Journal of KIIT. Vol. 23, No. 8, pp. 37-44, Aug. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.8.37

비주얼 텍스트 검색을 위한 글로벌-로컬 다중 벡터 융합

배용진*, 배경만**

Global-Local Multi-Vector Fusion for Visual Text Retrieval

Yongjin Bae*, Kyoungman Bae**

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-II220369, (4세부) 전문지식 대상 판단결과의 이유/근거를 설명가능한 전문가 의사결정 지원 인공지능 기술개발)

요 약

멀티모달 정보의 활용이 확대되면서, 텍스트와 이미지가 혼합된 복합 문서에서 정확한 정보 검색을 수행할 수 있는 모델에 대한 수요가 점차 증가하고 있다. 본 논문에서는 멀티벡터 기반 이미지-텍스트 표현 구조를 확장하여, 시퀀스 전체의 문맥 정보를 효과적으로 반영하는 WGF(Weighted Global Fusion) 방법을 제안한다. 기존 표현력을 유지하면서 시퀀스 문맥 정보를 효과적으로 반영한 점이 본 연구의 주요 특징이다. 제안하는 방법은 self-attention을 통해 시퀀스 내 각 토큰의 상대적 중요도를 계산하고, 이를 기반으로 전체 시퀀스를 대표하는 글로벌 임베딩을 생성한 뒤, 기존 토큰 임베딩과 선형적으로 융합하였다. 실제 업무 환경에서 수집한 멀티모달 문서 컬렉션을 기반으로 실험한 결과, 베이스라인 대비 Top-1 정확도가 약 25% 향상되었다.

Abstract

As the use of multimodal information continues to expand, there is a growing demand for models capable of accurately retrieving information from complex documents that combine text and images. This paper proposes a Weighted Global Fusion (WGF) method that extends a multi-vector image-text representation structure to effectively incorporate sequence-level contextual information. A key contribution of this work is its ability to preserve fine-grained token-level representations while integrating global semantic context. The proposed approach computes the relative importance of each token via self-attention, constructs a global embedding that summarizes the entire sequence, and linearly fuses it with the original token embeddings. Experimental results on a multimodal document collection sourced from real-world office environments show that the proposed method improves Top-1 accuracy by approximately 25% compared to the baseline model.

Keywords

information retrieval, deep learning, multimodal retrieval, question answering system

* 한국전자통신연구원 선임연구원(교신저자)
- ORCID: http://orcid.org/0000-0002-0227-8933
** 한국전자통신연구원 책임연구원
- ORCID: http://orcid.org/0000-0001-9007-4027

· Received: Jun. 17, 2025, Revised: Jul. 15, 2025, Accepted: Jul. 18, 2025
· Corresponding Author: Yongjin Bae
Language Intelligence Research Section, ETRI, Yuseong-gu, Daejeon, Republic of Korea
Tel.: +82-42-860-6879, Email: yongjin@etri.re.kr

Ⅰ. 서 론

최근 디지털 문서의 다양화와 함께, 텍스트와 이미지가 혼합된 복합 문서에서의 정보 검색에 대한 수요가 지속적으로 증가하고 있다. 특히 스캔 된 PDF, 발표 슬라이드, 표와 이미지가 포함된 보고서 등은 다양한 시각적 요소와 언어 정보를 동시에 포함하고 있으며, 이들 간의 복합적인 의미를 이해하고 검색 결과에 반영하는 것이 실제 업무 환경에서도 중요한 기술적 과제로 부상하고 있다. 이러한 멀티모달 검색 환경에서 핵심적인 구성 요소는 입력 정보를 효과적으로 표현할 수 있는 임베딩 구조이며, 그 표현력이 검색 성능을 좌우한다.

기존 멀티모달 검색 모델들은 주로 입력 시퀀스를 단일 임베딩 벡터로 요약하거나, 모든 토큰 임베딩을 동등하게 취급하는 구조를 사용해 왔다. 이와 같은 단순 멀티 벡터 기반 표현 방식은 토큰 단위의 세밀한 정보를 보존할 수 있다는 장점이 있으나, 각 토큰의 상대적 중요도를 반영하지 못하고, 시퀀스 전체의 문맥을 종합한 글로벌 표현(Global representation)이 부재하다는 한계를 가진다. 특히 문장 수준의 의미 이해나 중요 정보의 시각적 요소를 강조해야 하는 멀티모달 환경에서 이러한 구조적 제약이 검색 성능을 저해하는 요인으로 작용할 수 있다.

