 UNet 의 중간 층을 flatten 해서 Q에 사용하고 도메인 특화 인코더를 K,V에 사용

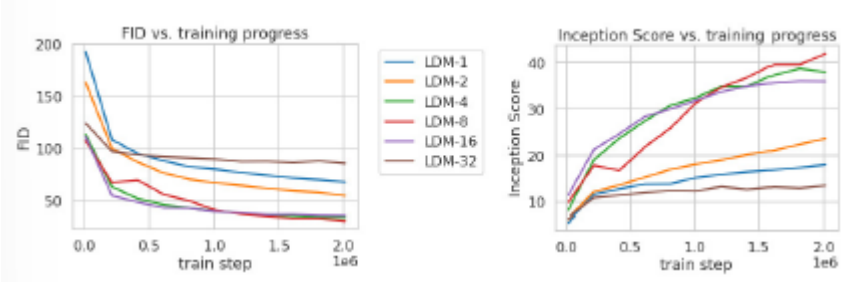
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$



4. **Experiments**

LDM은 다양한 이미지 합성에 대해 유연하고 추적 가능한 모델링이다. 모델은 픽셀 기반 모델과 비교하면 더 나은 샘플 품질을 달성하는 것을 확인할 수 있다.

4.1 On Perceptual Compression Tradeoffs

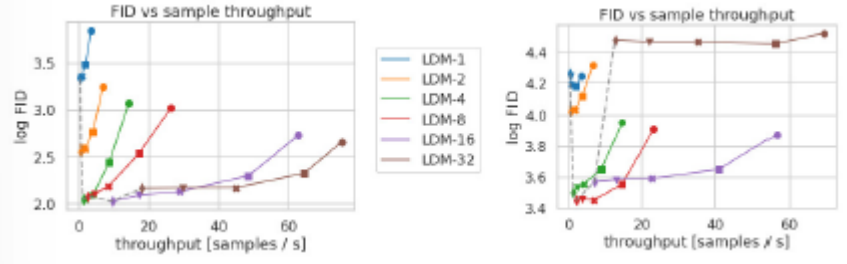


ImageNet으로 클래스 조건부 모델을 학습할 때 step에 대한 샘플 품질

factor가 작을 때 학습이 느려짐

factor가 너무 크면 step이 적을 때부터 품질 향상이 더딤

factor가 4에서 16 정도일 때 FID와 Inception 점수가 적절한 균형을 유지함.



