

이미지-텍스트 멀티모달 Dense Retrieval를 위한 생성모델 기반 데이터 증강

2023.06.20

Presentation by: Sung-Min Lee

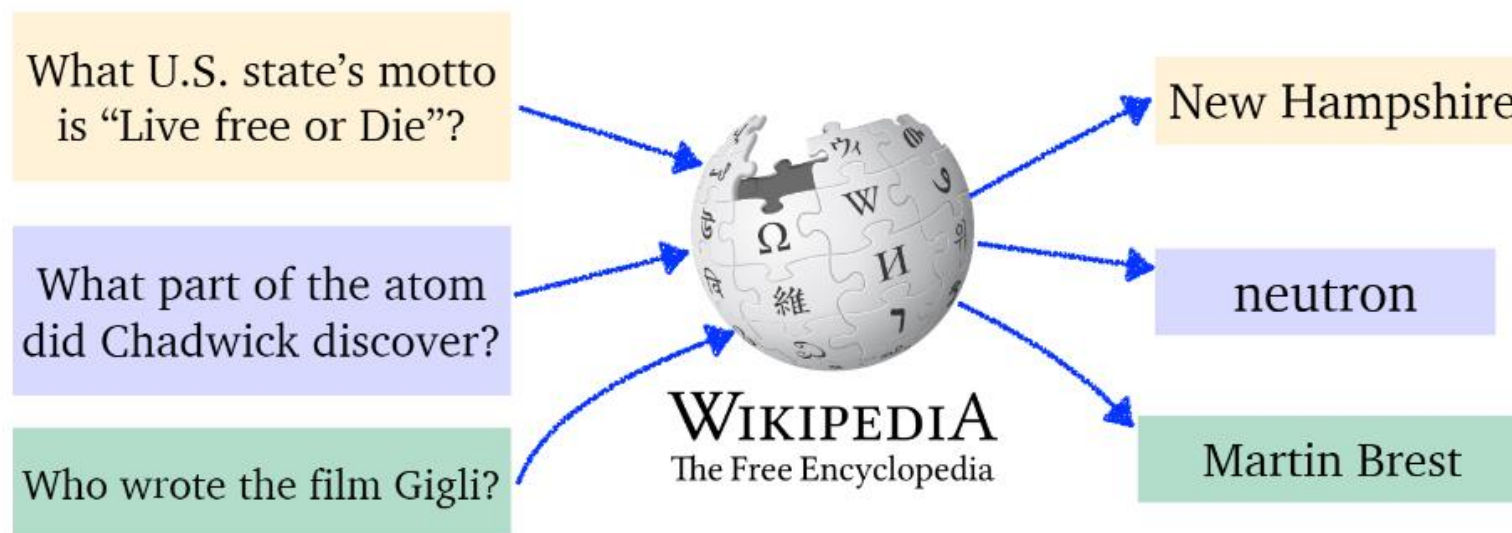
Mail: cap1232@jbnu.ac.kr



NAVER

Open Domain Question Answering

- **Input:** question Q , D = English Wikipedia (~5 million documents)
- **Output:** answer A



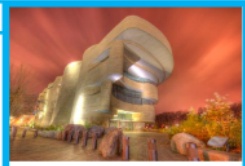
주어진 질의에 대해 연관된 지식을 필요로 하는 Task

Problem Formulation

WebQA (Multihop and Multimodal Open Domain QA over Image and Text)


Question: Are both the National Museum of the American Indian in Washington, D.C. and the Xanadu House in Kissimmee, Florida the same color?

Pos




'title': National Museum of the American Indian in Washington, D.C.
'caption': National Museum of the American Indian in Washington, D.C.

Neg



'title': Rechterswil Isler-Schale 02 09
'caption': Rechterswil Isler-Schale 02 09 Isler shell, concrete dome roof of a building of the former company Kilcher in Rechterswil, built 1965; Solothurn, Switzerland.



'title': Xanadu-House-in-Kissimmee-Florida-1985
'caption': Xanadu-House-in-Kissimmee-Florida-1985 A photo of the Xanadu House that was located in Kissimmee, Florida, showing the exterior of the house.

'title': 'Booker T. Washington', 'fact': 'In 1946, he was honored on the first coin to feature an African American, the Booker T. Washington Memorial Half Dollar, which was minted by the United States until 1951. On April 5, 1956, the hundredth anniversary of Washington's birth, the house where he was born in Franklin County, Virginia, was designated as the Booker T. Washington National Monument.'

