Ko-METER: Dual-Stream 구조의 Vision-Language Foundation Model을 이용한 한국어 시각 정보 질의응답

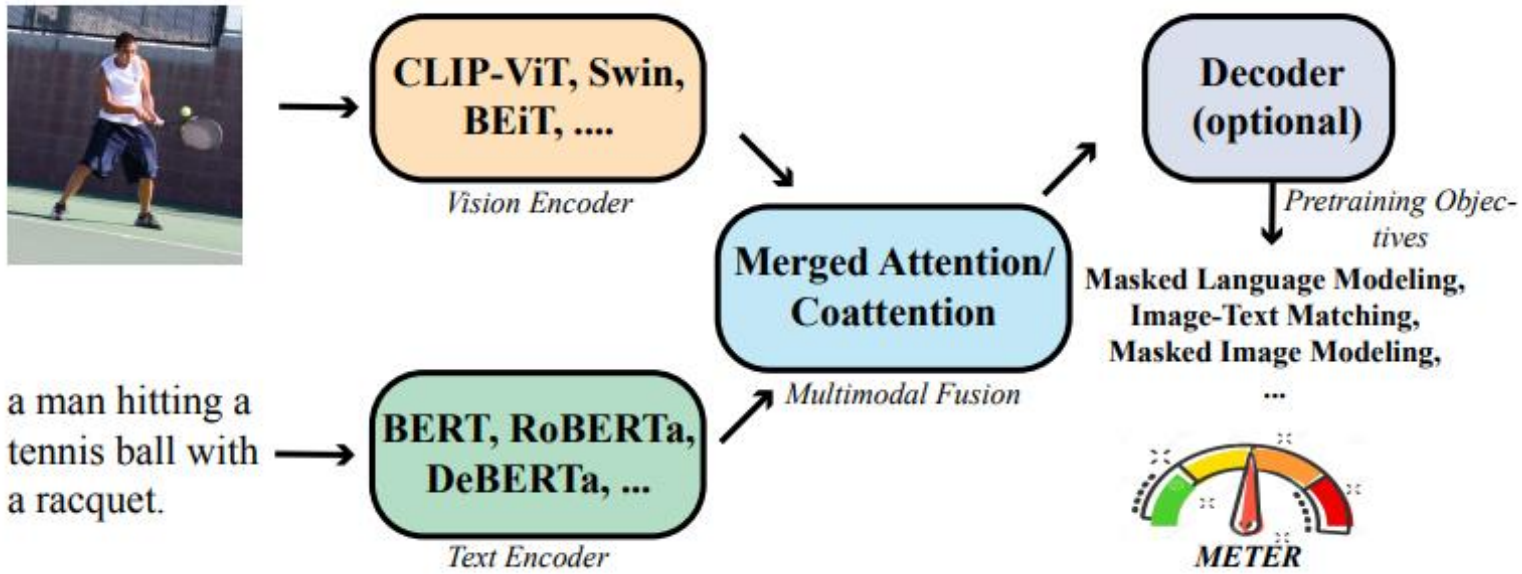
이성민, 나승훈
전북대학교

서대룡, 김선훈, 강인호
네이버

NAVER

KCC 2022

Vision-language Pre-training (METER)



METER:

이미 학습된 Vision Encoder, Language Encoder을 베이스로 Multimodal Fusion 을 통해 통합된 Vison-Language 임베딩을 얻어내는 모델.

Ko-METER Pre-training

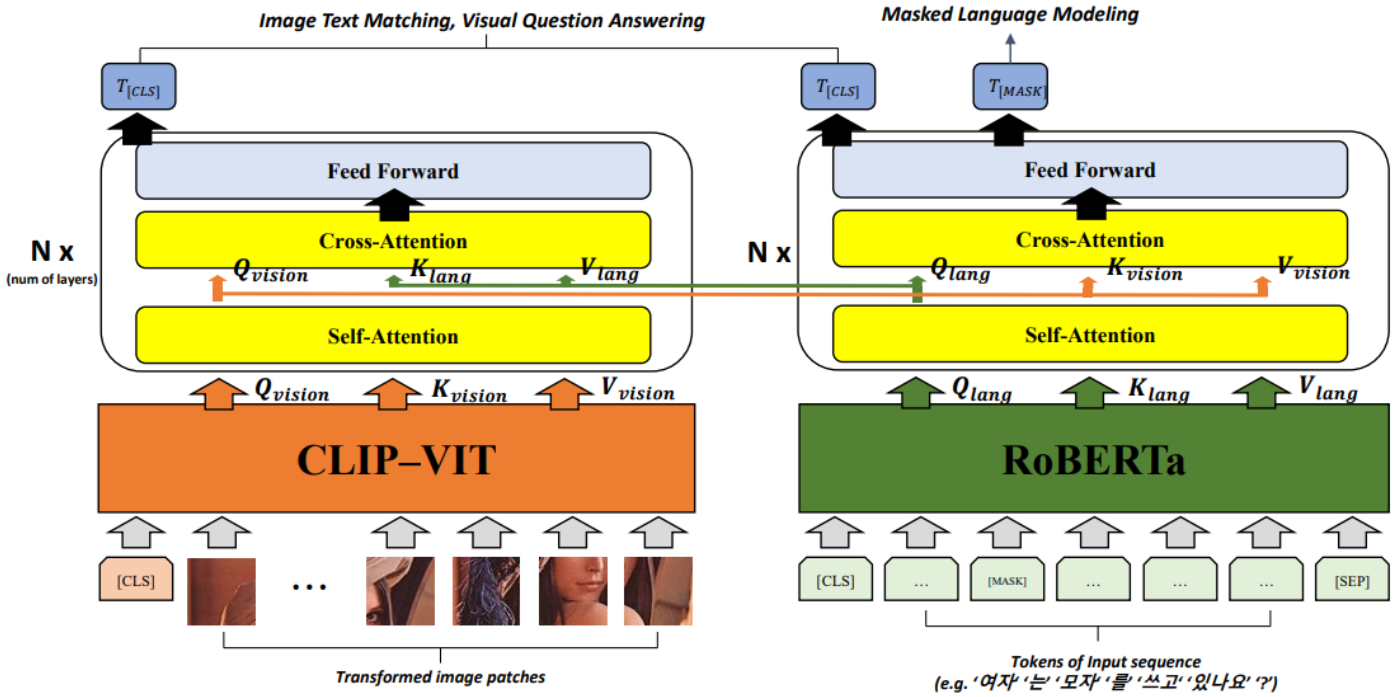
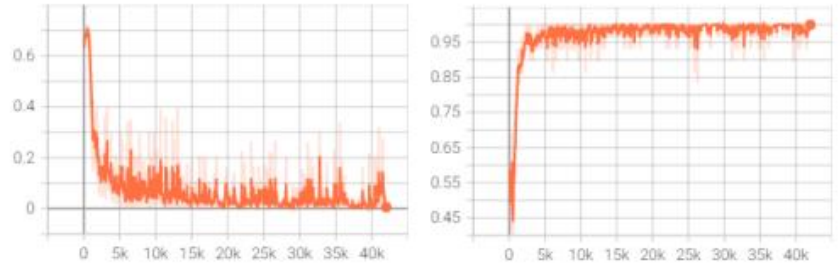


