DALL-E

TNT 24-1 CV

Min Seo Lee

2024/03/26

TRAIN AND TEST

1. 개요

DALL-E

- DALL_E는 120억개의 파라미터를 가진 GPT-3 기반의 모델로, 2,5억개의 데이터(텍스트,이미지)쌍으로 학습
- 사물을 의인화 하는 것이 가능하며, 서로 관련없는 두개의 컨셉을 합치는 것 또한 가능
- 복잡한 아키텍쳐나 추가적인 레이블 정보 없이 매우 우수한 성능을 보임
- 잘 학습된 DALL-E를 이용하여, zero-shot상황에서도 매우 우수한 성능을 보임

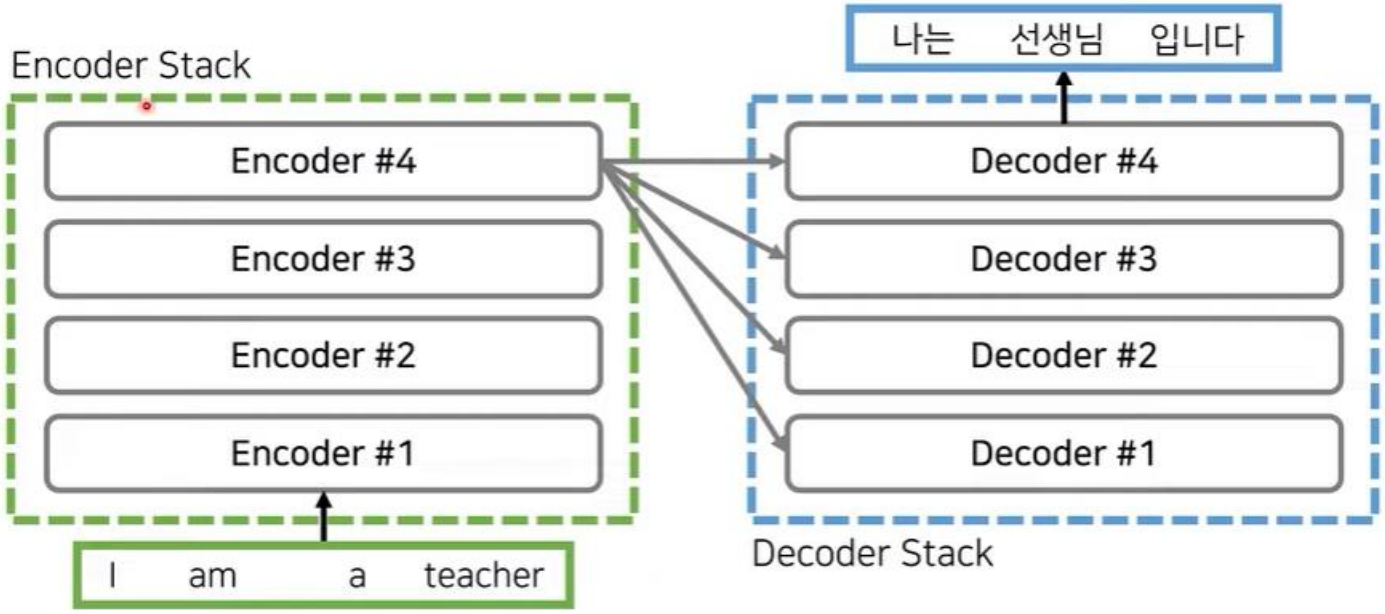
TRAIN AND TEST

2. 배경지식

Transformer

- 전통적인 트랜스포머에서는 마지막 인코더 레이어의 출력이 모든 디코더 레이어에 입력됨

레이어 개수가 4개일때 트랜스포머 아키텍쳐의 예시는 다음과 같음