본 연구에서는 이러한 문제점을 해결하기 위해, 멀티벡터 기반 구조를 유지하면서도 글로벌 문맥 정보를 효과적으로 통합할 수 있는 구조를 제안한다. 이 방식은 로컬 표현의 정밀함을 유지하면서도 시퀀스 전반의 의미를 보완할 수 있도록 설계되었으며, 기존 구조에 비해 파라미터 추가 없이 간단하게 적용 가능하다는 장점을 가진다.

실험 결과, 제안한 방법은 기존 베이스라인에 비해 Top-1 정확도가 약 25% 이상 향상되었으며, 다른 비교 모델들과의 성능 비교에서도 가장 높은 정확도를 나타냈다. 특히 높은 정확도와 낮은 검색 실패율을 동시에 달성함으로써, 정밀성과 신뢰성을 모두 갖춘 멀티모달 검색 모델로서의 실용적 가능성을 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 간략히 소개하고, 3장에서는 제안하는 모델 구조와 글로벌 임베딩 융합 방법에 관해 설명한다. 4장에서는 데이터 셋 구성, 실험 환경 및 성능 비교를 통해 모델의 우수성을 검증하며, 5장에서는 본 논문의 결론을 제시한다.

Ⅱ. 관련 연구

2.1 전통적인 문서 검색 및 OCR 기반 접근

기존의 문서 검색 시스템은 대부분 텍스트 기반의 색인 및 검색 파이프라인에 의존하였다. 이들 시스템은 일반적으로 PDF, 슬라이드, 스캔 이미지와 같은 문서에서 OCR(Optical Character Recognition)을 통해 텍스트를 추출한 후 벡터 공간 모델, BM25, 혹은 Transformers 기반 임베딩 모델 등을 사용하여 검색을 수행하였다. 그러나 이러한 방식은 문서에 포함된 시각적 요소나 레이아웃 정보를 반영하지 못하며, OCR 품질에 심각하게 의존한다. 특히 저해상도 이미지, 복잡한 배치, 다양한 폰트가 혼재된 문서에서는 OCR 오류율이 50%를 초과하는 경우도 보고[1]되었다. 또한, 이러한 텍스트 중심의 접근은 표, 그림, 도식 등 중요한 시각적 요소가 포함된 복합 문서에 대한 질의 처리 능력이 제한적이며, 시각적 단서에 기반한 적합성 판단이 불가능하다는 근본적인 제약을 가진다[2][3].

2.2 비전-언어 기반 검색 모델

최근에는 이러한 한계를 극복하기 위해 이미지와 텍스트를 통합적으로 이해할 수 있는 비전-언어 모델(VLM, Vision-Language Model) 기반의 검색 기법이 활발히 연구되고 있다. 대표적인 선행 모델인 CLIP(Contrastive Language-Image Pre-training)[4]은 대규모 이미지-텍스트 쌍을 활용하여 텍스트와 이미지를 동일한 임베딩 공간으로 매핑하는 대조 학습 기반 구조를 제안하였다. 이 방식은 실세계의 사진이나 그림 이미지를 중심으로 학습되었음에도 불구하고, 다양한 크로스모달 검색 과제에서 강력한 성능을 보였다. 이후에는 다양한 도메인에 맞게 확장되었으며, 문서 이미지와 같은 복합 시각 데이터를

효과적으로 처리하기 위한 연구도 활발히 진행되었다. 대표적인 예로, PaLI[5] 모델은 다국어 비전-언어 모델로서, 텍스트 인식과 이미지 이해를 동시에 수행할 수 있는 능력을 갖추고 있다. PaLI는 복잡한 문서 레이아웃에서도 강건한 성능을 보이며, OCR을 명시적으로 수행하지 않고도 이미지 내 텍스트 의미를 파악할 수 있다. 또한, XDoc 모델[6]은 대규모 문서 이미지와 텍스트 쌍 데이터를 이용한 사전 학습을 통해 문서 인식 및 질의 응답에 특화된 성능을 보여주고 있으며, LiT 모델[7]은 사전 학습된 이미지 백본과 언어 모델을 연결하여 zero-shot 검색에서도 우수한 성능을 보였다.

