한국형 멀티모달 몽타주 앱을 위한 생성형 AI 연구

임 정 현 * · 차 경 애 ** · 고 재 필 *** · 홍 원 기 ****

목 차

요약

멀티모달 (multi-modal) 생성이란 텍스트, 이미지, 오디오 등 다양한 정보를 기반으로 결과를 도출하는 작업을 말한다. AI 기술의 비약적인 발전으로 인해 여러 가지 유형의 데이터를 종합적으로 처리해 결과를 도출하는 멀티모달 기반 시스템 또한 다양해지는 추세이다. 본 논문은 음성과 텍스트 인식을 활용하여 인물을 묘사하면, 몽타주 이미지를 생성하는 AI 시스템의 개발 내용을 소개한다. 기존의 몽타주 생성 기술은 서양인들의 외형을 기준으로 이루어진 반면, 본 논문에서 개발한 몽타주 생성 시스템은 한국인의 안면 특징을 바탕으로 모델을 학습한다. 따라서, 한국어에 특화된 음성과 텍스트의 멀티모달을 기반으로 보다 정확하고 효과적인 한국형 몽타주 이미지를 만들어낼 수 있다. 개발된 몽타주 생성 앱은 몽타주 초안으로 충분히 활용 가능하기 때문에 기존의 몽타주 제작 인력의 수작업을 획기적으로 줄여줄 수 있다. 이를 위해 한국지능정보사회진흥원의 AI-Hub에서 제공하는 페르소나 기반 가상 인물 몽타주 데이터를 활용하였다. AI-Hub는 AI 기술 및 서비스 개발에 필요한 인공지능 학습용 데이터를 구축하여 원스톱 제공을 목적으로 한 AI 통합 플랫폼이다. 이미지 생성 시스템은 고해상도 이미지를 생성하는데 사용하는 딥러닝 모델인 VQGAN과 한국어 기반 영상생성 모델인 KoDALLE 모델을 사용하여 구현하였다. 학습된 AI 모델은 음성과 텍스트를 이용해 묘사한 내용과 매우 유사한 얼굴의 몽타주 이미지가 생성됨을 확인할 수 있다. 개발된 몽타주 생성 앱의 실용성 검증을 위해 10명의 테스터가 사용한 결과 70% 이상이 만족한다는 응답을 보였다. 몽타주 생성 앱은 범죄자 검거 등 얼굴의 특징을 묘사하여 이미지화하는 여러 분야에서 다양하게 사용될 수 있을 것이다.

표제어: 멀티모달 생성 AI, 이미지 생성 시스템, VQGAN, KoDALLE, 몽타주

접수일(2024년 02월 19일), 수정일(2024년 03월 12일), 게재확정일(2024년 03월 19일)

* 제1저자, 대구대학교 AI학부 학부생, dalek1568@gmail.com
** 공동저자, 대구대학교 AI학과 교수, chaka@daegu.ac.kr
*** 공동저자, 국립금오공과대학교 컴퓨터공학과 교수, nonezero@kumoh.ac.kr
**** 교신저자, 대구대학교 컴퓨터정보공학부 교수, wkhong@daegu.ac.kr

1. 서론

현대 디지털 환경에서는 텍스트, 이미지, 음성 등 다양한 형태의 데이터가 동시에 존재한다. 이것을 효과적으로 처리함과 동시에 정보를 통합하기 위해 연구되고 있는 AI 분야 중 하나가 바로 멀티모달(multi-modal) 생성이다. 멀티모달 생성은 언어 생성과 이미지 생성 등을 통합하는 방식으로 이루어질 수 있는데, 이를 통해 시각적이고 언어적인 정보를 결합하여 더 풍부하고 의미 있는 결과물을 얻을 수 있다. 그 예로 이미지 캡션 생성, 텍스트 설명에 기반한 이미지 생성 혹은 음성과 이미지를 결합한 다양한 응용 등이 있다.

본 논문에서는 텍스트 및 음성과 이미지로 구성된 멀티모달을 입력으로 하여 몽타주를 생성하는 AI 어플리케이션의 개발 내용을 소개한다. 개발 앱은 한국어 텍스트 기반 영상 생성 (text to image generation) 작업을 할 수 있는 KoDALLE 을 기반으로 한다(KoDALLE, 2024). VQGAN(Esser et al., 2021)을 인코더로 하는 GAN 모델을 생성하였으며, AI-Hub에서 제공하는 페르소나 기반 가상 인물 몽타주 데이터를 활용하여 학습을 진행하였다.

