



[Improving optical generative n network]

학 번: 20230232
이 름: 반가운
연구 지도교수: 백승환
학 과: 컴퓨터공학과

연구 목적 (Problem statement)

주제

이번 연구의 중심은 기존 Nature 논문에서 제안되었던 Optical Generative Model 을 개선시키는 것이다. Optical Generative Model 은 기존의 GPU, NN 모델에서 요구되었던 막대한 양의 에너지 소모와 하드웨어 문제를 Optical System 으로 재설계함으로써 해결할 수 있는 Potential 을 제시하고 있다. 본 연구는 현재 상용되는 다양한 text-to-image diffusion model 들처럼 user 가 원하는 text prompt 를 입력하면, 그에 맞는 image 를 생성해주는 text-to-image optical generative model 을 구현하고자 한다.

연구의 중요성

기존 Diffusion Model, GAN 등의 인공지능 기반 Image Generator 는 연산 parameter 가 기하급수적으로 증가함에 따라, 실행에 요구되는 GPU 사양과 에너지 cost 또한 기하급수적으로 증가하고 있다. 이로 인해 여러 연구 현장과 Data center 등에서는 전력 소모와 냉각, 하드웨어 재고 문제 등 여러 실질적인 부담이 커지고 있다. Optical Generative Model 은 SLM 과 여러 Optical Diffraction system 을 사용하여 기존 NN 의 GPU 의존성을 감소시키고, 새로운 하드웨어의 패러다임을 열 수 있다. 이 연구의 발전은 기존 GPU 를 사용하던 모든 분야에서의 속도 상승을 기대할 수 있으며, 특히 AI content production, edge computing 등의 분야에 긍정적인 영향을 기대할 수 있다. 또한 기존에는 학습된 class label 에

해당하는 값의 image 만 생성할 수 있었던 반면, text-to-image model 을 구현하면 위와 같은 제약 없이 유저가 원하는 어떠한 새로운 image 도 제작이 가능하기에 활용성이 높다.

연구 목표

현재 Optical generative model 은 학습을 진행할 때 같이 학습시킨 data 의 class label 에 해당하는 이미지만 생성이 가능하다. 예를 들어 “red flower, blue flower”이라는 class label 을 가진 dataset 으로 학습을 진행하면 위의 두 이미지 밖에 생성을 못하는 것이다. 이번 연구에서는 현재 상용화된 여러 text-to-image diffusion model 들처럼, user 가 어떤 text prompt 를 입력했든 그에 맞는 image 를 생성하는 text-to-image optical generative model 을 구현하고자 한다.

연구 배경 (Motivation and background)

기존 연구 요약

Optical Generative Model 은 Nature 논문에서 제안된, 기존 GPU 기반의 NN 의 한계를 극복하기 위해 제안된 model 이다. 기존의 digital generation model 들은 점점 커지는 model 크기와 complexity 때문에 막대한 computational cost(GPU, 메모리 등)과 에너지 소비 문제를 안고 있다. 이에 대해 Optical generative model 은 2D Gaussian noise 를 phase 패턴으로 변환하여 SLM(Spatial Light Modulator)에 전달한 뒤, Optimized 된 Optical Decoder 를 통해 physical 하게 image 를 생성함으로써 기존 GPU 기반의 NN 의 역할을 Optical System 으로 대체한다. 논문의 연구진들은 MNIST, Fashion-MNIST, CelebA 등의 다양한 데이터셋 이미지를 Optical 하게 생성하였으며, 기존 digital generation model 에 필적하는 quality 의 이미지를 생성하였다. 이러한 Optical Generative Model 은 에너지 소비와 computational cost 에서 매우 큰 장점을 가지고 있다.

Text-to-Image Optical Generative Model

기존 Optical Generative Model 은 Dataset 에 class label 을 붙여 학습을 시켜서, class label 에 해당하는 이미지만 생성할 수 있었다. 그러나 현재 많이 사용되는 Image 생성 diffusion model 들은 text-to-image, 즉 user 가 원하는 어떠한 text prompt 라도 입력하면 해당 prompt 에 맞는 Image 를 생성해준다. Class label 에 해당하는 Image 만 생성할 수 있는 현재의 Optical Generative Model 은 이러한 diffusion model 에 비해 활용성이 떨어진다. 따라서 이번

연구에서는 여타 text-to-image diffusion model 들과 같이, text prompt 를 입력 받아 그에 맞는 Image 를 생성해주는 text-to-image Optical Generative Model 을 구현하고자 한다.

