DALL-E : Zero-Shot Text-to-Image Generation (2021 ICML)

Jiin Kim

Teaser

Text to image Generation

(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

(c) a neon sign that reads "backprop". a neon sign that reads "backprop". backprop neon sign

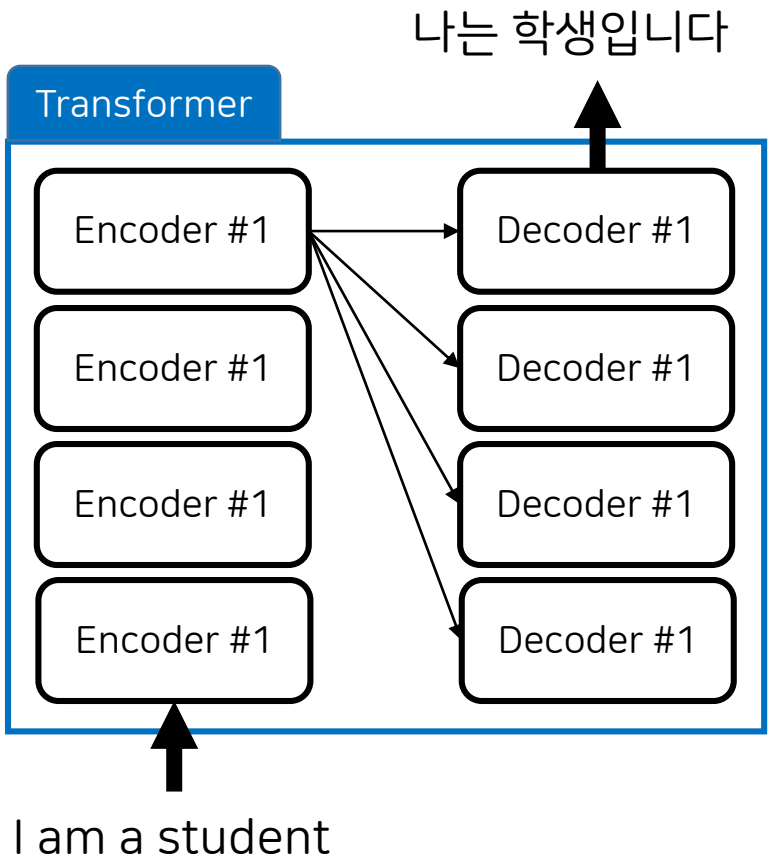
(d) the exact same cat on the top as a sketch on the bottom

Backgrond knowledge

Language Model (언어 모델)

- 언어 모델이란?
  - 자연어의 법칙을 컴퓨터로 모사한 것이다.
  - 주어진 단어들로부터 그 다음에 등장할 단어의 확률을 예측하는 방식으로 학습한다.
- Attention 기법이 개발되고 난 뒤로, 뛰어난 성능을 가진 많은 언어 모델들이 생겨났다.
- 기계번역 분야에서 Attention 만으로 구성된 Transformer 가 등장했다.

Transformer



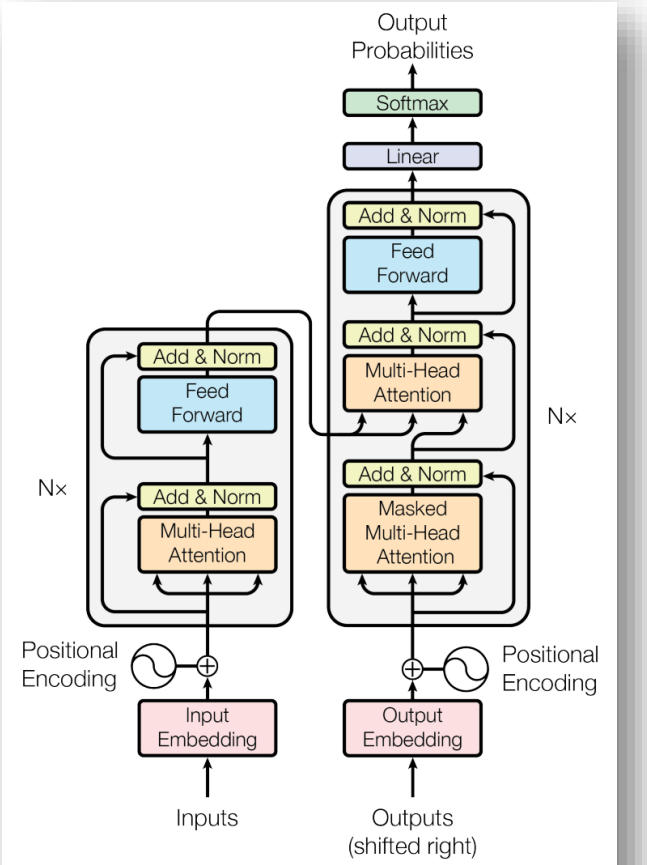
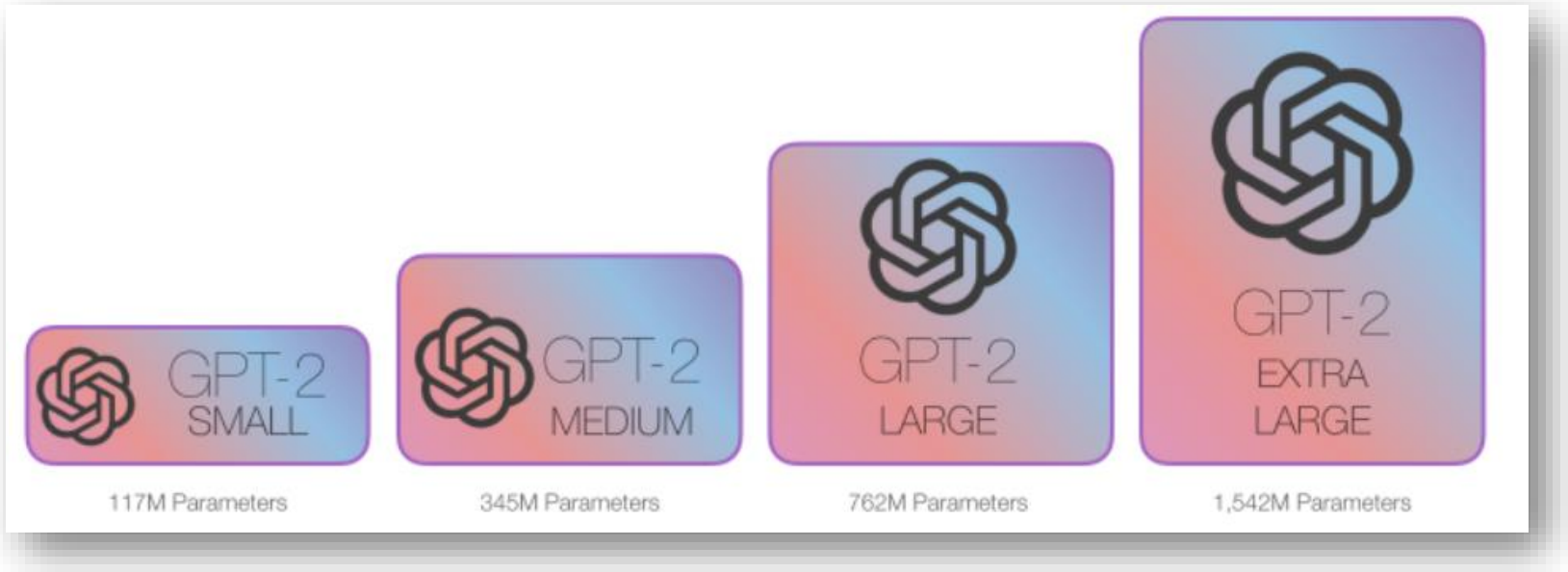


