멀티모달 VQA: XLM_RoBERTa와 ResNet50 기반의 한국어·일반 도메인 성능 분석

신수인(국어국문학과), 오석준(기계공학과), 이가연(전기전자공학부)

2025 CUAI 중앙대학교 인공지능 학회 하계 컨퍼런스
Proceeding of 2025 Chung-Ang University Artificial Intelligence Summer Conference

Abstract

본 연구는 한국어 기반 경량 멀티모달 VQA(Visual Question Answering)의 성능을 평가하고 그 한계를 규명하는 것을 목표로 한다.

해당 모델은 XLM-RoBERTa 텍스트 인코더와 ResNet50 이미지 인코더를 결합하고, 두 임베딩을 element-wise 곱을 사용해 융합한다. 한국어(A1) 및 일반(A2) 데이터셋을 활용한 검증 결과, Top-1 정확도는 각각 13.17%, 6.89%로 나타났는데, 이는 방대한 클래스 공간, 심각한 long-tail 분포, 단순한 융합 구조의 한계에서 기인한 것으로 분석했다.

이와 같은 한계점을 개선하고자 Cross-Attention 기반 융합, 데이터 정규화 및 증강, 외부 신호 결합을 통한 개선 가능성을 제시했다.

Introduction

최근 인공지능 연구에서 멀티모달 학습은 단일 모달리티의 한계를 극복하기 위한 방안으로 주목받고 있다. 그중 VQA는 이미지와 텍스트를 동시에 이해·추론해야 하는 대표적 과제로써 다양한 응용 가능성을 갖는다.

다만 기존 VQA 연구와 벤치마크는 영어권에 크게 편중되어 왔으며, 한국어를 포함한 비영어권에 대한 검증은 상대적으로 부족했다. (물론 최근에는 격차가 해소되는 중)

따라서 본 연구는 한국어 환경에서 VQA 모델이 어떻게 작동하는지에 집중하여 XLM-RoBERTa와 ResNet50을 사용한 경량 구조의 멀티모달 VQA 모델을 구현하였다. 이후 한국어(A1)와 일반(A2) 데이터셋을 구분하여 성능을 평가하고, 정량적 지표와 시각화 분석(Grad-CAM, Self – Attention)을 수행했다.

이 연구를 통해 한국어와 일반 도메인에서의 VQA 성능을 실질적으로 검증하고, 데이터 불균형 및 오분류 패턴을 분석하여 향후 개선 가능한 모델구조와 학습 전략을 제안하고자 하였다.

Aim

본 연구의 목적은 세 가지로, 모델 구현, 성능 검증, 해석·분석이다.

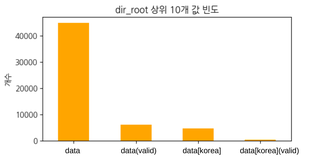
먼저 XLM-RoBERTa와 ResNet50을 결합한 경량 멀티모달 VQA를 설계한다. 그후 성능 검증을 위해 한국어(A1)와 일반(A2) 도메인 데이터셋에서 정확도(accuracy)와 확신도(confidence)를 측정하며, Grad-CAM 과 Self-Attention 시각화를 활용해 모델의 추론 근거를 정성적으로 분석한다.

Methods

1) 데이터셋 및 전처리

데이터셋의 각 샘플은 img_path, domain, set(train/test), answer, question 등의 필드로 구성된다. 본 연구에서는 검증을 위해 A1(한국어, N=372)과 A2(일반, N=6051)를 사용했다. 질문은 XLM-RoBERTa 토크나이저로 변환해 최대 50 토큰까지 제한을 두어 Padding 및 Truncation을 진행했고, 이미지는 224×224로 크기 조정 후 ImageNet 표준 값으로 정규화했다. 정답(Answer) 공간은 학습/검증 데이터의 답변 텍스트를 정규화하여 빈도순으로 리스트를 구성했고 (미등록 라벨은 '<unk>'로 매핑해 처리), 학습 시 CrossEntropyLoss의 타깃 인덱스로 사용했다.

