일반논문 (Regular Paper)
방송공학회논문지 제30권 제5호, 2025년 9월 (JBE Vol.30, No.5, September 2025)
https://doi.org/10.5909/JBE.2025.30.5.784
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

경량화 전문가 혼합 모델 기반 위성 영상 분석 VQA 알고리즘

정 시 연[a)], 권 오 설[a)‡]

A Lightweight Mixture-of-Experts Framework for Remote Sensing Visual Question Answering

Si-Yeon Jung[a)] and Oh-Seol Kwon[a)‡]

요 약

최근 원격탐사 분야에서 멀티모달(Multi-modal) 모델을 이용한 시각적 질의응답(Visual Question Answering, VQA) 연구가 활발히 진행되고 있으나, 방대한 데이터와 복잡한 추론 과정으로 인해 연산 효율성 및 성능 개선의 한계가 있다. 본 논문은 이러한 한계를 극복하기 위해 경량화된 원격탐사 이미지 기반 시각적 질의응답(Remote Sensing Image-Visual Question Answering, RSI-VQA)을 위한 전문가 혼합 모델(Mixture-of-Experts, MoE) 프레임워크를 제안한다. 비전과 언어 작업에 특화된 각각의 3개의 전문가 모델에게 데이터 특성에 따른 적응형 가중치를 동적으로 적용함으로써, 모든 전문가에게 동일한 가중치를 부여하던 기존 방식 대비 경량화된 구조로 성능을 향상시켰다. 다양한 원격탐사 데이터셋에 실험을 진행한 결과, 제안된 방법은 기존 모델 대비 우수한 문장 생성 능력과 연산 효율성을 가짐을 확인하였다.

Abstract

Research on Visual Question Answering (VQA) using multi-modal models has recently been active in the field of remote sensing. However, challenges exist in improving computational efficiency and performance due to the vast amounts of data and complex reasoning processes involved. This paper proposes a lightweight Mixture of Experts (MoE) framework for Remote Sensing Image-Visual Question Answering (RSI-VQA) to overcome these limitations. We enhance performance with a lightweight architecture compared to conventional methods that apply uniform weights to all experts. Our method dynamically applies adaptive weights to each of the three expert models specializing in vision and language tasks based on data characteristics. Experimental results on various remote sensing datasets confirm that the proposed method has superior sentence generation capabilities and computational efficiency compared to existing models.

Keyword : Mixture of Experts, Visual Question Answering, Deep Learning, Remote Sensing, Multi-modal

a) 국립창원대학교 지능로봇융합공학(Department of Intelligent Robotics and Convergence Engineering, Changwon National University)
‡ Corresponding Author : 권오설(Oh-Seol Kwon)
E-mail: osk1@changwon.ac.kr
Tel: +82-55-213-3669
ORCID: https://orcid.org/0000-0002-1077-9615

※ This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP2025-RS-2024-00438409) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00555758)
· Manuscript August 25, 2025; Revised September 8, 2025; Accepted September 8, 2025.

Ⅰ. 서 론

원격탐사는 지리적 특성을 탐지하고 모니터링하는 기술로, 특히 방위 산업 분야에서 인프라 손상 및 환경 변화를 객관적으로 파악하기 위해 활용되고 있다[1]. 최근 이미지 캡셔닝 모델을 원격탐사 이미지에 적용하는 연구가 증가하고 있는 추세이지만 실제 현장에서는 시설의 개수나 종류 등의 구체적 정보를 요구하는 반면, 이미지 캡셔닝 모델은 이미지 전체의 포괄적인 설명만을 제공한다는 한계가 있다. 또한 원격탐사 이미지는 넓은 영역을 포함하므로 관심 영역(Region of Interest, ROI)이 명확하지 않은 경우가 많아, 사용자가 원하는 특정 부분에 대한 선택적 정보 추출이 어렵다. 이러한 문제를 해결하기 위해 이미지와 질문 두 가지 입력으로 이미지에서 원하는 정보를 선택적으로 추출하는 VQA(Visual Question Answering) 모델을 원격탐사 분야에 적용한 연구가 진행되었다. 하지만 대부분의 원격탐사 이미지 기반 질의응답(RSI-VQA) 모델은 성능 향상을 위해 사전 학습된 모델의 모든 파라미터를 업데이트하여[2], 매우 많은 파라미터가 필요하고, 이로 인해 요구되는 막대한 계산 비용과 메모리가 요구되어 실제 현장에서 활용하기에는 한계가 있다.

따라서 이러한 문제를 해결하기 위해 전문가 혼합 모델(Mixture of Experts, MoE) 구조를 활용한 접근법이 주목받고 있다. MoE는 입력에 따라 특정 전문가 모듈만을 선택적으로 활성화하여 전체 모델 크기는 유지하면서도 실제 연산량을 크게 줄일 수 있는 효율적인 아키텍처이다[3]. 원격탐사 분야에서도 이러한 MoE 구조의 활용이 시도되고 있으며, 최근에는 해당 아키텍처를 활용해 여러 개의 LLM을 전문가로 사용하여 원격탐사 이미지 분석 모델에 적용하는 방법이 제안되었다[4]. 이 접근법은 LLM을 전문가로 활용함으로써 자연어 답변 생성을 가능하게 하였으나, 여전히 70억 개의 파라미터를 가져 계산 효율성과 성능을 동시에 확보할 수 있는 경량화된 MoE에 대한 연구가 필요하다.

Ⅱ. 관련 연구

1. 딥러닝 기반 원격탐사 이미지 분석

원격 탐사 데이터는 일반적인 자연 이미지와 다르게 광범위한 영역을 포착하므로 이미지의 규모가 방대하고, 같은 장면이라도 센서나 촬영 조건에 따라 기하학적 변형이 발생한다. 이러한 특징으로 인해 원격 탐사 분야에서의 정확하고 안정적인 이미지 분석에 대한 어려움이 있었다. 이를 해결하기 위해 최근 연구들은 원격탐사 이미지 처리에서 합성곱 신경망(Convolutional Neural Networks, CNN)과 비전 트랜스포머(Vision Transformer, ViT) 기반 딥러닝 아키텍처를 중심으로 발전하고 있다. 비전 트랜스포머(Vision Transformer, ViT)는 셀프 어텐션(self-attention) 메커니즘을 통해 입력 이미지의 패치 간 장거리 상호작용을 이해하는 데 효과적이지만, CNN이 가지고 있던 지역적 특성 학습 능력과 위치 변화에 대한 강건성이 부족하여 학습 데이터가 충분하지 않을 때 성능이 제한될 수 있다는 한계가 있다.

