일반논문 (Regular Paper)
방송공학회논문지 제28권 제6호, 2023년 11월 (JBE Vol.28, No.6, November 2023)
https://doi.org/10.5909/JBE.2023.28.6.733
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

한옥을 중심으로 한 한국형 재난 이미지 생성 결과 : 파인 튜닝 기술의 성능 비교 분석과 적용 가능성

최 민 지[a], 원 루 빈[a], 최 지 훈[b], 배 병 준[b]‡

Korean-style Disaster Image Generation Focused on Hanok: A Comparative Analysis of Fine-Tuning Techniques and their Applicability

Minji Choi[a], Ru-Bin Won[a], Ji Hoon Choi[b], and Byungjun Bae[b]‡

요 약

본 논문은 정보 인지가 취약한 계층에게 재난 상황을 효과적으로 알리기 위한 한국형 재난 이미지 생성의 새로운 접근법을 탐구한다. 현재까지의 재난 문자는 주로 텍스트를 이용하지만, 이는 고령층, 장애인, 외국인 등 다양한 사람들에게는 한계가 있음을 인식하고 있다. 이에 따라, 파인 튜닝 기술과 이미지 생성 모델을 이용해 한국의 특성과 지형에 맞는 재난 정보를 전달하는 방법을 연구한다. 특히 한옥이라는 한국의 대표적인 랜드마크를 중심으로 실험을 진행, 한국형 재난 이미지 생성을 위한 다양한 파인 튜닝 방법에 따른 결과를 확인하고, 성능을 분석한다. 결과적으로 각각의 파인 튜닝 기술들은 컴퓨터 사양, 프롬프트 설계, 연산 시간 등 추가적인 변수에 따른 성능 차이가 있어 추가 연구가 필요하다는 결론을 내린다. 본 논문의 연구는 국내뿐만 아니라 글로벌 재난 대응에도 적용할 수 있을 것으로 예상되며, 미래 재난 연구의 방향성을 제시한다.

Abstract

This paper explores a new approach to generating disaster images tailored for the Korean context, aiming to effectively communicate disaster situations to populations that have difficulty processing information. While disaster notifications have primarily utilized text so far, it has been recognized that this method has limitations, especially for the elderly, the disabled, and foreigners. Accordingly, this study investigates methods of conveying disaster information suited to the characteristics and topography of Korea by using fine-tuning techniques and image generation models. Experiments primarily focused on the 'Hanok', a representative Korean architectural landmark, were conducted to evaluate the results of various fine-tuning methods for creating Korean-style disaster images and analyze their performance. The findings indicate that the performance of each fine-tuning technique varies depending on computer specifications, prompt design, computation time, and other variables, suggesting the need for further research. The research presented in this paper is expected to be applicable not only domestically but also in global disaster response, providing direction for future disaster research.

Keyword : image generation, text to image model, Fine tuning, disaster image, prompt engineering

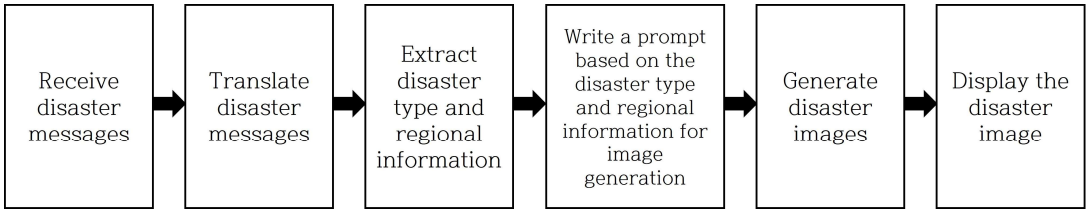

그림 1. 재난 문자 기반 재난 이미지 생성 순서도
Fig. 1. Disaster message-based disaster image creation flowchart

Ⅰ. 서 론

재난 문자는 국민에게 긴급한 재난 상황을 알리기 위한 서비스로, 현재는 텍스트 형식으로 제한되어 있다. 좀 더 재난 상황 인지를 향상하기 위해서는 텍스트와 더불어 이미지, 동영상, 음성(음향) 등의 멀티미디어를 함께 제공하는 것이 더 효과적이다. 특히, 시청각장애인, 노령층, 외국인과 같은 정보인지 취약계층에게는 이러한 멀티미디어 기반의 재난 정보 전달은 효과가 더욱 크다. 다양한 정보 수용 능력을 갖춘 국민과 외국인에게 효과적으로 재난 정보를 전달하기 위해서는 텍스트와 함께 멀티미디어 형태로 제공은 필수적이라고 볼 수 있다.

전 세계적으로 재난 정보를 멀티미디어로 제공하는 방법에 대한 다양한 연구가 진행되고는 있으나[1] 발생 지역, 시간 등 현재의 상세 정보가 필요한 긴급한 재난 상황이 발생할 시에는 멀티미디어 형태로 재난 정보를 전달하기가 어렵다. 최근에는 머신 러닝 기술들이 발전함에 따라서 머신 러닝을 활용하여 이러한 부분을 해결하고자 하는 노력이 있지만, 머신 러닝에서 활용되고 있는 학습 데이터셋이 대부분이 국외에서 만들어진 것이기 때문에 국내의 재난 상황을 표현하기에는 매우 부족하다[2].