연구 방법(Design and implementation)

고차원 설계 방식

1. Dataset 선정

먼저 학습할 Dataset 을 선정한다. 일반적인 text-to-image dataset 을 사용할 수도 있었으나, 이 경우 model 학습에 필요한 기간이 매우 길어(5 개월) 사용할 다른 dataset 을 찾아야 했다. 여러 dataset 을 찾아보고 “Oxford 102 Flower” dataset 을 사용하였다. 이 dataset 은 8189 장의 여러 종류의 꽃 사진과, " this flower is pink and white in color, with petals that are ruffled at the edges" 같은 사진을 묘사하는 text prompt 로 구성되어 있다.



Figure 1. Oxford 102 Flower Dataset

2. Text-to-Image Optical Generative Model 아키텍처 설계

시중의 여러 text-to-image diffusion model 의 구조를 살펴보며 어떻게 text prompt 를 입력 받아서 Image 를 생성하는지를 공부하였고, 그 결과 text encoder 와 cross attention 이 필요함을 느껴 이를 기존의 Optical Generative Model 과 결합하여 새로운 아키텍처를 설계하였다.

3. Evaluation

설계한 model 을 Oxford 102 Flower dataset 에 대해 학습시키고, 결과를 evaluation 했다. 다양한 text prompt 에 대해 이미지를 생성시켜 보았고, 이미지가 생성되는 동안 noise image 가 optical model 의 SLM 과 DOE 를 통과하며 어떤 식으로 변하는지도 확인했다.

실제 구현

1. Text-to-Image Optical Generative Model 학습 구조 소개

High level 에서 Text-to-Image Optical Generative Model 학습 구조는 다음과 같다. Dataset 으로 clean Image 와 description text 를 준비한다. Clip model 로 description text 를 text embedding 하여 model 에 clean Image 와 같이 input 한다. 이를 Model 에 통과시켜 여러 Timestep 별로 Image 를 생성한다. Timestep 에 따라 SNR (weight)를 다르게 하여 average mse loss 를 계산하고, 이 loss 를 반영하여 Digital Decoder 의 학습을 반복한다.

2. Text-to-Image Optical Generative Model 구조 소개

Random noise Image 와 Timestep, Text Embedding 을 Digital Encoder 에 통과시켜 Encoded Image 를 얻는다. 이를 SLM 에 통과시켜 Complex field 의 Image 를 생성하고 Digital Decoder 들에 이를 통과시켜 가며 최종 이미지를 생성한다.

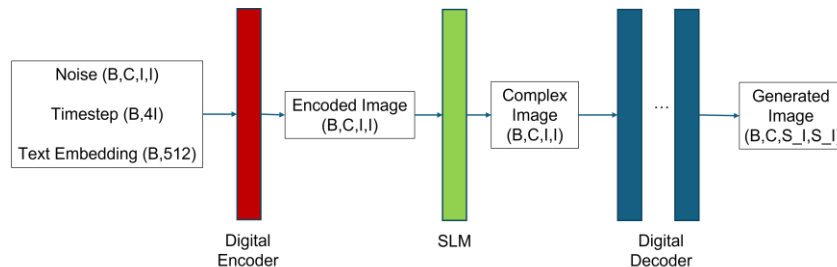


Figure 2. Text-to-Image Optical Generative Model Architecture ()안의 값은 dimension 을 의미한다. B 는 batch size, C 는 channel 개수, I 는 Image size (이미지 해상도), S_I 는 sensor 크기에 맞춘 Image size 이다.)

3. Digital Encoder 구조 소개

Text-to-Image 와 관련하여 가장 중요한 부분이 Digital Encoder 이다. 이전에 말했듯, Text-to-Image image generation 을 하려면 user 가 입력한 text prompt 를 clip model 을 통해 text embedding 을 진행해야하며, Image 와 이 들을 cross attention 하여 attention map 을 얻어야 한다. text embedding 을 통해 우리는 user 가 작성한 string 의 text prompt 를 vector 로 바꿀 수 있다. 또한 Cross attention 을 통해 우리는 attention map 을 얻어 Image 의 어떤 부분이 어떤 text 와 관련이 있는지를 얻을 수 있다. 이를 활용해 제작한 Digital Encoder 의 구조는 다음과 같다. timestep embedding 과 text embedding 을 합치고, 이 값과 Random noise Image 를 cross attention 하여 attention map 을 얻는다. attention map 과 noise image 를 torch.cat 한 값을 output 한다.