2.3 비주얼 문서 검색 특화 모델

복잡한 문서의 시각 구조를 더욱 세밀하게 반영하기 위한 연구로는 ColPali[8]가 있다. ColPali는 PaLI 기반의 VLM에 지연 상호작용(late interaction) 방법을 도입하여, 문서 이미지를 고정된 벡터가 아닌 지역 단위 임베딩으로 표현하며, 쿼리 임베딩과의 세밀한 다중 매칭을 수행한다. 이는 ColBERT[9] 방법에서 영감을 받은 구조로, 쿼리의 각 토큰 임베딩이 이미지 내의 개별 위치 임베딩들과 대응되는 구조를 통해 정밀한 검색을 가능하게 한다. 이 모델은 특히 OCR 기반 파이프라인 대비 높은 정확도와 낮은 응답시간을 모두 확보한 것으로 보고되었으며, 문서 이미지의 전처리 없이도 강건한 검색 성능을 제공한다. 또한, VisRAG[10]는 기존 텍스트 기반 RAG 시스템을 대체하는 비전 기반 RAG 구조를 제안하였다. 이 구조에서는 문서 전체를 이미지로 인코딩하고, 쿼리와의 의미적 유사성을 기반으로 관련 이미지를 검색한 후, VLM 기반 언어모델이 이를 기반으로 답변을 생성한다. VisRAG는 문서 레이아웃, 시각적 강조, 표와 같은 구성 요소를 그대로 보존하면서 질의응답 성능을 20~40% 향상시켰다. VDocRAG[11]는 문서 이미지를 기반으로 검색 및 응답하는 두 가지 주요 구성 요소를 제안하였고, 자기지도 학습에 기반하여 시각 정보와 텍스트 내용을 정렬하는 학습방법을 제안하였다. MDocAgent[12]는 기존의 멀티모달 모델은 텍스트나 이미지 중 하나의 모달리티에 초점을 맞추는 경향을 개선하기 위해 다양한 다양한 에이전트를 구성하여 협력적 추론 매커니즘을 제안하였다.

기존 연구들은 멀티모달 임베딩의 표현력을 향상시키기 위해 다양한 구조적 접근을 시도하였으나, 여전히 토큰 수준의 정밀한 정보와 시퀀스 전체의 문맥 정보를 동시에 통합하는 데에는 한계가 존재한다. 예를 들어, ColPali는 지역 단위 임베딩과 쿼리-문서 간 정밀 매칭을 통해 검색 성능을 높였지만, 각 토큰의 상대적 중요도나 문맥적 흐름을 충분히 반영하지 못하는 구조적 제약이 있다. 본 연구에서는 이러한 한계를 보완하고자, 멀티벡터 기반 표현 구조를 유지하면서도 시퀀스 전체의 문맥을 반영할 수 있는 WGF(Weighted Global Fusion) 기법을 제안한다. 제안하는 방법은 기존 지역 임베딩의 세밀함은 유지하면서도, 글로벌 문맥 정보를 효과적으로 융합함으로써 멀티모달 문서 검색의 표현력을 한층 강화하였다.

III. 제안 방법

3.1 멀티벡터 기반 표현 모델 구조

본 연구에서는 기존 ColPali 아키텍처에서 제안된 멀티벡터 기반 표현 방식을 기반으로, 멀티모달 검색 성능을 향상시키기 위한 글로벌-로컬 융합 기법을 추가로 도입한다. 제안하는 모델은 Transformers 인코더 아키텍처를 기반으로 설계되며, 입력된 텍스트와 이미지 정보를 하나의 시퀀스로 통합한 뒤 각 토큰별 임베딩을 생성하고 이를 검색 임베딩으로 활용하는 인코더 중심의 검색 특화 모델이다.

