미세먼지 예측 성능 개선을 위한 CNN-LSTM 결합 방법

CNN-LSTM Combination Method for Improving Particular Matter Contamination (PM2.5) Prediction Accuracy

저자
(Authors)

황철현, 신강욱
Chul-Hyun Hwang, Kwang-Wook Shin

출처
(Source)

한국정보통신학회논문지 24(1), 2020.1, 57-64(8 pages)
Journal of the Korea Institute of Information and Communication Engineering 24(1), 2020.1, 57-64(8 pages)

발행처
(Publisher)

한국정보통신학회
The Korea Institute of Information and Communication Engineering

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09298393

APA Style

황철현, 신강욱 (2020). 미세먼지 예측 성능 개선을 위한 CNN-LSTM 결합 방법. 한국정보통신학회논문지, 24(1), 57-64

JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 1: 57~64, Jan. 2020

미세먼지 예측 성능 개선을 위한 CNN-LSTM 결합 방법

황철현[1*] · 신강욱[2]

CNN-LSTM Combination Method for Improving Particular Matter Contamination ($PM_{2.5}$) Prediction Accuracy

Chul-Hyun Hwang[1*] · Kwang-Wook Shin[2]

[1*]Associate Professor, Department of Smart IT Software, Kyoung-Bok University, Gyeonggi, 12051 Korea
[2]Senior Researcher, K-Water Institute, Korea Water Resources Corporation, Daejeon, 34045 Korea

요 약

최근 IoT 센서의 확산과 빅데이터, 인공지능 관련 기술의 발전으로 인해 미세먼지 오염도에 대한 시계열 예측 관련 연구가 활발하게 진행되고 있다. 하지만 미세먼지 오염도를 나타내는 데이터가 급격히 변하는 특성(Extreme)을 가지고 있어 기존의 시계열 예측방법으로는 현장에서 사용할 수 있는 수준의 정확도를 내지 못하고 있다. 이 논문에서는 LSTM을 활용하여 미세먼지 오염도를 예측할 때 CNN을 통한 환경상황을 분류한 결과를 반영하는 방법을 제안한다. 이 방법은 LSTM과 CNN이 독립적이지만 인터페이스를 통해 하나의 네트워크로 통합되기 때문에, 응용 LSTM보다 이해하기 쉽다. Beijing $PM_{2.5}$ 데이터를 활용한 제안 방법의 검증 실험에서 예측 정확도와 변화 시기에 대한 예측력이 다양한 실험 case에서 일관되게 향상된 결과를 보였다.

ABSTRACT

Recently, due to the proliferation of IoT sensors, the development of big data and artificial intelligence, time series prediction research on fine dust pollution is actively conducted. However, because the data representing fine dust contamination changes rapidly, traditional time series prediction methods do not provide a level of accuracy that can be used in the field. In this paper, we propose a method that reflects the classification results of environmental conditions through CNN when predicting micro dust contamination using LSTM. Although LSTM and CNN are independent, they are integrated into one network through the interface, so this method is easier to understand than the application LSTM. In the verification experiments of the proposed method using Beijing $PM_{2.5}$ data, the prediction accuracy and predictive power for the timing of change were consistently improved in various experimental cases.

키워드 : 딥러닝, 컨볼루션 신경망, LSTM, 시계열 예측, 미세먼지

Keywords : Deep Learning, CNN, LSTM, IoT, $PM_{2.5}$

Received 7 November 2019, **Revised** 17 November 2019, **Accepted** 2 December 2019

*** Corresponding Author** Cheol-Hyun Hwang(e-mail:cheolhyun.hwang@gmail.com, Tel:+82-31-570-9608)
Associate Professor, Department of Smart IT Software, Kyoung-Bok University, Gyeonggi, 12051 Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.1.57 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

미세먼지 오염문제는 실생활에서 항상 접하기 때문에 건강에 직접적인 영향을 미치는 것 뿐 만 아니라 대다수 국민이 인지하고 있어 정책적으로 매우 관심도가 높은 주요한 환경 문제이다. 이에 따라 미세먼지 오염문제를 해결하기 위해 국가적 차원에서 데이터 수집과 분석 등 관련 연구를 지원하고 있다. 특히 미세먼지 오염도를 예측하는 것은 경보발령 등을 통해 미세먼지가 국민의 건강에 미치는 영향을 줄이는데 중요한 역할을 담당한다.