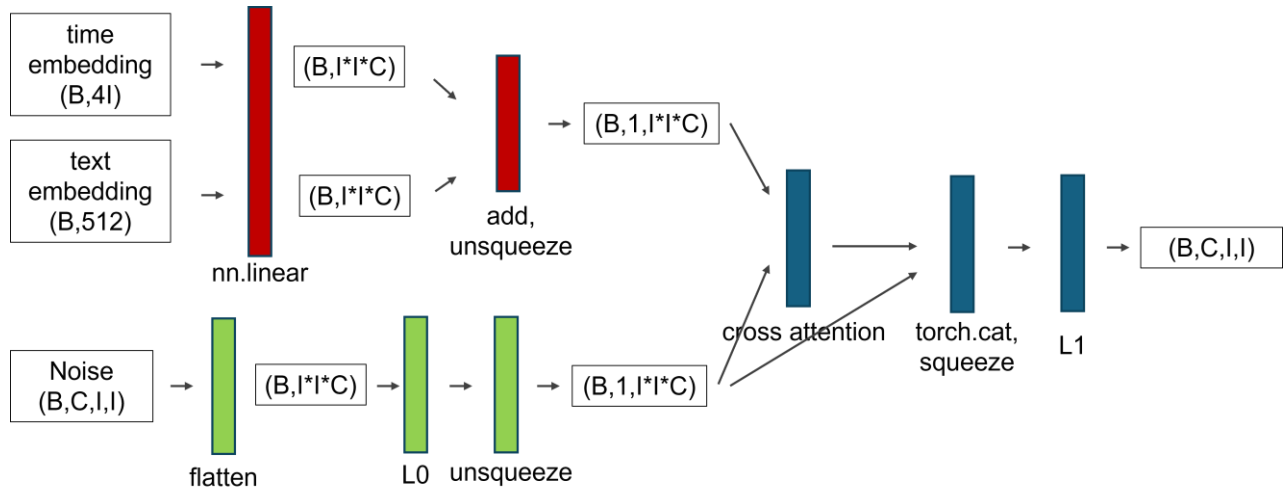


Figure 3. Digital Encoder Architecture

4. SLM 과 Digital Decoder

SLM 과 Digital Decoder 는 Optical element 로, 기존 Diffusion model 들과의 가장 큰 차이점이다. 기존 Diffusion model 에서는 NN 를 통해 Noise Image 를 단계별로 Denoising 해가며 새로운 Image 를 생성했기에 에너지 소모가 엄청났다. 하지만 Optical generative model 에서는 noise Image 를 SLM 과 Digital Decoder(DOE)를 통과시켜 이미지를 얻기에 Image Inference 시간이 짧고 에너지 소모가 적다.