2. 원격 영상 분석을 위한 비전-언어 모델

단순히 객체를 탐지하는 것을 넘어, 객체 간의 관계와 활동까지 이해하는 하이브리드 모델이 제안되자[5], 이를 이용해 자연어 응답을 생성하고자 이미지 캡셔닝을 원격탐사 분야에 적용한 연구(Remote Sensing Image Captioning, RSIC)가 주목받았다[6]. 이미지 캡셔닝의 일반적인 아키텍처는 인코더-디코더(Encoder-Decoder) 구조를 사용한다. 인코더는 주로 합성곱 신경망(CNN)과 비전 트랜스포머(ViT)를 활용하여 입력 이미지를 특징 벡터로 변환하는 역할을 하며, 순환 신경망(Recurrent Neural Network, RNN) 또는 트랜스포머(Transformer)를 디코더로 사용하여 인코딩된 특징 벡터를 기반으로 단어 시퀀스를 생성하여 문장을 생성한다. 초기 모델들은 이미지 전체의 특징을 하나의 전역 특징(global feature)으로 인코딩했으나[7], 이는 객체가 포함된 원격탐사 이미지의 복잡성을 충분히 담아내지 못하는 한계를 보였다. 이러한 모델의 성능을 향상시키는 데 있어 어텐션(Attention) 메커니즘이 핵심 기술로 적용되었다[8]. 어텐션은 디코더가 문장의 각 단어를 생성할 때, 이미지의 특정 영역에 집중하여 단어와 이미지 특징 간의 대응 관계를 학습하도록 하였다. 예를 들어, 모델이 “차량”이라는 단어를 생성할 때 이미지의 차량 부분에 더 높은 가중치

를 부여하는 방식으로 작동하여, 보다 정확하고 상세한 캡션을 생성하는 데 기여한다. 이처럼 이미지 캡셔닝은 원격탐사 분석에 있어 유의미한 성과를 보였지만, 모델이 학습 데이터에 존재하는 패턴을 모방하여 구체적인 표현이 부족하고, 문맥적 이해가 결여된 캡션을 생성한다는 문제가 있다. 또한, 캡셔닝 모델은 입력 받은 이미지에 대한 포괄적인 설명만을 출력하여 사용자가 이미지에 대한 특정 정보를 얻고자 할 때 답할 수 없다는 근본적인 한계가 존재한다.

시각적 질의응답(Visual Question Answering, VQA)은 이미지 캡셔닝의 수동적 설명 방식의 한계를 해결하고자 발전된 능동적 상호작용 기술이다[9]. 시각적 질의응답(VQA)은 주어진 이미지에 대한 자연어 질문을 받아들여 정확하고 맥락에 맞는 자연어 답변을 생성하기 위해 어텐션 메커니즘을 활용하여 질문에 관련된 이미지 영역에 집중하거나, 그래프 신경망(Graph Neural Network)을 사용해 객체 간의 암묵적인 관계를 추론하는 등 더욱 정교한 융합 기법을 사용한다. 하지만 원격탐사 VQA와 같은 복잡한 비전-언어 모델은 성능 향상을 위해 파라미터 수가 급격히 증가하는 경향이 있으며, 이러한 거대한 단일 모델(monolithic model)은 높은 연산 비용을 유발하고 개별 모달리티 간의 간섭(interference) 문제로 인해 언어 처리 능력이 저하될 수 있는 한계를 가진다.

3. 전문가 혼합 모델 기반 원격 영상 분석

최근 시각적 정보와 언어적 정보를 동시에 처리하기 위한 멀티모달(Multi-modal) 아키텍처가 제안되었고[10], 그 중, 하나의 거대한 문제 해결을 위해 여러 개의 작고 전문화된 모델(experts)이 협력하는 방식의 전문가 혼합 모델(Mixture of Experts, MoE)이 주목받았다. 전문가 혼합 모델(MoE)은 크게 전문가(Experts), 게이팅 네트워크(Gating Network), 조건부 연산(Conditional Computation)으로 구성된다. 전문가(Experts)는 전체 문제 중 특정 하위 집합이나 특정 작업에 특화되도록 훈련된 개별 신경망 모델이며, 라우터(Router)라고도 불리는 게이팅 네트워크(Gating Network)에서 어떤 전문가 모델이 해당 입력에 가장 적합한지를 판단하고 선택한다. 이 과정에서 전문가 혼합 모델(MoE)의 가장 중요한 특징은 모든 입력에 대해 전체 모델을 활성화하는 것이 아니라, 게이팅 네트워크의 판단에 따라 소수의 전문가만 선택적으로 활성화한다는 점이다[11]. 일반적으로 top-k 방식을 사용하여 상위 k개의 전문가만을 선택하며, 이때 k는 전체 전문가 수보다 훨씬 작은 값으로 설정된다. 이를 통해 모델의 전체 파라미터 수는 매우 크지만, 추론 시 실제로 사용되는 파라미터를 줄여 연산 효율성을 극대화한다. 이러한 전문가 혼합 모델을 원격 영상 분석에 적용하게 되면 각 전문가가 특정 분야에 특화되면서도, 게이팅 네트워크를 통해 하나의 모델로 통합되어 전반적인 성능을 최적화할 수 있다. 최근 제안된 약 147억 개 파라미터 규모의 원격탐사 멀티모달의 기초가 되는 모델은 특징 추출, 융합, 공유 전문가를 계층적으로 결합한 MoE 구조를 도입하였다[12]. 그 외에도 비전 트랜스포머의 순방향 신경망(feed-forward neural networks, FFN)을 희소 전문가 혼합 모델로 대체하여 성능 저하 없이 추론 계산을 크게 절감한 모델에 대한 연구가 진행되었다. 이렇듯 전문가 혼합 모델을 원격탐사 기반 VQA에 적용하려는 시도가 이어지고 있다[13]. 따라서 본 논문에서는 입력의 종류에 따라 전문가를 다르게 할당하여 각 출력을 하나의 정보로 결합하고 자연어 답변을 생성할 수 있는 경량화된 전문가 혼합 모델(MoE) 기반 원격탐사 시각적 질의응답(VQA) 프레임워크를 제안한다.

