다층 Co-Attention과 Question-aware prompt를 통한 Knowledge-based Visual Question Answering

박은주(이화여자대학교) | 김지원(연세대학교) | 이한결(홍익대학교) | 김태경(동국대학교) |
조유림*(성균관대학교)

deep daiv. 대한전자공학회 IEIE The Institute of Electronics and Information Engineers

ABSTRACT

In the field of Knowledge-Based Visual Question Answering (KB-VQA), where external knowledge is essential for accurate responses, significant progress has been made through the use of Large Language Models (LLMs). BLIP-2, a widely used multimodal LLM, employs a single-layer Q-Former for visual feature extraction and cross-modal interaction. However, it faces challenges in handling tasks requiring complex reasoning.

To address these limitations, we propose integrating the Multimodal Co-Attention Network (MCAN), which utilizes a multi-layered structure to enhance interactions between visual and textual inputs. Additionally, we introduce Question-Aware Prompts during fine-tuning, which combine Answer Candidates with confidence scores and Answer-Aware Examples from previous cases. This approach improves the model's ability to interpret questions and generate more contextually relevant answers.

Experimental results on KB-VQA datasets demonstrate a 6.9% accuracy improvement over baseline models, highlighting the effectiveness of our method in addressing complex multimodal reasoning tasks.

METHOD

Q-Former와 MCAN의 결합

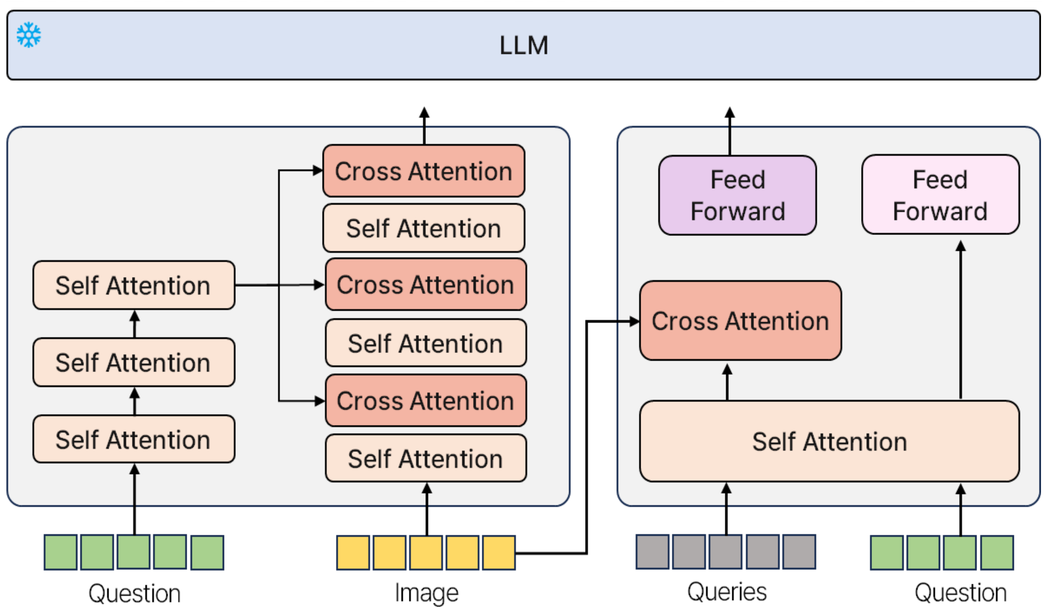

그림 1. 모델의 구조

- 본 연구는 텍스트와 이미지 간의 다층적 상호작용을 효과적으로 학습하기 위해 Q-Former와 MCAN을 결합한 새로운 구조를 제안
- 기존 Q-Former는 Self-Attention과 Cross-Attention 메커니즘의 스태킹 구조로 질문 특징이 충분히 정제되지 않은 상태에서 이미지와 상호작용을 수행하는 한계를 가짐
- 이를 보완하기 위해 MCAN의 인코더-디코더 구조를 도입, Self-Attention으로 질문 특징을 단계적으로 정제하고 최적화된 질문 표현을 생성
- 이어 Cross-Attention을 활용해 이미지 정보에 대한 주의를 효과적으로 집중시켜 질문의 의미를 충실히 반영하고 상호작용을 정교화
- 제안된 방법은 단일 레이어 구조에서 발생할 수 있는 정보 손실을 최소화하며 텍스트-이미지 간의 복합적 관계를 효율적으로 모델링

