스타일 특징 추출 기법 융합을 통한 미술 화풍 분류 RAG-LLM 시스템

오현찬, 김용화, 윤상필, 김영민
한국전자기술연구원

{hyeonchan.oh, yonghwa.kim, dkssyd000, rainmaker}@keti.re.kr

An RAG-LLM System for Art Style Classification through the Fusion of Style Feature Extraction Techniques

Hyeonchan Oh, Yonghwa Kim, Sangpil Yoon, Youngmin Kim
Korea Electronics Technology Institute(KETI)

요 약

본 논문은 특정 화가의 작품을 식별하고 분류하기 위한 시스템 구축을 위한 검색 증강 생성(RAG) 프레임워크를 활용한 작품 식별 및 식별된 작품의 선정 근거를 설명할 수 있는 시스템을 제안한다. 기존의 영상처리 기법은 이미지 및 영상이 가지는 의미론적 콘텐츠 정보 추출에 강력한 성능을 보여주었다. 그러나, 붓 터치, 질감, 색감, 구조 등의 작가의 정체성이라고 볼 수 있는 스타일적 특징(Stylistic features)을 추출 및 분류하기에는 어려움이 있다. 본 연구는 이를 극복하는 분류기 구성을 위하여 콘텐츠 및 스타일 특징 정보를 복합적으로 활용한 하이브리드 벡터 구조를 제안하였으며, 분류 근거를 설명하기 위한 거대언어모델(LLM)을 적용하여 최종적인 시스템 구조를 설계하였다. 최종적으로 기존의 이미지 특징 정보 추출 기법을 활용한 분류 방식보다 스타일 특징을 복합적으로 사용한 하이브리드 벡터 기법이 정확도(Accuracy)는 63.21%에서 77.23%로 약 14%p, F1-Score 는 53.65%에서 73.38%로 약 20%p 향상된 성능을 보이는 것을 확인하였다.

Ⅰ. 서 론

인공지능을 활용한 이미지 인식 및 분류 기법은 CLIP(Contrastive Language-Image Pre-Training)[1]과 같은 멀티모달 모델을 위한 정보 추출을 통하여 이미지와 텍스트를 함께 이해하는 방향으로 확대되고 있다. 또한, 거대언어모델(LLM)과 검색 증강 생성(RAG, Retrieval-Augmented Generation)과의 연계를 통한 데이터 분석 및 분류를 위한 방법이 대두되고 있다.

이 같은 검색 증강언어모델의 미술 작품 분석 및 분류 사용은 기존의 CLIP 과 같은 정보 추출로는 어려움이 있다. 기존의 CLIP 기반 시스템은 작품의 주제나 객체와 같은 콘텐츠 정보를 추출하는 데는 우수한 성능을 보이지만, 작가 고유의 화풍(artistic style)을 분류하기는 한계를 보인다. 이는 같은 콘텐츠를 작업하여도, 작가에 따라 화풍이 달라지는 미술 작품의 특성이 고려되지 않았기 때문이다.

따라서, 본 논문에서는 이러한 한계를 극복하고자 콘텐츠와 스타일 특징을 모두 고려하는 하이브리드 임베딩 RAG 시스템을 제안한다. 제안 모델은 CLIP 이미지 인코더 및 그람 행렬(Gram Matrix)을 활용하여 콘텐츠 정보와 스타일 특징 정보를 복합적으로 사용하며, 융합된 특징 정보를 기반으로 보다 정교한 화풍 분류 및 설명이 가능하도록 설계되었다.

Ⅱ. 본론

제안된 시스템은 입력된 이미지의 임베딩을 수행하는 하이브리드 임베딩 단계와 하이브리드 임베딩을 데이터베이스의 정보에서 검색하는 임베딩 검색 단계, 검색한 결과들을 분류하여 가장 적합한 결과를 도출하는 결과 투표 단계 그리고 최종 선정된 결과에 대한 근거를 도출하는 거대언어모델기반 분석 단계로 구성된다.

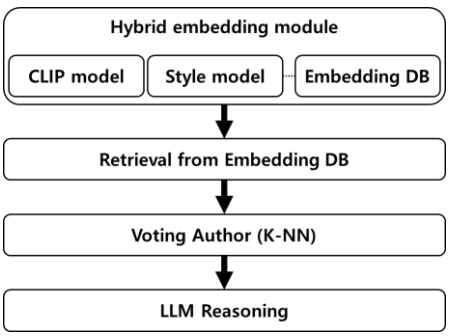

Figure 1. 제안 시스템의 전체 구조

하이브리드 임베딩 단계는 CLIP 인코더와 스타일 인코더를 기반으로 구성된다. CLIP 인코더는 총 768 차원의 벡터로 이미지가 내포한 작품의 주체 및 객체 정보를 담아낸다. 스타일 인코더는 VGG 기반의 특징맵(featuremap)으로부터 그람 행렬(Gram Matrix)[3]를 구성한다. 그람 행렬은 이미지의 질감, 색상, 패턴, 붓 터치와 같은 시각적 스타일 정보를 효과적으로 표현하기 위한 방법[3]으로 완전 연결

계층(Fully Connected Layer)를 이용하여 128 차원의 벡터로 표현된 스타일 정보로 변환된다. 완전연결계층을 통한 128 차원의 스타일 정보를 학습하기 위하여 Triplet loss 를 적용하여, 동일 작가의 작품과 그 외의 작가의 작품 분류를 위한 학습을 진행하였다. 이후, 두 벡터를 융합하여 896 차원의 하이브리드 임베딩을 구성하며, 이는 임베딩 데이터베이스의 구축에 활용된다.

