최신 Text-to-Image 생성 모델의 동향

한국과학기술원 ▌이상현·윤주열·박민호·형준하·주재걸

1. 서 론

2021년의 DALL-E [7] 생성 모델의 등장 이래로, 2022년은 대규모 데이터를 이용한 text-to-image 생성 모델의 시대라 해도 과언이 아닐 정도로 다양한 모델과 연구들이 시시각각 나오고 있다. 특히 최근 그림1의 Stable AI사에서 오픈소스로 공개한 Text-to-Image 생성 모델[11]은 AI전문가뿐만이 아니라 일반인들도 쉽게 접근이 가능하여, text-to-image 생성 모델의 큰 관심을 일으키고 있다.

최신 모델들의 이런 눈에 띄는 성과는 기존에 비해 더 커진 데이터셋 규모, 그리고 최신 생성 모델 기법 적용이라는 두 가지 측면에서 기인한다. 기존의 데이터 수집 방법인 사람이 이미지의 캡션을 직접 레이블

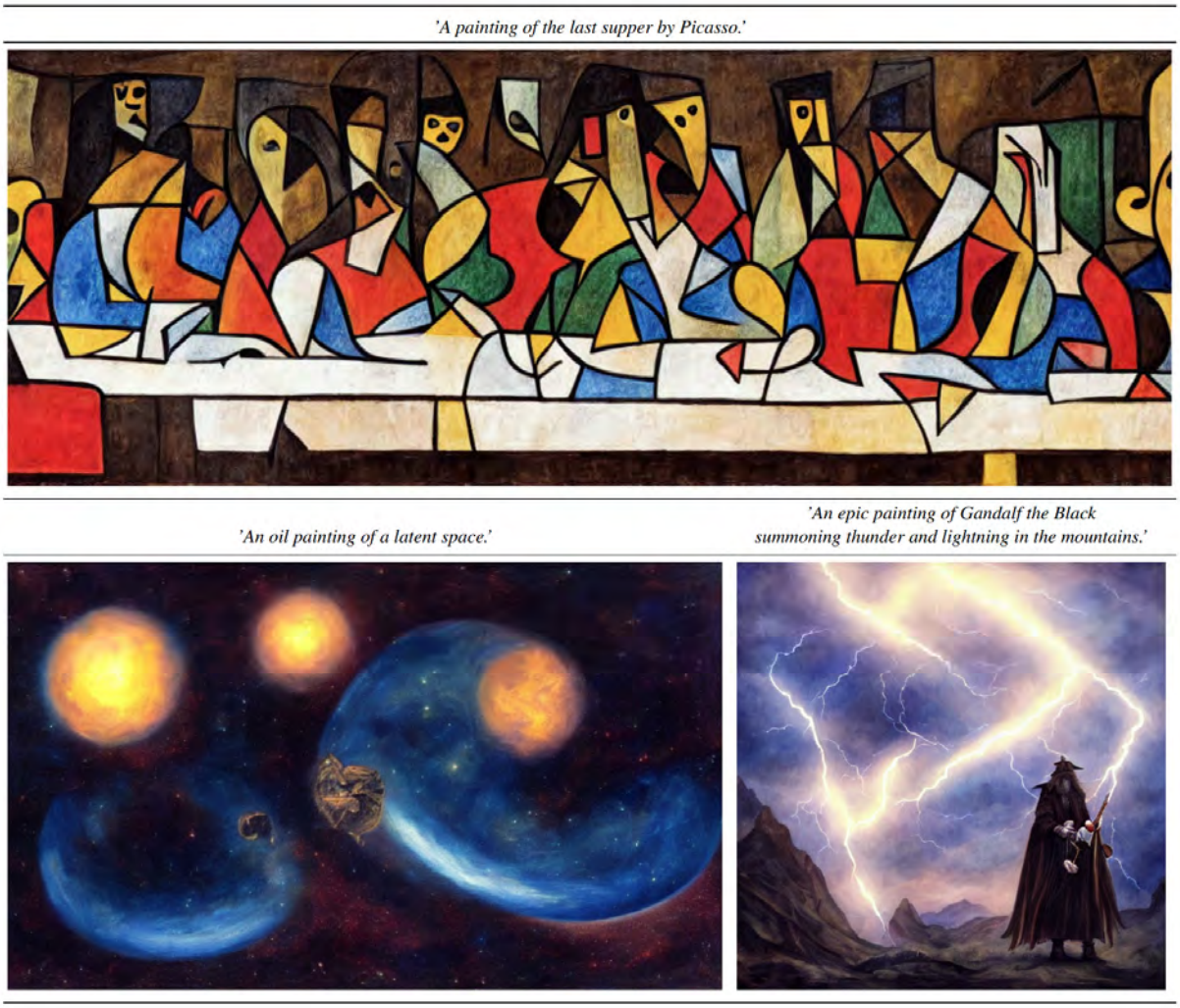

그림 1 Latent Diffusion Model[12]의 Text-to-Image 생성 결과.