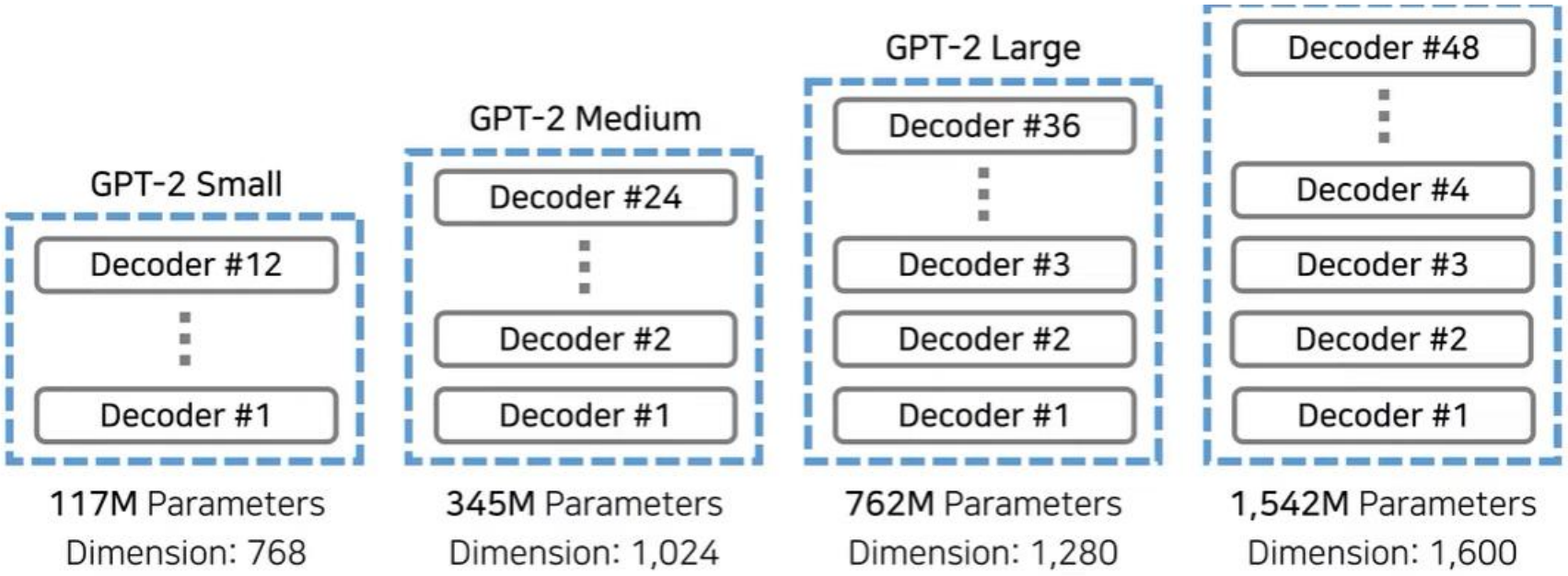


TRAIN AND TEST

2. 배경지식

GPT-2

- GPT-2는 트랜스포머의 decoder기반의 아키텍처로 대규모 데이터 세트로 학습된 대용량 언어모델
- 기계번역 목적이 아니므로 decoder만 사용해도 우수한 성능을 낼 수 있음

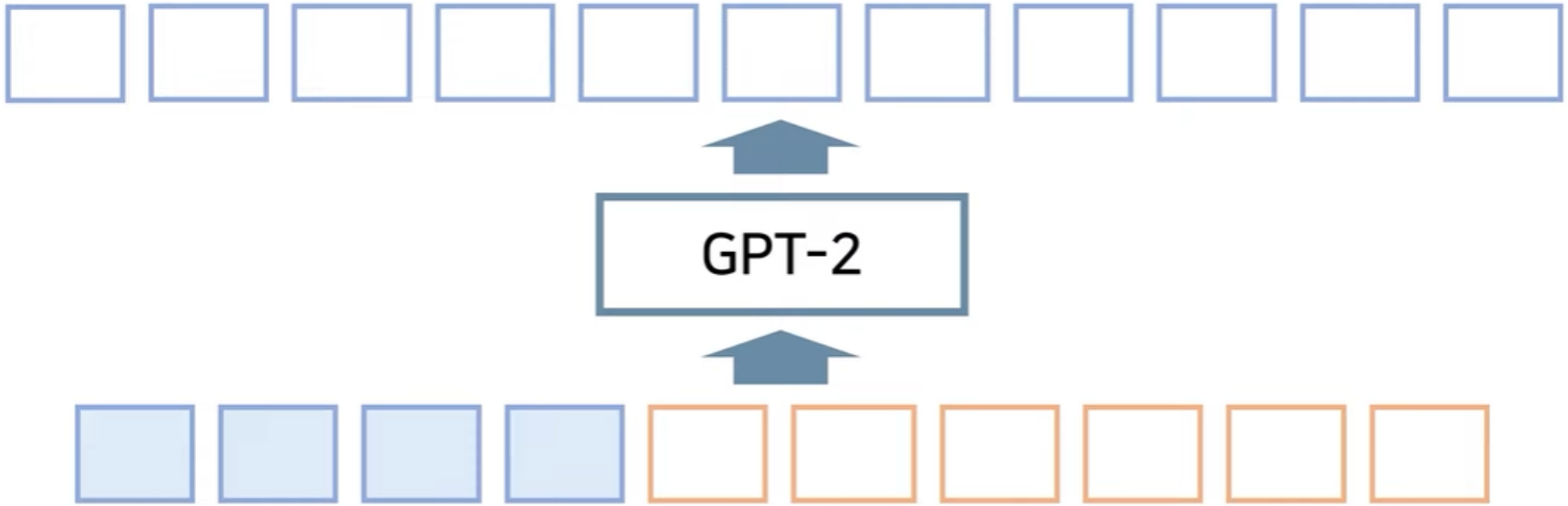


TRAIN AND TEST

2. 배경지식

GPT-2

- The GPT-2 autoregressively models the text tokens as a single stream of data(토큰을 입력으로 받았을때 다음 토큰을 예측하는 방식으로 모델링을 수행)
- 토큰(단어) sequence를 입력으로 넣으면 하나의 토큰이 출력되며 이를 다시 입력 sequence에 추가(그리고 다시 다음 토큰 예측하는 방식을 반복)



