IMMU: IMage to MUsic 시퀀스를 이용한 SNS 플랫폼 기술

김민성(*), 김태환(**), 윤수완(***), 이기성(****)
(*, **, ***) 중앙대학교 AI학과, {maruniverse, thwan11, swyoon0312}@cau.ac.kr
(****) 중앙대학교 인문콘텐츠연구소, goory@cau.ac.kr

1. 연구 배경

현시점 SNS 플랫폼의 트랜드는 음악과 함께 사진이나 영상을 포스트할 수 있는 플랫폼이다. 다만 기존의 SNS 서비스에서는 이미 존재하는 음악만 쓸 수 있다는 한계가 있어, 사용자가 포스팅 중 사진의 분위기를 온전히 담는 음악을 찾는 것은 쉬운 일이 아니다. 음원 서비스에서는 1억 개 이상의 곡들을 하나씩 찾아가며 가장 잘 맞는 곡을 찾아 포스팅하는 것은 사용자들로 하여금 노래를 첨부하는 것을 꺼리게 만들 수 있다. 이에 기존에 사진이나 영상을 통해 음악을 추천하는 서비스가 존재하지만, 우리는 기존의 음악 추천 시스템을 넘어, 사용자에게 신선한 경험을 제공하고 더 개성 있고 자신의 개성이 가미된 포스트를 올릴 수 있도록 자동으로 새로운 음악을 만들어 주는 시퀀스 IMMU 를 제시한다.

2. 연구 내용

IMMU 프로젝트의 핵심은 사진을 업로드하면 그에 어울리는 음악을 생성해 내는 시퀀스와 이를 활용할 SNS를 제작하는 것이다. 이러한 SNS를 구현하려면 사진에 어울리는 음악을 생성하는 인공지능을 구현하고, SNS 형식의 어플리케이션을 디자인해야 한다. 우리는 사진에 어울리는 음악을 생성하는 인공지능을 구현하기 위해 BLIP[1], GPT[2], Riffusion을 이용했다.

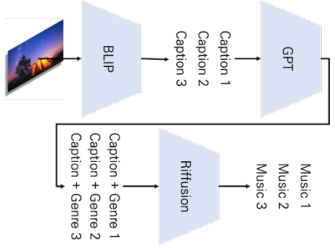

그림 1: 음악 생성 프로세스

2.1 BLIP, GPT

BLIP은 Image Captioning에서 SOTA를 기록했던 모델이다. 우리는 이미지의 다양한 정보를 추출하도록 BLIP을 이용해 이미지에 해당하는 캡션 3개를 출력하도록 했으며, GPT를 이용해 3개의 캡션을 하나의 문장으로 합치고 이에 어울리는 음악 장르 3개도 합치도록 했다.

2.2 Riffusion

오디오 데이터는 Short Time Fourier Transform을 이용해 Spectrogram으로 표현할 수 있고 그 역도 성립한다. x축은 시간을 나타내고 y축은 주파수를 나타낸다. 각 픽셀의 색상은 행과 열에 의해 주어진 시간과 주파수에서 오디오의 진폭을 나타낸다. Stable Diffusion[3]을 파인튜닝하면 텍스트 입력으로 Spectrogram을 생성할 수 있게 되며 이것을 다시 오디오로 변환하는 모델을 Riffusion이라 한다. IMMU 시퀀스 에서는 앞서 생성된 3개의 문장을 Riffusion에 넣어 3종류의 음악을 생성 및 저장하도록 했다.

그림 2: Python환경 실행 화면

Python 환경에서 BLIP, GPT, Riffusion을 이용하여 사진에서 캡션과 장르를 추출하고, 추출한 정보를 통해 음악을 생성하는 모델을 개발했다.

그림 2에서 붉은색 최상단 박스 부분은 BLIP을 통해 이미지에서 3개의 캡션을 추출한 결과물이다. 그 바로 아래 박스 부분은 GPT를 통해 3개의 캡션을 합치고 이에 어울리는 음악 장르 3개를 추출한 결과물이며, 최하단의 박스 부분은 3개의 문장을 Riffusion에 넣어 3종류의 음악을 생성한 결과물이다.

2.3 SNS 플랫폼 프로토타입

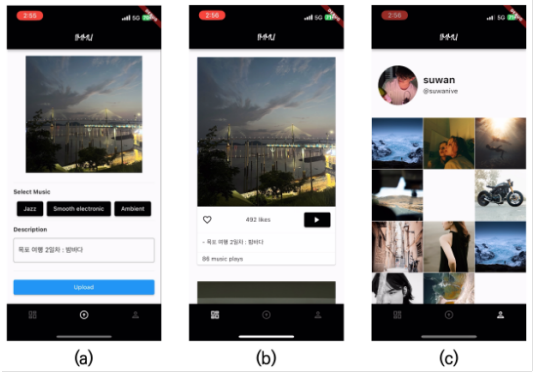

그림 3: (a) 업로드 페이지, (b) 피드, (c) 프로필

우리는 IMMU 기술을 이용한 SNS 플랫폼을 개발하기 위해서 Flutter를 사용했다. Flutter는 구글에서 개발한 크로스 플랫폼 앱 개발 언어이다. 우리는 이를 통해 IOS, Android, 웹 등에서 구동이 가능한 SNS 플랫폼을 구현하였다.

3. 현실적 제한 조건에 대한 검토

Stable Diffusion은 픽셀 기반 접근 방식에 비해 계산 요구 사항을 상당히 줄여주지만, 여전히 순차적인 샘플링 과정이 GANs보다 느리다. 이는 SNS와 같이 사용자와 빠른 소통을 요구하는 환경에서 걸림돌이 될 수 있다.

또한 현재 연구에서는 이미지에서 캡션을 추출하는 방식을 사용했지만, 감성 분석을 통해 이미지의 분위기나 감정을 직접적으로 파악하는 방법을 추가한다면 더 감각적이고 세련된 음악을 생성하는 데 도움이 될 수 있다.

4. 기대효과 및 활용

본 연구의 결과로 IMMU 시퀀스를 통한 사진-음악 생성 시스템이 제안되었으며, 이 기술은 사용자가 더욱 창의적이고 개성 있는 콘텐츠를 제작할 수 있도록 돕는 새로운 SNS 경험을 제공한다. 또한, 해당 기술은 시각장애인을 위한 새로운 형태의 감각 경험을 제공할 수 있는 잠재력을 가지며, 다양한 분야에서의 응용 가능성을 지닌다. 이러한 연구 결과는 향후 IMMU 시퀀스 및 관련 기술이 SNS 플랫폼뿐만 아니라 다양한 콘텐츠 생성 도구로 확장될 수 있는 가능성을 제시한다.

5. 결론

본 연구에서는 사진과 어울리는 음악을 자동으로 생성하는 IMMU 시퀀스와 이를 활용한 SNS 플랫폼을 제안하였다. IMMU 시퀀스는 이미지에서 다양한 정보를 추출하여 음악을 생성하는 일련의 과정을 통해 기존 SNS 플랫폼에서의 음악 선택의 한계를 극복하고자 한다. 이를 위해 BLIP, GPT, Riffusion을 각각 캡션 생성, 캡션 통합 및 음악 장르 추출, 그리고 음악 생성 부분에 활용하였으며, 콘텐츠 생성 도구로의 확장 가능성을 제시했다.

참고 문헌 (참고자료)

[1] Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International conference on machine learning. PMLR, 2022.

[2] Brown, Tom B. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.