몽타주는 수사에 활용되는 중요한 수사 도구 중 하나이다(Joh and Park, 2018). 그러나 기존의 몽타주 생성 기술은 사람이 직접 목격자의 증언에 따라 용의자의 몽타주를 그려나가는 방법으로 시간과 공간의 제약을 많이 받게 된다. 목격자가 조금이라도 용의자의 얼굴을 또렷하게 기억할 때 몽타주를 그려야 하는데 기존의 방법으로는 목격자의 기억이 흐려질 수 있다는 치명적인 단점이 있는 것이다. 이에 반해 본 논문에서 소개하는 몽타주 생성 앱은 위에서 기존 몽타주 생성 방법의 단점인 시간과 공간에 제약받지 않도록 하기위해 어플리케이션을 활용하여 목격자가 용의자의 얼굴을 확인했을 때 바로 몽타주를 생성할 수 있도록 개발하였다. 물론 기존에도 이러한 시도가 없었던 것은 아니지만 이것이 힘들었던 이유는 동양인의 얼굴 데이터베이스 한계로 인해 목격자의 기억을 정확하게 재현하는데 어려움이 많았기 때문이다. 하지만 페르소나 기반 가상 인물 몽타주 데이터와 한국어에 특화된 음성, 텍스트의 멀티모달을 기반으로 보다 정확하고 효과적인 한국형 몽타주 이미지를 만들어낼 수 있었다. 이를 통해 국내 범죄자 검거뿐만 아니라 실종자 수색, 안면 추정, 가상 얼굴 생성 등의 분야에서도 활용할 수 있는 안면 인식 및 재구축 기술의 발전을 가져올 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어를 기반으로 한 텍스트를 입력받아 이미지를 생성하기 위한 KoDDALE와 VQGAN을 중심으로 생성 AI에 대한 최근 연구를 소개한다. 3장에서는 몽타주 생성 AI를 위한 시스템을 소개하고 학습 데이터를 이용한 모델 학습 내용을 설명한다. 4장에서는 개발한 앱을 소개하고 실험 결과를 분석하며, 5장에서 결론을 맺는다.

2. 관련 연구

최근 자연어로 기술한 내용에 부합하는 영상을 자동으로 생성하는 딥러닝 모델이 주목 받고 있다. 대표적인 모델은 DALLE(Ramesh et al., 2021), Stable Diffusion(Rombach et al., 2022), Imagen(Saharia et al., 2022)이다. DALLE의 구조는 아래 그림 <Fig.2-1>과 같다. DALLE는 자연어 처리부와 이미지 생성부로 구성된다. 자연어 처리부에서 나온 결과는 이미지 생성부의 입력으로 들어간다. 자연어 처리부는 인간의 언어를 이미지 생성부가 다룰 수 있도록 코딩하여 벡터화한다. 이를 텍스트 임베딩(text embedding)이라고 부른다. 일반적으로는 임베딩 벡터라고도 하고 자연어처리 분야에서는 토큰이란 명칭으로 통용된다.

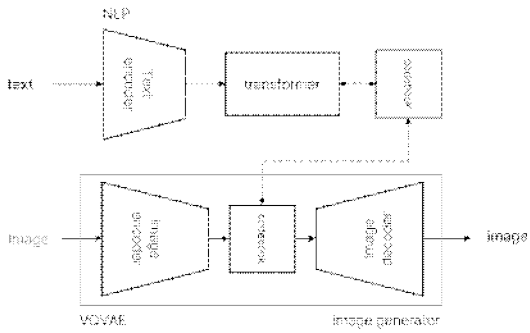

Fig. 2-1 DALLE Structure

텍스트 임베딩은 사전에 학습된 대형언어모델인 GPT-3(Brown et al., 2020)을 사용한다. 자연어로 생성된 토큰을 입력하여 영상을 생성하는 영상 생성부는 다시 두 부분으로 나뉜다. 첫번째는 텍스트 토큰 시퀀스를 입력으로 하여 이미지 토큰 시퀀스에 해당하는 이미지 패치 시퀀스를 생성하는 이미지 패치 생성부이다. 두번째는 이미지 패치 시퀀스를 하나의 영상으로 합성해 내는 이미지 생성부이다. 전자는 시퀀스 데이터에 특화된 트랜스포머(Vaswani et al., 2017) 구조로 설계되었으며, 후자는 대표적인 생성모델인 VAE(Kingma and Welling, 2013)를 개선한 모델인 VQVAE(Van Den Oord & Vinyals, 2017)를 사용한다.

한국어를 다루기 위해서는 한국어로 사전에 학습된 대형언어모델(Pre-trained Language Model; PLM)이 필요하다. 네이버, SKT, 카카오브레인에서 각각 KLUE RoBERTa(Park et al., 2021), KoGPT Trinity of SKT(KoGPT Trinity of SKT, 2024), KoGPT of Kakao Brain(KoGPT of Kakao Brain, 2024)를 공개하고 있다. 국내에서 가장 널리 사용되고 있는 모델은 KLUE RoBERTa이다. 이 모델은 다시 신경망 모델의 파라미터 수에 따라 small, base, large로 나뉜다.

2.1 KoDALLE

KoDALLE는 DALLE 구조에서 GPT-3를 KLUE RoBERTa-large 모델로 대체한 한국어 기반 영상 생성 모델이다. 아래 표 <Tab. 2-1>은 두 모델의 차이를 항목별로 비교한다. 주목할 점은 DALLE 대비 용량을 획기적으로 줄였다는 점이다. 이는 하드웨어 리소스가 충분하지 않아도 특정 분야 영상 생성에 한정된 전용 모델을 훈련시킬 수 있다는 것을 의미한다. 실제로도 표에서 제시한 KoDALLE의 특성은 AI-Hub의 패션 데이터(AI-Hub K-Fashion, 2024) 생성에 특화하여 훈련한 경우이다.