미세먼지 오염도를 예측하는 연구로는 로지스틱 회귀분석, SVM(Support Vector Machine)과 같은 통계적 접근 방법을 통한 전통적인 예측 모형이 주로 수행되며 최근에는 딥러닝과 같은 인공지능 관련 기술을 적용한 연구가 더욱 확대되고 있다.[1]

하지만 딥러닝 방법 중 시계열 예측에 주로 활용되는 LSTM(Long Short Term Memory)은 추세가 안정적으로 유지되고 데이터가 급격히 변하지 않을 때 유용하지만 미세먼지 오염도 같이 데이터 값이 급격히 변화하는 극단적인 데이터(Extreme Data)에는 적합하지 않다. 이러한 문제점을 해결하기 위해 하나의 예측모델에만 의존하지 않고 여러 예측모델을 결합한 모델에 대한 연구가 성과를 보이고 있다. [2][3][4]

이 논문은 극단적으로 변하는 미세먼지 오염도 데이터에서 시계열 예측 성능을 향상시키기 위해 딥러닝 방법인 CNN(Convolutional Neural Network)과 LSTM을 결합하는 방법을 제시하는 것을 목표로 한다. 이를 위해 우선, 미세먼지 오염도를 제외한 환경측정 데이터를 CNN에 입력하여 현재 환경상황을 분류하고, 과거 동일 환경상황에서 추출한 미세먼지 오염도 추세정보와 직전의 미세먼지 오염도 데이터를 활용하여 LSTM에 입력한 다음, 최종적으로 미래의 미세먼지 오염도 값을 예측하는 방법을 제안한다.

제안 방법의 효과를 검증하기 위해 공개된 Beijing $PM_{2.5}$를 실험데이터로 활용하여, 기존 LSTM과 제안 방법인 CNN-LSTM 결합 모델의 예측 정확도를 비교하는 실험과 함께, 미세먼지 오염도를 나타내는 $PM_{2.5}$ 값이 급격히 변화하는 시점을 예측하는 분류 정확도 시험을 수행하여 예측 모델이 급격히 변화하는 시계열 데이터에 얼마나 효과적으로 대응하는지를 실험하였다.

본 논문의 구성은 다음과 같다. 2장은 이론적 배경을 통해 미세먼지 오염도 예측을 위한 기존연구결과를 검토하고 한계를 제시한다. 3장은 논문의 제안 방법인 CNN-LSTM을 결합한 미세먼지 오염도 예측 방법에 대해 제시한다. 4장은 실험 환경과 결과 분석을 제시하고, 5장에서는 결론과 향후 연구 과제를 제시한다.

Ⅱ. 이론적 배경

미세먼지 오염도에 대한 예측 연구는 이미지 데이터 기반 분석, 시계열 데이터 기반 분석, 이미지와 시계열 데이터를 동시 고려한 분석의 3가지 측면에 서 시도 되어 왔다.

2.1. 이미지 분석을 통한 미세먼지 오염도 추정

디지털 촬영기기가 보편화되고 이미지를 축적하고 분석하는 기술의 발전으로 인해 위성사진이나 CCTV 등을 통해 촬영된 이미지를 활용하여 미세먼지 오염도를 추정하는 연구가 수행되었다. 이 연구는 촬영된 이미지에서 특징을 추출한 후 이미지 특징과 미세먼지 오염도와 상관관계를 구하거나, 이미지 분류를 위한 딥러닝을 활용하여 미세먼저 오염도를 추정하였다.

