디퓨전 모델에서의 전 범위적 이미지 조작을 위한 셀프 어텐션 제어 및 드래그 특징 반영 연구

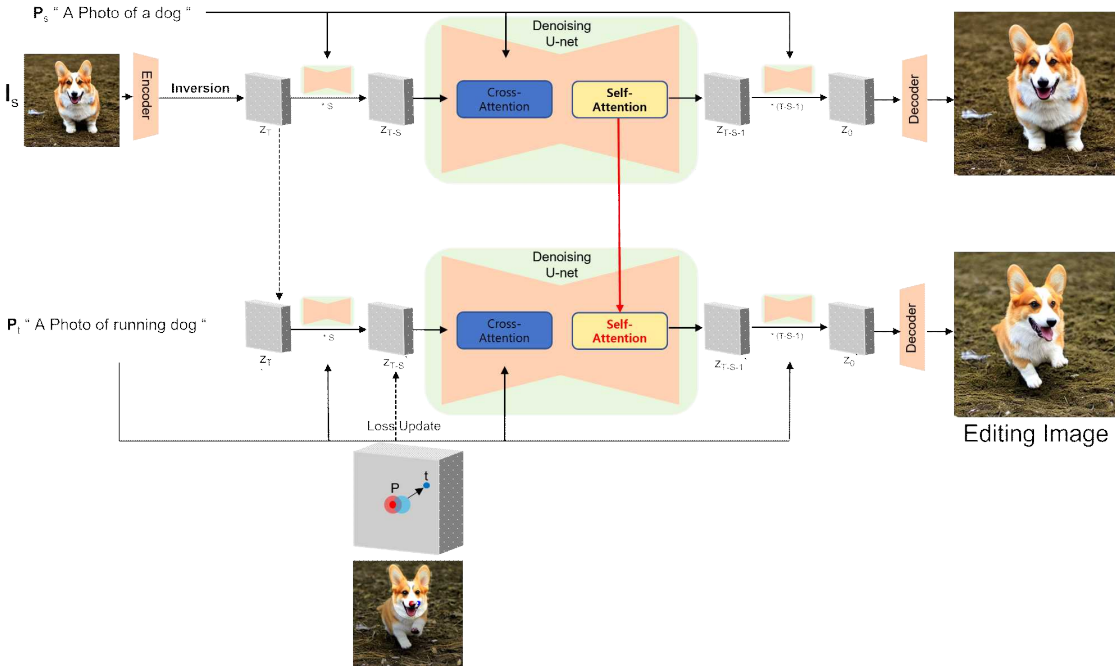
임성윤[1], 조영주[2], 이용주[2]
[1]숭실대학교 글로벌미디어학부 [2]한국전자통신연구원
tjddbs92dla@gmail.com, run.youngjoo@etri.re.kr, yongju@etri.re.kr

Image Manipulation in Diffusion Model with Drag Input using Self-Attention Control

SungYoon Lim[1], YoungJoo Jo[2], Yong-Ju Lee[2]
[1]Dept. of Global Media, Soongsil University
[2]Electronics and Telecommunications Research Institute, Korea

요 약

디퓨전 모델에서 생성한 이미지를 조작하는 기존 프롬프트 기반 방법과 포인트 기반 방법에는 각각의 단점이 있다. 프롬프트 기반은 프롬프트로만 조작이 가능하고 세세하지 못하다. 포인트 기반은 입력 이미지의 스타일을 보존하려면 파인튜닝이 필요하다. 본 논문은 디퓨전 생성 모델에 셀프 어텐션 제어와 드래그 조작을 통해, 파라미터 학습 없이, 이미지의 스타일을 보존하며 다양한 범위의 이미지 조작이 가능한 방법을 제안한다.