Tab. 2-1. DALLE vs KoDALLE

	DALLE	KoDALLE
Train Dataset Size	250 Million Pairs	0.8 Million Pairs
#Params	12 Billion	428 Million
#Layers	64 Layers	16 Layers
Computing Resource	1024 x V100 16GB	1 x V100 32GB
Text Encoder	16384 Vocab x 512 Dim BPE	32000 Vocab x 1024 Dim klue/robertalarge
Image Encoder	VQ-VAE	VQ-GAN
Optimizer	AdamW	AdamW

구체적으로 DALLE 대비 KoDALLE를 비교해 보면, 용량은 12억개 대비 4억2천8백만개에 불과하다. 대신 훈련 데이터의 수는 2억5천개 대비 80만개로 충분하다. 영상패치 생성을 위한 트랜스포머는 DALLE의 64개층과는 달리 16층으로 축소하였다. 한편, 컨텍스트가 강한 한국어에 걸맞게 임베딩 벡터의 크기는 512에서 1024로, 훈련한 단어의 종류는 16,384에서 32,000개로 각각 두 배 수준으로 증가하였다. 기술적인 면에서는 DALLE의 영상 생성부와는 달리 KoDALLE에서는 VQVAE대신

VQVAE와 GAN(Goodfellow et al., 2020)의 장점을 취한 VQGAN(Esser et al., 2021)을 채택하였다. 이 모델은 세밀한 표현면에서 더욱 우수하다.

2.2 VQGAN

VAE(Variational Autoencoder)는 생성 모델 중 하나이다. 주어진 학습데이터를 특정한 확률 분포로부터 샘플링하는 모델로, 이를 사용하여 새로운 데이터를 생성할 수 있게 한다. 하지만 디코더가 너무 강해 생성되는 데이터의 벡터값이 무시되는 문제점이 있어 이러한 문제점을 개선하여 향상된 모델인 VQGAN을 사용한다.

VQGAN은 두 가지 딥러닝 모델의 장점을 더하여 만들어진 모델이다. 아래 그림 <Fig. 2-2>에서 보는 것처럼 첫 번째는 CNN 모델이 가지는 특성으로, 주어지지 않은 입력의 출력을 예측하는 특성과 물체의 특징을 명확히 추출하는 장점과 두 번째로는 Transformer 모델의 글로벌 모델링이 가능하며 전역적인 관계를 잘 읽고 특징을 이해하는 장점을 결합하여 만들어진 모델이다.

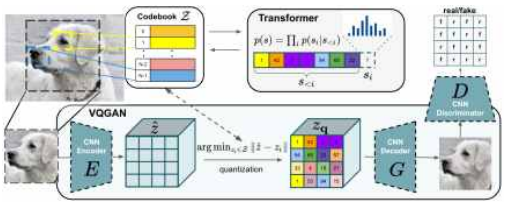

Fig. 2-2 VQGAN (Esser et al., 2021)

VQGAN은 주로 이미지나 음성과 같은 복잡한 데이터에서 좋은 학습 성능을 보이며, 생성적인 작업이나 데이터의 효율적인 표현을 위해 사용되는데 이러한 특징이 텍스트 기반 영상 생성 작업을 수행하는 KoDALLE에 적합하다.

2.3 몽타주 어플리케이션

기존에 2D 몽타주를 생성하는 방법은 대부분 PC 환경에서 숙련된 전문가들이 사용하는 프로그램을 통해 목격자들의 기억을 토대로 생성되는 방법을 이용한다. 일반 사용자가 아닌 전문가들이 프로그램을 다뤄야만 하는 이유는 몽타주를 생성하는 과정에서 얼굴의 특정 부위 미세 조정과 이로 인해 생겨난 다양한 얼굴들을 확인시켜 주며 목격자가 유사하다는 얼굴이 나올 때까지 반복해야 하는 등 하나의 몽타주를 얻기 위해 아래의 <Fig 2-3>과 같이 여러 가지 복잡한 작업이 이루어지기 때문이다.

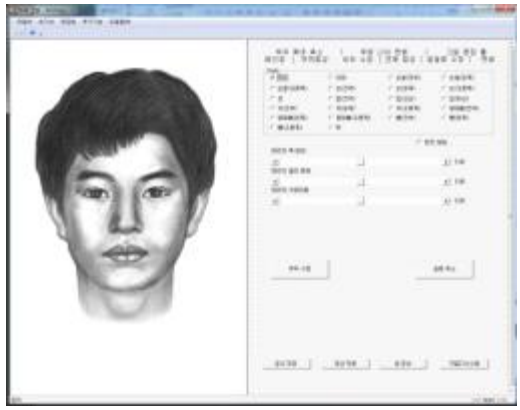

Fig. 2-3 Existing montage creation program (Park et al., 2014)

따라서, 목격자의 기억이 왜곡되기 전에 몽타주를 생성해야만 하는데 목격자가 전문가를 찾아가고 전문가를 통해 용의자의 인상을 설명할 때까지 그 기억이 왜곡되거나 오염될 가능성이 높다. 따라서 목격자의 기억이 왜곡되기 전에 목격자가 몽타주 초안을 만들 수 있도록 개발였다. 본 논문에서는 코틀린 기반의 Android Studio를 사용하여 UI/UX를 구성하고 훈련을 마친 몽타주 생성 모델을 Tensorflow Lite의 모델 생성 기능을 통해 어플리케이션에서 몽타주 생성을 가능하도록 구현한다.

