ERP 파노라마 왜곡 및 분절 완화를 위한 E2P 오버랩 전처리와 VICReg 결합 파인튜닝 기반 캡셔닝 성능 향상

우승우, 김지원, 염세경
동국대학교

woow0708@dgu.ac.kr, kimji1net@dgu.ac.kr, skyoum@dongguk.edu

Improving Captioning via Overlapped E2P Preprocessing and VICReg-Regularized Fine-tuning for ERP Images

Woo Seung Woo, Ji Won Kim, Youm Se Kyoung
Dongguk Univ.

요 약

본 논문은 본 논문은 ERP 기반 파노라마 이미지의 극점 왜곡과 다중 뷰 분해로 인한 분절 문제를 완화하기 위해, Equirectangular-to-Pinhole(E2P) 다중 뷰 변환과 인접 뷰 오버랩 전처리를 제안한다. 또한 캡셔닝 품질과 시각 표현의 일관성을 동시에 강화하기 위해 Auto-Regressive 생성 손실과 VICReg 기반 정규화 손실을 결합한 파인튜닝 전략을 설계하였다. QUIC-360 데이터셋에서 BLIP-2 를 베이스라인으로 실험한 결과, 제안 방법은 단순 리사이즈 기반 파인튜닝 대비 BLEU-4, METEOR, ROUGE-L, CIDEr 전 지표에서 성능 향상을 보였다..

Ⅰ. 서 론

파노라마 이미지는 수평 약 360°, 수직 최대 180°의 넓은 화각을 제공하여 로보틱스, 스마트홈 등에서 활용이 확대되고 있다. 그러나 파노라마는 구면 상 장면 정보를 등장방형도법(Equirectangular Projection, ERP)으로 저장하는 과정에서 극점 중심의 왜곡을 수반하며, 이는 일반적인 2D 이미지 분포를 가정하는 시각 모델의 표현 학습에 불리하게 작용한다.

한편, 시각-언어 모델(Vision-Language Model, VLM)은 이미지 캡셔닝을 포함한 다양한 멀티모달 과제에서 높은 잠재력을 보이며 빠르게 발전해왔다. [3, 4] 그럼에도 불구하고, 파노라마 이미지는 ERP 특유의 극점 왜곡과 전역 연속성(Continuous Spatial Context) 처리의 어려움으로 인해 VLM 기반 캡셔닝에서 활용이 제한적이다. 기존 연구에서는 ERP 이미지를 큐브맵(Cubemap)으로 변환하여 처리하는 접근이 제안되어 왔으나, 이 방식은 파노라마 장면을 하나의 연속된 공간으로 모델링하기보다 여러 개의 분절된 뷰(View)로 분해하여 독립적으로 인식하는 경향이 있어, 파노라마의 전역 맥락을 충분히 반영하지 못한다. 결과적으로 큐브맵 기반 처리는 다수의 부분 이미지를 처리하는 방식에 가깝다는 한계를 갖는다.

따라서 본 연구는 ERP 기반 파노라마 이미지가 갖는 왜곡과 뷰 분절 문제를 완화하고, 전역 맥락을 보존한 채 VLM 의 이미지 캡셔닝 성능을 향상시키기 위한 방법을 제안한다.

Ⅱ. 본론

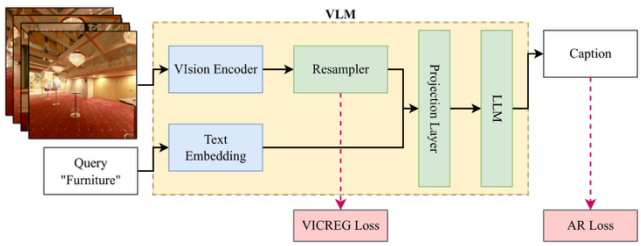

Figure 1: VLM Training Method Overview

본 논문에서는 표준 VLM[3]의 입출력 구조를 기반으로 파노라마 이미지에 적합한 전처리와 파인튜닝 전략을 설계하였다. VLM 은 사전학습 비전 인코더로 이미지 특징을 추출, 리샘플링한 뒤, 텍스트 임베딩과 결합하여 LLM 입력으로 정렬하고, 이를 조건으로 캡션을 생성한다.

E2P 는 구면 영상에서 중심 방향과 화각에 해당하는 영역을 단일 평면 뷰(pinhole view)로 근사하는 방식으로, ERP 에서 발생하는 왜곡을 효과적으로 줄일 수 있다. 다만 E2P 변환은 본질적으로 하나의 시야에 대한 단일 뷰만을 생성하므로, 파노라마 장면의 전역 정보를 포괄적으로 활용하기 위해서는 서로 다른 중심 방향을 기준으로 E2P 를 반복 적용하여 다중 뷰(view set)로 변환하는 과정이 필요하다. 그러나 이와 같이 파노라마를 다중 뷰로 분해하면 뷰 경계에서 불연속이 발생할 수

있으며, 결과적으로 전역 맥락이 단절되는 분절 문제가 남는다.

