creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

2023년 8월
박사학위 논문

멀티모달 AI를 적용한 웹툰 생성 연구

조선대학교 대학원

컴퓨터공학과

유 경 호

조선대학교
CHOSUN UNIVERSITY

멀티모달 AI를 적용한 웹툰 생성 연구

A Study on Webtoon Generation Using Multimodal AI

2023년 8월 25일

조선대학교 대학원

컴퓨터공학과

유 경 호

조선대학교
CHOSUN UNIVERSITY

멀티모달 AI를 적용한 웹툰 생성 연구

지도교수 김 판 구

이 논문을 공학박사학위 신청논문으로 제출함

2023년 4월

조선대학교 대학원

컴퓨터공학과

유 경 호

조선대학교
CHOSUN UNIVERSITY

유경호의 공학박사학위논문을 인준함

위원장 조선대학교 교 수 양 희 덕 (인)

위 원 조선대학교 교 수 전 찬 준 (인)

위 원 UST 교 수 황 명 권 (인)

위 원 William Paterson University 교 수 임 기 호 (인)

위 원 조선대학교 교 수 김 판 구 (인)

2023년 6월

조선대학교 대학원

조선대학교
CHOSUN UNIVERSITY

목 차

조선대학교
CHOSUN UNIVERSITY

조선대학교
CHOSUN UNIVERSITY

표 목 차

조선대학교
CHOSUN UNIVERSITY

그 림 목 차

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

A Study on Webtoon Generation Using Multimodal AI

Yu, Kyoungho
Advisor : Prof. Kim, Pankoo, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

In this thesis, I conducted research on generating webtoons using deep learning-based text-to-image generation technology to assist webtoon creators in their creative activities. The research methodology involved constructing a multimodal webtoon dataset by using publicly available datasets such as MSCOCO. The generated dataset consists of treatments (that is text descriptions) and their corresponding webtoon - treatment-webtoon dataset. Furthermore, continuous text data was also collected using ChatGPT. To generate webtoon, this thesis proposed to utilize a multilingual BERT model for feature extraction from the treatments, add noise to them, and input the noisy features into a DCGAN (Deep Convolutional Generative Adversarial Network). The experimental results showed relatively low performance with an inception score of 4.9 and FID (Fréchet Inception Distance) of 22.21.

To overcome the limitations of DCGAN, this thesis proposed to train the CLIP (Contrastive Language-Image Pretraining) model on the treatment-webtoon dataset by measuring the similarity between text and images and then use the diffusion model to generate webtoon. CLIP is a model that can learn a relationship between multimodal data (such as the treatment-webtoon dataset)

조선대학교
CHOSUN UNIVERSITY

by extracting features from each data modality. The goal is to bring similar data closer and dissimilar data farther apart in the same feature space, which is achieved by contrastive learning. In this thesis, the performance of CLIP trained on the treatment-webtoon dataset was evaluated using quantitative metrics such as measuring similarity between bilingual treatments and images, treatment-based search queries for similar images, and zero-shot classification.

For generating webtoons based on the diffusion model, a desired text along with its CLIP features and the image with the most similar CLIP features were inputted into a pretrained depth-to-image model. In the experiments, webtoons were generated using both single text and continuous text inputs. The results showed that when using continuous text inputs for webtoon generation, the inception score improved from 4.9, as in the case of DCGAN, to 7.14 and the generated images were of higher quality.

The technology developed in this thesis can be used by webtoon creators by inputting their desired text to generate webtoons more efficiently and in a timely manner. However, one of the main limitations of this work is that currently it cannot generate webtoons from multiple sentences and images or maintain consistent artistic style throughout the generated images. Therefore, further research is needed on diffusion models that can handle multiple sentences as inputs and generate images with consistent artistic styles when continuous text inputs are provided.

조선대학교
CHOSUN UNIVERSITY

Ⅰ. 서 론

A. 연구의 배경 및 목적

딥러닝 기술이 발전하면서 이미지를 컴퓨터에 학습시키고 컴퓨터 스스로 이미지를 생성하는 연구가 활발히 진행되고 있다[1]. 이미지 생성 기술은 생성자와 판별자의 적대적 학습을 통해 원본과 유사한 가짜 이미지를 생성하는 적대적 생성 신경망을 시작으로 하여 발전하였으며 현재에는 확산모델을 이용한 이미지 생성은 실제 사람이 그리는 것과 비교하였을 때 구분할 수 없을 정도로 사실적인 이미지를 생성하고 있다. 최근 딥러닝 기반의 이미지 생성기술은 사람이 그린 것과 유사한 이미지를 생성할 수 있기 때문에 가상 인물 생성, 애니메이션 등과 같은 엔터테인먼트 분야에 활용되고 있다[2,3].

이미지 생성기술과 더불어 사람이 문장을 보고 이미지를 연상하는 과정을 컴퓨터가 할 수 있도록 하는 텍스트에서 이미지 생성(Text-to-Image)기술 또한 발전되고 있다[1,2]. 딥러닝 기반의 텍스트에서 이미지 생성은 생성하고자 하는 이미지에 대한 설명을 딥러닝 모델에 입력하였을 때 이미지를 생성하는 것을 말한다. 이것이 가능한 이유는 이미지 생성모델을 학습할 때 텍스트와 이미지 사이의 관계를 학습하기 때문에 텍스트를 입력으로하여 이미지를 생성할 수 있다. 텍스트에서 이미지 생성은 이미지와 텍스트를 함께 학습하는데, 이와같이 하나의 정보에 대해 여러 종류의 데이터로 표현된 것을 멀티모달(Multimodal)이라하고 멀티모달 사이의 관계(유사한 정도)를 학습하여 멀티모달 형태로 출력할 수 있는 것을 멀티모달 AI라 한다[4,5]. 예를 들어 강아지에 대한 정보에는 강아지의 이미지, 강아지의 울음소리, 이름과 같은 멀티모달로 나타낼 수 있다. 수많은 강아지의 멀티모달을 학습하여 강아지에 대한 설명을 입력하면 어떤 강아지인지 이미지를 생성할 수 있으며 또한 강아지의 이미지를 입력하면 어떤 강아지인지 분류할 수 있는 것을 멀티모달 AI라 한다.

조선대학교
CHOSUN UNIVERSITY

웹툰은 Web과 cartoon의 합성어로 인터넷을 통해 연재되는 만화이다. 웹툰의 창작 과정은 스토리 기획, 콘티/스케치, 채색, 배경 그리기로 나뉜다[6]. 웹툰 제작은 각각의 과정에 사람이 직접 개입하기 때문에 시간이 오래 걸리고 소요되는 비용이 많이 든다. 그렇기 때문에 사람이 직접 관여하는 단계를 최소화하기 위해 인공지능 기술을 활용하여 자동 채색, 자동 라인드로잉, 화풍 변환과 같은 기술을 사용하고 있다[7,8].

웹툰이 독자들에게 매력적으로 다가가기 위해서는 웹툰의 전반적인 스토리가 중요하다[9]. 웹툰이 진행됨에 있어 독자들이 흥미를 잃지 않게 하려면 긴장감을 주거나 해소 시키면서 몰입할 수 있도록 해야 한다. 독자들의 흥미와 몰입을 놓치지 않기 위해서 작가는 웹툰을 제작하기 전에 스토리를 기획하는 단계를 거친다. 스토리를 기획하는 단계에서는 작가는 장르와 캐릭터 그리고 세계관 등을 설정하게 된다. 그 후 웹툰을 그리기 전에 웹툰의 한 장면에 대해 자세한 서술을 글로 나타내는데 이것을 트리트먼트라 한다. 트리트먼트는 웹툰뿐만 아니라 영화나 드라마 등 콘텐츠를 제작할 때 사용하는 용어로서, 한 장면에 대해 시간과 장소에 따라 등장인물, 주요 사건 등 장면의 핵심을 글로 나타내어 영화나 드라마 촬영, 웹툰을 그릴 때 그릴 때 트리트먼트를 참고하여 제작한다. 이와 같이 트리트먼트는 최종 결과물인 웹툰의 장면에 대한 풍부한 정보를 담고 있기 때문에 딥러닝 기반의 텍스트에서 이미지 생성에 활용하기에 적합한 데이터이다.

딥러닝 기반 텍스트에서 이미지 생성 모델에 트리트먼트와 웹툰을 학습시킨다면 트리트먼트를 텍스트에서 이미지 생성 모델에 입력하였을 때 트리트먼트와 의미적으로 유사한 웹툰을 생성할 수 있다. 따라서, 본 연구에서는 웹툰 저작자가 웹툰 저작활동에 도움을 줄 수 있도록 딥러닝 기반으로 멀티모달 데이터를 이용하여 웹툰을 생성하는 것이 목적이다.

조선대학교
CHOSUN UNIVERSITY

B. 연구 내용

본 연구에서는 텍스트-이미지와 같은 멀티모달 데이터를 딥러닝 기반의 생성 모델에 학습하고, 학습이 끝난 후 텍스트를 입력하였을 때 웹툰 스타일의 이미지를 생성할 수 있도록 한다.

딥러닝 모델을 이용하여 웹툰을 생성하기 위해, 이에 적합한 멀티모달 데이터를 구축한다. 데이터 셋의 구축은 이미지와 이 이미지에 대한 설명 텍스트로 구성된 공개 데이터 셋인 MSCOCO 데이터 셋을 사용하며, 이미지는 CartoonGAN[10]을 이용하여 툰 스타일의 이미지로 변환한다. 그리고 연속된 웹툰 스타일의 이미지를 생성하기위해 텍스트 생성모델을 이용하여 하나의 텍스트에서 다음에 나올 수 있는 연속된 3개의 문장을 생성한다. 구축한 데이터 셋은 학습을 위한 텍스트-웹툰 데이터 셋, 웹툰 스타일 생성을 위한 4개의 연속된 텍스트 데이터 셋으로 구성된다.

구축한 멀티모달 데이터를 사용하여 딥러닝기반의 텍스트에서 이미지 생성 모델에 학습한다. 텍스트를 딥러닝 모델에 입력하였을 때 웹툰 스타일의 이미지를 생성하는 방법은 크게 적대적 생성 신경망 기반의 생성모델을 사용하는 방법과 확산 모델을 생성하는 방법이 있다.

적대적 생성 신경망 기반의 이미지의 생성은 생성자와 판별자의 경쟁적인 학습을 통해 진짜 이미지와 비슷한 가짜 이미지를 생성한다. 학습은 텍스트에서 특징을 추출하고 노이즈와 결합하여 적대적 생성 신경망의 생성자에 입력하여 학습한다. 그 후 생성자에서 생성한 가짜 이미지와 입력한 텍스트에 맞는 진짜 이미지를 판별자에 입력하여 입력한 이미지가 가짜 이미지인지, 진짜 이미지 인지 판별하여 생성자와 판별자의 weight를 업데이트 한다.

확산 모델(Diffusion model)을 이용한 이미지의 생성은 텍스트-웹툰 데이터

조선대학교
CHOSUN UNIVERSITY

셋의 CLIP(Contrastive Language-Image Pre-training)을 이용한 두 데이터 사이의 유사도를 측정하는 멀티모달 학습(Multimodal Learning), 그리고 텍스트를 CLIP에 입력하여 출력된 이미지를 확산 모델에 입력하여 이미지를 생성하는 단계로 진행된다. 본 연구에서 사용한 CLIP의 구조는 텍스트와 이미지의 특징을 추출하기위해 사전학습된 transformer 계열의 BERT와 Vision transformer 모델에 텍스트와 이미지를 하나의 잠재 공간에 위치시키기 위한 투영 계층을 추가한 구조이다. 텍스트와 이미지가 CLIP 모델의 인코더를 통과하면 동일한 차원에 위치시키게 되고 대조 학습(contrastive learning)을 통해 각각의 인코더의 가중치를 업데이트 한다. 이렇게 학습된 CLIP 모델에 텍스트나 이미지를 입력하면 동일한 차원에 위치한 특징 벡터를 출력할 수 있다. 따라서, 텍스트를 CLIP 모델에 입력하면 입력된 텍스트의 특징을 출력할 수 있으며, 사전에 CLIP 통해 출력된 학습 이미지의 특징벡터와 유사도를 측정하여 유사도가 높은 이미지를 얻을 수 있다. 마지막으로 웹툰 처럼 연속적인 이미지를 생성하기 위해 연속된 4개의 문장으로 구성된 텍스트를 CLIP에 입력하여 각각의 문장에 대해 출력된 유사한 이미지를 사전학습된 확산 모델 기반의 생성모델에 입력하여 웹툰을 생성한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구와 관련성이 높은 연구에 대해 설명하고, 제 3장에서는 멀티모달인 트리트먼트-웹툰 데이터 셋을 적대적 생성 신경망에 학습하여 웹툰을 생성한다. 제 4장에서는 트리트먼트-웹툰 데이터 셋을 CLIP를 사용하여 멀티모달 학습을 수행한다. 제 5장에서는 4장에서 학습한 CLIP와 확산 모델을 사용하여 웹툰을 생성한다. 마지막 제 6장에서는 본 연구의 결론과 제한점, 향후 연구에 대해 논한다.

조선대학교
CHOSUN UNIVERSITY

Ⅱ. 관련연구

A. 딥러닝 기반 텍스트에서 이미지 생성

1. 적대적 생성 신경망

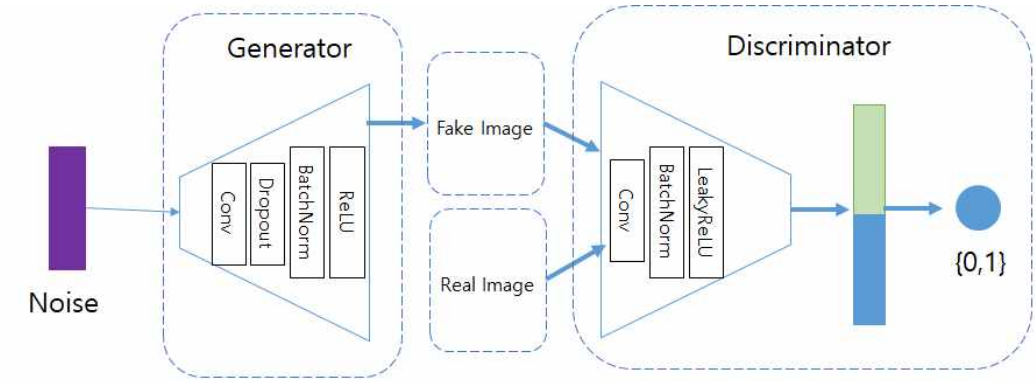


그림 1. Deep Convolutional Generative Adversarial Network의 구조

적대적 생성 신경망(Generative Adversarial Network)는 생성자(generator)와 판별자(discriminator) 두 가지 네트워크로 구성된 모델이다. 그림 1과 같이 생성자와 판별자라 불리는 두 가지 신경망의 경쟁적 학습을 통해 진짜 이미지와 유사한 가짜 이미지를 생성한다[11].