TRAIN AND TEST

2. 배경지식

GPT-2

$$Pr(x_1, x_2, \dots, x_n)$$

- x1,x2는 토큰이며 x1~xn까지의 sequence가 주어졌을때 이에 대한 확률값을 예측하도록 학습 진행(이때의 확률값은 chain rule을 통해 표현이 가능함)

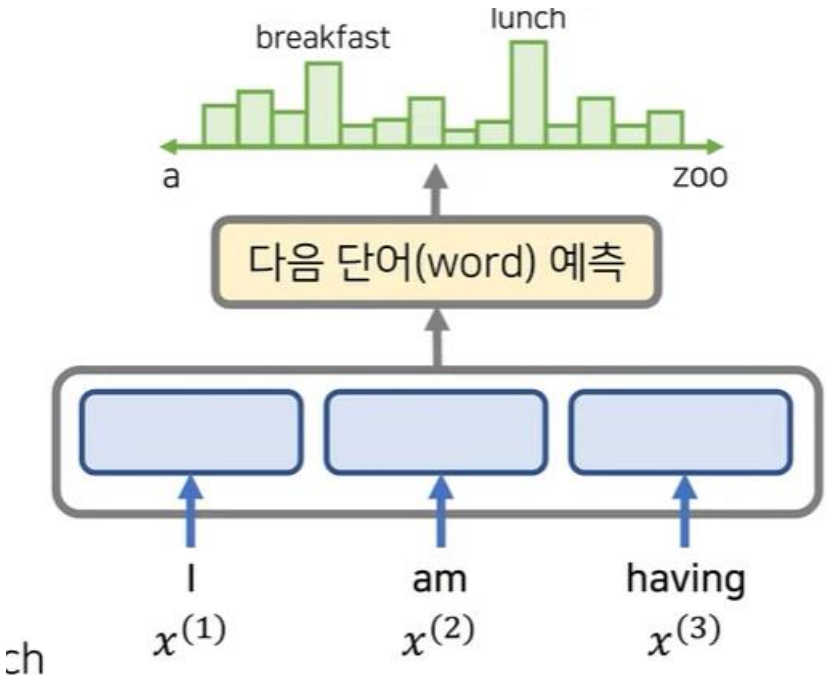
$$Pr(x_1, x_2, \dots, x_n) = Pr(x_1) * Pr(x_2|x_1) * Pr(x_3|x_1, x_2), \dots, * Pr(x_n|x_1, x_2, \dots, x_{n-1})$$

$$= \prod_{i=1}^{n} Pr(x_i|x_1, x_2, \dots, x_{i-1})$$

TRAIN AND TEST

2. 배경지식

GPT-2



[연쇄 법칙] $P(S) = P(I) \times P(am|I) \times P(having|I\ am) \times P(lunch|I\ am\ having)$