Figure 1: The Transformer - model architecture.

Language Model (언어 모델)

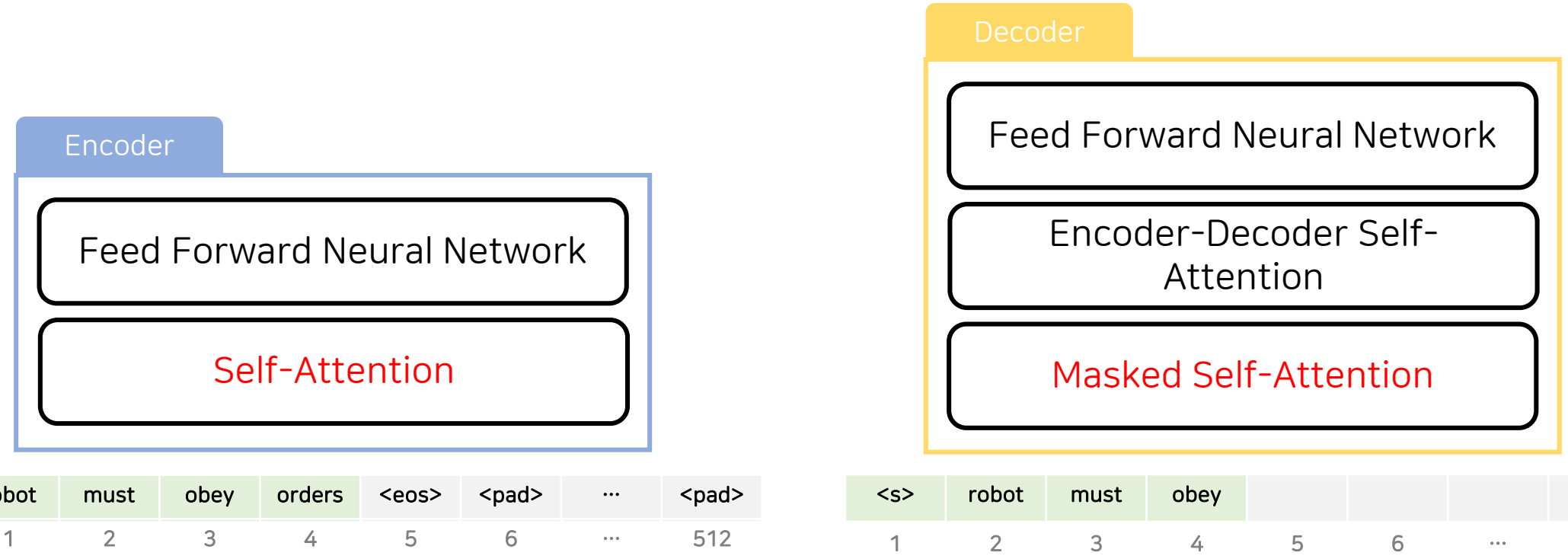
- 이에 그치지 않고, transformer 의 Encoder 혹은 Decoder 만 사용해야 한다는 주장들이 등장했다.
- 그리고, 등장한 모델이 그 유명한 **BERT** 와 **GPT-2** 이다.
- BERT 는 트랜스포머의 인코더 스택만 사용한 모델이고, GPT-2는 **디코더 스택**만 사용한 모델이다.