적대적 생성 신경망의 학습은 랜덤 노이즈 벡터를 생성자에 입력하는 것으로부터 시작된다. 그림 1과 같이 생성자는 랜덤 노이즈 벡터를 여러 개의 convolutional layer를 거쳐 가짜 이미지를 생성하고 판별자는 생성자로부터 출력된 데이터가 진짜 데이터인지 가짜 데이터인지 판별한다. 그 후 적대적 생성 신경망의 목적함수(식 1)을 통해 각각의 네트워크를 업데이트 하여 진짜와 같은 가짜 이미지를

조선대학교
CHOSUN UNIVERSITY

생성할 수 있다.

$$\min_{G}\max_{D} V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log\ 1 - D(G(z)))] \quad (1)$$

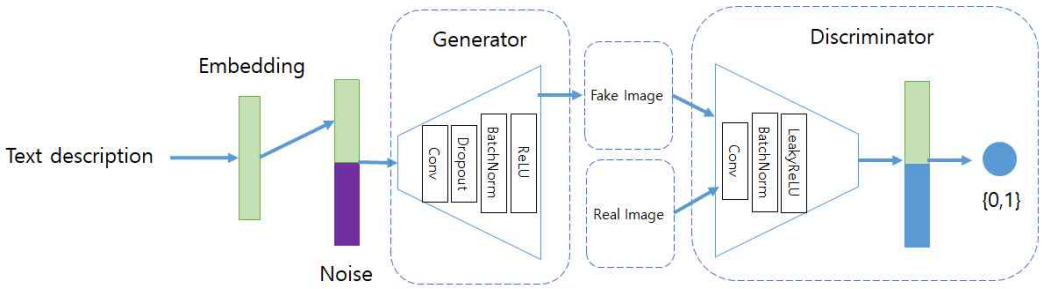


그림 2. DCGAN을 이용한 텍스트에서 이미지 생성 모델의 구조

적대적 생성 신경망 기반의 텍스트에서 이미지 생성은 텍스트에서 의미적 정보를 추출하는 단계와 이미지 생성 단계로 나눌 수 있다. 그림 2와 같이 의미적 정보를 추출하는 단계에서는 문장의 워드 임베딩을 transformer 구조에 입력하여 특징을 추출한다. 이미지 생성 단계에서는 적대적 생성 신경망에 텍스트의 의미적 정보와 생성하고자 하는 이미지 크기의 랜덤 노이즈 벡터를 입력하여 생성자를 통해 이미지를 생성한다. 판별자는 앞에서 설명한 것과 같이 생성자가 생성한 이미지가 가짜인지 아닌지 판별 한다.

조선대학교
CHOSUN UNIVERSITY

DCGAN 기반 텍스트에서 이미지 생성 모델은 문장의 특징 벡터를 조건으로 사용하여 이미지를 생성한다. 생성된 이미지는 입력된 문장의 의미와 어느 정도 유사한 이미지를 생성하지만 문장에서 각 단어의 문맥적 의미를 반영하지 못하고 저화질의 이미지를 생성하는 단점이 있었다. 이러한 단점을 극복하기 위해 이미지 생성에 어텐션 메커니즘 도입한 AttnGAN[12]은 문장의 특징벡터를 사용하여 이미지를 먼저 생성하고 다음 단계의 이미지를 생성할 때 단어의 어텐션 맵을 이미지 벡터와 결합하여 단계적으로 향상된 이미지를 생성한다(그림 3). AttnGAN은 DCGAN과 비교하였을 때 문장안에 있는 각 단어에 대한 세밀한 표현까지 가능하였으며 그 이후 입력된 문장의 의미를 더 세밀하게 표현하고 고해상도의 이미지를 생성할 수 있는 stackGAN[13], MirrorGAN[14], R-GAN[15]이 연구되었다.

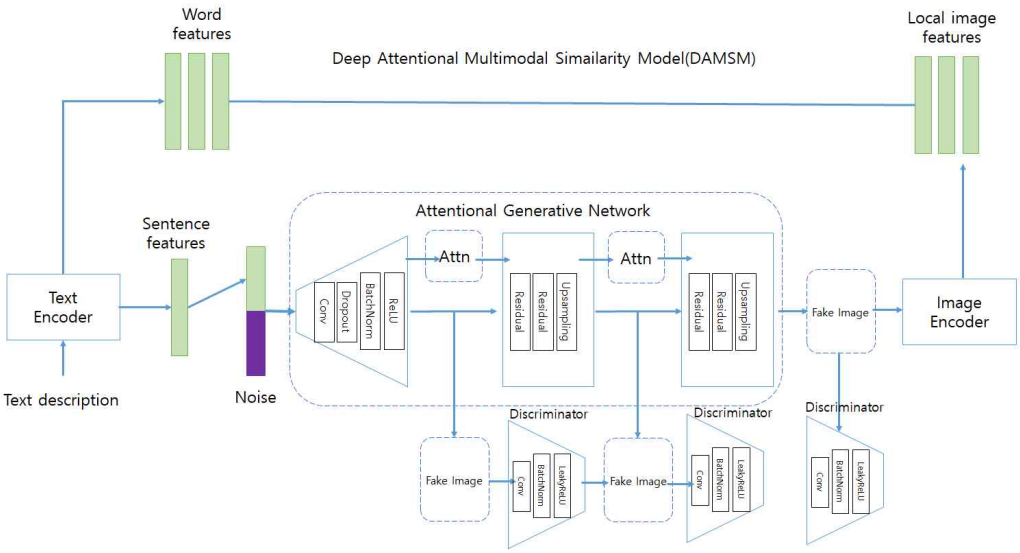


그림 3. AttnGAN 모델의 구조

조선대학교
CHOSUN UNIVERSITY

2. 확산 모델

확산 모델(diffusion model)은 이미지 생성에 사용되는 확률적인 모델링 방법이다. 이 모델은 이미지의 확률 분포를 모델링하고, 이를 통해 새로운 이미지를 생성하는 방법을 제공한다[16]. 확산 모델은 확률적인 접근 방법을 사용하기 때문에 이미지 생성 시 다양한 결과물을 얻을 수 있다는 장점이 있다. 그리고 이미지 생성 과정에서 학습된 정보를 재사용할 수 있기 때문에, 이미지 생성의 효율성을 높일 수 있다.

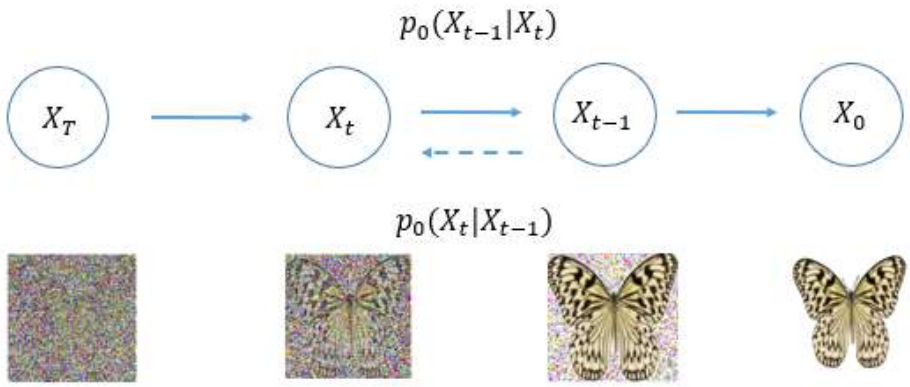


그림 4. 확산 모델의 학습 과정

DDPM(Denoising Diffusion Probabilistic Models)에서 사용되는 확산 모델의 확산 과정(diffusion process)는 크게 두 가지 구성 요소로 이루어져 있다[15]. 그림 4에서 볼 수 있듯이 하나는 정방향 확산 과정이고, 다른 하나는 역방향 확산 과정이다. 그림 4에서 오른쪽에서 왼쪽으로 진행되는 정방향 확산 과정은 이미지의 픽셀 값에 스텝마다 노이즈를 조금씩 추가하는 방법으로 이미지를 확산시키는 과정이다. 그림 4에서 왼쪽에서부터 오른쪽으로 진행되는 역방향 확산 과정은 이미지를 역으로 확산시켜 노이즈에서 원본 이미지를 복원하는 과정이다. 이 두 과정을 반복하면, 최종적으로 노이즈에서 새로운 이미지를 생성할 수 있다.

조선대학교
CHOSUN UNIVERSITY

확산 모델의 학습은 대규모 이미지 데이터셋을 사용하여 학습된다. 이 과정에서는 역방향 확산 과정을 통해 학습된 이미지와 실제 이미지의 차이를 최소화하는 방향으로 학습이 이루어진다. 정방향 확산 과정은 학습 과정에서는 사용되지 않지만, 이미지 생성 과정에서는 노이즈를 추가하는 과정으로 활용된다.

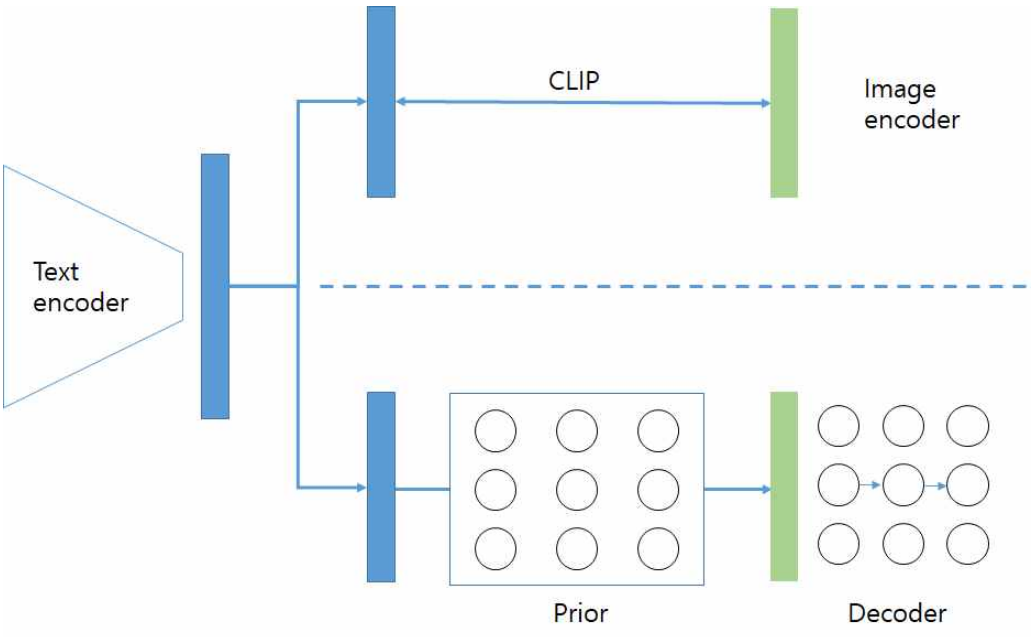


그림 5. Dalle-2(unCLIP) 모델의 구조

확산 모델을 기반으로 한 텍스트에서 이미지 생성 모델은 대표적으로 Dalle-2와 stable diffusion 모델이 있다[17,18]. Dalle-2는 unCLIP라고도 불리며 구조는 그림 5와 같다. Dalle-2는 CLIP, prior, 디코더 세부분으로 구성되어 있으며 학습은 대규모의 텍스트-이미지 쌍으로 구성된 웹 데이터셋을 CLIP에 학습하여 텍스트와 이미지 사이의 상호작용(유사도)을 기반으로 학습한다(그림 5의 점선 상단).

이미지를 생성하기 위해서 사전에 학습한 CLIP와 prior, 디코더 단계를 거쳐 이미지를 생성한다(그림 5의 점선 하단부분). 생성하고자 하는 이미지에 대한 텍스트를 CLIP 모델의 텍스트 인코더에 입력하여 CLIP 텍스트 임베딩을 생성한다. prior 단계에서는 확산 모델 기반의 prior을 사용하며 CLIP 텍스트 임베딩을 확산

조선대학교
CHOSUN UNIVERSITY

prior에 입력하여 CLIP 이미지 이미지를 생성한다. 디코더는 수정된 GLIDE(Guided Language to Image Diffusion for Generation and Editing)[19]를 사용한다. GLIDE는 확산 모델기반의 생성모델인데 Dalle-2에서는 인코딩된 텍스트, CLIP 텍스트 임베딩, noised CLIP 이미지 임베딩, timestep 임베딩 총 4가지를 입력으로 하여 이미지를 생성한다.

Stable diffusion 모델 또한 확산 모델을 기반으로 한 생성모델이다. 구성요소는 CLIP 기반의 텍스트 인코더, UNet과 scheduler로 구성된 정보 생성자, 이미지를 생성하는 디코더로 구성된다. Stable diffusion은 픽셀 공간(pixel space)에서 이미지를 생성하는 것이 아니라 잠재 공간(latent space)에서 이미지를 생성하기 때문에 빠른 이미지를 생성할 수 있다. 최근 딥러닝 기반의 생성모델은 텍스트에서 이미지 생성 뿐만아니라 다른 도메인에서도 확산 모델 기반의 생성모델 연구가 주를 이루고 있다[20-23].

조선대학교
CHOSUN UNIVERSITY

B. 멀티모달 학습

멀티모달(Multimodal) 학습은 하나의 정보를 여러 종류의 데이터로 표현하는 것을 말하며 이미지, 텍스트, 음성 등 다양한 모달리티를 고려하여 특징을 추출하고 학습하는 방법이다.