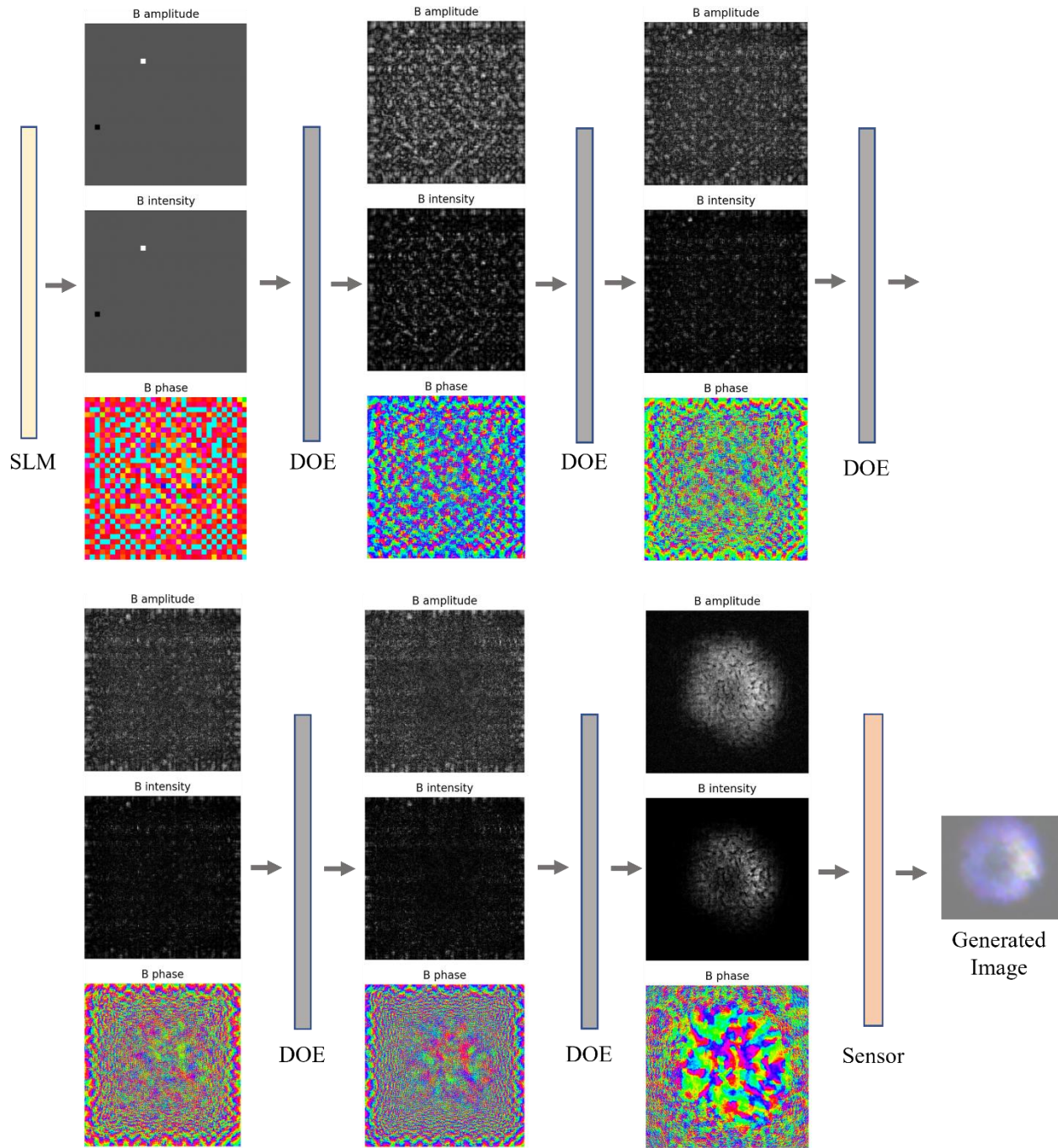


Figure 4. Noise image that changes over SLM and DOE SLM 과 DOE 를 지남에 따라 noise 의 Image 의 blue channel 에서 amplitude 와 intensity, phase 가 어떻게 변하는지를 나타낸다.

연구 결과 및 평가 (Methodology and evaluation)

Simulation 환경 설정

여러 text prompt 로 Image 생성을 진행했는데, 이때 model 의 여러 설정 값은 다음과 같다.

- Channel 개수: 3 (R,G,B)
- Image size(row 당, col 당 pixel 개수): 32
- 학습 epoch 개수: 100
- SLM 과 첫번째 Digital Encoder 사이 거리: 0.05m
- Digital Encoder 사이사이 거리: 0.01m
- Digital Encoder 와 Sensor 사이의 거리: 0.05m

