CUAI Multimodal 논문 리뷰 스터디팀

2024.03.26

발표자 : 김태환

스터디원 소개 및 만남 인증

스터디원 1 : 오규안 (AI)

스터디원 2 : 김태환 (AI)

스터디원 3 : 김태윤 (소프트웨어)

논문 리뷰 주제

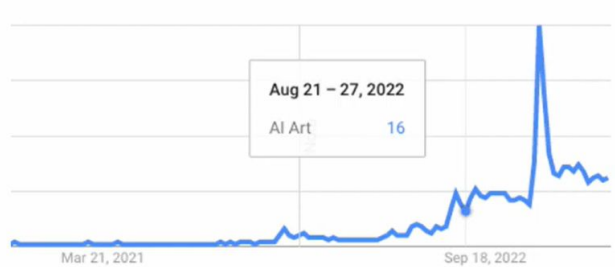
스터디원 논문 리뷰

1. 태윤님 : **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations**
2. 규안님 : **ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks**
3. 태환님 : **High-Resolution Image Synthesis with Latent Diffusion Models**

발표할 논문

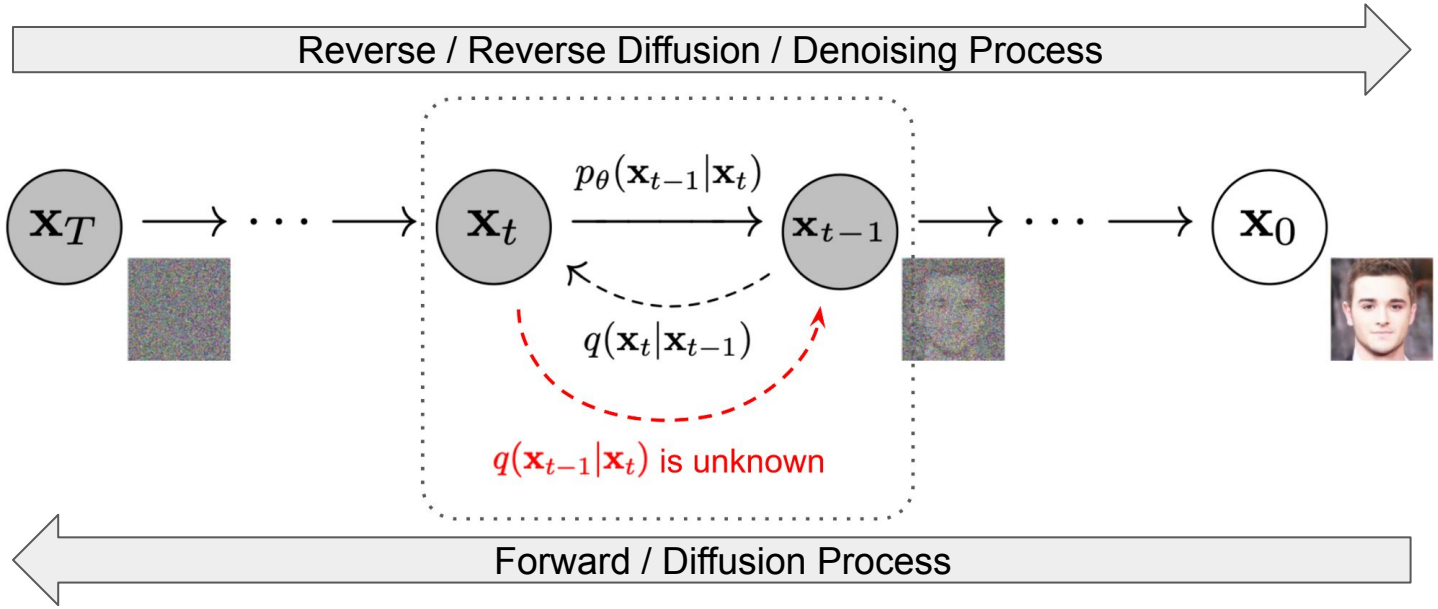
High-Resolution Image Synthesis with Latent Diffusion Models (2022)