Ⅲ. 제안한 전문가 혼합 모델을 이용한 원격 영상 분석 알고리즘

본 논문에서는 이미지 분석과 질문 분석을 위한 전문가를 별개로 두어 모델의 계산 효율성 증대를 통해 기존 모델의 한계를 개선하고자 한다. 이를 위해 그림 1과 같이 전문가 혼합 모델 구조를 설계하고, 제시한 아키텍처와 같이 두 개의 입력을 효율적으로 처리하기 위해 사전 훈련된 모델을 사용한다. 먼저, 위성 이미지와 자연어 질문을 멀티모달 입력으로 받아 라우터가 상황에 가장 적합한 전문가 조합을 선택하고, 그 중요도를 동적으로 산출한다. 선택된 3개의 전문가는 각자 특화된 영역의 입력을 처리한다. 여기서 YOLOv12는 객체와 클래스 분포를 추출해 장면의 구조적 단서를 제공한다. ViT는 전역 시각 패턴과 관계를 인코딩

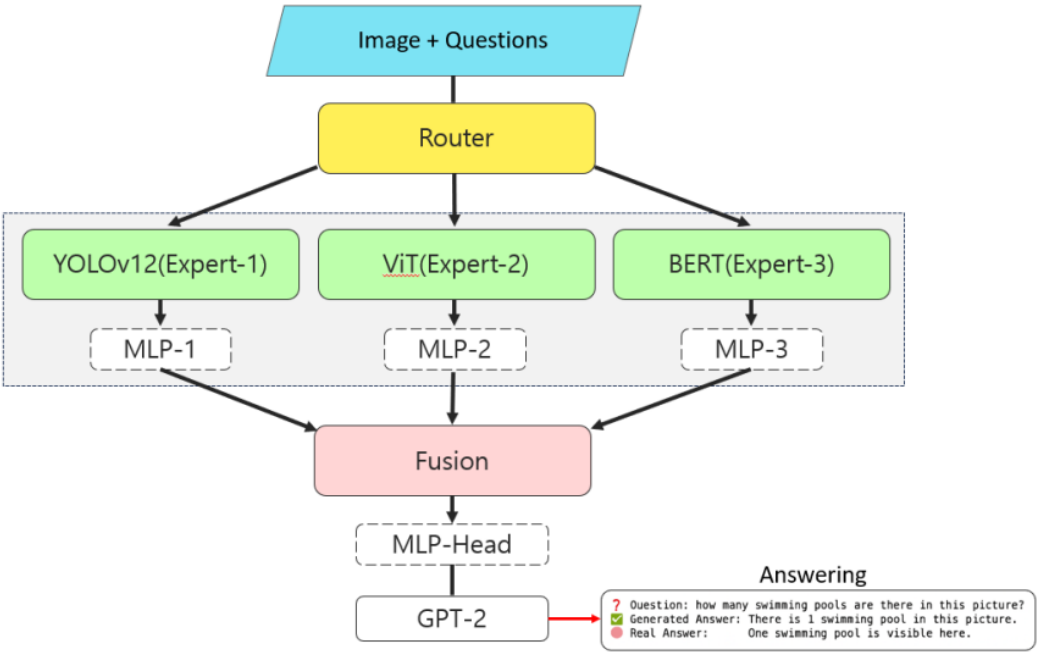

그림 1. 제안된 경량화 전문가 혼합 모델 기반 위성 영상 분석 VQA 아키텍처
Fig. 1. Proposed Lightweight Mixture-of-Experts Based Architecture for Remote Sensing Image Analysis VQA

하고, BERT는 질문의 의미를 문맥 임베딩으로 표현한다. 각 전문가의 출력은 개별 MLP를 통해 공통 임베딩 공간으로 변환 및 정규화되며, 이후 라우터가 산출한 가중치에 따라 3개의 출력이 가중합 방식으로 융합된다. 융합 벡터는 MLP-Head를 거쳐 언어 디코더가 활용하기 적합한 표현으로 변환되고, GPT-2가 이를 통해 자연어 답변을 생성한다.

1. 적응형 가중치 라우터(Router)

라우터(Router)는 전처리가 완료되어 입력으로 들어온 이미지-질문을 분석하여, 어떤 전문가(Expert)가 이 질문에 답하는 데 가장 적합할지를 계산한다. 모든 조건을 고려하기 위해 각각 768차원의 BERT, YOLO, ViT 출력과 15차원 YOLO 클래스 분포를 하나의 긴 2,319차원 벡터로 결합한다. 그 수식은 다음과 같다.

$$\tau_t = \frac{\tau_0}{1+\alpha \cdot max(p)}, (\tau_t \geq 0.5) \quad (1)$$

τ_0는 초기 온도 파라미터로 학습 시작 시 분포의 부드러움을 설정하여 값이 크면 여러 전문가에게 고르게 확률을 분배하고 작으면 특정 전문가에 집중한다. α는 조절 계수(scaling factor)로, 값이 크면 전문가 선택이 빨리 한쪽에 집중되므로 전문가 확률이 높아질 때 τ_t를 얼마나 빠르게 줄일지 결정한다. 어떤 전문가가 가장 적합한지를 나타내고 해당 값이 크면 "이미 확신이 큰 상태"임을 의미한다. τ_t는 시간 t에서 사용되는 소프트맥스 온도(softmax temperature)를 의미하며, softmax 함수에 적용되어 분포의 날카로움을 제어한다. 학습 초반에는 $\max(p)$ 값이 낮아 $\tau_t \approx \tau_0$이므로 여러 전문가가 동시에 고려되고, 학습이 진행되면서 특정 전문가 확률이 높아지면 τ_t 값이 줄어 softmax 분포가 뾰족해지며, 한 전문가에 집중하도록 활용된다. 따라서, 수식 (1)은 소프트맥스 함수의 파라미터를 동적으로 조절하여, 라우터가 가장 신뢰하는 전문가의 값인 $\max(p)$가 클수록 낮은 τ_t를 가지도록 하는 수식이다. 이를 통해 전문가 선택을 더 명확하게 함으로써 불필요한 전문가 활성화를 줄여 효율성을 높인다.