링 하는 방식은 사람의 노동력이 무한하지 않기에 대규모 데이터 구축에 한계가 있었다. 하지만 최근 모델들은 웹 크롤링을 통해 얻은 텍스트와 이미지 데이터 쌍을 이용하기에 대규모 데이터셋을 손쉽게 구축할 수 있다. 이를 통해 학습된 모델들은 기존의 작은 모델에서 보이지 않던 놀라운 결과물들을 보여준다. 한편 최신 생성 기법인 diffusion[2] 모델은 이런 대규모 데이터를 이용한 모델들의 이미지 생성 성능을 크게 향상했다.

본 논문에서는 text-to image 생성 분야의 소개 및 평가 방식에 관해 기술하고 (2장), 대규모 데이터를 이용한 모델들을 소개하고자 한다. 또한 이 분야와 관련된 다양한 논문들을 diffusion 모델 이전 (3장)과 이후 (4장)로 나누어 소개한다. 또한 마지막으로 현재 생성 모델의 한계 및 향후 연구 방향을 짚어본다 (5장).

2. Text-to-Image 생성 모델

Text-to-image 생성이란 텍스트가 입력으로 주어졌을 때 해당 텍스트 설명에 부합하는 이미지를 생성하는 분야이다. 높은 품질의 이미지를 생성하는 동시에 입력으로 주어진 텍스트를 알맞게 반영하는 것이 관건인 text-to-image 생성 모델들은, 2021년 1월 OpenAI가 발표한 DALL-E [7]의 등장부터 지금까지 가장 빠르게 발전하고 있는 분야 중 하나이다. 기존 text-to-image 생성 모델과 최근 각광받는 모델의 주요한 차이 중 하나는 바로 학습 데이터의 크기에서 나타난다. 최신 모델들은 text와 image 쌍으로 이루어진 초거대 데이터셋으로 학습되었기에 기존 생성 모델보다 다양한 장면을 생성할 수 있고 텍스트가 담고있는 의미도 알맞게 반영한다. 또한, text-to-image생성 모델의 비약적인 발전에 따라 이 모델들을 평가하는 방식도 구체화됐는데, text-to-image 생성 모델들을 소개하기에 앞서 해당 절에서는 이러한 모델의 학습에 사용된 데이터셋과 text-to-image 생성 결과물을 평가하는 지표에 대해 알아볼 것이다.

2.1 Text-to-Image 학습 데이터셋

기존 text-to-image 생성 모델의 학습 데이터 중 가장 규모가 큰 공개 데이터셋은 MS-COCO[1]로, 10만장의 이미지를 제공한다. MS-COCO에서 제공하는 text-image 쌍은 사람이 직접 라벨링한 image caption으로, 이미지를 묘사하는 문장으로 이루어져 있다. 이에 반해 최신 모델들이 사용하는 text-image 쌍은 인터넷에서 크롤링한 사진과 그 사진의 태그로 사용되었던 텍스트 정보로 이루어져있다. 인터넷에서 크롤링한 사진과 텍스트로 데이터셋을 구성하기 때문에, 데이터셋의 품질은 MS-COCO에 비해 낮을지 몰라도 그 규모를 MS-COCO에 비해 적게는 100배, 많게는 1000배 이상 늘릴 수 있게 되었다.

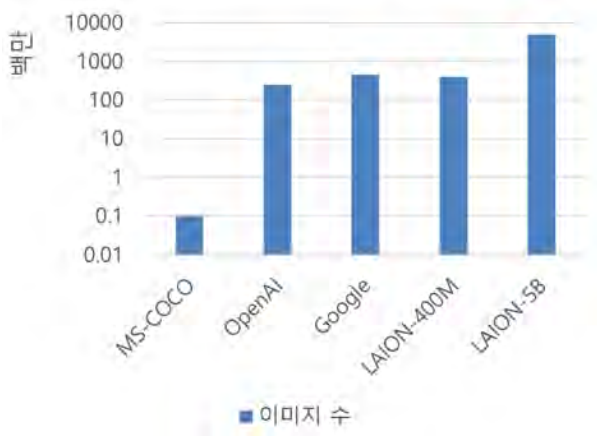

그림 2 Text-to-image데이터셋 별 이미지 개수.

OpenAI와 Google에서 자체적으로 수집한 text-image 쌍 데이터셋은 각각 2억장과 4억장 정도라고 발표하였지만 아직 공개되지 않은 비공개 데이터셋이다. 연구자들이 사용할 수 있는 공개 데이터셋 중에서 가장 많이 사용되고 있는 것은 LAION 프로젝트에서 수집한 LAION-5B[14] 데이터셋으로, 전체 image-text 쌍의 수는 50억장이다. 학습 목적에 따라 사용할 수 있는 50억장의 부분 집합 또한 공개가 되었는데, 예술 사진을 생성하기 위한 LAION-Art는 800만장, 높은 품질의 이미지 생성을 위한 LAION-Aesthetics는 1억2천만장, 그리고 고화질 생성을 위한 LAION-5B High-Res는 1억7천만장으로 LAION-5B보다는 규모가 작지만 기존 학습 데이터인 MS-COCO에 비해 1000배 가량 큰 초거대 데이터셋들이다.