Stable Diffusion이 여기서 제안한 **Latent Diffusion Model**의 일종



"AI Art" 의 Google Trends.
Stable Diffusion 공개일 기준으로 급격히 상승

Diffusion Model



Forward Process: 이미지가 완전한 Gaussian Noise가 되도록 점진적으로 Gaussian Noise를 추가

Reverse Process: Gaussian Noise에서 점진적으로 Gaussian Noise를 제거하여 원본 이미지 복원

Perceptual and Semantic Compression

일반적으로 생성모델은 이미지를 두 단계로 학습: Perceptual compression, Semantic compression

Perceptual Compression: High frequency detail들은 사라지지만 Semantic은 유지되는 구간
Semantic Compression: 실제 데이터의 본질이 Abstract하게 학습되는 구간

Perceptual and Semantic Compression

일반적으로 Diffusion Model은 Perceptual compression이 과도하게 오래 걸림
저자들은 이를 오토인코더로 수행하고, Diffusion이 Semantic compression을 하게 함
이에 일반적인 모델을 Pixel space model, Compression을 진행한 모델을 Latent space model이라 함

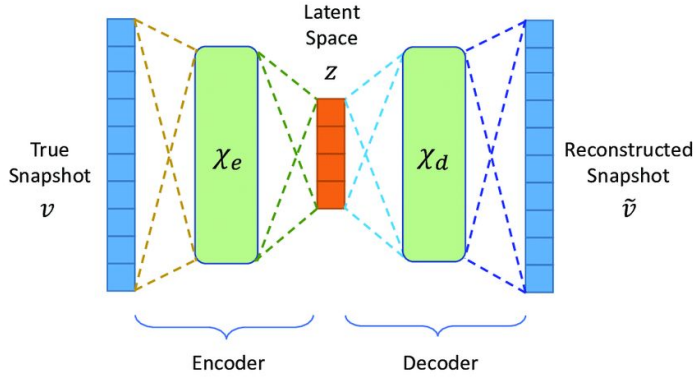
Input image

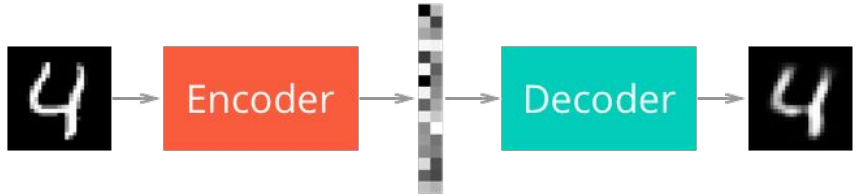
Latent representation

Autoencoder

입력이 들어왔을 때, 해당 입력 데이터를 최대한 compression 시킨 후,
compressed data를 다시 본래의 입력 형태로 복원 시키는 신경망

Encoder: 데이터를 압축하는 부분
Decoder: 데이터를 복원하는 부분
Latent vector: 압축 과정에서 추출한 의미 있는 데이터 Z