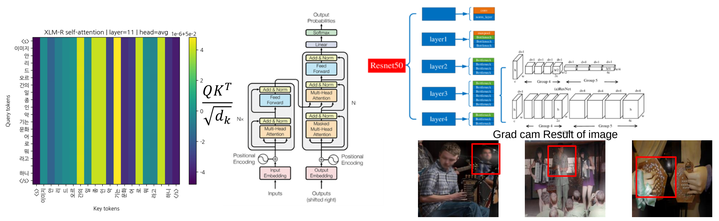
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56027 entries, 0 to 56026
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   img_path        56027 non-null  object
 1   dir_root        56027 non-null  object
 2   filename        56027 non-null  object
 3   key             56027 non-null  object
 4   domain          56027 non-null  object
 5   set             56027 non-null  object
 6   filename_label  56027 non-null  object
 7   answer_raw      56027 non-null  object
 8   answer          56027 non-null  object
 9   question        56027 non-null  object
 10  source_json     56027 non-null  object
 11  priority        56027 non-null  int64
```

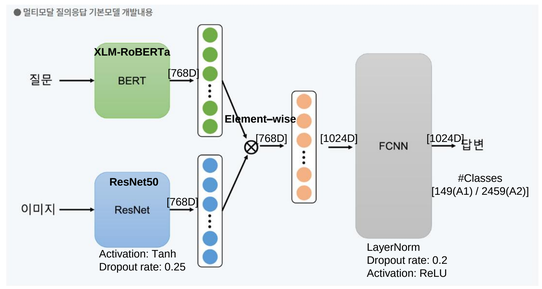


2) 모델 구조

텍스트 인코더로는 xlm-roberta-base, 이미지 인코더로는 timm의 ResNet50(pretrained)을 사용했다. 이미지 임베딩은 xlm-roberta의 텍스트 임베딩 크기에 맞추기 위해 FC 레이어를 사용해 768차원으로 매핑하고, 텍스트 임베딩(XLM-R의 pooler output)과 element-wise 곱으로 융합했다. 이후 ReLU와 Dropout을 거쳐 정답 분류기를 통과시켰다. 옵션으로 Transformer Encoder Layer(3층, 8-head)를 융합 임베딩에 추가할 수 있게 설계되어 있기는 하나, 본 연구에서는 해당 레이어를 사용하지 않은 베이스 모델을 구축했다.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad Q: query, K: key, V: value, d_k: key\ dimension$$





The outputs from BERT and ResNet are combined via element-wise multiplication before being fed into the FCNN

- LayerNorm: 이미지 · 텍스트 융합 feature의 분포를 안정화하여 학습 안정성을 확보
- Activation: XLM-R 임베딩과 스케일을 맞추기 위해 사용했으며, ResNet 출력을 Tanh로 [-1, 1] 범위 정규화한 후 FCNN에서 ReLU를 적용해 gradient 소실을 방지하고 학습 효율 강화
- Dropout: 이미지 branch(0.25)는 과적합 방지를 위해 강하게 설정했으며, 융합 hidden layer(0.2)는 정보 손실을 최소화하며 정규화를 유지하기 위해 사용

3) 학습 및 평가 설정

항목	설정값/방법
Optimizer	Adam(lr=3e-5, weight_decay=0.001
Loss Function	CrossEntropyLoss(기본), FocalLoss(선택적) [γ = 2.0]
Epoch, Batch Size	50 Epochs, 512
Regularization	Dropout(image:0.25, fushion:0.2), LayerNorm
Evaluation	Top-1 Accuracy, Top-5 Accuracy, Mean Confidence

Result & Discussion

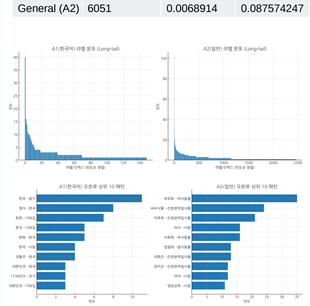
4) 결론

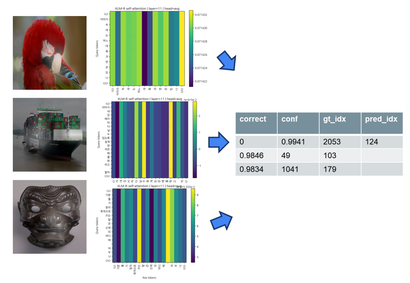
a) 정량 성능
정답일 때의 평균 확신도와 오답일 때의 평균 확신도 차이는 A1(한국어, 0.392 vs 0.177), A2(일반, 0.207 vs 0.079)로, 확신도가 예측 품질의 지표로 작동함을 나타낸다.

b) 클래스 공간과 장꼬리(long-tail) 분포
A1의 유효 라벨 수는 149개, A2는 2459개로 A2의 클래스 공간이 현저히 크다고 볼 수 있다. A1에서 최빈 50개 라벨은 전체의 68.0%를, A2에서는 15.9%를 차지했으며, 특히 A2에서 극심한 장꼬리 분포가 관찰됐다. 예측 라벨 다양성은 A1 28개, A2 116개로 모델이 빈도 상위 클래스로 수렴하는 경향이 나타났다.

