멀티모달 임베딩 정렬 기반 자연어-위성사진 유사 사례 검색*

강하운[1], 김윤희[01], 배세은[1], 이가현[1], 홍세원[1], 박하명[2]
국민대학교
gomtang3274@kookmin.ac.kr, yuni2821@kookmin.ac.kr, 1004bse@kookmin.ac.kr,
anna030608@gmail.com, hsw1805@kookmin.ac.kr, hmpark@kookmin.ac.kr

Multimodal Embedding Alignment-Based Retrieval of Similar Satellite Images from Natural Language

Ha-Woon Kang[1], Yun-Hui Kim[01], Se-Eun Bae[1], Ga-Hyun Lee[1], Se-Won Hong[1], Ha-Myung Park[2]
Kookmin Univ.

요 약

본 연구에서는 자연어를 통해 위성사진 유사 사례를 검색할 수 있는 멀티모달 임베딩 정렬 기법을 제안한다. 먼저, 멀티모달 LLM으로 사진으로부터 텍스트를 생성하여 데이터를 구축하고, 사전학습된 이미지 인코더 및 텍스트 인코더를 활용하여 임베딩을 추출하였다. 이후 대조학습 기반의 Triplet Loss를 적용해 두 임베딩 공간을 정렬하였다. 실험에는 천리안 1호 위성 데이터셋과 GPT-4o mini로 생성한 기상 텍스트를 활용하였다. 실험 결과, 제안 모델이 자연어 기반 기상 유사 사례 검색에서 효과적으로 작동함을 확인하였다.

1. 서 론

최근 기상 예보의 정확도를 높이기 위해 과거의 유사한 기상 상황을 탐색하는 유사 사례 기반 접근이 주목받고 있다. 특히, 과거의 위성사진 중 현재와 유사한 기상 상황을 가진 유사 사례를 검색의 경우, 예보관의 기상 예측 해석 보조 등 다양한 기상 분야에서 효과적인 도구로 활용될 수 있다.

다만 기존의 위성사진 유사 사례 검색은 위성사진 자체에만 의존하고 있으며, 자연어 기반의 질문을 통한 검색 기능은 거의 다루어지지 않았다. 따라서 자연어 텍스트를 이용해 위성사진을 검색하고자 할 때, 기존 시스템으로는 이를 수용하지 못하는 구조적 한계가 있었다. 또한 기존 연구에서는 멀티모달 모델 (CLIP [1] SigLIP [2] 등)을 기반으로 이미지-텍스트 정렬을 학습하였지만, 대부분은 일반적인 데이터에 한정되어 있어 위성사진 도메인에 바로 적용하기 어렵다.

따라서 본 연구에서는 다음과 같은 방법을 제안한다. 첫째, ChatGPT 기반의 멀티모달 LLM을 활용하여 위성사진에 대한 자연어 설명을 지역별로 생성한다. 둘째, 이미지와 텍스트 인코더를 활용하여 각각의 임베딩을 추출한다. 셋째, 두 임베딩 공간을 일치시키기 위해 Projection layer(MLP)를 각각 적용하고, 자연어로 위성사진 유사 사례를 검색할 수 있도록 대조학습 기반 정렬을 수행한다. 자연어 형태의 기상 텍스트를 입력하면 위성사진 중 유사 사례를 검색할 수 있으며, 다양한 기상 관련 업무에서 활용 가능한 위성사진 검색 도구를 제공할 수 있을 것이다.

2. 관련 연구

2.1. 이미지 임베딩 모델

이미지 임베딩은 이미지의 고차원 정보를 벡터공간에 효율적으로 표현하기 위한 핵심 기술로, 초기에는 Convolusional Neural Network(CNN) [3] 기반 모델이 주로 활용하였다. ResNet [4], EfficientNet [5] 등 대표적인 CNN 기반 백본 모델은 이미지 분류 및 검색 등의 다양한 분야에서 높은 성능을 보여주었으며, 임베딩 추출에서도 활용되고 있다. 최근에는 Vision Transformer(ViT) [6] 계열의 모델이 등장하면서, 이미지 내 장기적 의존 관계를 포착할 수 있는 self-attention 기반 임베딩이 가능해졌고, 일반적인 이미지 뿐만 아니라 도메인 특화 이미지에도 효과적으로 적용되고 있다.

