BLIP 기반 Knowledge Distillation 을 활용한 멀티모달 이미지 캡셔닝 모델의 경량화 및 성능 개선 연구

이승헌, 전윤모, 김웅섭*
동국대학교

leesh914@dgu.ac.kr, jumo0716@dongguk.edu, *woongsup@dongguk.edu

A Study on the Lightweighting and Performance Optimization of Multimodal Image Captioning Models via BLIP-Based Knowledge Distillation

SeungHeon Lee, YoonMo Jeon, Woongsup Kim*
Department of Information Communication Engineering, Dongguk Univ

요 약

본 연구에서 연산 자원이 제한된 엣지 디바이스에서도 효율적으로 동작할 수 있는 멀티모달 이미지 캡셔닝 경량화 모델을 설계하고자 한다. 이를 위해, 고성능 BLIP Large 모델의 출력 결과를 지도 신호로 활용하는 Knowledge Distillation 기반 Supervised Fine-tuning 기법을 적용하여, BLIP Base 모델을 효과적으로 학습시켰다. 학습 후에는 FP16 양자화를 통해 연산 효율성과 메모리 사용량을 추가로 최적화하였다. 5,000 장의 재난 이미지 데이터를 활용한 실험 결과, 제안된 모델은 CLIP Similarity 기반 정합성 평가와 Raspberry Pi 5 환경에서의 추론 성능 테스트에서 기존 모델 대비 우수한 성능을 기록하였으며, 엣지 디바이스 환경에서도 실용 가능한 멀티모달 모델 구조임을 입증하였다.

Ⅰ. 서 론

딥러닝 기반 멀티모달 모델은 이미지와 텍스트를 동시에 처리할 수 있어 다양한 작업에 활용되고 있으며, 특히 이미지 캡셔닝(image captioning)은 자율주행, 영상 요약, 접근성 지원 등에서 중요한 응용 분야로 주목받고 있다. 하지만 최신 고성능 모델들은 수십억 개 이상의 파라미터를 포함하고 있어, 연산 자원이 제한된 엣지 디바이스 환경에서는 실질적인 적용이 어렵다는 한계가 존재한다.

본 연구의 목적은 이러한 한계를 극복하고, 엣지 환경에서도 실행 가능한 경량 멀티모달 이미지 캡셔닝 모델을 구현하는 데 있다. 이를 위해 BLIP Large 모델을 Teacher 로, BLIP Base 모델을 Student 로 설정하고, Knowledge Distillation 기반 Supervised Fine-tuning 기법을 통해 Student 모델이 Teacher 의 표현력을 효과적으로 학습하도록 구성하였다[1]. 고성능 모델인 BLIP Large 는 이미지 캡셔닝, VQA, 이미지-텍스트 매칭 등 다양한 멀티모달 태스크에서 우수한 성능을 보이나, 연산량과 메모리 요구사항이 커 엣지 디바이스에서의 실용성은 제한적이다. 이에 반해 BLIP Base 는 연산 효율성과 모델 크기 측면에서 보다 경량화된 구조로, 디바이스 내 탑재 및 실시간 응용에 적합하다.

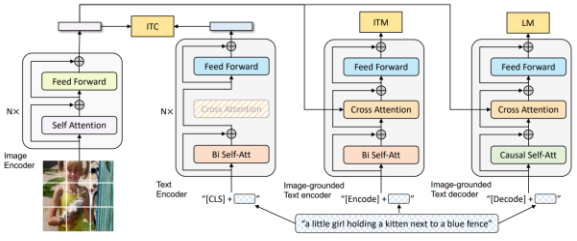
본 연구에서는, BLIP Large 가 생성한 고품질 캡션을 hard label 형태로 직접 지도학습에 활용함으로써 구조적 단순성과 학습 효율성을 동시에 확보하였다. 학습이 완료된 후에는 FP16 양자화를 적용하여 모델의 파라미터와 연산량을 효과적으로 줄였으며, 최종 모델은 Raspberry Pi 5 환경에서의 실험을 통해 추론 속도, 메모리 사용량, 정합성 등의 측면에서 BLIP Large 대비 경쟁력 있는 성능을 확보하였다.

본 연구는 멀티모달 모델의 실용적 경량화 가능성을 실험적으로 검증하였다는 점에서 의의가 있다.

Ⅱ. 본론

2.1 모델 구조 및 학습 방법

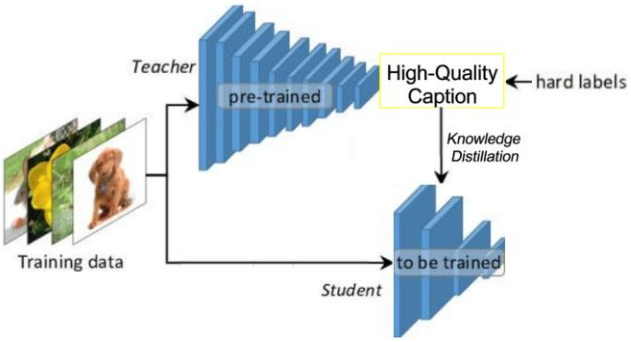
본 논문에서는 이미지와 텍스트의 멀티모달 입력을 동시에 처리할 수 있는 BLIP(Bootstrapping Language-Image Pre-training) 기반 구조를 도입하였다[2].