GPT-2



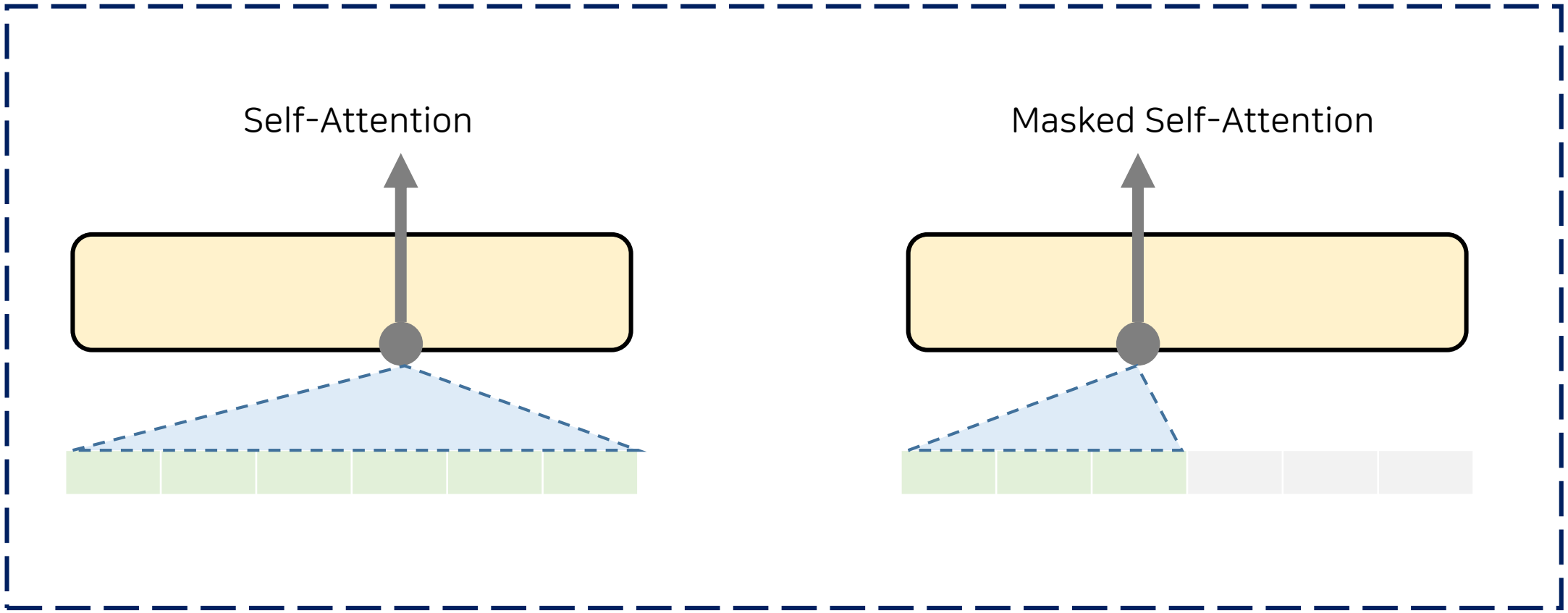
- GPT-2는 트랜스포머의 Decoder 기반의 아키텍처로, 대규모 데이터 세트로 학습된 대용량 언어 모델이다.
- DALL-E 에서 사용한 GPT 기반의 모델은 GPT-2 Extra large 보다도 훨씬 큰 모델이다. (GPT-3)

GPT-2 : Architecture



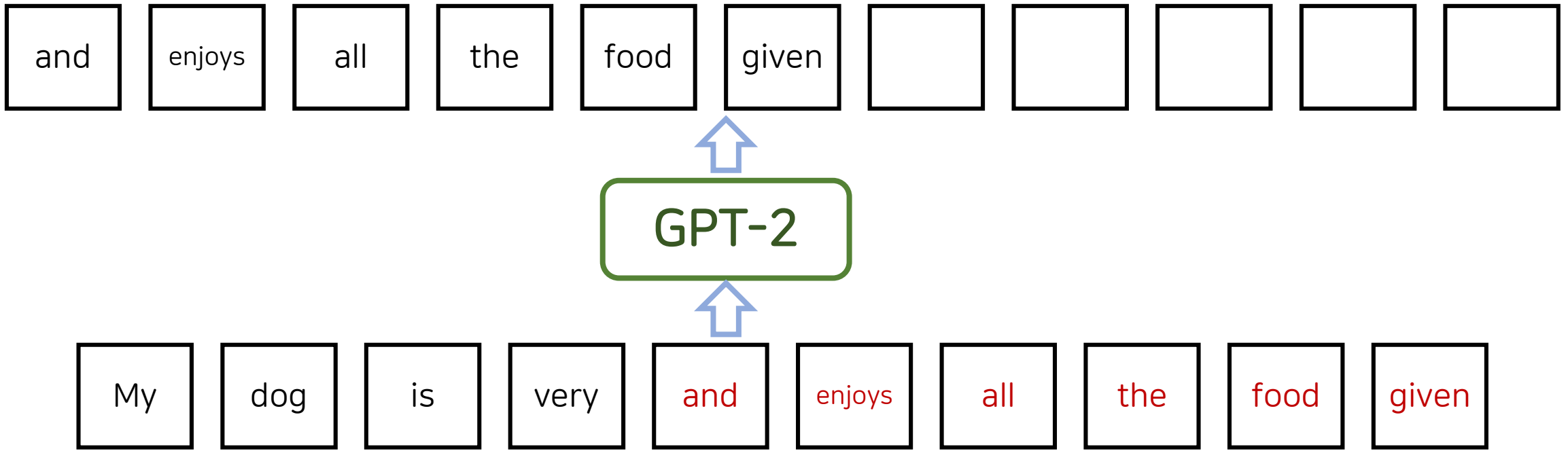
- Encoder 는 단순한 self-attention 을 사용하는 반면, Decoder 는 masked self attention 을 사용한다.
- 즉, GPT-2 는 self-attention 을 계산할 때 해당 스텝의 오른쪽에 있는 단어들은 고려하지 않는다.