먼저 상관관계를 이용하는 방법은 위성이나 CCTV로 촬영된 이미지의 특징(contrast, color, 날짜, 촬영 시간, 기상 상태)과 미세먼지 오염도와 상관관계를 통해 새로운 이미지에 대한 특징을 추출한 후 상관관계에 따른 미세먼지 오염도를 추정하는 방법이다.[5] 딥러닝을 활용하는 방법은 이미지 분류분야에서 뛰어난 성능을 보이고 있는 CNN을 활용하여 과거 이미지에 대한 미세먼지 오염도를 학습시켜 새로운 이미지의 미세먼지 오염도를 추정하는 방법이다.[1]

이미지 분석은 이미 설치된 촬영 장비에서 수집된 이미지를 활용하여 미세먼지 오염도를 추정하는데 효과를 보이지만, 최근 미세먼지 오염도에 대한 측정이 용이해 지고 센서 보급이 늘어나면서 추정만으로는 실효성이 없다는 문제 제기가 있다.

2.2. 시계열 예측을 통한 미세먼지 오염도 예측

데이터가 과거 수치 정보를 사용할 수 있고 과거 몇 가지 패턴이 미래에도 지속될 것이라는 가정 하에 주로

사용되는 방법이 시계열 예측 모델이다.

미세먼지 오염도 예측은 IoT센서로부터 $PM_{2.5}$, 기온, 강우, 습도와 같은 환경 정보를 일정한 간격으로 수집하여 기록하고 미세먼지는 환경의 영향에 따라 변하기 때문에 시계열 예측에 적합하다.

과거, 미세먼지 오염도와 같은 환경 영역의 시계열 예측은 지수 평할, ARIMA와 같은 통계적 모델을 주로 활용하였지만 최근에는 LSTM과 같은 딥러닝 기술을 활용한 예측이 더 높은 예측 성능을 제공하는 것으로 알려져 있다. 특히 LSTM은 환경 센서와 같이 IoT 기반의 정밀 데이터 영역에서 더욱 효과적이다.[6][7]

따라서 최근 미세먼지 오염도 예측에서 LSTM과 같은 딥러닝 모델을 사용하는 연구는 증가하고 있다. 하지만 LSTM이 추세를 유지하고 정밀한 시계열 데이터에 대해서는 좋은 성능을 보여주지만, 추세가 유지되지 않고 변화가 심한 데이터(extreme data)에 대해서는 예측값이 특정 상수 값으로 수렴하게 되어 정확도가 저하되는 문제를 보여 왔다. 또한 풍속, 강우량과 같은 환경 변수를 동시에 고려하지 못하는 단점이 있다.

2.3. CNN과 LSTM의 결합

앞 절에서 제시한 바와 같이 LSTM의 정확도가 저하되는 문제점과 주변 환경변수를 고려하지 못하는 문제점을 보완하기 위해 이미지 분류를 기반으로 한 CNN과 시계열 예측 방법인 LSTM을 결합하여 미세먼지 오염도를 예측하는 연구가 시도되고 있다. CNN과 LSTM을 결합하기 위해 모델 Layer 수준에서 결합하는 방법과 독립적인 모델의 결합 방법이 연구되었다.

첫 번째 방법은 CNN의 convolution layer와 LSTM을 결합하는 방식으로 환경 데이터에 대한 누적 값을 입력 데이터로 활용한다. 이 모델은 APNet이라는 명칭으로 발표되었으며 실험을 통해 다른 모델에 비해 높은 예측 정확도를 제공하는 것으로 증명되었다. 다음 그림 1은 CNN-LSTM 결합구조 사례인 APNet을 제시하였다. [2]

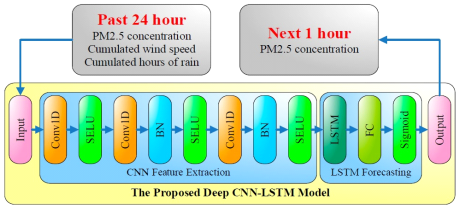

Fig. 1 The Architecture of APNet

하지만 APNet은 기존 기계학습 모델보다 좋은 성능을 보이는 장점을 가지지만 Neural Net의 Layer 차원에서 CNN과 LSTM이 결합되면서 구현이 복잡하고, 특정 데이터 환경에 맞춰져 응용이 힘들고, 장기 예측에 취약한 단점이 있다.

