다국어 질의-이미지 의미 정합성을 고려한 벡터 기반 이미지 검색 파이프라인 설계

장현준, 유재천*
성균관대학교 전자전기컴퓨터공학과

10hour@skku.edu, *yoojc@skku.edu

Design of a Vector-Based Image Retrieval Pipeline Considering Multilingual Query-Image Semantic Alignment

Jang Hyunjun, Yoo Jae-Chern*
Dept. of Electrical and Computer Engineering, Sungkyunkwan Univ.

요 약

스마트폰 보급과 SNS 확산으로 개인이 저장하는 사진은 수년 새 폭발적으로 증가했지만, 대부분의 갤러리 앱은 여전히 해시태그나 키워드 기반 필터에 의존하고 있다. 본 연구는 이러한 한계를 극복하기 위해, 이미지 캡셔닝 → 한국어 번역 → 문장 임베딩 및 검색의 세 단계를 하나의 흐름으로 결합한 파이프라인을 제안한다.

실험은 총 400 장의 개인 사진을 대상으로 진행되었으며, 기본적인 한국어 질의 이외에 영어·일본어 등 다국어 질의에 대해서도 높은 검색 성능을 확인하였다. 제안 시스템은 언어 장벽, 키워드 누락, 다중 객체 및 복합 행위 표현의 한계를 효과적으로 보완하며, 디지털 자산 관리 및 접근성 보조 기술로의 확장이 가능함을 보여준다.

Ⅰ. 서 론

스마트폰 카메라 성능이 고해상도·고주사율로 비약적으로 향상되고, 이미지 기반 소셜 네트워크(Instagram 등)가 일상의 기록 도구로 자리 잡으면서 개인이 보유한 사진은 한 해에 수만 장을 넘어선다. 그럼에도 사용자가 사진을 다시 꺼내 볼 때 의존할 수 있는 정보는 촬영 시 입력한 키워드나 해시태그처럼 제한적인 메타데이터에 머무는 경우가 많다. 이러한 방식은 ① 사용자가 장면에 등장하는 객체의 정확한 학명이나 브랜드명을 기억하지 못할 때, ② “아이에게 밥을 먹이는 장면”처럼 다중 객체와 행위가 얽힌 상황을 기술해야 할 때, 혹은 ③ 객체 탐지 모델이 특정 사물을 오인식, 미탐지할 때 치명적인 검색 누락을 초래한다. 결과적으로 사용자는 비슷한 키워드 조합을 반복해서 시도하거나 전체 썸네일을 수작업으로 훑어야 하는 불편을 감수한다.

이와 같이 사회·문화적으로는 이미지 기반 커뮤니케이션이 보편화되었고, 기술적으로는 멀티모달 인공지능이 급속히 발전했지만, 정작 개인 사용자의 사진 탐색 경험은 키워드 검색 패러다임에서 크게 벗어나지 못하고 있는 실정이다. 따라서 사진 속 시각 정보를 풍부한 자연어로 전환하고, 사용자가 친숙한 언어로 자유롭게 질문할 수 있는 고차원 검색 체계가 요구된다. 본 연구는 이러한 요구를 충족하기 위해, 최신 멀티모달 모델과 다국어 임베딩·벡터 검색 기술을 단계적으로 결합한 통합 파이프라인을 설계·검증하고자 한다.

Ⅱ. 본 론

2.1 벡터 간 관계와 의미 공간

문장을 임베딩 모델 f:S→R 로 투사하면, 의미가 유사한 두 문장 s1, s2 는 벡터 공간에서 코사인 유사도

$$\mathrm{sim}(s_1, s_2) = \frac{f(s_1) \cdot f(s_2)}{\|f(s_1)\| \|f(s_2)\|}$$

가 1 에 가까워진다. 모든 벡터를 L2-정규화하면 내적(inner-product)이 곧 코사인 유사도와 같아지므로, FAISS IndexFlatIP 구조로 최근접 탐색 시 높은 효율성을 얻을 수 있다. 이 벡터 공간은 다국어 동시 학습으로 구축되었기 때문에, 동일 의미의 한국어·영어 문장은 자연스럽게 인접 위치에 매핑된다.

2.2 단계별 모델 및 처리 과정

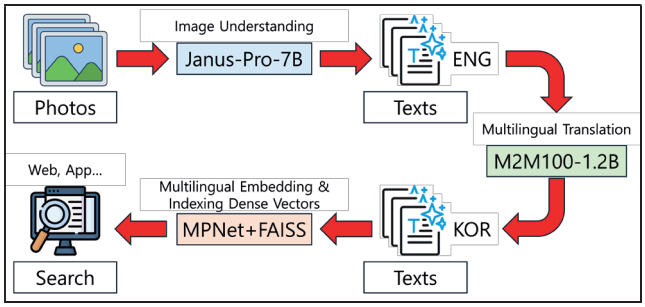
(1) 이미지 이해 — Janus-Pro-7B

Janus-Pro-7B 는 DeepSeek-AI 가 제안한 Unified Multimodal Understanding & Generation 모델로, 시각 인코더를 이해용(SigLIP-Large-Patch16-384)과 생성용(VQ-GAN 토크나이저, codebook = 16 384)으로 분리하여 두 과제 간 표현 충돌을 완화한다. 두 인코더에서 나온 시퀀스는 7 B 파라미터의 자가회귀 Transformer 에 입력되며, 이해 작업에서는 최대 384 × 384 해상도 이미지를 한 문장으로 기술한다.