c) 해석 시각화
Grad-CAM 결과는 모델이 질문과 연관된 시각 영역(예: 전통 건축의 지붕/기와, 의복의 특정 패턴)에 집중함을 나타낸다.
또한 Self-Attention 히트맵은 질문 내 일부 토큰에 상대적으로 높은 주의가 분포함을 보여준다.

domain	n	acc	Conf_mean
Korean (A1)	372	0.13172	0.2056656
General (A2)	6051	0.0068914	0.087574247





correct	conf	gt_idx	pred_idx
0	0.9941	2053	124
0.9846	49	103	
0.9834	1041	179	

5) 논의

전반적인 정확도는 A1 13.17%, A2 6.89%로 낮게 나타났다. Self-Attention 히트맵 또한 값이 비교적 균등하게 나오는 것을 통해 단어 간의 패턴이 잘 학습되지 않았다는 것을 확인할 수 있다. 주요 원인은 (1) 매우 큰 정답 클래스 공간, (2) 데이터의 long-tail 분포, (3) 텍스트-이미지 융합 시 단순 element-wise 결합 구조의 한계로 분석된다.

오분류는 주로 범주론적 위계(confusable taxonomy) 영역에서 집중적으로 발생했다.
일반 도메인에서는 세부 종 수준의 라벨이 상위 분류군으로 뭉뚱그려지는 경향이 관찰되었다. 예를 들어, '포유류/식육목'과 같은 세부 종이 '척삭동물'이라는 상위 분류군 라벨을 가졌다. 또 한국어 도메인에서는 문화적 연관성이 강한 라벨 간 혼동이 반복적으로 나타났다. 이 경우에는 '한국'을 '옹기', '기와집', '청자'로 잘못 예측하거나 반대로 일반화하는 문제가 나타났다.

6) 한계 및 개선 방법 제안

- 결합부 단순성: element-wise 곱은 멀티모달의 상호작용을 충분히 모형화하지 못하는 것으로 보여, Cross-Attention, FiLM, Co-Attention 등을 도입할 수 있다.
- 라벨 공간 축소 및 정규화: 동의어·상하위어를 통합하여 답변 공간을 축소할 수 있다.
- 데이터 증강 · 균형화: 도메인/클래스 균형 샘플링, MixUp·RandAugment, 질문 paraphrasing 등을 적용할 수 있다.
- 외부 신호 결합: OCR, 레이아웃, CLIP 임베딩 결합을 활용해 텍스트 포함 이미지와 개념 질문 대응력을 강화할 수 있다.
- 학습 기법: Focal Loss, class-balanced reweighting, label smoothing, temperature scaling(캘리브레이션) 등을 통해 클래스 불균형에 대응하고 개선할 수 있다. 다만 Focal Loss를 사용하는 경우, 희귀 클래스 성능은 올라가지만 전체 평균 정확도가 낮아질 수 있다는 점이 우려된다.

한편, 본 연구에서 관찰된 범주론적 위계 문제와 관련하여, [7]은 VQA 모델을 포함한 비전-언어 모델(VLLM)이 생물학적 분류 체계와 같은 계층적 시각 이해에 취약한 원인으로 LLM의 구조적 추론력 부족을 지적하고 있다. 이에 따라 향후 연구에서는 단순한 데이터나 파라미터 확장에 그치지 않고, 모델 설계와 학습 전략을 근본적으로 개선할 필요가 있다.

Reference

[1] Kim, M., Song, S., Lee, Y., Jang, H., & Lim, K. (2024, March). Bok-vqa: Bilingual outside knowledge-based visual question answering via graph representation pretraining. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 18381-18389).
[2] AI-hub 외부 지식 기반 멀티모달 질의응답 데이터.https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71357
[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on
[5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
[6] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., & Kalantidis, Y. (2019). Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217.
[7] Tan, Y., Qing, Y., & Gong, B. (2025). Vision LLMs Are Bad at Hierarchical Visual Understanding, and LLMs Are the Bottleneck. arXiv preprint arXiv:2505.24840.