국내에서 전송되는 재난 정보에는 국내 재난 발생 장소, 국내 재난 종류 등 국내 상황의 정보들이 상당수 포함되고 있다. 따라서 머신 러닝을 활용한 재난 멀티미디어를 생성하는 데 있어서 한국의 특성이 있는 학습 데이터셋을 통한 학습, 지형 또는 장소 등 한국 랜드마크 기반 표현 등에 대한 접근으로 재난 정보 전달의 효과를 극대화하는 것이 중요하다. 그림 1에서는 재난 문자를 기반으로 재난 이미지를 생성하는 순서를 나타낸다.

본 논문은 국내 지형과 재난 상황, 그리고 한국의 문화적 특성까지 고려한 다양한 면에서 재난 정보 전달 방법의 적용 가능성을 파인 튜닝 기술을 통해 제시하고자 한다. 다양한 파인 튜닝 방법을 적용하여 이미지 생성 모델들의 성능을 비교 분석하며, 그중에서 국내 지형과 재난 상황을 가장 정확하게 표현할 방안을 도출하기 위해서 다양한 조건의 실험을 수행한다. 한국의 많은 랜드마크 중에서 한국형 재난 이미지를 생성하기 위한 예시로 한옥을 선택한다. 한옥 이미지 데이터셋을 사용해 이미지 생성 모델들을 추가 학습하여 결과를 보고, 각 파인 튜닝 모델 결과의 원인을 분석한다.

Ⅱ. 이미지 생성 & 파인 튜닝 기법

이 논문에서는 네 가지의 파인 튜닝 기술을 사용하여 실험을 진행한다. 사용된 파인 튜닝 기술들은 이미지 생성 또는 추가 학습 관련 연구에서 자주 사용되고 비교되는 기술들로 선정하였다[3,4].

1. Stable Diffusion

Stable Diffusion은 diffusion 기반 Text to image 모델이다[5]. Stable Diffusion은 오픈 소스의 특성상 여러 연구가 진행되었고, 이를 통해 다양한 파인 튜닝 기술들을 실행할

a) 과학기술연합대학원대학교 정보통신공학전공(UST)
b) 한국전자통신연구원(ETRI)
‡ Corresponding Author : 배병준(Byungjun Bae)
E-mail: 1080i@etri.re.kr
Tel: +82-42-860-3888
ORCID: https://orcid.org/0000-0002-0872-325X
※이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2022-0-00083, 재난정보인지 취약계층을 위한 맞춤형 재난미디어 서비스 플랫폼 기술 개발).
· Manuscript November 2, 2023; Revised November 22, 2023; Accepted November 22, 2023.

수 있게 되었다. 본 논문에서 모든 실험은 Stable Diffusion을 간소화한 Stable Diffusion WebUI[6]에서 진행된다.

2. Textual inversion

Textual inversion[7]은 사전 훈련된 이미지 생성 모델의 텍스트 임베딩 공간에서 새로운 단어를 찾는 방식을 사용한다. 텍스트 인코딩 프로세스에서는 먼저 입력 문자열을 토큰 집합으로 변환한다. 이후, 각 토큰은 임베딩 벡터로 변환된다. 이 과정에서 'Pseudo word'를 사용하여 새로운 임베딩 벡터와 텍스트 쿼리를 구성한다. 이후 프롬프트를 입력하여 이미지를 생성하고자 할 때, Pseudo word가 들어간 문장을 설정하여 이미지를 재구성하는 것으로 이어지는 단일 단어 임베딩을 찾는다. 즉 Textual inversion은 텍스트 인코더의 임베딩 공간에서 새로운 Pseudo word를 찾아 어휘에 개념을 주입하는 방식으로 파인 튜닝을 진행한다.

3. Dreambooth

Dreambooth[8]는 두 가지 핵심 목표를 가지고 진행된다. 첫 번째 목표는 몇 장의 사진만으로 주제의 시각적 특징에 대한 높은 충실도를 유지하며 새로운 맥락의 사진을 합성하는 것, 두 번째 목표는 몇 장의 이미지로 Text to image 모델을 파인 튜닝 하면서 기존 semantic knowledge는 보존하는 것이다. 이 두 가지 목표를 이루기 위하여 class name과 class image를 사용한다. 개체를 표현하기 위한 몇 장의 이미지를 입력으로 주고 Text to image 모델을 파인 튜닝하여 고유 index와 입력 이미지를 함께 label 하는 방법을 학습한다. 이후 고유 index를 포함한 프롬프트를 입력하게 되면 입력 이미지의 피사체가 출력 이미지에 합성된다.

4. LoRA

LoRA(Low Rank Adaptation)[9]는 전체 파라미터 업데이트가 비효율적이라는 판단 하에 새롭게 연구되어 개발된 방법이다. LoRA를 사용하여 파인 튜닝을 진행할 경우 기존 사전 학습 모델의 가중치들은 업데이트 되지 않고, rank decomposition matrices들의 가중치만 업데이트 된다. 여기서 rank decomposition matrix는 두 개 이상의 하위 행렬의 곱으로 주어진 행렬을 근사값으로 나타내는 과정이다. LoRA는 Stable Diffusion의 cross attention layer에서 파인 튜닝을 진행한다. 기존의 cross attention layer의 가중치는 행렬로 정의되는데, 이러한 matrix에 가중치를 추가하여 모델을 파인 튜닝 하는 것이다.

