한국어 이미지 캡셔닝 향상을 위한 유창성 개선 모듈

유용상‡, 이기훈‡ ,임형준‡
롯데이노베이트
{4n3mone, kihoon.lee, hyoungjun.li}@lotte.net

Improving Korean Image Captioning with a Fluency Improvement Module

YongSang Yoo‡, KiHoon Lee‡, Hyoungjun Lim‡
Lotte Innovate

요 약

최근 Vision-Language Model(VLM)을 활용한 이미지 캡셔닝 연구가 활발히 진행되고 있으나, 대부분의 VLM은 이미지를 한국어로 디테일한 설명, 생성된 한국어의 다양한 맞춤법 처리 성능이 미흡한 상황이다. 이러한 문제를 해결하기 위해, 본 연구에서는 VLM이 생성한 이미지 설명을 보정하는 유창성 개선 모듈(FIM, Fluency Improvement Module)을 제안한다. FIM 기법은 VLM이 생성한 초기 설명을 sLLM(smaller Large Language Model)을 활용하여 한국어 문법에 맞게 재작성함으로써 보다 정확하고 자연스러운 이미지 캡션을 제공한다. 제안된 FIM 기법은 IC2024 데이터셋에서 기존 방법에 비해 최대 57.11%의 성능 향상을 보이며, 다양한 이미지 캡셔닝 응용 분야에 적용 가능하고 효율적인 자원으로 효과적인 보정이 가능함을 확인하였다.

주제어: 이미지 캡셔닝, 시각 언어 모델, 텍스트 교정
‡These authors contributed equally to this work

1. 서론

최근 멀티모달 모델이 급속히 발전하면서, 이미지를 텍스트로 설명하는 이미지 캡셔닝(Image Captioning)[1] 기술이 중요한 연구 주제로 부상하고 있다. 이미지 캡셔닝 기술은 의료 이미지 주석이나 산업군에서의 품질 관리, 교통 관리, 시각 정보를 활용하는 인공지능 챗봇 등 다양한 분야에 접목될 수 있다. 이러한 이미지 캡셔닝은 일반적으로 Vision-Language Model(VLM)[2]과 같은 사전 학습된 모델을 활용하여 이루어지며, VLM은 이미지와 텍스트 간의 관계를 학습하여 주어진 이미지에 적절한 설명을 생성하는 데 탁월한 성능을 보인다.

대부분의 VLM은 영어,중국어와 같은 주류 언어에 최적화되어 있기 때문에 한국어와 같이 소수가 사용하는 언어에 대해서는 성능이 미흡하다[3]. 이는 한국어를 사용하는 다양한 응용 프로그램에서 정확하고 자연스러운 이미지 설명을 생성하는 데 한계를 초래한다. 특히 한국어의 고유한 문법 구조와 어휘적 특성을 충분히 반영하지 못하는 문제점이 있으며, 이는 VLM이 생성한 캡션이 종종 부정확하거나 어색한 문장으로 이어질 수 있음을 의미한다.

이러한 문제를 해결하기 위해, 본 연구에서는 VLM에서 생성된 초기 이미지 설명을 smaller Large Language Model(sLLM)[4]을 활용하여 재작성하는 유창성 개선 모듈(FIM, Fluency Improvement Module)을 제안한다. FIM 기법은 VLM이 생성한 초안을 기반으로 sLLM이 보다 자연스럽고 한국어 문법에 맞는 최종 캡션을 생성하도록 돕는다. 이를 통해 VLM의 한계를 보완하고, 보다 정확하고 유창한 한국어 이미지 설명을 제공할 수 있다.

