위성사진 유사사례 검색을 위한 시계열 특성 반영 임베딩 모델 파인튜닝 기법*

강하운[0,1], 김윤희[1], 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

A Temporal Contrastive Fine-Tuning Approach to Learning Satellite Image Embeddings for Similarity Search

Ha-Woon Kang[0,1], Yun-Hui Kim[1], Se-Eun Bae[1], Ga-Hyun Lee[1], Se-Won Hong[1], Ha-Myung Park[2]
Kookmin Univ.

요 약

기상 예보의 정확도 향상을 위해 과거의 유사한 기상 상황을 탐색하는 유사사례 기반 접근이 주목받고 있다. 특히 위성사진은 대기 상태의 공간적 패턴을 직관적으로 표현할 수 있어 유사사례 검색에 효과적인 수단이 된다. 그러나 위성사진 간 시각적 차이가 미미하고 명확한 정답 레이블이 존재하지 않아, 기상학적 유사성을 효과적으로 반영하는 임베딩 생성을 위한 학습에는 어려움이 따른다. 본 연구에서는 위성사진의 시계열 연속성에 기반한 triplet loss 학습 기법을 통해, 기상학적 유사성을 반영한 임베딩 생성기법을 제안한다. 사전학습된 Vision Transformer(ViT) 모델을 기반으로 파라미터 효율성을 위해 Low-Rank Adaptation(LoRA)을 적용하였다. 실험에는 천리안 1호 위성의 9년치 자료를 활용하였으며, 임베딩 간 유사도를 UMAP 시각화와 대표적 이미지 유사도 지표(LPIPS, PSRN, SSIM)를 통해 정성적·정량적으로 평가하였다. 그 결과, 제안한 기법이 대조군인 기존의 사전학습 ViT 임베딩 및 SimCLR 파인튜닝 임베딩보다 시각적 및 기상학적으로 더 유사한 사례를 효과적으로 검색함을 확인하였다.

1. 서 론

기상 예보의 정확도를 높이기 위해 과거의 유사한 기상 상황을 탐색하는 유사사례 기반 접근이 주목받고 있다. 특히, 위성사진을 활용한 유사사례 검색은 관측된 대기 상태의 공간적 패턴을 직관적으로 반영할 수 있어, 예보관의 의사결정 지원 및 데이터 기반 분석에 효과적인 도구로 활용될 수 있다. 이러한 유사사례 검색의 성능은 위성사진 간 유사도를 정량적으로 표현할 수 있는 고품질의 임베딩의 확보에 크게 의존한다.

위성사진 임베딩은 일반적인 이미지 임베딩과는 다른 고유의 도전과제를 지닌다. 첫째, 지도학습을 위한 명확한 정답 레이블(label)이 존재하지 않아, 기존의 분류 기반 임베딩 학습 기법을 직접적으로 적용하기 어렵다. 둘째, 위성사진은 주로 구름이나 대기의 흐름을 포착하며, 이는 형태가 뚜렷한 객체 중심의 일반 이미지와 달리 대부분의 사진이 서로 유사하게 보이는 시각적 특성을 지닌다. 이러한 특성은 위성사진이 내포하는 기상학적 특성을 효과적으로 구분할 수 있는 임베딩 생성에 어려움을 준다.

Convolutional AutoEncoder, SimCLR 등과 같은 비지도학습 혹은 자기지도학습 모델을 통해 정답 레이블이 없어도 사진의 특성을 반영하도록 임베딩을 생성하거나, Vision Transformer(ViT) 등 사전학습된 임베딩 모델을 활용할 수 있다. 하지만, 대부분의 사진이 서로 미묘한 차이만을 보이기 때문에 임베딩의 유사도가 기상학적 유사도를 잘 반영하지 못한다.

이러한 한계를 극복하기 위해서는, 위성사진의 시각적 유사성과 기상학적 유사성을 동시에 반영할 수 있는 임베딩 학습 전략이 필요하다. 본 연구에서는 인접한 날짜의 위성사진이 유사한 기상 상황을 나타낼 가능성이 높다는 점에 착안하여, 시계열적 연속성을 임베딩 학습에 반영하고자 한다. 이를 위해, 시각적 표현력이 뛰어난 사전학습 Vision Transformer(ViT) 모델을 기반으로 하여, 시계열 정보와 시각적 정보를 효과적으로 통합하는 학습 방식을 제안한다.