3. 멀티모달 기반 몽타주 생성 AI 개발

본 논문은 KoDALLE를 이용하여 한국어로 얼굴의 특징을 묘사하면 이에 부합하는 몽타주 영상을 생성한다. 몽타주 생성을 위해 AI-Hub에제 제공하는 몽타주 데이터(AI-Hub Montage, 2024)를 사용하여 KoDALLE를 새롭게 훈련한다.

3.1 시스템 구성

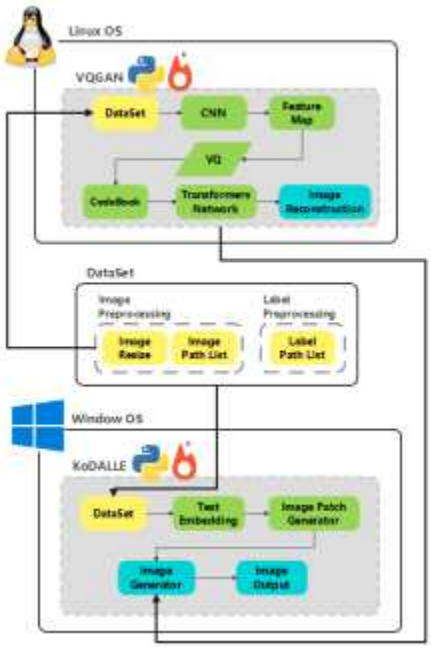

Fig. 3-1 Learning Model System Diagram

위 그림 <Fig. 3-1>은 학습 모델의 시스템 구성도이다. 전체적인 순서는 모델 학습을 진행하기 전 확보한 데이터를 256X256 크기로 재조정하고 라벨 데이터는 클리닝하였다. 전처리 된 이미지 데이터를 VQGAN의 입력으로 사용해 이미지 재건축 작업을 반복 학습하여 페르소나 몽타주 이미지의 화풍을 학습하고 KoDALLE의 이미지 생성기의 역할을 하는 디코더로 사용한다.

KoDALLE에서는 이미지 데이터와 라벨 데이터를 모두 사용하여 이미지의 부분적인 특징은 라벨 데이터를 통해 파악하고 이를 패치로 생성한다. 이렇게 생성된 패치를 페르소나 몽타주 이미지의 화풍을 학습 완료한 VQGAN에서 사용하게 되고 최종적으로는 음성 및 텍스트에 해당하는 몽타주 이미지를 만드는 작업을 학습한다.

아래 그림 <Fig. 3-2>는 학습이 완료된 KoDALLE를 사용자 UI인 어플리케이션으로 구현하기 위한 구성도이다. Android Studio에서 KoDALLE를 최적화하여 사용하기 위해 Tensorflow Lite로 모델을 변환한다. AodbeXD로 제작한 어플리케이션 화면과 Tensorflow Lite로 변환한 KoDALLE 모델을 사용해 어플리케이션의 형태로 제작하였다.

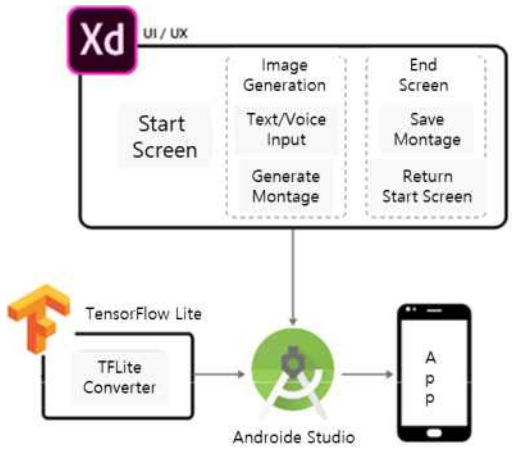

Fig. 3-2 App UI Module Diagram

3.2 학습 데이터 구축

데이터셋은 AI 허브에 있는 '페르소나 기반의 가상 인물 몽타주 데이터' 를 사용한다. 데이터셋은 한국인 가상 인물 이미지(8,071장)를 기반으로 스케치 된 이미지들을 사용하고 자세한 정보는 아래 표 <Tab. 3-1>과 같다.

설명문 기반 몽타주 스케치에는 스케치의 퀄리티

를 High, Medium, Low 3가지로 나누어 작업된 스케치 이미지들이 있다. 따라서 하나의 몽타주에 대한 화풍을 학습할 때 퀄리티가 다른 3개의 스케치를 학습시킬 수 있게 데이터셋을 사용할 수 있다. 이에 따라 사용된 데이터셋은 설명문 기반 몽타주 스케치 이미지 High(8,071장), Medium(8,071장), Low(8,071장)에 각각 퀄리티 별로 가상 인물 육안 관찰 인물 스케치를 추가해주어 총 48,426장의 이미지 데이터 및 라벨 데이터를 확보하여 사용한다. 이를 8대 2의 비율로 학습 데이터에 약 38,700장, 테스트 데이터에 약 9,600장을 나눠 사용한다.

