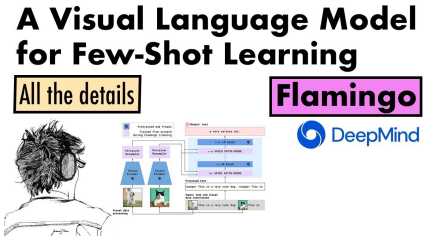
Generative Deep Learning

13 Multi Modal

Yeseul Oh

Multi Modal AI?

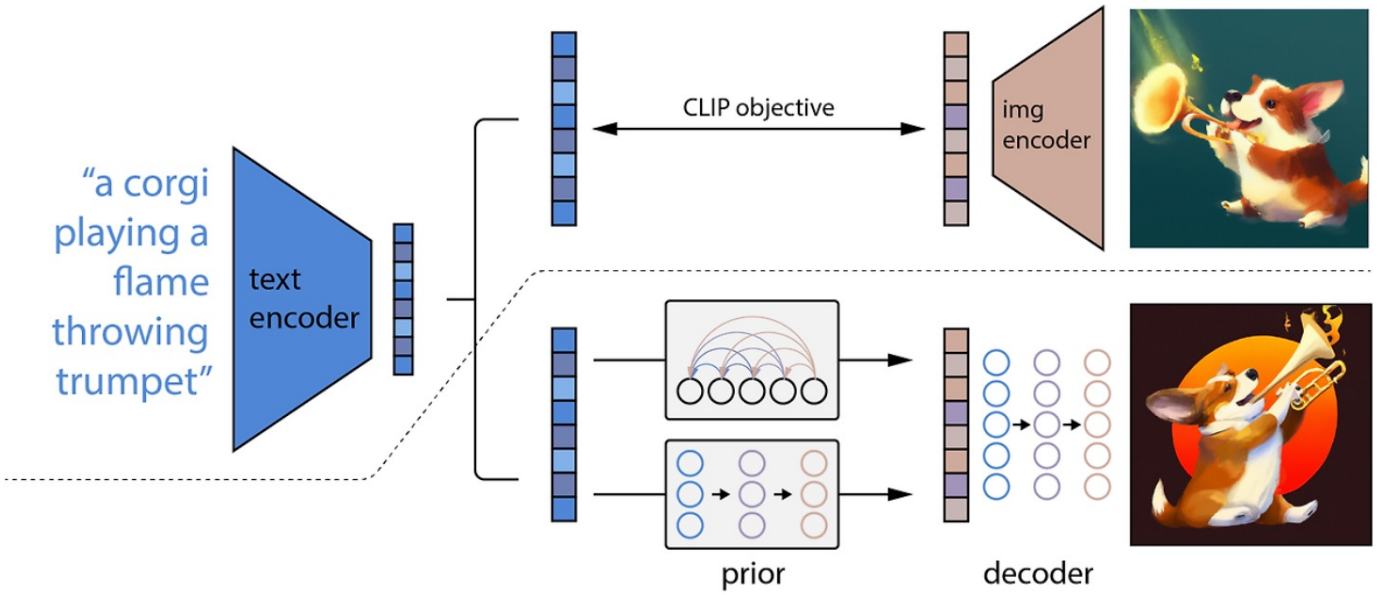
- 멀티모달 모델(Multimodal Model)은 텍스트, 이미지, 오디오, 비디오 등 다양한 유형의 데이터(모달리티)를 함께 고려하여 서로의 관계성을 학습 및 처리하는 AI



Dall-E 2

- Salvador Dalí + WALL·E
- OPEN AI
- Text to Image
- DALL:E 이전에는 GAN(Generative Adversarial Network) 모델을 이용한 접근법들이 있었으나 부자연스러운 결과가 대부분 (논리적이지 않은 개체 배치, 물체가 왜곡되는 현상 등)

Dall-E 2



A high-level overview of unCLIP

CLIP(Contrastive Language-Image Pre-training)

- Large-scale, paired dataset
  - ImageNet: 1,400만개의 텍스트설명-이미지 쌍
- Contrastive learning
  - 두 가지 다른 데이터(예: 텍스트와 이미지)의 연관성을 학습하는 방식
  - 주로 유사한 것끼리 가까이, 다른 것끼리 멀리 배치하는 학습 방식으로 사용됨
  - Cosine Similarity
    - 같은 방향 : 1
    - 직각 : 0
    - 방대방향 : -1
  - Positive / Negative pair