(그림 1) 입력 이미지를 셀프 어텐션 제어 및 드래그 조작하는 프로세스

1. 서론

최근 Stable Diffusion[1]을 이용해 생성한 이미지를 조작(Manipulation)하는 다양한 연구들이 이뤄지고 있다. 크게 두 가지로 나누면 프롬프트 기반(Prompt-Based)의 이미지 조작 방법과 포인트 기반(Point-Based)의 이미지 조작 방법이 있다. 그중 프롬프트 기반인 MasaCtrl[3]은 연구에서 제안한 Mutual Self-Attention Control이라는 방법을 통해 별도의 파라미터 학습 없이 입력 이미지의 스타일을 보전하면서 특정 단어를 추가해 이미지를 조작한다. 하지만 프롬프트 기반이기에 프롬프트로만 조작이 가능하고, 섬세한 조작은 불가능하다는 것이 단점이다. 포인트 기반의 DragDiffusion[2]은 DragGAN[4]의 Motion Supervision과 Point Tracking을 통한 이

미지 조작 방식을 디퓨전 모델(Diffusion Model)에 적용한 연구로 입력 이미지의 스타일을 보전하기 위해 파인 튜닝(Fine Tuning) 해 따로 학습이 필요하다는 단점이 있다.

본 논문은 셀프 어텐션 제어와 Motion Supervision, Point Tracking을 연구해 프롬프트 기반 조작 방법과 포인트 기반 조작 방법이 모두 가능한 새로운 방법을 제안한다. 이는 기존의 학습이 필요했던 DragDiffusion[2]의 약점과, 프롬프트 기반으로 세세한 조작이 불가능했던 MasaCtrl의 단점을 보완한다.

2. 선행연구

2.1 Stable Diffusion Model

Stable Diffusion[1]은 Latent Diffusion 모델의 일종으로 CLIP, UNet, VAE(Variational Auto Encoder)라는 세 가지 인공신경망으로 이루어져 있다. 사용자의 입력 텍스트를 텍스트 인코더(CLIP)가 토크나이저를 이용해 인코딩하여 텐서로 변환하고 이 변환된 텍스트 임베딩 텐서를 UNet에 전달한다. UNet은 스케줄러(Scheduler)에 따라 반복하여 랜덤 잠재 벡터(random latent vector)를 디노이징하며 이때 전달 받은 텍스트 임베딩에 의해 조건화(Conditioning)된다. 디노이징된 잠재 벡터는 최종적으로 VAE의 디코더를 통하여 이미지로 변환된다.

2.1 MasaCtrl

Prompt-to-Prompt[5]는 생성된 이미지의 공간 레이아웃과 기하학적인 요소는 cross-attention map에 따라 달라진다는 것과 이미지의 구조는 diffusion process의 초기 step에서 이미 결정되어 있음을 밝혔다. MasaCtrl[3]은 이를 통해 초기 step 이후 나머지 디노이징 step에서 원본 이미지의 셀프 어텐션 레이어의 (Self Attention Layer)의 Key, Value값을 수정된 타겟 프롬프트로 합성된 Key, Value값과 치환하여 입력 이미지의 스타일을 보존하며 타겟 프롬프트로 조작된 이미지를 생성한다. 하지만 타겟 프롬프트만으로 조작하기에는 세세한 조작이 불가능하다는 점과, 입력 이미지에서 포함되지 않은 내용은 조작할 수 없다는 단점이 있다.

2.2 DragDiffusion

포인트 기반 편집을 위해 DragGAN[4]에서 제안된 Motion Supervision과 Point Tracking이라는 방법을 도입한다. Motion Supervision의 목적 함수는 아래의 수식과 같다.

$$L(z_t^k) = \sum_{i=1}^{n} \sum_{q \in \Omega(h_i^k, r_1)} \| F_{q+d^i}(\hat{z}_t^k) - sg\left(F_q(\hat{z}_t^k)\right) \|_1 + \lambda \| (\hat{z}_{t-1}^k) - sg\left(F_q(\hat{z}_{t-1}^0)\right) \odot (1-M) \|_1,$$

k번째 motion supervision 반복에서 n개의 핸들 포인트(handle point)는 $h_i^k = (x_i^k, y_i^k : i = 1, \cdots, n)$, 타겟 포인트(target point)는 $g_i = (\tilde{x}_i, \tilde{y}_i : i = 1, \cdots, n)$로 표시한다. 입력 이미지는 z_0, t번째 step의 latent는 z_t가 된다. $F(z_t)$는 z_t가 입력일 때 UNet의 14번째 블록의 특징 맵이다. 또한 특정 위치 $h_i^k = (x_i^k, y_i^k)$의 feature 벡터를 $F_{h_i^k}(z_t)$로 표시한다. M은 이진 마스크, $\Omega(h_k^i r_1)$는 (h_i^k) 를 중심으로 한 변의 길이가 $2r_1+1$인 정사각 모양의 패치다. sg는 stop gradient로 특정 항이 역전파를 하지 않는 것을 표현한다. 마지막으로 $F_{q+d_i}(\hat{z}_t^k)$는 $q+d_i$의 요소가 정수가 아니기 때문에 양선형 보간을 통해 얻는다. 최종적으로 목적 함수를 최소화하기 위해 경사 하강법을 사용하여 아래와 같이 업데이트한다.