'title': 'Xanadu Houses', 'fact': 'The Xanadu house in Kissimmee, Florida used an automated system controlled by Commodore microcomputers. The house had fifteen rooms; of these the kitchen, party room, health spa, and bedrooms all used computers and other electronic equipment heavily in their design.'

'title': 'Xanadu Houses', 'fact': 'Construction of the Xanadu house in Kissimmee, Florida, began with the pouring of a concrete slab base and the erection of a tension ring 40 feet (12 m) in diameter to anchor the domed roof of what would become the "Great Room" of the house.'

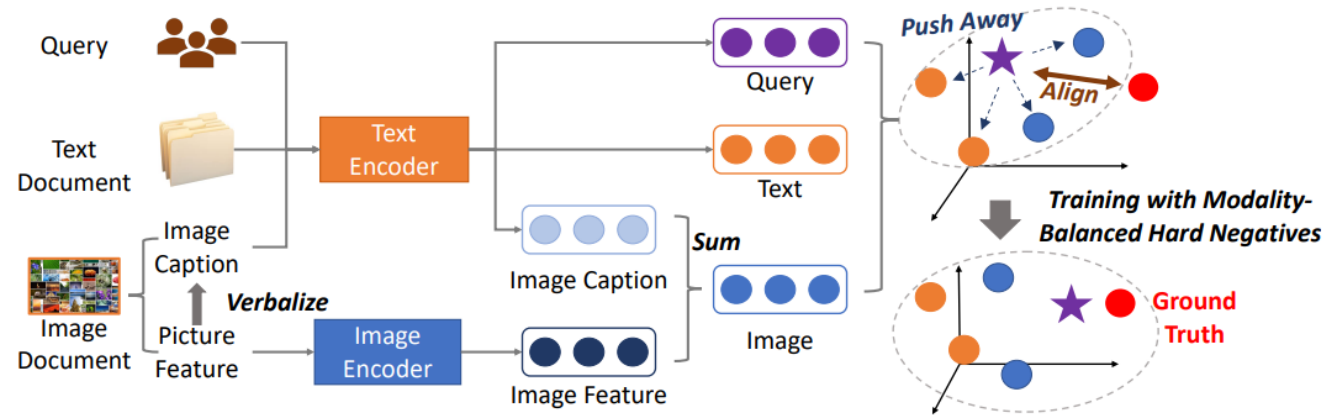
'title': 'Xanadu Houses', 'fact': 'The interior of the house was cave-like, featuring cramped rooms and low ceilings, although it is not clear whether these accounts describe the same Xanadu House with a thirty-foot dome. The interiors used a cream color for the walls, and a pale green for the floor.'

Answer: Yes, both the National Museum of the American Indian in Washington, D.C. and the Xanadu House in Kissimmee, Florida are beige.

- 정답을 도출하기 위해 이미지, 텍스트 sources 사이에서 positive sources를 검색하는 것이 필요. (즉, Multimodal retrieval가 필요하다.)
- 두 가지 세팅이 있는데, restricted ($n \approx 40$)와 Full ($n \approx 900K$) 세팅이 있음.

Chang, Yingshun, et al. "Webqa: Multihop and multimodal qa." CVPR 2022

Prior works: UniVL-DR



$$L = -\log \frac{e^{f(q, d^+)/\tau}}{e^{f(q, d^+)/\tau} + \sum_{d^- \in \mathcal{D}^-} e^{f(q, d^-)/\tau}}$$

$$= \underbrace{-f(q, d^+)/\tau}_{L_{\text{Align}}} + \log(e^{f(q, d^+)/\tau} + \underbrace{\sum_{i=1}^{k_1} e^{f(q, d_{\text{Image}}^{i-})/\tau}}_{L_{\text{Image}}} + \underbrace{\sum_{j=1}^{k_2} e^{f(q, d_{\text{Text}}^{j-})/\tau}}_{L_{\text{Text}}}),$$