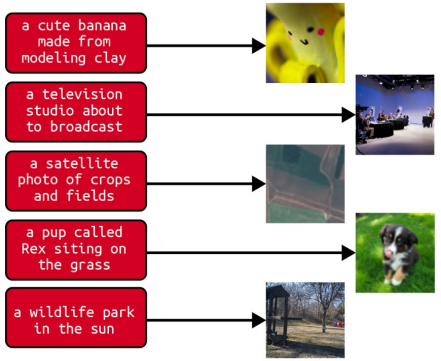


Figure 13-3. Examples of text–image pairs

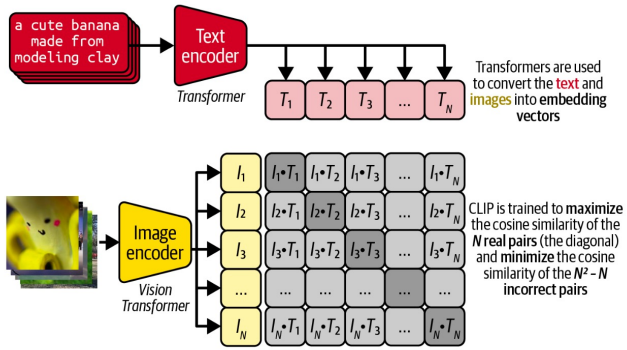


Figure 13-4. The CLIP training process

CLIP

- Prompt engineering
  - 레이블 문장 변환: label → a photo of a {label}
  - 1.3% up
- Zero-shot prediction
  - Fine-tuning이나 재학습 x
  - 학습x 새로운 이미지에 대해 레이블 예측 가능

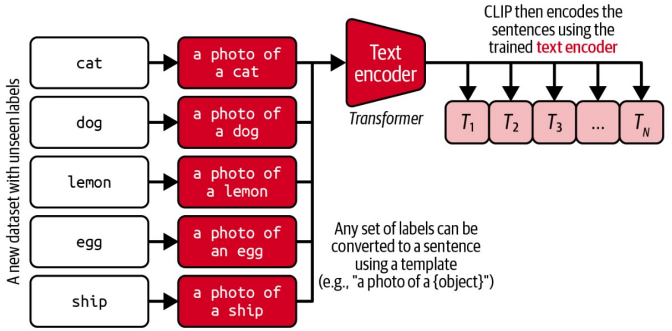


Figure 13-5. Converting labels in a new dataset to captions, in order to produce CLIP text embeddings

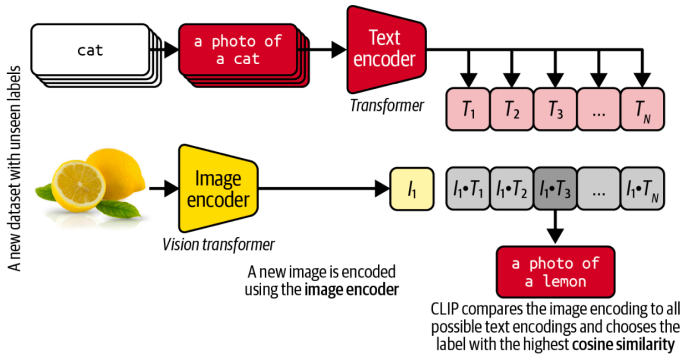


Figure 13-6. Using CLIP to predict the content of an image

CLIP

- CLIP의 수행능력
  - 다양한 이미지 변형이나 도전적인 데이터셋에서 기존 모델보다 높은 성능으로 레이블 예측 가능.
  - 일반화 능력 (아래 4)

