문서 이미지 품질에 따른 비전-언어 모델의 어텐션 및 응답 분석

문준렬, 임완수*
*성균관대학교

wnsfuf008@skku.edu, *wansu.lim@skku.edu

Impact of Document Image Quality on Vision-Language Models

Moon Junryeol, Lim Wansu*
*Sungkyunkwan University

요 약

비전-언어 모델(Vision-Language Model, VLM)은 문서 질의응답, 정보 추출, 문서 이해와 같은 다양한 분야에서 활용되고 있다. 그러나 실제 문서 환경에서는 스캔 품질 저하, 해상도 감소, 블러와 같은 입력 이미지 품질 열화가 빈번히 발생하며, 이러한 시각적 요인이 VLM의 내부 추론 과정과 최종 언어 응답에 미치는 영향은 충분히 분석되지 않았다. 특히 기존 연구들은 주로 대규모 데이터셋을 활용한 파인튜닝이나 모델 구조 개선을 통한 성능 향상에 초점을 맞추어 왔으며, 모델 내부의 시각적 어텐션과 생성 응답 간의 관계를 체계적으로 분석한 연구는 상대적으로 부족하다. 본 연구에서는 문서 이미지 품질 변화가 VLM의 시각적 어텐션 분포와 생성된 언어 응답에 미치는 영향을 분석한다. 이를 위해 Qwen2-VL-7B-Instruct 모델을 대상으로 원본 이미지, 블러 처리된 이미지, 그리고 해상도 향상 기반 전처리를 적용한 이미지를 입력으로 사용하여 동일한 질의응답 실험을 수행하였다. 각 조건에서 VLM의 시각적 어텐션 맵과 생성 응답을 함께 분석하였다. 실험 결과, 이미지 품질 열화는 시각적 어텐션을 문서 전반으로 확산시키고 응답의 안정성을 저하시킨 반면, 해상도 향상 전처리는 관련 영역에 대한 집중도를 향상시키는 효과를 보였다. 또한 이러한 어텐션 안정화는 생성 응답의 일관성과 구체성 향상과 밀접한 관련이 있음을 확인하였다. 본 연구는 입력 이미지 품질이 VLM의 내부 시각적 어텐션 안정성과 응답 생성 특성에 밀접하게 연관되어 있음을 보이며, 향후 어텐션 맵을 활용한 학습 전략 설계의 필요성을 제시한다.

Ⅰ. 서 론

비전-언어 모델(Vision-Language Model, VLM)은 이미지와 텍스트 정보를 활용하여 질의응답, 문서 이해와 같은 다양한 태스크를 수행한다 [1]. 최근 VLM의 발전으로 문서에 대한 이해력이 향상되었으나, 실제 문서 환경에서는 다양한 시각적 열화 요인이 존재한다. 스캔 해상도 저하, 블러, 인쇄 품질 차이와 같은 문제는 입력 이미지 품질을 저하시켜 모델의 시각적 인식 과정에 영향을 미칠 수 있다.

기존의 VLM 연구들은 주로 모델 구조 설계나 대규모 데이터셋을 활용한 파인튜닝을 통해 성능을 향상시키는 데 초점을 맞추어 왔다 [2]. 이러한 접근은 전반적인 성능 개선에는 효과적이지만, 모델 내부의 시각적 어텐션이 실제로 어떤 근거를 바탕으로 언어 응답을 생성하는지에 대한 분석은 상대적으로 제한적으로 다루어져 왔다. 특히 입력 이미지 품질 변화가 VLM 내부의 어텐션 분포에 어떠한 변화를 유발하며, 이러한 변화가 최종 생성 응답의 안정성과 일관성에 어떻게 반영되는지에 대한 체계적인 분석은 충분히 이루어지지 않았다.

본 연구는 문서 이미지 품질 변화가 VLM의 시각적 어텐션과 생성 응답에 미치는 영향을 함께 분석함으로써, 문서 기반 VLM의 내부 추론 과정을 보다 깊이 이해하는 것을 목표로 한다. 이를 통해 어텐션 맵과 생성 응답 간의 연관성을 분석하고, 향후 VLM의 새로운 학습 전략 설계를 위한 시사점을 도출하고자 한다.