본 논문에서 제안한 FIM 기법은 IC2024 데이터셋에서 기

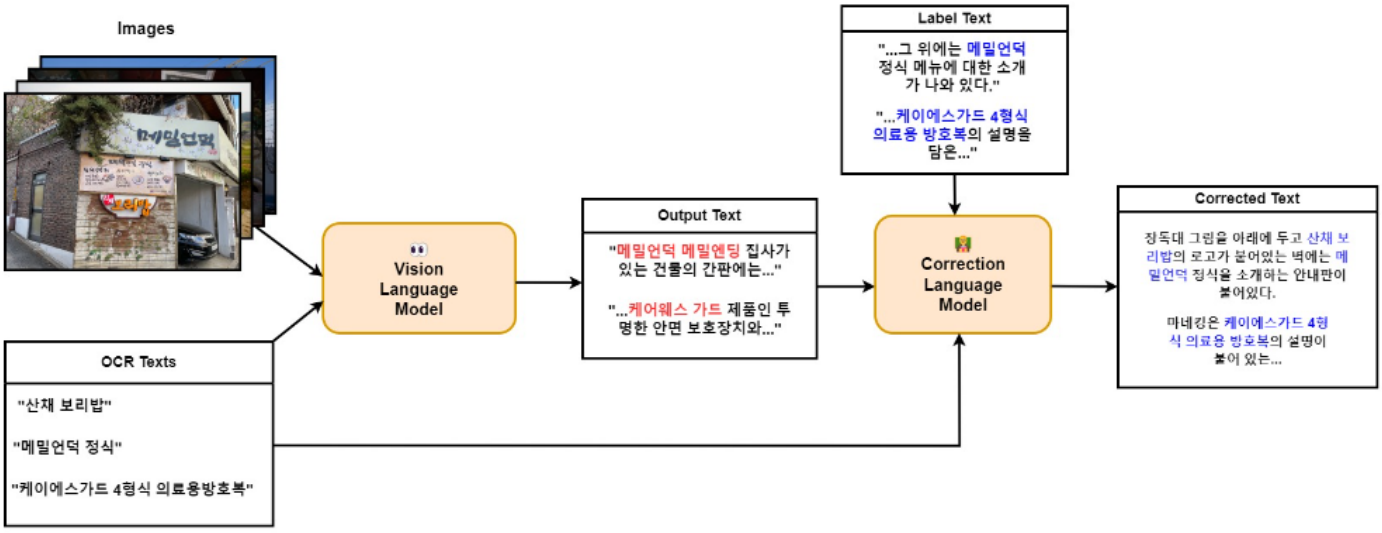

그림 1 : 유창성 개선 모듈을 포함한 이미지캡셔닝 진행 과정

존의 접근 방식에 비해 최대 57.11% 성능 향상을 보였다. 본 기술은 다양한 이미지 캡셔닝 응용 분야에 적용될 수 있으며 효율적인 자원으로 효과적인 보정이 가능함을 검증한다.

2. 관련 연구

2.1. 시각 언어 모델(VLM)

시각-언어 모델은 시각정 정보와 언어적 정보를 결합하여 학습한 모델을 의미한다. 이전까지의 시각-언어 모델은 단순히 단일 모달리티 모델을 결합하여 사용하거나, 이미지 특징 벡터를 학습한 LLM을 사용하여 사용자의 질의에 짧은 답을 생성할 수 있는 모델[5,6]이 주를 이루었다. 최근의 시각-언어 모델은 시각 정보를 포함한 지시문 데이터셋으로 학습한 시각-언어 모델들이 대거 등장[7,8,9,10]하였으며 본 학습 방식의 우수성을 검증했다. 한국어 지시문을 이해할 수 있는 시각-언어 모델은 한국어 LLM[11]을 언어 디코더로 사용한 KoLLaVA[12]와 다중 언어 데이터셋으로 학습한 언어 디코더를 사용하는 paligemma[13], idefics3[14] 등이 있다. 하지만 현재 공개된 모델들은 실질적인 한국어 이해 및 생성 능력이 부족한 것이 실증이며, 한국어 데이터를 주로 학습하여 한국인 사용자의 질의에 적절한 응답을 생성할 수 있는 시각 언어 모델의 개발이 필요한 상황이다.

2.2. 이미지 캡셔닝(Image Captioning)

이미지 캡셔닝은 컴퓨터가 시각적 데이터를 바탕으로 해당 내용에 대한 자연어 문장을 생성하는 과제이다[15]. 이 과제는 시각적 데이터와 자연어 데이터를 연결하는 작업으로, 두 가지 범주의 데이터에 대한 깊은 이해가 요구된다. 이미지 캡셔닝의 기본 구조는 주로 인코더-디코더(Encoder-Decoder) 방식으로 이루어진다. 이 구조에서 인코더는 이미지와 같은 시각 정보를 처리하고, 디코더는 이를 바탕으로 자연어 문장을 생성한다.

