CLIP 모델과 RAG 를 활용한 음성대화형 도슨트 서비스

정윤희 [1], 유강인 [1], 이민서 [1], 송예림 [1], 이정준 [2]
[1] 한국공학대학교 컴퓨터공학부 학부생
[2] 한국공학대학교 컴퓨터공학부 교수
{uniing, gangin0218, okok5317, shgfd5641, jjlee}@tukorea.ac.kr

A Voice-Interactive Docent Service Utilizing CLIP Model and RAG

Yun-Hee Jeong[1], Kang-In Ryu[1], Min-Seo Lee[1], Ye-Rim Song[1], Jeong-Joon Lee[2]
[1]Dept. of Computer Engieering, Tech University of Korea
[2]Dept. of Computer Engieering, Tech University of Korea

요 약

본 논문에서는 기존 전시 안내 서비스가 지닌 단방향성, 설명 부족, 그리고 소규모 전시에서의 접근성 한계를 극복하기 위한 도슨트 앱인 Artrip 의 설계·구현을 제시한다. Artrip 은 CLIP 기반 이미지 인식 모듈로 사용자가 촬영한 작품을 자동 식별하며, RAG 기반 생성형 질의응답을 통해 개인화된 작품 해설을 제공한다. 또한 관리자 페이지를 마련하여 작가 및 전시 주최자가 직접 작품 정보를 등록·갱신할 수 있도록 함으로써, 제공하는 설명의 최신성과 신뢰성을 보장한다. 이를 통해 Artrip 은 관람객에게는 깊이 있는 관람 경험을 제공하고, 전시 기획자와 작가에게는 효과적인 홍보·소통 채널을 제공하는 전시 안내 서비스를 제안한다.

1. 서론

최근 국내 문화예술에 대한 관심 증가로 전시 관람 수요가 확대되고 있으나[1], 대부분의 미술관 · 박물관은 단방향 오디오 가이드에 의존하고 있어 관람객의 개별 관심을 충족하지 못한다. 또한 오디오 가이드는 관람객의 자유로운 탐색을 방해하고 특정한 시각을 강요함으로써 관람 경험을 통제한다는 비판도 있다[2]. 특히 소규모 전시나 신진 작가의 전시에서는 해설 인력 부족으로 작품 이해가 더욱 어렵고, 작가 역시 홍보 수단이 제한된다.

본 논문에서는 이러한 문제를 해결하기 위해, 이미지 인식과 생성형 AI 기반의 음성 도슨트 앱 Artrip 을 제안한다. 제안하는 앱은 현장 도슨트에 비해 자유로운 질의응답이 가능하고, RAG 구조를 기반으로 특화된 대화내용을 쉽게 구성할 수 있어 소규모 전시장에 유용할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 주요 기능으로 이미지 인식 기반 작품 식별 방법, RAG 기반 생성형 질의응답의 구조를 설명하고, 이어서 관리자 콘텐츠 등록 방법을 설명한다. 제 3 장에서는 세부 설계 및 구현을 설명하고, 제 4 장에서 결론을 맺는다.

2. 주요 기능

본 논문에서 제안하는 Artrip 은 음성대화형 질의응답을 통해 깊은 사용자 경험을 제공하고, 이미지 인식 기반 작품 식별 기능을 통해 사용자가 정해진 동선에 얽매이지 않고 자유롭게 작품을 관람할 수 있도록 지원한다. 또한, 작가 및 전시 주최자가 직접 전시 및 작품 정보를 등록할 수 있는 관리자 페이지를 제공함으로써, 정보 접근성이 낮은 일반 관람객에게는 보다 깊이 있는 해설을 제공하고, 작가에게는 셀프 브랜딩과 홍보의 기회를 제공한다.

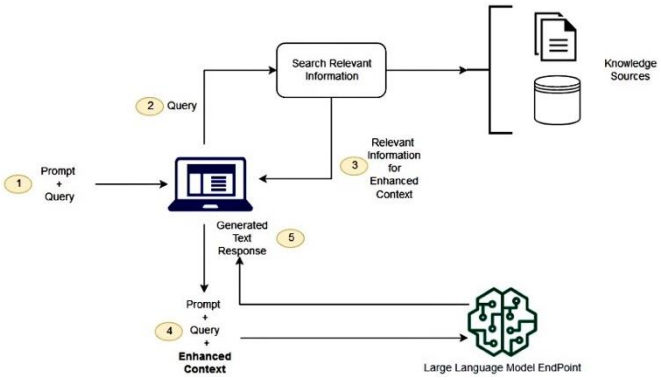
본 서비스의 주요 기능 구현에 쓰인 기술은 다음과 같다.

2.1 이미지 인식 기반 작품 식별

본 서비스는 사용자가 촬영한 작품을 자동으로 식별하기 위해 OpenAI 의 CLIP[3] 모델을 활용한다. CLIP 은 이미지와 텍스트를 동일한 임베딩 공간에 매핑하여 의미 유사도를 측정하며, 코사인 유사도를 기반으로 작품을 식별한다. 사전 등록된 이미지-텍스트 쌍은 임베딩을 거쳐 벡터 DB 에 저장되며, 촬영된 이미지와의 유사도를 비교해 작품을 식별한다.

