기상 데이터 이미지화를 통해 관측소 위치 정보를 활용한 ConvLSTM 기반 미세먼지 농도 예측 모델

김성윤○ 차영은 송교원 양희철
충남대학교 컴퓨터융합학부
sykim1106@naver.com, cye.dev@gmail.com, tim000519@o.cnu.ac.kr, hcyang@cnu.ac.kr

ConvLSTM-based Particulate Matter Concentration Prediction Model with Location of Weather Station by Meteorological Data Imagification

Seongyoon Kim○ Youngeun Cha Kyowon Song Heecheol Yang
Division of Computer Convergence, Chungnam National Univ.

요 약

본 논문에서는 ConvLSTM (Convolutional Long Short-Term Memory) 기반의 미세먼지 농도 예측 모델을 제안한다. 관측소의 위치 정보를 이미지화한 데이터를 통해 제안한 모델이 미세먼지 농도 데이터와 풍향, 풍속 데이터의 관측소 위치에 따른 영향을 학습할 수 있도록 한다. 이미지화한 데이터를 ConvLSTM 계층에 통과시키고, 배치 정규화를 진행함으로써 관측소 위치에 따른 영향과 시계열 데이터의 특성을 모두 학습할 수 있도록 하였다. 또한, 제안 모델에 대하여 이미지화한 풍향, 풍속 데이터 유무에 따른 성능을 비교하기 위해 기상 데이터의 종류별 성능 변화를 관찰하였다. 미세먼지 농도, 풍향, 풍속 데이터를 모두 학습한 모델이 미세먼지 농도 데이터만 학습한 모델에 비해 MSE (Mean Squared Error) 지표 기준 약 40% 낮은 수치를 보였다. 미세먼지 농도, 풍향 데이터만 학습한 모델과 미세먼지 농도, 풍속 데이터만 학습한 모델에 대해서도 모든 데이터를 학습한 모델이 약 18% 낮은 수치를 보였다.

1. 서 론

미세먼지는 눈에 보이지 않을 정도로 입자가 아주 작은 먼지로, 각종 질환을 유발하는 대기 오염물질이다. 세계보건기구(WHO)에서는 2013년 미세먼지를 발암물질로 규정하였다 [1].

미세먼지 농도를 예측하기 위해 머신러닝을 사용한 연구들이 진행되고 있다. 그 중, 선행연구 [2]에서는 미세먼지와 풍향, 풍속의 상관관계를 분석하고, LSTM (Long Short-Term Memory) 모델을 사용해 PM_{10} (Particulate Matter less than 10 microns) 농도를 예측하는 연구를 수행하였다 [3]. 선행연구 [4]와 [5]에서는 미세먼지 농도 정보를 이미지에 대응시켜 모델이 관측소의 위치를 고려하여 미세먼지 농도를 예측하도록 연구를 수행하였다. 선행연구 [4]에서는 대전시 PM_{10} 관측소 8곳에 대해서 3×3 그리드를 통해 상대적인 위치를 표현하였다. 선행연구 [5]에서는 90×70 그리드 이미지에 미세먼지 농도 데이터를 표현하였다.

본 논문에서는 풍향, 풍속과 미세먼지 농도 사이의 상관관계를 관측소 위치 정보까지 고려하여 활용하는 모델을 제시한다. 풍향, 풍속은 미세먼지 농도에 큰 영향을 주는 변수지만, 선행연구에서는 풍향, 풍속을 측정한 관측소들 사이의 위치 관계를 고려하지 않았다. 제안하는 모델은 데이터가 풍향, 풍속이 관측된 위치 정보도 포함할 수 있도록 이미지화시켜 학습에 활용하였다. 관측소 위치에 따른 영향과 시계열 데이터의 특성을 모두 학습할 수 있도록 ConvLSTM (Convolutional LSTM) 기반의 미세먼지 농도 예측 모델을 제안하였다. ConvLSTM은 LSTM과 CNN (Convolutional Neural Network)을 결합한 형태로, 입력 데이터의 공간적 구조를 인식하는 시계열 예측 모델이다 [6]. 제안한 ConvLSTM 기반 미세먼지 농도 예측 모델은 테스트 데이터셋에 대해 92.5의 MSE (Mean Squared Error)를 가져, 위치 정보를 고려하지 않은 RNN (Recurrent Neural Networks)과 LSTM보다 우수한 예측 성능을 보였다. 또한 풍향, 풍속을 모두 고려하지 않았을 때와, 풍향, 풍속 중 하나만 고려하였을 때에 비해 우수한 성능을 보여 풍향, 풍속 데이터가 성능 향상에 도움이 되었음을 확인하였다.