입력 시퀀스는 텍스트 및 이미지 토큰으로 구성되며, 모델은 시퀀스 길이에 상응하는 개별 히든 상태(Hidden states)를 출력한다. 이 히든 상태는 각 토큰의 문맥 정보를 반영한 표현으로, 토큰 단위의 세밀한 정보 보존을 가능하게 한다. 출력된 히든 상태는 선형 변환 모듈을 거쳐 고정된 임베딩 차원으로 투영되며, 이후 벡터 간 크기 정규화를 위해 L2 정규화가 적용된다.

이와 같은 베이스라인 구조는 각 토큰의 정보를

개별적으로 보존하고 활용할 수 있는 멀티벡터 표현 방식을 채택함으로써, 세밀한 매칭이 요구되는 멀티모달 검색 환경에서 우수한 성능을 발휘할 수 있는 기반을 제공한다. 모델의 출력은 정규화된 다수의 토큰 임베딩으로 구성되며, 아래의 수식으로 계산된다. 식 (1)은 $P \in R^{L \times d}$을 정의하며, 이는 정규화된 토큰 임베딩 $\widetilde{p_i}$의 집합으로 구성된다. 식 (2)는 어테션 마스크 $m_i \in 0,1$를 적용하여 패딩 토큰에 해당하는 임베딩은 제거하고, 유효한 토큰만 유지하는 과정이다. 식 (3)은 각 토큰 벡터 p_i를 L2 노름으로 정규화하여 단위 벡터 $\hat{p_i}$를 생성하는 과정으로, 벡터 간 유사도 계산의 일관성을 화보하기 위해 사용된다.

$$P = [\widetilde{p_1}, \widetilde{p_2}, \cdots \widetilde{p_L}] \in R^{L \times d} \tag{1}$$

$$\widetilde{p_i} = m_i \bullet \hat{p_i} \tag{2}$$

$$\hat{p_i} = \frac{p_i}{\| p_i \|_2} \tag{3}$$

그러나, 선행 연구의 방법은 각 토큰 임베딩이 동등한 중요도로 취급되며, 시퀀스 전체의 문맥 정보를 충분히 반영하지 못한다는 한계를 가진다. 다음 장에서 제안하는 방법에 대해 설명한다.

3.2 제안방법

베이스라인 구조는 멀티벡터 기반 표현을 통해 토큰 단위의 세밀한 의미를 포착하는 데 강점을 가지지만, 모든 토큰이 동일한 중요도로 처리된다는 한계를 내포하고 있다. 특히 시퀀스 전반의 의미를 집약하거나, 문장 수준의 의도와 문맥을 포괄적으로 이해해야 하는 상황에서 성능 저하가 발생할 수 있다. 이를 보완하기 위해 본 연구에서는 WGF 기법을 제안하며 그림 1과 같다. WGF는 셀프 어텐션을 기반으로 시퀀스 내 각 토큰의 상대적 중요도를 학습 가능한 방식으로 계산하고, 이를 바탕으로 전체 시퀀스를 대표하는 글로벌 임베딩을 구성한 뒤, 기존의 로컬 임베딩과 선형 보간 방식으로 융합한다.

우선, 정규화된 각 토큰 임베딩 $\hat{p_i} \in R^d$에 대해 선형 프로젝션을 적용하여 스칼라 점수를 계산한다. 식 (4)는 이 과정을 수식으로 나타낸 것으로, $w_a \in R^d$는 학습 가능한 가중치 벡터이다.

$$s_i = w_a^T \hat{p_i} \tag{4}$$

이후 식 (5)는 softmax 함수를 통해 전체 시퀀스에 걸친 상대적 중요도를 나타내는 어텐션 가중치 a_i를 산출하며, 이는 다음과 같이 계산한다.

$$a_i = \frac{\exp(s_i)}{\sum_{j=1}^{L} \exp(s_j)} \tag{5}$$

최종적으로, 시퀀스를 대표하는 글로벌 임베딩 $g \in R^d$는 각 토큰의 정규화된 임베딩에 softmax 가중치를 곱한 결과의 합으로 계산하며, 식 (6)은 앞서 식 (5)에서 정의한 softmax 기반의 어텐션 가중치이며, $\hat{p_i}$는 정규화된 토큰 임베딩이다..