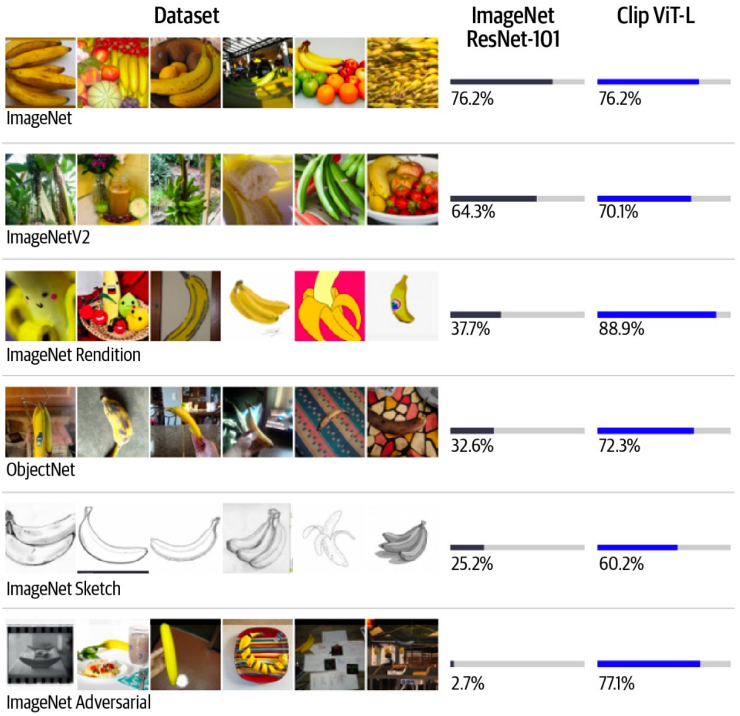


Figure 13-7. CLIP performs well on a wide range of image labeling datasets (source: Radford et al., 2021)

Prior

- 자기회귀 사전

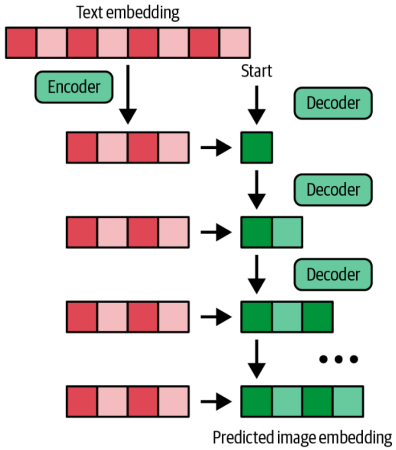


Figure 13-8. A simplified diagram of the autoregressive prior of DALL.E 2

<

- 확산 프라이어

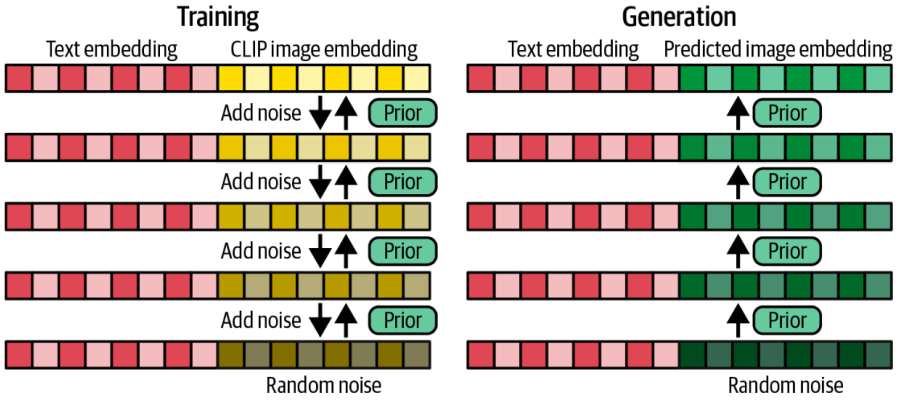


Figure 13-9. A simplified diagram of the diffusion prior training and generation process of DALL.E 2

Decoder

- GLIDE(Guided Language-to-Image Diffusion for Generation and Editing)
- DALL-E 2