$$P_s = softmax\left(\frac{1}{\tau_t}\right) \quad (2)$$

수식 (2)의 $softmax\left(\frac{1}{\tau_t}\right)$에서 $softmax(\ )$는 입력값을 확률 분포로 변환하는 함수로 큰 값은 더 높은 확률을, 작은 값은 낮은 확률을 부여한다. $\frac{1}{\tau_t}$는 온도 스케일링(temperature scaling)을 의미하며 τ_t가 작을수록 softmax 결과는 더 날카로워져 특정 전문가 확률이 커지고, τ_t가 크면 분포가 평탄해져 여러 전문가에 비슷하게 분배된다. P_s는 라우터(Router)가 계산한 시점 t에서 각 전문가가 선택될 확률을 나타내며, 합은 1이 된다. 이처럼 수식 (1)에서 조정된 τ_t 값을 이용해 소프트맥스의 확률 분포를 계산하고, 이는 소프트 맥스의 분포를 좁혀 특정 전문가에게 높은 가중치를 부여하여 퓨전(Fusion) 단계에서 어떤 전문가의 출력을 더 크게 반영할지 결정하는 기준을 제공한다. 계산 결과에 따라 전문가(Expert) 단계에서 입력을 분석하는데, 이를 위해 Top-p와 Max-k 2개의 상위 전문가 선택 필터링 알고리즘을 결합하여 보다 유연하고 안정적인 전문가 선택을 수행한다.

$$w = \frac{P_s \odot m}{\sum(P_s \odot m) + \in} \quad (3)$$

수식 (3)은 모든 전문가의 기여도를 정규화하여 합이 1인 가중치 벡터 생성하는 수식이다. 여기서 $P_s = [P_y, P_v, P_b]$, $m = [m_y, m_v, m_b], w = [w_y, w_v, w_b]$이며, y, v, b는 각각 YOLO, ViT, BERT에 대한 기호임을 의미한다. 이를 스칼라로 나타내면 아래와 같이 표현된다.

$$w_y = \frac{P_y m_y}{P_y m_y + P_v m_v + P_b m_b + \in} \quad (3\text{-}1)$$

$$w_v = \frac{P_v m_v}{P_y m_y + P_v m_v + P_b m_b + \in} \quad (3\text{-}2)$$

$$w_b = \frac{P_b m_b}{P_y m_y + P_v m_v + P_b m_b + \in} \quad (3\text{-}3)$$

위 세 개의 식은 공통된 분모를 가지며, 한 번의 정규화 과정으로 세 전문가의 가중치 값이 동시에 도출된다. 여기서 P_s는 수식 (2)에서 라우터(Router)가 계산한 각 전문가의 선택 확률을 의미하고 이를 모델이 해당 전문가를 얼마나 활용해야 하는지 나타내는 m과 곱하여 기여도를 산출한다. 해당 값을 전체 전문가의 합으로 정규화하면 확률 분포 형태가 만들어지는데, 이때 ϵ는 분모가 0이 되는 것을 방지하기 위한 매우 작은 상수로 사용된다. 이후 도출된 결과 w는 각 전문가의 최종 기여도 분포가 되며, 전체 합이 1이 되도록 조정된다. 이러한 과정을 통해 특정 전문가에 대한 편향을 방지하고 모든 전문가의 기여도를 균형있게 반영할 수 있다.

수식 (4)는 최종 가중치의 엔트로피를 정규화하는 수식이다. 전문가 가중치 분포의 엔트로피 H는 다음과 같이 정의된다.

$$H = -\sum_{i=1}^{N} w_i \log(w_i) \quad (4)$$

여기서 N은 전문가의 총 개수를 의미하며, w_i는 수식 (3)에서 도출된 i-번째 전문가의 최종 가중치이다. 이때 $0 \le w_i \le 1$이고, $\sum_{i=1}^{N} w_i = 1$의 조건을 만족한다. 모든 전문가 가중치에 대해 엔트로피를 계산하여 가중치 분포가 균등할수록 엔트로피 값이 증가하며, 특정 전문가에 집중될수록 엔트로피 값이 감소한다. 위 과정을 통해 특정 전문가에 지나치게 의존하는 것을 줄여 모델의 일반화 성능을 향상시키고 라우터가 계산한 가중치 분포가 일정 수준의 다양성을 유지하도록 조정한다. 또한, 계산된 정규화 항을 학습에 반영함으로써 라우터는 객체 분포, 질문의 의미와 전역적 시각 특징을 적합한 전문가에게 안정적으로 가중시킨다.

2. 전문가 모듈(Experts)

본 논문에서는 사전 학습된 세 개의 모델 YOLOv12x, ViT(Vision Transformer), BERT(Bidirectional Encoder Representations from Transformers)를 전문가(Experts)로

최종 VQA 프레임워크에 적용하였다. YOLOv12x는 객체 탐지에 특화되어 입력된 RGB 이미지를 20차원의 특징 벡터로 변환하며, 이 중 15차원은 객체 분포 정보로서 별도로 보존되어 이후 fusion 과정에서 추가적으로 활용된다. 이처럼 객체 분포 정보를 보존함으로써 장면 구조와 맥락 정보를 추론할 수 있는 기반을 제공하게 된다.

비전 트랜스포머(Vision Transformer, ViT)는 이미지 태스크를 위해 트랜스포머 구조의 인코더를 변형한 모듈이다[14]. 본 연구에서 ViT는 YOLOv12x가 추출한 객체의 세부 정보를 보완하여 이미지 전체의 전반적인 분위기와 객체 간 상호작용을 분석하여 768차원의 이미지 특성을 추출한다. 이를 통해 모델은 단일 객체 수준이 아닌 장면 전체의 문맥적 의미를 학습할 수 있고, 복잡한 환경에서의 추론 성능을 높인다.