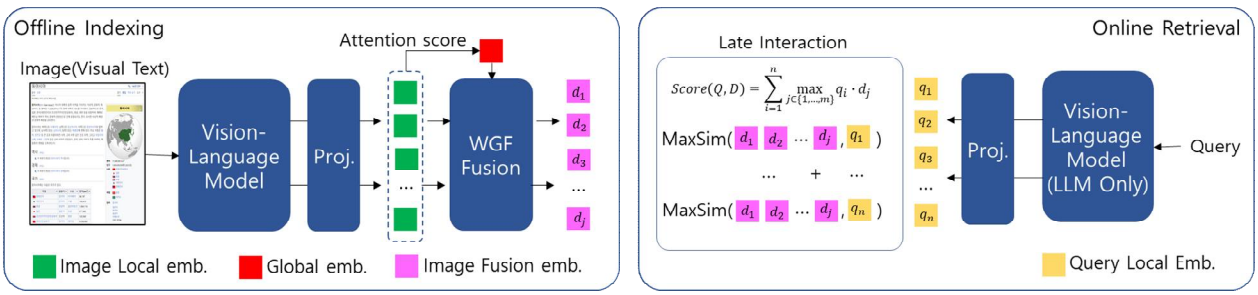

그림 1. WGF 기반 비주얼 텍스트 검색을 위한 색인 및 검색 구조
Fig. 1. Indexing and retrieval pipeline for WGF-based visual-text retrieval

$$g = \sum_{i=1}^{L} a_i \hat{p_i} \quad (6)$$

이 글로벌 벡터는 전체 시퀀스의 의미를 집약한 표현으로 해석되며, 이후 기존의 각 토큰 임베딩과 결합된다. 본 연구에서는 로컬 임베딩과 글로벌 임베딩의 조화를 위해 선형 보간 방식의 가중 융합을 사용하며, 식 (7)은 로컬 임베딩 $\widetilde{p_i}$와 글로벌 임베딩 g를 가중치로 조절하며, 최종적으로 다음과 같은 토큰 임베딩 $\widetilde{p_i}$를 생성한다.

$$\widetilde{p_i} = \lambda \widetilde{p_i} + (1-\lambda) g, \ \ \forall i \in 1,...,L \quad (7)$$

여기서, $\lambda = [0,1]$은 하이퍼파라미터로, 본 실험에서는 0.95로 설정하였다.

이러한 WGF 기반 글로벌-로컬 융합 구조는 기존 베이스라인 모델의 간단한 확장만으로도, 전체 문맥 정보를 반영하여 보다 정교한 표현을 생성할 수 있게 해준다. 특히 문장 및 전체 이미지 수준에서의 의미 이해, 중요 토큰의 강조, 세밀한 의미 적합성 판단 등 고차원적 표현이 요구되는 멀티모달 검색에서 안정적이고 효과적인 성능을 기대할 수 있다.

Ⅳ. 평가 결과

4.1 컬렉션 및 학습/평가 데이터

본 연구에서는 실제 업무 환경을 반영한 멀티모달 문서 컬렉션을 구축하여 모델의 학습 및 평가에 활용하였다. 해당 컬렉션은 표, 도표, 이미지, 그리고 한국어 기반의 본문 텍스트로 구성된 오피스 문서들로 구성되어 있으며, 실사용 문서의 구조와 내용을 충실히 반영함으로써 실제 응용 가능성을 고려한 데이터셋이라는 점에서 의의가 있다. 문서 유형은 주로 사내 규정, 지침, 운영 방침 등을 포함하며, 총 1,424페이지 분량의 문서를 이미지 형태로 변환하여 실험용 컬렉션을 구성하였다.

모델 학습을 위해 질문-이미지 쌍 형태의 데이터셋을 구성하여 활용하였다. 해당 데이터셋은 오피스 문서를 기반으로 한 질문-이미지 쌍 7,278건과 프레젠테이션(PPT) 문서를 기반으로 한 질문-이미지 쌍 1,037건으로 구성되어 있다. 질문은 주로 문서 내 특정 정보나 시각적 요소에 대한 이해를 요구하도록 설계되었으며, 이를 통해 모델이 실제 문서 검색 및 질의응답 환경에서 필요한 이해 및 추론 능력을 학습할 수 있도록 구성되었다.