2.2 텍스트 임베딩 모델

Sentence-BERT(SBERT) [7]는 문장 간 의미적 유사도 학습에 최적화된 구조로, 효율적이고 강건한 문장 임베딩을 제공한다. SBERT는 자연어 질의와 이미지 간 의미 정렬을 효과적으로 수행할 수 있어, 검색 및 분류 작업에서도 우수한 성능을 보인다.

2.3. 멀티모달 임베딩 기법

이미지-텍스트 간 의미를 학습하기 위해서는, 비지도 대조학습 기반 멀티모달 임베딩 기법이 활용되고 있다. 이 방식은 각 임베딩을 projection layer를 통해 정렬하고, 대조학습을 적용하여 positive pair는 가깝게, negative pair는 멀어지도록 학습한다. 대표적으로 CLIP [1]은 대규모 이미지-텍스트 쌍을 기반으로 공통 임베딩 공간을 구축해, 성능을 크게 향상시켰다. SigLIP [2]은 sigmoid 기반 손실 함수를 적용해 학습 안정성과 임베딩 정렬 성능을 개선하였다.

3. 제안 방법

본 연구에서는 이미지 임베딩과 텍스트 임베딩을 동일한 임베딩 공간으로 사상하고, 자연어 질의를 통해 기상 유사 사례를 검색할 수 있는 모델을 제안한다. 그림 1은 우리의 모델 전체 구조를 나타낸 것이다.

---

[0] 발표자
[1] 공동 제1저자
[2] 교신저자

* 본 연구는 2022년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2022-0-00964).

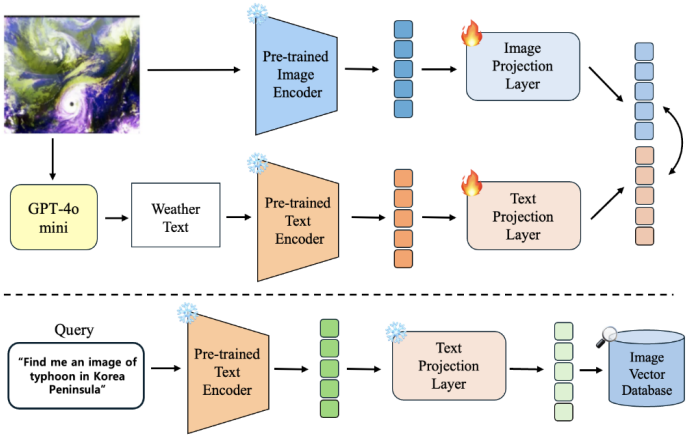

그림 1: 모델 구조

먼저 위성사진 데이터에는 명확한 정답 레이블이 없어, 위성 이미지에 대한 자연어 설명을 자동 생성하기 위해 멀티모달 LLM으로 생성하였다. 아래의 그림 2는 위성사진 설명용 텍스트를 생성하기 위해 LLM에 입력한 프롬프트를 나타낸 것이다.

```
You are a weather analysis expert who provides clear answers to
all questions. The given image is an infrared channel image
captured by the COMS satellite over East Asia in November 2017.
Respond by identifying the observed weather phenomena in the
image for each specified region as a "Region - Weather
Phenomenon" pair, using bullet points. Answer only bullets. Do
not explain the progress.

List of regions: Korean Peninsula, Mongolia, Northern China,
Eastern China, Southern China, Manchuria and Primorsky Krai,
Honshu (Japan), Hokkaido (Japan), Kyushu (Japan), East China
Sea, Philippine Sea, South China Sea, and the Pacific coast of
Japan.
```