기존의 이미지 캡셔닝 연구들은 주로 COCO Captions[16], SCICAP[17]과 같은 데이터셋을 학습하여 이미지의 내용을 한 문장으로 간단히 설명하는 방식에 집중되었다. 이러한 초기 연구들은 단순한 이미지 설명에 그쳤으나, 모델 구조의 발전과 데이터셋의 다양화로 인해 이미지 캡셔닝의 응용 범위가 크게 확장되었다. 예를 들어, DocVQA[18]와 같은 문서 이미지 설명 및 분석 작업에서는 텍스트 인식뿐만 아니라 문서의 레이아웃과 문맥적 의미를 파악하는 능력이 요구된다. 또한, 연역적 추론이 가능한 이미지 캡셔닝 모델은 단순한 설명을 넘어서서, 이미지의 숨겨진 의미나 상호 연관된 요소들을 추론하는 데 사용된다[19]. 하지만 한국어 기반의 이미지 캡셔닝은 아직 초기 단계에 머물러 있다. 또한 공개된 한국어 멀티모달 데이터셋은 영어 기반 데이터셋 대비 아주 미비한 수준이다. 따라서 이미지 내에서의 한국어 인식 능력이 향상되어야 하며, 자연스러운 한국어 기반의 이미지 캡셔닝이 필요한 상황이다.

3. 유창성 개선 모듈 기법

본 연구에서는 한국어 이미지 캡셔닝의 성능을 향상시키기 위해 유창성 개선 모듈(Fluency Improvement Module,

이하 FIM)을 제안한다. FIM은 기존의 VLM이 생성한 초기 이미지 설명을 보완하여, 문법적으로 정확하고 자연스러운 한국어 주석을 생성하는 것을 목표로 한다. 특히 한국어의 문법적 복잡성과 다양한 표현 방식을 고려할 때, 학습을 제외한 다른 일반적인 접근 방식으로는 자연스러운 캡셔닝을 작성하는 데 한계가 있음을 인식하고, 이를 극복하기 위해 추가적인 모듈을 설계하였다.

본 논문에서 제안하는 FIM 방안은 크게 세 단계로 이루어지며, 전체 구조는 그림 1과 같다. FIM 방안은 시각 인식 단계, 비전 언어 통합 단계 , 정밀 보정 단계로 차례대로 진행된다.

첫 번째 단계인 시각 인식 단계에서는 이미지와 해당 이미지 내의 텍스트 정보를 추출하는 OCR(Optical Character Recognition) 기법을 사용하여 텍스트 데이터를 추출한다. 추출된 텍스트 데이터는 리스트 형식으로 이미지 데이터와 함께 다음 단계로 넘어간다.

두번째 단계인 비전 언어 통합단계에서는 추출된 이미지와 텍스트 데이터를 VLM에 입력한다. VLM은 이를 바탕으로 초기 이미지 캡션을 생성한다. 이때 생성된 초기 캡션은 종종 문법적으로 부정확하거나 부자연스러운 표현을 포함할 수 있다. 이러한 문제점을 해결하기 위해, 세번째 단계인 정밀 보정 단계를 수행한다.

정밀 보정 단계는 본 논문에서 제안하는 FIM 기법을 적용하는 단계로, 초기 캡션을 보다 자연스럽고 문법적으로 정확한 문장으로 수정한다. 정밀 보정 단계에서 사용되는 FIM은 VLM이 생성한 초기 캡션을 입력으로 사용하며, 목표하는 최종 문장을 라벨로 설정하여 학습을 진행한다. 이 과정에서 보정 모델은 초기 캡션의 문법적 오류를 수정하고, 의미적으로 더 적절한 표현으로 변환하는 능력을 습득하게 된다. 또한 OCR 정보를 통해 누락된 단어를 추가하는 역할을 한다.

효율적인 자원으로 복잡한 한국어 문법을 교정하기 위해 4비트로 양자화된 Qwen-2 모델[4]을 사용하였다. 또한 unsloth 라이브러리를 사용하여 QLoRA 경량화 학습 및 추론을 진행하여 적은 리소스로 매우 빠르고 효과적인 보정이 가능하다.

본 방안은 한국어 이미지 캡셔닝의 품질을 전반적으로 향상시키며, 다양한 실제 응용 시나리오에서의 활용 가능성이 높다고 판단된다. 특히, 문법적으로 정확하고 자연스러운 문장 생성을 통해 한국어 캡션의 질을 높이는 데 기여할 수 있다.

4. 실험 및 결과

본 실험은 Linux 환경에서 NVIDIA A100 80GB GPU 1대를 활용하여 수행되었다. 기반 모델로는 PaliGemma-3b-pt-869와 KoLLaVA를 채택하였으며, PaliGemma의 경우 896x896 해상도의 입력 이미지 처리와 512 토큰의 입력 및 출력 텍스트 시퀀스를 지원하는 사전 학습된 모델이다. PaliGemma-3b-pt-869 모델은 단일 턴 작업에 최적화되어 있으며, 다양한 과제에서 미세 조정 후 우수한 성능[13]을 보이는 것으로 알려져 있다. KoLLaVA의 경우 기존의 이미지 캡셔닝 모델인 LLaVA를 한국어 데이터로 추가 학습시킨 모델이다. 두 모델 모두 본 연구의 목적에 부합한다고 판단하고 실험을 진행하였다. 두 모델 모두 OCR 정보의 유무 및 보정 모델의 유무를 구분하여 진행하여 최적의 FIM을 만드는 것을 목적으로 하였다.

