DALL.E

120억 개의 파라미터를 가진 GPT-3기반의 모델, 2.5억 개의 이미지 텍스트 데이터를 쌍으로 학습

결과

- 사물을 의인화하는 것이 가능하며, 관련 없는 두 개의 컨셉을 합치는 것이 가능

- 추가적인 레이블 정보 없이 우수한 성능

- ZERO-SHOT 상황에서도 매우 우수한 성능을 보임

* ZERO-SHOT이란, TRAIN데이터에 없는 클래스를 예측하는 것.

DALL.E에 필요한 배경 지식

Transformer : Attention is All You Need(NIPS 2017)

- Encoder와 decoder로 구성
- Positional encoding 사용

GPT-2 : Language Models are Unsupervised Multitask Learners(OpenAI 2019)

- Autoregressively한 모델 output으로 나온 결과값을 다시 input으로 집어넣음
- Transformer 의 decoder
- 대규모 데이터 세트로 학습된 대용량 언어 모델

Auto-Encoder

- 데이터 인코딩을 효율적으로 학습할 수 있는 뉴럴 네트워크
- 학습할 때 입력 데이터와 출력 데이터를 동일하게 설정
- Bottleneck에 해당하는 중간 latent vector z로 변환 후 복원
- 입력 이미지는 압축된 정보로 표현된다는 장점
- Data manifold를 학습. 데이터를 잘 표현하며, 차원 축소를 해줌.

Variational Auto-Encoder(VAE)

- Decoder는 latent code가 사전에 정해 놓은 분포를 따른다고 가정

VQ-VAE: Vector Quantised-Variational AutoEncoder(NIPS2017)

- 256*256 이미지를 입력으로 받아, 32*32개의 토큰들을 계산

VQ-VAE2: Generating Diverse High-Fidelity Images with VQ-VAE-2(NIPS 2019)

- dalle에서 실제로 사용된 모델

논문의 연구 동기

다양한 TEXT 대규모 모델 및 데이터 세트 연구가 이루어졌지만, TEXT-TO-IMAGE TRANSLATION 연구는 대규모 데이터 모델 및 데이터 세트를 이용해 연구가 이루어지지 않음.

DALL-E 학습 과정

VQ-VAE-2와 유사하게 two-stage training procedure를 사용

1. 256 * 256 이미지를 32 * 32 grid의 이미지 토큰들로 압축함.
2. 256개의 BPE-encoded text token들과 1024개의 image token들이 연속적으로 입력
3. 학습 과정은 ELBO (Evidence Lower Bound)를 maximizing하는데, lower bound의 구성은 dVAE디코더 (이미지 토큰을 토대로 결과 이미지 예측), dVAE인코더 (입력 이미지를 토대로 이미지 토큰 예측), Transformer (텍스트와 이미지 토큰에 대한 joint distribution 예측)
4. Stage One – dVAE 인코더와 Dvae 디코더를 학습
5. Stage Two – the prior distribution(transformer)를 학습
6. Dalle는 하나의 text에 대하여 N개의 다양한 이미지를 생성하고 생성한 뒤 이미지를 비교하기 위해 CLIP(OpenAI 2021)을 사용해 이미지를 선택