Proposal Title: Text-guided Image blending

Name: 공대현

Abstract

- About 800 words in Korean or 1200 words in English

- Representative figure for explaining the key idea of the proposal should be included.

최근 딥러닝을 이용한 Generative models은 4가지로 분류할 수 있다. Adversarial training을 이용한 GAN, variational lower bound maximizing을 이용한 VAE, Invertible transform of distributions를 이용한 Flow-based models 그리고 점진적으로 gaussian noise를 추가하고 그것을 reverse하는 것을 학습함으로써 latent space를 학습하는 diffusion model이 있음.

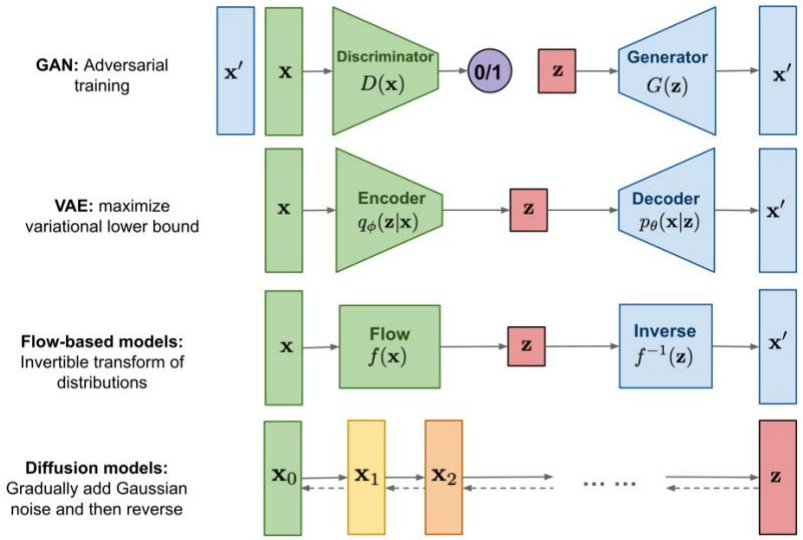

Figure 1 네 종류의 생성모델

Diffusion model과 CLIP의 pretrained text representation로 image generation, image inpainting을 하는 GLIDE[1] 라는 논문이 최근(3월 8일)에 arxiv에 업로드 되었고, FID score에서 SOTA의 성능을 보여주었다. 우리는 이 GLIDE 혹은, 이 방식을 활용하여, CVPR 2021에서 새롭게 소개된 분야인 image blending[2]에 text representation을 추가하여 기존 논문의 blending보다 좋은 성능을 얻는 것이 목적임

Image blending[2]은 아래 그림과 같이 UNet과 LSTM encoder, decoder를 이용한 네트워크 구조로 서로 다른 두 이미지를 연결하는 가운데 부분을 생성하여 최대한 자연스럽게 두 이미지를 잇는 것이 목적임.

Image blending은 GT가 없기 때문에, PSNR, SSIM 등의 refrence가 필요한 metric을 사용하지 않고, FID, IS 등의 Non reference score를 사용함. 이 논문[2]에서 다양한 image inpainting, outpainting 논문을 그대로 가져와서 성능을 비교해본 결과 이 네트워크 방식의 성능이 가장 좋았음.

Figure 2. text-gudied diffusion model의 output

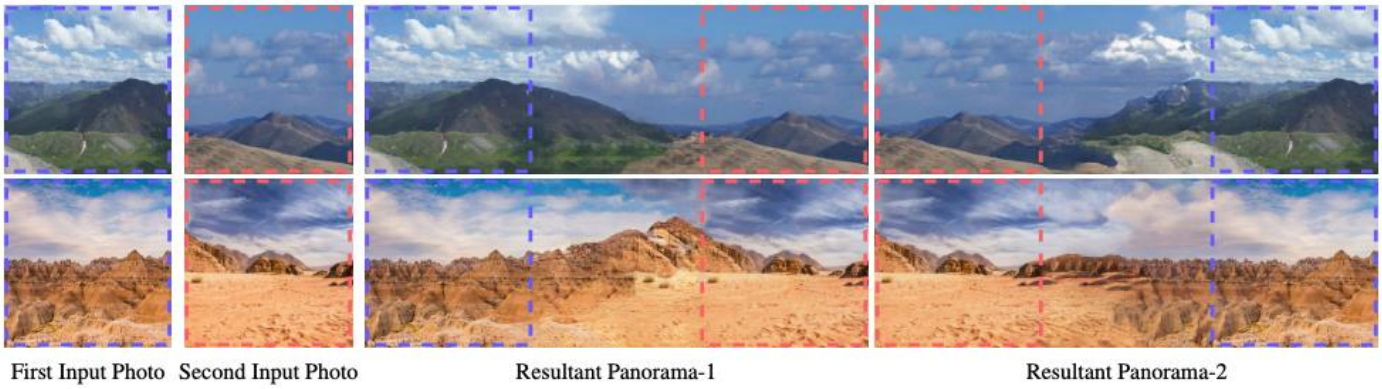

Figure 3. image blending model의 output

FID 성능이 가장 좋은 text-guided diffusion model과 image blending model을 적절히 조합하여, text-guided image blending network를 만드는 것이 목적임. Contribution으로는 image blending을 위해 합쳐질 두 대상 이미지의 caption을 적절히 조합하여, 가운데 부분을 위한 좋은 text-guide를 생성하여, FID 성능을 높이는 것임.

Caption은 CLIP-pretrained text encoder를 이용한 captioning network[3]를 이용할 것임.

Caption blending은 rule-based로 할 지, text blending network(?)를 이용할 지 고민 중

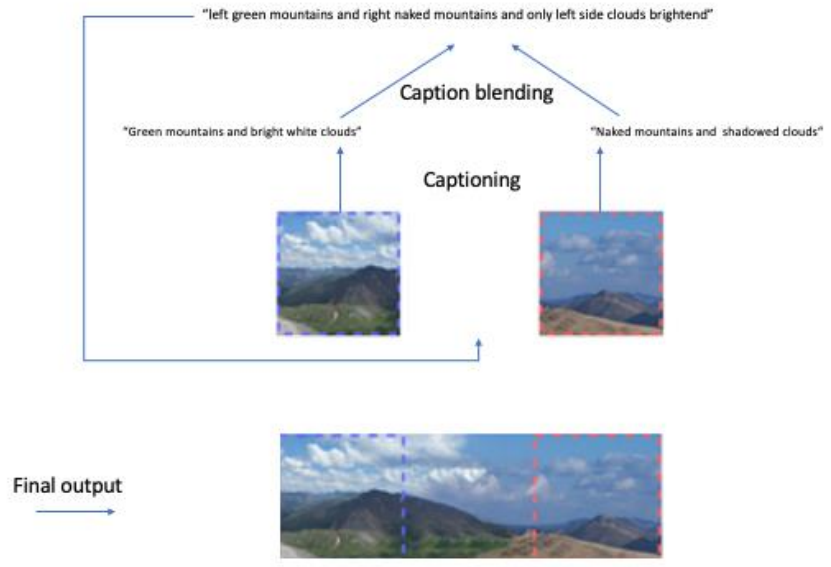

Figure 4. Proposal concept description

- Reference

[1] GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models (arxiv 2022)

[2] Bridging the Visual Gap: Wide-Range Image Blending (CVPR 2021)

[3] ClipCap: CLIP Prefix for Image Captioning (arxive 2021)

Expected results

- Image만을 이용한 Unet기반의 [2]보다 FID, IS score가 더 잘 나오고, text-guided + text blending의 효과가 주 원인으로 작용하는 것