마지막 전문가인 BERT(Bidirectional Encoder Representations from Transformers)는 문장 의미 파악에 특화된 모델로서 자연어 질문을 벡터로 변환하는 역할을 수행한다[15]. 데이터 전처리 과정에서 image_num : question? answer. 형식의 질문 데이터셋을 기호 :, ?를 기준으로 파싱(parsing)하여 이미지 번호, 질문, 정답을 구분한다. 이 과정을 통해 분리된 형태에 따라 각 샘플을 저장하고 구조화하여 문장의 맥락을 잘 이해할 수 있도록 하고, 전처리가 완료된 2개의 32차원 출력을 BERT 모델의 입력으로 사용하여 질문 문장을 CLS 토큰 기반의 768차원 임베딩 벡터로 변환한다.

3. MLP 모듈

세 전문가 모듈에서 추출된 특징들은 서로 다른 차원을 가지므로, 연산의 편리함을 위해 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 통한 차원 통일 과정을 거친다. YOLOv12x에서 출력된 20차원 벡터는 모델에 적합하도록 MLP를 거쳐 처리된다. 추출한 특징을 하나의 공간에 조합하기 위해 20차원 벡터를 512차원으로 선형 변환하고, 비선형 함수인 ReLU를 통해 복잡한 패턴을 학습하도록 한다. 학습 시에 발생하는 과적합을 방지하기 위해 드롭 아웃(Dropout)에 0.1을 적용하여 10%의 확률로 일부 뉴런이 비활성화되도록 하였다. 해당 512차원의 벡터를 전체 VQA 모델에 사용될 수 있도록 최종 768차원의 임베딩으로 변환하였다.

ViT와 BERT는 트랜스포머의 인코더 형태로서, YOLOv12x와 달리 768차원의 출력을 가지며, 이들 역시 동일한 MLP 구조를 통해 처리된다. 이처럼 세 전문가의 출력 차원을 통일함으로써 융합 과정에서 정보가 효과적으로 결합되도록 하였다. 세 전문가의 768차원 출력은 라우터의 입력으로 결합되어 전문가별 소프트 가중치(soft weight)를 산출한다. 이어서 융합 모듈에서 각 전문가 출력에 해당 가중치를 적용해 가중합된 표현을 생성하게 되고, 생성된 표현은 프리픽스(Prefix) MLP를 거쳐 GPT-2 디코더에 입력함으로써 최종 자연어 답변을 생성하는 데 활용된다.

4. 퓨전(Fusion) 모듈

퓨전(Fusion) 모듈은 서로 다른 전문가 모듈에서 추출된 특징을 통합하여 최종적으로 유의미한 멀티모달 표현을 형성한다. 3.2에 언급된 3개의 전문가와 같이 성격이 다른 출력은 차원이 상이하므로, 먼저 3.3의 다층 퍼셉트론(MLP)을 통해 동일한 임베딩 차원으로 투영된다. 이 과정을 통해 개별 전문가의 출력은 동일한 공간에서 결합이 가능해진다. 결합을 위해 본 모델에서는 수식 (5)와 같이 특정 가중치를 곱하여 합산하는 방식을 적용한다.

$$f = w_y \cdot (y + y_{dass}) + w_v \cdot (v) + w_b \cdot (b) \tag{5}$$

여기서 y, v, b는 각각 YOLO를 통해 얻은 객체 정보 벡터, ViT 그리고 BERT를 통해 인코딩된 질문 의미 벡터이며, y_{dass}는 1의 라우터에서 언급된 15차원의 YOLO 클래스 분포이다. 라우터는 입력 데이터의 특성과 질문 유형을 기반으로 각 전문가의 기여도를 반영하는 가중치 $w_{y,}\ w_{v,}\ w_b$를 산출하며, 이를 통해 불필요한 정보는 제외되고 중요한 특징이 강조된다. 융합된 표현은 이후 MLP-Head를 거치며 비선형 변환, 정규화, 드롭아웃 등의 과정을 통해 잡음을 줄이고 일반화를 향상시킨다. 이렇게 후처리된 벡터는 GPT-2 디코더에 입력되어 자연어 응답 생성을 위한 프리픽스 임베딩(prefix embedding)으로 활용된다. 이를 통해 모델은 원격탐사 이미지 기반 질의응답에

서 보다 정교하고 일관성 있는 답변을 생성할 수 있게 된다.

5. 프리픽스 튜닝(prefix-tuning)을 통한 GPT-2 디코더

본 모델에서는 사전 학습된 GPT-2 small 모델을 사용하여 전체 파라미터는 동결 상태를 유지하고, 프리픽스 임베딩(prefix-embedding)을 생성하는 Prefix MLP만 학습한다. 이때 YOLO, ViT, BERT, GPT-2 모두 768차원으로 맞춰져 있어 모듈 간 연산이 안정적으로 이루어지게 된다. 이 768차원의 입력을 1,024차원으로 확장시켜 비선형 함수인 GELU를 통해 손실을 줄이고 복잡한 패턴의 표현을 가능하게 한다. 이후 LayerNorm을 통해 1,024차원 피처를 중심으로 평균과 분산을 계산해 표준화하고 다시 학습이 가능한 형태로 선형 변환 후, 과적합(Over-Fiting)을 방지하기 위해 드롭아웃(Dropout)의 파라미터 값에 0.1을 적용한다. 이후 위의 과정을 한번 더 반복함으로써 표현이 안정화되도록 하고, 도출된 5개의 1,024차원 출력을 한 벡터에 표현하기 위해 3,840차원의 벡터로 선형 변환한다. 선형 변환한 벡터는 특정 형태의 텐서(Tensor)로 표현되고, 프리픽스 토큰을 5개로 분산하여 표현의 유연함을 주기 위해 최종 형태로 변환한다. 최종 변환된 출력은 이후 GPT-2의 입력으로 사용하여 자연어 응답을 생성한다. 이와 같이 본 모델에서는 크기 대비 적은 파라미터를 추가함으로써, 모델 경량화와 계산 효율성의 균형을 유지한다.