예: “A child eating rice in a sun-lit kitchen.”

(2) 다국어 번역 — M2M100-1.2B
Facebook AI 의 M2M100-1.2B 는 100 개 언어 간 direct many-to-many 번역을 지원하는 12 억 파라미터 인코더-디코더 모델이다. 9900 개에 달하는 언어 방향(Language Direction)을 별도 중간 언어 없이 직접 번역하며, 영어→한국어에서도 BLEU·COMET 지표가 기존 영어 중심 모델보다 우수하다.
(1)의 예문은 아래와 같이 자연스럽게 변환된다.
예: “햇빛이 들어오는 주방에서 밥을 먹는 아이.”

(3) 임베딩·검색 — MPNet + FAISS
해당 MPNet 모델(paraphrase-multilingual-mpnet-base-v2)은 50+ 개 언어 병렬 말뭉치로 학습된 768 차원 문장 임베더로, 문장·단락을 동일 의미 공간으로 매핑한다. 모든 벡터를 정규화 후 FAISS IndexFlatIP 에 저장하면 코사인 유사도에 상응하는 내적값을 O(d×N) 연산으로 최근접 벡터를 수 ms 내 반환한다.



(그림 1) 제안하는 이미지 검색 파이프라인의 전체 구조

2.3 실험 결과
우선, 기존 파이프라인 위에 웹 기반 인터랙티브 UI 를 얹어 사용자 친화적인 검색 환경을 구현하였다. 입력 창에 한글자씩 타이핑할 때마다 백엔드에서 임베딩·FAISS 탐색이 즉시 수행되고, 클라이언트 측 그리드 갤러리에 결과 썸네일이 실시간으로 교체되도록 설계하였다.
이를 통해 서론에서 제기한 ‘모호한 객체 명칭·행위 중심·다국어 표현’ 같은 키워드 기반 검색의 한계를 실제로 극복할 수 있는지를 검증하고자 했다. 구체적으로는 ① 명칭이 모호한 캐릭터 상품을 질의한 경우, ② 행위를 중심으로 한 문장을 입력한 경우, ③ 한국어·영어·일본어 다국어 질의를 입력한 경우의 세 시나리오를 설정해 테스트를 진행하였다.

(그림 2) “사랑스러운 동물 인형”이라는 모호한 객체를 포함한 질의어에 대한 이미지 검색 결과.

(그림 3) “걸어가는 중”이라는 행위를 묘사한 질의어에 대한 이미지 검색 결과.


(그림 4) “탄수화물이 풍부한 음식”이라는 의미의 질의어를 한국어(좌), 영어(중), 일본어(우)로 입력한 검색 결과.

Ⅲ. 결 론

테스트 결과, 세 시나리오 모두에서 상위 30 개 이미지가 평균 0.8 초 이내에 렌더링되었으며, 검색 품질 또한 대체로 만족스러웠다. 특히 다국어 질의의 경우, 동일 대상을 서로 다른 언어로 표현했을 때 일부 랭킹 차이는 있었으나 의미가 유사한 결과가 안정적으로 반환되었다. 다만, 문장 임베딩 과정에서 핵심 단어에 대한 중요도를 반영하지 못해, 질의 의미가 다소 분산되는 한계도 확인되었다.
그럼에도 불구하고, 제안된 시스템은 기존 키워드 중심의 단편적 검색 방식에서 벗어나, 이미지의 맥락적 의미와 사용자 질의의 자연어 표현을 효과적으로 연결함으로써 높은 실용성과 확장 가능성을 입증하였다. 향후 연구에서는 클라이언트 단 추론을 위한 경량화 기법과 토큰 수준의 가중치 제어 기법을 통해 이러한 한계들을 보완하고, 개인화된 검색 경험으로 발전시킬 수 있을 것으로 기대된다.

참 고 문 헌

[1] Chen, Xiaokang, et al. "Janus-pro: Unified multimodal understanding and generation with data and model scaling." *arXiv preprint arXiv:2501.17811* (2025).

[2] Fan, Angela, et al. "Beyond english-centric multilingual machine translation." *Journal of Machine Learning Research* 22.107 (2021): 1-48.

[3] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).

[4] Song, Kaitao, et al. "Mpnet: Masked and permuted pre-training for language understanding." *Advances in neural information processing systems* 33 (2020): 16857-16867.