그림 2: 기상 텍스트 생성 프롬프트

다음으로, 이미지 인코더와 텍스트 인코더를 사용하여 학습 데이터에서 제공된 이미지와 텍스트 임베딩을 추출한 후, 대조학습을 통해 두 임베딩을 동일한 공간으로 매핑한다. 이 과정에서 이미지와 텍스트 간 의미적 유사도를 학습하여, positive pair는 가까이, negative pair는 멀리 위치하도록 학습한다. 특히, 본 연구에서는 대조학습의 일환으로 같은 날짜의 이미지와 텍스트는 positive pair로 묶고 그 외의 경우는 negative pair로 설정하였다. 이후 Triplet Loss를 사용하여, 임베딩 공간에서 anchor와 positive 간의 거리는 최소화하고, anchor와 negative 간의 거리는 최대화하는 방향으로 모델을 학습하였다. Triplet Loss는 다음과 같이 정의된다.

$$L_{triplet} = \max(0, d(a,p) - d(a,n) + \alpha)$$

여기서 *a, p, n*은 각각 anchor, positive, negative의 임베딩 벡터를 의미하며, $d(\cdot)$는 코사인 거리이다. α는 양의 상수로, positive-negative 간 최소 거리 차이를 조절하는 하이퍼파라미터이다.

이후, 모델에 자연어 질의가 입력되면, 학습이 완료된 텍스트 인코더와 Projection layer를 거쳐서 질의의 임베딩을 생성하고, 이를 이미지 임베딩 공간으로 매핑한다. 이후 임베딩 간의 코사인 유사도를 계산하여 상위 N개의 유사 이미지를 검색한다. 이를 통해 자연어 질의와 의미적으로 유사한 위성사진을 효율적으로 검색할 수 있다.

4. 실험

4.1. 데이터셋

실험에는 대한민국의 천리안 1호 위성으로 2011년 2월부터 2020년 3월 31일까지 3시간 간격으로 촬영된 총 25,744장의 위성사진 데이터셋을 사용한다. 각 위성사진은 적외채널(IR), 단파적외채널(SWIR), 수증기 채널(WV) 등 세 개의 독립적인 채널로 구성되어 있다. ViT의 입력 형식에 맞도록 원본 위성사진은 224x224 크기로 변환하고, 각 픽셀에 해당하는 값은 평균이 0이고 표준편차가 1이 되도록 범위를 조정한다.

텍스트 데이터의 경우, GPT-4o mini를 사용하여 13개 지역에 대한 프롬프트를 통해 지역별/시간별로 다양한 기상 상황을 포괄하도록 생성하였다.

4.2. 학습 설정

이미지 인코더 모델로는 사전학습된 ViT-B/16을 사용하였으며, 텍스트 인코더 모델로는 사전학습된 all-mpnet-base-v2를 사용하여 학습을 진행하였다. 하이퍼파라미터의 경우, 학습률은 0.001, 배치 크기는 2048로 설정하였고, 에폭 수는 3000으로 설정하였다.

4.3. 평가 기준

사용자 자연어 질의와 검색된 상위 20개 이미지-텍스트 쌍 간의 일치도를 다음 두 가지 기준으로 평가하였다.

- 기상 현상만으로 질의했을 때, 상위 20개 결과 중 해당 기상 현상 텍스트가 13개 지역 중 몇 개가 나왔는지를 계산한다.
- 지역과 기상 현상을 함께 질의했을 때, 상위 20개 중 해당 지역과 기상 현상을 포함하는 결과의 개수를 측정한다.

4.4. 실험 결과

본 내용에서는 제안한 모델의 성능을 다양한 방식으로 분석한 결과를 제시한다.

4.4.1 정량적 분석

앞서 정의한 두 가지 평가 기준인 기상 현상의 지역 분포 범위와 지역·기상 일치도를 바탕으로 모델의 검색 성능을 정량적으로 평가한다.

표 1: 기상 현상 질의에 대한 해당 기상 현상의 등장 횟수 평균 (Top-20 기준)

기상 현상	Snow	Clear	Cloudy	Rain	Typhoon
검색 결과	3.3	2.75	8.45	4.95	2.3
학습데이터 평균	0.78	2.6	8.62	2.49	0.2

표 1은 기상 현상만을 질의했을 때, 검색된 이미지에 대한 텍스트 설명들 중 해당 기상 현상을 포함하는 지역의 평균 개수를 나타낸다. 일반적인 평균 기상 현상 수치보다 본 모델의 기상 현상 검색 결과가 더 높게 나타나, 모델이 기상 개념을 보다 정확히 학습함을 확인할 수 있다. 단, 'Cloudy'는 데이터 평균보다 검색 결과가 낮은 수치를 보이는데, 이는 동아시아 지역에서 구름 낀 날씨가 일반적인 기상 현상이기 때문이다.