두 번째 방법은 CNN과 LSTM의 변형 모델인 ConvLSTM을 상호 독립적인 모델로 구성한다. 이 방법은 측정 지점 주위의 공간적인 특징을 모델링한다. 다음 그림 2는 공간 특징을 고려한 CNN-LSTM 결합 구조 사례를 제시하였다. [3]

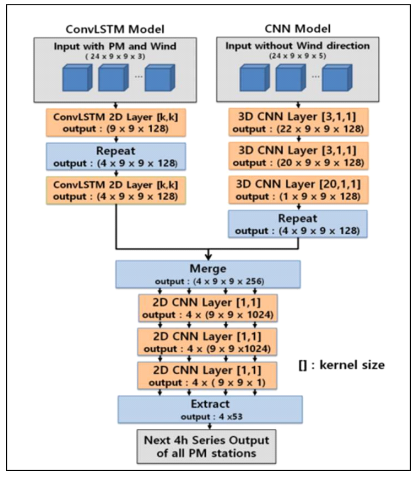

Fig. 2 Hybrid Model of ConvLSTM and CNN

이 방법은 측정하고자하는 주변 지역의 환경 측정값을 고려하여 대기 환경 분야의 특성을 반영하고, 정확도를 향상시킨 장점이 있으나 공간적 특성이 명확하고 분류 가능한 넓은 지역에만 활용할 수 있는 단점을 가진다. 도심과 같이 주변 측정소에서 유사한 측정값을 가진 환경에서 정밀한 예측을 수행하는데 제약이 있다.

Ⅲ. 제안 CNN-LSTM 결합 방법

제안한 CNN-LSTM 결합 방법은 CNN과 LSTM의 기본 모델을 사용하여 상호 독립적으로 구성하고 인터페이스를 통해 두 모델이 하나의 네트워크로 동작한다. 또한 LSTM에 입력되는 가장 최근의 데이터에 임의의 CNN에서 예측된 값을 추가 입력하여 미세먼지 오염도 데이터의 급격히 변화하는 특성을 반영한다.

3.1. 제안 결합 방법의 구조

제안된 CNN-LSTM 결합 방법은 상호 독립적으로 구성된 CNN과 LSTM의 기본 모형을 활용하여 Neural Network의 계층(Layer) 차원의 결합 없이 인터페이스로 연결된 네트워크 구조이다.

다음 그림 3은 제안한 CNN-LSTM 결합 방법에 대한 구조를 제시하였다.

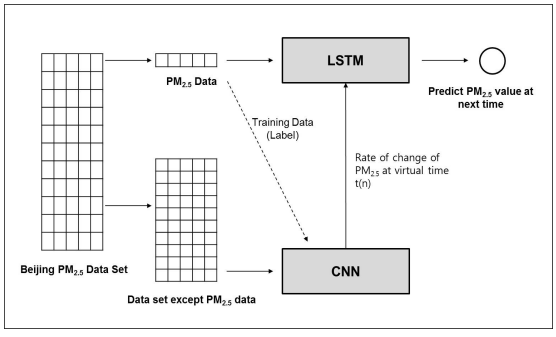

Fig. 3 The Architecture of Proposed CNN-LSTM Combination Method

제안된 결합 방법에 대한 수행 절차는 다음과 같다.

첫째, 환경 측정기기에서 수집한 환경 데이터 가운데 미세먼지 오염도를 제외한 모든 데이터를 CNN의 입력 데이터로 구성한다. 이 데이터는 측정 시점에 축정기기가 설치된 장소의 환경 상황 정보를 제공한다.

둘째, 환경 상황 정보를 기록한 데이터를 CNN에 입력하고 해당시간의 미세먼지 오염도 증감 추세를 Label 정보로 매칭하여 CNN을 학습시킨 후 환경 상황 정보만을 입력하여 환경 상황과 가장 매칭되는 미세먼지 오염도 증감 추세를 도출한다.

셋째, 가장 최근의 미세먼지 오염도를 측정한 시계열 값과 CNN의 출력 값을 미세먼지 오염도로 변환한 값을 활용하여 LSTM 입력 데이터를 생성한다. 생성된 입력 데이터를 활용하여 LSTM을 학습하고 예측한다.