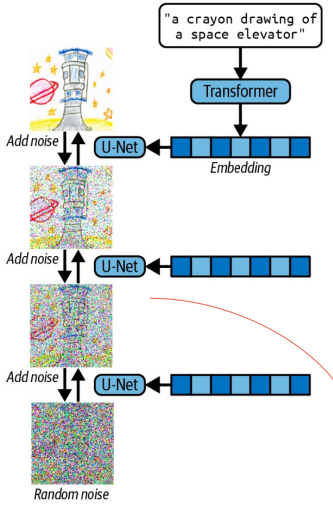


Figure 13-11. The GLIDE diffusion process

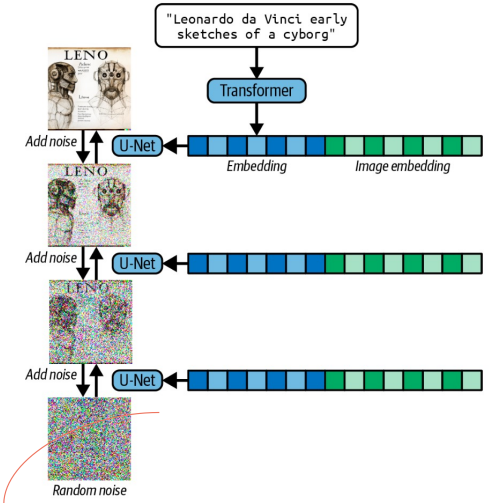
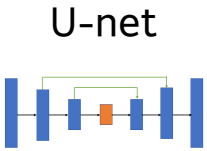


Figure 13-12. The DALL.E 2 decoder additionally conditions on the image embedding produced by the prior



Decoder

- Upsampler
  - 두개의 개별 diffusion upsampler model 사용
  - 64 x 64 → 256 x 256
  - 256 x 256 → 1024 x 1024

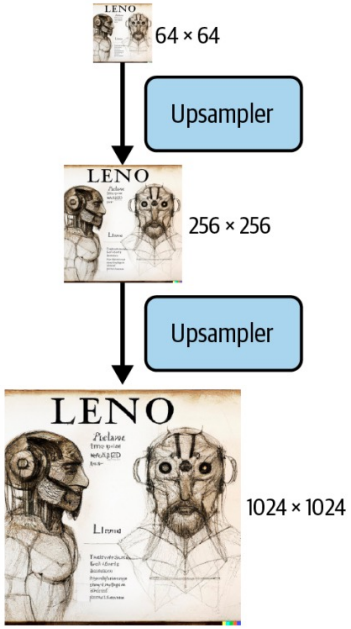


Figure 13-13. The first Upsampler diffusion model converts the image from 64 × 64 pixels to 256 × 256 pixels while the second converts from 256 × 256 pixels to 1,024 × 1,024 pixels

Imagen

- Dall-e2 출시 한달 뒤 출시
- Google Brain
- Text to Image model
- Transformer 기반 텍스트 인코더 + Diffusion 모델 디코더

Imagen

- Frozen text encoder
  - 대규모 encoder-decoder transformer인 pre-training된 T5모델
- Diffusion decoder
  - Efficient U-net
  - Upsampler
    - Diffusion model
    - 64x64
    - 1024x1024

An overview of the Imagen architecture is shown in Figure 13-17.

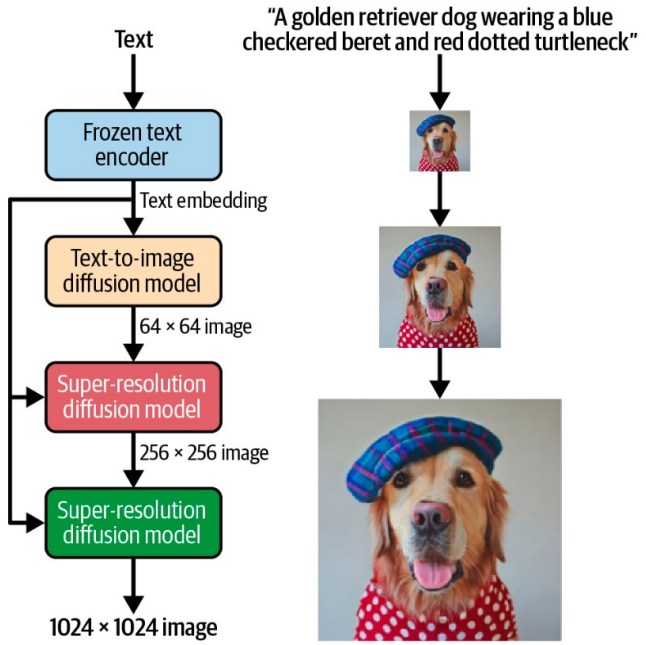
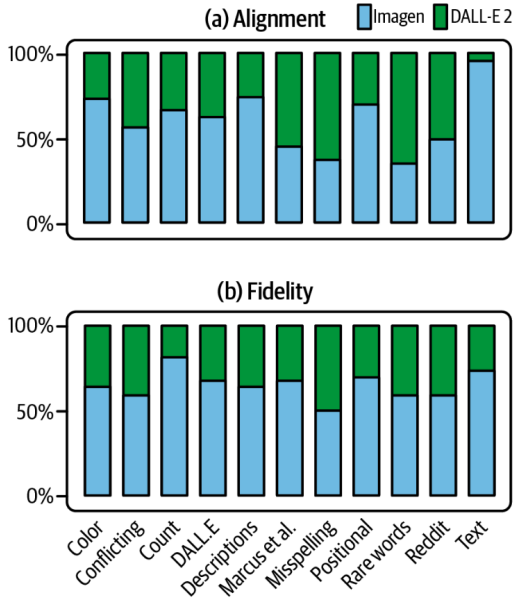