OpenAI에서 발표한 CLIP 모델은 멀티모달 학습 모델의 종류 중 하나이다[24]. CLIP는 이미지와 텍스트 두 가지 모달리티를 함께 학습하는데, CLIP는 이미지와 텍스트의 특징을 동일한 차원의 잠재 공간에 위치시키고 특징 사이의 연관성 분석을 위해 대조 학습을 수행한다. 대조적 학습은 데이터를 비교하여 유사한 데이터의 특징은 가까운 거리에 위치시키고, 서로 다른 데이터의 특징은 멀리 위치시키도록 학습한다. 이러한 대조 학습을 통해 서로 다른 도메인의 데이터 사이의 관계를 알 수 있다.

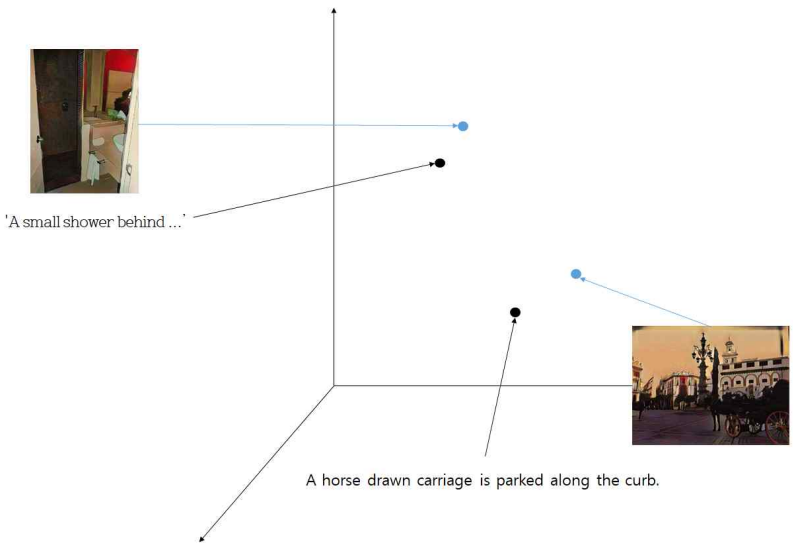


그림 6. CLIP를 이용한 멀티모달 데이터의
joint embedding space

조선대학교
CHOSUN UNIVERSITY

CLIP의 장점은 멀티모달 데이터를 동일한 잠재 공간에 위치시킴으로써 멀티모달 데이터 사이의 관계를 학습할 수 있다. 이를 통해 CLIP는 이미지와 텍스트를 연결시키고, 유사도를 측정할 수 있다. 이러한 장점을 바탕으로 OpenAI에서 발표한 Dalle-2와 Google에서 발표한 Dreambooth 같은 모델들은 CLIP를 활용하여 텍스트와 이미지 간의 유사도를 측정하거나, 텍스트나 이미지의 특징을 추출하는 인코더로 사용된다.

```
# normalized features
image_embeds = image_embeds / np.linalg.norm(image_embeds, axis=-1)
text_embeds = text_embeds / np.linalg.norm(text_embeds, axis=-1)

# cosine similarity as logits
logits_per_text = matmul(text_embeds, image_embeds.T) * logit_scale
logits_per_image = logits_per_text.T

# symmetric loss function
labels = np.arrange(n)
loss_i = cross_entropy_loss(logits_per_image, labels, axis=0)
loss_t = corss_entropy_loss(logits_per_text, labels, axis=1
loss = (loss_i + loss_t) / 2
```

그림 7. CLIP의 대조 학습 알고리즘

CLIP의 구조는 이미지와 텍스트의 특징을 추출하는 각각의 인코더에 같은 차원의 투영계층이 추가된 구조이다. 텍스트가 CLIP에 입력되면 텍스트 인코더를 통과하여 특징을 추출하고 마지막 투영계층을 거쳐 임베딩 된다. 이미지 또한 텍스트와 동일한 방법으로 이미지 인코더를 통해 특징을 추출하고 투영계층을 거쳐 임베딩 된다. CLIP의 학습은 그림 7과 같이 대조 학습을 통해 학습하는데, 각각의 인코더를 통해 같은 차원의 특징을 추출한 뒤 코사인 유사도를 구하기 위해 정규화 한다. 그 후 텍스트와 이미지 임베딩 벡터 사이의 내적을 계산한다. 위 코드에서 ‘logits_per text’는 텍스트별 이미지 사이의 유사도 값으로 이루어진 행렬이며, ‘logits per image’는 이미지별 텍스트 사이의 유사도 값으로 이루어진 행렬이다. 이후 cross entropy를 계산하여 loss를 각각의 네트워크를 업데이트하며 학습한다.

조선대학교
CHOSUN UNIVERSITY

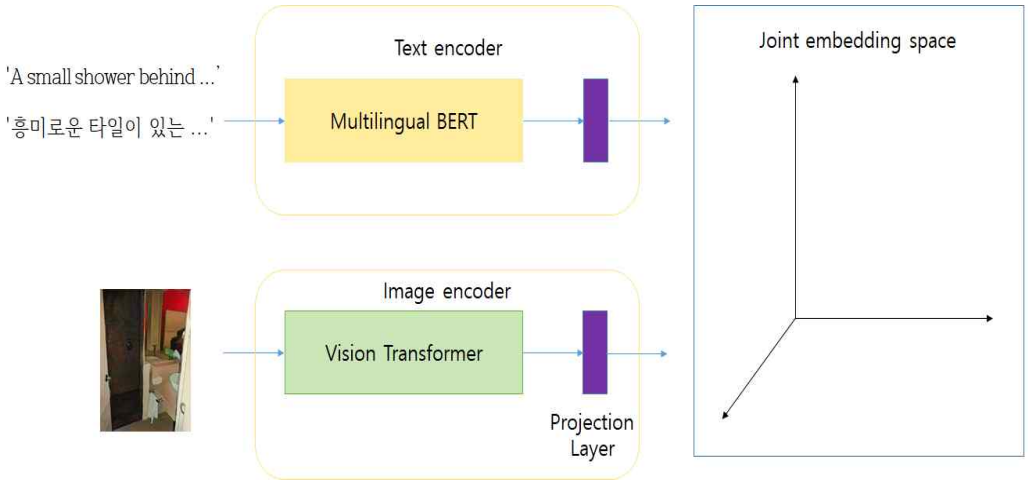


그림 8. 텍스트, 이미지 인코더로 구성된 CLIP 모델의 구조

학습이 끝난 후 텍스트와 이미지의 임베딩 벡터(특징 벡터)를 각각 독립적으로 계산할 수 있다. 예를 들어 '한 남자가 길을 걷고 있다' 라는 문장의 CLIP 임베딩 벡터를 얻기 위해서는 CLIP의 텍스트 인코더에 입력하여 입베딩 값을 계산할 수 있으며 이와 유사한 방식으로 이미지 또한 CLIP의 이미지 인코더로 계산할 수 있다.

조선대학교
CHOSUN UNIVERSITY

Ⅲ. 적대적 생성 신경망 기반의 멀티모달 데이터를 이용한 웹툰 생성

A. 적대적 생성 신경망을 이용한 텍스트에서 웹툰 생성 방법

본 절에서는 영어와 한국어로 구성된 다국어 텍스트를 텍스트에서 이미지 생성 모델에 입력하였을 때 웹툰을 생성할 수 있도록 다국어 BERT와 DCGAN으로 이루어진 텍스트에서 이미지 생성 모델을 사용한다[25]. 다국어 BERT와 DCGAN을 이용한 텍스트에서 이미지 생성 모델은 영어와 한글로 이루어진 텍스트를 텍스트에서 이미지 생성 모델에 입력하였을 때, 입력한 텍스트에 유사한 웹툰을 생성하는 것을 목적으로 한다.

첫번째 단계에서는 다국어 텍스트-웹툰 데이터 셋을 구축하기 위해 벤치 마크 데이터 셋인 MSCOCO 데이터 셋을 CartoonGAN을 사용하여 웹툰 데이터 셋을 구축한다.

그 다음 학습 단계에서는 구축한 다국어 트리트먼트-웹툰 데이터 셋의 다국어 트리트먼트를 다국어 BERT 모델을 사용하여 문장 벡터를 추출한 뒤 DCGAN에 문장 벡터를 조건으로 주어 학습한다.

Original	Hayao	Hosoda	Parika	shinkai

그림 9. CartoonGAN을 이용한 이미지 변환의 예

조선대학교
CHOSUN UNIVERSITY

딥러닝 기반의 텍스트에서 이미지 생성 모델에 트리트먼트를 입력하였을 때 웹툰 이미지를 생성하기 위해서는 트리트먼트와 웹툰 데이터 셋을 텍스트에서 이미지 생성 모델에 학습시키는 것이 필요하다. 실제 트리트먼트-웹툰으로 이루어진 대용량의 데이터 셋을 구축하는 것은 어렵기 때문에 생성모델에서 자주 사용되는 벤치마크 데이터 셋인 MSCOCO 데이터 셋[26]을 사전학습된 CartoonGAN을 사용하여 카툰이미지로 변형시킨다. MSCOCO 데이터 셋은 Microsoft에서 공개한 설명문-실사 이미지 데이터 셋으로 학습, 평가 이미지 123,287장의 이미지로 구성되어 있으며 각 이미지당 5개의 설명문으로 이루어져 있다. 공식적으로 제공하는 데이터 셋은 영어만을 제공하기 때문에 다국어 트리트먼트 데이터 셋을 구축하기 위해서 한국의 AI hub에서 MSCOCO 데이터 셋을 번역한 한국어 MSCOCO 데이터 셋을 추가로 사용하였다[27]. 한국어 MSCOCO 데이터 셋 또한 한 이미지당 5가지의 한국어 설명문으로 구성되어 있어, 다국어 트리트먼트는 한 이미지당 10개의 문장(영어 5문장, 한국어 5문장으로 구성된다. 사전학습된 CartoonGAN은 그림 9와 같이 4가지 화풍으로 학습되어있다.

따라서 하나의 화풍에 대해 1,232,870개의 다국어 트리트먼트-웹툰 데이터 쌍을 구축할 수 있으며 4가지 화풍에 대해 4,931,480개의 트리트먼트-웹툰 데이터 쌍을 구축할 수 있다.

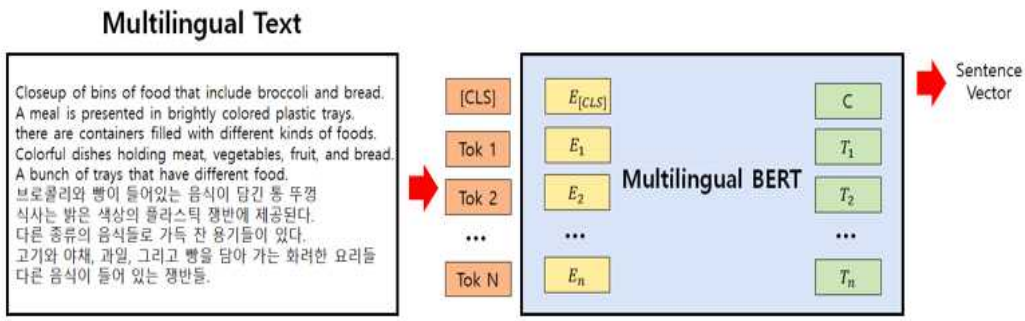


그림 10. 다국어 BERT를 이용한 sentence vector 추출

조선대학교
CHOSUN UNIVERSITY

딥러닝 기반의 텍스트에서 이미지 생성 모델은 텍스트를 입력으로 하였을 때 두 단계를 통해 이미지를 생성한다. 첫번째 단계는 텍스트의 특징을 추출하는 단계이고 두번째는 추출된 텍스트의 특징을 노이즈와 결합하여 적대적 생성 신경망 모델에 입력하여 학습하는 단계이다. 이 중 텍스트의 특징을 추출하는 단계가 중요한데, 텍스트에 내포된 문맥적 특성을 반영하는 특징을 추출하여 학습할수록 생성된 결과가 크게 달라진다. 다시 말하면, 이미지와 텍스트 사이에서 핵심적인 단어에 대한 특징을 추출하지 못한다면, 학습된 텍스트에서 이미지 생성 모델에서 생성되는 이미지는 텍스트와 동떨어진 이미지를 생성할 수 있다.

따라서, 본 연구에서는 그림 10과 같이 다국어 트리트먼트를 NLP에서 높은 성능을 보인 사전학습된 다국어 BERT를 사용하여 문장의 특징 벡터를 추출한다[28]. 문장의 특징 벡터는 BERT 모델에서 마지막 히든 스테이트를 가리키는 토큰을 문장벡터로 사용한다.

추출한 다국어 트리트먼트의 특징에서 웹툰을 생성하기 위해 적대적 생성 신경망 기반의 DCGAN 모델을 사용한다. 그림 11은 DCGAN을 이용한 텍스트에서 이미지 생성 모델의 구조이다. 적대적 생성 신경망은 생성자와 판별자라 불리는 두 가지 신경망의 경쟁을 통해 원본데이터와 유사한 이미지를 생성한다. 생성자는 입력 데이터의 분포를 반영하여 학습하고 랜덤 벡터로부터 가짜 데이터를 생성하고 판별자는 생성자가 생성한 데이터를 원본 데이터와 비교하여 진짜인지 가짜인지 판별하게 된다. 다국어 트리트먼트에서 웹툰을 생성하기 위해 다국어 BERT를 통해 추출한 문장 벡터를 노이즈와 결합하여 DCGAN에 입력하여 학습한다.

생성자는 convolutional layer, dropout layer, batch normalization layer, relu layer로 구성된 block이 총 6개로 구성되어 있으며 block을 거칠수록 upsampling 하게 된다.

마지막 계층을 거친 후에는 3 x 64 x 64 크기의 이미지를 생성한다. 판별자는 convolutional layer, batch normalization layer, leaky relu layer로 구성된 block, 총 6개로 구성된다. Block을 거칠수록 down sampling 하며 마지막에 텍스트의 특징과 결합하여 활성함수인 시그모이드(sigmoid)를 통해 진짜인지 가짜인지 판별한다.

