CLIP 기반 복수 물체 이미지 스타일 변환

박상영, 이민식
한양대학교
parksy1029@hanyang.ac.kr, mleepaper@hanyang.ac.kr

Multi-object image style transfer based on CLIP

Sangyoung Park, Minsik Lee
Hanyang Univ

요 약

스타일 변환은 이미지, 음성 혹은 비디오의 주된 형태는 유지하되 전체적인 스타일을 원하는 방향으로 변경하는 기법이다. 최근 방대한 이미지와 텍스트 쌍을 통해 멀티모달 임베딩 공간을 학습한 CLIP을 이용하여 텍스트를 통해 이미지 스타일을 지정하는 다양한 방법들이 소개되었다. 하지만 이들 대부분은 이미지 전체를 하나의 텍스트 조건을 이용하여 스타일을 변환하여 이미지 내의 다수의 물체를 각각 다른 스타일로 변환하고 싶은 경우에는 적용하기 어렵다는 문제가 있다. 이를 위해 한 이미지 내에서 원하는 물체를 분할한 후 각각의 물체에 대해 원하는 스타일로 변환하는 과정을 제안하며, 추가적인 계산 과정을 통해 더 자연스러운 스타일 변환을 수행한다. 또한 이 모든 과정은 텍스트 입력을 통해 이루어지므로 사용자 친화적인 사용이 가능하도록 하였다.

Ⅰ. 서 론

이미지 스타일 변환은 computer vision의 중요한 응용 과제 중 하나로서 content image와 style image를 입력으로 하여 content image의 object와 같은 주된 형태는 유지한 채 style image의 스타일을 이용하여 이미지의 뉘앙스를 변경한다. 이는 변경하고자 하는 스타일의 이미지를 가지고 있어야 한다는 점에서 실용성이 떨어질 수 있다.

따라서 최근 연구에서는 CLIP [1]을 이용하여 style 조건을 text로 입력해 이미지 스타일을 변환하는 연구들이 진행되고 있다 [2, 3]. CLIP은 약 4억 개의 image, text pair를 통해 multi-modal embedding 공간을 학습하였으며, text를 이용한 image segmentation, style trasnfer, image classification 등 다양한 downstream task에 이용되고 있다.

하지만 위와 같은 방법들은 input image 전체를 스타일 변환하기 때문에 한 이미지 내에서 서로 다른 스타일 변환을 원하는 object가 다수일 때는 적용하기 힘들다는 단점이 있다. 따라서 본 논문에서는 CLIP 기반의 segmentation 모델과 스타일 변환 모델을 사용하여 다수의 object를 스타일 변환하는 과정을 제안한다.

CLIP 기반 모델을 사용하여 입력 이미지에서 스타일 변환을 위한 object를 지정하는 것과 해당 object의 변환하고자 하는 스타일 모두를 text로 입력받기 때문에 직관적인 사용이 가능하다. 그림 1에 본 논문의 스타일 변환 과정을 그림으로 나타내었다. Segmentation model은 학습되어 있는 모델을 이용하여 inference 하고 style transfer model은 하나의 object 마다 200 iteration 동안 학습한다.

Ⅱ. 본론

본 논문에서는 CLIPSeg [4]를 이용하여 segmentation을 진행하고, CLIPstyler [3]를 이용하여 스타일 변환을 진행한다. 그림 1과 같이 스타일 변환을 원하는 object를 text로 입력하여 segmentation confidence

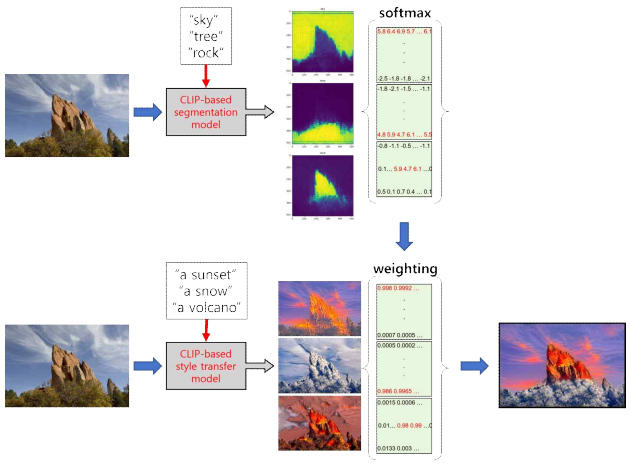

그림 1. 본 논문의 스타일 변환 과정. Segmentation을 진행 후 해당 결과를 이용하여 스타일 변환 된 이미지에 weighted sum을 적용하여 최종 이미지를 출력한다.

map을 얻는다.