표 2: 복합 질의에 대해 지역명과 기상 현상을 모두 포함하는 검색 결과 수 (Top-20 기준)

지역 / 기상 현상	Snow	Clear	Cloudy	Rain	Typhoon
Korean Peninsula	13	14	19	6	2
Mongolia	1	13	3	0	0
Northern China	5	1	19	9	0
Eastern China	1	0	14	12	7
Southern China	0	3	12	9	1
Manchuria and Primorsky Krai	14	3	17	2	0
Honshu (Japan)	11	1	12	12	9
Hokkaido (Japan)	20	12	3	1	0
Kyushu (Japan)	0	5	15	16	7
East China Sea	0	1	10	7	4
Philippine Sea	0	4	14	0	3
South China Sea	0	8	17	6	1
Pacific coast of Japan	2	0	10	18	1

표 2는 '기상 현상 + in + 특정 지역' 형태로 자연어 질의를 했을 때, 상위 20개의 이미지 검색 결과 중 텍스트 설명에 해당 지역과 기상 현

상이 모두 포함된 이미지의 개수를 나타낸 것이다. 전반적으로 'Cloudy'는 상대적으로 흔한 기상 현상이라 대부분의 지역에서 일치도가 높게 나타났으며, 'Snow'는 해양 지역에서는 자연적으로 발생하기 어려운 현상이기 때문에 대부분 0으로 나타난 것이 특징이다. 지역별로는 Hokkaido가 'Snow' 질의에서 20건 모두 해당 기상 현상을 포함해 가장 뛰어난 성능을 보였고, Korean Peninsula 또한 'Cloudy', 'Clear', 'Snow'에서 높은 정확도를 보였다. 반면, Mongolia와 같은 내륙 지역은 강수나 태풍처럼 실제 발생 빈도가 낮은 기상 현상에서 일치도가 낮았으며, 해양 지역은 'Snow'처럼 지리적으로 발생이 어려운 현상에서 0건을 기록하는 등, 지역 특성과 기상 현상의 자연적 연관성이 모델 성능에 영향을 준 것으로 보인다. 이러한 결과는 모델이 실제 기상 현상의 지리적 분포를 반영하며, 자연어 기반의 복합 질의에서도 효과적으로 검색 결과를 반환할 수 있음을 보여준다.

4.4.2 질의 예시 및 정성적 분석

자연어 질의에 따라 반환되는 이미지 결과를 시각적으로 확인함으로써, 해당 기상 현상과 지역이 실제로 상위 결과에 잘 반영되는지를 정성적으로 분석한다.

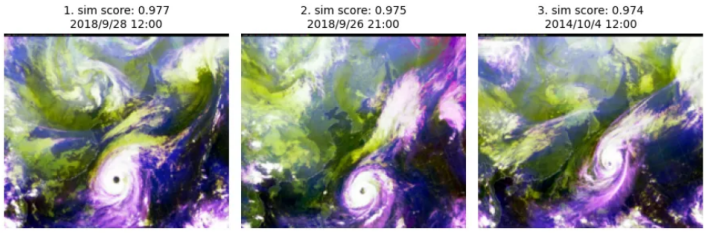

그림 3: 검색된 상위 3개 이미지

그림 3은 "Honshu, Japan is experiencing a typhoon."라는 문장으로 검색한 결과를 나타낸다. 검색 결과 상위 3개 이미지 모두 일본 혼슈 지역이 태풍의 영향을 받고 있는 위성사진이며, 각 이미지에 해당하는 지역 기상 텍스트는 각각 "Affected by the edge of a typhoon, heavy clouds", "Impacted by a typhoon", "Typhoon effects"이다.

이는 모델이 입력된 질의의 지역 정보와 기상 현상을 정확히 반영하여, 관련성 높은 이미지가 검색됨을 보여준다.