Tab. 3-1 Training Dataset

데이터 유형	이미지
데이터 형식	- 한국인 가상 인물 이미지, 가상인물 육안 관찰 인물 스케치: png - 설명문 기반 몽타주 스케치: jpg
데이터 구축량	- 한국인 가상 인물 이미지: 8,071장 - 가상 인물 육안 관찰 인물 스케치 : 8,071장 - 설명문 기반 몽타주 스케치 : High 8,071장, Medium 8,071장, Low 8,071장 -> 총 24,213장
라벨링 유형	- 바운딩 박스 (이미지) - 텍스트 (자연어)
라벨링 형식	- json 형식

라벨 데이터는 얼굴에서 볼 수 있는 특징들 위주로 라벨링되었으며 아래 표 <Tab. 3-2>에서 자세한 정보를 확인할 수 있다. 라벨 데이터는 KoDALLE에서 이미지 데이터와 함께 입력 데이터로 사용되며 하나의 이미지마다 각 라벨에 해당하는 부분의 패치를 생성하고 이를 토큰화하는 작업을 거쳐 VQGAN이 학습한 화풍으로 최종적인 몽타주 이미지를 생성하는 과정을 학습한다.

Tab. 3-2 Definition of Data Input Label

얼굴	유형	특징	수염(턱수염, 콧수염)
	크기		구렛나룻
	이마 유형		보조개
	이마 크기		흉터
	턱 유형		점
	턱 크기		주근깨
	볼		잡티
	얼굴 서술		문신
			화장
			특징서술
헤어	앞머리 길이	눈	크기
	옆머리 길이		유형
	가르마		눈 사이 거리
	헤어 서술		눈꼬리
			눈 서술
눈썹	유형	코	크기
	짙음 정도		길이
	길이		콧대
	두께		코끝 모양(콧망울)
	미간 넓이		콧볼 넓이(콧날개)
	눈썹 서술		코밑 길이(인중 길이)
			코 서술
인상	인상분류	입	유형
			크기
			입술 모양
	인상 서술		입술 두께
			입꼬리
			인중선
			입 서술
주름	이마 주름	목	목 서술
	미간 주름		
	눈주름		
	팔자주름		
	광대 주름		
	입술 주름		
	목주름		
	주름서술		

3.3 VQGAN과 KoDALLE 모델 학습

학습 모델 시스템 구성도에서 제시한 순서에 따라 KoDALLE의 디코더인 VQGAN을 사전에 전처리한 페르소나 몽타주 이미지 데이터 약 48,000장으로 학습을 진행하기 위한 컴퓨팅 환경을 설정한다. 컴퓨팅 환경은 VQGAN의 내부 참조 라이브러리의 일부가 Linux 운영체제에서만 동작하기 때문에

VQGAN의 학습은 Linux에서 진행하였다. VQGAN 학습을 위한 컴퓨팅 환경과 학습 파라미터는 각각 아래 표 <Tab. 3-3>과 <Tab. 3-4>와 같다.

Tab. 3-3 Computing Environment for training VQGAN

운영체제	Ubuntu 20.04.6 LTS
CPU	Intel(R) Core(TM) i5-10400
RAM	32.0GB
GPU	NVIDIA GeForce RTX 2070 SUPER with 8GB GDDR6
Python 버전	3.10
Pytorch 버전	1.7.0

Tab. 3-4 Learning Parameter for VQGAN

이미지 데이터량	48,426
학습 횟수 (epoch)	20
image size	256 X 256
batch size	8
num workers	4
학습 소요 시간	약 100시간

Linux에서 VQGAN을 충분히 학습시킨 후, KoDALLE의 학습은 Window 운영체제에서 진행하였다. KoDALLE의 학습을 진행한 컴퓨팅 환경과 학습 파라미터는 각각 아래 표 <Tab. 3-5>와 <Tab. 3-6>과 같다.

학습이 진행된 결과로 생성된 각 모델의 성능 측정을 시각화하기 위해 Weight & Biases의 AI 개발자 플랫폼(Weight&Biases, 2024)을 사용하였으며, 아래 그림 <Fig. 3-3>와 같은 그래프를 얻었다. Weight & Biases로 생성된 그래프에서 각 모델의 손실율을 확인할 수 있는데 이 값이 0에 수렴할수록 고품질의 학습 결과를 출력하는 것으로 볼 수 있다. VQGAN은 화풍을 학습하는 단순 이미지 재건축 작업을 진행하기에 손실율이 0에 잘 수렴하는 것을 보여준다.

Tab. 3-5 Computing Environment for training KoDALLE

운영체제	Window 11 Pro (64bit)
CPU	Intel(R) Core(TM) i5-10400
RAM	32.0GB
GPU	NVIDIA GeForce RTX 2070 SUPER with 8GB GDDR6
Python 버전	3.8
Pytorch 버전	1.6.0

Tab. 3-6 Learning Parameter for KoDALLE

이미지 데이터량	48,426
라벨 데이터량	48,426
학습 횟수 (epoch)	5
이미지 크기	256 X 256
number of text tokens	10000
text sequence length	256
filter threshold	0.9
학습 소요 시간	약 10시간

아래 그림 <Fig. 3-4>를 입력 원본 이미지로 사용하였을 때 아래 그림 <Fig. 3-5>와 같이 재건된 생성 결과를 통해 학습 모델의 성능을 확인할 수 있었다. 세 가지 형태의 여성 이미지와 한가지 형태의 남성 이미지를 학습한 결과 입력 이미지와 유사한 화풍으로 이미지를 생성하고 있음을 확인할 수 있다. KoDALLE는 텍스트 임베딩을 거쳐 패치를 생성해 이를 토대로 이미지를 생성하는 더 복잡한 작업을 하다 보니 손실율이 학습을 거치며 점진적으로 0에 수렴하는 것을 볼 수 있었고 학습 횟수 5회 이상에서는 더 이상 유의미한 변화는 없었기에 학습 횟수 5회에서 훈련을 마쳤다.