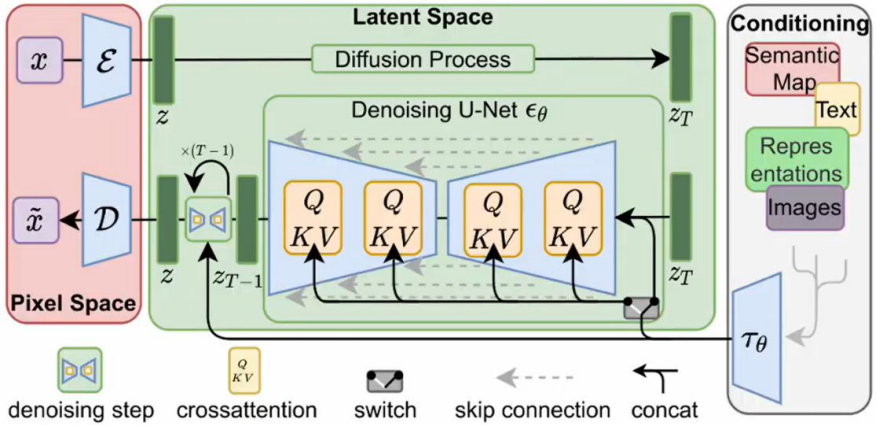




Architecture of LDM

트랜스포머의 텍스트 인코더 블럭 사용, 기존 Diffusion의 UNet에 Cross-Attention으로 Condition 주입
Stable Diffusion은 CLIP의 텍스트 인코더 사용

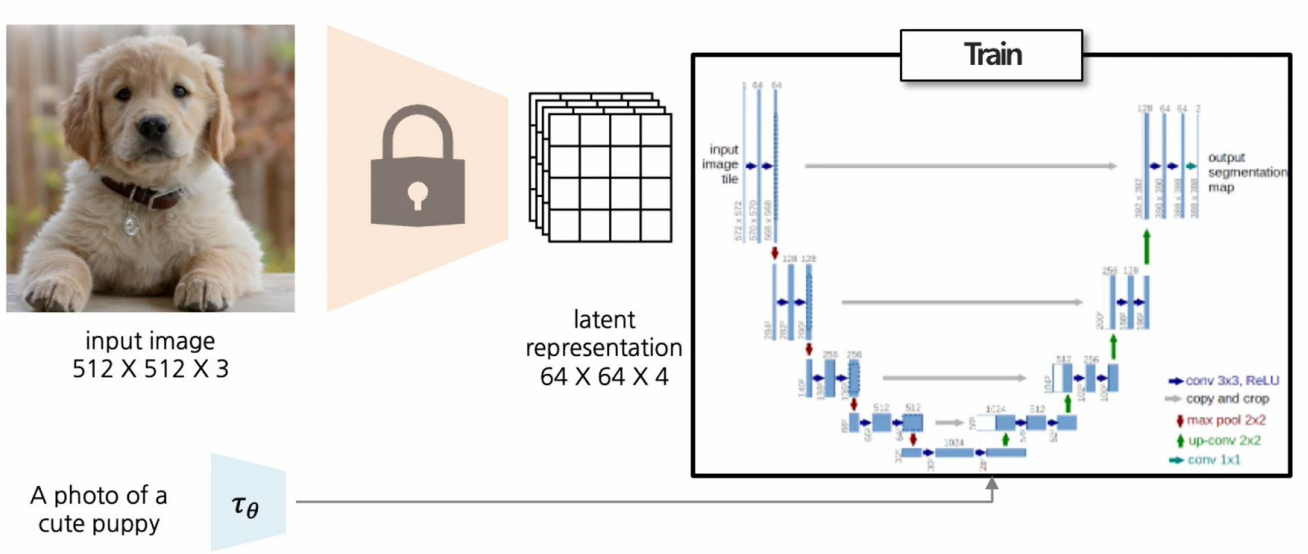
$$x \in \mathbb{R}^{H \times W \times 3}, z \in \mathbb{R}^{h \times w \times c}$$
$$z = \mathcal{E}(x)$$
$$\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$$