5. Hypernetworks

Hypernetwork[10]는 Stable Diffusion 내부 U-Net의 cross attention에 삽입되어 파인 튜닝을 진행하는 선형 네트워크이다. Hypernetwork 두 개를 cross attention 모듈 이전에 삽입한다. Attention 매커니즘에 존재하는 key, query, value 벡터 중 key, value 두 개의 벡터를 변형한다. Hypernetwork 자체는 dropout과 activation이 있는 fully connected layer network이고, 이 네트워크를 사용하여 변환된 key, value 벡터는 cross attention에서만 적용된다.

Ⅲ. 한옥 이미지의 중요성 및 선택 배경

1. 한옥 이미지 중요성 및 선택 배경

대다수의 재난 장소로써 산, 바다 및 하천 등의 자연 지형이 주로 언급되지만, 대부분의 이미지 생성 모델은 대규모 데이터셋을 기반으로 학습되기 때문에 이러한 자연 장소의 이미지 생성은 상대적으로 자연스럽게 나타난다. 반면, 아파트와 같은 도시 지역도 마찬가지의 현상을 보이지만, 특정 아파트 혹은 자연 지형에서의 재난을 구체적으로 표현하기는 어렵다.

본 연구는 한옥이라는 독특한 건축 양식에 주목하여 이미지 생성 모델의 효율성과 한계점을 탐구하였다. 주요 이유는 아래와 같다:

1) 대다수의 이미지 생성 모델이 한옥 이미지의 생성에 어려움을 보인다. 이는 주로 사전 학습 데이터셋에 한옥 이미지가 부족하거나 존재하지 않기 때문으로 파악된다.

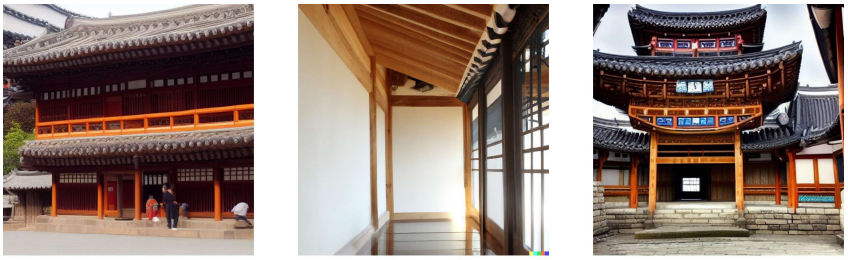

그림 2. Text to Image 모델의 한옥 이미지 출력 결과
Fig. 2. Hanok image output results of the Text to Image models

2) 그림 2에서는 여러 Text to Image 모델들이 생성한 한옥 이미지 결과를 제시하고 있다. 관찰 결과, Stable Diffusion, DALL • E 2, Playground 모델에서는 한옥이 아닌 중국 또는 일본의 전통 가옥 스타일의 이미지가 출력되었다. 이러한 경향은 기존 Text to image 모델 학습 과정에서 중국 및 일본의 전통 건축물에 대한 데이터셋이 한옥 데이터셋에 비해 상대적으로 많다는 사실을 반영하는 것으로 해석된다.

표 1에서는 한·중·일의 전통 가옥에 대한 각각의 특징을

표 1. 한·중·일의 전통 가옥 별 특징
Table 1. Characteristics of traditional houses in Korea, China, and Japan

country		Characteristics
Korea		- Monochrome or wooden-colored main gates and pavilions are numerous - Curvature, simplicity - The eaves have a gentle curve - Mostly single-story buildings
China		- Features shades of red/green/gold - Flamboyant, intense, linear - Large in size and majestic
Japan		- Low-saturation main gates - Understated, linear - Many two-story structures - White and gray buildings - Many structures have gardens

설명한다. 한옥은 그 자체로 고유한 특성을 보인다[11,12,13]. 본 연구의 결과는 국내 건축물과 지형의 특성을 반영한 이미지 생성은 정보 전달의 새로운 수단으로 활용될 수 있을 것으로 기대된다.

2. 파인 튜닝 과정에서의 특징 및 변형

Stable Diffusion에서의 파인 튜닝 과정에서는 학습 대상을 특정하기 위한 고유 변수 할당이 필수적이다. 이 변수는 "initialization text"로 지칭된다. Initialization text는 고유한 벡터로 작동하여, 생성되는 임베딩 공간을 해당 벡터로 초기화한다. 중요한 점은 이 initialization text가 기존 이미지 생성 모델 내부에 존재하지 않는 단어여야 한다는 점이다. 흔하거나 이미 존재하는 단어를 사용하면, 그 단어가 기존에 학습된 개념과 혼동될 위험이 있다. 본 연구에서는 이 initialization text를 특정하지 않은 단어로 설정하였으며, 본문에서는 이를 '*'로 표기하였다.

또한, 네 가지 파인 튜닝 모델을 적용할 때, 이미지 전처리 과정에서는 BLIP captioning을 활용하여 각 이미지에 대한 캡션을 자동 생성한다. 이러한 캡션들은 각 이미지와 함께 학습 과정에 포함된다. 일반적으로 한옥 이미지에 대한 캡션은 "A building with a roof"라는 표현으로 생성된다. 본 연구에서는 한옥의 이러한 표현을 초기에 설정한 initialization text, 즉 “*” 치환하였다. 이러한 접근은 "A building with a roof"라는 기존의 표현을 initialization text “*”로 변환하며, 주어진 한옥 이미지의 개념을 학습하게 하는 목적을 가지고 있다.