GPT-2 : Attention

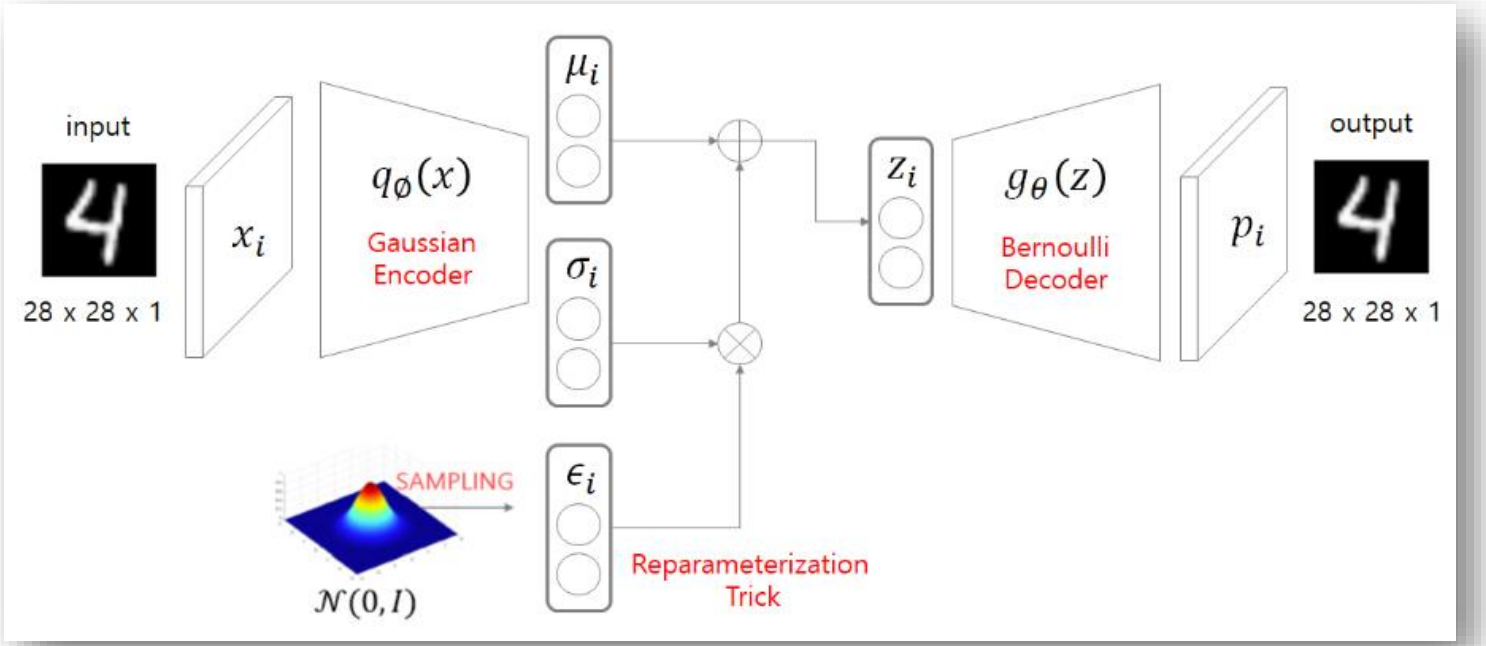


GPT-2 : Auto-Regressive

- GPT-2는 auto-regressive model이다.
- Auto-regressive model 이란? 이전의 출력이 다음의 입력이 되는 모델을 의미한다. (like RNN)



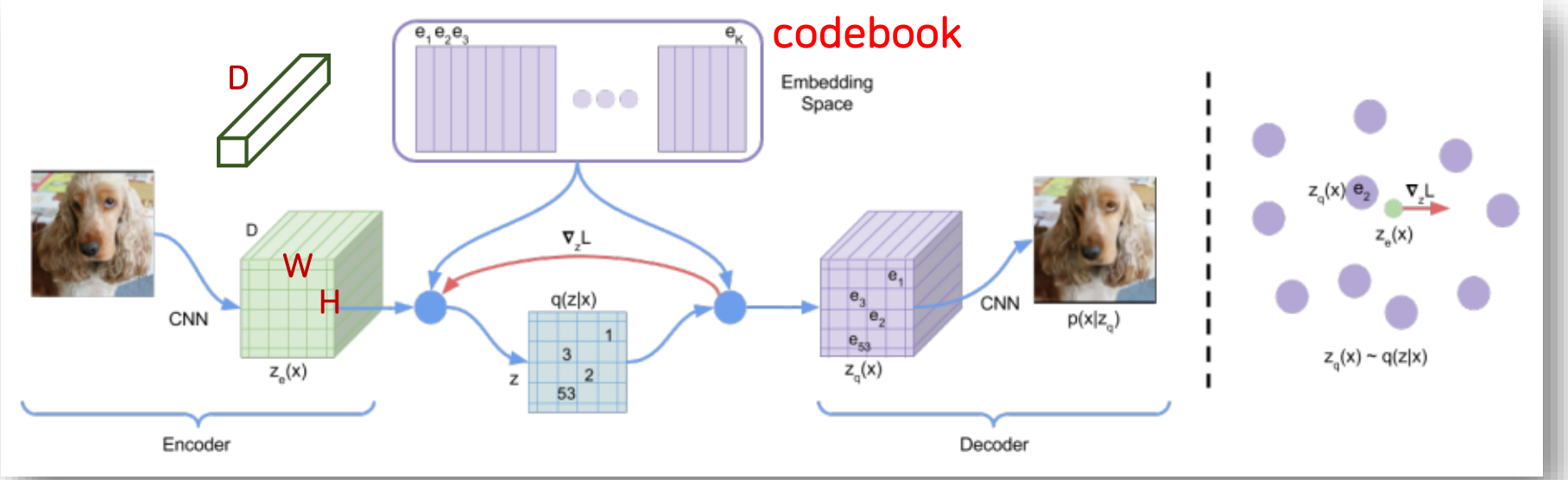
VAE



- Input image를 Encoder 에 넣으면 latent variable 의 평균(μ)과 분산(σ)이 출력된다.
- ϵ을 정규분포에서 랜덤하게 뽑고, 그 ϵ을 σ^2과 곱하고 μ 을 더해서 z를 만든다.
- z 를 Decoder 에 넣어 output image 를 생성한다.

