시각-언어 모델을 이용한 엣지-서버 간 협력 추론

송수창, 김용준*
포항공과대학교

{ssc6351, yongjune}@postech.ac.kr

Edge-Server Collaborative Inference Using Vision-Language Model

Soochang Song, Yongjune Kim*
POSTECH

요 약

스마트폰 등의 엣지 디바이스에서 수집된 데이터를 활용하여 시각-언어 모델(vision-language model, VLM)의 추론을 수행하기 위해서는 해당 데이터를 서버로 전송해야 하며, 이 과정에서 높은 통신 비용이 발생한다. 이러한 문제를 해결하기 위해 최근에는 시맨틱 통신(semantic communication)에 기반한 엣지-서버 간 협력 추론(collaborative inference)이 활발히 연구되고 있다. 협력 추론에서 엣지 디바이스는 중요한 의미를 담은 정보를 VLM 의 토큰 형태로 추출한 뒤, 이를 선택적으로 서버로 전송함으로써 추론 정확도를 유지하면서도 통신 비용을 효과적으로 절감할 수 있다. 본 논문에서는 이러한 협력 추론에서 사용되는 주요 기법을 소개하고, 그 적용 결과를 실험적으로 분석한다.

Ⅰ. 서론

최근 VLM 의 발전으로 이미지와 텍스트를 동시에 활용하는 객체 탐지, 이미지 분할, 이미지 캡셔닝(image captioning) 등의 다양한 인공지능 태스크가 가능해졌고 여러 분야에서 활용되고 있다[1]. 그러나 시각 언어 모델의 추론을 스마트폰과 같은 엣지 디바이스의 이미지 데이터를 이용해 수행하고자 할 경우, 저장된 데이터를 서버에 전송해야 하는데, 이 과정에서 상당한 양의 통신 자원이 요구된다. 이는 대역폭이 제한된 환경이나 에너지 효율이 중요한 모바일 환경에서 실용적인 문제로 작용한다. 이러한 문제를 해결하기 위해 최근에는 시맨틱 통신에 기반한 엣지 디바이스와 클라우드 서버 간 협력 추론이 활발하게 연구되고 있다[2]. 시각 언어 모델의 협력 추론에서 엣지 디바이스는 비전 트랜스포머(ViT) 기반의 경량 인코더가 탑재되어, 이미지로부터 시맨틱하게 중요한 정보만을 추출하고 이를 서버로 전송한다. 서버는 사전 학습된 LLM 디코더를 통해 해당 정보를 기반으로 추론을 수행하며, 최종 추론 결과를 엣지 디바이스로 전송한다. 본 논문에서는 이러한 협력 추론에서 이미지 정보를 효율적으로 추출하고 전송하기 위해 벡터 양자화(vector quantization) 방법을 사용하며, 제안한 방법의 추론 정확도와 통신 효율을 평가한다.

Ⅱ. 본론

본 논문은 이미지 캡셔닝을 위한 협력 추론에 중점을 두며 전체적인 프레임워크는 Fig. 1 에 제시되어 있다. 해당 프레임워크에서는 엣지 디바이스에 탑재된 경량화된 이미지 인코더를 활용하여 입력 이미지로부터 시각적 특징(feature)을 추출한다. 이후, 추출된 특징은 의미론적으로 중요한 정보를 선별하는 특징 선택(feature selection) 모듈을 통과하여, 토큰의 수를 효과적으로 감소시킨다. 이렇게 선택된 시각 정보는 무선 통신 채널을 통해 사용자의 텍스트 입력과 함께 서버로 전송된다. 서버에 위치한 LLM 디코더는 수신된 데이터를 기반으로 이미지에 대한 캡션을 생성하고, 생성된 캡션은 다시 엣지 디바이스로 전송되어 사용자에게 제공된다.