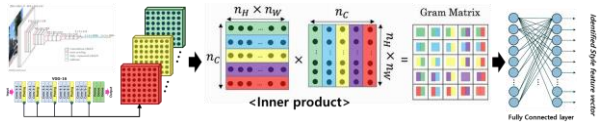

Figure 2. 그람 행렬 기반 스타일 인코더

검색 단계에서는 임베딩 기반 데이터베이스에서 가장 유사한 화풍을 검색하며, 단순히 최적 결과만을 도출하는 것이 아닌 상위 20 개의 정보를 검색한다. 검색된 상위 20 개 결과는 K-최근접 이웃(k-Nearest Neighbors) 분류 알고리즘 기반의 투표 시스템을 통하여, 가장 적합한 화풍 정보를 검출한다.

선정된 화풍 정보를 거대언어모델에 전달하여, 최종적으로 해당 화풍이 선정된 근거 도출 및 자연어 형태로 생성한다. 생성된 설명은 모델이 주목한 시각적 특징과 결과 상의 인과관계를 보여준다.

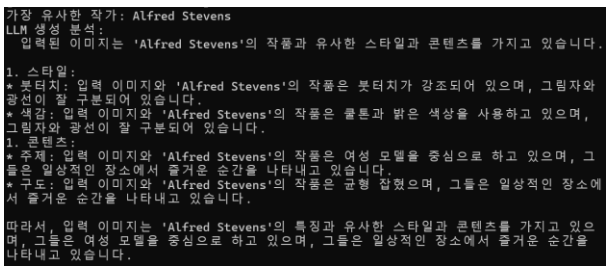

Figure 3. 유사 작가 분류 근거 설명

실험은 퍼블릭 도메인 작품(Public Domain Works)을 기반으로 구성된 총 27 명의 주요 화가(빈센트 반 고흐, 클로드 모네, 구스타프 클림트 등)의 작품 이미지로 구성된 데이터셋을 대상으로 수행하였다.

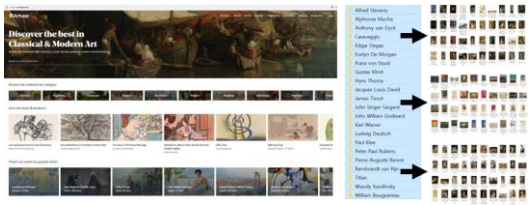

Figure 4. Public domain works from Web[2]

성능 평가를 위해 약 100 여 장의 이미지를 무작위로 추출하여 CLIP 단독 모델과 제안된 하이브리드 모델을 비교하였다. 성능 평가는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score 의 네 가지 지표를 기준으로 수행하였다.

CLIP 단독 모델은 정확도 63.21%, 정밀도 55.28%, 재현율 62.35%, F1-Score 53.65%를 기록하였다. 반면 하이브리드 모델은 정확도 77.23%, 정밀도 78.08%, 재현율 76.23%, F1-Score 73.38%로 모든 지표에서 CLIP 단독 모델을 크게 능가하였다. 특히 종합 성능을 나타내는 F1-Score 는 약 20%p 향상되었는데, 이는 스타일 정보를 추가함으로써 모델의 변별력이 크게 향상되었음을 보여준다. 또한 기존의 CLIP 이미지 인코더 기반의 분석 모델이 종종 유사한 주제를 그린 다른 화가의 작품을 혼동한 반면, 스타일적 특징(Stylistic features)를 복합적으로 사용한 하이브리드 모델은 이러한 오류를 상당 부분 보완하는 것을 확인하였다.

Table 1. CLIP 및 하이브리드 모델의 분류 성능

모델 구분	CLIP 단독	하이브리드	성능향상
정확도	63.21%	77.23%	22.18% (+14.02%p)
정밀도	55.28%	78.08%	41.24% (+22.80%p)
재현율	62.35%	76.23%	22.26% (+13.88%p)
F1-Score	53.65%	73.38%	36.78% (+19.73%p)

Ⅲ. 결론

본 논문에서는 CLIP 기반 화풍 분류의 한계를 극복하기 위해 스타일 특징 추출기를 결합한 하이브리드 임베딩 기반 RAG-LLM 시스템을 제안하였다. 실험 결과, 제안 모델은 CLIP 단독 모델 대비 모든 성능 지표에서 크게 향상된 성능을 보였으며, 특히 스타일 정보를 반영한 접근법이 작품의 화풍 분류에 효과적임을 확인하였다. 이 연구는 미술사 연구, 예술 교육 등 다양한 분야에서 응용될 수 있으며, 향후 Transformer 기반의 스타일 추출기를 적용 및 LoRA 과 같은 LLM 의 정밀 조정(Fine-tuning)의 추가적인 연구를 수행할 예정이다. 또한 하이브리드 임베딩 개념은 미술 작품뿐만 아니라 패션, 디자인, 건축 등 스타일이 중요한 다른 도메인에도 확장 가능성이 크다.

ACKNOWLEDGMENT

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024 년도 문화기술 연구개발 사업으로 수행되었음 (과제명: 멀티모달 생성형 AI 모델의 데이터셋 저작권 핵심 기술 개발, 과제번호: RS-2024-00333068, 기여율: 100%)

참 고 문 헌

[1] A. Radford et al., “Learning transferable visual models from natural language supervision,” arXiv:2103.00020, 2021.

[2] Artvee, “Artvee | Discover Classical Art,” https://artvee.com/, 2025.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423, doi:10.1109/CVPR.2016.265.