조선대학교
CHOSUN UNIVERSITY

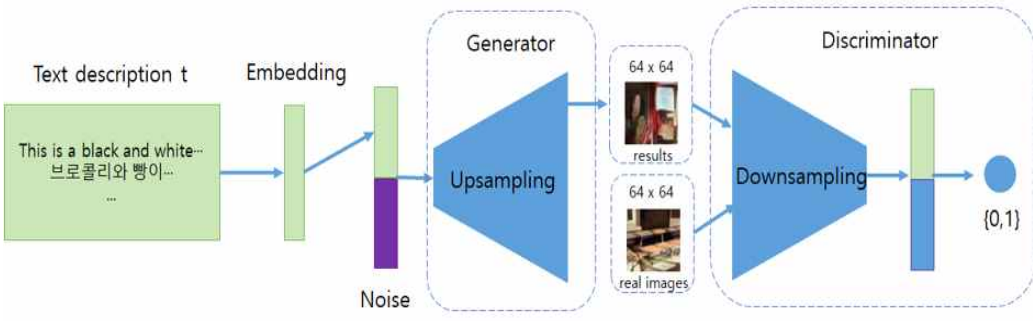


그림 11. DCGAN을 이용한 텍스트에서 이미지 생성 모델의 구조

DCGAN을 활용한 텍스트에서 이미지 생성의 loss fuction은 식 2와 같다.

식 2에서 G는 생성자, D는 구별자를 뜻한다. x는 입력 벡터이고 z는 랜덤 벡터, ϕ(t)는 입력된 텍스트의 특징벡터를 뜻한다. 그리고 판별자는 진짜 데이터와 가짜데이터를 구별할 수 있도록 식 2의 값이 큰 값을 갖도록 학습하며 생성자는 진짜 데이터와 비슷한 가짜 데이터를 생성하도록 식 2의 값이 작은 값을 갖게 학습한다.

본 연구에서 사용한 DCGAN의 판별자는 총 3가지 타입의 데이터를 입력 받아 학습하는데, 실제 이미지-실제 텍스트는 참으로 판단하고 실제 이미지-거짓 텍스트와 거짓 이미지-거짓 텍스트는 거짓라고 판단하게 학습한다. 과적합 방지를 위해 label smoothing, feature matching을 사용하여 실험을 하였다.

$$\min_{D}\max_{G} V(D,G) = E_{(x,t)\sim p_{data}(x,t)}[\log D(x,\varphi(t))] + E_{z\sim p_z(z),t\sim p_t(t)}[\log\ 1-D(G(z,\varphi(t)))] + E_{x\sim p_x(x),\hat{t}\sim p_t(t)}[\log\ 1-D(G(z,\varphi(\hat{t})))] \quad \text{식(2)}$$

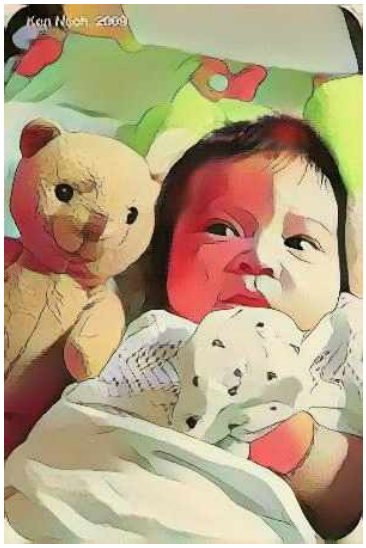
조선대학교
CHOSUN UNIVERSITY

B. 적대적 생성 신경망을 이용한 텍스트에서 웹툰 생성 결과

본 절에서는 구축한 멀티모달 데이터인 트리트먼트-웹툰 데이터 셋을 적대적 생성 신경망에 학습하고 그 결과를 확인한다. 다국어 웹툰 데이터 셋을 구축하기 위해 Microsoft에서 공개한 MSCOCO와 한국에서 이를 번역하여 공개한 AI hub의 한국어 MSCOCO 데이터 셋을 사전학습된 CartoonGAN을 사용하여 카툰 이미지로 변형하였다. 원래의 MSCOCO 데이터 셋은 학습 데이터 82,783장, 평가 데이터 40,504장의 이미지가 있으며 이미지당 5개의 영어로된 설명문이 있다. 여기에 한국어 MSCOCO 데이터 셋을 추가하여 표 1과 같이 한 이미지당 10개의 설명문(영어 5개, 한국어 5개)으로 구성하였다. CartoonGAN은 4가지 화풍으로 이미지를 변형시킬 수 있어 1개의 이미지와 1개의 트리트먼트를 1쌍으로 하였을 때 총 1,232,870쌍의 다국어 웹툰 데이터를 구축하였다.

조선대학교
CHOSUN UNIVERSITY

표 1. 트리트먼트-웹툰 데이터 셋의 예

이미지	텍스트
	'A shower stall with interesting tile is the focal point.'
	'A full perspective of a washroom with a sink.'
	'A white bathroom sink sitting under a mirror.'
	'I picture of a bathroom with a stand up shower stall and a person's reflection in the mirror.'
	'A small shower behind a small bathroom sink.'
	'흥미로운 타일이 있는 샤워 부스가 초점이다.'
	'싱크대가 있는 화장실의 전체적인 전망'
	'거울 밑에 하얀 욕실 싱크대가 있다.',
	'샤워 부스가 서 있고 거울에 사람이 비치는 모습이 그려진 욕실 사진입니다.'
	'작은 욕실 싱크대 뒤로 작은 샤워기가 있다.'
	'A baby is laying down with a teddy bear.'
	'A baby laying in a crib with a stuffed teddy bear.'
	'A baby wearing gloves, lying next to a teddy bear'
	'A baby stares to the left while taking a picture with a teddy bear lays beside it .'
	'A baby lies on blue and green bedding next to a teddy bear.'
	'아기가 테디 베어와 함께 누워 있다.'
	'테디 베어 인형을 채운 아기가 침대에 누워 있었다.',
	'장갑을 낀 아기가 테디 베어 옆에 누워 있다.'
	'한 아기가 곰 인형이 옆에 누워 있는 사진을 찍으며 왼쪽을 응시하고 있어요.'
	'아기는 테디 베어 옆의 푸른 색과 초록색 침구 위에 누워 있다.'

조선대학교
CHOSUN UNIVERSITY

다국어 텍스트를 딥러닝 기반의 텍스트에서 이미지 생성 모델에 입력하여 웹툰을 생성하기 위해 다국어 웹툰 데이터 셋을 DCGAN모델에 학습한다. 다국어 텍스트의 특징 벡터는 DCGAN이 데이터를 학습할 때 실시간으로 사전학습된 다국어 BERT에서 추출한다면 학습시간이 오래 걸릴 것을 감안하여 학습전 미리 특징벡터를 추출하여 학습이 진행될 때에는 저장된 특징벡터를 불러오는 방식으로 실험하였다. 학습에 사용한 데이터는 구축한 다국어 웹툰 데이터 셋에서 하나의 스타일에 대해서 실험하였으며 불완전한 데이터를 제외한 820,752개의 다국어 텍스트-웹툰 데이터 쌍을 학습하였다. 평가 데이터는 405,040 개의 데이터이고 이것을 8:2의 비율로 분할하여 평가 데이터와 텍스트 데이터로 사용하였다. 실험에 사용한 이미지의 크기는 3 x 64 x 64이며, noise는 100차원, batch size는 32로 설정하였다. 그리고 학습율은 생성자, 판별자 모두 0.0002이며 optimizer는 adam을 사용하였다.

그림 12는 학습 횟수가 증가함에 따라 평가 데이터의 텍스트를 생성자에 입력하였을 때 생성하는 이미지를 나타내는 표이다. 그림 12의 (a)와 같이 학습 횟수가 1일 때에는 어렴풋이 이미지를 생성하는 것을 볼 수 있으며, 그림 12의 (e)와 같이 100회까지는 학습이 진행될수록 물체의 형상을 그리는 것을 볼 수 있다. 하지만 그 이상 학습을 하게 되면 생성자가 더 이상 이미지를 생성하지 못하는 것을 확인할 수 있다.

따라서 생성자와 판별자의 손실과 평가 데이터를 사용하여 출력한 이미지를 볼 때 75회의 뉴런이 본 논문에서 구축한 웹툰 데이터에 적합하게 학습된 뉴런이라 볼 수 있다.

조선대학교
CHOSUN UNIVERSITY

그림 12. 학습 횟수에 따른 평가 데이터를 이용한 이미지 생성

조선대학교
CHOSUN UNIVERSITY

그림 13~15는 75회 학습된 생성자를 사용하여 테스트 데이터의 텍스트를 DCGAN에 입력하였을 때 생성되는 웹툰이다. 그림 13과 그림 14는 각각 한글, 영어로 작성된 트리트먼트를 입력하였을 때 출력되는 웹툰이며 그림 15는 같은 의미의 한글과 트리트먼트를 입력하였을 때 생성되는 웹툰이다. 생성된 웹툰을 보았을 때 원래의 이미지와의 유사성은 떨어지나 텍스트에 표현되어있는 형상에 대해 어느 정도 그리는 것을 확인할 수 있다. 객체별로 분석해 보면 사람이나 동물과 같은 객체는 잘 표현하지 못하는 것을 볼 수 있지만 의자와 같은 사물은 비교적 잘 표현하는 것을 실험 결과 알 수 있다. 그림 15를 보면 같은 의미의 한국어, 영어로 된 텍스트를 입력하였을 때 생성되는 웹툰을 보면 서로 생성된 웹툰은 다를지라도 트리트먼트에 표현된 단어의 의미적인 형상이 서로 유사하게 표현된 것을 확인할 수 있다.

표 2는 테스트 데이터 셋을 본 연구에서 제안하는 다국어 텍스트에서 이미지 생성 모델에 입력하였을 때 생성되는 이미지를 inception score와 FID(Frechet Inception Distance) score를 사용하여 평가한 점수이다. Inception score는 생성된 이미지의 품질과 다양성을 평가하는 점수이며, FID score는 실제 이미지와 생성된 이미지의 특징값의 평균과 공분산값을 비교한 점수이다. 다국어 텍스트에서 이미지 생성 모델의 inception score는 4.992이며, FID score는 22.212이다. MSCOCO 데이터 셋을 학습한 DCGAN 모델의 inception score는 7.88점이며, 다국어 웹툰 데이터 셋을 학습한 DCGAN 모델의 점수가 낮게 나온 것은 카툰화 작업을 진행하였을 때 이미지에 표현된 형상이 일그러지기 때문에 inception score가 낮게 나온 것으로 판단된다.

표 2. 다국어 텍스트에서 이미지 생성 모델의 성능평가

데이터 셋	Inception score	FID score
다국어 웹툰 데이터 셋의 테스트 데이터	4.992	22.212

조선대학교
CHOSUN UNIVERSITY

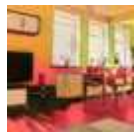
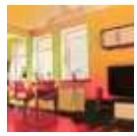
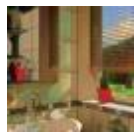
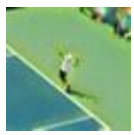
텍스트	실제 이미지	생성된 가짜 이미지
의자와 테이블과 여자가 있는 방		
텔레비전과 테이블이 있는 생활 공간		
치즈 브로콜리와 닭고기가 들어 있는 흰색 접시		
거울 아래에 있는 욕실 싱크대		
한 남자가 경기에서 테니스 공을 서브하기 위해 준비한다		

그림 13. 한국어 텍스트를 DCGAN에 입력하였을 때 생성되는 웹툰

조선대학교
CHOSUN UNIVERSITY

텍스트	실제 이미지	생성된 가짜 이미지
youth surfing on a body of water outside		
This is a black and white picture of a chester bench		
A smiling person holding a snowboard standing on a snow covered hill		
A turkey dinner shows corn, peas, mashed potatoes and biscuits all on one plate		
A business called Ray's Tavern with motorcycles sitting outside it		

그림 14. 영어 텍스트를 DCGAN에 입력하였을 때 생성되는 웹툰

조선대학교
CHOSUN UNIVERSITY

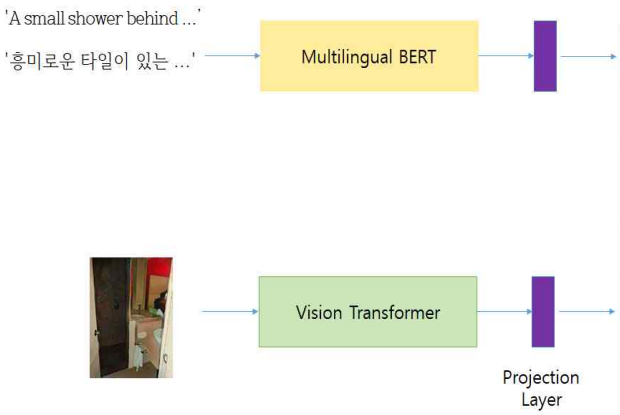
텍스트	실제 이미지	생성된 가짜 이미지
한 남자가 거품이 이는 바다에서 서핑을 하고 놀았다		
A man surfs and plays in the foamy ocean		
두툼한 빵 위에 야채와 치즈를 얹은 샌드위치		
A plate of food with bread, grape tomatoes, cheese, cucumbers and sauce on it		
밝은 색과 초록색 그리고 흰색 커튼이 있는 방에 있는 두개의 침대		
Two beds in a room with a light and green and white curtain		

그림 15. 같은 의미의 텍스트를 DCGAN에 입력하였을 때 생성되는 웹툰

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 트리트먼트-웹툰 데이터 셋의 멀티모달 학습 방법

A. CLIP 모델을 이용한 멀티모달 학습 방법



	T_1	T_2	T_3	T_4	T_5
I_1	0.2	0.3	0.5	0.1	0.4
I_2	0.8	0.2	0.1	0.4	0.6
I_3	0.3	0.1	0.7	0.2	0.5
I_4	0.6	0.4	0.3	0.8	0.9
I_5	0.1	0.5	0.4	0.7	0.2

그림 16. CLIP의 구조와 텍스트-이미지 쌍의 logits 행렬의 예

본 절에서는 트리트먼트-웹툰 데이터 셋을 멀티모달 학습을 통해 서로 다른 도메인의 데이터 사이의 유사도를 측정한다. CLIP 모델은 텍스트와 이미지의 특징을 추출할 수 있는 transformer 기반의 다국어 BERT와 Vision transfromer를 인코더로 사용하고 각각의 인코더의 마지막 계층에 동일한 512차원의 완전 연결계층인 투영 계층이 추가된 구조이다.