모델 성능 평가는 별도로 구축된 테스트 셋을 기반으로 수행하였다. 평가 데이터는 오피스 문서를 기반으로 구성된 총 800건의 질문과 정답 이미지 쌍으로 구성하여 평가에 사용하였다.

4.2 실험 환경 및 평가 측정 방법

모델은 총 3 에폭에 걸쳐 학습하였고, 학습률은 4e-4에서 8e-4 범위로 조정하면서 성능 최적화를 위한 실험을 수행하였다. 옵티마이저는 AdamW를 사용하였고, warmup 비율은 전체 스텝의 10%로 설정하였다. 학습에 사용된 배치 사이즈는 64로, 연산의 효율성과 학습 안정성을 모두 고려하여 결정하였다. 모든 실험은 동일한 데이터 셋과 설정 조건 하에서 일관되게 수행되었다.

비교 실험으로는 대조군 구조인 AGF(Additive Global Fusion) 방식을 적용하였다. AGF는 global 임베딩 g를 L2 정규화한 뒤, 이를 각 토큰 임베딩에 단순히 덧셈하여 결합하는 방식이다. 이 방식은 별도의 중요도 가중치를 도입하지 않고, 모든 토큰에 동일한 글로벌 정보가 일괄적으로 적용된다는 점에서 구조는 단순하지만 계산 비용이 낮고 구현이 직관적이라는 장점이 있다.

또한, 글로벌 융합 기법의 효과를 정량적으로 평가하기 위해, ColPali 아키텍처에서 로컬 임베딩만을 활용하는 구조를 베이스라인 모델로 설정하고, 동일한 구조에 도메인 특화된 멀티모달 학습 데이터를 활용하여 파인튜닝한 모델과의 성능 비교 실험도 함께 수행하였다. 이를 통해 제안 기법의 실질적인 효과와 성능 향상 폭을 종합적으로 분석하였다.

검색 성능 평가는 BR(Binary Recall)@TopN 지표를 사용하였다. BR@TopN은 상위 N개의 검색 결과 내에 정답 단락이 포함되어 있는지를 기준으로 성능을 측정하며, 그 정의는 식 (8)과 같다.

$$BinaryRecall@TopN \quad (8)$$
$$BinaryRecall = \frac{\#of\ Q\ including\ correct\ passage}{\#of\ total\ \ Q}$$

4.3 평가 결과

제안한 모델의 검색 성능을 평가하기 위해 BR@TopN 기준의 정확도 지표를 사용하였으며, 그 결과를 표 1에 제시하였다. 성능 비교는 총 네 가지 실험 구성에 대해 수행되었으며, 각각은 다음과 같다: 1) 토큰 임베딩만을 활용한 베이스라인(baseline), 2) 동일한 구조에 글로벌 임베딩을 적용하지 않고 도메인 데이터로 파인튜닝만 수행한 모델(baseline+), 3) 비교 실험으로 AGF 모델, 4) 본 논문에서 제안한 WGF 모델이다.

표 1. 모델 별 비주얼 텍스트 검색 성능 비교 평가
Table 1. Comparison of visual-text retrieval performance according to different model architectures

	baseline	baseline+	AGF	WGF
BR@1	65.88%	86.88%	85.50%	91.25%
BR@5	90.63%	98.38%	98.80%	99.13%
BR@10	94.25%	99.75%	99.75%	99.63%
miss	5.75%	0.25%	0.25%	0.38%