2.2 Text-to-Image 생성 모델 평가 방식

기존 text-to-image 생성 모델을 평가하는 방식은 이미지 품질을 평가하는 데에 주안점이 있었다. 생성된 이미지와 학습 데이터셋인 MS-COCO간의 Frichet Inception Distance (FID) [3]를 계산하면 생성된 이미지가 학습 데이터셋의 분포와 얼마나 유사한지 측정할 수 있다. 이는 일반적으로 이미지 품질을 의미하는 지표로 해석된다.

그러나 text-to-image의 목적은 텍스트 묘사에 부합하는 이미지를 생성하는 것이기 때문에 텍스트를 얼마나 잘 반영했는지 진단하는 것 역시 중요하다. OpenAI의 CLIP[9] 모델이 공개된 후, 텍스트와 이미

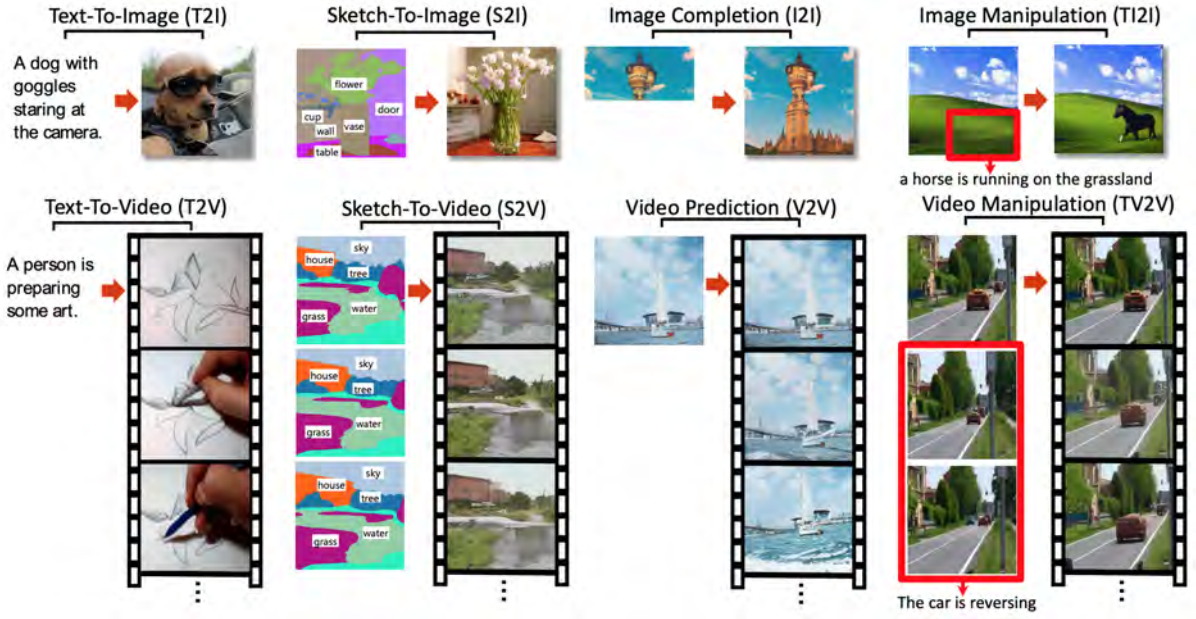

그림 5 NUWA[15]에서 진행할 수 있는 다양한 종류의 이미지/비디오 합성.

수록 이미지의 차원이 기하급수적으로 올라간다는 것이다. 이에 따라 Visual Codebook이라는 것을 도입하는데, 이는 이미지의 차원이 높기 때문에 이를 압축하여 나타내기 위한 일종의 문자의 집합으로 이후에 이미지를 나타낼 때 Visual Codebook을 문자로 한 문자열로 나타내게 될 예정이다. 이렇게 만듦으로써 이미지의 차원을 낮추어 생성하기 쉽게 만들 뿐 아니라, 이전에 GPT-3에서 거두었던 성공을 이미지에서 재현할 수 있도록 구조한 것이다. 본 학습 과정은 수많은 이미지를 통해 학습하는데 이미지를 재현하는 복원 로스와 중간에 우리가 사용할 hidden feature를 그룹화 또는 양자화 하는 트릭을 사용하여 연속적인 이미지를 비연속적인 단어들의 문자열로 나타낼 수 있도록 하였다. 이후 이렇게 양자화된 표현들을 모아 놓은 Codebook을 Visual Codebook이라고 부르는 것이다.

Prior 학습: Prior 학습이라고 불리는 두 번째 단계는 OpenAI가 GPT-3의 구조와 같은 구조를 채택하였다. 이미지를 앞서 학습된 Visual Codebook을 이용하여 2차원의 문자열로 해석하여 왼쪽 위부터 오른쪽 아래까지 쭉 늘여 놓은 후 이 구조를 학습하는 단계이다. 즉 이미지들이 Visual Codebook의 형태로 나타냈을 때 어떠한 형태를 띄는지 학습하며 이를 이용하여 Visual Codebook의 형태인 임의의 이미지를 생성할 수 있도록 2차원의 문자열의 분포를 학습하도록 한다. 구체적으로는 길이 N의 문자열이 있다고 했을 때 i번째 문자까지 준 후 i+1번째 문자를 예측하라는 방식을 진행된다. 이렇게 진행된 후에는 처음부터 문자를 N번째까지 생성하는 방식으로 이미지를 생성할 수 있게 된다. 또한 이렇게 생성되는 과정에서 다른 조건을 주입함으로써 조건별로 생성을 진행할 수도 있다. 위 방식이 이전 GPT-3에서 사용했던 방식이며 엄청나게 많은 데이터를 이용하면 최종 모델이 생성하는 문자열이 실제 문자열과 같이 자연스럽게 생성된다는 것을 확인할 수 있다.