3. 화재 시뮬레이션의 선택 배경

본 연구에서 주목한 재난 유형인 '화재'는 그 자체의 시각적 특성, 즉 독특한 붉은색 및 연기 발생 등으로 인해 다른 자연재해 시나리오보다 상대적으로 빠르게 인식될 수 있는 장점을 가진다. 이러한 화재의 특성은 이미지 생성 모델에게 명확한 시각적 가이드라인을 제시해줌으로써, 더욱 정확한 재난 시뮬레이션 생성에 기여한다. 추가로, 건축물과 관련된 재난 발생 현황 분석 결과에 따르면, 대형 화재가 건축물 재난의 32%를 차지하며, 이는 건축물에서 가장 주요한 재난 유형 중 하나임을 시사한다[14]. 이에 기초하여 본 논문은 특히 한옥이라는 전통 건축물에서 발생 가능한 화재 시나리오를 중점적으로 고려하고, 이에 대한 이미지 생성 결과 분석을 진행하였다.

4. 데이터셋 장수의 선택 이유와 중요성

본 연구에서는 이미지 데이터셋의 장수를 5장, 50장, 100장으로 구분하여 실험을 진행하였다. 이러한 분류의 기준은 각 기술별로 최적의 성능을 발휘하는 데 필요한 이미지의 장수가 서로 다르기 때문이다. Textual inversion과 Dreambooth는 3~5장의 이미지만을 활용하여도 우수한 성능을 나타내었다는 연구 결과가 있으며, LoRA와 Hypernetwork는 훨씬 많은 장수의 이미지 학습이 필요하다고 한다. 이러한 분류를 통해 다양한 데이터셋 규모에서의 성능 차이를 분석하였으며, 추후 한국형 재난 이미지 생성 모델의 구축 과정에서 필요한 데이터셋의 규모를 결정하는 데 큰 도움을 제공할 것으로 기대된다.

Ⅳ. 입력 프롬프트 별 이미지 생성 결과

본 연구는 NVIDIA A6000 GPU 환경의 Stable Diffusion WebUI를 사용하여 네 가지 파인 튜닝 기술을 실험하였다. 실험에 사용된 지형 이미지 데이터셋은 AI HUB의 한옥 이미지 데이터 셋이다[15]. 네 가지의 파인 튜닝 기술에서 세 종류의 프롬프트를 입력하여 이미지 데이터셋 장수 별 결과를 얻었다. y축은 파인 튜닝 과정에서 활용된 이미지의 수를 나타내며, x축은 각 파인 튜닝 모델을 표현한다.

1. A 프롬프트(“*, fire”) 입력 결과

그림 3은 A 프롬프트(“*, fire”)에 따른 다양한 파인 튜닝 모델의 출력 결과를 시각화한 것이다. Textual inversion 모델은 한옥 건물의 형상을 정확히 재현하는 데에 성공하였으나, 화재와 관련된 시각적 요소는 검출되지 않았다. 또한, 50장 학습 결과에서 일본의 전통 가옥과 유사한 2층 건축물을 보였다. Dreambooth 모델은 화재의 형상을 명확하게 캡

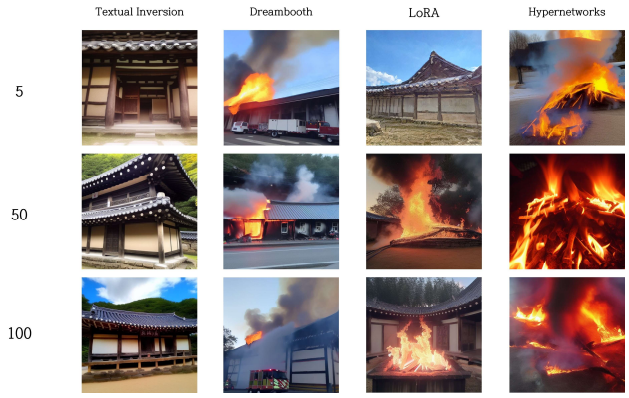

그림 3. 데이터셋 장수 별 A 프롬프트 입력 결과
Fig. 3. Results of entering prompt A by dataset size

처하였으나, 100장의 이미지 학습 시 한옥의 특징을 정확하게 반영하지 못하였다. LoRA 모델은 5장의 학습 데이터로는 화재의 형상을 제대로 표현하지 못하였으며, 50장 및 100장의 학습 데이터를 활용한 결과에서도 건물과 불 사이의 연결성이 미흡하게 나타났다. 마지막으로, Hypernetwork 모델은 전체 학습 데이터에 걸쳐 한옥의 형태를 제대로 표현하지 못하였지만, 화재와 관련된 형상은 두드러지게 표현되었다.