[그림 1] BLIP 모델의 구조

본 연구는 BLIP Large 의 출력 결과를 지도 신호로 활용하여, 경량화된 BLIP Base 모델을 지도학습(Supervised Fine-tuning) 방식으로 학습시켰다. 이 과정은 전통적인 Knowledge Distillation 에서 사용하는 soft label 이나 temperature scaling 같은 손실 기반

기법을 사용하지 않고, BLIP Large 가 생성한 캡션 결과를 hard label 로 간주하여 직접 학습에 활용하는 구조로 설계되었다. 이는 별도의 복잡한 손실 함수 설계 없이도 Teacher 모델의 표현 능력을 Student 가 자연스럽게 학습할 수 있도록 하며, 구조적으로 단순하면서도 효과적인 성능 향상을 기대할 수 있다. 특히, 실제 label 확보가 어려운 멀티모달 데이터셋 환경에서 활용도가 높은 전략이다.



[그림 2] BLIP-Base 모델의 학습과정

학습에는 총 5,000 장의 재난 이미지(화재, 구조, 붕괴 등)를 활용하였으며, 각 이미지에 대해 생성된 Teacher 의 캡션을 정제된 학습 데이터로 사용하였다. 학습이 완료된 이후, FP16(Floating Point 16-bit) 양자화 기법을 적용하여 모델의 weight 와 activation 을 32-bit 에서 16-bit 로 변환함으로써, 메모리 사용량과 연산량을 약 50% 절감하였다. 이는 추론 속도 향상뿐만 아니라, 실제 엣지 디바이스 적용 가능성을 확보하는 데 중요한 역할을 하였다.

2.2 평가지표 및 실험결과

경량화된 BLIP Base 모델의 성능을 평가하기 위해 다음과 같은 환경과 지표를 사용하였다.

첫번째는, 정합성 평가 (Semantic Alignment Evaluation)로 이미지와 모델이 생성한 텍스트 캡션 간의 의미적 정합성을 정량적으로 평가하기 위해, OpenAI 의 CLIP 모델(vit-base-patch32)를 활용하였다. CLIP 은 이미지와 텍스트의 임베딩 공간에서의 유사도를 계산할 수 있는 멀티모달 모델로, 본 연구에서는 이미지-캡션 쌍 간의 cosine similarity score 를 통해 모델이 생성한 캡션의 의미적 타당성을 평가하였다. 본 평가는 총 100 장의 테스트 이미지를 기반으로 수행되었으며, 각 이미지에 대해 하나의 캡션을 생성하고 이에 대한 평균 CLIP Similarity 를 측정하였다.

두번째는, 추론 성능 평가 (Inference Performance Evaluation)로 모델의 실용성을 검증하기 위해, Raspberry Pi 5 (8GB RAM, Quad-core Cortex-A72 1.5GHz) 환경에서 실제 추론을 수행하였다. 모델 입력으로 동일한 100 장의 테스트 이미지를 사용하였으며, 각 이미지에 대해 텍스트 캡션을 생성하는 데 걸리는 평균 추론 시간(초)을 측정하여 비교하였다.

모델명	모델 크기(MB)	CLIP Similarity(%)	라즈베리파이 추론 시간(초)
BLIP-Large	1791.89	30.60	18.84
BLIP-Base	943.81	29.48	13.06
Ours	565.81	31.65	11.71

[그림 3] 성능 결과 비교

본 논문에서 제안한 모델(Ours)은 세 모델 중 가장 높은 경량화 수준(565.81MB)을 달성하면서도, 추론 속도와 정합성(CLIP Similarity) 모두에서 가장 우수한 성능을 보였다. 이는 BLIP Large 의 범용 학습 방식과 달리, 제안된 모델이 도메인 특화 학습을 통해 입력 데이터에 보다 정밀하게 적응한 결과로 해석된다.

마지막으로, 결과물을 정성적으로 분석한 결과, 생성된 캡션은 시각적 장면을 자연스럽고 일관성 있게 기술하고 있어, 인간 관점에서도 충분히 수용 가능한 수준의 품질을 보였다.

[그림 4] Ours, Base, Large 모델의 캡셔닝 결과 비교

Ⅲ. 결론

본 논문에서는 BLIP 기반의 Knowledge Distillation 기법을 활용하여, 연산 자원이 제한된 엣지 디바이스 환경에서도 실행 가능한 경량 멀티모달 이미지 캡셔닝 모델을 제안한다. BLIP Large 모델이 생성한 캡션을 hard label 로 활용하여 BLIP Base 모델을 지도학습 방식으로 학습시켰으며, 이후 FP16 양자화를 적용하여 연산 효율성과 메모리 사용량을 더욱 향상시켰다.

실험 결과, 제안한 모델은 정합성 평가(CLIP Similarity) 및 추론 속도 측면에서 기존 모델 대비 우수한 성능을 나타냈으며, 정성적 분석에서도 시각적 장면을 자연스럽고 일관성 있게 설명하는 캡션을 생성하여 실제 응용 가능성을 입증하였다. 특히, 지식 증류 기반의 경량화 전략은 복잡한 손실 함수 없이도 효과적인 성능을 구현할 수 있음을 보여준다.

결과적으로, 제안된 BLIP Base 기반의 경량 모델은 제한된 하드웨어 환경에서도 안정적인 추론 속도와 정합성을 유지하며, 멀티모달 모델의 실용적 경량화 방안으로서 의미 있는 사례를 제시한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로

정보통신기획평가원-학석사연계 ICT 핵심인재 양성 지원을

받아 수행한 연구임 (IITP-2024-00436744).

참 고 문 헌

[1] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network,” arXiv:1503.02531, Mar. 2015. https://doi.org/10.48550/arXiv.1503.02531

[2] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” International Conference on Machine Learning. PMLR, pp12888-12900 2022.