Figure 13-17. The Imagen architecture (source: Saharia et al., 2022)

Imagen

- DrawBench
  - Text to Image를 평가하는 200개의 text prompt 모음
  - Counting(지정된 개수의 객체를 생성하는 능력), Description, Text등 11가지 범주
- 두 모델을 비교 방법
  - 각 모델에 DrawBench text prompt를 전달하고 평가자에게 출력을 제공하여 두개의 측정 지표 평가
  - 정렬(일치도) / 충실도(품질)

Prompts	Category
A red colored car.	Colors
A black colored car.	Colors
A horse riding an astronaut.	Conflicting
A pizza cooking an oven.	Conflicting
One car on the street.	Counting
Two cars on the street.	Counting
…	…



대체적으로 ● > ●

Figure 13-18. Comparison of Imagen and DALL.E 2 on DrawBench across alignment and image fidelity (source: Saharia et al., 2022)

Imagen examples

Figure 13-19. Example Imagen generations (source: Saharia et al., 2022)

Stable Diffusion

- Stability AI
- Text-to-Image 모델
- 기본 생성 모델로 latent diffusion model (LDM)사용

Stable Diffusion

- 구조

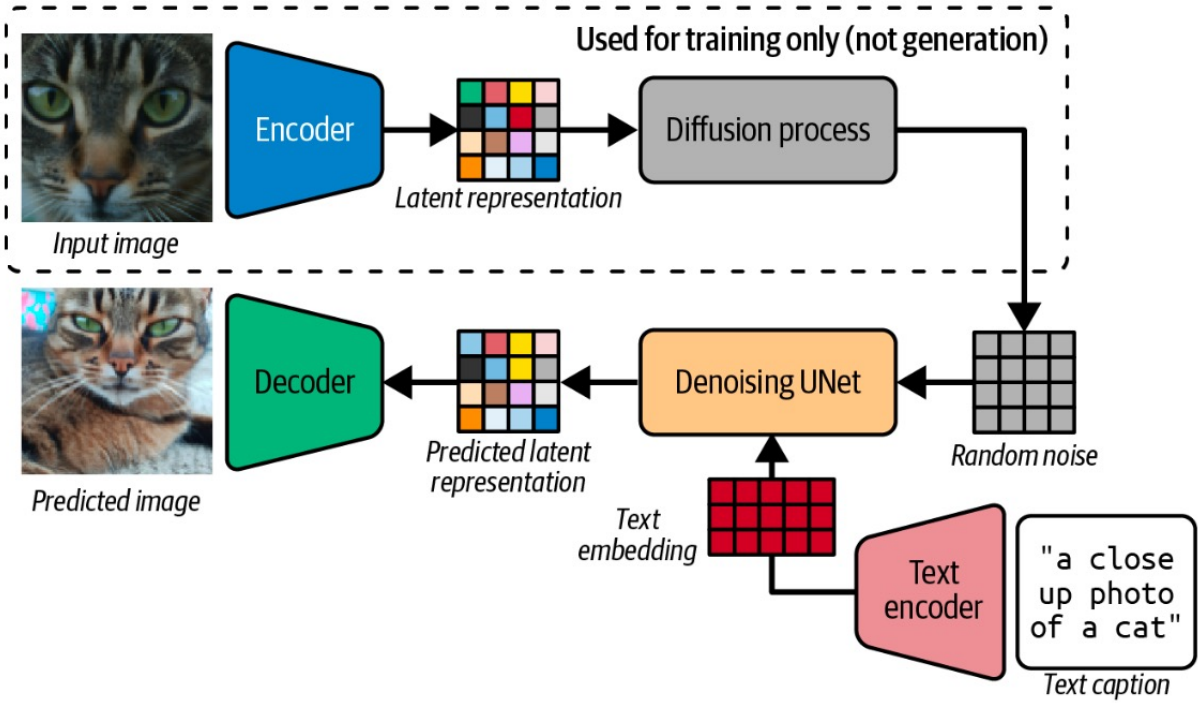


Figure 13-20. The Stable Diffusion architecture

Stable Diffusion examples

"an insect robot preparing a delicious meal"

