

OFA 기반 멀티모달 이미지-텍스트 오픈도메인 질의응답

이성민, 박은환, 서대룡, 전동현, 강인호, 나승훈

cap1232@jbnu.ac.kr

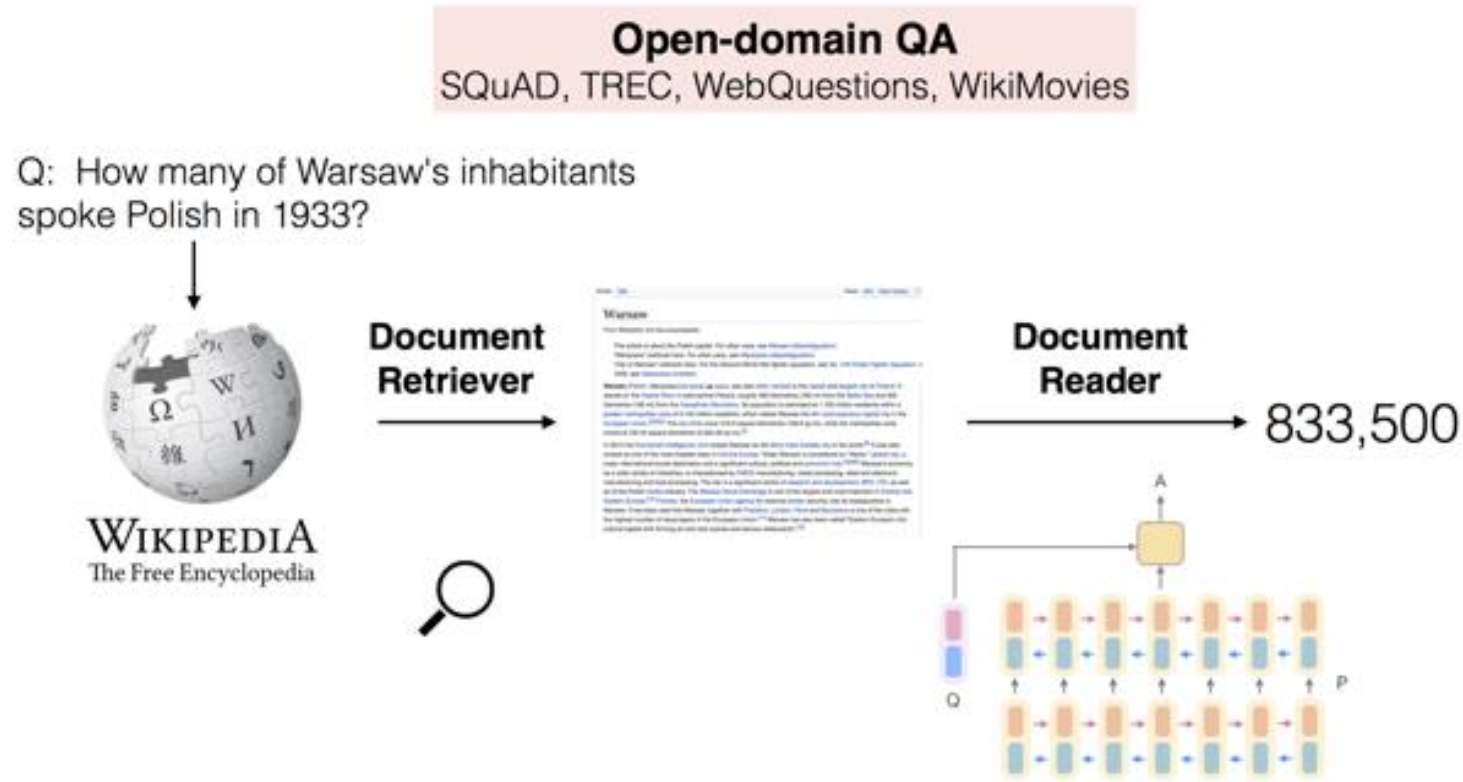


NAVER

Introduction

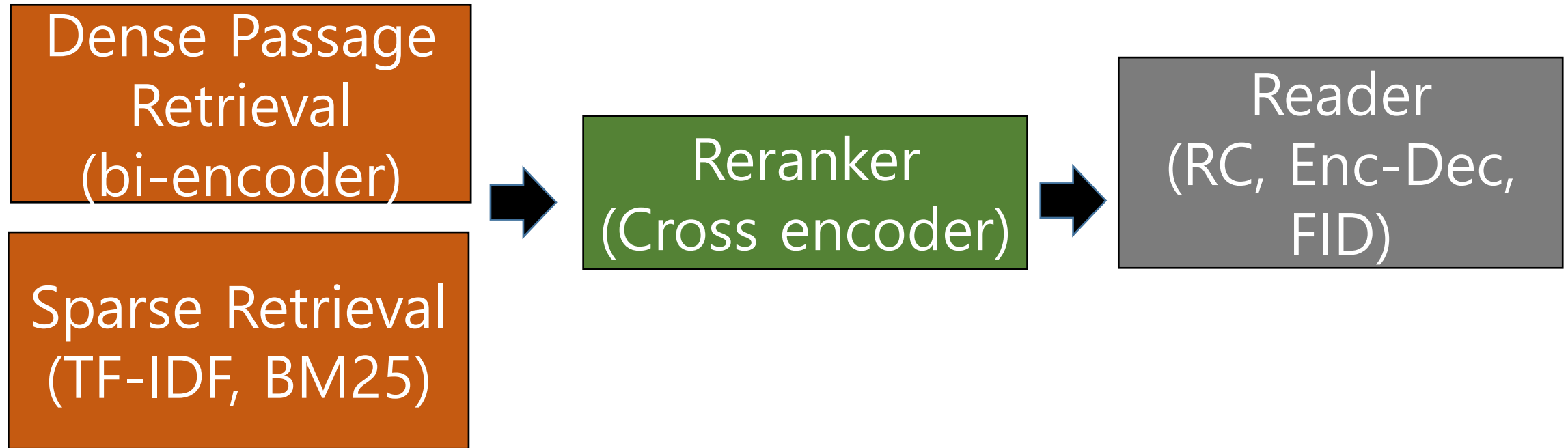
Open Domain Question Answering

오픈 도메인 질의 응답은 명시적인 단서가 제공되지 않아 위키피디아나 웹, 지식베이스 같은 데이터 자원을 기반으로 자연어 질문에 답변하는 Task (e.g. IRQA, KBQA)

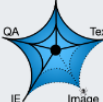


Retrieval based Open Domain Question Answering

- Retrieval-Reranker-Reader



Dataset - WebQA

WebQA 

Yingshan Chang Yonatan Bisk Mridu Narang
Guihong Cao Hisami Suzuki Jianfeng Gao

WebQA is a new benchmark for multimodal multihop reasoning in which systems are presented with the same style of data as humans when searching the web: Snippets and Images. The system must then identify which information is relevant across modalities and combine it with reasoning to answer the query. Systems will be evaluated on both the correctness of their answers and their sources.