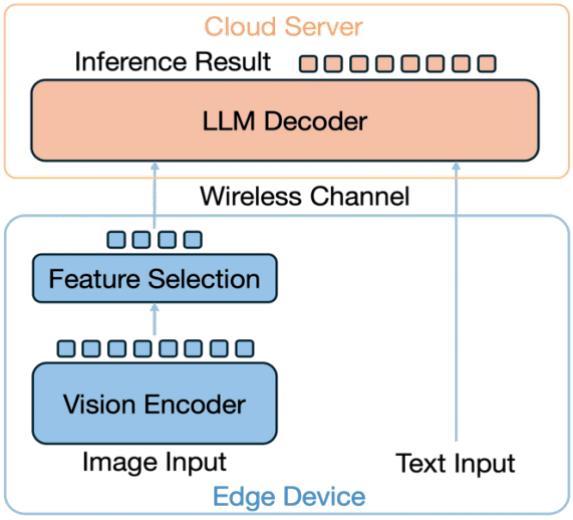

Fig. 1. 이미지 캡셔닝을 위한 시각 언어 모델 기반의 협력추론 프레임워크

협력 추론 프레임워크의 특징 선택 모듈에는 다양한 기법을 적용할 수 있으며, 본 연구에서는 k-평균 군집화(k-means clustering)기반의 벡터 양자화 방법을 활용한다. k-평균 군집화는 초기화 단계에서 임의로 k 개의 중심점을 설정한 뒤, 각 데이터 포인트와 중심점(centroid) 간의 거리를 계산하여 가장 가까운 중심점에 데이터를 할당하는 방식으로 동작한다.

$$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2$$

여기서 $c^{(i)}$는 데이터 포인트 $x^{(i)}$가 속한 군집의 레이블이고 μ_j는 j 번째 군집의 중심점 벡터이다. 이후 각 클러스터에 할당된 데이터 포인트들의 평균값을 이용하여 중심점을 갱신한다.

$$\mu_j := \frac{\sum_{i=1}^{n} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} 1\{c^{(i)} = j\}}$$

위의 과정을 중심점이 수렴할 때까지 반복함으로써 최종 군집 구조를 도출한다.

비전 인코더를 통과한 이미지 특징 벡터들에 대해 k-평균 군집화를 수행하여 시맨틱하게 유사한 특징 벡터들을 동일한 군집으로 묶을 수 있다. 이 때 군집의 중심점 벡터와 각각의 특징 벡터가 속한 군집의 인덱스를 무선 채널을 통해 전송함으로써 전송하는 데이터의 양을 효과적으로 줄일 수 있다. LLM 디코더에서는 전송받은 인덱스와 중심점 벡터를 바탕으로 원래의 특징 벡터를 대체하는 새로운 입력을 만들어 추론을 진행한다. Fig. 2 는 원본 이미지가 비전 인코더를 거친 후의 특징 벡터와, 벡터 양자화 과정을 거친 후의 특징 벡터에 대해 L2-Norm 값을 기반으로 시각화한 결과를 나타낸다. 하단의 두 그림에서는 동일한 색상으로 표시된 특징 벡터가 동일한 군집에 속함을 의미하며, 이는 해당 특징 벡터들이 시맨틱하게 유사한 정보를 포함하고 있음을 나타낸다.