"a high tech solarpunk utopia in the the Amazon rainforest"

"a small cabin on top of a snowy mountain in the style of Disney, artstation"

Figure 13-21. Example outputs from Stable Diffusion 2.1

Flamingo

- Google Deepmind
- Visual Language Model (VLM)
- 시각 데이터와 텍스트로 구성된 input → 텍스트 output
- 다양한 Vision-Language task에서 적은 수의 example로 학습해 fine-tuned model의 SotA에 가까운 성능

Flamingo

- 구성
  - Vision encoder
  - Perceiver resampler
  - Language model

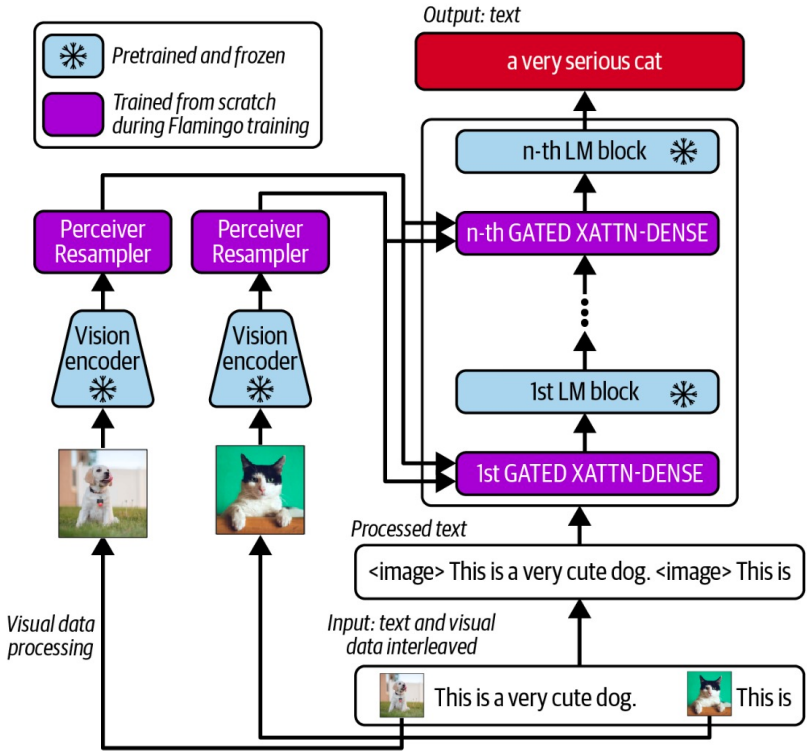
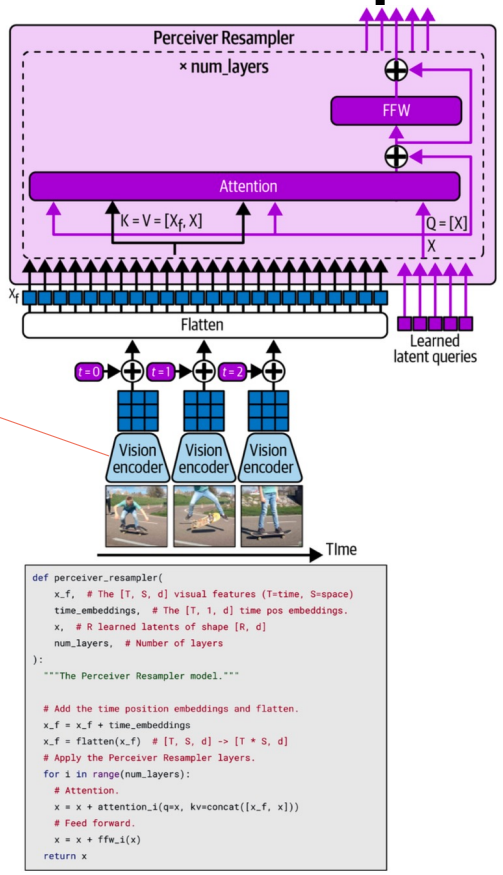