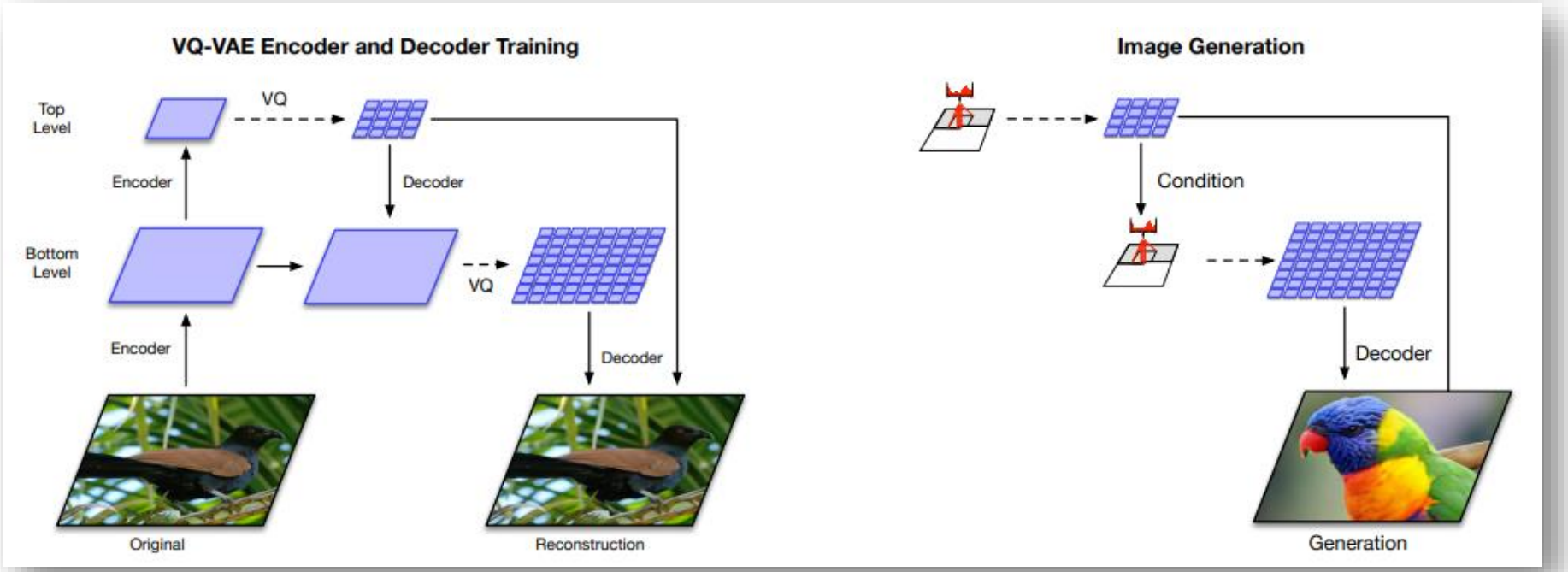
VQ-VAE : Vector Quantised-Variational AutoEncoder (2017 NIPS)

- VQ-VAE 는 discrete latent representation 을 학습한다.
- 이미지가 CNN을 통과하면 H x W개의 그리드로 나눠지고 (각 위치마다 D차원), 각 위치마다 codebook 의 e_1부터 e_k까지 중에서 가까운 1개로 변환된다.
- $Quantization$ 방법 : $z_q(x) = e_k \; where \; k = \underset{j}{\operatorname{argmin}} \left\| z_e(x) - e_j \right\|$



VQ-VAE-2 (2019 NIPS)

- Stage 1. 계층적인 VQ-VAE 를 학습한다. (local patterns 와 global information 을 분리하는 방식)
- Stage 2. latent discrete codebook 에 대한 확률(prior distribution)을 계산한다.
  - Self-attention autoregressive model 을 사용해 다음 토큰(codebook에서 하나의 code)를 예측한다.



DALL-E

Intro

- 규모가 큰 생성모델을 사용하면 성능이 향상된다는 것을 보였다.
- 구체적으로 모델의 크기, 연산 능력, 데이터의 크기를 신중하게 확장한 autoregressive transformers (Vaswani et al., 2017) 는 텍스트 (Radford et al., 2019), 이미지 (Chen et al., 2020), 오디오 (Dhariwal et al., 2020) 와 같은 여러 영역에서 인상적인 결과를 얻었다.

Intro

- 지금까지 Text-to-image generation 은 MS-COCO, CUB-200 와 같은 비교적 소규모 데이터셋에서 계산되어 왔다. DALL-E는 **대규모의 데이터 셋**과 **대규모의 파라미터**를 가진 모델을 사용하는 것이 성능을 높이기 위한 돌파구가 될 것이라고 생각했다.
- 그래서 2.5억 개의 이미지-텍스트 쌍을 학습 데이터로 사용했고 120억개의 파라미터를 가진 autoregressive transformer (GPT-3) 을 훈련시켰다.