2. 배경지식

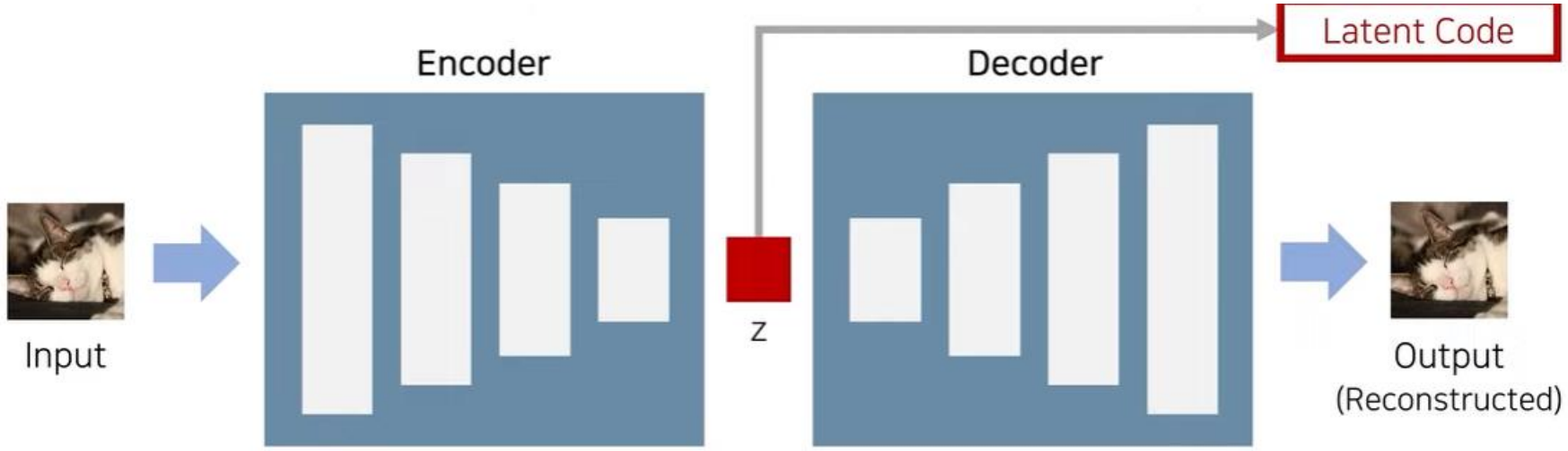
Auto Encoder

•데이터의 인코딩(data encoding)을 효율적으로 학습할 수 있는 뉴럴 네트워크

학습할때 입력 데이터와 출력데이터를 동일하게 설정

•모든 입력 이미지는 bottleneck에 해당하는 중간 latent vector z로 변환되었다가 복원됨

입력 이미지는 압축된 정보(latent code)로 표현될 수 있다는 장점이 있음



TRAIN AND TEST

2. 배경지식

Auto Encoder

- 일반적으로 픽셀 공간에서 강아지와 새의 이미지를 선형 보간하면 부자연스럽지만 잠재 공간에서 강아지의 latent 벡터와 새의 벡터를 선형보간할 경우에 상대적으로 자연스러운 이미지를 얻을 수 있음

[그림] 입력(픽셀) 공간에서의 interpolation

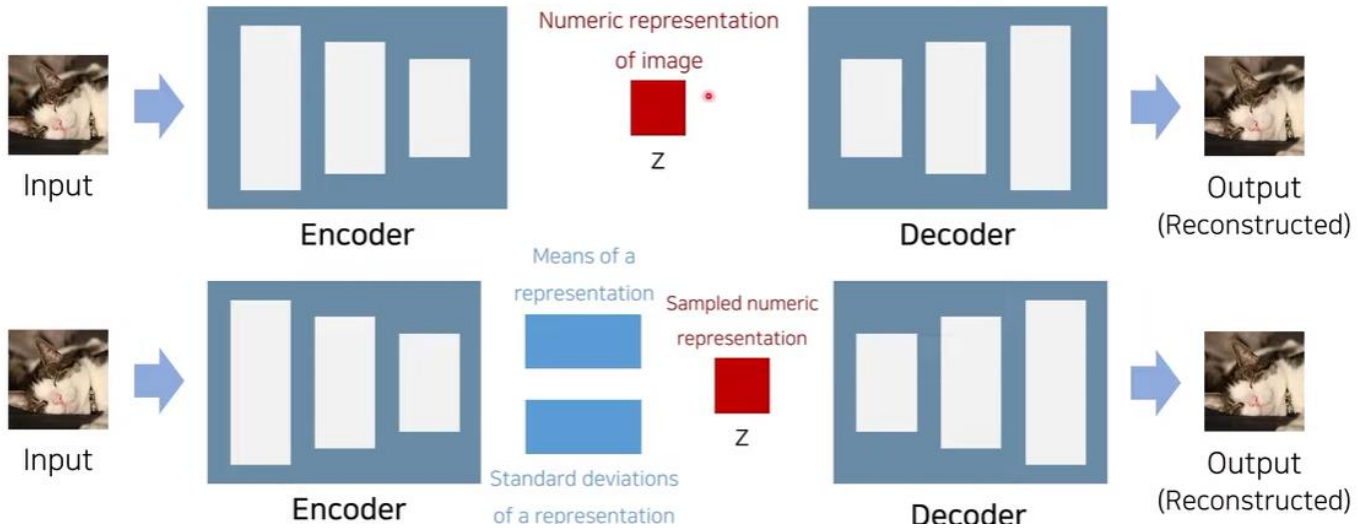
[그림] 잠재(latent) 공간에서의 interpolation

TRAIN AND TEST

2. 배경지식

VAE

•VAE의 decoder는 latent code가 사전에 정해놓은 분포를 따른다고 가정



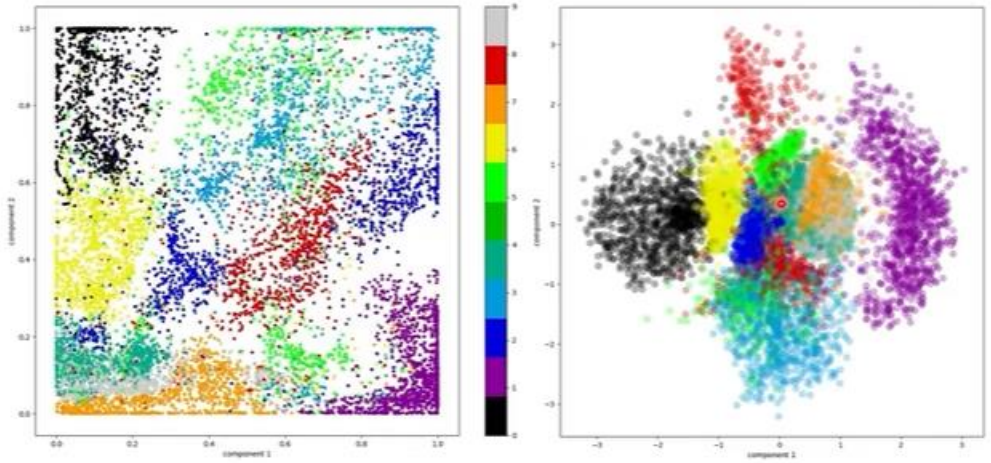
encoder를 거치면 mean 값과 variation을 얻는데 이때 mean과 variation으로 구성되는 distribution에서 z를 샘플링하도록 한다.