Generated Images

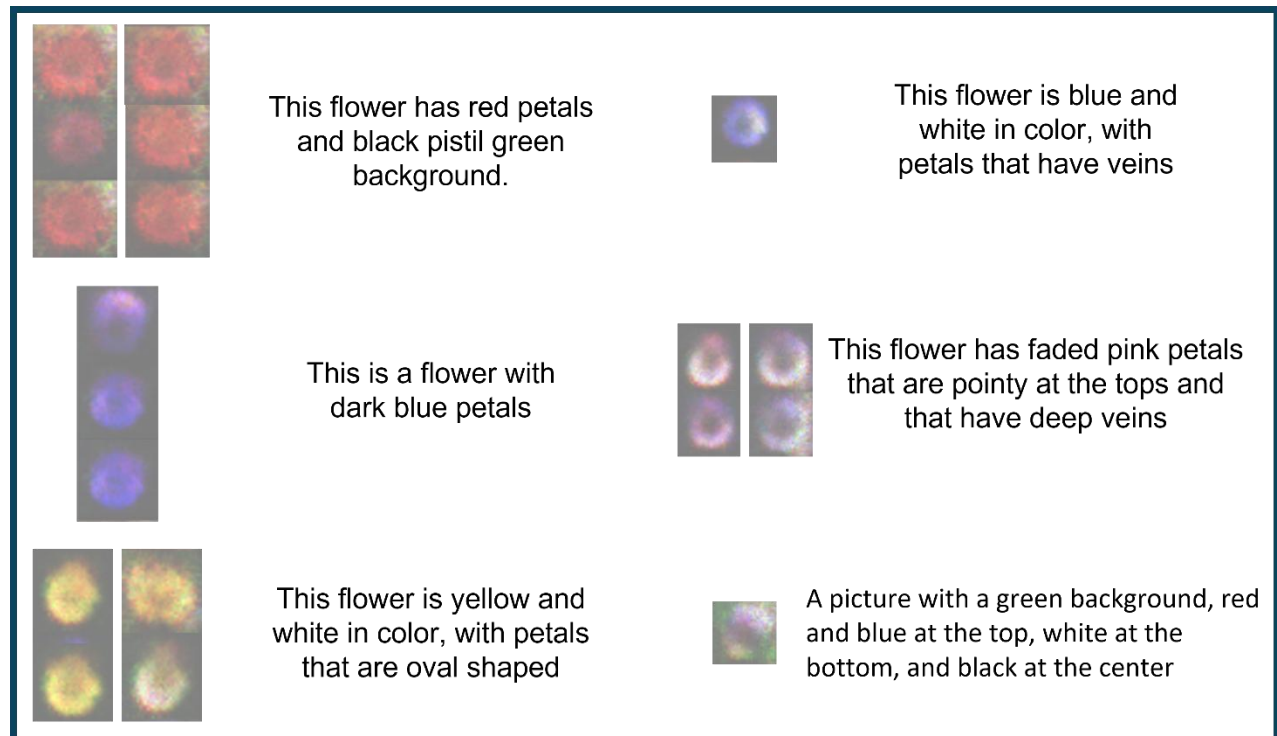


Figure5. Generated Images & Text Prompt

토론 및 전망 (Discussion and future work)

연구 내용 최종 정리

본 연구는 기존 Nature 논문의 Optical Generative Model 을 확장하여 text-to-image 생성 기능을 구현하였다. 기존 model 이 미리 학습된 class label 에 해당하는 이미지만 생성할 수 있었던 한계를 극복하고, user 가 입력한 임의의 text prompt 에 대응하는 이미지를 optical 하게 생성하는 시스템을 개발했다. 현대 diffusion model 과 GAN 은 parameter 수 증가로 인해 GPU 사용량, 전력 소비가 기하급수적으로 증가하는 문제를 안고 있다. 이에

SLM(Spatial Light Modulator)와 DOE(Diffractive Optical Element)를 활용한 optical 연산으로 GPU 의존성을 감소시키고 에너지 효율을 획기적으로 개선할 수 있는 optical system 에 주목했다.

여러 꽃 사진과 description text 로 이루어진 Oxford 102 Flower dataset 을 사용하여 model 을 학습하여 model 학습시간을 현실적으로 관리하였다. 기존 text-to-image diffusion model 을 분석하여 text encoder 와 cross-attention mechanism 이 필수임을 파악하고, 이를 optical system 과 결합하여 새로운 아키텍처를 설계하였다. 학습 과정에서는 CLIP model 로 text 를 embedding 하고, text embedding 과 clean Image 를 함께 input 하여 여러 timestep 별로 이미지를 생성한 후, SNR 기반 weighted average MSE loss 로 Digital Decoder 를 학습시켰다.

Model 의 핵심은 Digital Encoder 와 Optical element 의 결합이다. Digital Encoder 에서는 timestep embedding 과 text embedding 을 결합한 후, 이를 noise image 와 cross attention 처리하여 attention map 을 생성한다. 이 attention map 은 이미지의 어느 부분이 어떤 text 와 연관되는지를 표현하며, noise image 와 concatenate 된다. 이후 SLM 이 encoded image 를 phase pattern 으로 변환하여 complex field 를 생성하고, Digital Decoder (DOE)의 optical diffraction 와 sensor 를 통과해 최종 image 를 생성한다. 기존 NN 기반 denoising 과 달리 optical 하게 처리하여 image inference 시간을 단축하고 에너지 소비를 대폭 감소시켰다.

보완되어야 할 점

생성된 image 를 보면 noisy 한 것을 알 수 있는데, 이는 model 을 학습시킬 때 pixel 의 개수를 32x32 로 학습시켰기 때문이다. 메모리와 시간 부족 문제 때문에 resolution 을 매우 낮게 하여 모델 학습과 이미지 생성을 진행하였는데, 후에 더 높은 resolution 으로 image 를 생성해보고자 한다.

또한 현재는 flower dataset 으로 학습하여 꽃과 관련된 이미지 만을 생성할 수 있는데, 더 많은 dataset 으로 model 을 학습시켜 novel ai 같은 diffusion model 처럼 어떤 text prompt 를 입력하든 그에 맞는 이미지를 생성하도록 하고 싶다.

[Code, dataset, result](#)

참고 문헌(References)

- Shiqi Chen et al., “Optical generative models” Nature(2025)
- Amir Hertz et al., “Prompt-to-Prompt Image Editing with Cross Attention Control”