멀티모달 임베딩과 Knowledge Graph 기반 하이브리드 검색 시스템 설계 및 구현

권병헌, 박지은, *신병철

인텔렉투스

bh.kwon@int2.us, jieun.park@int2.us *bc.shin@int2.us

A Hybrid Search Framework Integrating Multimodal Embeddings and Knowledge Graphs: Design and Implementation

Byeongheon Kwon, Jieun Park, *Byungchul Shin

Intellectus Corp.

요 약

최근 데이터 활용은 키워드 검색을 넘어 의미적 연관성과 맥락을 반영하는 시맨틱 검색으로 발전하고 있다. 그러나 기존 Knowledge Graph(KG) 기반 검색은 자연어 질의와의 정합성이 낮고, 벡터 임베딩 기반 검색은 개체 간 관계 표현에 한계가 있으며, CLIP·BLIP과 같은 멀티모달 모델도 복잡한 맥락 반영에는 부족하다.

본 연구는 이러한 한계를 극복하기 위해 CLIP, BLIP, SentenceTransformer를 결합하고, Knowledge Graph 기반 그래프 요약 및 구조적 필터링을 통합한 하이브리드 검색 시스템을 제안한다. 제안 시스템은 텍스트-이미지 매칭, 자동 캡션 기반 의미 보강, 그래프 요약 임베딩을 결합하여 장면의 구조적 맥락까지 반영한다.

Ⅰ. 서 론

최근 데이터 활용 패러다임은 단순 키워드 검색을 넘어, 의미적 연관성을 파악하고 복잡한 맥락을 반영할 수 있는 시맨틱 검색으로 발전하고 있다. OECD는 공공데이터 활용의 성숙도를 높이기 위해 데이터 기반 행정 전환의 중요성을 강조하며, 데이터의 품질과 활용성 제고가 핵심 과제임을 지적하였다 [1][2].

이와 같은 흐름 속에서 Knowledge Graph(KG)는 개체와 관계를 명시적으로 표현하여 시맨틱 검색을 지원하는 핵심 기술로 주목받고 있으며, 추천 시스템을 포함한 다양한 분야에서 성능과 해석 가능성을 높이는 것으로 보고되었다 [3]. 또한 KG의 불완전성과 노이즈를 개선하기 위한 정제(refinement) 연구도 활발히 이루어지고 있다 [4]. 한편, 벡터 검색(Vector Search)은 대규모 사전학습 언어모델의 발전과 함께 부상하여, Sentence-BERT [5]와 SimCSE [6]는 문장을 고차원 벡터로 임베딩해 유사도 기반 검색 정확도를 크게 향상시켰다. 최근에는 이미지와 텍스트를 동시에 임베딩하는 멀티모달 접근이 활발히 연구되고 있으며, CLIP은 이미지와 자연어를 동일한 잠재 공간에서 학습하여 이미지-텍스트 검색의 가능성을 확장하였고 [7], BLIP은 이미지 캡션 생성을 통해 텍스트와 시각 정보의 연결성을 강화하였다 [8].

그러나 이러한 기술적 진보에도 불구하고 여전히 현실에는 몇 가지 한계가 존재한다. 첫째, KG 기반 검색은 구조적 관계를 잘 반영하지만 자연어 질의와의 정합성이 낮다. 반대로 벡터 검색은 의미적 유사도 계산에는 강점이 있으나 개체 간 구체적 관계를 드러내지 못한다. 둘째, CLIP과 BLIP 같은 멀티모달 모델은 이미지-텍스트 검색 성능을 개선하였으나, 장면 내 복잡한 객체 관계나 맥락적 상황(예: 환경 조건, 주행 맥락)을 충분히 반영하기 어렵다. 셋째, 실제 응용 환경(자율주행, 스마트 시티 관제 등)에서는 단순히 “차량”이나 “보행자”를 찾는 수준을 넘어, “비 오는 교차로에서 차량이 보행자와 근접해 있는 상황”과 같은 복합 조건 검색이 요구되지만 기존 시스템은 이를 효과적으로 처리하지 못한다.