Contributions

- 1) modality-balanced hard negatives
- 2) Image verbalization method

Prior works: PAQ

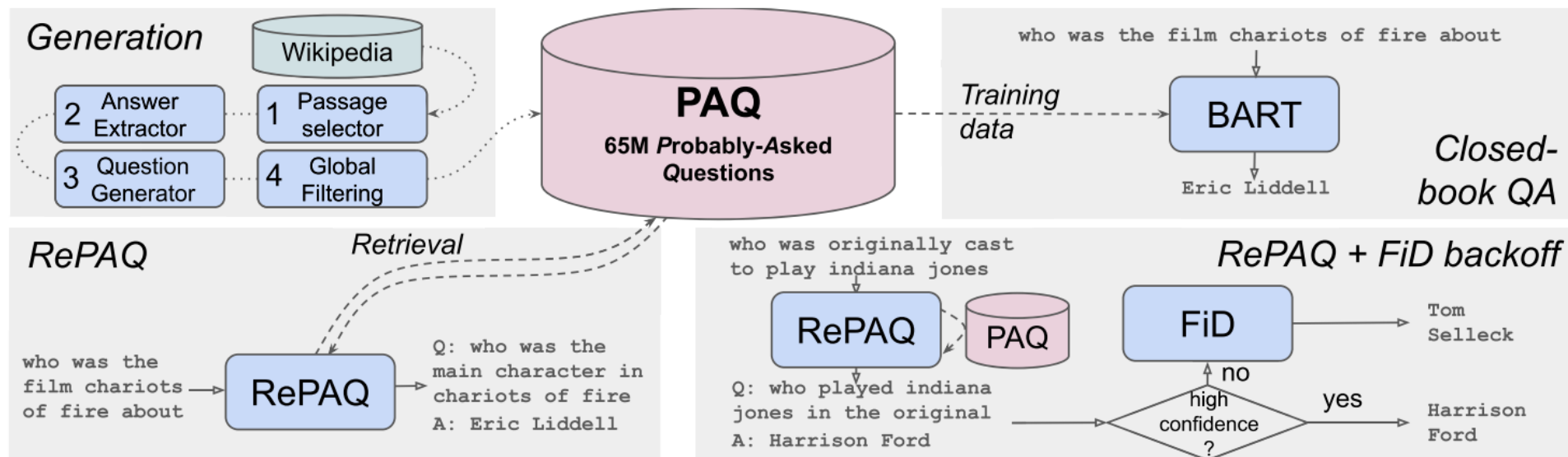
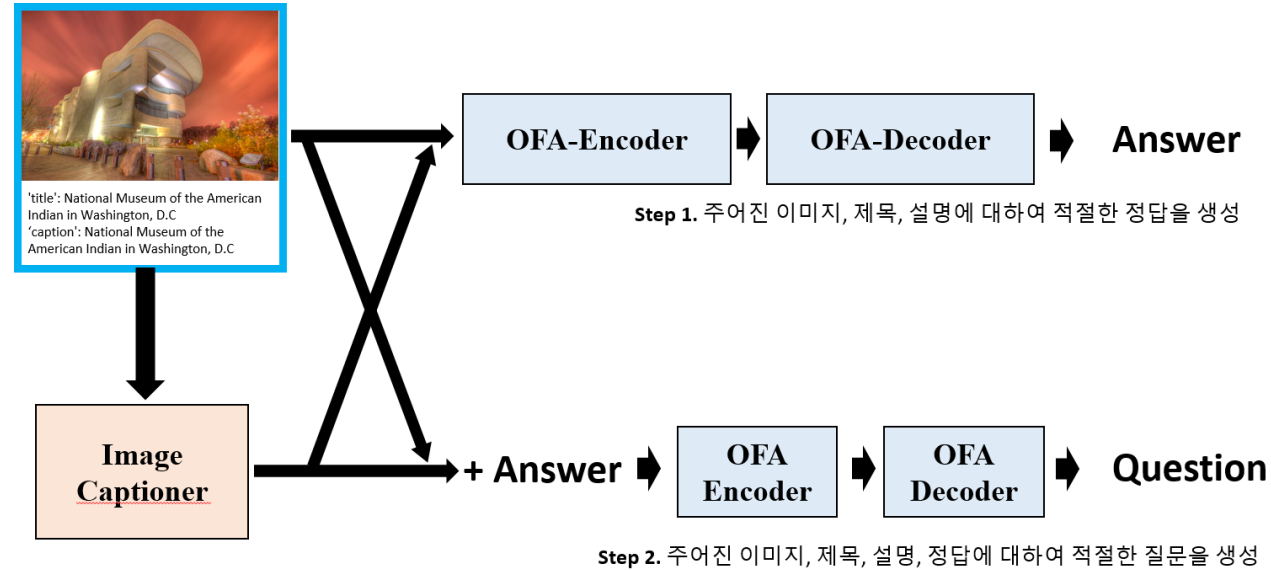


Figure 1: Top Left: Generation pipeline for QA-pairs in PAQ. Top Right: PAQ used as training data for CBQA models. Bottom Left: RePAQ retrieves similar QA-pairs to input questions from PAQ. Bottom right: RePAQ's confidence is predictive of accuracy. If confidence is low, we can defer to slower, more accurate systems, like FiD.