Architecture

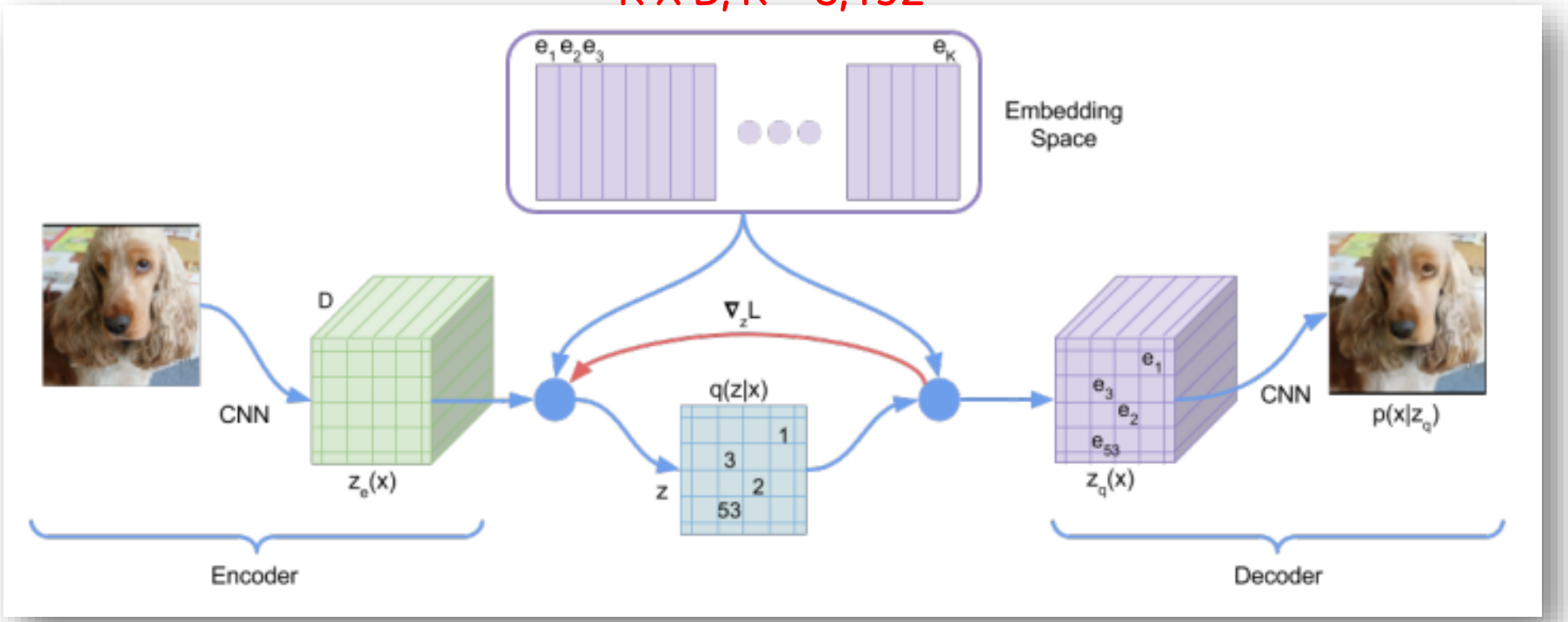
DALL-E

= dVAE(Stage 1) + Transformer(Stage 2)

Stage-1: dVAE

- dVAE 를 학습하여 256 X 256 RGB 이미지를 32 X 32 그리드의 이미지 토큰들로 압축한다.
  - 이러한 압축을 통해 transformer 가 처리해야 하는 크기를 192배 압축하면서, visual quality 는 유지할 수 있다.
- Transformer 를 고정한 상태로 dVAE 인코더 q_ϕ 와 dVAE 디코더 p_θ 를 학습한다. 이 때 초기 prior transformer p_ψ 는 uniform categorical distribution 으로 설정한다.