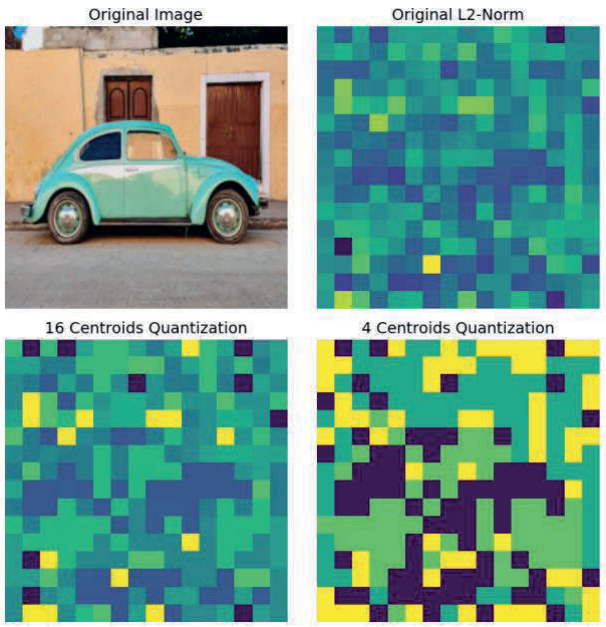

Fig. 2. 중심점의 갯수에 따른 이미지 특징의 시각화

Fig. 3 은 제안한 프레임워크에 대표적인 VLM 인 PaliGemma-3B[3] 모델을 사용했을 때 BLEU (bilingual evaluation understudy) 점수와 통신 효율 사이의 트레이드-오프를 나타낸다. 캡셔닝 결과의 정확도를 나타내는 BLEU 점수는 정답인 문장과 생성된 문장 사이에 일치하는 단어 조합의 갯수가 많을수록 높게 나타난다. 통신 효율은 전송한 중심점의 수를 전체 특징 벡터의 수로 나누어 구하였고 함께 전송한 인덱스는 통신량이 적어 계산에서 제외하였다. 결과를 보면, 무작위로 특징 벡터를 선택하여 전송한 경우보다, 벡터 양자화를 통해 각 군집의 중심점만을 전송한 경우에 LLM 기반 추론의 정확도가 더 높게 나타남을 확인할 수 있다. 특히 전체 이미지 특징 벡터 수의 약 3%에 해당하는 8 개의 중심점만을 전송하였을 때에도, 전체 특징을 모두 전송한 경우의 약 80%에 해당하는 추론 정확도를 달성하였다. 이는 무작위로 40%의 특징 벡터를 선택하여 전송했을 때와 유사한 수준의 성능으로, 군집 기반 중심점 선택이 시맨틱한 정보를 효과적으로 전달함을 의미한다.

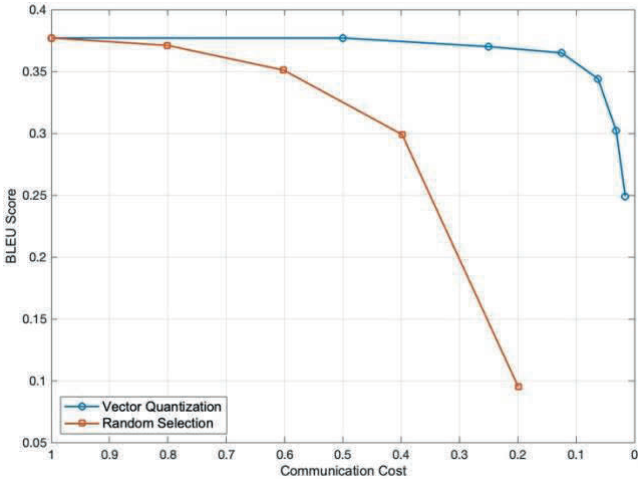

Fig. 3. 벡터 양자화를 이용한 협력 추론의 정확도 비교

Ⅲ. 결론

본 논문에서는 이미지 캡셔닝을 위한 VLM 기반의 협력 추론(collaborative inference) 프레임워크를 제안하고, 엣지 디바이스에서 경량의 비전 인코더와 k-평균 군집화를 활용하여 이미지 특징을 추출하는 방식의 성능을 분석하였다. 실험 결과, 전체 특징 벡터 대신 군집의 중심점과 인덱스만 전송해도 높은 수준의 추론 정확도를 유지할 수 있었다. 이는 엣지-서버 간 통신 효율을 크게 향상시키는 동시에 모델의 추론 성능을 효과적으로 유지할 수 있는 가능성을 보여준다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00212103).

참 고 문 헌

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625-5644, Aug. 2024.

[2] C. Yuan et al., "Task-oriented feature compression for multimodal understanding via device-edge co-inference," *arXiv preprint arXiv:2503.12926*, Mar, 2025.

[3] L. Beyer et al., "Paligemma: a versatile 3b VLM for transfer, " *arXiv preprint arXiv:2407.07726*, 2024.