비디오 배경음악 엔드투엔드 프레임워크: 비디오 콘텐츠를 위한 자동 캡셔닝 및 음악 생성

이재건 [1],권태준 [2],최장훈 [3]
[1] 경북대학교 데이터사이언스 석사과정
[2] 경북대학교 데이터사이언스 석사과정
[3] 경북대학교 데이터사이언스 교수

leejken530@knu.ac.kr, ktjmamamoo0629@knu.ac.kr, jhchoi09@knu.ac.kr

Video-to-Background Music End-to-End Framework: Automatic Captioning and Music Generation for Video Content

Jaegun Lee[1], Taejun Kwon [2], Janghoon Choi[3]
[1]Dept. of Data Science, Kyungpook National University
[2]Dept. of Data Science , Kyungpook National University
[3]Dept. of Data Science, Kyungpook National University

요 약

본 논문에서는 비디오 콘텐츠에 대한 자동 캡셔닝과 그에 맞는 배경음악 생성을 위한 통합 프레임워크를 제안한다. 우리의 접근 방식은 최첨단 비디오 이해 모델인 VideoLLaMA3 와 텍스트 기반 음악 생성 모델인 YuE 를 결합하여 비디오의 시각적 내용을 분석하고 이에 적합한 배경음악을 자동으로 생성한다. 제안된 워크플로우는 입력 비디오를 VideoLLaMA3 모델에 전달하여 비디오의 내용, 분위기, 감정을 설명하는 캡션을 생성하고, 이 캡션을 YuE 모델에 입력하여 비디오의 시각적 내용과 조화를 이루는 배경음악을 생성한다. 다양한 장르와 분위기의 비디오에 대한 실험 결과, 우리의 프레임워크는 비디오 내용에 적합한 고품질 배경음악을 생성할 수 있음을 보여주었다. 이 연구는 비디오 콘텐츠 제작, 자동 멀티미디어 콘텐츠 생성, 그리고 접근성 향상을 위한 응용 프로그램에 기여할 수 있다.
Code : https://github.com/2JAE22/Video-to-BackgroundMusic
Project_page: https://2jae22.github.io/projects/CV/Video-to-BackgroundMusic/

1. 서론

멀티미디어 콘텐츠의 생성과 소비가 급증함에 따라, 비디오 콘텐츠를 자동으로 이해하고 이에 적합한 오디오 요소를 생성하는 기술에 대한 수요가 증가하고 있다. 비디오 캡셔닝은 비디오의 시각적 내용을 자연어로 설명하는 작업으로, 콘텐츠 검색, 접근성 향상, 그리고 자동 메타데이터 생성에 중요한 역할을 한다. 한편, 비디오에 적합한 배경음악은 시청자의 몰입감을 높이고 감정적 반응을 유도하는 데 필수적이다. 최근 대규모 비전-언어 모델(VLM)의 발전으로 비디오 내용을 정확하게 이해하고 설명하는 능력이 크게 향상되었다. 특히 VideoLLaMA3[1] 와 같은 모델은 비디오의 시간적, 공간적 정보를 효과적으로 처리하여 상세하고 정확한 캡션을 생성할 수 있다. 또한, 텍스트 기반 음악 생성 모델인 YuE[2] 는 자연어 설명을 바탕으로 다양한 스타일과 감정을 표현하는 음악을 생성할 수 있는 능력을 보여주고 있다. 본 연구에서는 이러한 두 기술을 통합하여 비디오 콘텐츠에 대한 자동 캡셔닝과 그에 맞는 배경음악 생성을 위한 엔드투엔드 프

레임워크를 제안한다. 우리의 접근 방식은 비디오의 시각적 내용을 분석하여 캡션을 생성하고, 이 캡션을 바탕으로 비디오의 분위기와 내용에 적합한 배경음악을 자동으로 생성한다. 이러한 통합 프레임워크는 비디오 콘텐츠 제작자들에게 시간과 비용을 절약할 수 있는 도구를 제공하고, 자동화된 멀티미디어 콘텐츠 생성 시스템의 발전에 기여할 수 있다.

2. 관련연구

2.1 비디오 캡셔닝

비디오 캡셔닝은 컴퓨터 비전과 자연어 처리의 교차점에 위치한 연구 분야로, 비디오의 시각적 내용을 자연어로 설명하는 작업이다. 초기 접근 방식은 주로 CNN[3] 과 RNN[4] 을 결합하여 비디오의 특징을 추출하고 이를 텍스트로 변환하는 방식이었다. 최근에는 Transformer[5] 기반 아키텍처가 도입되면서 비디오의 시간적, 공간적 정보를 더 효과적으로 처리할 수 있게 되었다. VideoLLaMA, BLIP[6], LLaVA[7] 와 같은 대규모 언어-비전 모델들은 방대한 양의 이미지-텍스트 쌍과 비디오-텍스트 쌍으로 사전 학습되어 비디오 내용에 대한 상세하고 정확한 설명을 생성할 수 있게 되었다. 특히 VideoLLaMA3 는 비디오의 시간적 다이내믹스를 효과적으로 포착하고, 장면 전환, 객체 움직임, 인간 행동 등을 정확하게 설명할 수 있는 능력을 보여주고 있다.