4.1 성능 평가

Model	Rouge-1	Rouge-L	BLEU	Average
Baseline	31.80	24.08	30.36	28.75
PaliGemma-3b-pt-869	37.26	29.31	32.48	33.02
+OCR	38.39	29.90	33.95	34.08
+ Correction Model	44.83	34.36	41.07	40.09
+OCR + Correction Model	**46.28**	**35.04**	**43.21**	**41.51**

표 1: Paligemma-3b 기반 IC2024에서의 모듈별 성능 비교

Model	Rouge-1	Rouge-L	BLEU	Average
KoLLaVA	28.48	23.29	27.01	26.26
+OCR	48.28	39.40	43.29	43.66
+Correction Model	40.54	33.83	33.09	35.82
+OCR +Correction Model	**51.52**	**40.02**	**46.41**	**45.98**

표 2: KoLLaVA 기반 IC2024에서의 모듈별 성능 비교

실험 결과는 표 1, 2와 같으며, IC2024 데이터셋을 기반으로 평가되었다. 평가는 생성한 이미지의 캡셔닝과 IC2024에서 제공하는 평가셋을 Rouge-1[20], Rouge-L, BLEU[21]의 평균을 기준으로 하고 있다. PaliGemma-3b-pt-869, KoLLaVA 각각의 모델에 이미지만을 입력

- (교정 전): 미화원 탈의실이라고 적힌 팻말이 붙어 있는 곳은 탈의실 입구 옆이다.

↓

- **(교정 후):** 미화원 탈의실이라고 적힌 파란색 안내문이 부착되어 있는 곳은 회색 문이다.

- (교정 전): 노란색 입간판에는 명랑 쌀 핫도그라고 적혀 있다.

↓

- **(교정 후):** 명랑 쌀 핫도그라고 적힌 간판은 가게의 상단에 걸려 있고, 그 아래에는 식당 내부를 볼 수 있다.

그림 2: IC2024 평가 데이터에서의 FIM 기법 차이 비교

으로 제공하고 미세 조정을 수행한 결과, 각각 33.02, 26.26점의 성능을 기록하였다. 두 모델 모두 시각적 인식을 할 수 있는 이미지 인코더와 LLM인 언어 디코더를 결합하여 사용하는 방식으로 기존 베이스 라인 보다 우수한 성능을 확인할 수 있다. 본 논문에서 제안한 유창성 개선 모듈(FIM)을 적용한 결과 성능이 각각 41.51, 45.98점으로 향상되는 것을 확인할 수 있었다. 이는 단순 VLM 학습 및 추론 대비 약 25.71%, 57.11% 추가 성능 향상을 의미한다. 특히, 정성적으로 분석했을 경우에도 FIM 기법은 한국어 문법 오류와 문맥적 일관성 능력을 효과적으로 교정하고 보다 자연스러운 표현으로의 전환 능력을 크게 향상시켰다.

4.2 효율성 분석

unsloth 기반의 학습 및 추론 기법 적용 결과, 제한된 컴퓨팅 리소스 환경에서도 신속하고 효과적인 보정 작업 수행이 가능했다. 실험 결과, 기존의 미세 조정된 VLM 대비 VRAM 사용량이 약 9GB 증가하고, 처리 속도가 건당 약 1초 증가하였으나, 약 21.6% 이상의 성능 향상을 달성하였다. OCR 정보 활용 시 텍스트 위치 정보를 입력하여 학습을 진행하였으나, 오히려 성능 저하가 관찰되어 위치 값은 학습에서 제외하였다.

본 연구에서 제안한 방법론의 일반화 가능성을 검증하기 위해, KoLLAVA 모델을 기반으로 한 추가 실험을 수행하였다. 표 2는 KoLLAVA에 FIM 적용 유무에 따른 성능 차이를 나타낸다. 표 1과 동일한 실험 조건에서, 이미 높은 이미지 캡셔닝 성능을 보이는 모델에도 FIM 적용 시 추가적인 성능 향상이 관찰되었다. OCR 정보를 포함한 보정 모델 사용 시, 기존 대비 5.3% 향상된 45.98점을 기록하며 IC2024 작업에서 최고 성능(State-of-the-Art)을 달성하였다.