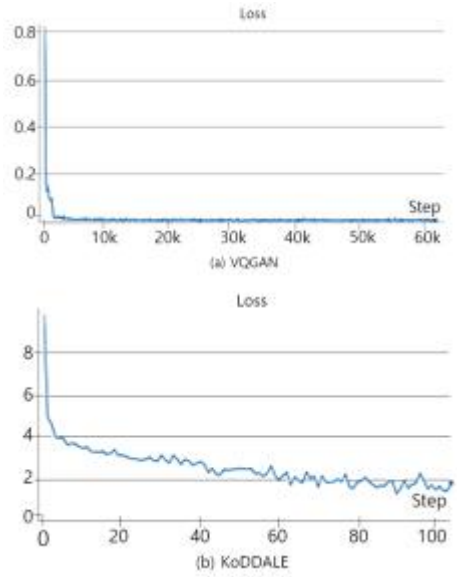

Fig. 3-3 Loss Rate of Learning Outcome

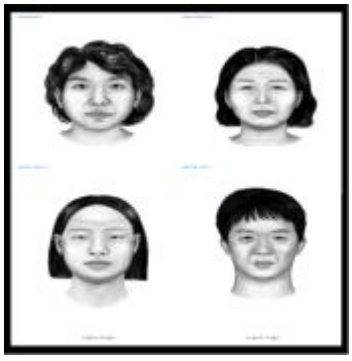

Fig. 3-4 Source Image for VQGAN learning

학습이 완료된 모델의 생성 성능을 아래 표 <Tab. 3-7>에서 확인할 수 있다. 여성과 남성의 몽타주 설명을 한 개씩 넣어 테스트한 결과를 봤을 때 몽타주 설명과 많은 부분이 일치하는 모습을 보임이 확인된다. 여성의 경우 눈매와 입매가 부드럽고 유순한 성격으로 보이는 텍스트에 대해서 이와 유사한 느낌의 얼굴 형상이 생성되었다. 또한 남성의 경우도 이마라인과 M자형, 가르마 등의 서술되 내용과 흡사한 얼굴 형상이 생성되었다.

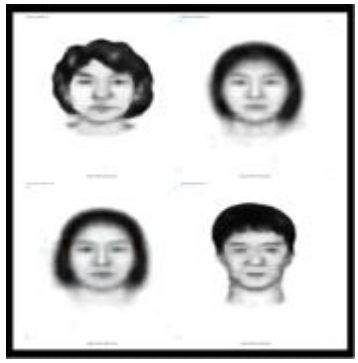

Fig. 3-5 Reconstructed Image on VQGAN

Tab. 3-7 Montage generated by KoDALLE

입력(텍스트)	출력(몽타주)
눈매와 입매가 부드럽고 미소 짓고 있는 모습으로 성품이 착하고 유한 사람으로 보인다. 약간 여성스러운 성격일 수도 있을 것 같으며 누구에게나 친절하고 어진 모습으로 대해주는 편안한 사람으로 느껴진다.	
이마 라인 가운데가 아래로 살짝 내려와 있어 M모양을 띄고 있으며 가르마의 구분 없이 모두 올려져 있으나 헤어라인 가운데에서 살짝 오른편으로 헤어젤을 바른 듯 조금 나뉜 모습도 보인다. 옆머리는 왼쪽은 귓바퀴 시작 지점까지 내려와 있으며 오른쪽은 귀에 바로 닿을 듯 내려와 있다.	

4. 실험 결과

훈련한 학습 모델이 플랫폼 상에서 제대로 동작하는지 확인 및 실험하기 위해 3.3절에서 기술한 환경을 기반으로 학습한 KoDALLE 모델을 적용하여 몽타주 생성 앱을 개발하였다. 아래 그림 <Fig. 4-1>, <Fig. 4-2>와 <Fig. 4-3>은 개발한 몽타

주 생성 앱의 실행 화면을 보여주고 있다. <Fig. 4-1> 어플리케이션의 시작 화면에서 사용자는 두 개의 버튼 중 시작 버튼으로는 몽타주 생성을 시작할 수 있고 앨범 버튼으로는 생성했던 몽타주 이미지들을 확인할 수 있다. 시작 버튼을 눌렀을 때 나오는 몽타주 생성 화면에서는 음성인식 또는 키보드 타자를 통한 방법으로 몽타주 묘사를 입력할 수 있다.