2.2 텍스트 기반 음악 생성

텍스트 기반 음악 생성은 자연어 설명을 바탕으로 음악을 자동으로 생성하는 기술이다. 초기 연구는 주로 규칙 기반 시스템이나 통계적 모델을 사용했으나, 최근에는 딥러닝 기반 접근 방식이 주류를 이루고 있다. 특히 Transformer 기반 모델과 생성적 적대 신경망(GANs)을 활용한 연구가 활발히 진행되고 있다. YuE 와 같은 모델은 텍스트 프롬프트를 입력으로 받아 해당 설명에 맞는 음악을 생성할 수 있다. 이러한 모델들은 장르, 분위기, 악기 구성, 템포 등 다양한 음악적 요소를 텍스트 설명에서 추출하여 이에 맞는 음악을 생성한다. 최근 연구에서는 감정, 장면 묘사, 스토리텔링 등 더 복잡한 텍스트 프롬프트를 처리할 수 있는 능력이 향상되고 있다.

2.3 멀티모달 콘텐츠 생성

멀티모달 콘텐츠 생성은 텍스트, 이미지, 오디오, 비디오 등 여러 모달리티를 결합하여 새로운 콘텐츠를 생성하는 연구 분야이다. 최근 연구에서는 서로 다른 모달리티 간의 변환과 조화를 위한 다양한 접근 방식이 제안되고 있다. 비디오와 음악의 조화에 관한 연구는 주로 비디오의 시각적 특성과 음악의 청각적 특성 간의 상관관계를 모델링하는 데 중점을 두고 있다. 이러한 연구들은 비디오의 리듬, 색상, 움직임 등의 시각적 요소와 음악의 템포, 음색, 멜로디 등의 청각적 요소 간의 매핑을 학습하는 방향으로 진행되고 있다.

3. 방법론

본 연구에서 제안하는 프레임워크는 비디오 캡셔닝과 텍스트 기반 음악 생성을 통합하여 비디오 콘텐츠에 적합한 배경음악을 자동으로 생성한다. 전체 워크플로우는 크게 두 단계로 구성된다

(1) VideoLLaMA3 를 사용한 비디오 캡셔닝
(2) YuE 를 사용한 텍스트 기반 음악 생성.

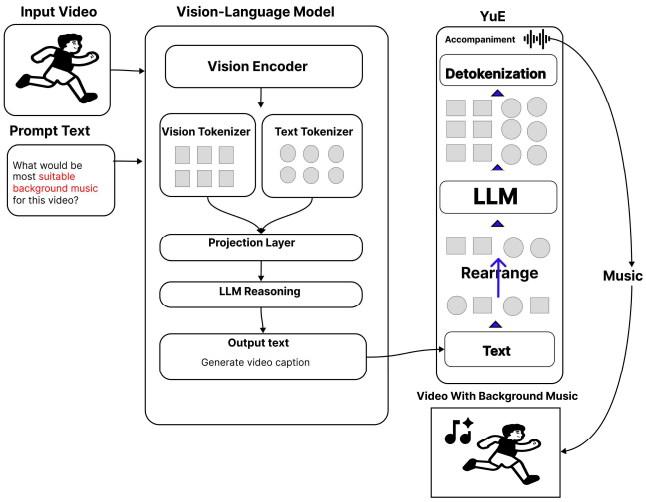

그림 1. 입력 비디오의 내용을 이해하여 가장 어울리는 배경음악을 자동으로 생성하고 삽입하는 Video-to-Music 파이프라인

3.1 시스템 아키텍처

우리의 시스템 아키텍처는 첨부된 이미지에 나타난 것처럼 두 개의 주요 모듈로 구성된다. 첫 번째 모듈은 입력 비디오를 처리하고 캡션을 생성하는 VideoLLaMA3 모델이다. 두 번째 모듈은 생성된 캡션을 바탕으로 배경음악을 생성하는 YuE 모델이다. 입력 비디오는 먼저 VideoLLaMA3 모델에 전달되어 비디오의 내용, 분위기, 감정을 설명하는 캡션이 생성된다. 이 캡션은 두 가지 주요 요소로 구성된다(1) 비디오의 시각적 내용을 반영한 적절한 텍스트 생성 (Caption) (2) YuE 에 적합한 음악 태그 생성 (Music Tag). 생성된 캡션은 YuE 모델에 입력되어 비디오의 시각적 내용과 조화를 이루는 배경음악이 생성된다. 본 모델은 VLM(Vision-Language Model)과 YuE 모듈로 구성되어 있으며, VLM 은 비디오 캡션을 텍스트로 생성하고, YuE 모듈은 이 텍스트를 바탕으로 적절한 배경음악을 생성한다.