Ⅱ. 본론

2. 1. 이론적 배경 및 선행연구

문서 이미지 처리 분야에서는 OCR 성능 향상을 목적으로 해상도 향상, 대비 강화, 노이즈 제거와 같은 다양한 전처리 기법들이 제안되어 왔다 [3]. 이러한 연구들은 주로 문자 인식 정확도 개선이나 문서 구조 보존에 초점을 두고 있으며, 최근에는 주로 대규모 데이터셋을 활용한 파인튜닝을 통해 문서 이해 성능을 향상시키는 방식이 널리 사용되고 있다. 그러나 이러한 방식들은 최종 성능 개선에 집중되어 있어, 입력 이미지 품질 변화가 VLM의 내부 시각적 추론 과정에 미치는 영향까지는 상대적으로 제한적으로 다루어졌다 [2].

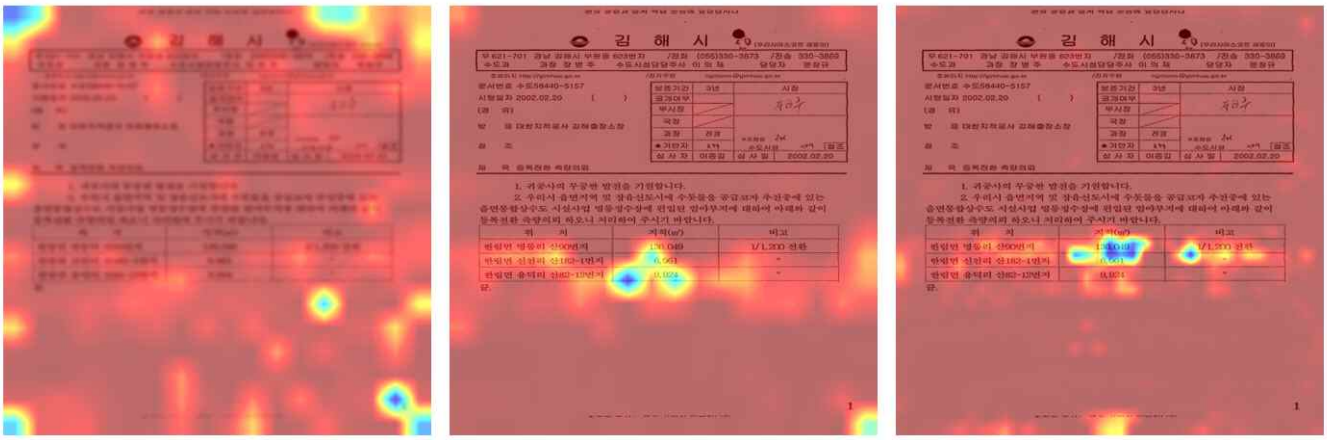
한편, 비전-언어 모델 연구에서는 어텐션 시각화를 통해 모델의 시각적 근거를 분석하려는 시도가 증가하고 있다. 그러나 입력 이미지 품질 변화에 따른 어텐션 분포 변화와 생성 응답 특성 간의 관계를 함께 분석한 연구는 제한적이다. 본 연구는 문서 이미지 품질 변화가 VLM의 시각적 어텐션과 생성 응답에 미치는 영향을 함께 분석함으로써, 어텐션 집중도와 응답 품질 간의 연관성을 분석하고자 한다.

2. 2. 실험 설정

본 연구에서는 Qwen2-VL-7B-Instruct 모델을 사용하여 문서 기반 질의응답 실험을 수행하였다 [4]. 입력 문서 이미지는 AIHub 공공행정문서 OCR로 구성되며, 동일한 문서에 대해 다음 세 가지 전처리 조건을 적용하였다:

- Original: 원본 문서 이미지
- Blur: Gaussian blur를 적용하여 시각적 정보가 열화된 이미지
- Upscaling: Bicubic 업샘플링과 unsharp masking을 결합한 해상도 향상 전처리 이미지

Upscaling은 딥러닝 기반 모델을 사용하지 않고, 해상도 증가와 고주파 성분 강조를 통해 텍스트 획과 표 경계를 선명하게 만드는 방식이다. 이는 의미 정보를 변경하지 않으면서 이미지 품질을 향상시키기 위함이다. 각 전처리 조건에서 동일한 질의를 입력하고, 모델이 생성한 언어 응답과 VLM의 시각적 어텐션 맵을 추출하였다.



Blur Original Upscaling

그림 1 이미지 품질에 따른 Qwen2-VL-7B-Instruct 모델의 어텐션 맵 예시

Query	이미지	응답
한림면 명동리 산90번지의 지적은?	Blur	한림면 명동리 산90번지의 지적은 **2**입니다.
	Original	한림면 명동리 산90번지의 지적은 **136,049**입니다.
	Upscaling	한림면 명동리 산90번지의 지적은 **136,049 m²**입니다.