2. 관련 연구

2.1. 이미지 임베딩 모델

이미지 임베딩은 이미지의 고차원 정보를 벡터공간에 효율적으로 표현하기 위한 핵심 기술로, 주로 Convolusional Neural Network(CNN) [1] 기반 모델이 초기부터 널리 활용되어 왔다. ResNet [2], EfficientNet [3] 등 대표적인 CNN 기반 백본 모델은 이미지 분류, 탐지, 검색 등의 다양한 비전 과제에서 뛰어난 성능을 보여주었으며, 임베딩 추출에서도 널리 활용되고 있다. 최근에는 Vision Transformer(ViT) [4]계열의 모델들이 등장하면서 이미지 내 장기적인 의존 관계를 포착할 수 있는 self-attention 기반 임베딩이 가능해졌고, 일반적인 이미지뿐만 아니라 도메인 특화 이미지에도 성공적으로 적용되고 있다.

2.2. 비지도 학습 기반 임베딩 기법

명확한 정답 레이블이 존재하지 않는 위성사진 도메인에서는, 비지도 학습 기반의 임베딩 생성 방식이 주로 활용되어왔다 [5],[6]. 대표적으로 AutoEncoder [7]는 이미지 복원을 통해 잠재 공간의 특성을 학습하는 방식으로, 위성사진의 시각적 특징을 반영한 임베딩 생성을 시도한다. 한편, 최근에는 대조학습 방식이 주목받고 있으며, SimCLR [8], MoCo [9], BYOL [10] 등은 증강을 통해 생성된

---

[0] 발표자
[1] 공동 제1저자
[2] 교신저자
* 본 연구는 2022년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2022-0-00964)