2. B 프롬프트("* in the fire") 입력 결과

그림 4는 B 프롬프트("* in the fire")에 따른 다양한 파인 튜닝 모델의 출력 결과를 제시한다. Textual inversion 모델은 한옥 건물의 형상을 충실히 재현하였으나, "fire"에 해당하는 시각적 요소는 감지되지 않았다. Dreambooth 모델은 모든 학습 데이터셋에서 한옥 형태의 건축물이 화재와 함께 나타나는 특징을 보였다. LoRA 모델은 5장의 학습 결과에서 화재와 관련된 형상을 포착하지 못하였으나, 50장의 결과에서는 한옥과 모닥불 형상이, 그리고 100장의 결과에서는 두 형상이 조화롭게 통합된 이미지가 나타났다. 반면, Hypernetwork 모델은 건물로 추정되는 형상이 누락되었고, 대신 모닥불 형상이 주로 출력되었다.

3. C 프롬프트("A fire broke out in *") 입력 결과

그림 5는 C 프롬프트("A fire broke out in *")를 기반으로 다양한 파인 튜닝 모델들의 반응을 보여준다. Textual inversion 모델은 전반적으로 한옥 형상을 잘 포착하였으나, 50장의 학습 결과에서 화재와 관련된 요소는 누락되었다. Dreambooth 모델은 모든 학습 데이터셋에서 화재가 발생한 건물의 형상을 출력하였는데, 특히 100장의 학습 결과에서는 한옥이 아닌 다른 형태의 건물이 주로 나타났다. LoRA 모델은 5장의 학습 결과에서는 한옥 형상만을 출력하였으나, 50장 및 100장의 학습 결과에서는 한옥과 화재가 조화롭게 통합되어 나타났다. 반면, Hypernetwork 모델은 모든 학습 데이터셋에서 화재가 발생한 건물의 형상을 보였으나, 그 형상이 한옥을 정확하게 반영하지는 않았다.

본 연구에서 첨부된 그림을 통해 각 파인 튜닝 기술에 따른 학습 이미지 장수 별 학습 소요 시간을 살펴보았다. 그림 6에서 x축은 학습 이미지의 장수를, y축은 학습에 소

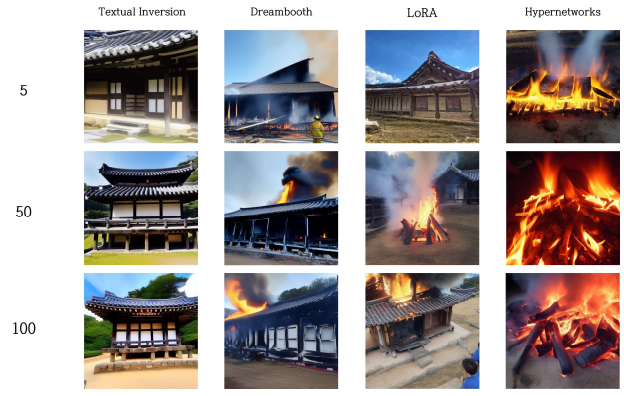

그림 4. 데이터셋 장수 별 B 프롬프트 입력 결과
Fig. 4. Results of entering prompt B by dataset size

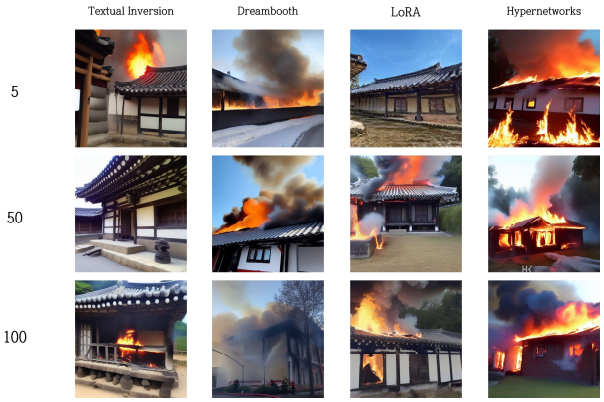

그림 5. 데이터셋 장수 별 C 프롬프트 입력 결과
Fig. 5. Results of entering prompt C by dataset size

요된 시간을 나타낸다.

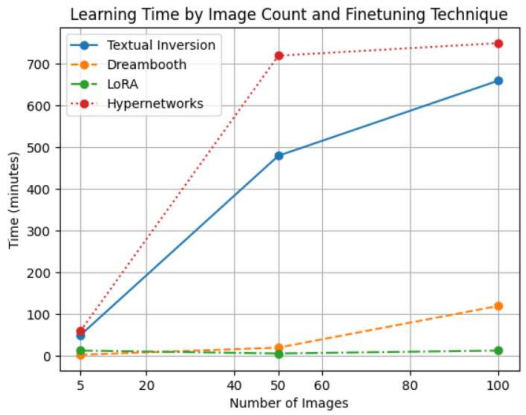

그림 6. 학습 이미지 장수 별 학습 소요 시간
Fig. 6. Learning time required by the number of training images

결과를 분석하면, 학습 소요 시간이 가장 오래 걸린 파인 튜닝 기술은 Hypernetworks로 나타났다. LoRA는 전반적으로 가장 빠른 학습 시간을 보였으나, 5장의 이미지를 학습할 때는 Dreambooth가 미세하게 더 짧은 시간을 보였다. 특히, 5장의 이미지를 학습할 때, 모든 기술은 큰 시간 차이를 보이지 않았다. 그러나 50장의 이미지 학습부터는 명확한 차이가 나타나기 시작하였다. LoRA와 Dreambooth는 상대적으로 더 빠른 학습 시간을 보였다. 반면에, Hypernetworks와 Textual inversion은 이들보다 학습 시간이 길게 걸렸다.