조선대학교
CHOSUN UNIVERSITY

```
# normalized features
image_embeds = image_embeds / np.linalg.norm(image_embeds, axis=-1)
text_embeds = text_embeds / np.linalg.norm(text_embeds, axis=-1)

# cosine similarity as logits
logits_per_text = matmul(text_embeds, image_embeds.T) * logit_scale
logits_per_image = logits_per_text.T
```

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

그림 17. CLIP 슈도코드에서의 코사인 유사도 계산 방법

그림 18는 logits를 구하는 방법은 배치 크기가 5일 때 logits를 구한 예이다. 여기서 logits은 텍스트와 이미지 사이의 코사인 유사도이다. 그림 17에서 볼 수 있듯이 텍스트와 이미지 사이의 코사인 유사도의 측정은 텍스트의 특징 벡터와 이미지의 특징 벡터를 정규화하고 행렬곱을 통해 구할 수 있다. 코사인 유사도는 -1부터 1까지의 값을 갖게 되며 1에 가까울수록 가장 유사함을 뜻한다.

	T_1	T_2	T_3	T_4	T_5
I_1	0.2	0.3	0.5	0.1	0.4
I_2	0.8	0.2	0.1	0.4	0.6
I_3	0.3	0.1	0.7	0.2	0.5
I_4	0.6	0.4	0.3	0.8	0.9
I_5	0.1	0.5	0.4	0.7	0.2

```
labels = np.arange(len(logits))
acc_i = ((np.argmax(logits, 1) == labels)).mean()
acc_t = ((np.argmax(logits, 0) == labels)).mean()
acc = (acc_i + acc_t) / 2
```

그림 18. CLIP 슈도코드에서 정확도 계산 방법

정확도의 측정은 텍스트의 특징와 이미지의 특징 사이의 logits을 구한 뒤 레이블과 비교하여 측정할 수 있다. 그림 18에서 오른쪽은 앞서 계산한 logits을 사용하여 정확도를 구하는 슈도코드이다. 정답 레이블은 numpy의 arrange 함수를

조선대학교
CHOSUN UNIVERSITY

사용하여 구할 수 있다. 예를 들어 그림 18에서 logits는 5이기 때문에 np.arragne 함수를 실행하였을 때 [0, 1, 2, 3, 4]의 행렬을 얻을 수 있다. 이 행렬의 뜻은 각 행에서 0번째, 1번째, 2번째, 3번째, 4번째 인덱스에 있는 값을 뜻한다.

np.argmax함수는 각 행에서 가장 큰 값의 인덱스를 찾는 함수이며 두 번째 인자값이 1일때는 행, 0일때는 열에서 가장 큰 값의 인덱스를 찾는다. acc_i에서 np.argmax(logits, 1)의 결과값은 [2, 0, 2, 3, 3]과 같은 배열이다. 이 뜻은 각 행에서 2번째, 0번째, 3번째, 3번째 인덱스에 있는 값이 가장 크다는 것을 뜻한다. 레이블 배열인 [0, 1, 2, 3, 4]와 argmax의 결과 배열 [2, 0, 2, 3, 3]을 비교하여 같으면 True, 다르면 False를 반환하게 된다. 따라서 '==' 연산자를 수행하였을 때의 결과는 [False, False, True, True, False]로 나타나며 예측값과 정답이 일치하는 비율은 0.4이다. 여기서 레이블의 배열이 [0, 1, 2, 3, 4]인 이유는 그림 18에서 같은 인덱스의 logits(왼쪽 표의 파란색 부분)은 학습 시 명시적으로 서로 유사한 데이터이기 때문에 확률적으로 1에 가까운 코사인 유사도를 나타내야 한다. 그렇기 때문에 레이블의 배열 인덱스에 있는 logits은 각 행에서 가장 큰 값을 나타내야 한다. 따라서 정확도는 학습하며 예측한 logits과 정답 레이블과의 일치 비율을 나타낸다.

CLIP loss를 구하기 위해 앞서 계산한 logits을 행 방향, 열 방향 각각의 cross entropy를 구한 후 평균을 구해 CLIP loss를 계산하게 된다. 마지막으로 계산된 loss를 텍스트, 이미지 각 인코더에 업데이트하여 CLIP 잠재 공간에서 서로 유사한 데이터의 특징은 가깝게, 서로 다른 데이터의 특징은 멀어지도록 학습하게 된다.

조선대학교
CHOSUN UNIVERSITY

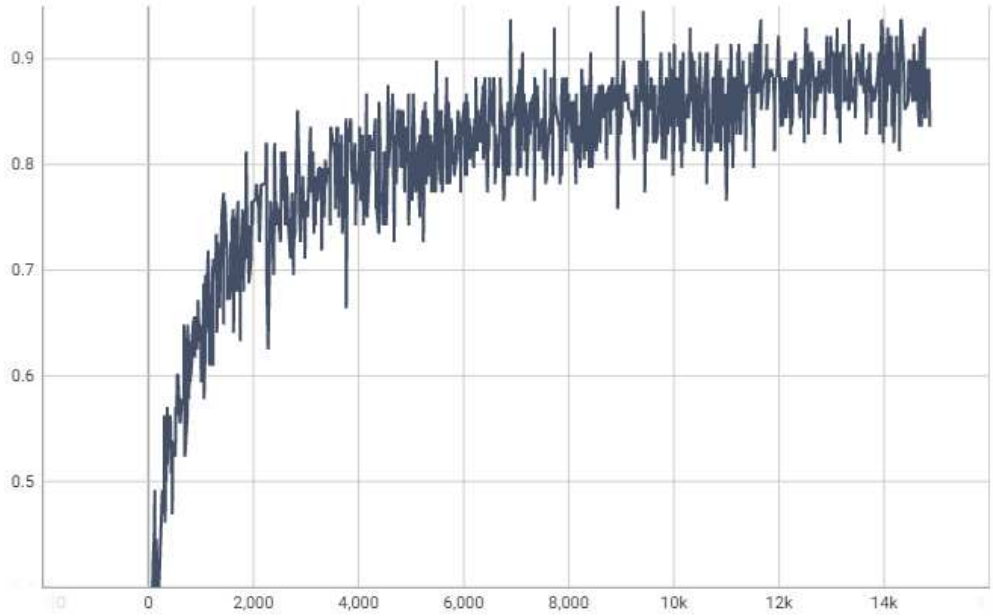
B. CLIP 모델의 실험 결과

본 절에서는 트리트먼트-웹툰 데이터 셋을 CLIP 모델에 학습하고 그 결과를 확인한다. CLIP모델은 사전학습된 다국어 BERT와 vision transformer을 텍스트와 이미지의 인코더로 사용하고 각 모델의 마지막 출력층에 512차원의 완전연결계층으로 구성된 투영계층을 추가한 모델이다[32]. 학습시 각 인코더의 계층을 얼리지 않고 전체 계층을 학습하였다. loss 함수는 대조 loss를 사용하였는데, 대조 loss는 텍스트와 이미지 벡터를 차원이 같은 투영계층에 위치시킨 후 벡터끼리 코사인 유사도를 구하고 이 유사도를 cross entropy를 통해 loss를 구하게 된다. 학습 시 사용한 주요 하이퍼파라미터로는 학습횟수 25회, 학습률은 0.00001, 배치 크기는 4개의 GPU에서 각 GPU당 32로 총 128 크기로 학습하였다.

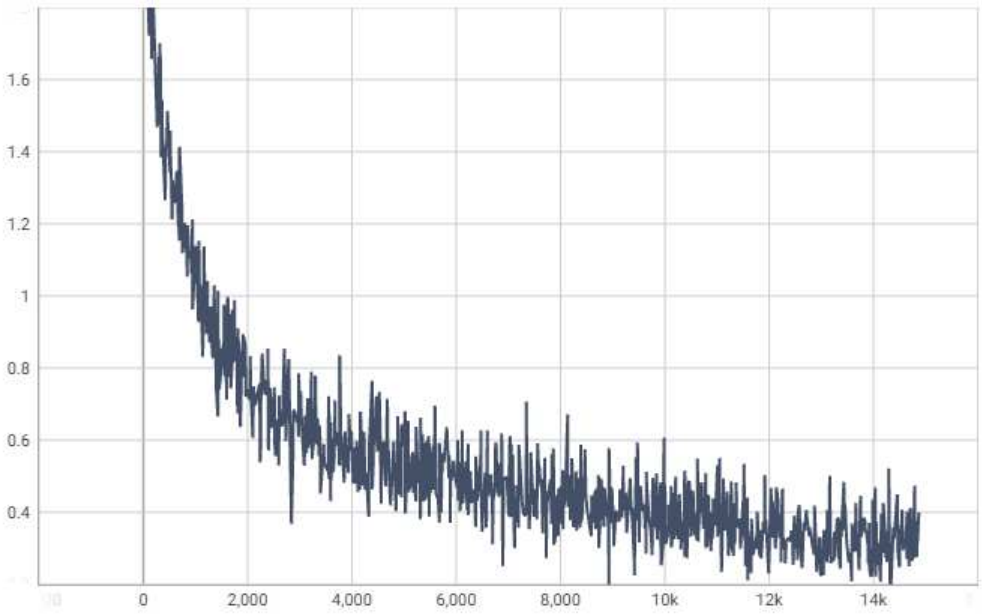
CLIP 모델을 학습할 때 모델의 성능평가로 정확도와 loss를 사용하여 모델의 학습 성능을 평가한 결과를 확인한다. 그림 19, 20은 트리트먼트-웹툰 데이터 셋을 CLIP 모델에 학습하면서 측정한 정확도와 loss 이다. 학습횟수는 25회 진행하였으며, 그림 19의 (a)에서 볼 수 있듯이 학습횟수 25회에서 학습 데이터 셋의 정확도는 88%를 나타내었으며 그림 20의 (a)와 같이 평가 데이터의 정확도는 60%를 나타내었다. 그림 19와 20에서 볼 수 있듯이 학습횟수가 증가할수록 accuracy가 높아지고 loss가 감소하는 것을 확인할 수 있으며 점차 상승 및 감소 폭이 낮아지는 것을 확인할 수 있다.

본 연구에서 사용하는 CLIP 모델은 25회 학습한 모델을 사용하였다.

조선대학교
CHOSUN UNIVERSITY



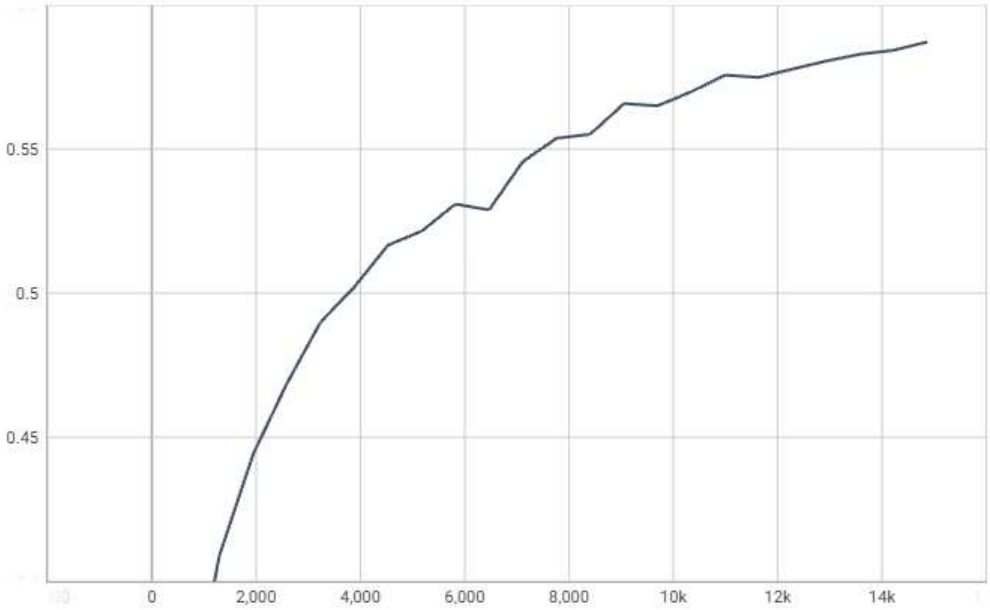
(a) 학습 데이터 셋의 정확도



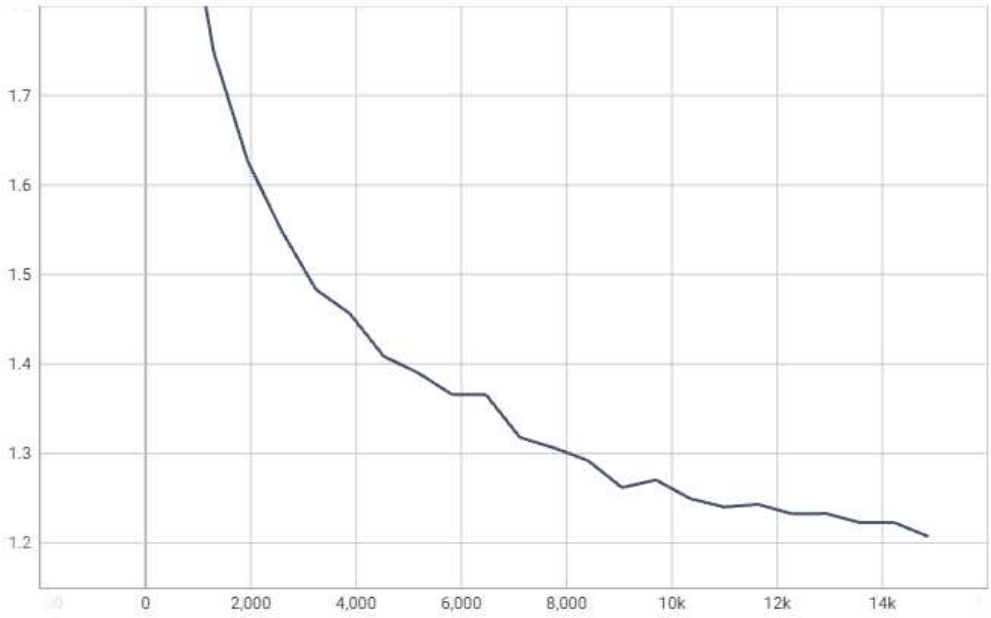
(b) 학습 데이터 셋의 loss

그림 19. 학습 데이터 셋의 CLIP 정확도와 손실 그래프

조선대학교
CHOSUN UNIVERSITY



(a) 평가 데이터 셋의 정확도



(b) 평가 데이터 셋의 손실

그림 20. 평가 데이터 셋의 CLIP 정확도와 손실 그래프

조선대학교
CHOSUN UNIVERSITY

1. 텍스트-웹툰 사이의 유사도 측정

본 절에서는 여러개의 다국어 문장과 하나의 이미지를 CLIP 모델에 입력하였을 때 출력되는 임베딩값을 사용하여 PCA를 통해 임베딩한 결과를 확인하고 유사도를 측정한다. 표 3은 실험에 사용한 데이터이며 첫 번째, 두 번째 문장은 이미지에 대한 올바른 설명이며 다른 문장은 이미지와 의미가 다른 문장이다.