Task Formulation: Given a question Q , and a list of sources $S = \{s_1, s_2, \dots\}$, a system must **a)** identify the sources from which to derive the answer, and **b)** generate an answer as a complete sentence. Note, each source s can be either a snippet or an image with a caption. A caption is necessary to accompany an image because object names or geographic information are not usually written on the image itself, but they serve as critical links between entities mentioned in the question and the visual entities.

Evaluation: Source retrieval will be evaluated by F1. The answer quality will be measured by Fluency (BARTScore) and Accuracy (keywords overlap).

Q: At which festival can you see a castle in the background: Oktoberfest in Domplatz Austria or Tanabata festival in Hiratsuka, Japan?

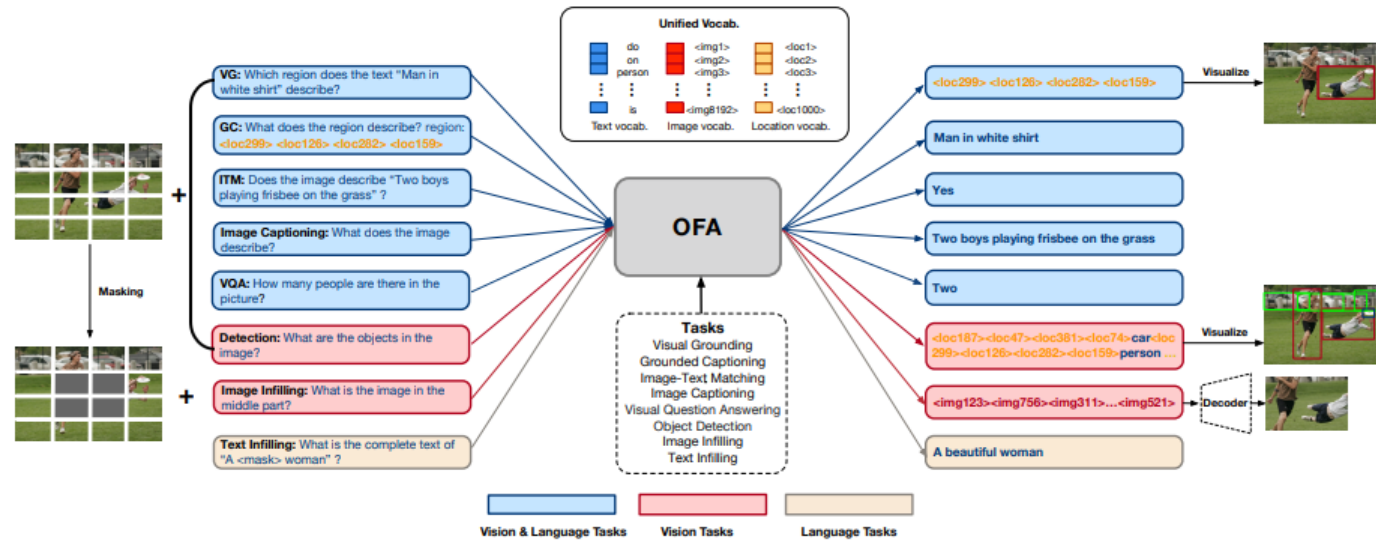
A: You can see a castle in the background at Oktoberfest in Domplatz, Austria