2. 제안 방법

본 절에서는 관측소의 위치 정보를 고려하기 위한 기상 데이터셋 전처리 기법과, 제안하는 ConvLSTM 기반의 미세먼지 농도 예측 모델에 대해서 자세히 기술한다.

2.1 데이터셋

본 연구에서는 기상청 기상자료개방포털 [7]에서 제공하는 방재기상관측과 황사관측 데이터셋을 사용하였다. 방재기상관측정보의 경우 전국 510개 관측소의 측정치를

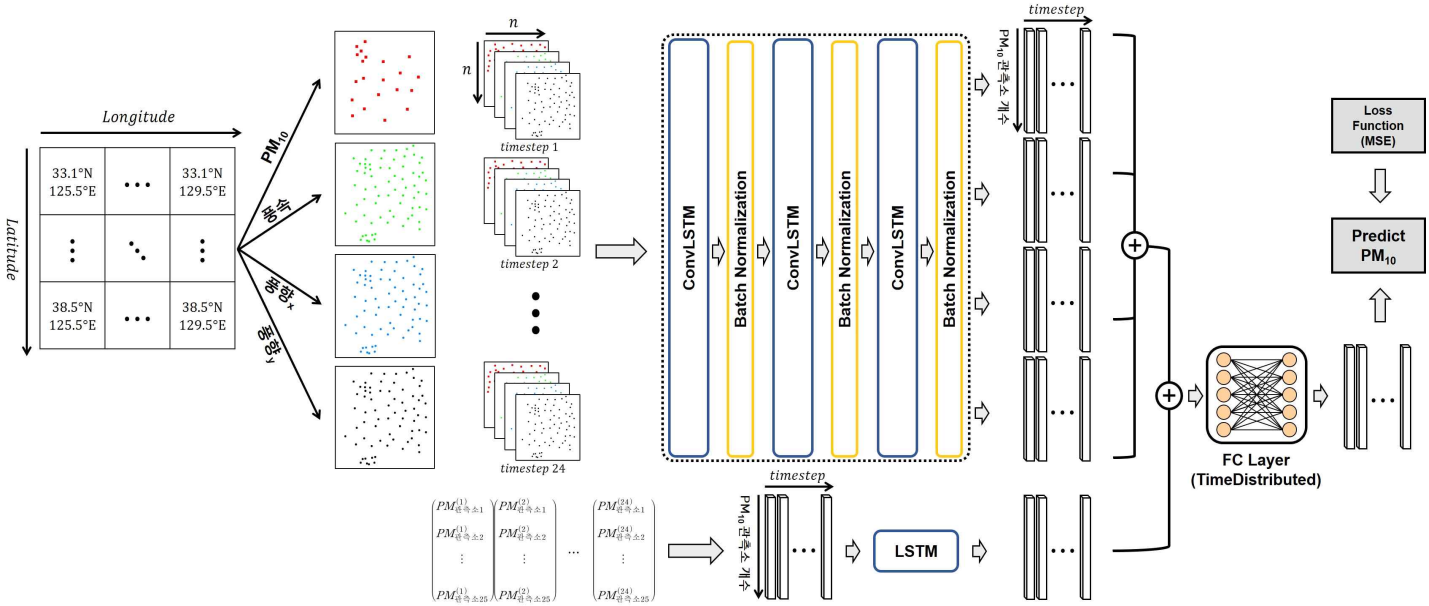

그림 1. 기상 데이터 이미지화를 위한 전처리 과정 및 미세먼지 농도 예측을 위한 모델 구조도

제공하며, 황사관측의 경우 전국 25개 관측소의 측정치를 제공한다. 2019년 1월 1일 0시부터 2022년 12월 31일 23시까지 총 4년간 1시간마다 측정된 데이터를 활용하였다.

2.2 데이터 전처리

방재기상관측 데이터 중 풍향, 풍속 데이터를 사용하였다. 또한, 선형보간법을 활용하여 결측치를 처리하였다.