표 3. 유사도 측정을 위한 비교 데이터

“싱크대가 있는 화장실의 전체적인 전망”	
“A full perspective of a washroom with a sink.”	
“일몰에서의 커플”,	
“a couple at sunset”,	
“고속도로에서 달리는 말”,	
“a horse on the highway”,	
“양배추, 여섯 명의 친구, 그리고 한 잔의 물”	
“a cabbage, six friends, and a glass of water”	
“주차된 차들이 있는 집”	
“a house with parked cars”	

조선대학교
CHOSUN UNIVERSITY

표 4. 트리트먼트와 툰이미지의 CLIP 인코딩 값

입력	출력된 특징의 형태
	torch.Size([1, 512])
‘A shower stall with interesting tile is the focal point.’	torch.Size([1, 512])
‘흥미로운 타일이 있는 샤워 부스가 초점이다.’	torch.Size([1, 512])

트리트먼트-웹툰 데이터 셋을 학습한 CLIP에 텍스트와 이미지를 입력하여 인코딩 하였을 때 출력되는 임베딩 값(특징 벡터)는 표 4와 같다. CLIP 의 인코딩은 데이터가 입력되면 각각의 텍스트 인코더와 텍스트 인코더를 통해 출력된 값이 투영계층을 통해 같은 차원의 특징값으로 출력된다. 그렇기 때문에 본 논문에서 데이터를 입력했을 때 출력되는 특징의 형태는 [배치 사이즈, 투영 차원]이다. 표 4에서 볼 수 있듯이 하나의 데이터를 입력하여 출력하였기 때문에 임베딩 값의 형태는 [1, 512]이다.

조선대학교
CHOSUN UNIVERSITY

표 5. CLIP 임베딩 값을 이용한 유사도 측정

입력		유사도
“싱크대가 있는 화장실의 전체적인 전망”		52 %
“A full perspective of a washroom with a sink.”		53 %
“일몰에서의 커플”		2 %
“a couple at sunset”		-3 %
“고속도로에서 달리는 말”,		-15 %
“a horse on the highway”		-18 %
“양배추, 여섯 명의 친구, 그리고 한 잔의 물”		-10 %
“a cabbage, six friends, and a glass of water”		-2 %
“주차된 차들이 있는 집”		-4 %
“a house with parked cars”		-2 %

표 5는 여러개의 문장과 이미지를 CLIP에 입력하여 출력한 임베딩값을 코사인 유사도를 측정한 결과이다. 실험결과 이미지에 대한 올바른 설명이 정확도가 높은 것을 알 수 있다.

조선대학교
CHOSUN UNIVERSITY

그림 21. 표5 데이터의 CLIP 임베딩의 PCA 결과

그림 21은 표 3의 데이터를 CLIP에 입력하여 출력한 임베딩 벡터의 PCA를 수행한 결과이다. PCA란, Principal Component Analysis의 약자로서 높은 차원의 데이터를 차원축소를 통해 저차원의 공간으로 투영하는 것이다. 그림 21에서 주황색 점은 입력한 텍스트의 임베딩 벡터이고 파랑색은 입력한 이미지의 벡터이다.

실험결과 표 5에서 볼 수 있는 정확도가 높은 텍스트가 입력한 이미지와 가장 가까운 거리에 위치하는 것을 확인할 수 있다. 그리고 동일한 의미를 지닌 한국어, 영어 트리트먼트들은 비슷한 공간에 위치하는 것을 확인할 수 있다.

조선대학교
CHOSUN UNIVERSITY

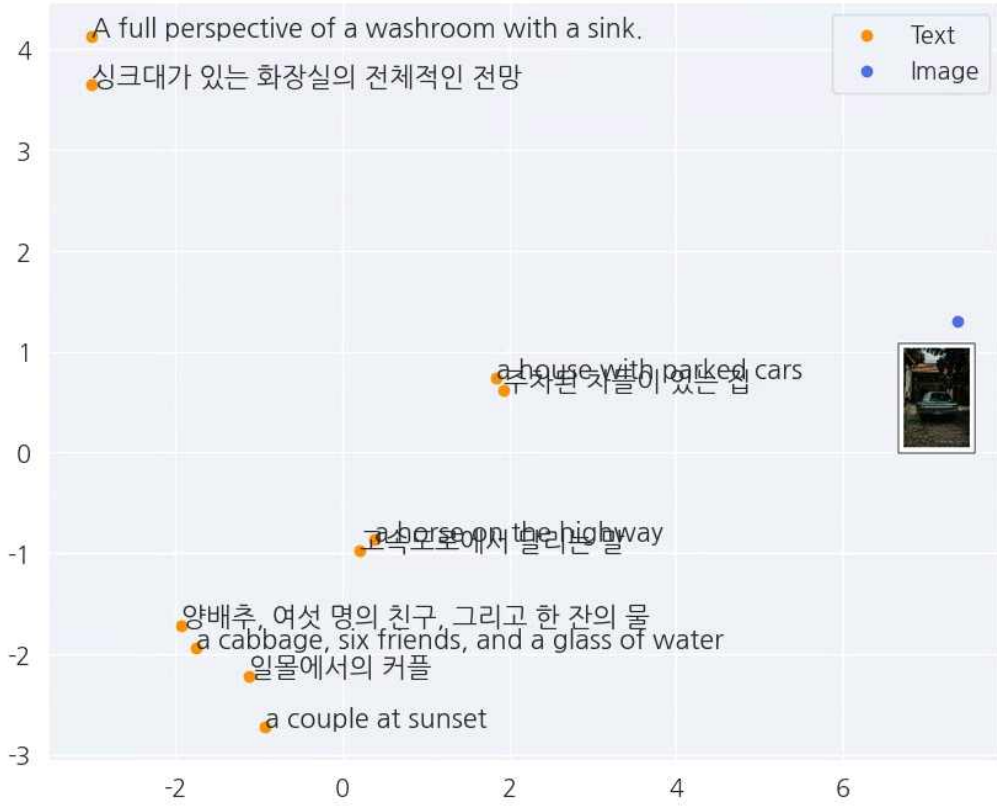


그림 22. 다른 이미지를 입력하였을 때 CLIP 임베딩의 PCA 결과

그림 22는 표 3의 한국어-영어 트리트먼트에 다른 이미지를 입력하였을 때의 PCA 결과이다. 이 이미지에 해당하는 올바른 트리트먼트는 '주차된 차들이 있는 집'이다. 그림 23에서 볼 수 있듯이 다른 문장들보다 올바른 문장의 임베딩 값이 이미지의 임베딩 값에 가장 가까운 거리에 있는 것을 확인할 수 있다.

조선대학교
CHOSUN UNIVERSITY

2. 텍스트와 유사한 웹툰 검색

다음으로, 텍스트를 쿼리로 하였을 때 구축한 데이터 셋에서 가장 유사한 이미지를 출력하도록 한다. 데이터 셋에서 유사한 이미지를 찾는 방법은 쿼리로 텍스트가 입력되면 CLIP 모델에 인코딩하여 텍스트의 임베딩 값을 얻고, 사전에 데이터 셋의 이미지 CLIP 임베딩값 사이의 코사인 유사도를 측정하여 유사도가 높은 N개의 이미지를 출력한다. 데이터 셋의 이미지에 대한 CLIP 임베딩 값을 사전에 출력하여 저장한 이유는 실시간으로 이미지의 임베딩 값과 유사도를 측정할 수 있지만 그렇게 한다면 대용량의 이미지의 임베딩 값을 계산하는데 시간이 오래 걸리기 때문이다.

표 6-9는 평가 데이터 셋에 있는 텍스트를 쿼리로 하였을 때 원래의 쿼리에 해당하는 이미지와 이미지 검색을 통해 유사도가 높은 Top 4의 이미지를 나타낸 표이며 표 6과 7은 영어 트리트먼트를 쿼리로 검색하였을 때, 표 8과 9는 한글 트리트먼트를 쿼리로 하였을 때의 결과이다. 이미지 검색 실험결과 쿼리에 사용한 트리트먼트에 의미상으로 유사한 이미지를 찾을 수 있는 것을 확인할 수 있었으며 원래 트리트먼트에 맞는 이미지와도 상당부분 유사한 이미지를 찾는 것을 확인할 수 있었다.

조선대학교
CHOSUN UNIVERSITY

표 6. 영어 트리트먼트를 입력하였을 때 CLIP를 이용한 데이터 셋 이미지 검색-1

입력 문장	실제 이미지
A horse drawn carriage is parked along the curb.	
이미지 검색 - Top 1	이미지 검색 - Top 2
	
이미지 검색 - Top 3	이미지 검색 - Top 4
	

조선대학교
CHOSUN UNIVERSITY

표 6. 영어 트리트먼트를 입력하였을 때 CLIP를 이용한 데이터 셋 이미지 검색-2

입력 문장	실제 이미지
A black bear sitting on rock with mossy patches.	
이미지 검색 - Top 1	이미지 검색 - Top 2
	
이미지 검색 - Top 3	이미지 검색 - Top 4
	

조선대학교
CHOSUN UNIVERSITY

표 8. 한글 트리트먼트를 입력하였을 때 CLIP를 이용한 데이터 셋 이미지 검색-1

입력 문장	실제 이미지
많은 사람들이 멋진 날을 즐기고 있어요, 큰 잔디밭에서 연을 날리면서 말이죠.	
이미지 검색 - Top 1	**이미지 검색 - Top 2**
	
이미지 검색 - Top 3	**이미지 검색 - Top 4**
	

조선대학교
CHOSUN UNIVERSITY

표 9. 한글 트리트먼트를 입력하였을 때 CLIP를 이용한 데이터 셋 이미지 검색-2

입력 문장	실제 이미지
주차장의 오토바이를 타고 옆에 서 있는 사람이 있다.	
이미지 검색 - Top 1	이미지 검색 - Top 2
	
이미지 검색 - Top 3	이미지 검색 - Top 4
	

조선대학교
CHOSUN UNIVERSITY

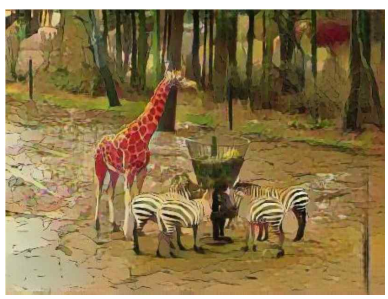
3. 제로샷 분류

제로샷 분류(Zero-shot classification)는 모델이 훈련 과정에서 새로운 클래스에 대한 레이블을 학습한 적이 없어도 새로운 클래스의 이미지를 분류할 수 있는 것을 말한다. CLIP는 제로샷 분류에서 장점을 갖고 있는데, 이것이 가능한 이유는 텍스트와 이미지를 하나의 공간에 매핑하여 유사한 것은 가깝게 유사하지 않은 것은 멀어지도록 하는 대조 학습과 명시적인 레이블 없이 데이터 그 자체를 레이블로 학습하기 때문이다. 만약 새로운 이미지를 CLIP에 입력하여 제로샷 분류를 하게 된다면, 입력한 이미지와 입력한 레이블을 함께 동일한 임베딩 공간에 매핑하게 된다. 그 후 임베딩 공간에서 가장 가까운 레이블 선택함으로써 제로샷 분류를 할 수 있다.

CLIP를 이용한 제로샷 분류는 앞서 실험한 텍스트를 쿼리로 하는 이미지 검색 방법과 유사한 방법으로 진행된다. 입력한 클레스와 이미지를 각각 CLIP에 입력하여 유사도를 측정하여 유사도가 가장 높은 레이블을 이미지의 레이블이다. 표 10과 11은 텍스트를 label로 하여 이미지를 분류한 결과이다. 실험에 사용한 레이블은 장면에 관한 레이블 일부를 사용하였다. 제로샷 분류를 수행한 결과 이미지에 유사한 레이블로 분류하는 것을 확인할 수 있다.

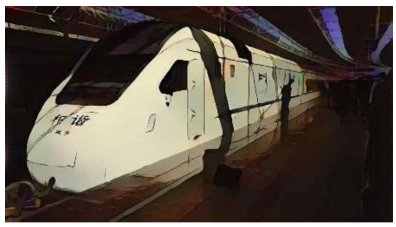
조선대학교
CHOSUN UNIVERSITY

표 10. CLIP를 이용한 제로샷 분류의 결과 - 1

입력 텍스트(레이블)	입력 이미지	제로샷 분류
'Bedroom', 'Office', 'Kitchen', 'Living room', 'Beach', 'Mountain', 'Stadium', 'Restaurant', 'Forest'		**Predicted label: Kitchen** --- Bathroom: 0.188 Bedroom: -0.046 Office: 0.130 **Kitchen: 0.384** Living room: 0.063 Beach: -0.031 Mountain: -0.059 Stadium: 0.017 Restaurant: 0.121 Forest: 0.043
'Bathroom', 'Bedroom', 'Office', 'Kitchen', 'Living room', 'Beach', 'Mountain', 'Stadium', 'Restaurant', 'Forest'		**Predicted label: Forest** --- Bathroom: -0.131 Bedroom: -0.047 Office: 0.006 Kitchen: 0.076 Living room: -0.037 Beach: 0.006 Mountain: -0.015 Stadium: -0.033 Restaurant: -0.062 **Forest: 0.198**

조선대학교
CHOSUN UNIVERSITY

표 11. CLIP를 이용한 제로샷 분류의 결과 - 2

입력 텍스트(레이블)	입력 이미지	제로샷 분류
'Beach', 'City street', 'Forest', 'Kitchen', 'Mountain', 'Office', 'Restaurant', 'Skatepark', 'Swimming pool', 'Train station'		Predicted label: **Train station** --- Beach: 0.139 City street: 0.071 Forest: 0.114 Kitchen: 0.134 Mountain: 0.010 Office: 0.221 Restaurant: 0.191 Skatepark: 0.194 Swimming pool: 0.083 **Train station: 0.525**
'Beach', 'City street', 'Forest', 'Kitchen', 'Mountain', 'Office', 'Restaurant', 'Skatepark', 'Swimming pool', 'Train station',		Predicted label: **Forest** --- Beach: 0.119 City street: -0.086 **Forest: 0.352** Kitchen: 0.142 Mountain: 0.133 Office: 0.086 Restaurant: -0.012 Skatepark: -0.014 Swimming pool: 0.000 Train station: -0.246

조선대학교
CHOSUN UNIVERSITY

MRR@K 지표는 랭킹모델을 평가하기 위한 지표로서 질의에 대해 상위 순위의 항목의 순위 역수의 평균을 계산한다. MRR은 0과 1 사이의 값을 나타내며 값이 1에 가까울수록 상위 순위에 좋은 항목이 포함되었다는 것을 의미하며 높은 값일수록 좋은 성능을 나타낸다. 본 연구에서 학습한 CLIP 모델과 mscoco-it 데이터를 사용하여 정량적 성능을 나타낸 결과 MRR@1에서 0.23, MRR@5에서 0.47, MRR@10에서 0.51의 지표는 나타내었다.