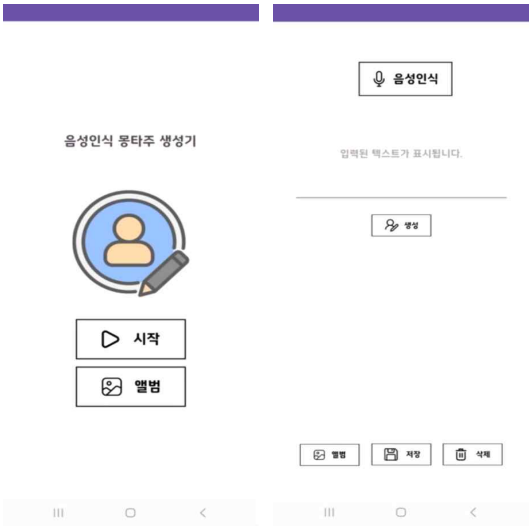

Fig. 4-1 Start Up Screen of App

입력을 완료한 후 생성 버튼을 눌렀을 때는 <Fig. 4-2>와 같이 입력받은 사람의 생김새를 토대로 몽타주 이미지가 약 3~5초 후 생성된다. 생성된 몽타주에 대한 이미지는 해당 갤러리에 정상적으로 저장되었음을 <Fig. 4-3>에서 확인할 수 있으며 이를 통해 SNS 공유, 초기화, 메인 버튼 등 기능적인 버튼 모두 오류 없이 정상적으로 동작한다. 따라서 음성과 텍스트 인식을 통해 몽타주 이미지 초안을 생성함으로써 몽타주 제작에 효율성을 가져올 수 있다.

Fig. 4-2 Input Screen for Image Description

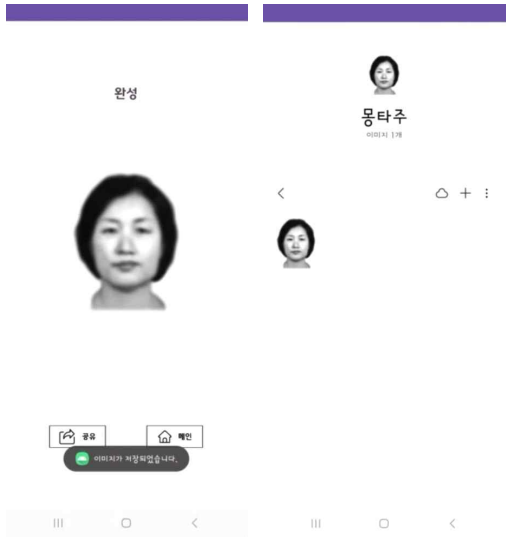

Fig. 4-3 Generated Montage

이렇게 실험이 완료된 어플리케이션을 테스터들을 모집하여 설문 방식으로 성능 조사를 실시하였다. 테스트 참가 인원은 총 10명으로 테스트 방법은 학습에 사용되지 않은 몽타주 데이터를 참가자들에게 나눠주고 다음 몽타주 이미지의 생김새를 음성

또는 텍스트로 묘사하여 몽타주 이미지를 생성했을 때 자신이 봤던 몽타주 이미지와 얼마나 유사하게 나왔는지 1~10점의 점수로 표현하는 방법을 사용했다. 10명 중 5명의 참가자들에게는 최대한 자세하게 묘사를 요청했고, 나머지 5명의 참가자에게는 얼굴의 특징만 묘사하도록 요청했다.

그 결과 최대한 자세하게 묘사를 요청한 그룹의 점수는 평균 7.6점, 적당히 묘사를 요청한 그룹의 점수는 평균 4.4점을 보여줬다. 이후 참가자들에게 별도로 테스트에 사용한 이미지의 라벨 데이터를 사용한 결과를 보여줬을 때 이미지의 만족도를 조사한 결과 매우 만족 1명, 만족 3명, 보통 4명, 불만족 2명, 매우 불만족 0명으로 확인되었다. 대체로 묘사한 특징을 반영한 몽타주를 보여줬다고 평가하였으나, 이미지 생성 후 묘사 텍스트의 일부를 다시 수정했을 때 반영이 제대로 되지 않는 점을 불만족의 이유로 지적하였다.

5. 결론

본 논문에서는 인공지능의 학습 능력으로 페르소나 몽타주 이미지의 화풍을 학습하고 음성과 텍스트를 입력해 사람을 대신하여 몽타주를 생성하는 AI 모델을 제안하고 생성하였다. 몽타주 생성 과정은 VQGAN을 이미지 생성기로 사용하는 KoDALLE를 통해 입력된 몽타주 설명에서 패치를 찾아내 몽타주 이미지를 생성하게 된다.

사용자 UI로 사용자가 목격한 용의자의 몽타주를 시간과 장소에 제약받지 않고 빠른 시간 내에 몽타주로 만들 수 있도록 어플리케이션으로 제작하였다. 위의 4장에서 참가자들의 테스트 결과를 통해 몽타주 생성 어플리케이션의 실용성에 대해 검증을 해볼 수 있었고 범죄자 검거 등 얼굴의 특징을 묘사하여 이미지화하는 여러 분야에서 다양하게 사용될 수 있는 가능성을 확인할 수 있었다.

본 논문의 모델 훈련에 사용된 데이터 셋은 AI 허브에서 제공하는 데이터로 제한되어 있다. 따라서, 충분한 양의 페르소나 가상인물 데이터 셋을 추가 확보하여 몽타주 이미지의 화풍 및 정확도 개선이 필요하다. 또한, 입력한 텍스트로 생성된 몽타주 이미지에 대해 특정 부위를 수정 묘사한 텍스트를 반영한 이미지 성능 개선 연구를 추후 진행할 계획이다.