K X D, K = 8,192

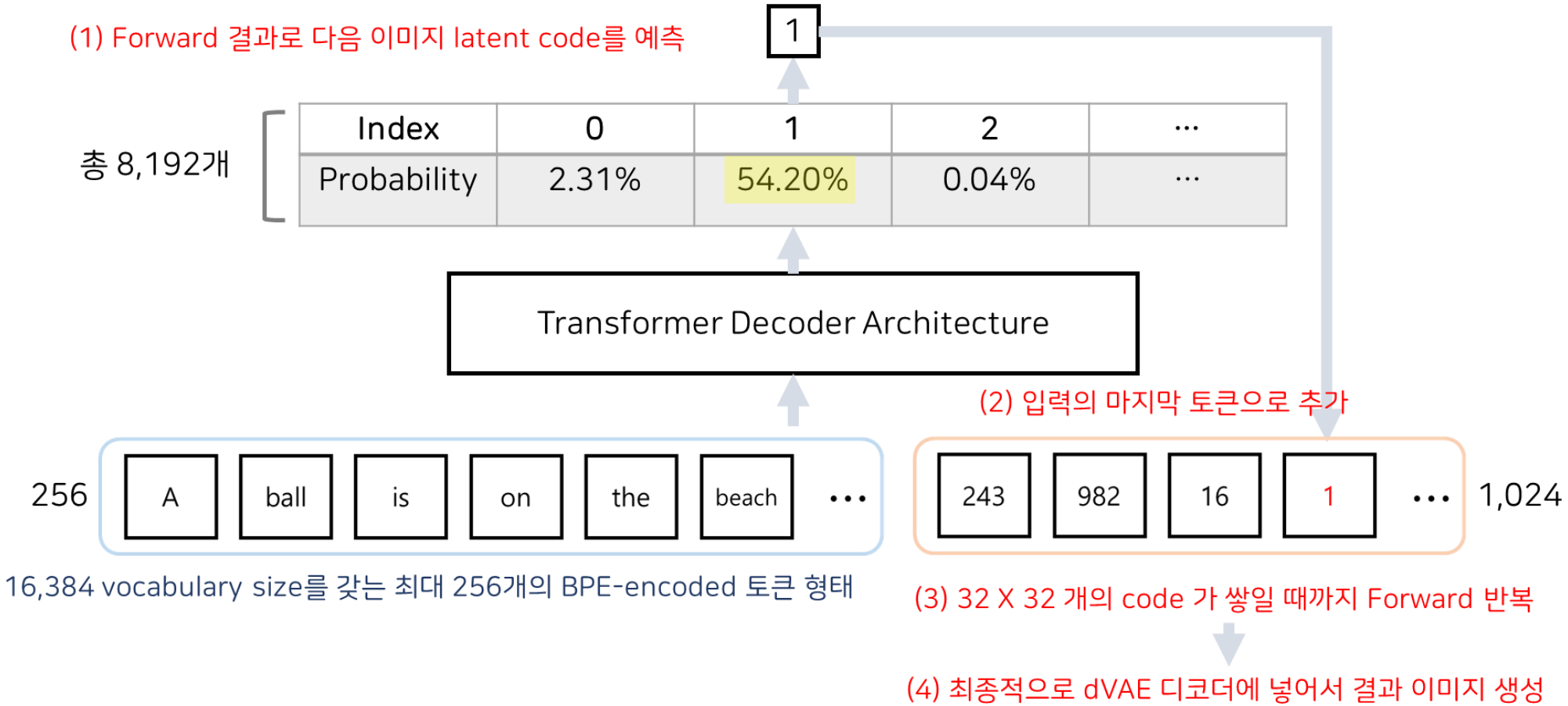


Stage-1: dVAE

- Original images (top) 와 discrete VAE (bottom) 으로부터 reconstructions 한 것이다.
- 고양이 털의 texture, 가게 앞의 글자, 삽화의 가는 선과 같은 디테일들이 때때로 손실되거나 왜곡되지만, 주요한 특징들은 일반적으로 여전히 알아볼 수 있다.
- 본 논문은 8,192 크기의 큰 vocabulary size (codebook vector의 개수)를 사용한다.

Stage-2: Transformer

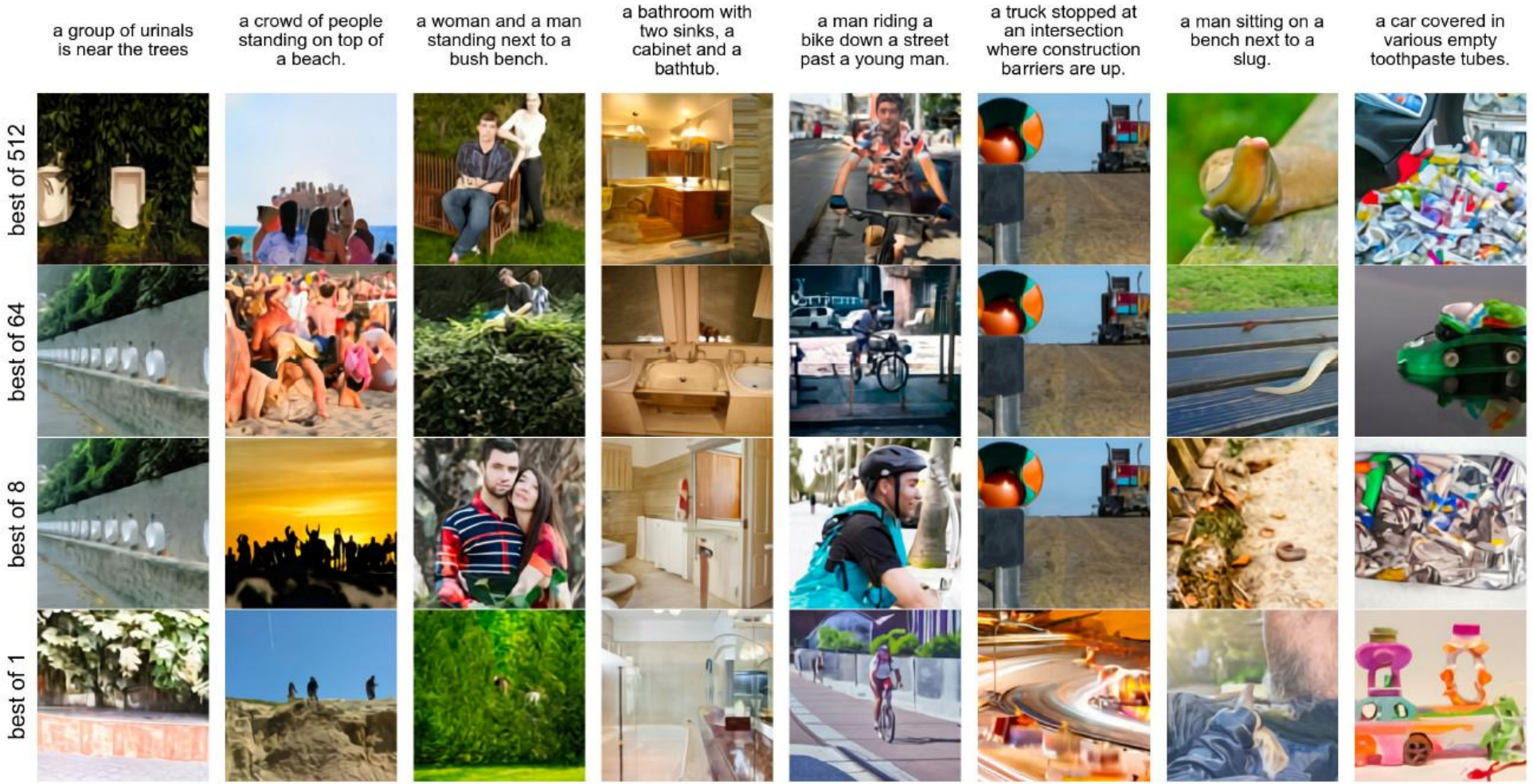
- 256개의 BPE-encoded text token 들과 1,024개의 image token들을 concat 하여 트랜스포머에 입력한다. 텍스트와 이미지 토큰에 대한 결합 확률 분포를 학습한다.