Our approach: Generative Multi-Modal Data Aug



- 현재 시스템에서는 Answer extractor, Question generator 두가지로 구성.
- Image captioner는 Fine-tuning된 OFA 모델 사용
- Image selector, Filtering을 추가로 도입할 필요가 있음.

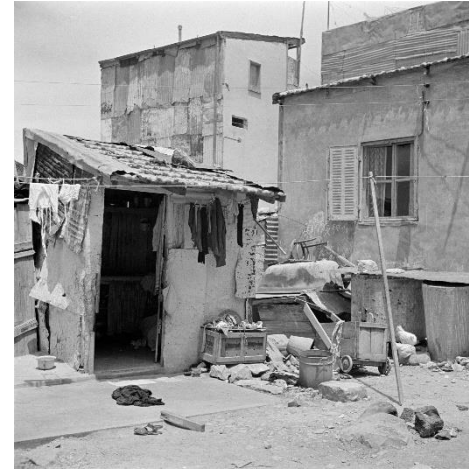
Experimental results

생성된 데이터 샘플들



A: The mouth is open.

Q: Is the mouth of the Burmese python open or closed?



A: The roof of the orphanage in Krotwoning is sloped.

Q: Is the roof of the orphanage in Krotwoning flat or sloped?



A: Yes, there is a phone number in the window of the Ritz Cinema.

Q: Is there a phone number in the window of the Ritz Cinema?

Experimental results

Model	MRR@10	NDCG@10	MRR@20	NDCG@20	Rec@20	Rec@100
CLIP (Zero-Shot)	10.59	8.69	10.80	9.52	14.32	20.21
VinVL-DPR	38.14	35.43	38.74	37.79	53.89	69.42
CLIP-DPR	48.83	46.32	49.34	49.11	69.84	86.43
UniVL-DR	62.40	59.32	62.69	61.22	80.37	89.42
CLIP-DPR(ours)	48.12	45.59	48.65	48.30	68.72	85.64
CLIP-DPR(after pre-training)	50.65	48.09	51.21	51.05	72.13	87.26
UniVL-DR(ours)	60.59	57.42	60.88	59.27	77.90	87.23
UniVL-DR(after pre-training)	62.75	59.72	63.05	61.87	81.53	90.86

표 1: 전체 실험 결과

Model	MRR@10	NDCG@10	MRR@20	NDCG@20	Rec@20	Rec@100
CLIP-DPR(ours)	48.12	45.59	48.65	48.30	68.72	85.64
CLIP-DPR(MMAug + PAQ)	50.56	48.10	51.05	50.83	71.79	87.81
CLIP-DPR(MMAug + WebQA text)	50.65	48.09	51.21	51.05	72.13	87.26

표 3: 사전학습 데이터셋 종류에 따른 추가 분석 실험 결과

- VQA 데이터 추가로 이용해서 사전학습 진행했었으나, 추가 성능 향상은 이루지 못함.
- LAION 데이터셋 고려했지만 보안문제로 진행X

MRR: 검색 결과의 순위를 역수로 변환하여 평균을 구한 지표.
 NDCG: 검색 결과의 관련성 점수를 할인 계수를 적용해 누적한 값.
 Recall: 검색된 관련 문서 중 실제 관련 문서의 비율.

Experimental results

Model	MRR@10	NDCG@10	MRR@20	NDCG@20	Rec@20	Rec@100
CLIP-DPR(fine-64)	48.12	45.59	48.65	48.30	68.72	85.64
CLIP-DPR(fine-512)	48.00	45.57	48.52	48.27	69.02	85.83
CLIP-DPR(pre-64, fine-64)	49.94	47.52	50.48	50.38	71.35	86.84
CLIP-DPR(pre-512, fine-64)	50.65	48.09	51.21	51.05	72.13	87.26

표 2: 배치사이즈에 따른 추가 분석 실험 결과

Model	MRR@10	NDCG@10	MRR@20	NDCG@20	Rec@20	Rec@100
CLIP-DPR	60.36	61.30	60.79	63.45	84.31	94.82
UniVL-DR	64.93	65.95	65.29	67.72	87.69	94.74
CLIP-DPR(ours)	57.78	59.02	58.20	61.26	83.13	93.89
CLIP-DPR(after pre-training)	60.99	61.91	61.43	64.26	85.68	94.58
UniVL-DR(ours)	63.13	64.16	63.44	65.92	85.74	93.51
UniVL-DR(after pre-training)	65.97	66.36	66.33	68.38	87.71	95.02

표 3: 이미지에 대한 검색 성능평가 결과