Ⅴ. 평가 결과

본 연구에서는 텍스트와 이미지 간의 관계성을 평가하는 기준으로 CLIP score를 채택하여 결과 이미지의 평가를 진행하였다. 기존에는 FID, IS, SSIM과 같은 다양한 이미지 생성 평가 지표들이 활용되었으나, 이러한 지표들은 원본 이미지와 생성된 이미지 간의 비교나 이미지의 품질 및 다양성 평가에 초점을 맞추었다. 따라서 기존 모델의 데이터셋에 없거나 비교 대상 이미지가 부재할 경우, 그 평가 결과가 신뢰성을 떨어뜨릴 가능성이 있었다[16]. 이러한 한계점을 고려하여, 본 연구에서는 입력 프롬프트와 생성된 이미지 간의 연관성을 정밀하게 측정하기 위해 CLIP score를 도입하여 평가를 시행하였다.

표 2. A 프롬프트 입력 시 이미지 평가 결과
Table 2. Evaluation results for image upon entering prompt A

Prompt A	Textual inversion	Dreambooth	LoRA	Hypernetworks
5	0.24	0.30	0.24	0.26
50	0.25	0.30	0.28	0.27
100	0.25	0.30	0.28	0.27
Average	0.247	0.30	0.267	0.267

표 3. B 프롬프트 입력 시 이미지 평가 결과
Table 3. Evaluation results for image upon entering prompt B

Prompt B	Textual inversion	Dreambooth	LoRA	Hypernetworks
5	0.26	0.28	0.26	0.24
50	0.25	0.30	0.26	0.26
100	0.23	0.27	0.29	0.25
Average	0.247	0.283	0.27	0.25

표 4. C 프롬프트 입력 시 이미지 평가 결과
Table 4. Evaluation results for image upon entering prompt C

Prompt C	Textual inversion	Dreambooth	LoRA	Hypernetworks
5	0.29	0.28	0.21	0.29
50	0.21	0.30	0.27	0.28
100	0.25	0.30	0.30	0.28
Average	0.25	0.293	0.26	0.283

표 2, 3, 4는 다양한 프롬프트를 기반으로 한 CLIP score의 평가 결과를 제시하고 있다. 관찰 결과, Dreambooth가 모든 평가 결과에서 가장 좋은 score를 보였으며, 반대로 Textual inversion은 모든 평균 결과에서 가장 낮은 score를 받았다.

Ⅵ. 결과 분석

다음으로 파인 튜닝 기술별 학습 결과의 특징과 원인을 분석한다.

1. Textual inversion

Textual inversion 기술은 적은 수의 데이터셋으로도 한옥의 특징을 잘 나타낸다. 그러나 일본 혹은 중국의 전통

가옥과 혼동할 수 있는 이미지를 생성하기도 했다. 이런 현상의 원인은 Textual inversion의 특징에서 알 수 있다. Textual inversion은 파인 튜닝하고자 하는 개념을 나타낼 수 있는 최적의 임베딩을 찾는 기술이다. 즉 생성되는 이미지의 결과가 기존 모델 출력 도메인의 내부에서만 생성되도록 제한되어 있고, 기존 모델이 아예 학습한 적 없는 개념을 추가로 생성할 수 없으므로 기존에 다수 학습되어 있던 해외의 전통 가옥 데이터셋과 유사하게 결과가 출력되는 것으로 추측된다. 게다가 'fire'라는 단어를 입력 프롬프트에 넣었을 때, 총 9개의 결과 이미지에서 C 프롬프트의 5장, 100장 학습 결과에서만 화재의 형상이 나타났다. 이러한 원인은 Textual inversion이 프롬프트가 길어지거나 더 복잡하게 구성될 경우 추론 능력이 떨어지는 현상이 발생하기 때문으로 예상된다[17]. 재난 이미지 생성을 위하여 국내 지형과 재난 상황을 Textual inversion을 통해 파인 튜닝해야 한다면, 기존 모델의 데이터셋에서 최적의 결과 이미지를 출력할 수 있는 프롬프트 엔지니어링 기술을 추가적으로 연구해야 한다.

Dreambooth는 학습 장수와 관계없이 대부분의 결과에서 화재에 대한 뛰어난 생성 결과를 보였다. 또한, CLIP score에서 네 가지 파인 튜닝 기술을 비교했을 때 가장 좋은 성능을 보였다. Dreambooth는 어떠한 스타일이나 개념을 학습시킬 경우 다른 상황을 추가로 입력해도 피사체의 본질과 세부 사항을 비교적 잘 유지하면서 결과를 생성한다. 하지만 Diffusion 모델 내부 U-Net과 텍스트 인코더의 모든 가중치를 파인 튜닝하기 때문에 크기와 속도의 면에서의 단점이 존재한다[18]. 실제로, Stable Diffusion 모델을 기준으로 Dreambooth를 사용하여 파인 튜닝할 경우 1GB 이상의 파라미터가 조정되며, 1,000회의 training iteration에 약 5분이 소요된다. 따라서 Dreambooth를 사용하기 위해서는 최소 12GB 이상의 VRAM과 높은 성능의 GPU가 필요하며, 이와 관련하여 용량과 속도를 축소하는 연구가 필요하다. 또한, 최적의 데이터셋 조사와 프롬프트 엔지니어링 연구가 필요할 것으로 보인다.