TRAIN AND TEST

2. 배경지식

VAE

•이처럼 latentvector가 가우시안과 같은 특정한 분포를 띄고 있다고 가정하기에 training data에 속하지 않는 새로운 데이터를 생성할때 유리



[그림] Auto-encoder과 Variational Auto-encoder의 MNIST

TRAIN AND TEST

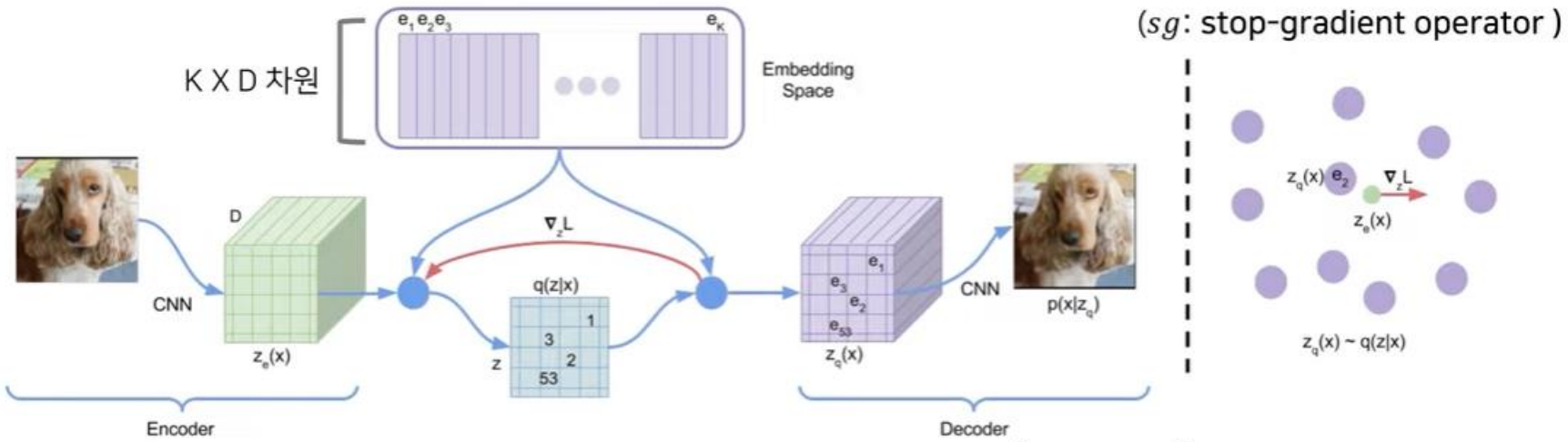
2. 배경지식

VQ-VAE

•기존의 VAE는 latent vector가 가우시안 분포를 따른다고 가정하였기에 값이 연속적으로 나왔으나 VQ-VAE는 미리 정해진 몇개의 값중 하나만 가질 수 있도록 함

$$L = \underbrace{\log p(x|z_q(x))}_{\text{Reconstruction loss}} + \underbrace{\|\text{sg}[z_e(x)] - e\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|z_e(x) - \text{sg}[e]\|_2^2}_{\text{Commitment loss}}$$

(sg: stop-gradient operator)



$$Quantization \text{ 방법: } z_q(x) = e_k \; where \; k = \arg\min_j \|z_e(x) - e_j\|_2$$

3. 논문의 연구 동기

DALL-E

•최근 대규모의 생성모델(large-scale generative models)을 활용한 연구들은 향상된 모델의 규모, 데이터 크기, 연산능력을 기반으로 한 auto-regressive transformer가 다양한 도메인에서 매우 뛰어날 수 있음을 보였음.

•그러나 Text-to image translation분야에서는 상대적으로 소규모 데이터 셋을 이용해 평가하는 경우가 많았기에 대규모 데이터 셋과 모델 아키텍처를 사용하는 것이 성능상의 돌파구가 될 수 있음을 보이는 것이 목적.(GPT-3와 2.5억개의 학습 데이터쌍을 활용한다)