4.4.3 자연어 질의의 임베딩 위치 시각화

그림 4에서 전체 이미지 임베딩을 t-SNE로 시각화하였으며, 그 중 대표적인 기상 현상 5개(Typhoon, Cyclone, Rain, Snow, Clear)에 해당하는 이미지만 색상으로 구분하였다. 나머지 이미지는 배경처럼 회색으로 표현하였으며, 두 개의 질의 임베딩을 함께 사영하여 어떤 기상 카테고리와 가까운지 확인하였다.

- 질의 1: "Please find a picture of snow falling in Korea"
- 질의 2: "A typhoon is affecting Honshu, Japan."

각 질의가 각각 'Snow', 'Typhoon' 이미지가 밀집된 영역에 위치하였고, 이는 모델이 질의의 의미를 적절히 파악하고 기상 현상을 올바르게 임베딩 공간에 반영했음을 보여준다.

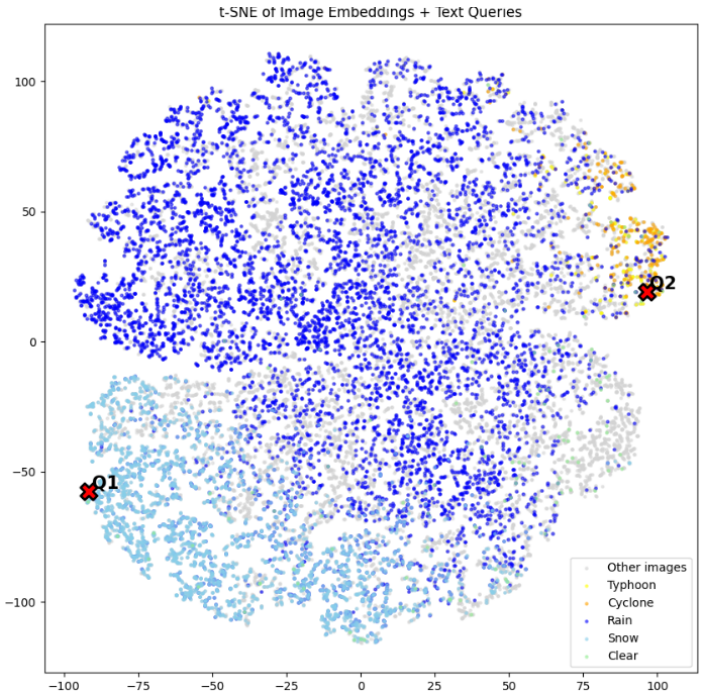

그림 4: 기상 카테고리와 질의 임베딩의 t-SNE 시각화

5. 결론

본 연구에서는 기존 기상 유사 사례 검색 시스템이 사용자 직관에 기반한 자연어 질의를 효과적으로 처리하지 못하는 한계를 지적한다. 이를 해결하기 위해, 사전학습된 이미지 및 텍스트 인코더를 활용하고 Triplet Loss 기반의 대조학습을 통해 두 임베딩 공간을 정렬함으로써, 자연어 질의를 이용한 위성사진 유사 사례 검색 모델을 구현하였다. 실험 결과, 자연어 질의를 넣었을 때 그에 맞는 위성사진이 다수 검색되었다. 향후 연구에서는 다양한 기상 수치 데이터와의 융합을 통해 검색 성능을 더욱 향상시킬 수 있을 것이라 기대한다.

참 고 문 헌

[1] Alec Radford et al., Learning Transferable Visual Models From Natural Language Supervision, PMLR, 2021.

[2] Xiaohua Zhai et al., Sigmoid Loss for Language Image Pre-Training, ICCV, 2023.

[3] Y. LeCun et al., Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation Vol. 1, No.4, pp. 541-551, 1989.

[4] Kaiming He et al., Deep Residual Learning for Image Recognition", CVPR, 2016.

[5] Mingxing Tan, Quoc Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, PMLR, 2019.

[6] Alexey Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recignition at scale, ICLR, 2021.

[7] Nils Reimers, Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, EMNLP, 2019.