[References]

[1] AI-Hub K-Fashion(2024), https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=51

[2] AI-Hub Montage(2024), https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=618

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D.(2020), Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901

[4] Esser, P., Rombach, R., and Ommer, B.(2021), Taming Transformers for

High-Resolution Image Synthesis, *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873-12883

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.(2020), Generative adversarial networks. *Communications of the ACM*, 63(11), pp. 139-144

[6] Joh, H., and Park, B.S.(2018), A Comparative Study of Montage investigation and portrait investigation. *가천법학*, 11(3), pp. 235-264 (조현진, 박병식(2018), 몽타주 수사와 초상화 수사의 비교 연구, *가천법학*, 제11권 3호, pp.235-264)

[7] Kingma, D.P., and Welling, M.(2013). Auto-encoding variational bayes. *arXiv preprint arXiv*:1312.6114

[8] KoDALLE (2024), *https://github.com/KR-HappyFace/KoDALLE*

[9] KoGPT of Kakao Brain (2024), https://github.com/kakaobrain/kogpt

[10] KoGPT Trinity of SKT (2024), https://github.com/SKT-AI/KoGPT2

[11] Park, B., Nam, S., Chang, H. and Choi, C. (2013), EsFit - A facial composites methodology to help eyewitness, *Annual Conference of IEIE*, 1393-1396 (박보훈, 남상준, 장희정, 최창석. (2013). EsFit-목격자 진술을 최소로 하는 몽타주 작성 방법, *2013년도 대한전자공학회 하계종합학술대회*, 1393-1396)

[12] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.-W., and Cho, K. (2021), Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*

[13] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I.(2021), Zero-shot text-to-image generation, *In International Conference on Machine Learning,* 8821-8831

[14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022), High-resolution image synthesis with latent diffusion models, *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 10684-10695

[15] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M.(2022), Photorealistic text-to-image diffusion models with deep language understanding, *Advances in Neural Information Processing Systems,* 35, 36479-36494

[16] Van Den Oord, A., and Vinyals, O.(2017),

Neural discrete representation learning, *Advances in neural information processing systems,* 30

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017), Attention is all you need. *Advances in neural information processing systems*, 30

[18] Weight&Biases(2024), *https://wandb.ai/site*

Lim Jeounghyun (dalek1568@gmail.com)

Jeounghyun Lim graduated from Daegu University, Department of AI, majoring in AI Entertainment. He is currently pursuing studies related to Artificial Intelligence and application development.

Kyung-Ae Cha (chaka@daegu.ac.kr)

Kyung-Ae Cha is a professor in the Department of Artificial Intelligence at Daegu University. She received her Ph.D. in Computer Science from Kyungpook National University in 2003. Her research areas include artificial intelligence and smart applications.

Jaepil Ko (nonezero@kumoh.ac.kr)

Jaepil Ko is a professor in the Department of Computer Engineering at Kumoh National Institute of Technology. He received his Ph.D. in Computer Science from Yonsei University in 2004. His research areas include Computer Vision and Machine Learning.

Won-Kee Hong (wkhong@daegu.ac.kr)

Won-Kee Hong is a professor in the Department of Information Security at Daegu University. He received his Ph.D. in Computer Science from Yonsei University in 2001. His research areas include Information Security and Artificial intelligence.

Research on Generative AI for Korean Multi-Modal Montage App

Lim, Jeounghyun* · Cha, Kyung-Ae** · Koh, Jaepil*** · Hong, Won-Kee****

ABSTRACT

Multi-modal generation is the process of generating results based on a variety of information, such as text, images, and audio. With the rapid development of AI technology, there is a growing number of multi-modal based systems that synthesize different types of data to produce results. In this paper, we present an AI system that uses speech and text recognition to describe a person and generate a montage image. While the existing montage generation technology is based on the appearance of Westerners, the montage generation system developed in this paper learns a model based on Korean facial features. Therefore, it is possible to create more accurate and effective Korean montage images based on multi-modal voice and text specific to Korean. Since the developed montage generation app can be utilized as a draft montage, it can dramatically reduce the manual labor of existing montage production personnel. For this purpose, we utilized persona-based virtual person montage data provided by the AI-Hub of the National Information Society Agency. AI-Hub is an AI integration platform aimed at providing a one-stop service by building artificial intelligence learning data necessary for the development of AI technology and services. The image generation system was implemented using VQGAN, a deep learning model used to generate high-resolution images, and the KoDALLE model, a Korean-based image generation model. It can be confirmed that the learned AI model creates a montage image of a face that is very similar to what was described using voice and text. To verify the practicality of the developed montage generation app, 10 testers used it and more than 70% responded that they were satisfied. The montage generator can be used in various fields, such as criminal detection, to describe and image facial features.

Keywords: Multi-modal Generative AI, Image Generation System, VQGAN, KoDALLE, Montage

* First Author, Department of Artificial Intelligence, Daegu University.
** Coauthor, Professor, Department of Artificial Intelligence, Daegu University.
*** Coauthor, Professor, Department of Computer Engineering, Kumoh National Institute of Technology.
**** Corresponding Author, Professor, Department of Information Security, Daegu University.