LoRA의 결과를 종합적으로 분석하면, 대부분 한옥 형태의 건물을 성공적으로 생성했다. 그러나 'fire'에 관련한 표현이 일관되지 않았고, 모닥불 형태의 불이 나타나는 문제가 발생했다. 그림 3, 4, 5에서 학습 결과를 보면 한옥 건물이 비슷한 위치와 모습, 그리고 색감으로 나타나는 과적합 문제가 확인되었다. 이러한 현상은 LoRA가 기존 모델의 파라미터를 변형하지 않고 rank decomposition matrices를 독점적으로 학습하여 훈련할 수 있는 파라미터를 줄이는 것을 목표로 하기 때문이다. 이 방식은 더 큰 행렬에 대해서는 확장되지 않아 특성 공간에서의 특성 변화를 효율적으로 학습할 수 없다[19]. 따라서 재난 이미지를 생성하기 위해 지형 학습에 LoRA를 사용하게 된다면, 과적합에 주의하며 일관적으로 지형과 재난 상황을 잘 표현할 수 있는 프롬프트 엔지니어링 연구가 필요할 것으로 예상된다.

Hypernetworks의 학습 결과를 분석하였을 때, 대부분의 이미지에서는 한옥의 독특한 특징을 명확히 파악하지 못했다. 'fire'라는 프롬프트 입력 시, 그림 3, 4에서는 건물과 불이 조화롭게 표현되기보다는 모닥불과 같은 형상으로 독립적으로 표현되는 경향이 있었다. 그림 4에서는 건물과 불의 조화를 보였으나 건물의 형상이 한옥처럼 보이지 않는다. 이러한 결과는 Hypernetworks가 입력 프롬프트에 크게 영향을 받으며, 학습된 지형 형태를 구체적으로 반영하지 못한다는 것을 시사한다. Hypernetworks는 전체 모델을 파인 튜닝하지 않고 cross attention 부분에만 영향을 주기 때문에 다른 기술들과 비교했을 때, 스타일 혹은 주제에 대한 학습이 어렵다[20]. 따라서 Hypernetworks를 활용하여 재난 이미지를 성공적으로 생성하기 위해서는 학습 데이터셋의 최적화와 프롬프트의 세밀한 조정 등의 추가 연구가 필요하다.

최종적으로 학습 시간과 학습 결과의 특징을 종합적으로 분석한 결과, Dreambooth가 특정 지형과 재난 상황을 표현하는데 가장 적합하다고 보인다. 그러나 본 연구에서 NVIDIA A6000 GPU를 사용한 사실을 고려해야 하며, 이는 Dreambooth의 특징인 뛰어난 GPU와 메모리가 필요하다는 점을 충족시키기 때문에 빠른 속도와 높은 파인 튜닝 성능을 보여준 것으로 판단된다. 또한, CLIP 모델에서 한옥에 대한 충분한 학습이 이루어지지 않았기 때문에 본 실험 결과에서 제시된 CLIP score가 연구 맥락에서 제한된 가치를 가질 수 있다.

Ⅶ. 결 론

본 논문은 정보인지가 어려운 취약계층을 위하여 한국 지형에 특화된 재난 이미지 생성 방법을 탐구한다. 국내 재난 이미지 생성의 가장 큰 제약 중 하나는 대규모 범용 데이터셋을 기반으로 한 대부분의 이미지 생성 모델이 특정 국가나 지형, 재난 상황에 특화되지 않았다는 것이다. 이를 해결하기 위해 본 논문에서는 한국의 전통가옥인 한옥을 대상으로 한 이미지 학습과 그 성능의 비교 분석을 하였다. 다양한 규모의 국내 지형 이미지 5장, 50장, 100장을 활용하여 파인 튜닝을 진행하고, 한옥과 화재 상황의 조화를 잘 나타내는 방법을 조사했다.

분석 결과, 특정 파인 튜닝 방법이 국내 재난 상황과 지형에 더 적합한 이미지를 생성하는 경향이 있음을 확인하였다. 그러나 이 기술들 역시 프롬프트 설계, 연산 시간, 컴퓨터의 사양, 그리고 파라미터 설정 등의 변수에 따라 성능이 달라질 수 있음을 확인했다. 이러한 이유로 추가적인 연구가 요구된다.

향후 연구에서는 본 논문의 한계를 극복하고 다양한 파인 튜닝 기술과 프롬프트 설계 전략 등 다양한 변수를 고려하여 한국형 재난 이미지 생성 방법의 완성도를 높이고자 한다. 한국의 랜드마크 학습 연구를 지속함으로써 더욱 효과적인 재난 정보 전달 방법을 제공할 수 있을 것으로 기대한다. 또한, 이 연구는 단순히 국내 재난 상황에 적용될 뿐만 아니라 다양한 지역과 상황에 활용할 수 있는 글로벌 재난 대응과 대규모 이미지 생성 기술의 발전에도 기여할 것으로 예상된다.

참 고 문 헌 (References)

[1] Federal Emergency Management Agency, “The IPAWS Symbol Set”, FEMA, 2021, IPAWS TIP #36, 2021.

[2] C. Schuhmann, R. Beanumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "Laion-5b: An open large-scale dataset for training next generation image-text models," Advances in Neural Information Processing Systems, Vol.35, pp.25278-25294, 2022.