표 1 프롬프트 유형에 따른 Qwen2.5, Mistral, Llama3.1 모델 추론 결과

2. 3. 실험 결과 및 분석

그림 1과 표 1은 서로 다른 이미지 품질 조건에서의 시각적 어텐션 맵과 생성 응답 예시를 보여준다. 블러 처리된 이미지의 경우, 문서 내 텍스트 및 표의 경계 정보가 손실됨에 따라 시각적 어텐션이 문서 전반에 분산되는 경향을 보였다. 특히 정답이 포함된 표 영역에 대한 집중이 거의 이루어지지 않았으며, 그 결과 모델은 질의에 대해 실제 문서 내용과 무관한 값을 응답하는 모습을 보였다. 이는 입력 이미지 품질 저하로 인해 모델이 문서의 의미 정보를 충분히 인식하지 못하였음을 보인다.

원본 이미지에서는 주요 텍스트 및 표 영역에 대한 어텐션이 비교적 명확하게 형성되었으며, 질의에 대한 핵심 수치 정보를 올바르게 추출하는 데에는 성공하였다. 그러나 어텐션이 정답 영역뿐만 아니라 주변 텍스트 및 배경 영역에도 함께 분포하여, 응답에는 단위 정보와 같은 부가적인 문맥 요소가 포함되지 않는 한계가 관찰되었다.

반면, Upscaling 전처리를 적용한 경우 어텐션의 전반적인 공간 분포는 원본 이미지와 유사하게 유지되면서도, 정답이 포함된 표 셀과 해당 텍스트 영역에 대한 집중도가 더욱 명확하게 증가하는 경향을 보였다. 이러한 어텐션 집중도 향상은 생성 응답에도 직접적으로 반영되어, 단순한 수치 값뿐만 아니라 표에 명시된 단위 정보(m²)까지 함께 포함한 보다 완전한 형태의 응답이 생성되었다. 이는 해상도 향상 전처리가 텍스트 획과 표 구조의 가독성을 개선함으로써, 모델이 문서의 세부 의미를 보다 정확하게 이해하도록 돕는 역할을 수행했음을 의미한다.

이러한 결과는 입력 이미지 품질이 단순히 응답의 정답 여부에만 영향을 미치는 것이 아니라, 문서 내 정보의 맥락적 이해 수준과 응답의 완성도에도 중요한 영향을 미친다는 점을 보여준다. 특히 시각적 어텐션이 정답 영역에 안정적으로 집중될수록, 생성 응답 또한 보다 구체적이고 문서 맥락을 반영한 형태로 나타나는 경향을 확인할 수 있었다.

Ⅲ. 결론

본 연구에서는 문서 이미지 품질 변화가 비전-언어 모델의 시각적 어텐션 분포와 생성 응답 특성에 미치는 영향을 분석하였다. 실험 결과, 이미지 품질 열화는 어텐션의 확산과 응답 불안정성을 유발하는 반면, 해상도 향상 기반 전처리는 정답과 관련된 영역에 대한 어텐션 집중도를 높이고 응답의 일관성을 개선하는 효과를 보였다.

이러한 결과는 문서 기반 VLM 분야에서 입력 이미지 품질이 모델의 시각적 추론 과정과 응답 생성에 중요한 요소임을 시사한다. 특히 어텐션 안정성과 응답 품질 간의 밀접한 연관성을 확인함으로써, 단순한 데이터 확장이나 파인튜닝만으로는 VLM의 시각적 집중 메커니즘을 충분히 제어하기 어렵고 이는 성능 향상에 제약으로 작용할 수 있음을 보인다. 향후 연구에서는 내부 어텐션 맵을 활용하여 핵심 영역에 대한 집중을 유도하는 학습 전략을 통해 보다 안정적이고 해석 가능한 VLM을 설계하는 방향으로 연구를 확장할 계획이다.

ACKNOWLEDGMENT

본 연구는 산업통상자원부(MOTIE)와 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구 과제입니다. (No. RS-2022-KP002701)

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원에 의하여 이루어진 것임 (No. RS-2025-02223417)

참 고 문 헌

[1] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou., “LayoutLM: Pre-training of Text and Layout for Document Image Understanding,” PProceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)., 2020.

[2] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei., “LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,” Proceedings of the 30th ACM International Conference on Multimedia (MM '22), 2022.

[3] C. Tensmeyer and T. Martinez, “Document Image Binarization with Fully Convolutional Neural Networks,” ICDAR, 2017.

[4] Wang, Peng, et al., “Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.” ArXiv abs/2409.12191, 2024.