Ⅳ. 실험 및 결과

1. 하이퍼파라미터

제안된 모델은 총 3개의 전문가를 활용하며, 각 전문가의 출력은 768의 투영 차원으로 공통된 특징 공간을 형성한다. 이 과정에서 512의 은닉 차원을 갖는 MLP가 라우터와 퓨전 모듈의 표현을 개선한다. GPT-2 디코더로 전달되는 융합 벡터는 5개의 프리픽스 토큰(prefix_length)으로 변환되어 정보 전달력을 높였다. 또한, 모델의 과적합을 억제하기 위해 0.1의 드롭아웃(dropout)을 적용하였다. 라우터의 입력 차원은 768차원의 YOLO, ViT, BERT의 특징 벡터와 15개의 YOLO의 클래스 수를 결합한 총 2,319차원이다. 동적 온도 스케일링을 위해 τ_0를 5.0, α=2.0을 사용하여 라우팅의 안정성을 확보하였다. 학습은 두 단계에 걸쳐 진행되어 첫 번째 학습에서는 5 epoch 동안 모델의 기본 구조를 안정화하고, 2단계에서는 30 epoch에 걸쳐 모델 일부를 파인튜닝하였다. 성능 최적화를 위해 Adam 옵티마이저(Optimizer)를 사용하였으며, YOLO 전문가 3e-4, ViT/BERT 전문가 5e-4, GPT-2 디코더 1e-4의 모듈별 학습률을 적용하였다. 모델의 추론 단계에서는 GPT-2 디코더가 최대 50 토큰을 생성하도록 하였으며, 빔 서치(num_beams) 값을 5로 설정하여 생성된 응답의 품질을 향상시켰다. 생성 다양성을 조절하기 위해 온도(temperature)를 0.8로 사용하였고, 반복적인 표현을 억제하기 위해 반복 페널티(penalty) 값에 1.2를 적용하였다.

2. 실험 결과

제안한 방법으로 응답을 생성했을 때 성능 향상을 확인하기 위해 약 13,000개의 이미지-질문 쌍으로 구성된 RSI-VQA 데이터셋을 대상으로 모델을 평가하였다. 훈련, 검증, 테스트 세트로 구성된 이 데이터셋은 원격탐사 이미지에 대한 자연어 질문에 답하는 시스템 개발을 위한 벤치마크로 이용되며, 각 줄은 이미지 번호, 질문, 답변을 포함하고 세 부분이 콜론과 물음표로 구분된다. 표 1은 기존 모델들과의 성능 비교를 위해 무작위로 배정된 쌍을 사용하여 실험을 진행한 결과이다.

표 1. RSI-VQA 데이터셋의 예/아니오 및 기타 질문에 대한 정확도(%)
Table 1. Accuracy (%) on Yes/No and Others Questions (RSI-VQA Dataset)

Method	Yes/No	Others
IMG+SVM	48.95	62.45
BOW+SVM	81.74	9.82
IMG+BOW+SVM	84.08	65.28
IMG+SOFTMAX	40.37	65.49
BOW+SOFTMAX	81.74	7.19
IMG+BOW+SOFTMAX	84.01	70.95
IMG+GoogleNews+SOFTMAX	80.06	72.68
IMG+GloVe+SOFTMAX	84.09	75.10
IMG+BERT+SOFTMAX	84.92	73.30
MAIN-740M[16]	92.82	54.50
Proposed Model-383M	95.58	90.52

제안 모델은 Yes/No 질문에 대해 95.58%의 정확도를 달성하여 기존 최고 성능 모델인 MAIN-740M의 92.82%보다 2.76% 향상된 성과를 보였다. 특히 Others 항목에서는 MAIN-740M보다 36.02% 개선된 결과를 나타냈다. 최신 BERT 기반 방법인 IMG+BERT+SOFTMAX과도 비교했을 때 10.66%, 17.22%의 향상된 성능을 달성하여 트랜스포머 기반 방법에서도 우수한 결과를 가짐을 확인하였다. 또한, 대부분의 기준 모델들이 Yes/No와 Others 항목 간에 상당한 성능 편차를 보인 반면, 제안한 모델은 두 항목 모두에서 높고 안정된 결과를 보였다. 예를 들어, BOW+SVM과 BOW+SOFTMAX는 Yes/No에서 각각 81.74%의 동일한 성능을 기록하였지만, Others에서는 9.82%, 7.19%로 낮은 성능을 보였다. 하지만 제안한 모델은 두 항목 간 성능 차이가 5.06%에 불과하여 강건한 추론 성능을 확인하였다. 이러한 결과를 통해 제안한 모델이 RSI-VQA에서 우수한 계산 효율성과 안정성을 가짐을 알 수 있다. 그림 2는 이를 시각화한 결과이다.

표 2는 제안한 적응형 가중치 연산을 적용한 라우터(CONFIDENCE ROUTER)의 문장 생성 능력을 확인하기 위해 기존의 RSI-VQA 데이터셋의 형태를 문장형으로 바꾼 RSI-FullVQA 데이터셋으로 학습을 진행한 결과이다. 현재 RS-VQA 분야에서는 문장형 표준 벤치마크가 부족해 ChatGPT로 생성한 전체 문장형 답변이 포함된 데이터셋을 사용하였다. 이러한 이유로 기존 모델과의 비교에 한계가 있어, 본 논문에서는 생성된 텍스트와 정답 텍스트 간의 단어 정밀도와 유사성을 측정하기 위해 BLEU-1~4 및 ROUGE-L 점수를 사용하였다. 해당 평가에는 기존 라우터, 제안한 라우터 총 2개의 모델에 대해 문장 생성 능력을 평가하였다.