$$\hat{z}_t^{k+1} = \hat{z}_t^k - \eta \cdot \frac{\partial L \hat{z}_t^k}{\partial \hat{z}_t^k}$$

Motion Supervision의 업데이트가 $\hat{z}_t^k$를 변경하기 때문에, 핸들 포인트의 위치도 이동을 해야한다. 따라서 잠재 변수를 최적화한 후 핸들 포인트를 업데이트하기 위해 Point Tracking을 수행해야 한다. 핸들 포인트의 새 위치는 아래 수식과 같이 패치 내에서 최근접 이웃 검색(Nearest Neighbor Search)로 설정한다.

$$h_i^{k+1} = \arg \min_{q \in \Omega(h_i^k, r_2)} \| F_q(\hat{z}_t^{k+1}) - F_{h_i^k}(z_t) \|_1$$

3. 본론

3.1 학습 없는 드래그 이미지 조작

본 연구에서는 셀프 어텐션 제어와 포인트 기반의 조작 방법을 통해, 별도의 학습이 없이 스타일을 보존한 포인트 기반의 조작 방법을 제안한다. 전체적인 과정은 (그림 1)과 같다. 입력 이미지를 T step만큼 DDIM Inversion을 진행해 노이즈 z_T와 동일한 노이즈 z_T'를 생성한다. 디노이징 프로세스에

서는 DDIM Sampling을 통해 초기 step T-S에서 이미지 구조가 결정되어 있음을 전제로 Noise T-S 번째 step의 z_{T-S}'를 Motion supervision loss로 업데이트한다. T-S step 이후에서는 업데이트된 z_{T-S}'와 입력 이미지의 z_{T-S}를 합쳐서 (concatenate) 셀프 어텐션 제어로 z_{T-S}의 Key, Value 값을 z_{T-S}'의 Key, Value값에 대입한다. DDIM Sampling 단계를 거친 z_0'를 VAE의 Decoder를 거쳐 이미지로 변환한다.

3.2 전 범위적 이미지 조작

본 논문에서는 프롬프트 기반의 조작과 포인트 기반의 조작을 동시에 진행 가능한 방법을 제안한다. 또한 입력 이미지를 조작하거나, Stable Diffusion으로 생성한 이미지 둘 다 조작이 가능하다. 생성 이미지 조작은 랜덤 잠재 코드(random latent code) z_T에 원본 프롬프트와 수정 프롬프트를 임베딩(embeding)하고, 디노이징 단계를 거치며, 셀프 어텐션 제어를 통해 원본 프롬프트로 생성했을 때의 이미지의 스타일을 보존한 조작된 이미지로 생성한다. 그 후 수정된 이미지의 특정 T-S step에서의 z_{T-S}를 Motion Supervision 업데이트한 후 나머지 step만큼 디노이징 한다. 입력 이미지를 제공할 경우, 무작위 잠재 코드와 원본 프롬프트 대신 원본 이미지가 Inversion되어 노이즈 z_T를 만들어 같은 과정을 수행한다.

3.3 DDIM Sampling Step 보정

DDIM의 Inversion에서는 T step 만큼 반복해 입력 이미지에 노이즈를 더해서 T step의 노이즈 z_T를 생성한다. 기존의 DDIM의 Inversion의 Forward process에서는 입력 이미지 z_0에서 z_1로 노이즈를 더하는 첫 번째의 step이 누락 되어 있다. 따라서 이 과정을 추가하여 Sampling Step을 보정했다.

4. 실험 결과

4.1 Real Image Editing

입력 이미지를 제공하고, 셀프 어텐션 제어를 통해 프롬프트 기반의 조작을 진행했다. 그 후 포인트 기반의 조작(고개를 오른쪽으로 회전)을 진행했다. 결과는 (그림 2)와 같다.