Training Procedure

- 전체 학습과정은 joint likelihood 에 대한 ELB (Evidence Lower Bound)를 maximizing 하는 것으로 볼 수 있다.
- $p_{\theta,\psi}(x,y,z) = p_\theta(x \mid y,z)p_\psi(\mathrm{y},\mathrm{z})$, x : images, y : captions, z : tokens
- 이때 lower bound 는 다음과 같이 산출된다. 기본적인 VAE 의 ELB 부등식과 유사하다.
- $\ln p_{\theta,\psi}(x,y) \geq \mathbb{E}_{z\sim q_\phi(z \mid x)}\left(\ln p_\theta(x \mid y,z) - \beta D_{KL}(q_\phi\,(y,z \mid x), p_\psi(y,z))\right)$
  - q_ϕ : dVAE Encoder (입력 이미지를 토대로 이미지 토큰 예측)
  - p_θ : dVAE Decoder (이미지 토큰을 토대로 결과 이미지 예측)
  - p_ψ : Transformer (텍스트와 이미지 토큰에 대한 결합 확률분포 예측)

Result



생성한 뒤에는 우수한 결과를 고르기 위해 CLIP (OpenAI 2021)을 사용해 주어진 text 와 k번째로 similarity 가 높은 이미지를 선택할 수 있다. (현재 $k = 1$)

Figure 6. Effect of increasing the number of images for the contrastive reranking procedure on MS-COCO captions.

Comparison

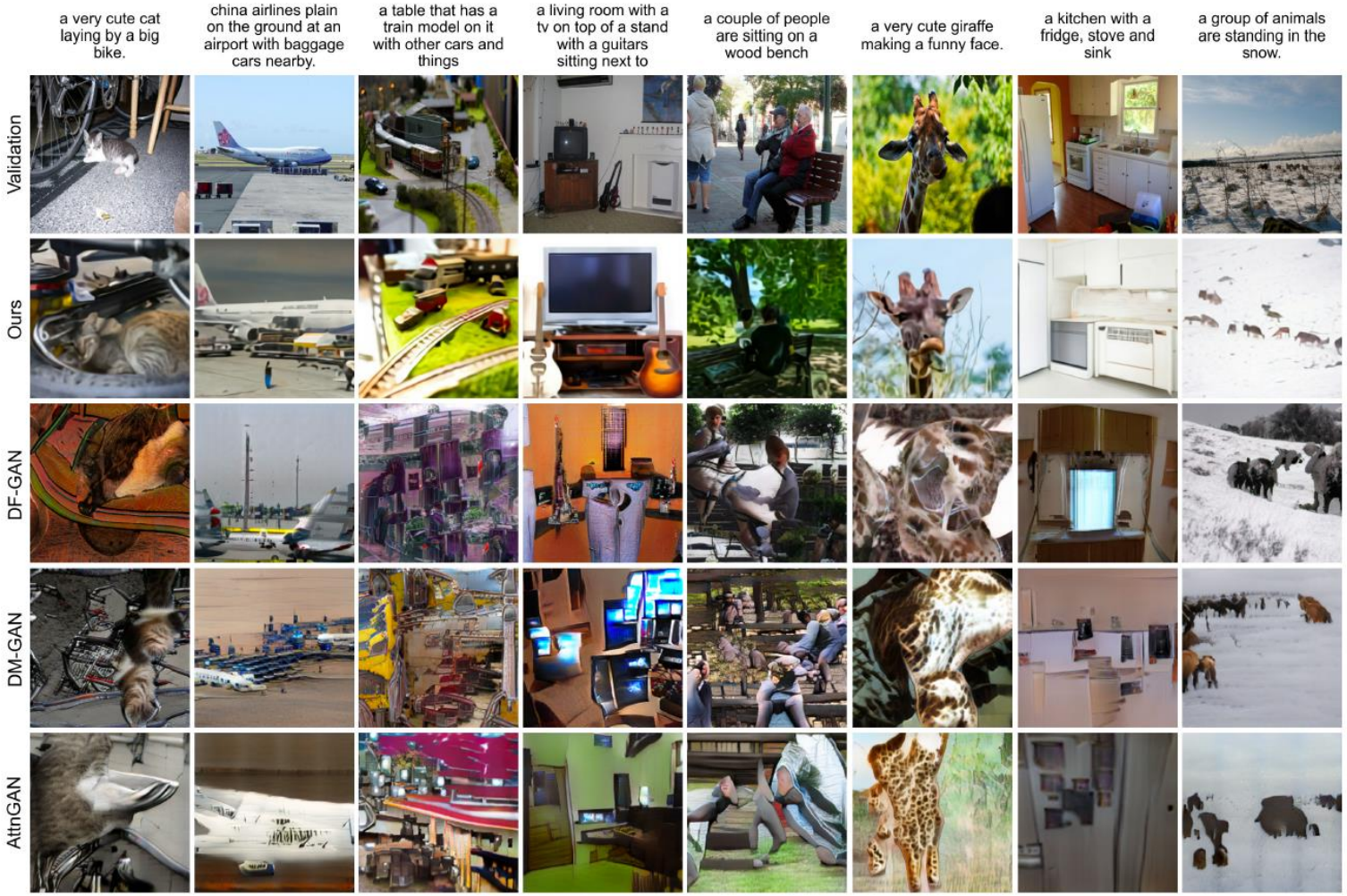


Figure 3. Comparison of samples from our model to those from prior approaches on captions from MS-COCO. Each of our model samples is the best of 512 as ranked by the contrastive model. We do not use any manual cherrypicking with the selection of either the captions or the samples from any of the models.