

Self-supervised Graph Learning을 통한 멀티모달 기상관측 융합

전현주, 강전호, 권인혁

차세대수치예보모델개발사업단 자료동화실

hjeon@kiaps.org, jhkang@kiaps.org, ihkwon@kiaps.org

Multi-modal Meteorological Data Fusion based on Self-supervised Learning for Graph

Hyeon-Ju Jeon, Jeon-Ho Kang, In-Hyuk Kwon
Dept. of Data Assimilation, Korea Institute of Atmospheric Prediction Systems

요약

현재 수치예보 시스템은 항공기, 위성 등 다양한 센서에서 얻은 각종 관측 데이터를 동화하여 대기 상태를 추정하고 있지만, 관측변수 또는 물리량이 서로 다른 관측들을 처리하기 위한 계산 복잡도가 매우 높다. 본 연구에서 기존 시스템의 계산 효율성을 개선하여 관측을 평가하거나 전처리하는 데에 효율적으로 활용하기 위해, 각 관측의 특성을 고려한 자기 지도학습 방법을 통해 멀티모달 기상관측으로부터 실제 대기 상태를 추정하는 방법론을 제안하고자 한다. 비균질적으로 수집되는 멀티모달 기상관측 데이터를 융합하기 위해, (i) 기상관측의 heterogeneous network를 구축하여 개별 관측의 위상정보를 표현하고, (ii) pretext task 기반의 self-supervised learning을 바탕으로 개별 관측의 특성을 표현한다. (iii) Graph neural network 기반의 예측 모델을 통해 실제에 가까운 대기 상태를 추정한다. 제안하는 모델은 대규모 수치 시뮬레이션 시스템으로 수행되는 기존 기술의 한계점을 개선함으로써, 이상 관측 탐지, 관측의 편차 보정, 관측영향 평가 등 관측 전처리 기술로 활용할 수 있다.

1. 서론

기상 예측은 운송업, 제조업 및 환경 등의 분야에서 중요한 역할을 하고 있다[1, 2]. 기상 예측을 위해 지상 관측소에서는 바람, 온도, 습도, 기압 등의 대기 상태를 관측하고 있다. 지상 관측소에서 수집되는 정보는 특정 지역의 정밀한 대기 상태를 추정하는 데 유용하지만, 그 관측 범위가 제한적이다. 한편 remote-sensing 기술이 발전함에 따라 더 넓은 범위에 걸쳐 관측할 수 있는 레이더의 반사도나 위성의 복사량 관측이 수집되고 있다. 다양한 센서는 전 지구 환경을 설명하기 위해 여러 규모의 multi-modal 관측 데이터를 지속적으로 제공하므로 대규모 관측 데이터는 하루에 페타바이트 이상의 전송속도로 수집되고 있다[3].

기존의 수치예보 모델 시스템은 운동량, 질량, 엔탈피 등의 물리 및 역학적 관점에서 대기를 설명하도록, 결합 편미분 방정식으로 구성된 대규모 수치 시뮬레이션이다[4]. 이러한 방법은 대기 상태를 높은 정확도로 예측하여 산업 분야 전반에 걸쳐 효율적으

로 활용되고 있지만, 물리적 모델의 복잡성과 많은 도메인 지식의 필요성으로 인해, 급증하는 관측의 영향을 실시간으로 평가하기에는 어려움이 있다.

최근, 그래프 기반의 신경망 모델을 활용하여 대기의 시공간적 특징을 학습하는 연구가 수행되고 있다[5]. 일반적으로 기상 네트워크는 여러 지역에 분포하는 기상 관측소를 node로, 관측소 간의 인접 정보를 edge로 표현해 관측의 위상정보를 구조한다. 여기에 graph convolutional networks와 같이 비정형 데이터를 학습하는 기법을 적용하여 태양 복사량[6], 지진파[7], 해수면 온도[8] 등 다양한 기상 변수를 예측하는 연구가 다양하게 수행되고 있다. 특히, Lam et al.[9]은 글로벌 예측 시스템을 모델링하기 위해, 수치예보 모델의 3차원 격자 구조를 네트워크로 표현하여 대기의 공간 상호작용을 추출한다. 하지만, 실시간 관측정보가 반영되지 않아 예측 정확도 측면에서의 한계점이 있었다.

본 연구에서는 비균질적으로 분포한 멀티모달 기상관측 데이터를 융합하여 실제에 가까운 대기 상태

를 추정하기 위해 그래프 신경망 기반의 예측 모델을 제안하고자 한다. 특히, 관측 종별로 서로 다른 특성을 반영하기 위해 pretext task 기반의 그래프 자기 지도학습(Self-Supervised Learning, SSL) 프레임워크를 활용한다.