이러한 분절 문제를 완화하기 위해 인접 뷰 간 오버랩(Overlap)을 도입하여, 각 뷰가 일정 비율의 중복 영역을 공유하도록 구성한다. 오버랩은 뷰 경계에서의 정보 손실을 줄이고, 모델이 인접 뷰 사이의 연속성을 학습할 수 있도록 돕는다. 더 나아가 파인튜닝 단계에서는 생성(캡셔닝) 품질과 시각 표현의 일관성을 동시에 강화하기 위해, 표준 Auto-Regressive(AR) 생성 손실과 VICReg[4] 기반 정규화 손실을 결합하여 최적화한다. 이때 VICReg 기반 손실은 이미지 인코딩(시각 특징 추출) 단계에 대해서만 그래디언트를 적용하여 표현 공간의 안정성과 불변성을 유도하고, AR 생성 손실은 출력 토큰에서부터 비전 인코더를 포함한 전체 경로로 그래디언트를 전파함으로써 최종 캡션 생성 성능을 직접적으로 최적화한다.

Figure 2: E2P-Overlap Method

실험은 QUIC-360[5] 데이터셋의 Test Split 에서 수행하였으며, 쿼리 조건부 캡셔닝 성능을 평가하였다. 베이스라인은 BLIP-2 의 공개 사전학습 가중치로 초기화한 후 파인튜닝하였고, 입력 전처리는 ERP 단순 리사이즈(Resize)와 제안하는 E2P 다중 뷰와 오버랩 전처리 및 VICReg 결합 학습으로 구성하여 비교하였다. 성능 평가는 BLEU-4, METEOR, ROUGE-L, CIDEr 지표를 사용하였으며, 정량적으로 검증하였으며, 성능 지표는 백분율로 표기되었다.

Table 1: Results in QUIC-360

	BLEU-4	METEOR	ROUGE-L	CIDEr
Resize	00.02	04.20	07.13	10.80
E2P	**01.26**	**12.98**	**16.20**	**12.81**

제안하는 방법(w. VICReg)은 기존 BLIP-2 의 전처리 및 파인튜닝 방식 대비 BLEU-4 에서 6,200% 개선을 보였으며, METEOR(209.05%), ROUGE-L(127.21%), CIDEr(18.61%)에서도 성능 향상을 확인하였다. 이러한 개선은 단순 리사이즈 입력이 고해상도 ERP 파노라마의 세부 단서를 크게 손실시켜 시각 토큰 표현이 약화되고, 결과적으로 QUIC-360 의 쿼리 조건부 캡셔닝에서 정렬(Alignment)이 저하되는 문제와 연관된다. 반면 제안 방법은 E2P 다중 뷰와 오버랩으로 왜곡 및 경계 단절을 완화하고, VICReg 정규화를 통해 뷰 변화에 대해 일관된 표현을 유도함으로써 시각- 언어 정렬을 강화해 성능 향상에 기여한다.

Ⅲ. 결론

본 연구는 표준 VLM 구조를 유지하면서 ERP 기반 파노라마 이미지에서 발생하는 극점 왜곡과 다중 뷰 분해로 인한 분절 문제를 완화하기 위해, E2P 다중 뷰 변환과 인접 뷰 오버랩 전처리, 그리고 AR 생성 손실과 VICReg 정규화 손실을 결합한 파인튜닝 전략을 제안하였다. QUIC-360 데이터셋에서 BLIP-2 를 기반으로 평가한 결과, 제안 방법은 단순 리사이즈 입력 대비 BLEU-4, METEOR, ROUGE-L, CIDEr 전 지표에서 성능 향상을 보여 파노라마 캡셔닝에서 전처리- 학습 결합 설계의 효과를 확인하였다. 향후에는 전처리 하이퍼파라미터의 최적화와 계산 비용과 성능의 트레이드오프 분석, 그리고 구면 연속성을 더 직접적으로 반영하는 토큰 결합/인코딩 방식으로의 확장을 통해 일반화 성능과 효율성을 추가로 고도화할 예정이다.

참 고 문 헌

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

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," Advances in neural information processing systems, vol. 36, pp. 34892- 34916, 2023.

[3] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 46, no. 8, pp. 5625- 5644, 2024.

[4] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," arXiv preprint arXiv:2105.04906, 2021.

[5] K. Maeda, S. Kurita, T. Miyanishi, and N. Okazaki, "Query-based image captioning from multi-context 360cdegree images," in Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 6940- 6954.