멀티모달 Differentiable Search Index에 기반한 텍스트-이미지 오픈도메인 질의응답

Multimodal Differentiable Search Index for Text-Image Open-domain QA

전북대학교 JEONBUK NATIONAL UNIVERSITY

윤준호 · 이성민 · 이건희 · 나승훈 (전북대학교), 서대룡 · 전동현 · 강인호 (네이버)

{hoho5702, cap1232, dkghszkfhs, nash}@jbnu.ac.kr {daeryong.seo, donghyeon.jeon, once.ihkang}@navercorp.com

Ⅰ. 서론

연구 목적

정보 검색을 위해 잘 정립된 패러다임인 DPR이 In-batch negative, Hard negative를 이용한 대조 학습 기법을 토대로 높은 성능을 보이면서 최근에는 하나의 인코더-디코더 모델을 이용해서 MIPS와 같은 고정된 검색 절차 없이 표준 모델 추론을 통해 사용자 쿼리와 문서를 직접적으로 연결하는 DSI 검색 아키텍처가 제안되었다.

또한 최근 텍스트와 이미지를 검색하는 멀티 모달 검색에 많은 연구가 집중되면서 서로 다른 모달 간의 차이를 줄이기 위한 노력이 이어지는데, 기존의 텍스트 문서 검색을 위한 DSI를 멀티 모달 검색으로 확장하여 DSI가 멀티 모달 지식을 하나의 통합된 공간에 잘 형성하는지 파악하고자 한다.

실험 세부사항

데이터셋으로는 open-domain 멀티 모달 QA 벤치마크인 WebQA를 이용하며 각 문서의 임베딩은 AltCLIP을 통해서 추출하고 K=30 그리고 C=30으로 계층적 K-Means 알고리즘을 통해 문서 식별자를 부여한다. 텍스트 문서와 관련된 쿼리 생성의 경우는 DocTTTTTQuery 모델을 이용해 최대 길이 64를 갖도록 하고 이미지 문서와 관련된 쿼리는 Faster-RCNN을 통해서 객체 감지 후 WebQA에서 제공하는 캡션과 VinVL에 입력으로 주어 MLM을 통해 학습시킨다. 모델은 사전 학습된 T5 텍스트 인코더와 SwinV2 이미지 인코더를 이용하고 디코더는 랜덤 초기화된 T5를 이용하며 언어모델 이외에도 멀티 모달 인코더-디코더인 OFA를 바탕으로 추가 실험을 진행하며 OFA의 인코더를 DPR로 학습시켜 그 성능을 비교한다.

Ⅱ. 실험 설계

모델 구성

멀티 모달 DSI의 프레임워크는 그림1 과 같다.

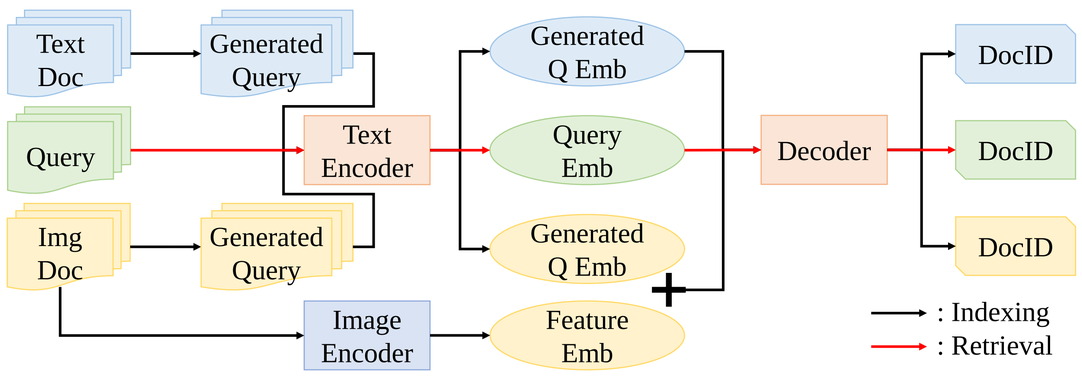

그림1: 멀티 모달 DSI Overview

멀티 모달 DSI도 기존의 DSI와 마찬가지로 Indexing과 Retrieval 과정을 통해 쿼리와 문서를 입력 받아 해당되는 문서 식별자를 생성한다. Indexing은 문서를 입력으로 받아 문서 식별자를 생성하는 과정을 의미하며, Retrieval은 입력으로 쿼리가 주어지면 자동 회귀 생성을 통해 순위가 매겨진 문서 식별자들을 반환하는 과정을 의미한다.

기존의 DSI는 Indexing과 Retrieval에 있어 다양한 기법을 제안하는데, 본 논문에서는 Direct Indexing을 통한 Inputs2Targets로 Indexing을, Retrieval을 위한 문서 식별자 표현 방법으로는 Semantically Structured Identifiers 기법을 선택하였다.