2.2.1 관측소 위치 정규화

관측소의 기상 데이터를 이미지화하기 위해, 관측소의 실제 위도와 경도를 이미지 픽셀 위치에 대응이 되도록 MinMax 정규화를 진행하였다. 관측소의 개수가 더 많은 방재기상관측소를 기준으로 정규화하였다. 방재기상관측소 s_i의 실제 위도를 $lat(s_i)$, 경도를 $lon(s_i)$라 하고, 관측소 s_i의 위도와 경도의 최댓값, 최솟값을 각각 $lat_{\max}$, $lat_{\min}$, $lon_{\max}$, $lon_{\min}$라 하였을 때, 이미지 픽셀 수에 맞게 정규화한 위도 $norm_lat(s_i)$와 경도 $norm_lon(s_i)$는 다음과 같이 표현할 수 있다.

$$norm\_lat(s_i) = \left\lfloor \frac{lat(s_i) - lat_{\min}}{lat_{\max} - lat_{\min}} \times H_{img} \right\rfloor$$

$$norm\_lon(s_i) = \left\lfloor \frac{lon(s_i) - lon_{\min}}{lon_{\max} - lon_{\min}} \times W_{img} \right\rfloor$$

위 식에서 H_{img}, W_{img}는 각각 모델에 입력할 이미지의 높이, 너비를 의미한다. 방재기상관측소를 기준으로 $(W_{img}, H_{img}) = (224, 224)$일 때 관측소의 위치가 각 픽셀에 겹치지 않고 대응되어 해당 값을 사용하였다.

2.2.2 기상 데이터 이미지화

관측소별 기상 데이터를 각 관측소의 위치에 기반하여 224×224 크기로 이미지화하였다. 시간별 기상 관측소의 기상 데이터는 4개의 채널로 표현된다. 채널1은 PM_{10}의 데이터, 채널2는 풍속의 데이터를 저장한다. 풍향의 경우, 0°부터 360°까지의 각도로 표현되어 있어 이를 좌표평면상에 변환 후 x축 값은 채널3에, y축 값은 채널4에 저장한다.

2.3 미세먼지 농도 예측 모델

본 연구에서는 ConvLSTM, LSTM을 이용한 미세먼지 농도 예측 모델을 설계하였다. 평가 지표는 MSE를 사용하여 모델의 예측 성능을 측정하였다.

우선, 앞서 정규화한 관측소 위치에 기반해 이미지화한 기상 데이터를 제안하는 모델의 학습 데이터로 활용하였다. 예측 모델은 전날 24시간의 기상 데이터를 입력으로 다음 날 24시간의 미세먼지 농도를 예측한다. 전체 관측소의 시간별 기상 데이터를 2.2.1에서 소개한 관측소 위치 정규화를 통해 224×224 크기의 이미지에 대응하였다. 이미지는 4개의 채널로, 각각 PM_{10}, 풍속, x축 방향 풍향, y축 방향 풍향으로 구성된다. 예측 모델은 24개의 이미지를 입력받는 ConvLSTM 서브 모델과 지역별 24시간의 PM_{10} 수치 데이터를 입력받는 LSTM 서브 모델로 구성되며, 이를 그림 1로 표현하였다. ConvLSTM 서브 모델에는 $(\text{시간}, H_{img}, W_{img}, \text{채널}) = (24, 224, 224, 4)$ 크기의 시계열 이미지 데이터셋이 입력되며, ConvLSTM 계층과 배치 정규화 계층이 연결된 블록 3개로 구성된다. 서브 모델의 출력의 크기는 $(25, 96)$으로, 미세먼지 농도 관측소 25개에 대한 24시간의 예측 결과를 출력하는 것으로 설정하였다. ConvLSTM 서브 모델에는 4개 채널의

데이터를 모두 입력으로 가지므로, LSTM 서브 모델과의 비중을 고려하여 4배의 출력 크기를 가지도록 하였다. LSTM 서브 모델에는 (관측소,시간)=(25,24) 크기의 지역별 24시간의 PM_{10} 수치 데이터가 입력되며, 같은 크기의 출력을 가지도록 하였다. 각 서브 모델의 출력을 (25,120)의 크기로 결합한 후, TimeDistributed 계층을 통과하여 (25,24)의 크기로 출력되도록 하였다. 전체 모델의 출력은 25개의 관측소의 24시간의 PM_{10} 예측 농도를 의미한다.

3. 실험 결과

학습 데이터로는 2019년부터 2021년까지 총 3년 동안의 데이터셋을 사용하였고, 테스트 데이터는 2022년 1년 동안의 데이터를 사용하였다.