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

Training

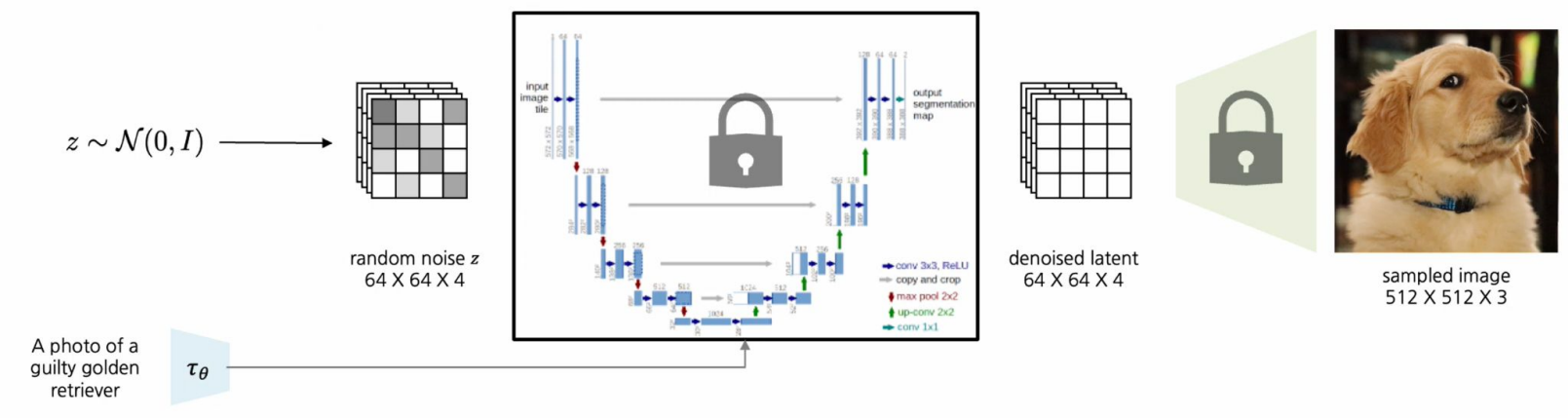
학습된 Autoencoder를 사용하여 Latent representation 획득
Latent representation으로 Diffusion model 학습



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

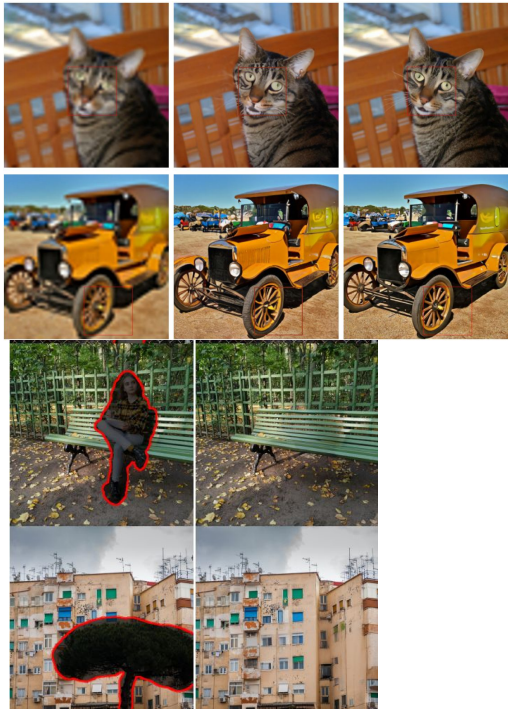
Sampling

Gaussian에서 Latent와 같은 사이즈의 Random noise 획득
텍스트 프롬프트를 텍스트 인코더에 입력
Random noise로 Sampling 진행 후 Decoding
텍스트를 얼마나 반영할지는 CFG 이용



Other Conditions

텍스트 프롬프트 외에도 다양한 컨디션으로 학습하여 수행 가능
Semantic synthesis, Super resolution, Layout to image synthesis, Inpainting



Conclusion

오토인코더의 Latent space에서 DDPM을 학습하면 throughput이 좋아지고 학습 속도가 빨라진다.

적어진 계산 비용 때문에 고해상도 이미지 생성이 가능하다.

Task별 별도의 아키텍처 없이 광범위한 Conditional Image Synthesis task에 SOTA 모델들과 비교해도 손색이 없는 모델을 만들 수 있다.

먼저 읽으면 좋은 논문

U-Net: Convolutional Networks for Biomedical Image Segmentation (2015)

Attention is All You Need (2017)

DDPM: Denoising Diffusion Probabilistic Model (2020)

Diffusion Models Beat GANs on Image Synthesis (2021)

Learning Transferable Visual Models From Natural Language Supervision (2021)

감사합니다