이상과 같이 제시된 결합 방법을 수식으로 나타내면 다음 식 1과 같다.

$$P_{t+1} = L(P_{t..tn-1} \bullet f_{conv}(C(E_{t..tn-1} \bullet f_{cat}(P_{t..tn-1})) \bullet P_t)) \quad (1)$$

- f_{conv} : CNN 결과와 최종 $PM_{2.5}$ 값을 이용하여 $PM_{2.5}$ 값 형태로 변환
- f_{cat} : $PM_{2.5}$ 시계열 데이터의 기울기를 활용하여 CNN에서 활용할 Label을 추출
- P_t : 시점 t에서 $PM_{2.5}$ 값
- $P_{t..tn-1}$: t에서 tn-1까지의 $PM_{2.5}$ 구간 시계열 값
- $E_{t..tn-1}$: t에서 tn-1까지의 환경변수들의 구간 시계열 값 (2차원 행렬)
- L, C : LSTM, CNN의 학습과 예측(분류) 과정

3.2. CNN을 활용한 환경 상황 분류

제안된 방법에서 CNN은 예측 시점의 환경 상황을 판단하는 역할을 수행한다. 환경 상황이 유사한 과거 사례를 학습하여 새로운 환경 상황에서 미세먼지 오염도에 대한 증감 추세를 예측한다. 즉, 과거에 수집된 강우, 풍속 등의 시계열 환경 데이터를 입력하여 미세먼지 오염도에 대한 시계열 변화율을 산출한다.

미세먼지 오염도에 대한 시계열 변화율은 CNN 특성을 고려하여 범주형 변수(예 : 매우 높음, 높음, 유지, 낮음, 매우 낮음)로 산출한다. 즉 CNN 출력 값은 미세먼지 오염도가 아니라 Label 정보인 변화율 구간 정보를 산출한 값이다.

다음 그림 4는 제안한 결합 방법에서 CNN의 구조와 처리 절차를 제시하였다.

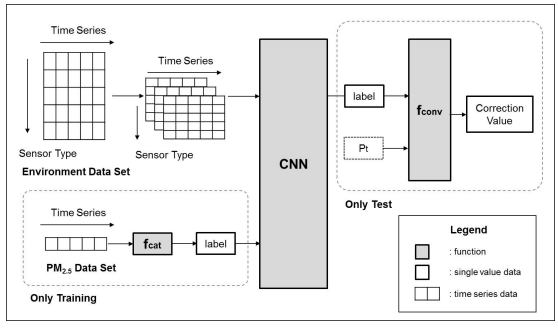

Fig. 4 Process Design of Proposed CNN Model

3.3. LSTM 모델을 활용한 시계열 예측 모델

제안 방법에서 LSTM은 가장 최근에 측정된 미세먼지 오염도 데이터를 학습하여 다음 주기의 미세먼지 오염도 데이터를 예측하는 역할을 수행한다.

최근 발생한 look_back-1개의 미세먼지 오염도 데이터와 CNN을 통해 산출된 1개의 추정 값을 결합하여

LSTM 예측에 필요한 look_back 데이터를 생성한다. CNN의 추정 값은 CNN의 출력 값인 추세 정보를 최근 발생한 미세먼지 오염도 값을 고려하여 최종 입력 데이터 으로 변환한 값이다.

다음 그림 5는 제안 방법에서 사용된 LSTM 모델의 구성을 제시하였다.

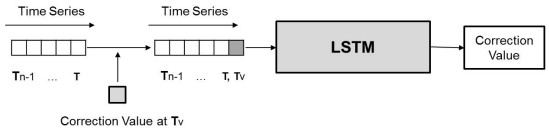

Fig. 5 Process Design of Proposed LSTM Model

Ⅳ. 실험 및 검증

4.1. 실험 방법 및 절차

제안 방법 검증을 위해 Beijing $PM_{2.5}$ 데이터를 활용하여 LSTM 단독모델로 구성된 기존 방법과 제안 결합 방법의 예측 성능을 비교하는 실험을 수행하였다.

실험에 사용된 Beijing $PM_{2.5}$ 데이터는 2010년 1월 1일부터 2014년 12월 31일까지 5년간 $PM_{2.5}$, 강우, 풍속 등 7개의 환경 측정 항목을 매시간 측정한 자료로 다수의 연구에서 사용된 공개된 데이터 세트이다.