본 논문에서는 멀티모달 임베딩(CLIP(Contrastive Language-Image Pre-training), BLIP(Bootstrapping Language-Image Pre-training), SentenceTransformer)과 Knowledge Graph 기반 요약을 결합한 하이브리드 검색 시스템을 설계 및 구현하는 것이며, 이를 통해 텍스트-이미지-그래프 간 검색의 정밀도와 유연성을 향상시키고자 한다.

Ⅱ. 본론

2.1 CLIP(Contrastive Language-Image Pre-training)

CLIP은 이미지와 자연어를 동일한 공간으로 임베딩하기 위해 대규모 이미지-텍스트 쌍 데이터를 활용하여 학습된 멀티모달 모델이다. 이미지 인코더(ViT-L/14)와 텍스트 인코더를 통해 서로 다른 모달리티를 벡터로 변환한 후, 코사인 유사도 기반의 대조학습(contrastive learning)을 수행한다.

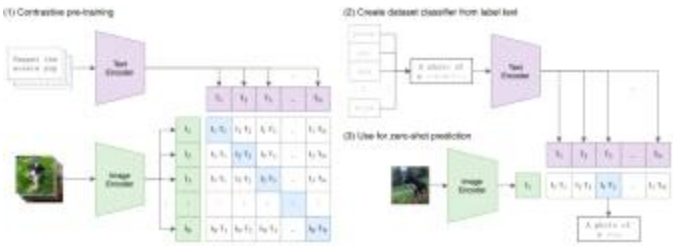

Fig. 1 CLIP 구조

2.2 BLIP(Boostrapping Language-Image Pre-training)

BLIP은 이미지와 언어를 동시에 이해하고 생성할 수 있는 멀티모달 사전학습 모델로, 특히 이미지 캡션 생성과 질의응답 태스크에서 우수한 성능을 보인다. 본 연구에서는 BLIP을 이용하여 이미지로부터 자동으로 캡션을 생성하고, 이를 CLIP 텍스트 인코더를 통해 임베딩하여 이미지 표현을 보강하였다. 이는 단순히 이미지 픽셀 특징을 임베딩하는 것보다 풍부한 의미 정보를 제공하여 검색 랭킹 산출의 성능을 높이는 데 기여한다.

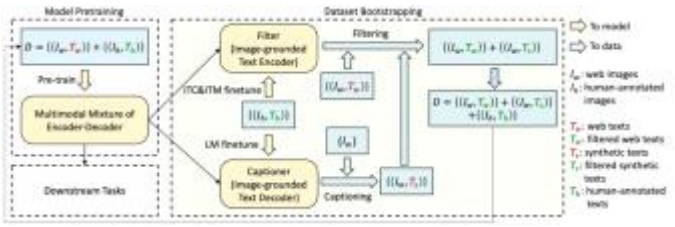

Fig. 2 BLIP 구조

2.3 ST(SentenceTransformer)

SentenceTransformer는 문장을 고차원 벡터로 변환하여 의미적 유사도를 계산할 수 있도록 하는 모델로, 본 연구에서는 intfloat/e5-base-v2를 적용하였다. 특히 Knowledge Graph 기반 장면 요약문(graph summary)을 입력으로 하여 벡터화함으로써 장면의 구조적 맥락을 반영할 수 있도록 하였다.

2.4 임베딩 가중합 및 랭킹 산출

본 연구에서 활용된 CLIP, BLIP, SentenceTransformer 세 가지 임베딩을 단일 벡터로 결합하기 위해 가중합 방식(weighted sum)을 적용하였으며, 각 임베딩의 유사도 점수를 S_{clip},S_{blip}, S_{st}, 가중치를 w_{dip}, w_{blip}, w_{st}라 할 때, 최종 검색 점수 S_{tinal}은 다음과 같다.

$$S_{final} = w_{dip} \cdot S_{dip} + w_{blip} \cdot S_{blip} + w_{st} \cdot S_{st}, (w_{dip} + w_{blip} + w_{st} = 1)$$

2.5 검색 성능 테스트

KITTI Dataset의 10개 대표 질의를 대상으로, CLIP, BLIP, SentenceTransformer를 활용한 다양한 가중치 조합에 대해 검색 성능을 평가하였다. 각 질의에 대해 텍스트-이미지 간 유사도 점수를 산출하고, 가중합 방식으로 최종 Score를 계산하였다.