4. DALL-E의 학습과정

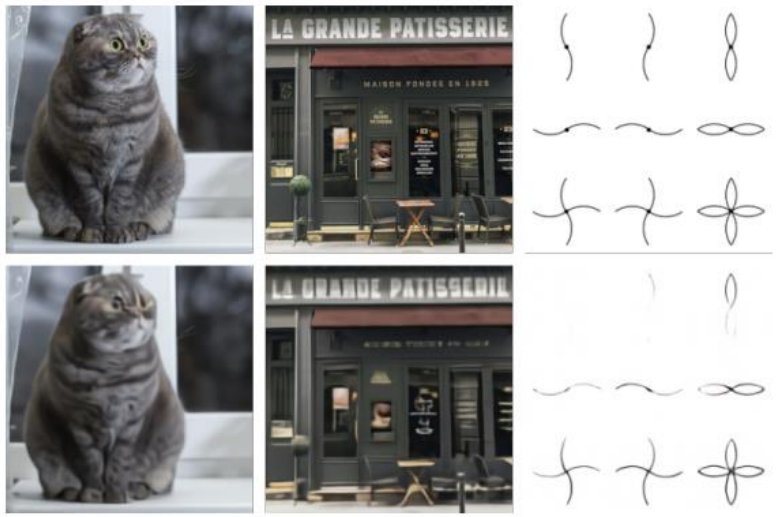
•본논문에서는 VQ-VAE-2와 유사하게 two-stage 방식을 사용

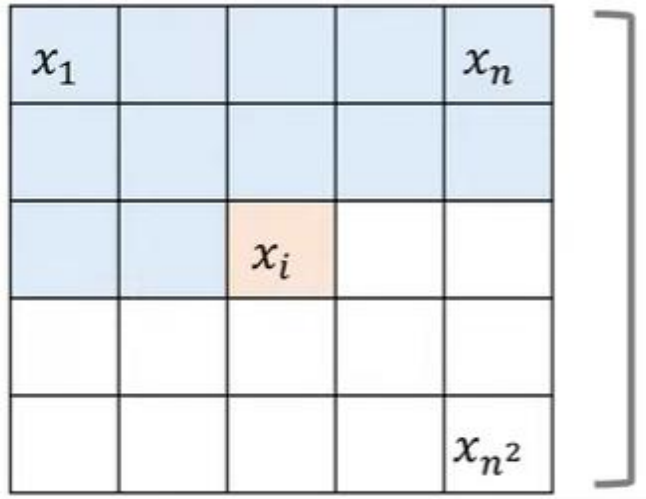
1.256x256 이미지를 32x32 grid의 이미지 토큰들로 압축한다.(각 토큰은 8192개의 코드중 하나로 배정)

이로써 큰 quality 손실 없이 transformer의 contect size를 8x8x3배 만큼 적게 만들수 있음

2. 어떤 이미지를 생성할지의 설명의 text가 토큰화 되어 총 256개 들어갈 수 있고 이미지는 1026개의 이미지 토큰의 sequence형태로 입력될 수 있도록 함

만약 픽셀 전체를 일일이 하나씩 연달아 예측할 경우 너무 비효율적이기에 32x32grid 형태로 예측하도록 함.