총 41,757개의 데이터 가운데 70%인 29,229개를 학습 데이터로 활용하였고 30%인 12,528개는 검증 데이터로 활용하였다. 특히 실험 횟수를 증가시키기 위해 검증 구간을 5개로 분할하여 예측 실험을 수행하였고, $PM_{2.5}$를 제외한 6개의 환경데이터를 활용하였다. $PM_{2.5}$ 데이터에 대한 증감추세를 분류하기 위해 총 8개 증감 구간으로 구분하였고, 제안한 변환 과정을 거쳐 LSTM의 입력 값으로 사용하였다.

다음 표 1은 실험에 사용된 딥러닝 네트워크의 hyper parameter를 제시하였다.

Table. 1 Hyper Parameters of Deep Learning Networks

Network	Layer	Hyper Parameter
LSTM	-	Hidden Layer : 4 epoch : 50 Batch Size : 4 Loss Function : MSE Optimizer : Adam
CNN	Conv2D	Filter No : 32 (64) Kernel Size : 2*2 Activation : ReLu
	Pooling	Method : Max Stride : 2 Dropout : 0.1
	Dense	Node : 100(4) Activation : ReLu(SoftMax)

4.2. 실험 결과

실험 결과에 대한 예측 성능을 측정하기 위한 방법으로 RMSE(Root Mean Square Error) 활용하였다. RMSE는 시계열 데이터 예측에서 가장 보편적으로 사용되는 검증 방법이다.

다음 표 2에서는 예측 실험 단위인 5개 구간에 대한 기존 LSTM 방법과 제안 방법의 RMSE를 비교하여 제시하였다.

Table. 2 Comparison of Experiment Results

Section No.	Exist LSTM (RMSE)	Proposal Method (RMSE)
Section 1	17.0817	8.4291
Section 2	27.0759	9.4679
Section 3	23.6574	8.1427
Section 4	14.4274	9.2430
Section 5	20.9401	8.8565

표 2에서 제시된 바와 같이 실험에 적용된 5개 구간 모두에서 기존 LSTM을 수행했을 때 보다 제안 방법을 활용했을 때 RMSE가 낮게 산출 되었다. 따라서 예측 정확도는 개선 된 것으로 볼 수 있다. 개선 결과를 명확하게 알 수 있도록 표 3에서는 시험 구간에서 실제 발생한 값에 대한 기존 LSTM과 제안 방법에 의해 예측된 결과를 각각 비교하여 제시하였다.

Table. 3 Comparison of predicted values for the first 15 data occurrences

No.	Real Value	Exist LSTM	Proposal Method
1	104	90.87053	115.0198
2	118	104.1914	130.4518
3	124	117.9622	129.1657
4	116	122.3993	118.9944
5	99	112.883	100.262

No.	Real Value	Exist LSTM	Proposal Method
6	96	95.09254	92.49257
7	75	94.18356	79.52666
8	96	71.03513	107.2014
9	117	97.16846	112.1436
10	122	117.5845	118.6947
11	124	119.6792	127.0537
12	107	122.5249	108.5176
13	102	102.912	104.7821
14	70	99.70883	69.12551
15	71	64.74248	74.03382

다음은 제안 방법이 기존의 LSTM 예측에서 문제로 제시되었던 극단적인 데이터 환경에서도 좋은 예측 성능을 보여주고 있는지를 살펴보기 위한 분석을 제시하였다. 다음 그림 6은 극단적인 데이터가 발생한 시점을 표시하고 이때 제안 방법이 변화를 예측했는지를 시각적으로 표현하였다. 그림 6에서 보는 바와 제안 방법에 의 한 예측 결과는 극단적으로 변하는 데이터에 적응된 결과를 보여준다.