3.2 NUWA

NUWA[15]는 text-to-image뿐 아니라 거의 모든 종류의 이미지/비디오 합성을 진행할 수 있는 모델이다. 그 예시로는 아래 그림처럼 sketch-to-image, 이미지 완성, 문자열을 이용한 이미지 합성, text-to-video, sketch-to-video, 이후 비디오 프레임을 예측하는 비디오 완성, 마지막으로 문자열을 이용한 비디오 합성까지 거의 모든 종류의 합성을 지원한다. 본 논문에 있는 예시 외에도 더 다양한 작업이 데이터가 존재한다면 학습하여 사용할 수 있는 장점을 가지고 있다.

NUWA의 구조는 Transformer[4] 구조를 적극적으로 활용하였다. 아래 그림에 있는 encoder, decoder 각각의 구조는 모두 Transformer이다. Transformer 구조의 장점 중 하나는 연산되는 방식이 입력에 따라서 다르다는 부분에 있어 단순 행렬곱보다 적응가능하고 유연한 연산이 지원될 수 있다는 장점이 있다. 이 때문에 한동안 많은 행렬곱 연산들을 Transformer의 구조로 바꾸는 노력들이 존재하였다. 하지만 여기에 더하여 Transformer의 큰 장점 중 하나는 다양한 종류의 입력을 받을 수 있다는 것이다. 이를 통하여 NUWA

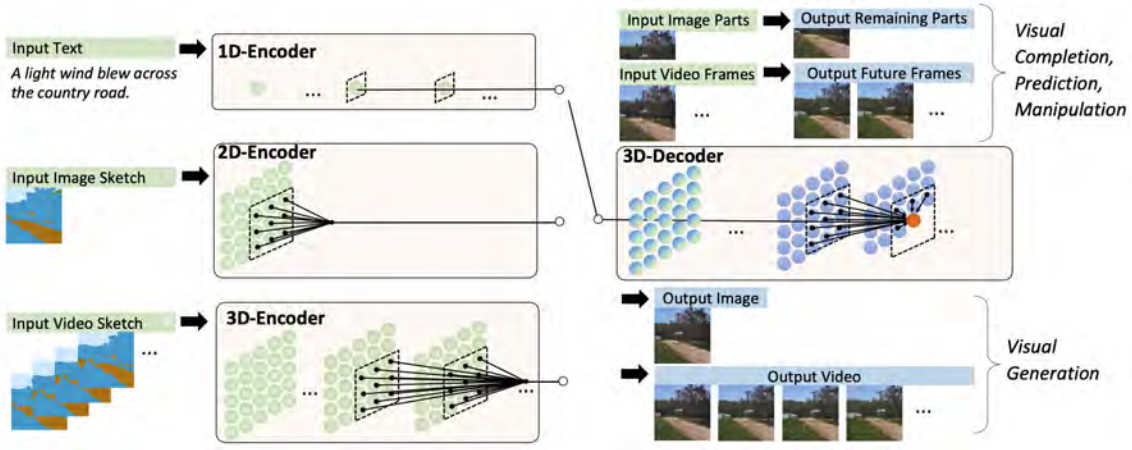

그림 6 NUWA [15]의 모델 구조. 이미지 합성을 하는 상황에 있어서 출력은 항상 이미지 또는 비디오의 형태이지만 입력은 다양한 종류의 입력을 받을 수 있다.

"Oil-on-canvas painting of a blue night sky with roiling energy. A fuzzy and bright yellow crescent moon shining at the top. Below the exploding yellow stars and radiating swirls of blue, a distant village sits quietly on the right. Connecting earth and sky is a flame-like cypress tree with curling and swaying branches on the left. A church spire rises as a beacon over rolling blue hills."

"A close-up high-contrast photo of Sydney Opera House sitting next to Eiffel tower, under a blue night sky of roiling energy, exploding yellow stars, and radiating swirls of blue"

그림 7 Parti [16]의 생성 결과.

는 문자열, 이미지, 비디오 모두를 입력으로 받을 수 있는 구조를 만들었고 이를 통해 하나의 3D Decoder를 학습함으로써 다양한 데이터가 모두 활용된 3D Decoder를 얻을 수 있었다. 학습되는 과정은 입출력이 모두가 가능하기 때문에 비교적 단순한데, 주어진 이미지의 가려진 부분을 예측하거나 비디오의 다음 프레임을 예측하는 것을 기본으로 하며 이후에 추가로 text-image 쌍이나 text-video 쌍 등 또 다른 쌍들이 존재한다면 이를 각각 입출력으로 학습하는 지도학습을 채택하였다.

