음악 가사와 배경 이미지 매칭 모델

서가연[1], 박수연[2], 신형환[2], 조준수[2], 강세이[2], 정재준[2], 서보경[1], 김승원[3]
[1]전남대학교 인공지능학부 학부생
[2]전남대학교 인공지능융합학과 대학원생
[3]전남대학교 인공지능융합학과 교수
215001@jnu.ac.kr, suyb1234@jnu.ac.kr, gudghks@jnu.ac.kr, whwnstn@jnu.ac.kr,
200793@jnu.ac.kr, triplej@jnu.ac.kr, sbk0301@jnu.ac.kr, Seungwon.Kim@jnu.ac.kr

Music Lyrics and Background Image Matching model

Gayun Suh[1], Su-Yeon Park[2], Hyeong-Hwan Shin[2], Jun-Su Cho[2], Sei Kang[2],
Jae-Joon Jeong[2], Bo-Gyeong Seo[1], Seung-Won Kim[3]
[1]Dept. of Artificial Intelligence, Chonnam National University
[2]Dept. of Artificial Intelligence Convergence, Chonnam National University
[3]Dept. of Artificial Intelligence Convergence, Chonnam National University

요 약

본 연구는 음악 가사와 배경 이미지를 매칭하는 시스템을 개발하는 데 초점을 맞추고 있다. GPT-4o를 활용하여 배경 이미지에 어울리는 음악 가사를 생성해 데이터셋을 구축하였으며, Long-CLIP 모델을 미세 조정하여 음악 가사와 배경 이미지의 임베딩 벡터를 비교함으로써 가장 적합한 배경 이미지를 추천하는 시스템을 구현하였다.

1. 서론

디지털 콘텐츠 제작 분야에서 시각적 요소와 청각적 요소의 조합은 감정적 영향력을 극대화하고, 사용자 경험을 풍부하게 만드는 중요한 방법론으로 자리 잡고 있다. 특히, 음악 가사와 배경 이미지를 매칭하는 시스템은 이러한 융합적 접근의 핵심적인 도구로서, 콘텐츠의 몰입감을 강화하고 시청자와의 감정적 연결을 깊게 하는 데 이바지한다.

본 논문에서는 GPT-4o(Generative Pre-Trained Transformer-4o)를 활용하여 배경 이미지와 어울리는 음악 가사를 생성해 음악 가사-배경 이미지 매칭 데이터 세트를 구축했다. 이를 바탕으로 음악 가사 텍스트와 배경 이미지의 임베딩 벡터를 구하고 유사도를 비교하여 입력된 음악 가사에 어울리는 배경 이미지를 추천해주는 시스템을 구현했다.

2. 관련 연구

기존의 BERT(Bidirectional Encoder Representations from Transformers)[1]와 같은 신경망 기반 검색 모델은 주로 문서-질의 사이의 유사성을 계산하는 데 초점을 맞췄다. 하지만 기존의 신경만 기반 검색 모델은 같은 도메인(텍스트-텍스트) 간의 유사도를 다루기 때문에 텍스트-이미지와 같은 서로 다른 도메인의 유사도를 계산하는 데는 한계가 있다.

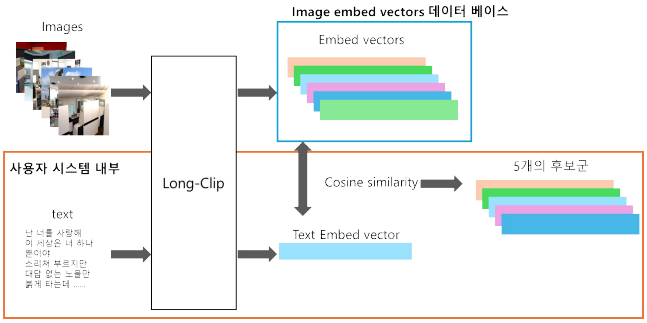
CLIP(Contrastive Language-Image Pre-training)[2] 모델은 텍스트와 이미지를 대규모 데이터 세트로 함께 학습하여 자연어 설명과 이미지를 같은 벡터 공간으로 매핑시켜 서로 다른 도메인 간의 유사도를 계산하는 문제를 해결하였다. 하지만 CLIP 모델은 일반적으로 간결한 문장이나 프롬프트를 활용해 학습시켰기에 음악 가사와 같은 긴 텍스트 문장을 처리하는 데 한계가 있다.