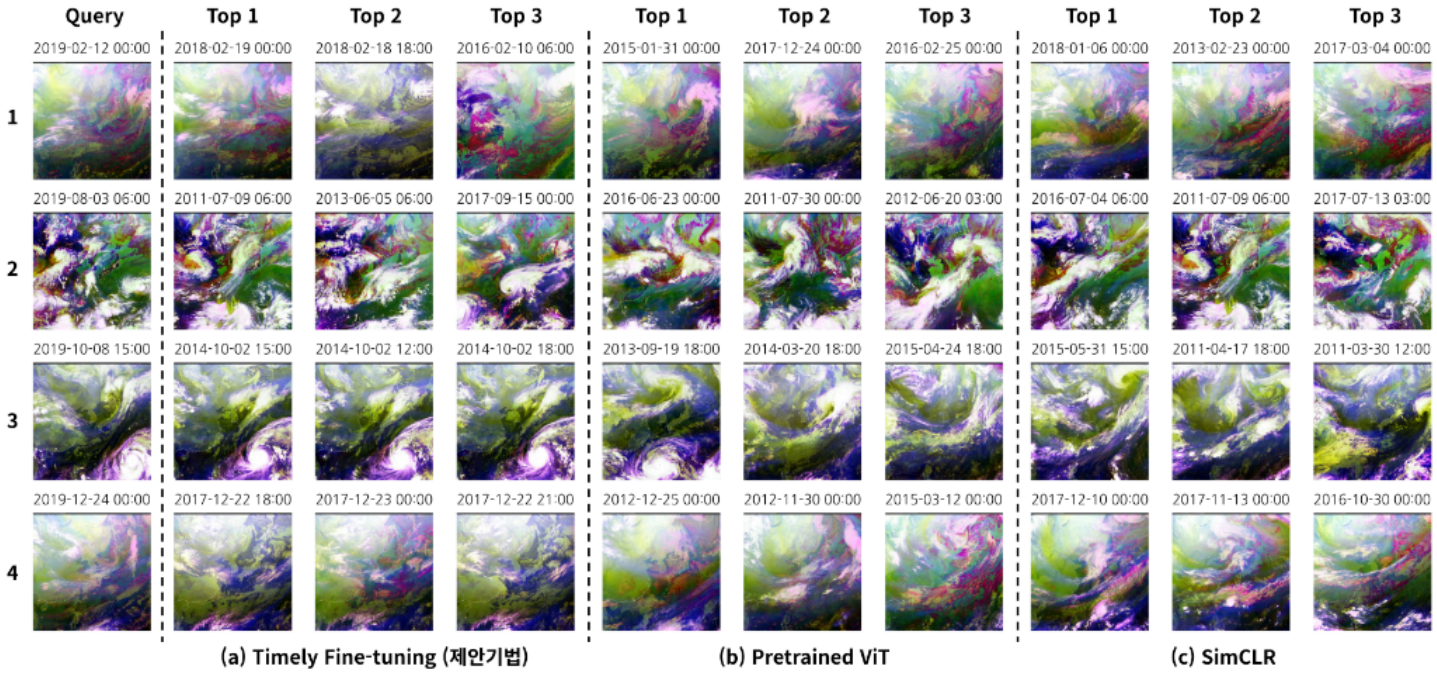


그림 1: 각 기법별 유사사례 질의 결과. 제안기법이 Pretrained-ViT 및 SimCLR 파인튜닝 방법에 비해 시각적으로 질의와 유사한 결과를 도출함. 각 사진은 IR, SWIR, WV 채널을 R, G, B 색상 채널로 매핑하여 시각화함.

positive 쌍과 negative 쌍 간의 상대적 유사도를 학습함으로써 강건한 학습을 가능하게 한다. 하지만 대부분의 위성사진이 유사한 시각적 특성을 가지기 때문에 앞서 언급한 방법으로는 기상학적 의미를 효과적으로 임베딩에 반영하기 어렵다.

2.3. Low-Rank Adaptation

Low-Rank Adaptation(LoRA) [11]는 사전학습된 대형 모델을 효율적으로 파인튜닝하기 위한 기법으로, 기존 모델의 가중치를 고정한 채 저랭크 행렬을 삽입하여 학습 파라미터 수를 크게 줄이면서도 성능 저하 없이 도메인 적응을 가능하게 한다. 특히, ViT와 같은 파라미터 수가 많은 모델에 적합하며, 제한된 계산 자원과 학습 데이터 환경에서도 효과적인 파인튜닝이 가능하다는 장점이 있다. 본 연구에서는 ViT모델에 위성사진의 시계열 정보를 반영하기 위해 LoRA를 적용한다.

3. 제안 방법

본 연구에서는 위성사진 간의 기상학적 유사성을 반영할 수 있는 임베딩을 생성하기 위해, 시계열 정보를 활용한 triplet 기반 학습 방법을 제안한다. 전체 구조는 사전학습된 Vision Transformer(ViT) 모델을 기반으로 하며, 파인튜닝 과정에서는 Low-Rank Adaptation(LoRA)을 적용하여 파라미터 효율성을 확보한다.

제안하는 임베딩 모델은 사전학습된 ViT를 활용한다. 입력 위성사진은 ViT의 입력 형식에 맞게 전처리 한다. ViT의 출력 중 [CLS] 토큰에 해당하는 벡터를 위성사진의 대표 임베딩으로 사용한다. 파인튜닝 과정에서 ViT의 주요 가중치는 고정하고, self-attention 블록 내 query, key, value projection layer에만 LoRA를 적용하여 효율적이고 효과적인 파인튜닝이 가능하도록 구성한다. 이러한 구성은 ViT 전체 파라미터는 변함없이 유지하므로 과적합 방지에도 효과적이다.

지도 학습용 레이블이 존재하지 않는 상황을 고려하여, 학습 샘플은 위성사진의 촬영 날짜 정보를 기반으로 구성된 triplet으로 정의된다. anchor는 특정 일시의 사진, positive는 인접 시간의 사진 (예: 3시간 후의 사진), negative는 임의 일시의 사진으로 설정한다. 이는 시간적으로 가까운 사진일수록 유사한 기상 패턴을 나타낼 가능성이 높다는 도메인 지식을 반영한 구성이다.

모델 학습은 triplet loss를 사용하여 임베딩 공간에서 anchor와 positive간 거리를 줄이고, anchor와 negative 간 거리는 멀어지도록 유도한다. Triplet loss $L_{triplet}$는 다음과 같이 정의된다.