3.2 Parti

Parti [16] 역시 NUWA와 비슷한 부분이 많은데 text-to-image에 보다 초점을 맞추어 어떻게 하면 이를 더 큰 규모로 학습할 수 있는지에 대해서 다룬 논문이다. 단순히 입력과 출력만 다양한 입력을 받을 수 있도록 구조한 NUWA와 다르게 Parti는 구체적으로 어떤 구조가 나은지 자세히 제안하여 성공적으로 text-to-image를 생성한 모델이다.

Parti는 기본적으로 text-to-image의 작업을 마치 문자열에서 번역의 과정 (sequence-to-sequence)로 바라

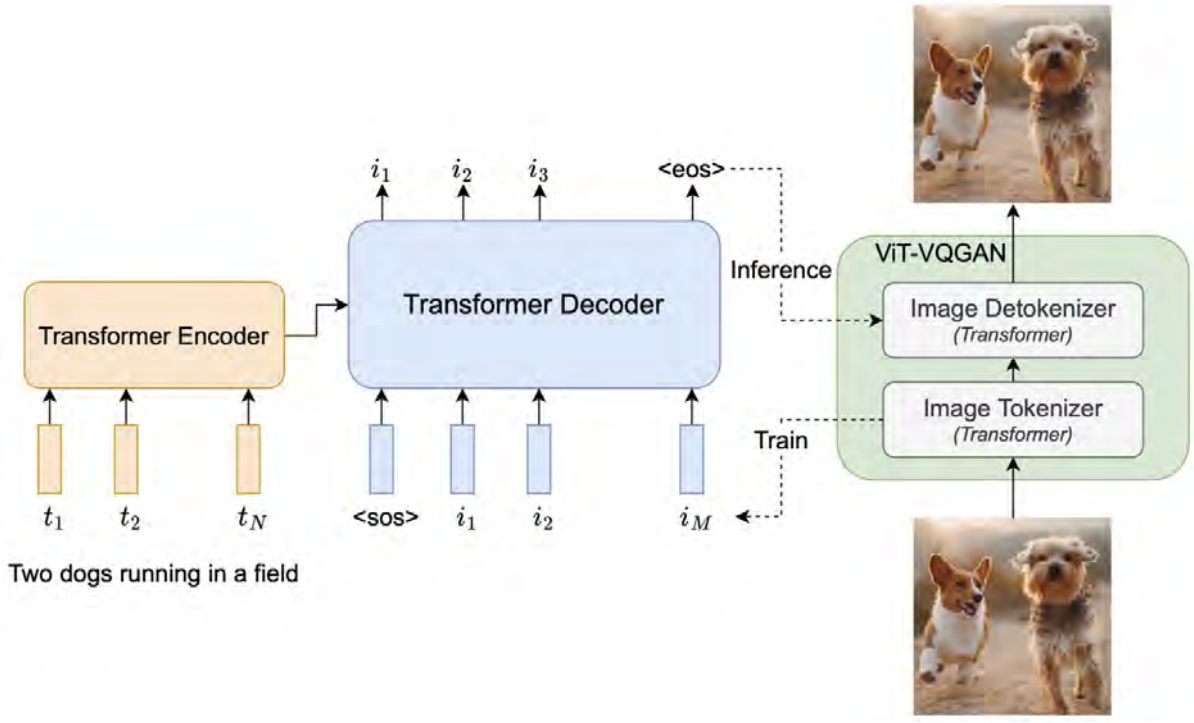

그림 8 Parti의 구조. Transformer Encoder와 Transformer Decoder를 이용하여서 문자열과 이미지를 일종의 번역의 형태로 보았으며 이미지를 문자열의 형태로 양자화하는 과정은 DALL-E에서 사용하였던 것과 유사한 형태지만 좀 더 발전된 형태로 사용하였다.

보았다. 즉 text의 형태로 되어 있는 정보를 이미지의 형태로 바꾸기만 하면 된다는 것이다. 이에 따라서 Parti는 DALL-E 1에서 사용한 Visual Codebook을 다시 사용하였다. Visual Codebook을 활용하면 이미지 역시 문자열로 볼 수 있기 때문에, text-to-image를 표현간 번역으로 볼 때 꼭 필요한 요소임을 알 수 있다. 번역에 사용된 구조는 일반적으로 서로 다른 언어 간 번역에 사용되었던 Transformer[5] 구조를 채택하였다. 최종적으로 자연어처리에서 번역을 잘하는 것으로 입증된 구조와 큰 데이터셋을 활용함으로써 다른 text-to-image 생성 모델들과 다르게 굉장히 어려운 문자열에 대해서도 이해를 잘하여 이미지를 생성하는 것을 확인할 수 있다 (위 그림 1행 참고. 67개의 단어로 이루어진 문자열에 대해서도 이미지를 잘 생성하는 것을 확인할 수 있다).

학습되는 과정으로는 text-image 쌍이 항상 필요로 하며 번역에서 진행하던 것과 같이 text-image 쌍이 있으면 text를 주어 이미지 토큰이 하나씩 나올 수 있도록 학습하는 구조를 채택하였다. 이후에도 text-to-image의 생성 과정에서 이미지를 잘 생성하는 구조, 문자열을 잘 이해하는 구조를 각각 더 발전시키기 위한 노력들이 이어졌다.