2.2 RAG 기반 생성형 질의응답

기존 GPT 기반 언어 모델은 최신 전시나 작가 정보에 대한 비사실적 응답을 생성할 수 있다. 이를 해결하기 위해 본 논문에서는 아래 (그림 1)과 같이 RAG[4]를 도입하였다. 사용자의 질문은 벡터화되어 전시 정보가 저장된 벡터 DB 와 비교되며, 검색된 관련 문서를 GPT 의 입력 맥락으로 전달함으로써 최신성 · 정확성을 보장하는 응답을 생성한다.



(그림 1) RAG 기반 질의응답 시스템 구조도[5]

2.3 관리자 기능: 콘텐츠 등록

본 서비스의 관리자 페이지는 작가 및 전시 주최자가 AI 모델의 지식 베이스를 직접 제어할 수 있는 인터페이스를 제공한다. 사용자가 전시 해설 문서나 작품 이미지를 업로드하면 텍스트는 임베딩되어 Pinecone 벡터 DB 에, 추출된 이미지 벡터는 CLIP 을 통해 Faiss 인덱스에 각각 저장된다. 이를 통해 새로 추가된 전시 · 작품 정보가 RAG 검색과 이미지 인식 과정에 즉시 반영된다.

이처럼 관리자 페이지는 콘텐츠 제공자와 AI 시스템 간의 동적 데이터 파이프라인 역할을 수행하며, 해설 콘텐츠의 최신성·신뢰성을 확보하고, 작가에게는 셀프 브랜딩과 홍보 수단을 제공한다.

3. 세부 설계 및 구현

본 시스템은 Android 앱, Django 기반 서버, 관리자 웹 인터페이스, 외부 GPT API, 벡터 DB 로 구성된다. 사용자가 촬영한 이미지는 CLIP 모델을 통해 임베딩 벡터로 변환되며, 사전에 등록된 작품 벡터와의 코사인 유사도 계산을 통해 가장 유사한 작품 ID 로 매핑된다. 이렇게 식별된 작품 ID 는 RAG 모듈의 질의 입력으로 전달되며, 해당 작품과 관련된 문서를 벡터 DB(Pinecone)에서 검색한다. 검색된 문서는 GPT 모델의 입력 맥락으로 전달되어, 사용자의 질문에 맞춘 생성형 응답이 생성된다. 생성된 답변은 TTS 모듈을 통해 음성으로 변환되어 사용자에게 제공된다. 데이터는 RDS 와 MongoDB 에 저장되며, 서버는 AWS EC2 환경에서 배포 · 운영된다.

구현 코드는 GitHub 저장소(https://github.com/Artrip-docent)에서 확인할 수 있으며, 구현한 앱의 화면은 아래 (그림 2)와 같다.

(그림 2) 구현 결과(좌: 작품 촬영 화면, 우: 질의응답 화면)

4. 결론

본 논문에서는 기존 전시 안내 시스템의 한계를 보완하기 위해 이미지 인식과 생성형 AI 를 결합한 도슨트 서비스 Artrip 을 기획 · 설계 · 구현하였다. 제안된 서비스는 CLIP 기반 작품 식별을 통해 관람자가 손쉽게 작품 정보를 확인할 수 있도록 하고, RAG 구조를 적용한 질의응답 모듈을 통해 최신성과 정확성이 강화된 맞춤형 해설을 제공한다. 또한, 신진 작가 및 소규모 전시 주최자가 직접 전시 정보를 등록할 수 있는 페이지를 제공함으로써, 정보 격차를 완화하고 홍보 채널을 확대할 수 있음을 보였다.

본 연구의 의의는 Artrip 이 기존 오디오 가이드 시스템이 제공하지 못한 사용자 중심성, 상호작용성, 확장성을 실현했다는 점에 있다. 관람객은 작품과의 실시간 상호작용과 개인 맞춤형 해설을 통해 깊이 있는 문화예술 경험을 누릴 수 있으며, 작가는 관리자 기능과 리뷰 생태계를 통해 셀프 브랜딩과 피드백 기반의 창작 활동을 이어갈 수 있다. 향후 연구에서는 미술 도메인 특화 CLIP 모델 학습 및 인물 · 신화 주제 인식 향상을 위한 Fine-tuning 을 진행할 예정이다.

참고문헌

[1] 조소현, "국내 미술시장 성장의 동력 미술 전시 관람객 현황 분석", ASQUARE, 2024.

[2] 신승철, “미술관 권력과 오디오 가이드”, 서울아트가이드 Vol.209, 2019.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, “Learning Transferable Visual Models From Natural Language Supervision”, Proc. Int. Conf. on Machine Learning (ICML), 2021.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, Proc. NeurIPS, 2020.

[5] What is RAG (Retrieval-Augmented Generation)?, Amazon Web Services, 2024. [Online]. Available: "https://aws.amazon.com/ko/what-is/retrieval-augmented-generation/" (accessed Sep. 16, 2025)