Top-1 정확도 기준으로는, 베이스라인이 65.88%의 성능을 보인 반면, 파인튜닝만 적용한 구성에서는 86.88%, AGF는 85.50%, 제안한 WGF는 가장 높은 91.25%를 기록하였다. 이는 글로벌 문맥 정보를 토큰 임베딩에 통합하는 방식이 성능 향상에 유의미하게 기여함을 나타낸다. 반면, AGF 모델은 글로벌 임베딩을 각 토큰 임베딩에 직접 더하는 단순한 구조를 적용하였으나, baseline+보다 낮은 85.50%의 Top-1 정확도를 기록하였다. 이는 글로벌 임베딩을 모든 토큰에 동일하게 일괄 적용하는 방식이 각 토큰의 문맥적 특성과 상대적 중요도를 충분히 반영하지 못하고, 오히려 기존 멀티벡터 임베딩의 표현력을 저해할 수 있음을 시사한다. 또한, 이러한 결과는 멀티벡터 기반 표현 자체가 강력한 자질임을 보여주며, 글로벌 임베딩 정보는 이를 대체하기보다는 보조적으로 작용하여 문맥 정보를 보완하는 방식으로 활용될 때 더욱 효과적임을 의미한다. 특히 문서 내 특정 영역의 정보가 질의와 국소적으로 정합성을 갖는 멀티모달 검색 환경에서는, 각 토큰 수준의 세밀한 표현을 보존하면서 그 위에 전역 문맥 정보를 정교하게 융합하는 전략이 더 적합함을 실험 결과가 뒷받침한다.

추가적으로 글로벌 및 로컬 임베딩의 융합 비율을 조정하는 하이퍼파라미터 λ값을 조정하면서 실험한 결과는 표 2로 확인할 수 있다. 실험 결과 로컬 정보가 여전히 검색 정확도의 핵심적인 자질로 작용함과 동시에, 적절한 수준의 글로벌 정보 보완이 성능 향상에 실질적인 기여하는 것을 알 수있다. 특히 λ=0.95일 때 가장 높은 BR@Top1 정확도를 보여 융합 전략이 가장 효과적임을 뒷받침한다. 반면 λ가 낮아지는 경우(예: 0.8), 전역 정보가 지나치게 강조되면서 로컬 표현의 정밀성이 희석되어 오히려 성능 저하가 발생하는 경향을 확인할 수 있었다.

표 2. λ 값 변화에 따른 WGF 모델의 비주얼-텍스트 검색 성능 비교
Table 2. Visual-text Retrieval Performance of WGF Model under Different λ Settings

λ	0.8	0.85	0.9	0.95
BR@1	88.13%	88.25%	89.38%	91.25%
BR@5	99.00%	99.00%	98.88%	99.13%
BR@10	99.75%	99.88%	100.00%	99.63%
miss	0.25%	0.13%	0.00%	0.38%

한편, 정답을 하나도 찾지 못한 검색실패 비율을 살펴보면, baseline은 5.75%로 비교적 높은 오류율을 보인 반면, baseline+, AGF, WGF는 각각 0.25%, 0.25%, 0.38%로 모두 1% 미만의 검색 실패 비율을 나타냈다. 특히 WGF는 낮은 검색 실패 비율과 높은 Top-1 정확도를 동시에 달성함으로써, 정확성과 신뢰성을 모두 충족하는 검색 성능을 보여주었다.

V. 결론 및 향후 과제

본 논문에서는 멀티모달 검색 환경에서 표현력의 한계를 극복하고자, 새로운 임베딩 융합 기법인 WGF 방식을 제안하였다. 기존의 멀티벡터 기반 표현 모델은 각 토큰의 로컬 정보를 세밀하게 반영할 수 있다는 장점이 있으나, 시퀀스 전체의 문맥을 종합적으로 표현하는 데에는 한계가 존재하였다.

이를 보완하기 위해, 본 연구는 셀프 어텐션 메커니즘을 활용하여 입력 시퀀스로부터 글로벌 임베딩을 생성하고, 이를 기존 토큰 임베딩과 선형 보간 방식으로 융합하는 구조를 설계하였다.

제안한 WGF 구조는 기존 베이스라인 및 비교 모델들과의 실험을 통해, 도메인 특화 멀티모달 데이터셋 상에서 우수한 검색 성능을 입증하였다. 특히 Top-1 정확도 측면에서 가장 높은 성능을 달성하였으며, 검색 실패율 또한 낮은 수준을 기록함으로써 정밀도와 안정성 측면 모두에서 탁월한 성능을 보였다. 이는 시퀀스 전반의 문맥 정보를 효과적으로 통합하는 구조가 멀티모달 검색에서 실질적인 성능 향상으로 이어질 수 있음을 시사한다.