[3] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman. "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models." 2023. doi: https://doi.org/10.48550/arXiv.2307.06949

[4] M. Arar, R. Gal, Y. Atzmon, G. Chechik, D. Cohen-Or, A. Shamir, and A. H. Bermano. "Domain-agnostic tuning-encoder for fast personalization of text-to-image models," 2023. doi: https://doi.org/10.48550/arXiv.2307.06925

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp.10684-10695, 2022. doi: https://doi.org/10.1109/cvpr52688.2022.01042

[6] Stable Diffusion WebUI, https://github.com/AUTOMATIC1111/stable-diffusion-webui (accessed Aug, 2022).

[7] Y. Alaluf, R. Gal, Y. Atzmon, O. Patashnik, A.H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. doi: https://doi.org/10.48550/arXiv.2208.01618

[8] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouber, BC, Canada, pp. 22500-22510, 2023. doi: https://doi.org/10.1109/cvpr52729.2023.02155

[9] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. doi: https://doi.org/10.48550/arXiv.2106.09685

[10] NovelAI, NovelAI Improvements on Stable Diffusion, https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac (accessed 2022).

[11] SK. Shon and MS. Kim, " A Study on Modernization Elements of Traditional Hanok and Character of a New One.” *Journal of the Korean Housing Association*, Vol.24, pp.77-87, 2013. doi: https://doi.org/10.6107/JKHA.2013.24.2.077

[12] ES. Cho and YS. Park, "A Comparative Study on the Image characteristics in Traditional Palaces of Korea, China and Japan," *Archives of Design Research*, Vol.18, No.59, pp.27-38, 2005.

[13] SH. Lee, "Comparative Evaluation of Gate and Wall Color in Traditional Village-A Case of Traditional Housing on Hahe Village in Korea and Makabe-machi in Japan," *Archives of Design Research,* Vol.24, No.2, pp.299-308, 2011.

[14] H. Park and W. Yi, "Proposed Improvements to the Safety Inspection System: An Analysis of the Current Status of Building Disaster Accidents and Safety Inspection Systems," *Journal of the Korean Society of Hazard Mitigation*, Vol.19, No.5, pp.11-21, 2019. doi: https://doi.org/10.9798/KOSHAM.2019.19.5.11

[15] AI HUB, Utilization of Metaverse for Traditional House Learning Data, https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71460 (accessed 2022).

[16] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," 2023.

doi: https://doi.org/10.48550/arXiv.2303.11305
[17] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 4, pp. 1-10, 2023.
doi: https://doi.org/10.1145/3592116
[18] Y. Tewel, R. Gal, G. Chechik, and Y. Atzmon, "Key-locked rank one editing for text-to-image personalization," *Proceeding of ACM SIGGRAPH 2023 Conference*, pp. 1-11, 2023.
doi: https://doi.org/10.1145/3588432.3591506
[19] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-All: Generalized LoRA for Parameter-Efficient Fine-tuning," 2023.
doi: https://doi.org/10.48550/arXiv.2306.07967
[20] R. Gal, M. Arar, Y. Atzmon, A.H. Bermano, G. Chechik, and D. Cohen-Or, "Encoder-based domain tuning for fast personalization of text-to-image models," *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 4, pp. 1-13, 2023.
doi: https://doi.org/10.1145/3592133

저 자 소 개

최 민 지

- 2022년 8월 : 홍익대학교 과학기술대학 전자·전기공학과 학사
- 2022년 9월 ~ 현재 : 과학기술연합대학원대학교(UST) 정보통신공학 석사과정
- ORCID : https://orcid.org/0009-0007-5851-6535
- 주관심분야 : 컴퓨터 비전, Deep Generative model, Text-to-Image generation

원 루 빈

- 2022년 11월 : University of Alberta Computer Science 학사
- 2023년 3월 ~ 현재 : 과학기술연합대학원대학교(UST) 정보통신공학 석사과정
- ORCID : https://orcid.org/0009-0007-2664-0507
- 주관심분야 : Deep Generative model, Text-to-Image generation, Diffusion model

최 지 훈

- 1999년 2월 : 경희대학교 전자공학과 학사
- 2001년 2월 : 경희대학교 전자공학과 석사
- 2001년 3월 ~ 현재 : 한국전자통신연구원 미디어연구본부 미디어지능화연구실 책임연구원
- ORCID : https://orcid.org/0000-0002-3402-1921
- 주관심분야 : UHD 방송 기술, 재난정보미디어 서비스, AI 미디어 처리 기술

배 병 준

- 1997년 2월 : 경북대학교 전자공학과 석사
- 2006년 8월 : 경북대학교 전자공학과 박사
- 1997년 2월 ~ 2000년 10월 : ㈜엘지전자 주임연구원
- 2000년 11월 ~ 현재 : 한국전자통신연구원 미디어연구본부 미디어지능화연구실 책임연구원
- 2012년 3월 ~ 현재 : 과학기술연합대학원대학교(UST) ETRI스쿨 통신미디어공학 주임교수
- 2017년 1월 ~ 현재 : 한국정보통신기술협회(TTA) WG8028(시스템및코덱정합테스트실무반) 의장
- ORCID : https://orcid.org/0000-0002-0872-325X
- 주관심분야 : UHD 방송 기술, 재난정보미디어 서비스, AI 미디어 처리 기술