4.3 정성적 생성 문장 비교

그림 2는 FIM 기법 적용 전후의 결과를 비교 분석한다. 우리가 제안한 방식은 주어진 이미지의 맥락을 정확히 파악하면서 OCR 결과를 포함하는 문장을 생성하는 것을 확인할 수 있다. 어 가독성과 이해도가 향상된 문장으로 재구성되었다. OCR 정보는 보정 모듈 및 시각-언어 모델에서 중요한 역할을 수행하였으며, 누락된 단어나 부정확한 문구를 교정하는 데 크게 기여하였다. 이로 인해 최종 캡션의 정확성과 자연스러움이 현저히 개선되었다.

5. 결론

본 논문에서는 Vision-Language Model(VLM)이 생성하는 한국어 이미지 캡션의 품질을 향상시키기 위해 유창성 향상 모듈(FIM)을 제안하였다. FIM 기법은 기존 VLM이 생성한 초기 이미지 설명을 sLLM(smaller Large Language Model)을 활용해 문

법적으로 정확하고 자연스러운 한국어로 교정함으로써, 이미지 캡셔닝의 성능을 효과적으로 개선할 수 있음을 보였다.

IC2024 데이터셋을 사용한 실험 결과, 본 논문에서 제안하는 FIM 기법은 기존 방법 대비 최대 57.11%의 성능 향상을 달성하였다. 특히 한국어의 복잡한 문법 구조와 다양한 표현 방식을 고려한 보정 과정이 최종 이미지 캡션의 자연스러움과 정확성을 높이는 데 중요한 역할을 하는 것을 검증하였다. 또한, 본 기법은 효율적인 자원 사용으로 다양한 응용 분야에서 활용 가능하며, 한국어 이미지 캡셔닝 기술의 발전에 기여할 수 있을 것으로 기대된다.

참고문헌

[1] Ke, L., Pei, W., Li, R., Shen, X., & Tai, Y. W. (2019). Reflective decoding network for image captioning. Proceedings of the IEEE/CVF International Conference on Computer Vision, 8888-8897.

[2] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2018). Multi-modal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[3] Romero, D., Lyu, C., Wibowo, H. A., Lynn, T., Hamed, I., Kishore, A. N., ... & Aji, A. F. (2024). CVQA: Culturally-diverse multilingual visual question answering benchmark. arXiv preprint, arXiv:2406.05967.

[4] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., ... & Fan, Z. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

[5] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. International conference on machine learning, PMLR, 19730-19742.

[6] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, 23716-23736.

[7] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Thirty-seventh Conference on Neural Information Processing Systems.

[8] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Thirty-seventh Conference on Neural Information Processing Systems.

[9] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966.

[10] Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision- language understanding with advanced large language models. arXiv preprint arXiv :2304.10592.

[11] Maywell. (2023). Synatra-7B- v0.3-dpo. Hugging Face. https://huggingface.co/maywell/S ynatra-7B-v0.3 -dpo.

[12] Tabtoyou. (2023). KoLLaVA: Korean Large Language-and -Vision Assistant (feat. LLaVA). GitHub.github.c om/tabtoyou/KoLLaVA.

[13] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., ... & Zhai, X. (2024). PaliGemma: A versatile 3B VLM for transfer. arXiv preprint arXiv:2407.07726.

[14] Laurencon, H., Marafioti, A., Sanh, V., & Tronchon, L. (2024). Building and better understanding vision-language models: insights and future directions. arXiv preprint arXiv:2408.12637.

[15] Ke, L., Pei, W., Li, R., Shen, X., & Tai, Y. W. (2019). Reflective decoding network for image captioning. Proceedings of the IEEE/CVF international conference on computer vision, 8888-8897.

[16] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[17] Hsu, T. Y., Giles, C. L., & Huang, T. H. K. (2021). SciCap: Generating captions for scientific figures. arXiv preprint arXiv:2110.11624.

[18] Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). Docvqa: A dataset for vqa on document images. Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2200-2209.

[19] Zhang, Y., Bai, H., ZHANG, R., Gu, J., Zhai, S., Susskind, J. M., & Jaitly, N. (2024). How Far Are We from Intelligent Visual Deductive Reasoning?. ICLR 2024 Workshop: How Far Are We From AGI.

[20] Lin, Chin-Yew.(2004), "Rouge: A package for auto matic evaluation of summaries." Text sum marization branches out.

[21] Papineni, Kishore, et al., (2002), "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

감사의 글

본 연구는 롯데이노베이트의 자원과 지원을 통해 수행되었습니다. 또한 본 논문은 중앙대학교 첨단영상대학원 최종원 교수의 지도를 받아 작성되었습니다.