Figure 13-22. The Flamingo architecture (source: Alayrac et al., 2022)

Vision encoder / Perceiver resampler

- Vision encoder
  - 입력에 포함된 시각 데이터를 임베딩 벡터로 변환
  - Contrastive learning
  - NFNet(Normalizer-Free ResNet)
- Perceiver resampler
  - 긴 입력 시퀀스를 효율적으로 처리할 목적으로 설계
  - Cross attention




```
def perceiver_resampler(
    x_f,  # The [T, S, d] visual features (T=time, S=space)
    time_embeddings,  # The [T, 1, d] time pos embeddings.
    x,  # R learned latents of shape [R, d]
    num_layers,  # Number of layers
):
  """The Perceiver Resampler model."""

  # Add the time position embeddings and flatten.
  x_f = x_f + time_embeddings
  x_f = flatten(x_f)  # [T, S, d] -> [T * S, d]
  # Apply the Perceiver Resampler layers.
  for i in range(num_layers):
    # Attention.
    x = x + attention_i(q=x, kv=concat([x_f, x]))
    # Feed forward.
    x = x + ffw_i(x)
  return x
```

Figure 13-23. The Perceiver Resampler applied to video input (source: Alayrac et al., 2022)

Language model

X: Vision data
Y: Language data

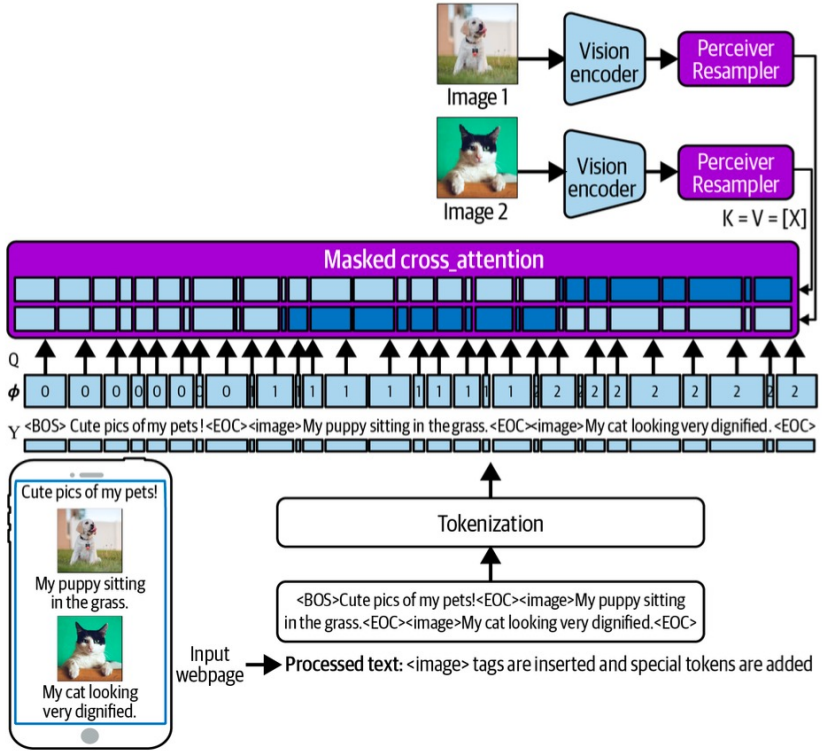


Figure 13-24. Masked cross-attention (XATTN), combining vision and text data—light blue entries are masked and dark blue entries are nonmasked (source: Alayrac et al., 2022)

- Chinchilla

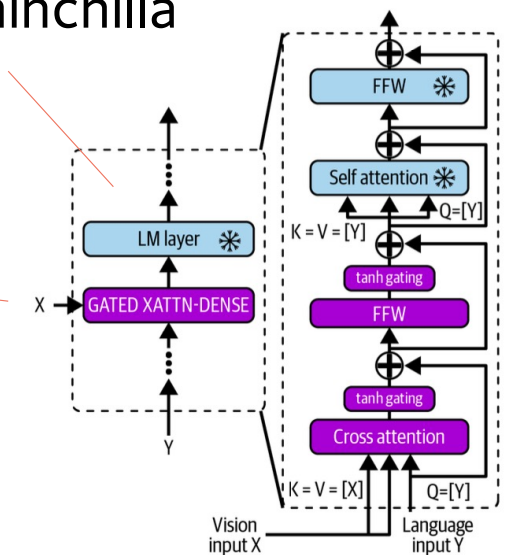


Figure 13-25. A Flamingo Language Model block, comprising a frozen language model layer from Chinchilla and a `GATED XATTN-DENSE` layer (source: Alayrac et al., 2022)