	Model	MSE
Baseline	전날 데이터로 예측	2455.27
	RNN(3 layers, PM_{10}+풍향+풍속)	265.60
	LSTM(3 layers, PM_{10}+풍향+풍속)	277.97
제안 모델	ConvLSTM(PM_{10})	130.07
	ConvLSTM(PM_{10}+풍향)	112.28
	ConvLSTM(PM_{10}+풍속)	109.25
	ConvLSTM(PM_{10}+풍향+풍속)	**92.49**

표 1. 모델별 성능 비교

각 모델로부터 MSE를 산출한 결과를 표 1에 나타내었다. Baseline 모델로 전날의 PM_{10} 농도를 다음 날의 농도와 같다고 예측하는 모델, RNN 모델, LSTM 모델로 설정하였고, 입력은 Convolutional Layer를 활용하지 않는 모델이므로, 이미지가 아닌 시계열 데이터로 하였다.

제안 모델에 대해 기상 데이터의 조합별 성능을 확인하였고, 모두 baseline 모델보다 우수한 성능을 보였다. 기상 데이터 조합별 MSE는 PM_{10}만을 입력하였을 때 130.07로 가장 높았고, PM_{10}, 풍향, 풍속 데이터를 모두 입력하였을 때 92.49로 가장 낮았다. 따라서 풍향, 풍속 정보를 모두 반영하였을 때 예측 성능이 가장 우수함을 확인하였다. 제안 모델의 기상 데이터 조합별 테스트 데이터에 대한 epoch별 예측 성능은 그림 2와 같다.

4. 결 론

본 연구에서는 미세먼지 농도 예측을 위해 기상 데이터 이미지화를 통해 관측소 위치 정보를 활용한 ConvLSTM 기반 미세먼지 농도 예측 모델을 제안하였다. 미세먼지 농도를 관측소의 위치에 기반하여 이미지화하는 것이 예측 성능 향상에 기여함을 알 수 있었다. 또한 풍향, 풍속을 조합한 데이터셋에 대해 가장 우수한 성능을 보여 풍향, 풍속 정보를 관측소의 위치 정보와 결합하여 이미지화하는 것이 미세먼지 농도 예측에 유용하게 활용될 수 있음을 보였다.

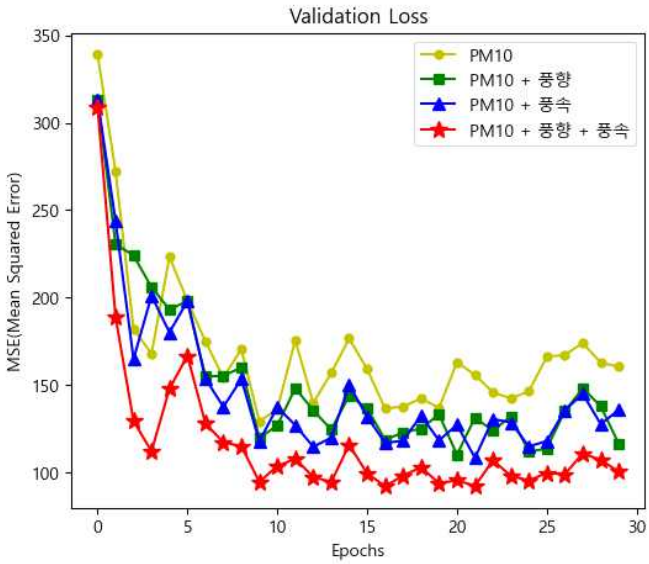

그림 2. 기상 데이터 조합별 예측 성능

참 고 문 헌

[1] the International Agency for Research on Cancer (IARC), “IARC: Outdoor air pollution a leading environmental cause of cancer deaths”, 17 October 2013

[2] 조수현, 정미리, 이진향, 오일석, 한영태 , “풍향풍속과 미세먼지의 상관관계 분석과 LSTM을 이용한 미세먼지 예측”, 한국정보과학회 학술논문발표집, pp.1649-1651, 2020.

[3] Hochreiter, S., & Schmidhuber, J., “Long Short-Term Memory”, Neural Computation, 9(8), pp.1735-1780, 1997.

[4] 이홍석, 부이 칵 남, 선충녕, “도심지 교통흐름 및 미세먼지 예측을 위한 딥러닝 LSTM 프레임워크”, 정보과학회논문지, 47권 3호, pp.292-287, 2020.

[5] 이준민, 김경태, 최재영, “미세먼지 농도 예측을 위한 어텐션 기반 합성곱 장단기 메모리 모델”, 한국멀티미디어학회논문지, 26권 8호, pp.911-924, 2023.

[6] S. H. I. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting“, Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, pp.802-810, 2015.

[7] 기상자료개방포털, https://data.kma.go.kr