4. Diffusion 이후의 대규모 Text-to-Image 생성 모델

4.1 Diffusion Models

Diffusion 모델 [2]은 훌륭한 생성 결과와 안정적인 학습이 가능해 이미지 생성 분야에서 많은 주목을 받고 있다. 안정적인 학습을 바탕으로 정밀한 이미지를 생성할 수 있을 뿐만 아니라, 대규모의 데이터로 학습이 가능하기 때문에 GAN 모델과 달리 폭넓은 범위의 이미지 생성이 가능하다. Diffusion 모델은 non-equilibrium thermodynamics이론에서 영감을 받은 방법으로써 2015년 처음 제안되었다. Diffusion 모델에서는 주어진 데이터의 분포를 서서히 파괴하는 forward process를 정의한 후, reverse diffusion process를 학습해 random noise로부터 데이터를 복원시킨다. 이 방식을 통해 복잡하고 고차원인 이미지 데이터의 분포를 모사하고, 분포로부터 이미지를 샘플링 할 수 있다. 이미지 생성 과정에서 condition을 주고 생성을 할 수도 있는데, 이를 guided diffusion이라 한다. Condition으로 class 정보를 줄 수도 있고, text description을 줄 수도 있다.

그림 9 Diffusion[2] 모델의 생성 과정.

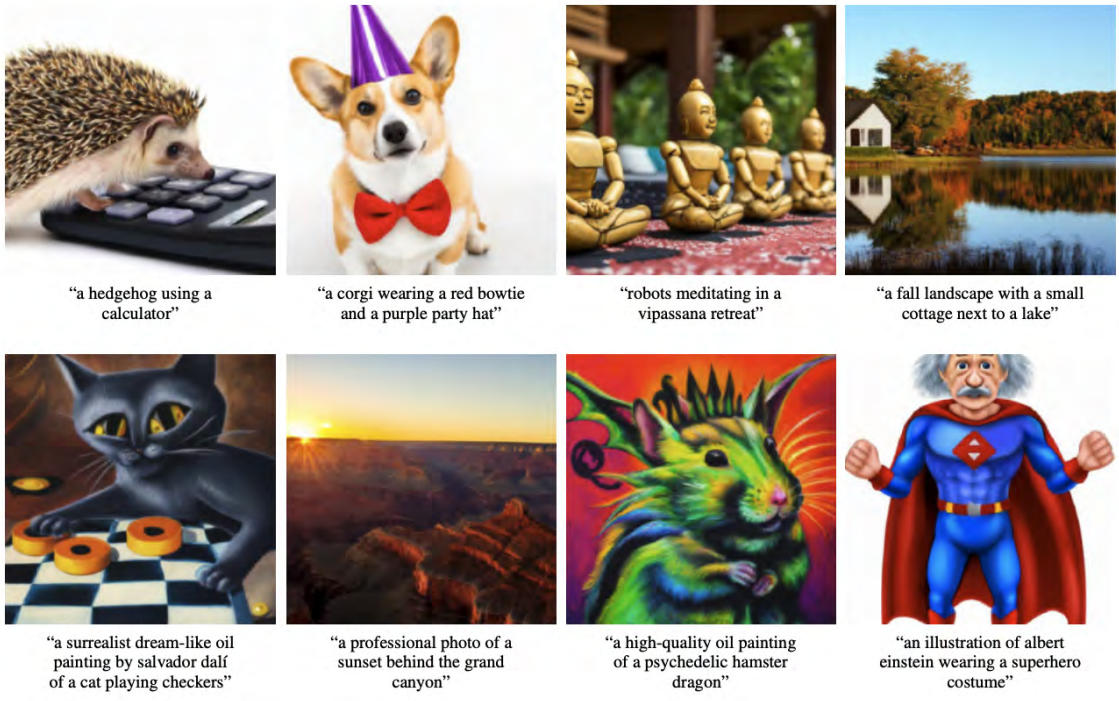

그림 10 GLIDE [9]의 Text-to-Image 생성 결과물들.

4.2 GLIDE

GLIDE [9]는 대규모 text-conditional diffusion 모델로서, text-to-image generation 문제에서 새로운 가능성을 보여준 모델이다. 구체적으로, GLIDE는 35억 개의 변수를 가진 text-conditional diffusion 모델로 64x64 이미지를 생성한 후, 15억 개의 변수를 가진 text-conditional upsampling diffusion 모델로 256x256 이미지로 upsampling을 한다. Text condition을 입력으로 주기 위해 CLIP이 사용되는데, diffusion 모델의 경우 중간 과정의 생성 결과에는 noise가 섞여 있기 때문에 별도의 noised CLIP모델을 학습시켜 사용한다. 또 추가로 생성 성능을 높이기 위해 classifier-free guidance 등의 기법이 사용된다.

4.3 DALL-E 2

DALL-E 2 [11]는 GLIDE와 비슷한 text-conditional diffusion모델로서, "prior" 모델을 추가해 성능을 향상시켰다. CLIP text embedding을 직접 사용하지 않고, prior 모델을 통해 image embedding으로 변환한 후 condition으로 사용하는 것이 특징이다. CLIP의 text embedding을 이미지로 변환한다는 이러한 과정 때문에 DALL-E 2는 unCLIP이라고 불리기도 한다. GLIDE와 비슷하게 고화질 이미지를 생성하기 위해 upsampling 모듈을 사용하는데, 최종적으로 1024x1024의 고화질 이미지를 생성할 수 있는 것이 특징이다.