본 연구는 일반적인 이미지 묘사를 위한 간결한 문장보다 상대적으로 훨씬 긴 길이의 음악 가사 텍스트 입력을 다룬다. 이를 해결하기 위해 Long-CLIP[3] 모델을 도입하였다. Long-CLIP 모델은 기존의 CLIP 모델보다 더 긴 입력을 받을 수 있도록 설계된 모델로, 길이가 긴 텍스트와 이미지 간의 매칭 정확도를 높일 수 있다.

3. 구현

3D 배경 모델의 섬네일 이미지와 GPT-4o 모델을 활용하여 훈련 데이터 세트와 테스트 데이터 세

트를 구축하였다. 배경 데이터의 이미지를 “이미지와 어울리는 음악 가사를 작성해줘”라는 프롬프트와 함께 모델의 입력으로 넣어 데이터 세트를 구축하였다. 이렇게 생성된 데이터 세트는 총 267개의 이미지와 생성된 음악 가사로 구성되었다. 해당 데이터 세트를 활용하여 사전 학습된 Long-CLIP 모델을 GeForce RTX 3090 그래픽 카드를 사용하여 100에폭 동안 학습시켰다.



(그림 1) 전체 프레임워크.

음악 가사 입력에 따른 배경 이미지 추천시스템은 그림 1과 같이 구현하였다. 사용자 입력에 따른 적합한 배경 이미지를 추천하기 위해서 미세 조정된 Long-CLIP 모델로 사전에 이미지 임베딩 벡터를 계산한 데이터베이스를 구축하였다. 이에 따라 사용자가 시스템에 음악 가사를 입력하면 텍스트의 임베딩 벡터를 계산하고, 구축한 이미지 임베딩 데이터베이스와의 코사인 유사도를 비교하여 가장 유사한 이미지 5개를 사용자에게 제공한다.

4. 결론

음악 가사와 배경 이미지를 매칭하는 시스템을 개발하였다. 하지만 제안하는 시스템에 몇 가지 한계점이 존재한다. 첫 번째로, 구축한 데이터 세트의 규모가 상대적으로 작다. 더 다양한 종류의 배경 이미지와 음악 가사로 작성된 데이터 세트를 사용하면 더 좋은 성능을 달성할 수 있다. 두 번째로, 긴 입력을 처리하기 위해 Long-CLIP 모델을 사용하였음에도 음악 가사를 전부 입력으로 사용하기에는 토큰 길이가 충분하지 못하다.

향후 연구에서 더 많은 이미지와 다양한 장르의 음악 가사에 대한 데이터 세트를 구축하여 학습을 진행하고 제안한 모델에 대한 다각적인 성능을 평가할 필요가 있다. 또한, Long-CLIP 모델보다 더 긴 텍스트 입력을 받을 수 있는 모델을 고안하면 긴 음악 가사의 텍스트에 최적화를 통해 정확도를 높일 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

이 논문은 광주정보문화산업진흥원의 재원으로 아시아 문화기술 실증센터 운영기관 구축사업 내의 2023 실감 콘텐츠 데이터응용 서비스개발 지원사업의 지원을 받아 수행된 연구임. 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음 (IITP-2023-RS-2023-00256629).

참고문헌

[1] Jacob Devlin, Wing-Wei Chang, Kenton Lee, Kristina Toutanova “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019, p.2.

[2] Alect Radford, Jong Wook Kim, Chris Hallacy et al. “Learning Transferable Visual Models From Natural Language Supervision” Proceedings of the 38th International Conference on Machine Learning, Virtual, 2021, pp. 8748-8763.

[3] Beichen Zhang, Pan Zhang, Xiaoyi Dong et al. “Long-CLIP:Unlocking the Long-Text Capability of CLIP” arXiv preprint, 2024, arXiv:2403.15378.