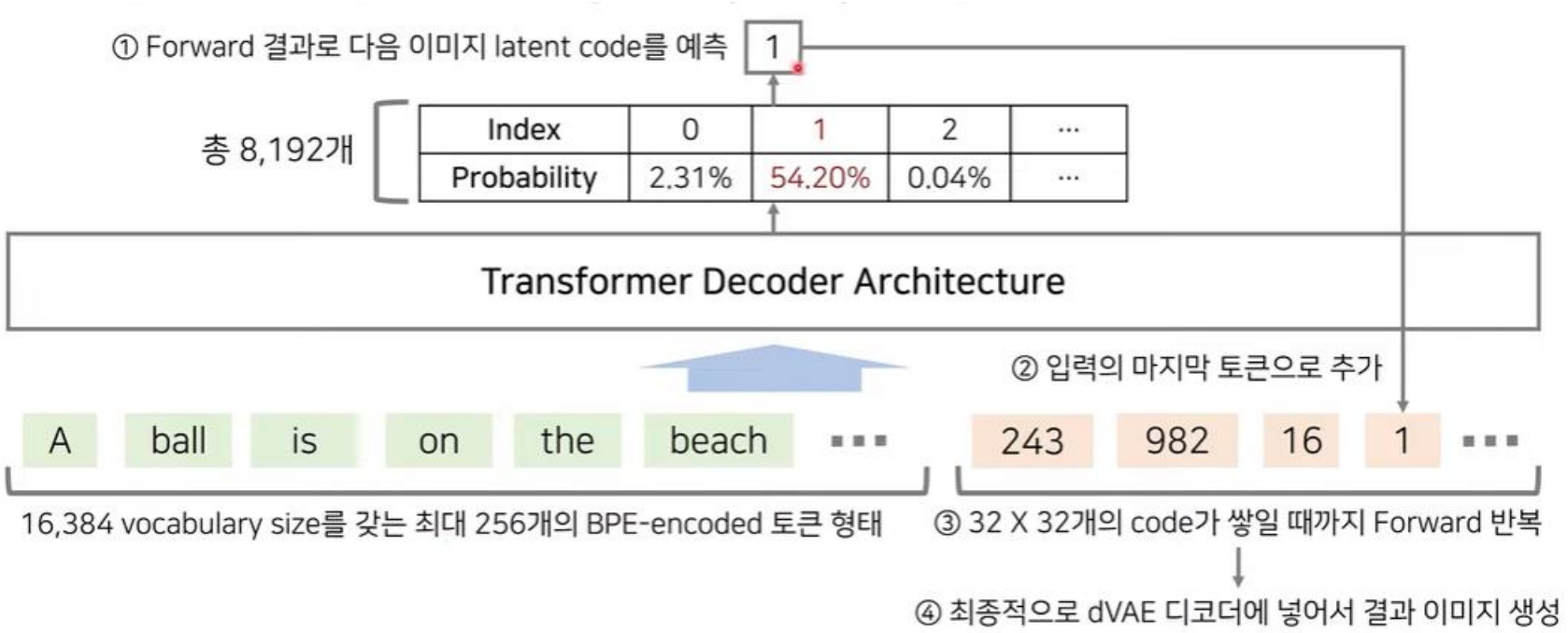




TRAIN AND TEST

4. DALL-E의 학습과정- overview

•실제로 모델을 사용할때는 text만 넣거나 text+image를 넣어서 결과 이미지를 생성할 수 있음.



TRAIN AND TEST

4. DALL-E의 학습과정

$$p_{\theta,\psi}(x,y,z) = p_\theta(x \mid y,z)p_\psi(y,z)$$

$$\ln p_{\theta,\psi}(x,y) \geqslant \mathop{\mathbb{E}}_{z\sim q_\phi(z\mid x)} \left(\ln p_\theta(x \mid y,z) - \beta\, D_{\mathrm{KL}}(q_\phi(y,z \mid x), p_\psi(y,z))\right)$$

p_θ: dVAE 디코더 (이미지 토큰을 토대로 결과 이미지 예측)

q_φ: dVAE 인코더 (입력 이미지를 토대로 이미지 토큰 예측)

p_ψ: Transformer (텍스트와 이미지 토큰에 대한 joint distribution 예측)

TRAIN AND TEST

4. DALL-E의 학습과정(1)

transformer는 고정하고 visual code book을 학습한다.(dVAE를 학습한다)

특정 입력 이미지가 주어졌을때 결과 이미지가 재대로 reconstructure될 수 있도록 한다.

Problem

DALL-E에서는 discrete problem을 gumbel softmax realtion을 이용해 해결

→ 학습을 진행할때는 argmax말고 softmax를 사용하여 gradient를 계산할 수 있도록 함

$$y_i = \frac{e^{\frac{g_i + log(q(e_i|x))}{\tau}}}{\sum_{j=1}^{k} e^{\frac{g_j + log(q(e_j|x))}{\tau}}} \quad z = \sum_{j=1}^{k} y_j e_j$$

TRAIN AND TEST

4. DALL-E의 학습과정(2)

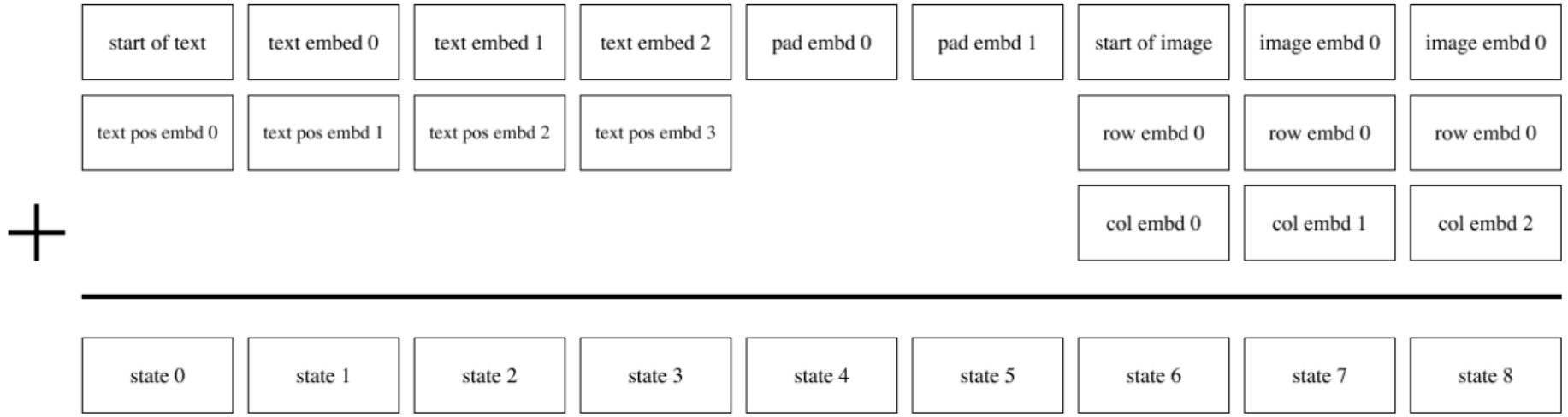


Figure 10. Illustration of the embedding scheme for a hypothetical version of our transformer with a maximum text length of 6 tokens. Each box denotes a vector of size $d_{\text{model}} = 3968$. In this illustration, the caption has a length of 4 tokens, so 2 padding tokens are used (as described in Section 2.2). Each image vocabulary embedding is summed with a row and column embedding.