표 2의 평가 결과에 따르면, 제안한 라우터가 기존 라우터보다 모든 평가 지표에서 향상된 성능을 보였다. BLEU-1에서 0.7758점을 기록하여 기존 ROUTER보다

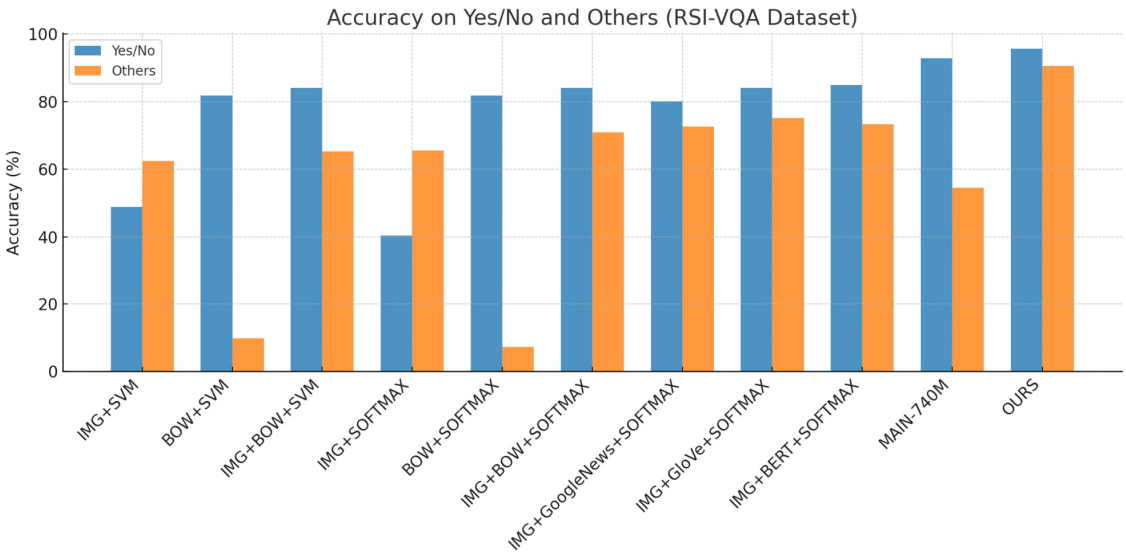

그림 2. RSI-VQA 데이터셋의 예/아니오 및 기타 질문에 대한 정확도(%) 비교
Fig. 2. Accuracy (%) comparison for Yes/No and Others questions on the RSI-VQA dataset

표 2. 문장형 RSI-FullVQA 데이터셋으로 학습하여 문장 생성 능력을 평가한 결과
Table 2. The evaluation of sentence generation capability after training on the sentence-formatted RSI-FullVQA dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
CONVENTIONAL ROUTER	0.6801	0.6563	0.5625	0.5170	0.7050
PROPOSED ROUTER	0.7758	0.7321	0.6304	0.5820	0.7912

14.07% 향상된 성능을 보였고, 모든 평가 지표에서 일관되게 높은 성능을 보였지만 특히 ROUGE-L 지표에서 제안 모델이 0.7912점으로 최고 성능을 기록하였다. 우수한 BLEU 점수는 생성된 답변의 정확성과 자연스러움이 크게 개선되었음을 의미하며, ROUGE-L 점수의 향상은 정답과의 의미적 유사성이 증대되었음을 보여준다.

그림 3은 기존 RSI-VQA 데이터셋의 단답형 응답을 문장형 응답으로 변환한 RSI-FullVQA 데이터셋을 사용해 생성된 자연어 응답 결과이다. (a)에서는 이미지 내 비행기 존재 여부에 대해 “No, there is no plane present here.”라는 문장형 응답을 출력하였으며, 정답 “No, a plane is not present here.”과 의미적으로 일치하였다. 또한 (b)에서는 소형 차량의 개수를 묻는 질문에 “There are 217 small vehicles in this picture.”라는 문장형 응답을 생성하였으며, 실제 정답과 일치하였다. (c)~(f) 역시 객체의 개수, 존재 여부 질문에 대해 실제 정답과 일치한 답변을 생성한 것을 알 수 있다. 이를 통해 제안한 적응형 가중치 연산을 적용한 라우터가 기존의 라우터보다 자연스럽고 서술적인 응답을 생성하여, 모델이 생성하는 답변의 해석 가능성을 높이는 것을 확인하였다.

표 3은 본 연구의 MoE 모델 구조의 연산 효율성을 평가하기 위해 GFLOPs(GIGA Floating-point Operations per second)를 측정한 결과이다. 실제 모델의 모듈 구성과 데이터 흐름을 유지하면서 FLOPs 측정을 안정화하기 위해 단

그림 3. RSI-VQA의 짧은 답변을 문장형 답변으로 변환한 데이터셋인 RSI-FullVQA를 활용한 실험 결과
Fig. 3. Results Using RSI-FullVQA, a Dataset that Converts Short Answers in RSI-VQA to Sentence-Formatted Responses

일 샘플 기준(Batch=1) 프록시 모델(Proxy MoE)을 사용하였으며, 측정은 Python 3.11 및 PyTorch 2.8.0(CPU 빌드) 환경에서 fvcore FlopCountAnalysis 라이브러리를 적용하여 진행하였다.

표 3. 프록시 모델을 활용해 모듈별 FLOPs 측정한 결과
Table 3. Results by Module using a Proxy Model

Module	Input Size	GFLOPs
YOLO (Expert-1)	RGB 512×512	2.530
ViT (Expert-2)	RGB 224×224	0.039
BERT (Expert-3)	32 tokens	0.018
ROUTER+Fusion	2319	0.003
GPT-2 Decoder	50 tokens	2.748
Total		5.338

프록시 모델의 GFLOPs는 각 전문가 모듈의 연산량을 개별적으로 측정한 후 합산되었다. 먼저 YOLO 전문가는 특징 추출 백본, 투영 MLP, 객체 수의 연산량을 포함하며, ViT 전문가는 ViT 백본과 차원 변환 연산량을 의미한다. 또한, BERT 전문가의 GFLOPs는 단일 레이어 트랜스포머 인코더와 특징 벡터의 차원 변환 연산량을 나타낸다. 라우터 및 퓨전 모듈은 게이팅 및 융합 MLP의 연산량으로 계산되며, 이는 모델 전체 연산에서 매우 작은 비중을 차지하였다. 마지막으로 GPT-2 디코더는 융합된 특징 벡터를 기반으로 텍스트를 생성하는 프리픽스 MLP, 디코더 블록을 통해 도출되었다. 제안된 모델의 GFLOPs 결과를 통해 라우터 및 퓨전 모듈이 낮은 연산량으로 복잡한 계산 없이 다수의 전문가를 효율적으로 통합하고 있음을 확인할 수 있다. 이는 개별 전문가들이 각자의 특화된 분야(시각, 텍스트)를 독립적으로 처리하고, 통합 모듈은 그 결과를 최소한의 연산으로 결합하여 모델의 전체적인 효율성을 증가시켰음을 나타낸다.