Task

- 1) Source Retrieval (Question에 대응되는 sources(either Text or Image-description))
- 2) Question Answering (question과 검색된 sources 기반하여 answer generation)

Our approach

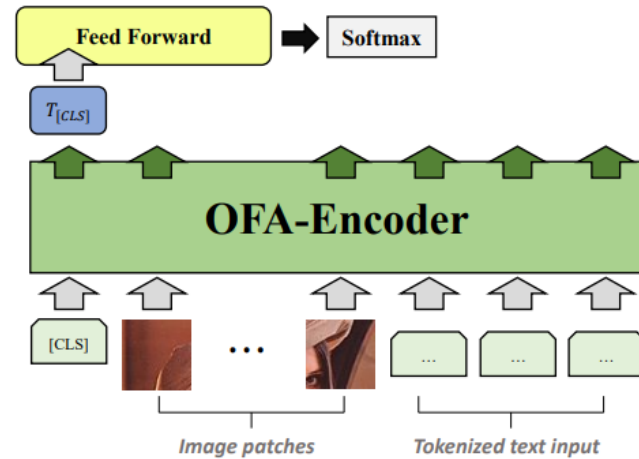
Pre-trained model



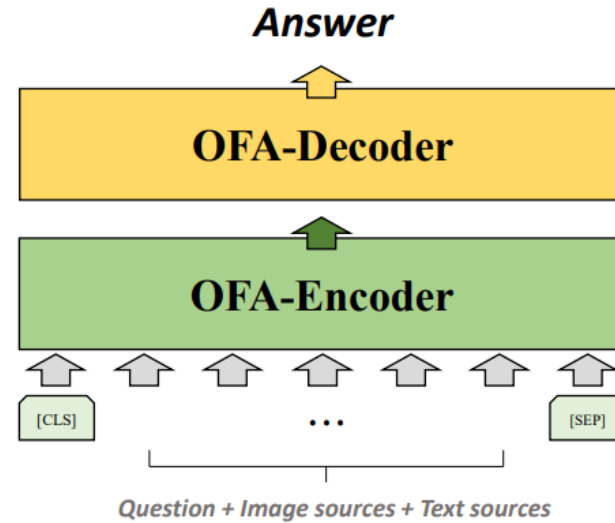
OFA (ICML 2022, accepted)는 Text-Image 통합 Encoder-Decoder 모델로, Tiny ~ Large Size 까지 지원하고 있고,

기존 Baseline의 Foundation model인 VLP(2020 AAAI)와 성능이 크게 10%까지 차이가 나서 OFA를 Foundation model로 채택했다.

Model Architecture



Multi-Modal Source Retrieval



Multi-Modal Answer Generator

- 1. Source Retrieval:** Question 마다 Source가 20~40개 정도 주어지는데, 그 중에 question에 답하기 위한 적 절한 source를 찾아내야 한다. 따라서 다음과 같이 Source retrieval를 구성했다. Positive class probability를 rank score로 활용한다.
- 2. Question answering module:** Question에 대해 검색된 Image-description sources와 Text sources를 취합해 answer를 생성한다.

Hard negative mining

1. 모델을 학습 후 학습된 모델을 이용해서 negative sample들을 추론하고, negative confidence가 낮게 나오는 sample들을 hard negative sample으로 간주한다.
2. Multi-Modal Source Retrieval 모델을 다시 학습시키는데, 추출된 hard negative를 우선으로 batch에 포함하여 학습을 진행한다.

Experiment - results

Evaluation Metrics	Distractor			
	Retr	FL	Accuracy	Overall
Question-Only	-	34.9	22.2	13.4
VLP (Oscar)	68.9	42.6	36.7	22.6
VLP + ResNeXt	69.0	43.0	37.0	23.0
VLP + VinVL	70.9	44.2	38.9	24.1
MURAG	74.6	55.7	54.6	36.1
Ours	77.42	41.75	44.5	24.43
Ours (+hard negative)	82.46	48.11	47.97	28.12

- WebQA test dataset에 대한 실험결과 -

Conclusion

- 본 논문에서는 OFA 모델을 이용하여 Multi-Modal 오픈도메인 QA를 위한 시스템을 제안했다. 또한 hard negative mining을 통한 추가적인 성능 향상을 이루어 냈다.

Future works

- 대용량 Image-Text 멀티모달 데이터셋을 통한 사전학습을 통해 멀티모달 모델의 성능을 더욱 향상시킬 예정.