Question-Aware Prompts

Question-Aware Prompt 구조

Prompt info: "Your task is to answer a knowledge-based question. Please choose the correct answer in the choices according to the context, the question, the answer candidates. Each answer candidate is associated with a confidence score within a bracket. The true answer may not be included in the candidates."
===
Context: Inflated kites in various shapes float in the air.
Question: What chemical makes cats fly?
Candidates: air(0.68), helium(0.62), ...
Answer: helium
===
Context: The man is smiling at a birthday cake.
Question: What is he about to blow out?
Candidates: candle(0.99), birthday(0.02), fire(0.01)
Answer: candle
===
Context: a group of children stand around a cake.
Question: What fills the balloons?
Candidates: air(0.28), helium(0.07), wine(0.03)
Answer:

그림 2. Question-Aware Prompt

- 파인 튜닝 단계에서 모델이 질문의 문맥적 정보를 효과적으로 이해할 수 있도록 Question-Aware Prompt를 적용
- 이 Question-aware prompt는 Answer candidates와 Answer-aware examples로 구성
- 이를 통해 LLM이 더 정밀한 추론을 할 수 있을 뿐만 아니라, 필요에 따라 다양한 답변을 탐색할 수 있는 유연성 제공

Answer Candidates

- 실시간으로 제공되는 답변 후보를 바탕으로 모델이 정답을 선택
- 각 답변 후보에는 신뢰도 점수를 포함하여 모델이 정답에 가까운 답변을 우선적으로 선택할 수 있게 유도

Answer-Aware Examples

- 질문의 의미와 유사한 질문에 대한 과거 답변을 활용하여 추론 능력을 강화
- 이런 사례 기반 학습은 모델이 새로운 질문에 대해 더 정확한 추론을 가능하게 하여, 복잡한 질문-이미지 상호작용을 처리

RESULTS

Comparison Accuracy

	Only-Question Acc	Question-Aware Prompt Acc
Q-Former	**49.2**	**55.65**
MCAN	52.65	-
Ours (Q-Former + MCAN)	**50**	56.1

표 1. 기존 모델들의 성능 비교: Question-Aware Prompt 도입 여부에 따른 성능 차이

- 표1에서 확인할수 있듯이, Question-Aware Prompt 미사용 시 정확도가 50% 였으나, 문맥적 정보 반영 후 정확도는 56.1% 로 상승
- 해당 결과를 통해 질문의 배경 정보와 문맥적 정보가 성능 향상에 기여함을 확인
- 또한, MCAN 구조를 도임함으로써, 기존의 Q-Former보다 복잡한 이미지와 질문 간의 관계를 깊이 있게 학습하여 정확한 응답을 생성

Ablation Study

	OK-VQA Acc	Question-Aware Prompt Acc
Image Feature w/ Linear Projection to Q-Former	**38.83**	-
Image Feature to Q-former	**43.15**	-
Image Feature w/ Linear Layer to LLM	51.19	**53.8**
Ours (Image Feature to LLM)	**50**	56.1

표2. 성능 비교 실험: 다양한 설정에 따른 성능 비교

- MCAN의 이미지 특징 출력을 Q-Former에 입력한 경우 OK-VQA 데이터셋에서 정확도는 43.15%를 기록했으나, 이를 LLM에 입력했을 때는 50%로 더 높은 성능을 보임
- Question-Aware Prompt를 활용한 파인 튜닝 결과, 이미지 특징을 입력으로 사용하는 방식이 56.1%의 정확도로 가장 높은 성능을 보였으며, 선형 레이어 방식을 추가한 경우보다 성능이 우수함(53.8%)

CONCLUSION

- 본 연구는 외부지식이 필요한 KB-VQA 분야에서 성능을 개선하기 위해 Q-Former와 MCAN을 결합한 새로운 구조를 제안
- 이를 통해 텍스트와 이미지 간의 복잡한 관계를 학습하고, Question-Aware Prompt를 도입하여 질문의 문맥적 정보를 반영함으로써 정답 가능성이 높은 답변을 선택
- 제안된 모델은 OK-VQA와 AOK-VQA 데이터셋에서 기존 방법보다 높은 성능을 보이며, Question-Aware Prompt를 활용해 최종적으로 56.1%의 정확도를 달성
- 향후 연구에서는 다양한 모달리티를 통합하여 모델의 범용성과 성능을 극대화하고, 더 복잡한 질의응답 시나리오에서도 높은 성능을 유지할 수 있도록 연구를 확장할 예정

REFERENCE

[1] Li, J., "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv preprint arXiv:2301.12597, 2023.

[2] Yu, Z., Jiang, Y., Huang, Z., Yang, C., and Xu, W., "Deep Modular Co-Attention Networks for Visual Question Answering," arXiv preprint arXiv:1906.10770, 2019.

[3] Yu, J., Lin, K., Jiang, Y., Chen, Z., and Bansal, M., "Prophet: Prototypical Networks for Few-Shot Visual Question Answering," arXiv preprint arXiv:2303.01903, 2023.

[4] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.