그 결과, CLIP 중심 조합은 평균 0.681로 baseline 수준의 성능을 보였다. 그러나 SentenceTransformer 기반 그래프 요약 임베딩을 결합할 경우 평균 Score가 0.720으로 향상되었으며, SentenceTransformer의 비중을 더욱 높인 조합에서는 평균 0.767까지 상승하였다.

CLIP Image	**0.1**	**0.5**	**0.9**
CLIP Text	**0.17**	**0.3**	**0.05**
Sentence Transformer	**0.73**	**0.2**	**0.05**
A road with cars parked on both sides	0.782	0.667	0.664
A pedestrian crossing the street in front of a car	0.762	0.699	0.655
A cyclist riding along the right side of the road	0.776	0.684	0.65
A car driving straight on a two-lane road	0.766	0.689	0.661
Traffic lights at an intersection ahead	0.764	0.686	0.649
A pedestrian standing near the sidewalk	0.74	0.653	0.636
A few cars waiting at a traffic light	0.782	0.684	0.641
Buildings and shops along the street with parked cars	0.769	0.685	0.647
A car turning left at an intersection	0.767	0.683	0.651
A pedestrian and a cyclist crossing the same intersection	0.765	0.678	0.648

Tabel. 1 모델 가중에 따른 검색 유사도 결과

Ⅲ. 결론

본 연구에서는 CLIP, BLIP, SentenceTransformer(ST)를 활용한 하이브리드 검색 시스템을 제안하고, KITTI Dataset을 통해 성능을 검증하였다. 실험 결과, SentenceTransformer 단독 임베딩을 활용했을 때 평균 Score가 가장 높게 나타나, 그래프 요약 기반 표현이 검색 성능 향상에 크게 기여함을 확인하였다.

그러나 ST 단독 접근은 이미지 자체의 시각적 특징을 반영하지 못하고, 객체 탐지 및 그래프 요약 품질에 크게 의존한다는 한계가 있다. 반면, CLIP은 이미지-텍스트 간 직접적인 매칭을 지원하여 멀티모달 검색의 기본 축을 형성하며, BLIP은 이미지 캡션을 자동 생성함으로써 텍스트 임베딩과의 연결성을 강화한다. 따라서 세 가지 임베딩을 결합했을 때, 단일 모델 접근보다 다양한 질의 유형(텍스트-이미지 매칭, 캡션 기반 검색, 조건 기반 검색)에 안정적이고 범용적으로 대응할 수 있다.

즉, 성능 수치만 보면 ST가 우세하더라도, CLIP + BLIP + ST 결합은 실제 응용 환경에서 요구되는 복합 질의 처리와 신뢰성 측면에서 더 타당한 접근이다. 향후 연구에서는 더 다양한 데이터셋(nuScenes, Waymo Open Dataset 등)에서의 일반화 성능 검증과, Neo4j-FAISS 기반의 실시간 검색 시스템 구현을 통해 제안 방법의 실효성을 확대할 예정이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정부통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2023-00235293, 활용 목적에 따른 자율주행 데이터 제공을 위한 자율주행 빅데이터 가공/관리, 검색 및 공유 인터페이스 기술 개발)

참 고 문 헌

[1] OECD, Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact, OECD Publishing, 2019.

[2] OECD, The Path to Becoming a Data-Driven Public Sector, OECD Publishing, 2021.

[3] H. Wang, F. Zhang, X. Xie, and M. Guo, "A comprehensive survey of knowledge graph-based recommender systems," Information, vol. 12, no. 6, pp. 232, Jun. 2021.

[4] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," Semantic Web, vol. 8, no. 3, pp. 489–508, 2017.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proceedings of EMNLP-IJCNLP, pp. 3982–3992, 2019.

[6] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," Proceedings of EMNLP, pp. 6894–6910, 2021.

[7] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," Proceedings of ICML, pp. 8748–8763, 2021.

[8] J. Li, D. Li, C. Xiong, and S. C. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," Proceedings of ICML, pp. 12888–12900, 2022.