2. 기상 네트워크를 위한 자기 지도학습 방법

선행 연구[10]에서는 수치예보 모델의 공간정보를 표현하기 위해 3차원 그리드 형태의 convolutional neural networks(CNN) 기반의 모델을 활용하였다. 넓은 범위에서 다양한 기상현상의 패턴을 학습하는 모델은 단일 지역의 기상현상만을 학습하는 모델보다 예측 성능에 있어 6배 이상의 개선이 있다는 점을 실험적으로 검증하였다. n차원 텐서 형태의 입력자료가 요구되는 CNN 모델을 비균질한 관측 자료에 적용하기 위해서는 interpolation해야하나, 관측의 실제 위상정보와 특성이 크게 손실된다.

따라서, 본 연구에서는 비정형 데이터 구조를 그래프 형태로 표현하여 관측 주변의 기상현상을 표현한다. 그래프 구조로 관측과 수치예보 모델의 공간정보를 표현하면, 관측정보의 손실을 최소화할 수 있으며, 서로 다른 데이터 구조를 가지는 자료를 융합하기에도 효율적이다. 기상현상의 공간적 특성을 표현할 수 있는 그래프를 기상 네트워크(meteorological network)로 정의하고, $G^{hetero} = \langle V, E \rangle$ 로 수식화할 수 있다. V 는 개별 관측 또는 수치 모델의 격자점을 나타내며, E 는 관측 혹은 격자점의 인접 관계로 정의한다. 여기서 격자점과 관측점을 서로 다른 node type으로 표현하며, 관측점은 더 세분화하여 관측 종별로 node type을 구별한다. 따라서, G 는 heterogeneous graph로 정의할 수 있으며, 관측 종에 따라 구분되는 특성을 학습하기에 효율적이다.

관측변수의 개수와 물리량은 관측 종마다 서로 다르므로, 각 노드는 서로 다른 크기의 node feature를 가진다. 본 연구에서는 예측 모델이 pretext task를 해결함으로써, 고정된 크기의 node feature vector로 관측의 역할을 표현한다. 각 관측의 특징적인 역할을 이해할 수 있도록 설계된 pretext task를 통해 self-supervised learning(SSL)은 원본 데이터보다 정보를 풍부하게 학습하여 downstream task에서 더 개선된 일반화 성능 및 강건함을 달성할 수 있게 된다. 따라서, 본 연구에서는 관측의 종류에 따른 node classification을 pretext task로 수행하여 관

측 종별 특성을 pre-training 할 수 있다. 그래프 인코더를 f_θ , pretext 디코더를 p_ϕ 라 하면, 자기 지도 학습 방법은,

$$\theta^*, \phi^* = \arg \min_{\theta, \phi} L_{ssl}(f_\theta, p_\phi, D)$$

로 수식화할 수 있다. 여기서 D 는 그래프 G 의 degree를 나타내는 행렬이며, L_{ssl} 은 본 연구에서 설계된 pretext task에 따라 pretext 디코더의 출력을 결정하는 SSL 손실함수이다.

훈련된 그래프 인코더 f_θ 를 활용하여 생성한 node feature vector는 실제에 가까운 대기 상태를 추정하는 데에 사용된다. 대기 상태를 추정하는 이 작업을 downstream task로 정의하고, downstream 디코더를 q_ψ 라 하면, 다음과 같이 그래프 지도학습 task를 구성할 수 있다;

$$\theta^{**}, \psi^* = \arg \min_{\theta^*, \psi} L_{down}(f_{\theta^*}, q_\psi, G, y).$$

여기서 y 는 추정하고자 하는 대기 상태 값을 나타내며, L_{down} 은 모델을 downstream task에 대해 학습시키는 지도학습 손실함수이다.

3. 사전실험 결과

본 연구에서는 실제 기상청 현업에서 사용되는 multi-modal 기상관측 데이터와 수치예보 모델 데이터를 확보하였다. 수치예보 시스템에서 자료동화 모델의 결과인 분석장을 실제 대기 상태로 가정하였으며, 사용되는 관측값은 기상청의 관측 전처리 시스템인 KPOP으로 정제된 데이터를 활용하였다. 관측 종은 위성 및 지상관측인 AIRCRAFT, GPSRO, SONDE, AMV, AMSU-A, AMSR2, ATMS, CRIS, GK2A, IASI, MHS로 총 11종이 사용되었다. 학습에 사용된 기간은 2021. 5. 1. 06UTC-2021. 5. 7. 00UTC 이며, 검증에 사용된 기간은 2021. 5. 7. 06UTC-2021. 5. 10. 18UTC이다.