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

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

4. 실험

4.1. 데이터셋

실험에는 대한민국의 천리안 1호 위성으로 2011년 2월부터 2020년 3월 31일까지 촬영된 사진 데이터셋을 사용한다. 위성 사진은 3시간 간격으로 촬영되며, 적외채널(IR), 단파적외채널(SWIR), 수증기채널(WV)로 구성된다. 각 위성사진은 ViT 입력 형식에 맞게 크기보정 및 자르기를 통해 224×224 크기로 변환되고, 픽셀 값은 평균 0, 표준편차 1이 되도록 정규화된다. 전체 데이터는 학습용(80%)과 검증용(20%)로 분할한다.

4.2. 학습 설정

백본 모델로는 ImageNet-21k로 사전학습된 ViT-B/16을 사용하였으며, LoRA 파라미터는 rank=8, dropout=0.1, alpha=16으로 설정하였다. 학습은 AdamW 옵티마이저를 사용하였고, 하이퍼 파라미터는 lr=3e-4, BatchSize=96, epochs=10으로 설정하였다.

4.3. 대조군

제안 방법의 성능을 평가하기 위해 다음의 baseline과 비교를 수행한다.

- Pretrained ViT: 사전학습된 ViT의 [CLS] 토큰 출력 사용.
- SimCLR: 원본 이미지와 해당 이미지에 가우시안 블러를 적용한 이미지를 anchor-positive 쌍으로 구성하고, 이를 기반으로 triplet loss를 활용해 ViT 파인튜닝 진행.

4.4. 실험 결과

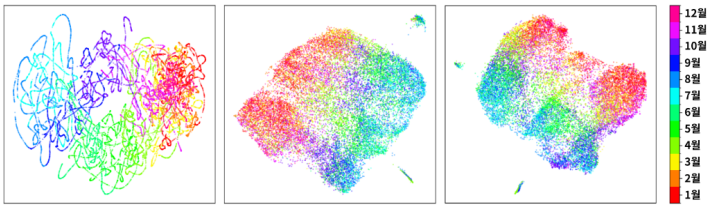
계절별 유사도 질의 결과: 그림 1는 봄(2019년 2월 12일), 여름(8월 3일), 가을(10월 8일), 겨울(12월 24일)의 위성사진을 질의로 하여 제안기법과 대조군인 SimCLR 임베딩과 Pretrained-ViT의 임베딩 유사도를 기준으로 가장 유사한 3개의 이미지를 순서대로 표시한 것이다. 시각적

		(a) Timely fine-tuning (제안기법)			(b) Pretrained ViT			(c) SimCLR		
질의	지표	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
1	LPIPS	0.387	0.444	**0.386**	0.387	0.444	0.386	0.461	0.432	0.447
	PSNR	**17.6**	14.5	17.5	17.6	14.5	17.5	15.3	15.3	15.1
	SSIM	**0.401**	0.355	0.401	0.401	0.355	0.401	0.330	0.365	0.327
2	LPIPS	**0.467**	0.500	0.543	0.563	0.546	0.572	0.519	0.467	0.514
	PSNR	**9.6**	8.3	8.3	8.5	8.9	8.1	8.8	9.6	9.1
	SSIM	**0.129**	0.086	0.090	0.108	0.099	0.087	0.115	0.129	0.107
3	LPIPS	0.447	0.455	**0.438**	0.508	0.489	0.507	0.523	0.532	0.497
	PSNR	**11.2**	11.1	11.1	9.2	9.6	8.9	8.7	8.4	10.4
	SSIM	0.214	0.208	**0.222**	0.178	0.203	0.179	0.135	0.149	0.172
4	LPIPS	0.400	**0.381**	0.419	0.430	0.421	0.432	0.462	0.499	0.489
	PSNR	15.7	**16.3**	15.9	15.5	14.9	15.5	13.2	13.7	12.8
	SSIM	0.344	0.350	0.345	**0.360**	0.349	0.281	0.293	0.288	0.282

표 1: 그림 1의 결과에 따른 이미지 간 LPIPS, PSNR, SSIM 유사도. 각 질의별로 가장 높은 유사도 수치를 두껍게 표시함.

으로 보았을 때 제안기법의 결과가 Pretrained-ViT의 결과 및 SimCLR의 결과보다 질의사진과 더 유사함을 알 수 있다. 특히 10월 18일의 질의의 경우 우측 하단에 태풍이 존재하는데, 제안기법의 경우 동일하게 우측 하단에 태풍이 존재하는 결과가 상위 3개 결과로 도출된 반면, 대조군의 경우 Pretrained-ViT의 첫 번째 사진만 위치가 다른 태풍이 존재하고 나머지 사진에는 태풍이 존재하지 않는다. 질의 결과의 이미지 유사도를 LPIPS, PSNR, SSIM을 사용하여 표 1로 수치화했다. 수치적으로도 제안기법이 더 유사한 사진을 찾아냄을 확인할 수 있다.