V. 결 론

본 논문은 전문가 혼합 모델 기반의 경량화된 원격탐사 영상 질의응답 프레임워크를 제안하였다. 제안한 방법은 각각의 전문가에게 적응형 가중치를 적용하여 불필요한 연산을 줄임으로써 답변의 효율성을 높였으며 RSI-FullVQA 데이터셋을 통해 문장형 답변을 생성하는 방법을 제안하였다. 제안한 모델의 성능을 평가하기 위해 표준 데이터셋으로 기존 모델들과 비교한 결과, 7.4억 개의 파라미터를 가진 기존 모델 대비 절반 수준의 크기(3.8억 개)로 향상된 성능을 달성하여 우수한 계산 효율성을 확인하였다. 이는 제한된 자원 환경에서도 원격탐사 VQA를 효과적으로 수행할 수 있음을 의미하며, 대규모 멀티모달 학습을 통한 일반화 성능 개선 및 실제 응용 환경에서의 실시간 추론 가능성을 확장할 수 있을 것으로 기대된다.

참 고 문 헌 (References)

[1] H. Xu, S. Barbot, and T. Wang, “Remote sensing through the fog of war: Infrastructure damage and environmental change during the Russian-Ukrainian conflict revealed by open-access data,” *Natural Hazards Research,* vol. 4, no. 1, pp. 1 – 7, Mar. 2024. doi: https://doi.org/10.1016/j.nhres.2024.01.006

[2] H. Touvron, L. Martin, K. Stone, et al., “LLaMA 2: Open foundation and fine-tuned chat models,” *arXiv preprint,* Jul. 2023, arXiv: 2307.09288. https://arxiv.org/abs/2307.09288 doi: https://doi.org/10.48550/arXiv.2307.09288

[3] N. Shazeer, A. Mirhoseini, K. Maziarz, et al., “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” *Proceedings of International Conference on Learning Representations (ICLR)*, Toulon, France, Apr. 2017. https://arxiv.org/abs/1701.06538 https://doi.org/10.48550/arXiv.1701.06538

[4] H. Lin, D. Hong, S. Ge, C. Luo, K. Jiang, H. Jin, and C. Wen, “RS-MoE: A vision-language model with mixture of experts for remote sensing image captioning and visual question answering,” *IEEE Trans. Geosci. Remote Sens.*, in press, Nov. 2024. https://arxiv.org/abs/2411.01595 doi: https://doi.org/10.48550/arXiv.2411.01595

[5] J. Chen, Z. Wang, and Y. Li, “P2FEViT: Plug-and-play CNN feature embedded hybrid vision transformer for remote sensing image classification,” *Remote Sensing*, vol. 15, no. 3, pp. 1 – 15, Mar. 2023. doi: https://doi.org/10.3390/rs15071773

[6] X. Li and H. Zhang, “Remote sensing image captioning using deep learning,” *Proceedings of the International Conference on Automation and Computing (AUTOCOM)*, Mar. 2024. doi: https://doi.org/10.1109/AUTOCOM60220.2024.10486178

[7] X. Lu, X. Zheng, and X. Li, “Exploring models and data for remote sensing image captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol.56, no.4, pp.2183 – 2195, Apr. 2018. doi: https://doi.org/10.1109/TGRS.2017.2776321

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” *Proc. ICML, Lille*, France, pp.2048 – 2057,

Jul. 2015. https://proceedings.mlr.press/v37/xuc15.html
[9] Y. Wang and P. Ghamisi, “RSAdapter: Adapting Multimodal Models for Remote Sensing Visual Question Answering,” *IEEE Transactions on Geoscience and Remote Sensing,* vol. 62, pp. 1 – 13, Jun. 2024. doi: https://doi.org/10.1109/TGRS.2024.3413174
[10] Y. Liu and T. Chen, “Unified transformer with cross-modal mixture experts (TCMME),” *Remote Sensing*, vol.15, no.19, pp.1 – 12, Sep. 2023. doi: https://doi.org/10.3390/rs15194682
[11] Y. Zhang, H. Wang, X. Li, et al., “A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications,” *arXiv preprint*, arXiv:2503.07137, Mar. 2025. doi: https://doi.org/10.48550/arXiv.2503.07137
[12] Bi, H., et al., “RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation,” *arXiv preprint*, arXiv:2504.03166, Apr. 2025. doi: https://doi.org/10.48550/arXiv.2504.03166
[13] C. Riquelme, J. Puigcerver, B. Mustafa, et al., “Scaling vision with sparse mixture of experts (V-MoE),” *Advances in Neural Information Processing Systems (NeurIPS)*, vol.34, pp.1 – 14, Dec. 2021. doi: https://doi.org/10.48550/arXiv.2106.05974
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An image is worth 16 x16 words: Transformers for image recognition at scale,” *in Proc. Int. Conf. Learn. Represent. (ICLR)*, Vienna, Austria, May 2021. https://openreview.net/forum?id=YicbFdNTTy
[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, Minneapolis, MN, USA, pp.4171 – 4186, Jun. 2019. doi: https://doi.org/10.18653/v1/N19-1423
[16] X. Zheng, B. Wang, X. Du and X. Lu, “Mutual Attention Inception Network for Remote Sensing Visual Question Answering,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022, Art no. 5606514. doi: https://doi.org/10.1109/TGRS.2021.3079918

저 자 소 개

정 시 연

- 2023년 3월 ~ 현재 : 국립창원대학교 지능로봇융합공학전공 학사
- ORCID : https://orcid.org/0009-0005-8151-3258
- 주관심분야 : 멀티모달, 컴퓨터 비전, 인공지능

권 오 설

- 2008년 8월 : 경북대학교 전자공학과 박사
- 2008년 9월 ~ 2010년 8월 : New York University, Psychology 박사후연구원
- 2010년 9월 ~ 2011년 8월 : 삼성전자 영상디스플레이사업부 책임연구원
- 2021년 12월 ~ 2022년 8월 : Princeton University, Electrical & Computer Engineering 방문교수
- 2011년 9월 ~ 현재 : 창원대학교 로봇제어계측공학과 교수
- ORCID : https://orcid.org/0000-0002-1077-9615
- 주관심분야 : 영상처리, 인공지능, 피지컬 AI