표 12. 트리트먼트-웹툰 CLIP의 MRR@K 성능평가

성능지표	MRR@1	MRR@5	MRR@10
MRR@K	0.23	0.47	0.51

조선대학교
CHOSUN UNIVERSITY

Ⅴ. 확산 모델 기반의 멀티모달 데이터를 이용한 웹툰 생성

A. 확산 모델을 이용한 텍스트에서 웹툰 생성 방법

본 절에서는 확산 모델을 이용하여 웹툰을 생성하기 위해 앞절에서 학습한 CLIP 모델과 stable diffusion의 depth-to-image 모델을 사용한다[29]. 웹툰의 생성 단계는 그림 21과 같이 첫 번째 단계에서는 생성하고자 하는 트리트먼트와 가장 유사한 이미지를 데이터 셋에서 찾고, 두 번째 단계에서는 가장 유사한 이미지를 depth-to-image 모델에 텍스트와 함께 입력하여 웹툰을 생성한다.

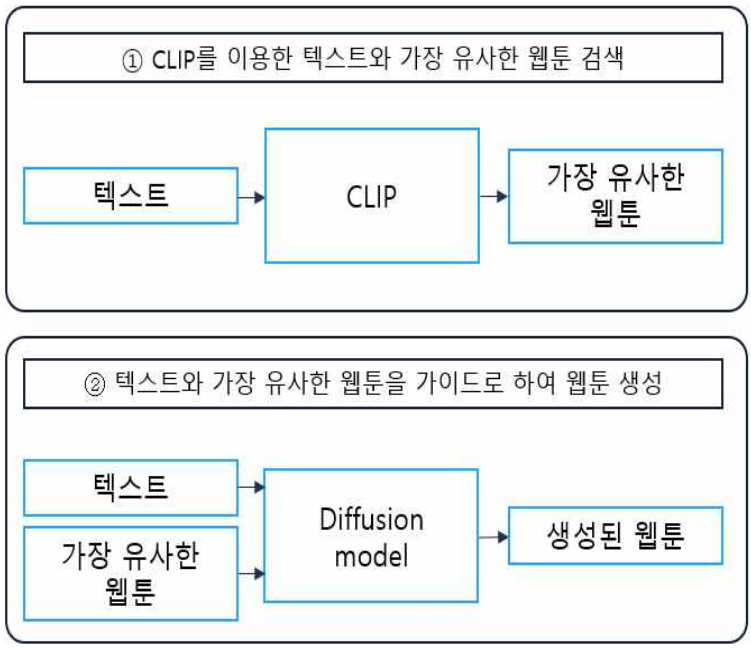


그림 21. CLIP와 확산 모델을 이용한 웹툰 생성의 과정

조선대학교
CHOSUN UNIVERSITY

텍스트를 CLIP 모델에 입력하였을 때 유사한 이미지를 찾는 것은 데이터 셋의 모든 이미지의 CLIP 임베딩 벡터와 코사인 유사도를 측정하여 가장 유사도가 가장 높은 임베딩 벡터의 인덱스 값으로 이미지를 찾는다. 모든 이미지에 대한 CLIP 임베딩 값의 계산은 오래 걸리기 때문에 사전에 이미지에 대한 임베딩 벡터를 모두 계산하여 저장해두어 트리트먼트의 CLIP 임베딩 벡터와 유사도를 계산할 때 저장된 전체 이미지 데이터의 임베딩 값을 불러와 유사도를 측정한다.

텍스트와 가장 유사한 이미지 데이터 셋에 찾고 이 이미지와 생성하고자 하는 트리트먼트를 depth-to-image 모델에 입력하여 이미지를 생성한다. depth-to-image는 입력된 이미지의 depth 정보를 초기 이미지로 사용하여 역방향 확산 과정을 통해 이미지의 노이즈를 제거해나가며 RGB 이미지를 생성하는 모델이다. 이러한 방법이 가능한 이유는 학습시 RGB 이미지와 이미지의 depth 이미지를 쌍으로 학습하고 그 결과 이미지 생성 시 depth 정보와 연관된 색상 정보를 예측하고 이를 기반으로 각 픽셀의 RGB값을 결정할 수 있다. depth-to-image 모델은 사실적인 이미지를 생성하는 모델이기 때문에 본 연구에서 생성하고자 하는 카툰 스타일의 이미지를 생성하기 위해 쿼리에 “webtoon”이라는 키워드를 더해 모델에 입력하여 웹툰을 생성한다.

조선대학교
CHOSUN UNIVERSITY

B. 확산 모델을 이용한 텍스트에서 웹툰 생성 결과

1. 하나의 텍스트를 입력으로 웹툰 생성

본 절에서는 트리트먼트-웹툰 데이터 셋을 학습한 CLIP와 사전학습한 확산 모델을 이용하여 웹툰을 생성한다. 이미지 생성의 방법은 트리트먼트를 CLIP에 입력하여 트리트먼트의 임베딩 값을 구하고 사전에 계산된 이미지 데이터 셋의 임베딩 값과 유사도를 측정한다. 그 후 유사도가 가장 높은 이미지 임베딩에 해당되는 이미지를 찾은 뒤 이 이미지와 트리트먼트를 사전학습된 확산 모델에 입력하여 이미지를 생성한다. 실험은 하나의 트리트먼트를 CLIP와 확산 모델을 이용하여 웹툰 스타일의 이미지를 생성하고, 언어모델을 사용하여 다음에 나올 수 있는 3개의 문장을 생성한 트리트먼트를 사용하여 하나의 트리트먼트를 사용하여 웹툰 스타일을 생성하였을 때와 동일한 방법으로 웹툰을 생성한다.

표 13-17은 본 연구에서 학습한 트리트먼트-웹툰 데이터 셋을 학습한 CLIP와 사전학습된 확산 모델인 depth-to-image 모델을 사용하여 하나의 트리트먼트를 입력하였을 때 웹툰을 생성한 결과이다. 각 표의 (a)는 CLIP에 쿼리로 입력한 트리트먼트이며 (b)는 트리트먼트-웹툰 데이터 셋의 이미지에서 가장 높은 유사도가 높은 이미지를 출력한 결과이다. (c)와 (d)는 CLIP에 쿼리로 주었던 트리트먼트 (a)와 가장 유사한 이미지 (b)를 확산 모델에 함께 입력하였을 때 생성된 이미지이다.

조선대학교
CHOSUN UNIVERSITY

표 13. CLIP와 확산 모델을 이용한 웹툰 생성 결과 - 1

(a) CLIP 입력 텍스트	(b) 가장 유사한 이미지
Six teens playing frisbee in a field of grass	
(c) 생성된 이미지-1	(d) 생성된 이미지-2
	

조선대학교
CHOSUN UNIVERSITY

표 14. CLIP와 확산 모델을 이용한 웹툰 생성 결과 - 2

(a) CLIP 입력 텍스트	(b) 가장 유사한 이미지
A person in a ski suit is skiing.	
(c) 생성된 이미지-1	(d) 생성된 이미지-2
	

조선대학교
CHOSUN UNIVERSITY

표 15. CLIP와 확산 모델을 이용한 웹툰 생성 결과 - 3

(a) CLIP 입력 텍스트	(b) 가장 유사한 이미지
A horse drawn carriage is parked along the curb.	
(c) 생성된 이미지-1	(d) 생성된 이미지-2
	

조선대학교
CHOSUN UNIVERSITY

표 16. CLIP와 확산 모델을 이용한 웹툰 생성 결과 - 4

(a) CLIP 입력 텍스트	(b) 가장 유사한 이미지
Electric appliances crammed on a counter in a kitchen.	
(c) 생성된 이미지-1	(d) 생성된 이미지-2
	

조선대학교
CHOSUN UNIVERSITY

표 17. CLIP와 확산 모델을 이용한 웹툰 생성 결과 - 5

(a) CLIP 입력 텍스트	(b) 가장 유사한 이미지
A woman cutting a large white sheet cake.	
(c) 생성된 이미지-1	(d) 생성된 이미지-2
	

조선대학교
CHOSUN UNIVERSITY

2. 연속된 텍스트를 입력으로 웹툰 생성

웹툰은 세로 방향으로 연속되어있는 이미지로 구성되어 있다. 따라서, 실제 웹툰과 유사하게 웹툰을 생성하기 위해 구축한 트리트먼트-웹툰 데이터 셋 중 텍스트 데이터를 활용하여 의미적으로 다음에 나타날 문장을 GPT-3.5기반의 ChatGPT[30]를 사용하여 3문장씩 생성하였다. 트리트먼트-웹툰 데이터 셋은 하나의 웹툰과 한국어 5문장, 영어 5문장으로 구성되 있다. 이 중 영어 5문장에서 각각의 문장에 대해 다음에 나타날 문장을 3개의 문장을 생성하여 하나의 트리트먼트 데이터 셋에서 총 4문장으로 구성된 5개의 연속적인 텍스트 데이터를 생성하였다. 이와 같은 방법으로 4개의 문장으로 구성된 연속된 텍스트 데이터 5000개를 구축하였다. 실험은 이전과 동일한 방식으로 하나의 텍스트를 CLIP에 입력하여 유사한 이미지를 찾은뒤 확산 모델에 입력하여 웹툰을 생성하였다.

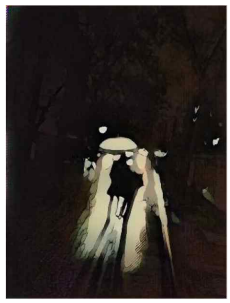
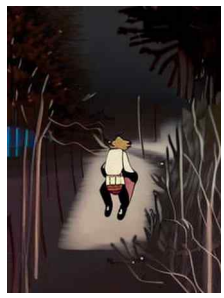
표 19~21은 연속된 텍스트데이터를 CLIP에 입력하여 유사한 이미지를 찾고 확산 모델에 입력하여 웹툰을 생성한 결과이다. 5000개의 연속된 텍스트 데이터 셋을 CLIP와 확산 모델에 입력하여 웹툰 스타일의 이미지를 생성하였으며, 정량적 성능평가를 위해 inception score를 측정하였다. 측정 결과는 아래 표 18과 같다.

표 18. 연속된 텍스트을 이용한 웹툰의 inception score

성능지표	Inception score
Inception score	7.14

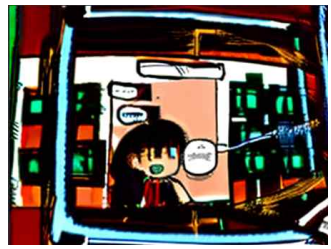
조선대학교
CHOSUN UNIVERSITY

표 19. 연속된 텍스트를 이용한 웹툰 생성 결과 - 1

CLIP 입력 텍스트	가장 유사한 이미지	생성된 이미지
Man riding a motor bike on a dirt road on the countryside.		
The wind whipped through the rider's hair as he navigated the twists and turns of the path.		
The sound of the engine echoed across the fields, disturbing the peaceful silence of the countryside.		
He couldn't help but feel a sense of freedom as he sped along on his trusty machine.'		

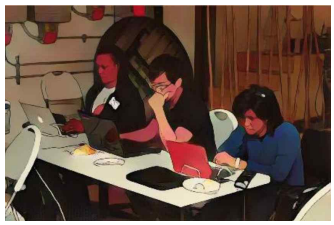
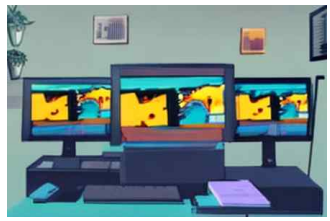
조선대학교
CHOSUN UNIVERSITY

표 20. 연속된 텍스트를 이용한 웹툰 생성 결과 - 2

CLIP 입력 텍스트	가장 유사한 이미지	생성된 이미지
A young boy stares up at the computer monitor.		
He seems to be fascinated by the image on the screen.		
After a few moments, he adjusts his glasses and begins to type furiously on the keyboard.		
It's clear that he's up to something important.		

조선대학교
CHOSUN UNIVERSITY

표 21. 연속된 텍스트를 이용한 웹툰 생성 결과 - 3

CLIP 입력 텍스트	가장 유사한 이미지	생성된 이미지
Children sitting at computer stations on a long table.		
They are all focused on their monitor.		
The soft hum of the computers fills the air.		
The room is filled with the sound of clicking keyboards and occasional laughter, creating a lively and bustling atmosphere.		

조선대학교
CHOSUN UNIVERSITY

Ⅳ. 결 론

본 연구에서는 웹툰 저작활동에 도움을 줄 수 있도록 딥러닝 기반의 텍스트에서 이미지 생성기술을 사용하여 텍스트를 딥러닝 모델에 입력하면 웹툰을 생성하는 연구를 수행하였다.