그림 1: Dual-Stream 구조의 Vision-Language Foundation model architecture

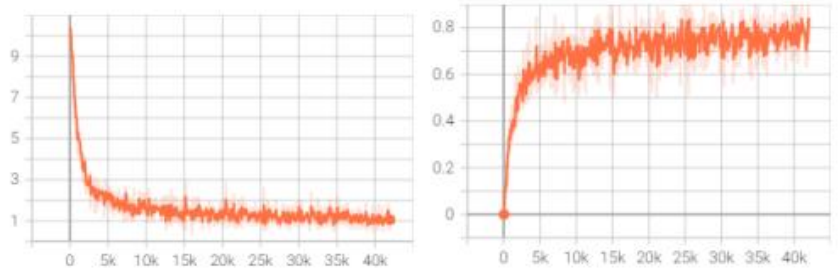
Co-attention 방식으로 Attention layer를 Vision, Language Encoder 위에 쌓아서,
(Vision Encoder – CLIP VIT, Language Encoder – KLUE RoBERTa-base)
한국어 COCO 데이터셋(image 12만개 * 이미지당 caption 5개)으로 Pre-training 진행.
Loss는 MLM, ITM 두가지 사용.

실험결과

그림 2: 사전학습에 활용된 Task들에 대한 Loss, Accuracy 그래프



(a) ITM Loss (b) ITM Accuracy



(c) MLM Loss (d) MLM Accuracy

학습을 빠르게 진행하기 위해, batch size 256, seq_len 40 으로 40000 Steps 학습을 진행하였다. (논문에서는 batch size 4096, 100000 Steps)

실험결과

AIhub 시각정보 질의응답 데이터 셋으로 Fine-tuning 진행.

제공되는 Dev 셋이 너무 커서, Train/Dev – 3900000/3000 으로 재구성. (Class개수 – 3811)

Pre-trained 모델의 성능을 비교하기 위해, Pre-trained 되지 않은 모델도 fine-tuning하여 Dev set에 대해 Step 별 성능을 비교해 보았다.

결과는 아래 표와 같다.

	3415 steps	6831 steps	10250 steps
Pre-training	0.532	0.567	0.5873
w/o Pre-training	0.4987	0.534	0.564

실험결과

AIhub 시각정보 질의응답 데이터 셋으로 Fine-tuning 진행.

제공되는 Dev 셋이 너무 커서, Train/Dev – 3900000/3000 으로 재구성. (Class개수 – 3811)

Pre-trained 모델의 성능을 비교하기 위해, Pre-trained 되지 않은 모델도 fine-tuning하여 Dev set에 대해 Step 별 성능을 비교해 보았다.

결과는 아래 표와 같다.

	예 or 아니오	숫자	알 수 없음	그 외
Pre-training	93.09	38.10	35.08	48.70
w/o Pre-training	95.68	38.83	35.08	42.98

실험결과

표 3 은 AIHUB 생활 및 거주환경 기반 VQA 데이터 셋을 이용한 미세조정 성능 비교표이다. Single-stream 구조의 VILT[2] 모델과 비교했을 때, Dual-Stream 구조인 본 논문의 모델이 더 좋은 성능을 보임을 알 수 있다.

표 3: Dev Accuracy on AIHUB 생활 및 거주환경 기반 VQA 데이터 셋

Model	Accuracy
Our Model	**85.24**
VILT[2]	79.46

결론

- 본 논문에서는 Co-attention layer를 이용해 이미지, 텍스트의 표 상을 통합하는 Dual-Stream 구조의 Vision-Language Model을 구 성하고 사전학습을 진행했다.
- 사전학습 완료된 모델을 이용 하여 한국어 시각 정보 질의응답 데이터 셋에 미세 조정을 진행해 사전학습된 모델의 성능을 확인하고, 그 결과, 사전학습 된 모델이 안 된 모델보다 더 좋은 성능을 보였고, VILT 모델과 비교했을 때 본 논문의 모델이 5.78% 높은 성능을 보였다.