GLIGEN 모델을 활용한 이미지 편집 성능 향상에 관한 연구

김민창, 이재희
동서울대학교
minchang1205@naver.com, ljh7314@du.ac.kr

Research on Improving Image Editing Performance with GLIGEN Models

Kim Min Chang, Lee Jae Hee
Dongseoul Univ.

요 약

본 논문은 GLIGEN(Grounded-Language-to-Image Generation)을 활용하여 GPT(Generative Pre-trained Transformer)-3.5 Turbo 모델과 DETR(Detection Transformer) 모델을 통합해 이미지 편집을 더욱 효과적으로 수행하는 개선된 방식을 제안하였다. 본 논문에서 제안한 방식은 입력된 프롬프트(Prompt)를 GPT를 통해 이미지의 표현을 더욱 세밀하게 표현하고 DETR을 통해 이미지를 분석해 원하는 객체를 정확하게 식별하여 이미지를 편집할 수 있도록 개선하였다. DETR을 이용하여 추출한 식별된 정보(Bounding Box)와 GPT로 재구성된 프롬프트를 GLIGEN 모델에 적용함으로써 새로운 이미지 생성하는 방식이다. 본 논문에서 제안한 방식을 기존의 GLIGEN 모델과 비교 평가하기 위해 COCO2017 데이터 셋을 이용하여 FID(Frechet Inception Distance) Score 값을 계산하였다. 실험 결과, 기존 GLIGEN 모델보다 제안한 방식을 통해 생성한 이미지가 더 다양하고 세밀한 표현이 가능하다는 것을 실험 결과를 통해 증명하였다.

Ⅰ. 서 론

최근의 발전에 따라 확산(Diffusion) 모델을 사용한 이미지 생성 기술은 매우 사실적이고 다양한 이미지를 생성하는 놀라운 결과를 가져왔다. 하지만 확산 모델은 입력 프롬프트를 정확하게 해석하는데 한계를 보였다. 이러한 문제를 해결하기 위해 LLM(Large Language Model)을 활용하여 프롬프트의 이해력을 높이기 위해 확산 모델과 LLM을 결합하였다[1]. 본 논문은 생성 모델을 실용적으로 활용하기 위해 GLIGEN의 개선된 방식을 제안해 Image-to-Image 성능을 향상 시키고자 한다. 이미지 편집 시 DETR 모델을 활용하여 편집하려는 객체를 정확하게 식별하고 GPT로 편집 내용을 기술하게 하여 GLIGEN 모델을 통해 이미지 내의 객체를 원하는 객체로 변경하는 새로운 이미지 편집 방식을 제안하였다.

Ⅱ. 본론

1. DETR(Detection Transformer)

DETR은 일반적인 CNN(Convolutional Neural Network)과 트랜스포머 아키텍처를 결합하여 최종 탐지 집합을 직접 예측한다. CNN Backbone에서 추출한 Feature Map과 이미지의 위치 정보를 Encoder에 입력하여 이미지 내의 객체를 검출한다. 이 아키텍처는 출력 개수 N개로 고정하고 Bipartite Matching 통해 Grouding Truth와 비교하여 Set Prediction Problem을 직접 해결하였기 때문에 객체 검출(Object Detection) 파이프라인이 크게 단순화되었다[2]. 그림 1은 DETR의 동작 절차도이다.