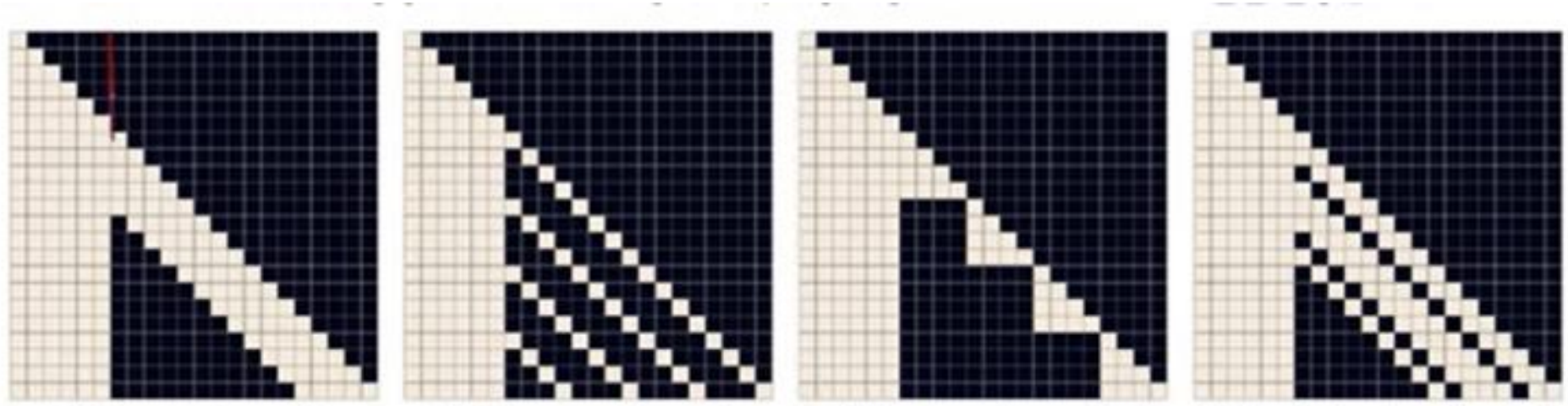
TRAIN AND TEST

4. DALL-E의 학습과정(2)

dAVE를 고정한채로 transformer를 학습

이미지 토큰은 dAVE인코더의 결과 logits에서부터 argmax sampling을 진행하여 생성함

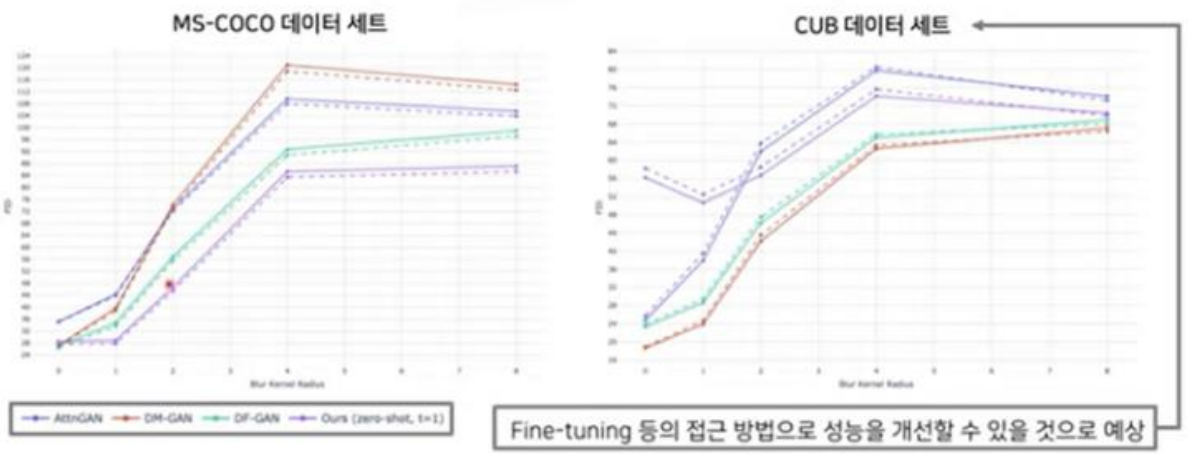
본 논문에서는 모든 테스트 토큰에 대하여 항상 attention을 하는 방식으로 다양한 attention mask를 사용한다.(각각의 상황에 따라 다양한 attention mask를 사용함)



(a) Row attention mask. (b) Column attention mask. (c) Column attention mask with (d) Convolutional attention mask.

TRAIN AND TEST

5. 성능 비교



TRAIN AND TEST

5. 성능 비교

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

TRAIN AND TEST

6. 결론

- 저자들은 **autoregressive transformer**를 기반으로 text-to-image generation task를 위한 **간단한 접근법**을 제안(특히 large scale에서)

- 그런 scale이 **zero-shot performance**와 **single generaive model** 등의 관점에서 **훌륭한 수준의 일반화 성능**을 보장했음을 보여줌

TRAIN AND TEST

TRAIN AND TEST