CelebA-HQ / ImageNet 데이터에서 LDM으로 학습했을 때 샘플링 속도와 FID

LDM-4,8이 200만번 학습 했을 때 성능이 가장 좋음

4.2 Image Generation with Latent Diffusion

샘플 품질과 커버하는 데이터 매니폴드, FID와 정밀도 recall 등으로 모델 평가

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

CelebA-HQ 데이터에서는 FID 결과가 다른 likelihood 기반 모델(GAN, LSGM)보다 좋은 결과를 보임

LDM이 DM, GAN 에 대해 더 적은 계산 자원을 사용하면서도 높은 성능을 보임

4.3 Conditional Latent Diffusion

Transformer encoders for LDMs

cross attention 기반의 LDM 성능 확인을 위해 다양한 대상으로 실험

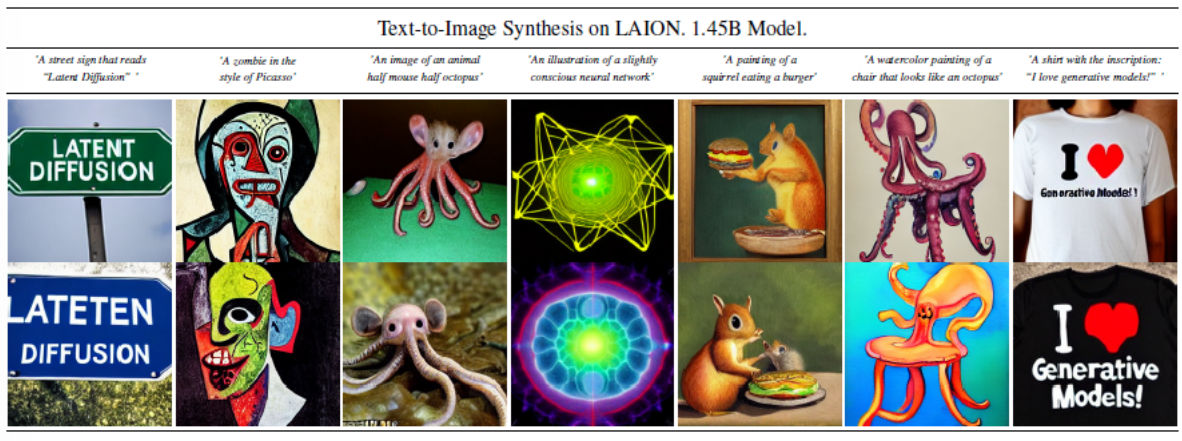


Figure 5. Samples for user-defined text prompts from our model for text-to-image synthesis, *LDM-8 (KL)*, which was trained on the LAION [78] database. Samples generated with 200 DDIM steps and $\eta = 1.0$. We use unconditional guidance [32] with $s = 10.0$.

텍스트→이미지 합성 결과

레이아웃→ 이미지 합성 결과

Convolutional Sampling beyong 256*256

256*256 크기로 학습한 데이터를 더 큰 이미지를 생성하도록 한 결과

4.4 Super-Resolution with Latent Diffusion

64에서 256으로 upscaling 이 가능하다

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑	Nparams	[samples/s](*)
Image Regression [72]	15.2	121.1	**27.9**	**0.801**	625M	N/A
SR3 [72]	5.2	**180.1**	26.4	0.762	625M	N/A
LDM-4 (ours, 100 steps)	2.8†/4.8‡	166.3	$24.4_{\pm 3.8}$	$0.69_{\pm 0.14}$	**169M**	4.62
emphLDM-4 (ours, big, 100 steps)	**2.4**†/**4.3**‡	174.9	$24.7_{\pm 4.1}$	$0.71_{\pm 0.15}$	552M	4.5
LDM-4 (ours, 50 steps, guiding)	4.4†/6.4‡	153.7	$25.8_{\pm 3.7}$	$0.74_{\pm 0.12}$	184M	0.38

Upscaling 할 때 기존 파라미터의 1/4배만 사용하면서도 좋은 뛰어난 성능을 보인다

4.5 Inpainting with Latent Diffusion

Inpainting: 이미지의 마스킹된 부분을 새로 채움. 미완성 이미지의 빈 부분을 채우거나 다른 것으로 대체

기존의 이미지 제거를 시도한 결과.

5. **Limitations & Societal Impact**

한계: 계산량을 확연히 줄여주지만 샘플링 과정이 GAN 모델보다 느리다. 높은 정확성을 보이지는 않음. 오토 인코더에 의한 이미지 품질 손실은 적지만 픽셀 공간에서 높은 정확도가 필요할 때는 reconstruction 과정이 병목으로 작용 가능

사회적 영향: 생성 모델은 양날의 검. 다양한 창의적인 곳에 사용될 수도 있지만 복제 데이터나 잘못된 정보가 쉽게 생성될 수 있다는 의미이기도 함. 데이터에 이미 존재하는 bias를 재현하거나 악화하는 경향이 있음.

6. **Conclusion**

LDM은 품질 저하 없이 학습하고 샘플링하는데 간단하고 효율적인 방법.

잠재 공간에서 diffusion으로 메모리 사용도 줄이고 고품질 이미지 효율적으로 생성.

다양한 이미지 생성에서 뛰어난 성능을 보이고 고해상도 이미지 생성에 매우 효과적.