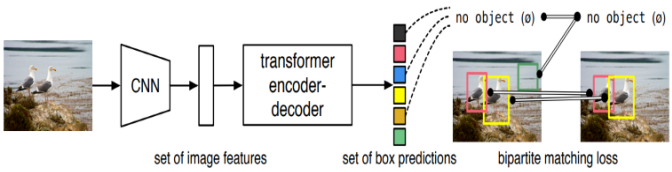

그림 1. DETR의 동작 절차도

2. Stable Diffusion

Stable Diffusion은 기존의 확산 모델의 비지각적(Non Perceptual) 한 부분을 학습하는 데에 초점을 맞춰 화소 값을 직접 예측하지 않고 VAE(Variational AutoEncoder)를 사용하였다. 먼저 CLIP(Contrastive Language-Image Pre-training) Text Encoder를 통해 입력 받은 텍스트 프롬프트를 잠재 벡터(Latent Vector) 형태로 변환한다. 이를 U-Net 에 전달하여 Text Embedding에 따라 조건화(Conditioning)된 채로 무작위 잠재 벡터(Random Latent Vector)를 n번 반복하여 Denoise 하는 과정을 거치게 된다. 이렇게 복원되어 나온 저해상도의 잠재 벡터를 VAE의 Decoder를 통해 고해상도의 그림을 만들어 주는 과정을 거치게 된다[3]. 그림 2는 Stable Diffusion의 구조도이다.

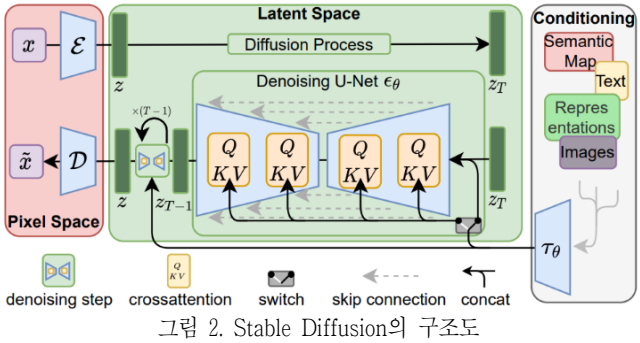

그림 2. Stable Diffusion의 구조도

3. GLIGEN(Grounded-Language-to-Image Generation)

GLIGEN은 Text-to-Image 확산 모델에 새로운 Grounding 조건부 입력을 제공하는 방법을 사용하였다. 기존 모델의 가중치는 동결 시키고 새로운 Grounding 정보를 주입하기 위해 Bounding Box와 Groudning Keypoint같은 새롭게 학습 가능한 Gated Self-Attention 레이어를 추가하였다. 이 레이어는 Visiual Token과 Grounding Token을 Concat 한 후 Attention을수행한다[4]. 그림 3은 GLIGEN의 추가 레이어의 동작도이다.

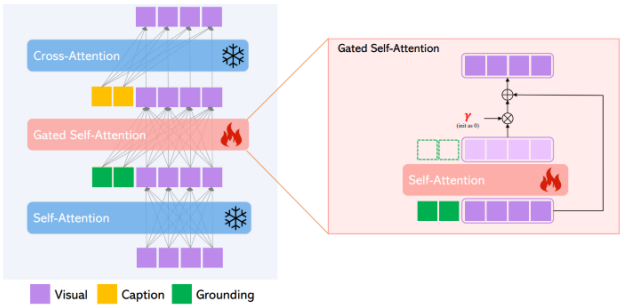

그림 3. GLIGEN의 추가 레이어 동작도

Ⅲ. 실험 결과

1. DETR을 활용하여 GLIGEN의 이미지 생성 결과

본 논문에서는 GLIGEN 모델을 활용해 DETR로 검출한 객체를 사용자가 원하는 객체 이미지로 생성하였다. 그림 4는 DETR를 통해 객체를 검출한 결과이다. 그림 5는 DETR의 검출 결과를 GLIGN에 적용해 이미지를 편집한 결과이다.

(a) 원본 이미지 (b) 객체들을 검출한 이미지

그림 4. DETR을 통해 객체를 검출한 결과

(a) 휴대폰 객체 검출 이미지 (b) 검출 객체에 대한 변경 이미지

그림 5. GLIGEN 수행 결과

2. GPT를 활용한 GLIGEN의 이미지 생성 결과

본 논문은 단순한 입력 프롬프트를 GPT를 통해 더 세밀하고 정확한 입력 프롬프트로 재구성하였다. 표 1은 GPT를 통해 재구성된 입력 프롬프트이다. GPT를 통해 입력을 요약 해주는 캡션과 해당 내용의 자세한 출력 Instruction을 사용하여 모델을 생성하게 된다. 이 때 Instruction은 유사한 의미를 가진 입력이라도 조금의 변경으로 다른 Instruction이 생성되기 때문에 사용자의 입력에 따라 다양한 이미지를 생성할 수 있게 된다.