실험을 위해 공개 데이터 셋인 MSCOCO을 사용하여 트리트먼트-웹툰 데이터 셋과 ChatGPT를 활용하여 연속적인 텍스트 데이터를 구축하였다. 적대적 생성 신경망 기반의 웹툰 스타일의 이미지 생성은 다국어 BERT를 사용하여 트리트먼트의 특징을 추출하고 노이즈와 결합하여 DCGAN에 입력하여 생성자와 판별자의 경쟁적 학습을 통해 웹툰 스타일의 이미지를 생성하였다. 실험 결과 inception score는 4.9, FID 22.21의 성능을 보여 비교적 낮은 성능의 결과를 보였다. 이러한 결과는 MSCOCO데이터의 실사 이미지를 CartoonGAN을 사용하여 툰 스타일 이미지로 변형할 때 이미지에 표현된 오브젝트의 형상이 일그러지며, 이러한 현상을 낮은 성능의 생성모델인 DCGAN이 학습할 때 형상을 정확히 학습하지 못한 것으로 예상된다. 또한 AttnGAN이나 CLIP, 확산 모델에서는 텍스트와 이미지 사이의 유사도를 학습할 때 반영하지만 본 연구에서 사용한 DCGAN은 유사도를 사용하지 않아 낮은 성능을 보인 것으로 예상된다. 비록 생성된 웹툰 스타일의 이미지의 품질이 낮긴하지만 실제 이미지에 표현된 형상을 표현할 수 있는 것을 확인할 수 있다

적대적 생성 신경망 기반의 웹툰 생성 결과의 단점을 극복하기 위해 트리트먼트-웹툰 데이터 셋을 CLIP에 학습하여 유사도를 측정한 뒤 확산 모델을 사용하여 웹툰 스타일의 이미지를 생성하였다.

CLIP는 트리트먼트-웹툰 데이터와 같은 멀티모달 데이터의 관계를 학습할 수 있는 모델로서 각각의 데이터의 특징을 추출하고 contrastive learning을 통해 같은 차원에서 특징이 유사한 데이터는 가깝게, 서로 다른 데이터는 멀도록

조선대학교
CHOSUN UNIVERSITY

학습한다. 트리트먼트-웹툰 데이터 셋을 학습한 CLIP의 성능 평가하기 위해 정량적 지표를 통해 확인하였고 한글과 영어로 이루어진 다국어 트리트먼트와 이미지 사이의 유사도 측정, 트리트먼트를 쿼리로 하여 데이터 셋의 유사한 이미지 검색, 제로샷 분류를 수행하였다. 학습 중 모델의 정량적 지표 중 하나인 train, eval 데이터의 학습 중 정확도와 손실은 학습 횟수가 증가함에 따라 증가하였지만 학습횟수 25회 이상이었을 때 감소하는 것을 확인할 수 있었다. 다국어 트리트먼트와 웹툰 이미지 사이의 유사도를 측정하였을 때 의미상으로 유사한 다국어 트리트먼트가 높은 수치의 유사도를 나타내었으며 PCA를 통해 시각화하였을 때 유사도가 높은 다국어 트리트먼트가 이미지와 가까운 거리에 임베딩된 것으로 나타났다. 뿐만아니라 의미가 같은 다국어 트리트먼트도 가까운 거리에 임베딩이 되는 것을 확인하였다. 트리트먼트를 쿼리로하여 데이터 셋에서 유사한 Top 4 이미지를 찾는 실험에서는 실제 이미지와 다른 이미지가 검색 되었지만 트리트먼트와 실제 이미지에 표현되어있는 특징이 반영된 이미지를 찾는 것을 확인할 수 있었으며 검색된 이미지들 사이에서도 유사한 형상을 나타내는 것으로 확인하였다. 제로샷 분류에서는 여러 가지의 텍스트를 레이블로 하여 이미지를 분류하였을 때 주어진 레이블 중 가장 근접한 레이블을 예측하는 것을 확인할 수 있다. CLIP 모델의 성능 지표는 MRR@K 지표를 사용하여 측정하였는데 MRR@1에서 0.23, MRR@5는 0.47, MRR@10은 0.51의 지표를 나타내어 준수한 성능을 나타낸 것을 알 수 있었다.

확산 모델을 기반으로 웹툰을 생성하기 위해 그리고자 하는 텍스트와 텍스트의 CLIP 특징과 가장 유사한 CLIP 이미지 특징을 갖는 이미지를 확산 모델 기반의 텍스트에서 이미지 생성 모델인 사전학습된 depth-to-image 모델에 입력하여 웹툰 이미지를 생성하였다. 확산 모델을 이용한 웹툰 생성은 하나의 텍스트를 이용한 웹툰 생성, 연속된 텍스트를 입력으로 웹툰을 생성하는 두 가지 방식의 실험을 수행하였다. 실험결과 연속된 텍스트를 입력으로 웹툰을 생성하였을 때 inception scroe는 7.14이며 적대적 생성 신경망을 사용하였을 때보다 높았으며 생성된 이미지의 품질 또한 높은 것을 확인할 수 있었다. 이러한 결과가 나타난 이유는 depth-to-image 모델은 모델에 입력한 이미지의 깊이 정보를 이용하여

조선대학교
CHOSUN UNIVERSITY

깊이 정보에 맞는 이미지를 생성한다. 또한, 생성할 때마다 다른 화풍의 이미지를 생성하는데, 이는 depth-to-image모델은 LAION-5B 데이터를 사용하여 사전학습하였기 때문에 다양한 이미지를 생성할 수 있는 것으로 예상된다.

본 연구에서 수행한 멀티모달을 이용한 웹툰 생성을 웹툰 작가들이 저작활동을 할 때 본 기술을 사용한다면 그리고자 하는 텍스트를 딥러닝 모델에 입력하면 웹툰을 생성할 수 있기 때문에 웹툰 저작시간을 단축할 수 있을 것으로 예상된다.

본 연구의 한계점으로는 여러개의 문장과 이미지를 고려하여 웹툰 스타일의 이미지를 생성하지 못하는 것이며, 일관성 있는 화풍의 이미지를 생성하지 못하였다. 추후 여러개의 문장을 입력할 수 있는 확산 모델에 대한 연구가 필요하며 연속된 문장이 입력되었을 때 일관성 있는 화풍의 이미지를 생성하는 연구가 필요하다.

조선대학교
CHOSUN UNIVERSITY

* 본 연구는 'A Study on Generating Webtoons Using Multilingual Text-to-Image Models' 논문을 발전시켰음(참고문헌 25).

참고문헌

[1] Agnese, J., Herrera, J., Tao, H., & Zhu, X., "A survey and taxonomy of adversarial neural networks for text-to-image synthesis", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(4), e1345.

[2] Wang, S., Zeng, W., Wang, X., Yang, H., Chen, L., Zhang, C., ... & Liu, J., "SwiftAvatar: Efficient Auto-Creation of Parameterized Stylized Character on Arbitrary Avatar Engines",InProceedings of the AAAI Conference on Artificial Intelligence, 2023, Vol. 37, No. 5, pp. 6101-6109.

[3] Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P., "A Survey on Deep Learning for Skeleton-Based Human Animation", InComputer Graphics Forum, 2022, Vol. 41, No. 1, pp. 122-157.

[4] Xu, P., Zhu, X., & Clifton, D. A., "Multimodal learning with transformers: A survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[5] Baltrušaitis, T., Ahuja, C., & Morency, L. P., "Multimodal machine learning: A survey and taxonomy", IEEE transactions on pattern analysis and machine intelligence, 2018, 41(2), pp. 423-443.

[6] 김가이, "웹툰전공 교과과정 개선방안에 대한 연구 : 서울·경기 소재 전문학사과정을 중심으로", 상품문화디자인학연구, 2022, 69, pp.141-150.

[7] Bhunia, A. K., Koley, S., Khilji, A. F. U. R., Sain, A., Chowdhury, P. N.,

조선대학교
CHOSUN UNIVERSITY

Xiang, T., & Song, Y. Z., “Sketching without worrying: Noise-tolerant sketch-based image retrieval”, InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 999-1008.

[8] Kang, S., Choo, J., & Chang, J., “Consistent comic colorization with pixel-wise background classification”, InProceedings of the NIPS, 2017, Vol. 17.

[9] 이순기, 스토리텔링 정형화 연구, 국내석사학위논문 세종대학교, 2019. 서울.

[10] Chen, Y., Lai, Y. K., & Liu, Y. J., “Cartoongan: Generative adversarial networks for photo cartoonization”, InProceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9465-9474.

[11] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H., “Generative adversarial text to image synthesis”, InInternational conference on machine learning, PMLR, 2018, pp. 1060-1069.

[12] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X., “Attngan: Fine-grained text to image generation with attentional generative adversarial networks”, InProceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1316-1324.

[13] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N., “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”, InProceedings of the IEEE international conference on computer vision, 2017, pp. 5907-5915.

[14] Qiao, T., Zhang, J., Xu, D., & Tao, D., “Mirrorgan: Learning text-to-image generation by redescription”, InProceedings of the IEEE/CVF Conference on

조선대학교
CHOSUN UNIVERSITY

Computer Vision and Pattern Recognition, 2019, pp. 1505-1514.

[15] Qiao, Y., Chen, Q., Deng, C., Ding, N., Qi, Y., Tan, M., ... & Wu, Q., "R-GAN: Exploring human-like way for reasonable text-to-image synthesis via generative adversarial networks", InProceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2085-2093.

[16] Ho, J., Jain, A., & Abbeel, P., "Denoising diffusion probabilistic models", Advances in neural information processing systems, 2020, 33, 6840-6851.

[17] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M., "Hierarchical text-conditional image generation with clip latents", 2022, arXiv preprint arXiv:2204.06125.

[18] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B., "High-resolution image synthesis with latent diffusion models", InProceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.

[19] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M., "Glide: Towards photorealistic image generation and editing with text-guided diffusion models", 2021, arXiv preprint arXiv:2112.10741.

[20] Kim, G., Kwon, T., & Ye, J. C., "Diffusionclip: Text-guided diffusion models for robust image manipulation", InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2426-2435.

[21] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B., "High-resolution image synthesis with latent diffusion models",

조선대학교
CHOSUN UNIVERSITY

InProceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.

[22] Dhariwal, P., & Nichol, A., "Diffusion models beat gans on image synthesis", Advances in neural information processing systems, 2021, 34, 8780-8794.

[23] Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., ... & Norouzi, M., "Palette: Image-to-image diffusion models", InACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1-10.

[24] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I., "Learning transferable visual models from natural language supervision", InInternational conference on machine learning, PMLR, 2021, pp. 8748-8763.

[25] Yu, K., Kim, H., Kim, J., Chun, C., & Kim, P., "A Study on Generating Webtoons Using Multilingual Text-to-Image Models", Applied Sciences, 2021, 13(12), 7278.

[26] MSCOCO, Available online: https://cocodataset.org/ (accessed on 1 January 2023).

[27] AI Hub, Available online: https://aihub.or.kr/ (accessed on 1 January 2023).

[28] Pires, T., Schlinger, E., & Garrette, D., "How multilingual is multilingual BERT?", 2019, arXiv preprint arXiv:1906.01502.

[29] stable-diffusion-2-depth, https://huggingface.co/stabilityai/stable-diffusion-2-depth (access on May 20)

조선대학교
CHOSUN UNIVERSITY

[30] GPT API, https://platform.openai.com/docs/guides/gpt (access on May 22)

조선대학교
CHOSUN UNIVERSITY

감 사 의 글

조선대학교에서 학사와 석사학위를 받고 짧으면 짧고 길다면 긴 시간이 흘러 박사학위를 마칠 수 있었습니다. 2009년 조선대학교 컴퓨터공학과에 입학하여 오랜 시간 동안 학교에 머물며 박사학위까지 받을 줄은 생각지도 못하였습니다. 학사과정을 마칠 즈음 게임프로그래밍에 관심이 있어 공부하였고 그 뒤 VR/AR에 관심이 있어 석사에 진학하였고 어려움이 있었지만, 인공지능에 관하여 석사학위를 받을 수 있었습니다. 석사학위가 계기가 되어 박사학위 또한 인공지능 분야, 특히 멀티모달 AI를 주제로 학위를 받게 되었습니다.

여려 모로 많은 일들과 고난이 있었지만 저를 생각해주시고 지도해주신 김판구 교수님께 감사드립니다. 교수님 덕분에 제안서 작성, 연구적 역량과 더불어 갖추어야 할 덕목에 대해 많은 것을 배울 수 있었습니다. 그리고 어려움이 많았던 석사과정의 지도교수에서부터 박사학위 심사위원장까지 맡아주신 양희덕 교수님께 감사드립니다. 교수님의 넓은 아량이 없었다면 지금까지 올 수 없었으리라 생각됩니다. 학교 세미나나 학회에서 최신 연구 동향에 대해 말씀해주시고 저의 학위논문에 대해서도 심도 있게 봐주신 황명권 교수님께 감사드립니다. 미국에서 유학생활을 할 때 지도해주시고 직접 찾아와주신 임기호 교수님께 감사드립니다. 마지막으로 함께 웹툰 관련 과제를 진행하며 저의 논문에 대해 상세하게 지적해주시고 연구에 대한 방향을 지도해주신 전찬준 교수님께 감사드립니다.

박사과정에서 희로애락을 함께한 분들이 있습니다. 노주현, 김준혁, 김강민 학생 그리고 연구실 선배인 김정인, 김형주, 홍택은 선배님들과 박사과정을 함께 하여 매우 기쁩니다. 그리고 석사과정부터 박사과정까지 항상 도움을 준 이자즈에게도 고마움을 전합니다.

지금까지 저를 아낌없이 지원해주시고 사랑해주신 존경하는 아버지 유규열, 어머니 이춘례에게 감사의 말씀을 드립니다. 부모님의 사랑이 없었더라면 많은

조선대학교
CHOSUN UNIVERSITY

고난 속에서 다시 일어설 힘이 없었을 것입니다. 한집에서 자라오며 곁에 있어준 누나 유은정, 유금영 그리고 귀여운 조카 김가윤, 김태윤에게 고마움을 전합니다.

제 옆에서 응원과 격려를 해준 미래를 함께하기로 약속한 김현정에게 고마움과 사랑을 전합니다. 현정씨를 만나면서 정신적인 안정을 찾았고 따뜻한 마음으로 남은 박사과정 기간을 보낼 수 있었습니다. 저의 박사학위논문이 앞으로 함께 할 날들에 있어 한 걸음 내디딜 수 있는 디딤돌이 됐으면 합니다.

위의 모든 분께 다시 한번 감사의 말씀을 드리며 이 논문에서 해결하지 못한 연구를 앞으로도 열심히 해 나가겠습니다.

감사합니다.

2023년 8월

유 경 호 올림