4.4 Imagen

Imagen [13] 또한 마찬가지로 GLIDE와 비슷한 구조를 가진 text-conditional diffusion 모델이다. DALL-E 2와 마찬가지로 1024x1024의 고화질 이미지를 생성할 수 있고, 현존하는 모델 중 가장 좋은 생성 성능을 보여준다. Imagen은 CLIP 대신 large language model (e.g. T5-XXL [6])를 사용하면 더 좋은 생성 모델을 얻을 수 있다는 사실을 보였다.

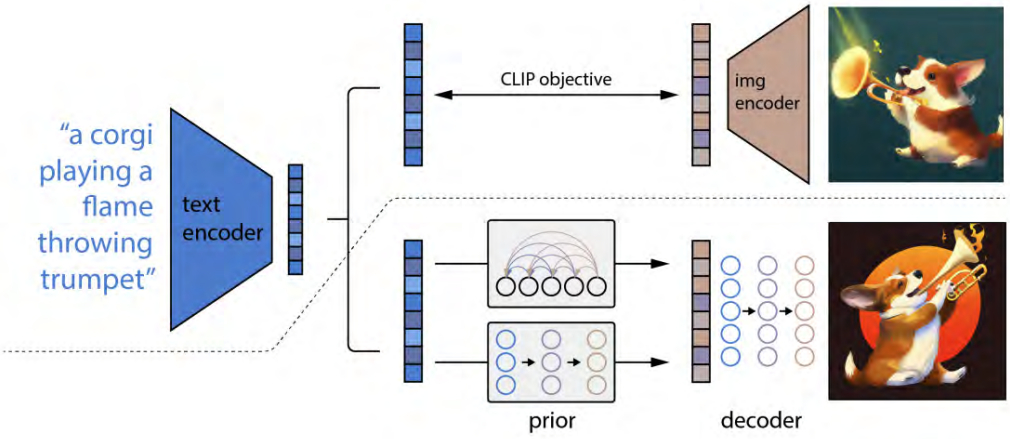

그림 11 DALL-E 2[11]의 모델 구조도.

그림 12 Imagen [13]의 Text-to_Image 생성 결과물.

4.5 Latent Diffusion Model

하나의 diffusion 모델로 고화질 사진을 생성하는 것은 연산량과 학습 난이도 모두 높다. 앞서 설명한 OpenAI의 GLIDE나 Google의 Imagen은 학습 난도를 낮추기 위해 두 개 이상의 diffusion 모델을 도입하여서 한 모델로 낮은 해상도의 이미지를 생성한 후 그 결과를 여러 번의 초해상화 (super resolution) 모델을 통과시켜 고화질 이미지를 생성하는 방식을 채택하였다. 그러나 여러 개의 diffusion 모델을 거쳐 이미지를 생성하는 것은 여전히 추론 시간(inference time)이 길다는 단점을 지니고 있다.

Latent Diffusion Model (LDM) [12]은 이미지의 디테일을 생성하는 것은 VQ-GAN에서 사용하는 디코더 구조로 충분하다고 주장하며 diffusion 모델은 디코더에 입력으로 사용될 latent code 를 생성하도록 학습시킨다. 구체적으로 LDM은 우선 초거대 데이터셋인 LAION-400M으로 오토인코더를 학습시켜 높은 품질의 이미지를 생성할 수 있는 디코더와, 특정 이미지를 디코더가 복원할 수 있는 latent code 로 변환해주는 인코더를 확보한다. 그리고 이미지 생성을 전담하는 diffusion 모델은 노이즈가 있는 이미지에서 깨끗한 이미지를 생성하는 것 대신 아니라 노이즈가 있는 latent code에서 깨끗한 latent code를 생성하도록 학습시킨다. Latent code의 차원은 이미지의 차원보다 훨씬 낮기 때문에 diffusion 모델의 추론 시간을 대폭 줄일 수 있게 된다. Latent code를 생성하기 위한 diffusion 모델의 구조는 기존 diffusion 모델과 같으므로 동일한 방법으로 텍스트 조건을 넣을 수 있고, LDM에서는 텍스트 조건 외에 그림* 에서처럼 추가적으로 다른 modality의 조건도 입력으로 받을 수 있는 모델을 제안하기도 했다.