Confidence map의 dimension은 $D_{seg} \in R^{N \times 1 \times 512 \times 512}$로서 N은 입력 text의 개수이며, 512는 각각 이미지의 세로와 가로 크기를 의미한다. 해당 confidence map의 score는 음수와 양수를 포함하는 normalize 되어있지 않은 상태로 출력되기 때문에 이를 softmax function을 하나의 pixel마다 object-wise로 적용하여, 각 픽셀이 어떤 object에 해당하는지에 대한 확률값을 구한다.

Segmentation object에 대해 스타일 변환을 하기 위해 CLIPstyler의 입력 순서대로 style text를 입력한다. CLIPstyler는 Adam optimizer를 통해 200 iteration 동안 학습되며, CLIP model은 고정시킨 상태로 U-net 기반의 스타일 변환 네트워크만을 학습시킨다. CLIPstyler의 output

그림 2. 스타일 변환 결과, 왼쪽과 오른쪽 이미지는 각각 원본과 스타일 변환 완료된 이미지이다. 화살표 왼쪽의 text는 segmentation 시 입력되는 text이며 오른쪽의 text는 해당 object를 스타일 변환하기 위한 text를 의미한다.

dimension은 $D_{sty} \in R^{N \times 3 \times 512 \times 512}$ 로서 N과 512는 segmentation model과 마찬가지로 각각 style text의 개수, 이미지의 크기를 의미하며, 3은 R, G, B 3개의 채널을 의미한다.

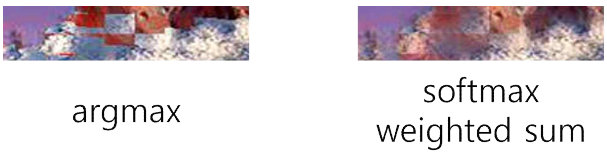

그림 3. segmentation confidence score가 높은 object의 pixel로 대체한 결과와 논문에서 사용된 softmax를 통한 weighted sum의 비교.

각 object의 부자연스러운 경계면을 해결하기 위하며 softmax를 적용한 D_{seg}의 값을 weighted sum 하여 smoothing 시킨다. 그 후, 원본 이미지의 크기로 resize 하여 최종 output image를 구한다. 그림 2에 스타일 변환 결과를 나타내었다. 첫 번째 열에는 real image에 대한 결과를, 두 번째 열에는 illustration image에 대한 결과를 나타내고 있으며, 두 경우 모두 segmentation object 별로 성공적으로 변환된 것을 확인할 수 있다.

추가적으로 그림 3을 통해 단순히 argmax function을 이용하여 segmentation confidence score가 큰 object의 pixel로 대체했을 때 보다 weighted sum을 이용한 smoothing을 이용하면 경계 부분 또한 자연스럽게 변환된 것을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 CLIP 기반의 segmentation model과 style transfer model을 이용한 multi-object style transfer 방법을 제안한다. 모든 과정은 text input을 통해 이루어져 사용자 친화적으로 사용할 수 있도록 하였다. 또한 스타일 변환의 결과를 weighted sum을 통해 smoothing 시켜 더욱 자연스러운 결과를 얻을 수 있도록 하였으며, 정성적인 실험을 통해 real과 illustration image 모두에 대해 스타일 변환이 성공적으로 이루어짐을 확인하였다.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA))

참 고 문 헌

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. .Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748 - 8763. PMLR, 2021.

[2] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In Pro- ceedings of the IEEE/CVF International Conference on Computer Vision, pages 2085 - 2094, 2021.

[3] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18062 - 18071, 2022.

[4] Timo L¨uddecke and Alexander Ecker. Image segmentation using text and image prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7086 - 7096, 2022.