표 1. GPT를 통한 입력 프롬프트 재구성

입력 프롬프트	GPT 생성 프롬프트
Makea duck character	Duck expressive eyes. He has a bright orange beak and feet, and a long, white feathered tail. His wings are small and often flapping excitedly.

그림 6은 DETR모델을 사용하여 객체 검출을 수행한 결과를 바탕으로 생성한 이미지이다. (a)는 단순한 프롬프트를 이용하여 GLIGEN모델로 생성한 결과이고 (b)는 GPT를 통해 재구성된 입력 프롬프트를 통해 GLIGEN 모델로 생성한 결과이다.

(a) GLIGEN의 결과 (b) 제안한 방식의 결과

그림 6. GLIGEN과 제안한 방식 간의 이미지 비교

3. 제안한 GLIGEN 개선 방식

본 논문에서 제안한 GLIGEN 개선 방식 구조는 그림 7과 같다. 기존의 GLIGEN모델에 DETR, GPT모델을 추가 조합하는 방식이다.

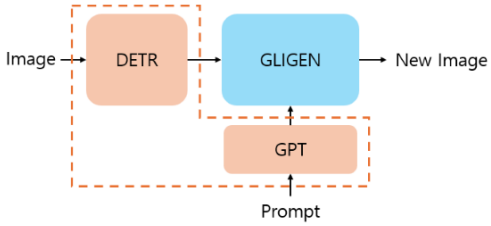

그림 7. 제안한 GLIGEN 개선 방식의 구조도

4. 성능 평가

성능 평가를 위해 MS(Microsoft)사의 COCO2017 데이터 셋의 원본 이미지 객체 중 임의의 객체를 랜덤으로 마스킹 한 후 평가를 진행하였다. 기존의 GLIGEN모델과 제안한 방식으로 생성한 이미지를 원본 이미지의 Annotation과 FID(Frechet Inception Distance) Score로 비교 평가하였다. 표 2의 결과를 보면 제안한 방식의 FID Score가 약 2.81정도 낮은 것을 알 수 있다. 제안한 방식으로 생성한 이미지가 원본 이미지와 더 유사하다는 것을 실험 결과로 알 수 있다.

표 2. 본 논문에서 제시한 방식과의 성능 비교

평가 방식 / 모 델	FID Score
GLIGEN	28.94
제안한 개선 방식	26.13

Ⅳ. 결론

생성 모델은 진취적인 발전 속도를 보였지만 실제 사용자가 활용하기에는 여전히 한계가 존재한다. 이를 극복하기 위해, 기존의 이미지를 더욱 효과적으로 편집하는 개선된 GLIGEN방식을 제안하였다. 객체 검출 방식이 아닌 객체 분할(Object Segmentation)방식을 통해 해당 객체를 더 세밀하게 검출할 수 있다면 더 우수한 이미지 편집이 가능할 것으로 판단된다. 향후 제안한 방식을 휴대폰에서 촬영한 이미지를 편집하는데 사용할 수있도록 온디바이스(OnDevice) AI로 구현하는 것을 진행하고자 한다.

참 고 문 헌

[1] Long Lian(2023)., LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models, arXiv:2305.13655
[2] Nicolas Carion(2020)., End-to-End Object Detection with Transformers., arXiv:2005.12872
[3] Robin Rombach(2022)., High-Resolution Image Synthesis with Latent Diffusion Models., arXiv:2112.10752
[4] Yuheng Li(2023)., GLIGEN: Open-Set Grounded Text-to-Image Generation., arXiv:2301.07093