Inputs2Target는 일정 길이의 문서 토큰을 입력으로 받아 문서 식별자를 생성하는 Seq2Seq 기법으로 텍스트 문서의 경우 본 논문에서는 단순히 문서를 입력으로 주지 않고, DSI-QG에서 제안한 것과 마찬가지로 문서와 관련된 pseudo 쿼리를 추출 후 입력으로 활용하여 Training과 Inference 시에 입력 형태 불일치로 인해 DSI의 성능 저하를 유발하는 Data Distribution Mismatch 문제를 완화해준다. 이미지 문서의 경우 단순히 이미지 특징만을 이용하지 않고 UniVL-DR에서 제안한 바와 같이 다음과 같은 시퀀스 X를 Vision-Language 모델에 입력으로 주어 이미지 I와 연관된 쿼리 q를 추가로 생성하여 입력으로 활용한다.

$$X = [\mathrm{CLS}]; q; [\mathrm{SEP}]; C; [\mathrm{SEP}]; \overrightarrow{O_1}; \ldots; \overrightarrow{O_l};$$

생성한 이미지 쿼리는 텍스트 인코더를 통해 이미지의 텍스트 임베딩을 추출하고, 이미지 인코더를 통한 이미지의 특징 임베딩을 얻어 텍스트 임베딩과 이미지 임베딩을 더하여 디코더에 통과시켜 문서 식별자를 생성한다.

DSI는 문서 식별자를 다양한 기법으로 표현하는데, 본 논문에서는 Semantically Structured Identifiers를 이용한다. 이는 계층적 K-Means 알고리즘을 적용하여 의미론적 문서 식별자를 부여해 비슷한 문서들은 서로 비슷한 문서 식별자를 가지도록 하는 것이다. Indexing 해야 하는 모든 문서의 임베딩을 추출 후, K개의 클러스터로 분류하고 문서가 C개를 초과하는 클러스터는 계층적 K-Means 알고리즘을 재귀적으로 적용시키며 문서가 C개 이하인 클러스터는 문서들에 0부터 C-1 까지의 식별자를 할당한다. 이를 통해 문서의 의미론적인 정보들이 디코딩 과정에 통합된다.

Ⅲ. 실험 결과

표 1: 모달별 검색 결과비교

	MMT5-DSI			OFA-DSI		
	TXT	IMG	ALL	TXT	IMG	ALL
R@1	18.85	23.09	**21.00**	12.70	19.87	16.33
R@5	30.95	40.30	**35.68**	21.58	33.09	27.40
R@10	35.39	45.91	**40.71**	24.60	37.87	31.31
R@20	40.65	51.73	**46.25**	28.43	43.01	35.80
R@50	45.41	58.98	**52.27**	33.93	49.90	42.00
R@100	48.96	64.83	**56.98**	37.43	56.19	46.91
MRR@100	14.58	23.41	**18.29**	10.11	19.54	14.07

표 2: DPR-DSI 검색 결과 비교

모델	사전학습	Recall					MRR
-	-	@1	@10	@20	@50	@100	@100
OFA-DPR	✓	14.51	40.52	**48.06**	**57.57**	**64.37**	**22.92**
	✗	06.39	20.67	26.02	33.40	38.89	15.59
OFA-DSI	✗	16.33	31.31	35.80	42.00	46.91	14.07
MMT5-DSI	✗	**21.00**	**40.71**	46.25	52.27	56.98	18.29

결과 분석

MMT5-DSI와 OFA-DSI 모두 사전 학습을 거치지 않은 OFA-DPR보다 높은 검색 성능을 보이며 MMT5-DSI의 경우 사전 학습된 OFA-DPR보다 높거나 비슷한 검색성능을 보인다.

추가로 멀티 모달 DSI는 몇 가지 성능 개성의 여지가 있다. 멀티 모달 DSI 학습 시 생성한 쿼리들의 질이 높지 않은데, 조금 더 문서와 관련된 질 높은 쿼리를 활용하고 OFA-DPR이 사전 학습을 통해 많은 성능 향상을 보인 것처럼 OFA-DSI도 사전 학습을 거치면 더 높은 성능을 보일 것으로 기대된다.

또한 문서 식별자에 있어 단순히 번호를 사용하는 것을 넘어 PLM의 역량을 더욱 발휘할 수 있는 문서 식별자를 사용한다면 더 우수한 성능을 보일 것으로 기대된다.

Ⅳ. 결론

결론

본 논문은 검색 분야에서의 새로운 아키텍처인 DSI를 멀티 모달 검색으로 확장하고 이를 WebQA 벤치마크에서 평가함으로써 DPR보다 더 나은 검색 성능을 달성할 수 있음을 보여주었다. 특히, DSI 아키텍처의 멀티 모달 확장을 통해 텍스트와 이미지, 비디오 등 다양한 모달을 모두 고려하여 정보 검색을 수행할 수 있다는 것을 확인하였다.

본 논문은 멀티 모달 정보 검색 분야에서 새로운 방향성을 제시하며, DSI 아키텍처를 활용하여 다양한 분야에서의 적용 가능성을 시사한다.