(그림 2) 원본 이미지 프롬프트 및 포인트 기반 조작

4.2 T2I Generate Image Editing

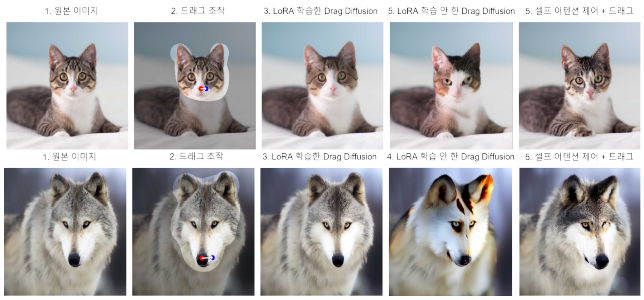
원본 프롬프트와 수정 프롬프트로 생성된 이미지를 셀프 어텐션 컨트롤을 통해 프롬프트 기반의 조작을 진행했다. 그 후 Drag Update를 통해 세부적인 조작(고개를 오른쪽으로 회전)을 시도했다. 결과는 (그림 3)과 같다.

(그림3) 생성 이미지 프롬프트 기반 및 포인트 기반 조작

4.3 DragDiffusion과의 비교

DragDiffusion[2]과 본 연구의 드래그 방식의 조작의 결과물을 비교한다. DragDiffusion[2]과 똑같은 조건에서 실험을 진행한다. 모델은 stable-diffusion 1.5v, DDIM Inversion과 Sampling, seed는 42, 셀프 어텐션 제어 step 10, step 10에서 드래그 업데이트를 진행한다.

그림[4]와 같이 LoRA 학습을 진행하지 않은 것과 비교했을 때, 학습 없이 스타일이 잘 보존된 것을 확인할 수 있다. LoRA 학습을 진행한 Drag Diffusion은 과적합으로 인해 변화가 적은 것에 비해 본 연구의 결과물은 스타일이 적용되면서 변화가 뚜렷하다.



(그림 4) Drag Diffusion과 본 연구 비교

4.4 셀프 어텐션 제어 step과 드래그 업데이트 step 간의 차이 비교

셀프어텐션 제어 Astep을 4 ~ 8, 드래그 업데이트 step을 4 ~ 8로 수치를 조절하며 실험을 했다. 모델은 stable-diffusion-1.5v, Inference timestep은 50, 학습율은 0.01로 설정했다. 드래그 업데이트 step이 낮을수록, 변화의 정도가 크고, 8step 이후부터는 고착화 되었다. 셀프 어텐션 제어 step은 증가할수록 타겟 프롬프트로 생성한 이미지의 구조를 강하게 반영됐다.

<표 1> 셀프어텐션 제어 step과 드래그 업데이트 step간의 차이 비교

5. 결론

본 연구에서 제안하는 방법을 통해 별도의 파라미터 학습 없이 이미지의 조작과 스타일 보존이 가능함을 확인하고, 다양한 이미지 조작 방법을 제안했다. 이를 통해 사용자들은 입력 이미지 및 생성된 이미지를 본인의 기호에 맞게 수정할 수 있다. 더불어 본 연구에서 제시한 방법론은 다양한 T2I(Text to Image) 모델에 활용될 수 있어 높은 활용성을 제공한다. 향후 연구에서 T2I뿐 만 아니라 I2V(Image to Video), T2V(Text to Video) 등 다양한 생성형 AI에도 적용하는 방법을 연구하고자 한다.

Acknowledgement

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2022-00187238, 효율적 사전학습이 가능한 한국어 대형 언어모델 사전학습 기술개발)

참고문헌

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer “High-Resolution Image Synthesis with Latent Diffusion Models” (CVPR), 2022, pp. 10684-10695

[2] M Cao, X Wang, Z Qi, Y Shan, X Qie, Y Zheng “MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing” arXiv preprint arXiv:2304.08465, 2023

[3] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, Song Bai “DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing” arXiv:2306.14435 [cs.CV]

[4] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, Christian Theobalt “Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold” SIGGRAPH '23: ACM SIGGRAPH 2023 Conference ProceedingsJuly 2023Article No.: 78Pages 1 - 11

[5] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, Daniel Cohen-Or “Prompt-to-Prompt Image Editing with Cross Attention Control” arXiv:2208.01626 [cs.CV]