향후 연구에서는 본 논문에서 고정된 형태로 설정한 글로벌-로컬 임베딩 간의 중요도 계수를 학습 가능한 파라미터로 확장하여, 입력 내용에 따라 적응적으로 융합 비율을 조절할 수 있는 적응형 융합 방식으로 발전시키고자 한다. 또한 다양한 도메인 및 멀티모달 시나리오에 대한 확장 적용을 통해, 실환경에서의 범용성과 실용성을 더욱 심층적으로 검증할 예정이다.

References

[1] A. Most, J. Winjum, A. Biswas, S. Jones, N. R. Ranasinghe, D. O'Malley, and M. Bhattarai, "Lost in OCR Translation? Vision-Based Approaches to Robust Document Retrieval", arXiv:2505.05666, May 2025. https://doi.org/10.48550/arXiv.2505.05666.

[2] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-free Document Understanding Transformer", Proc. of the ECCV 2022, Tel Aviv, Israel, pp. 498-517, Oct. 2022. https://doi.org/10.1007/978-3-031-19815-1_29.

[3] L. L. Oliveira, et al., "Evaluating and mitigating the impact of OCR errors on information retrieval", International Journal on Digital Libraries, Vol. 24, No. 1, pp. 45-62, Jan. 2023. https://doi.org/10.1007/s00799-023-00345-6.

[4] A. Radford, et al., "Learning Transferable Visual Models From Natural Language Supervision", Proc. of the ICML, Online, PMLR 139, pp. 8748-8763, Jul. 2021. https://doi.org/10.48550/arXiv.2103.00020.

[5] X. Chen, et al., "PaLI: A Jointly-Scaled Multilingual Language-Image Model", Proc. of the ICLR 2023, Kigali, Rwanda, May 2023. https://doi.org/10.48550/arXiv.2209.06794.

[6] J. Chen, T. Lv, L. Cui, C. Zhang, and F. Wei, "XDoc: Unified Pre-training for Cross-Format Document Understanding", Findings of EMNLP 2022, Abu Dhabi, UAE, pp. 1006-1016, Dec. 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.71.

[7] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "LiT: Zero-Shot Transfer with Locked-image Text Tuning", Proc. of the CVPR 2022, New Orleans, USA, pp. 18123-18133, Jun. 2022. https://doi.org/10.48550/arXiv.2111.07991.

[8] M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, and P. Colombo, "ColPali: Contextualized Late Interaction over PaLI for Document Retrieval", Proc. of the ICLR, Singapore EXPO, Singapore, Apr. 2025. https://doi.org/10.48550/arXiv.2407.01449.

[9] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT", Proc. of the SIGIR 2020, Xi'an, China, pp. 39-48, Jul. 2020. https://doi.org/10.1145/3397271.3401075.

[10] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu, and M. Sun, "VisRAG: Vision-based Retrieval-Augmented Generation", Proc. of the ICLR, Singapore EXPO, Singapore, Mar. 2025. https://doi.org/10.48550/arXiv.2410.10594.

[11] R. Tanaka, T. Iki, T. Hasegawa, K. Nishida, K. Saito, and J. Suzuki, "VDocRAG: Retrieval -Augmented Generation over Visually-Rich Documents", Proc. of the CVPR, Nashville, USA, pp. 24827-24837, Apr. 2025. https://doi.org/10.48550/arXiv.2504.09795.

[12] S. Han, P. Xia, R. Zhang, T. Sun, Y. Li, H. Zhu, and H. Yao, "MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding", arXiv:2503.13964, Mar. 2025. https://doi.org/10.48550/arXiv.2503.13964.

저자소개

배 용 진 (Yongjin Bae)

2011년 8월 : 목원대학교 컴퓨터교육과(학사)
2014년 2월 : 과학기술연합대학원(UST) 컴퓨터소프트웨어 및 공학(석사)
2014년 5월 ~ 현재 : 한국전자통신연구원 선임연구원
관심분야 : 정보검색, 질의응답, 딥러닝

배 경 만 (Kyoungman Bae)

2004년 2월 : 동아대학교 컴퓨터공학(학사)
2016년 2월 : 동아대학교 컴퓨터공학(공학박사)
2016년 8월 ~ 현재 : 한국전자통신연구원 책임연구원
관심분야 : 언어지능, 생성형 AI, AI Safety