Flamingo examples

- 다양한 용도
  - 이미지 및 비디오 이해
  - 대화식 프롬프트
  - 시각적 대화

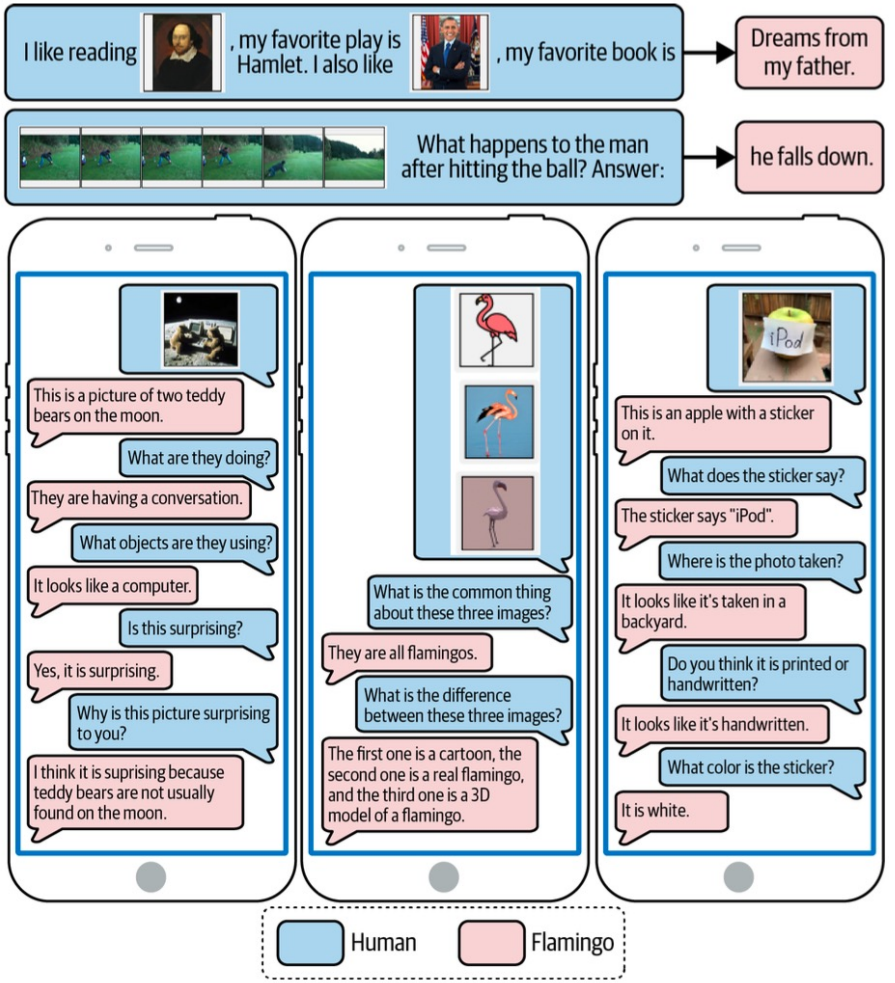


Figure 13-26. Examples of inputs and outputs obtained from the 80B parameter Flamingo model (source: Alayrac et al., 2022)

Multi Modal AI

- 멀티모달 모델(Multimodal Model)은 텍스트, 이미지, 오디오, 비디오 등 다양한 유형의 데이터(모달리티)를 함께 고려하여 서로의 관계성을 학습 및 처리하는 AI

DALL-E 2

Text to Image
CLIP

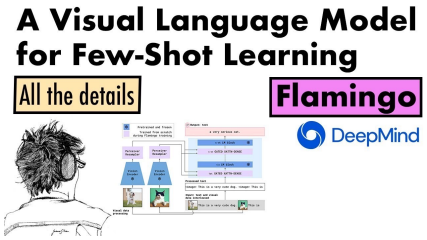
imagen

Text to Image
DrawBench

Stable diffusion

Text to Image
LDM

Flamingo



Text-Image to Text
VLM

Thank you