3.2 VideoLLaMA3 를 사용한 비디오 캡셔닝

VideoLLaMA3 는 비디오의 시간적, 공간적 정보를 효과적으로 처리하여 상세하고 정확한 캡션을 생성할

수 있는 대규모 언어-비전 모델이다. 이 모델은 비디오 프레임을 입력으로 받아 비디오의 내용을 자연어로 설명하는 캡션을 생성한다. 우리의 접근 방식에서는 VideoLLaMA3 모델을 사용하여 다음과 같은 두 가지 유형의 캡션을 생성한다:

- **Caption**: 비디오의 시각적 내용에 대한 객관적인 설명
- **Music**: Caption 과 YuE 에 형식에 맞는 텍스트 생성

이러한 다양한 유형의 캡션은 YuE 모델이 비디오의 내용과 분위기에 적합한 배경음악을 생성하는 데 필요한 정보를 제공한다.

3.3 YuE 를 사용한 텍스트 기반 음악 생성

YuE 는 텍스트 프롬프트를 바탕으로 음악을 생성하는 모델로, 장르, 분위기, 악기 구성, 템포 등 다양한 음악적 요소를 텍스트 설명에서 추출하여 이에 맞는 음악을 생성한다. 우리의 프레임워크에서는 VideoLLaMA3 에서 생성된 캡션을 YuE 모델에 입력하여 비디오의 내용과 분위기에 적합한 배경음악을 생성한다. YuE 모델은 두 가지 주요 컴포넌트로 구성된다:

- **VLM(Vision-Language Model)**: 이미지 캡션을 텍스트로 생성하는 모듈
- **YuE**: 텍스트 설명을 바탕으로 음악을 생성하는 모듈

우리의 접근 방식에서는 VideoLLaMA3 모델이 비디오의 내용을 설명하는 캡션을 생성하며, 이 캡션은 별도의 변환 과정 없이 직접 YuE 모델의 입력으로 전달된다. YuE 모델은 입력된 캡션을 바탕으로 음악의 장르와 분위기를 자동으로 결정하여 비디오의 시각적 내용에 적합한 배경음악을 생성한다. 현재의 데모 구현에서는 lyrics.txt 파일에 별도의 가사를 생성하지 않았으나, 향후 VideoLLaMA3 의 캡션 생성 프롬프트를 수정하여 가사 생성을 활성화할 수 있으며, 이를 통해 배경음악에 가사를 추가할 수 있다.

감사의 말

본 연구는 과학기술정보통신부의 재원으로 한국연구재단 및 정보통신기획평가원의 지원을 받아 수행되었음 (과제번호: NRF-2021R1C1C2095450, RS-2023-00242528, RS-2024-00437756)

참고문헌

[1] Zhang, Boqiang; Li, Kehan; Cheng, Zesen; Hu, Zhiqiang; Yuan, Yuqian; **et al. "VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding."** *arXiv preprint* arXiv:2501.13106, 2025. DOI: 10.48550/arXiv.2501.13106

[2] Yuan, Ruibin; Lin, Hanfeng; Guo, Shuyue; **et al. "YuE: Scaling Open Foundation Models for Long-Form Music Generation."** *arXiv preprint* arXiv:2503.08638, 2025. DOI: 10.48550/arXiv.2503.08638

[3] LeCun, Yann; Bottou, Léon; Bengio, Yoshua; Haffner, Patrick. **"Gradient-Based Learning Applied to Document Recognition."** *Proceedings of the IEEE*, **86**(11): 2278–2324, 1998. DOI: 10.1109/5.726791

[4] Elman, Jeffrey L. **"Finding Structure in Time."** *Cognitive Science*, **14**(2): 179–211, 1990. DOI: 10.1207/s15516709cog1402_1.

[5] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; **et al. "Attention Is All You Need."** *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, 2017, pp. 6000–6010. *(Introduced the Transformer architecture.)*

[6] Li, Junnan; Li, Dongxu; Xiong, Caiming; Hoi, Steven C. H. **"BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation."** *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, Baltimore, USA, 2022, pp. 12888–12900. *(Best-of-breed vision-language pre-training method.)* DOI: 10.48550/arXiv.2201.12086.

[7] Liu, Haotian; Li, Chunyuan; Wu, Qingyang; Lee, Yong Jae. **"Visual Instruction Tuning."** *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, USA, 2023. *(Introduces LLaVA, a vision-language instruction-tuned model.)* DOI: 10.48550/arXiv.2304.08485.