UMAP 시각화 결과: 그림 2은 Pretrained-ViT와 SimCLR 파인튜닝으로 생성한 임베딩과 제안기법으로 생성한 임베딩을 UMAP 알고리즘으로 2차원 시각화 한 것이다. 시계열 특성을 시각화 하기 위해 월마다 다른 색으로 표시하였다. Pretrained-ViT와 SimCLR의 경우 계절별로 좌우로 색이 스펙트럼을 이루고 있으며 왼쪽에 겨울, 오른쪽에 여름 데이터가 분포하고 있다. 제안기법의 결과는 실이 엉켜있는 것처럼 데이터가 분포하고 있고, 거시적으로 볼 때 도넛 모양으로 계절의 흐름이 반영되고 있음을 확인할 수 있다. 실처럼 데이터가 표현되는 것은 자연스러운 결과로 볼 수 있는데, 이유는 기상 위성사진들이 시간에 따라 연속적으로 촬영이 되고, 인접한 시간의 이미지들은 기상학적으로나 시각적으로나 매우 유사하기 때문이다.



(a) Timely Fine-tuning (제안기법) (b) Pretrained-ViT (c) SimCLR

그림 2: 각 임베딩 방법에 대한 UMAP 시각화

LPIPS, PSNR, SSIM 기준 성능 비교: 표 2은 각 임베딩 방법에 대해 대표적인 이미지 유사도 지표(LPIPS, PSNR, SSIM)를 기준으로 Recall@100을 계산한 결과를 나타낸다. 여기서 Recall@100은 해당 임베딩이 시각적으로 유사한 이미지를 얼마나 잘 검색해내는지를 나타내는 지표로, 값이 높을수록 성능이 우수함을 의미한다.

방법 (Recall@100)	LPIPS	PSNR	SSIM
Timely Fine-tuning (제안기법)	0.2096 ± 0.0736	0.2148 ± 0.0974	0.2210 ± 0.0749
Pretrained-ViT	0.0901 ± 0.0540	0.0597 ± 0.0452	0.0788 ± 0.0519
SimCLR	0.1191 ± 0.0650	0.0810 ± 0.0566	0.1045 ± 0.0647

표 2: 유사도 지표에 대해 Recall@100 계산 결과

실험 결과, 제안한 Timely Fine-tuning 기법은 대조군인 SimCLR 및 Pretrained ViT에 비해 전반적으로 우수한 성능을 보였으며, 특히 유사도 기준(LPIPS, PSNR, SSIM)에서 가장 높은 Recall@100을 기록해 이미지 간 시각적 유사성을 효과적으로 반영함을 입증하였다.

5. 결론

본 연구에서는 위성사진의 시계열적 특성을 활용하여, 시간적으로 인접한 이미지 간 유사성을 반영할 수 있는 triplet 기반 임베딩 학습 기법을 제안하였다. ViT에 LoRA를 적용하여 효율적인 파인튜닝이 가능하도록 했으며, 시계열 정보를 고려한 학습 샘플 구성 방식을 통해 기존 대비 더욱 시각적 의미가 보존된 임베딩을 생성하였다.

제안한 기법은 단순한 데이터 증강 기반 접근법보다 시계열 정보를 적극 활용함으로써, 위성사진 분석에 있어 구조적·시각적 유사성 파악 능력을 향상시킬 수 있음을 확인하였다.

참 고 문 헌

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

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

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

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

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

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

[7] Geoffrey Hinton and Ruslan Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science, 2006.

[8] Ting Chen et al., A Simple Framework for Contrastive Learning of
Visual Representations, ICML, 2020.

[9] Kaiming He et al., Momentum Contrast for Unsupervised Visual Representation Learning, CVPR, 2020.

[10] Jean-Bastien Grill et al., Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, NeurIPS, 2020.

[11] EJ Hu, Y Shen et al., Lora: Low-rank adaptation of large language models, ICLR, 2022.