실험 결과는 아래 표1과 같으며, 동아시아 지역의 중층 고도의 바람, 온도, 습도의 상태를 추정한 결과이다. 본 연구에서 제안하는 graph attention network(GAT) (w/ssl)은 mean square error(MSE)와 mean absolute error(MAE)가 다른 비교 모델(fully-connected neural network(FCN)와 graph neural network(GNN), GAT)보다 작으므로, 최우수 성능을 달성한 모델임을 확인할 수 있다. 비가장 높은 정확도를 달성한 비교모델인 GAT와 비교하면 MAE기준으로 10배의 성능 개선을 확인할 수 있다. 관측의 역할을 SSL 방식으로 학습한 관측의 feature

vector는 인접 지역의 정보를 단순히 aggregation 한 feature vector보다 예측에 효과적인 것으로 보인다. 인접 관측과 대기 상태를 바탕으로 관측의 역할을 선행 학습하는 모델은 대기 상태를 추정하는 데 있어 정확도를 개선하는 데에 효과적이다. 추가적으로, FCN 모델은 공간적 상관관계를 학습할 수 없어 정확도가 낮은 반면, GNN과 GAT 모델의 경우 관측과 인접한 지역의 대기 상태를 고려하므로 비교적 높은 성능을 유지하는 것으로 보인다. 관측을 포함한 격자점 주변의 대기 상태는 현재 대기 상태를 추정하는데 중요한 정보를 제공하는 것으로 생각된다.

<표 1> 각 모델의 정확도 비교

Models	MSE	MAE
FCN	0.159	0.245
GNN	0.020	0.196
GAT	<u>0.016</u>	<u>0.175</u>
GAT (w/ssl)	0.004	0.075

4. 결론

본 연구에서 제안하는 multi-modal 기상관측을 융합하는 그래프 자기 지도학습 방법은 실제로 가까운 대기 상태를 추정할 수 있는 새로운 접근법이다. 제안하는 모델은 기존의 전통적인 방법의 비용 및 시간 효율성을 개선함으로써, 이상 관측 탐지, 관측의 편차 보정, 관측 영향 추정 등 다양한 관측 전처리 기술에 활용할 수 있을 것이다. 특히, 관측이 예보에 미치는 영향을 평가하는 것은 관측 전처리 성능을 개선하는 데에 필수적이다. 본 연구에서 제안하는 SSL 학습 방법을 기상 네트워크에 적용하고, 네트워크에서 어떤 관측이 대기 상태를 추정하는데 높은 가중치로 참여했는지 설명가능한 인공지능 기법을 접목할 수 있으며, 이러한 방법을 통해 관측의 영향을 추정하는 방향으로 연구를 확장할 수 있을 것이다. 향후 연구에서는 제안하는 모델의 우수성을 다방면으로 검증하기 위한 실증을 설계하고자 한다.

사사

이 연구는 기상청 출연사업인 (재)차세대수치예보모델개발사업단의 4차원 고품질 기상분석을 위한 최신 자료동화기술 개발(KMA2020-02211)의 지원을 받아 수행되었음.

참고문헌

- [1] Konapala, G., Kumar, S. V., & Ahmad, S. K. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 163–173, (2021).
- [2] Wang, S., Chen, J., Rao, Y., Liu, L., Wang, W., & Dong, Q. Response of winter wheat to spring frost from a remote sensing perspective: Damage estimation and influential factors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 168, 221–235, (2020).
- [3] Maffei, C., Lindenbergh, R., & Menenti, M. Combining multi-spectral and thermal remote sensing to predict forest fire characteristics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 400–412, (2021).
- [4] Bauer, P., Thorpe, A., & Brunet, G. The quiet revolution of numerical weather prediction. *Nature*, 525(7567), 47–55, (2015).
- [5] Hoang, V. T., Jeon, H. J., You, E. S., Yoon, Y., Jung, S., & Lee, O. J. Graph Representation Learning and Its Applications: A Survey. *Sensors*, 23(8), 4168, (2023).
- [6] Jeon, H. J., Choi, M. W., & Lee, O. J. Day-Ahead Hourly Solar Irradiance Forecasting Based on Multi-Attributed Spatio-Temporal Graph Convolutional Network. *Sensors*, 22(19), 7179, (2022).
- [7] Yano, K., Shiina, T., Kurata, S., Kato, A., Komaki, F., Sakai, S. I., & Hirata, N. Graph partitioning based convolutional neural network for earthquake detection using a seismic array. *Journal of Geophysical Research: Solid Earth*, 126(5), e2020JB020269, (2021).
- [8] Yang, Y., Dong, J., Sun, X., Lima, E., Mu, Q., & Wang, X. A CFCC-LSTM model for sea surface temperature prediction. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 207–211, (2017).
- [9] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., ... & Battaglia, P. GraphCast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, (2022).
- [10] Jeon, H. J., Kang, J. H., Kwon, I. H., What Meteorological Characteristics Do Affect Weather Forecasting? In Proceedings of the 3rd International Conference on Human-centered Artificial Intelligence (Computing4Human 2022), Hanoi, Vietnam, 2022, 681–690.