LDM은 발표 당시 LAION-400M으로 학습한 모델을 공개하였으나 Stability AI에서 LDM의 구조를 약간 개선하여 LAION-5B으로 학습한 text-to-image 생성 모델, 일명 Stable Diffusion을 공개하였다. Stable Diffusion의 뛰어난 텍스트 이해도와 이미지 생성 능

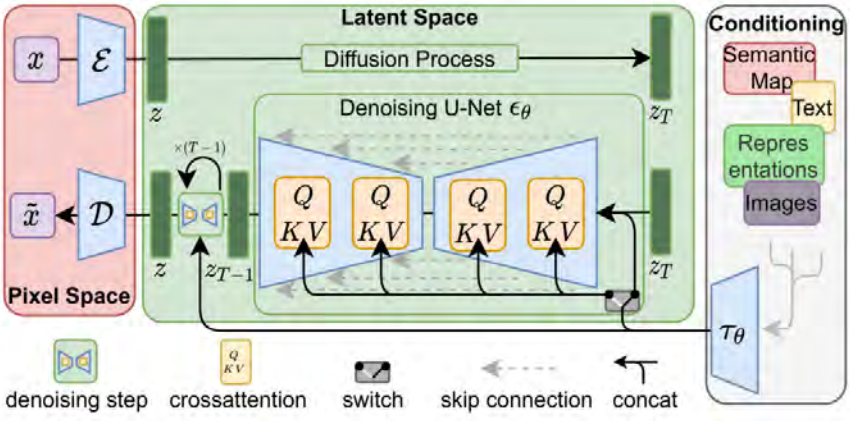

그림 13 Latent Diffusion [12] 모델의 구조도.

력으로 Image editing이나 image-to-image translation의 기반 모델로 많이 사용되고 있다.

5. 결 론

본 논문은 대규모 데이터셋을 이용해 학습한 최신 text-to-image 생성 모델들을 소개했다. 대규모 데이터를 통해 학습한 모델들의 생성 성능은 기존에 상상하기 어렵던 놀라운 품질의 이미지를 생성할 수 있었으나, 이미지 내의 디테일 한 부분의 생성이 어렵다는 점과 학습에 많은 GPU와 전력 자원을 사용해야 하기에 소규모 연구가 어렵다는 한계가 있다. 또한 diffusion모델을 이용한 모델들의 경우, diffusion 모델의 근본적인 생성 속도의 한계로 인해 이미지 생성이 느리다는 단점이 존재한다. 앞으로 Text-to-Image 생성 모델은 디테일에 대한 연구 및 모델 효율성 연구, 그리고 생성 속도가 개선된 Diffusion의 이용을 통해서 지금의 한계를 개선할 수 있을 것으로 기대된다.

6. 참고 문헌

[1] Lin, Tsung-Yi, et al., “Microsoft COCO: Common Objects in Context”, ECCV, 2014.

[2] Sohl-Dickstein, Jascha, et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics”, ICML, 2015.

[3] Heusel, Martin, et al., “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”, NeurIPS, 2017.

[4] Vaswani, Ashish, et al. “Attention is all you need.” NeurIPS, 2017.

[5] Brown, Tom, et al. “Language models are few-shot learners.“ NeurIPS, 2020.

[6] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140: 1-67, 2020.

[7] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." ICML, 2021.

[8] Croitoru, Florinel-Alin, et al., “Diffusion Models in Vision: Survey”, arXiv, 2022.

[9] Nichol, Alex, et al., “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models”, ICML, 2022.

[10] Radford, Alec, et al., “Learning Transferable Visual Models From Natural Language Supervision”, ICML, 2021.

[11] Ramesh, Aditya, et al., “Hierarchical Text-Conditional Image Generation with CLIP Latents”, arXiv, 2022.

[12] Rombach, Robin, et al., “High-Resolution Image Synthesis with Latent Diffusion Models”, CVPR, 2022.

[13] Saharia, Chitwan, et al., “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”, arXiv, 2022.

[14] Schuhmann, Christoph, et al., “LAION-5B: An open large-scale dataset for training next generation image-text models”, arXiv, 2022.

[15] Wu, Chenfei, et al. “Nüwa: Visual synthesis pre-training for neural visual world creation.” ECCV, 2022.

[16] Yu, Jiahui, et al. “Scaling autoregressive models for content-rich text-to-image generation.” arXiv, 2022.

약 력

이 상 현

2018 서울대학교 경제학부 졸업(학사)
2019~2021 한국과학기술원 김재철AI대학원 졸업(석사)
2021~ 현재 한국과학기술원 김재철AI대학원 석박통합과정
관심 분야: 컴퓨터비전, 계산사진학
Email: shlee6825@kaist.ac.kr

윤 주 열

2021 고려대학교 컴퓨터학과 졸업(학사)
2021~현재 한국과학기술원 김재철AI대학원 석박통합과정
관심 분야: 컴퓨터비전, 계산사진학
Email: blizzard072@kaist.ac.kr

박 민 호

2021 고려대학교 전기전자공학부 졸업(학사)
2021~현재 한국과학기술원 김재철AI대학원 석박통합과정
관심분야: 컴퓨터비전, 생성모델, 의미영역분할
Email: m.park@kaist.ac.kr

형 준 하

2016~2021 한국과학기술원 전기및 전자공학부, 전산학부 졸업(학사)
2021~현재 한국과학기술원 김재철AI대학원 석박통합과정
관심 분야: 컴퓨터비전
Email: sharpeeee@kaist.ac.kr

주 재 걸

2001 서울대학교 전기전자공학부 졸업(학사)
2013 미국 Georgia Institute of Technology 졸업(박사)
2015~2019 고려대학교 전기전자공학부 조교수
2019~현재 한국과학기술원 김재철AI대학원 부교수
관심분야: 인공지능, 기계학습, 컴퓨터비전, 자연어처리, HCI
Email: jchoo@kaist.ac.kr