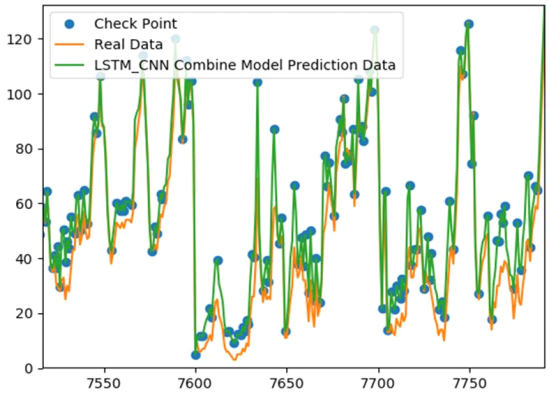

Fig. 6 Comparison of Extreme Data Occurrence

극단적으로 변하는 데이터에 대한 예측 성능을 그림으로 살펴본 결과를 제시하면 다음 표 4와 같다. 표 4는 실험 대상에서 극단적인 데이터가 발생된 시점에 대한 예측 성능을 살펴보기 위해 발생 횟수와 예측 횟수를 제시하였다.

시험 구간 전체에서 총 5,895개 시점에서 급격히 데이터가 변화하였고 이에 대해 기존 LSTM 모델은 2,995개를 예측하여 50.8%의 정확도를 보였으며, 제안 방법은 4,875개를 예측하여 82.7%의 정확도를 보여 제안 방법이 데이터가 급격히 변화하는 시점에 대한 예측력이 기존 LSTM보다 높다.

Table. 4 Extreme Data Generation and Prediction

Section No.	Number of All Count	Number of Hit Count (%)
Real Data	5,895	-
Exist LSTM	5,835	2,995 (50.8%)
Proposal Method	5,866	4,875 (82.7%)

다음 표 5은 예측 결과를 좀 더 구체적으로 비교하기 위해 전체 예측 구간에서 최초 15개 시점의 실제 값과 예측 값을 제시한다.

Table. 5 Comparison of predicted values for the first 15 data occurrences

No.	Real Value	Exist LSTM	Proposal Method
1	104	90.87053	115.0198
2	118	104.1914	130.4518
3	124	117.9622	129.1657
4	116	122.3993	118.9944
5	99	112.883	100.262
6	96	95.09254	92.49257
7	75	94.18356	79.52666
8	96	71.03513	107.2014
9	117	97.16846	112.1436
10	122	117.5845	118.6947
11	124	119.6792	127.0537
12	107	122.5249	108.5176
13	102	102.912	104.7821
14	70	99.70883	69.12551
15	71	64.74248	74.03382

4.3. 실험 결과 분석

$PM_{2.5}$에 대한 시계열예 측 결과, 전체 시험구간(5개)에서 제안방법의 예측정확도가 기존 LSTM 보다 높은 것으로 제시되었다. 따라서 제안 방법이 기존 방법에 비해 높은 예측 정밀도를 가진다고 할 수 있다.

다음 그림 7은 RMSE를 활용하여 기존과 제안방법에 대한 예측 정확도를 비교하여 제시하였다.

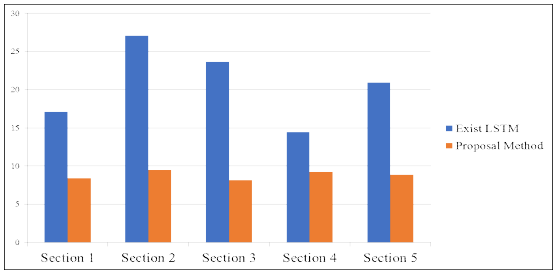

Fig. 7 Comparison of Exist and Proposal Method's RMSE

그림 8에서는 제안 방법이 극단적인 데이터가 발생한 시점에서 변화를 예측하였는지를 살펴보기 위해 정확도(Precision)와 재현율(Recall)에 근거한 F1 Score 산출결과를 기존 LSTM과 비교하여 제시하였다.

앞서 RMSE와 달리 F1 Score는 분류 정확도를 측정하는 평가기준으로서 급격한 변화가 발생하는 시점과 추세를 유지하는 통상적인 시점을 분류한 결과에 대한 예측성능을 측정하는데 적용된다. 이 성능 측정 방법은 급격히 변화하는 $PM_{2.5}$ 데이터 환경의 예측 성능을 산출하는데 유용하다.

F1 Score 산출 결과 기존 LSTM의 F1 Score는 0.51이며 제안 방법은 0.83으로 극단적인 데이터가 발생하는 시점을 예측하는 성능 역시 제안 방법이 우수하다.

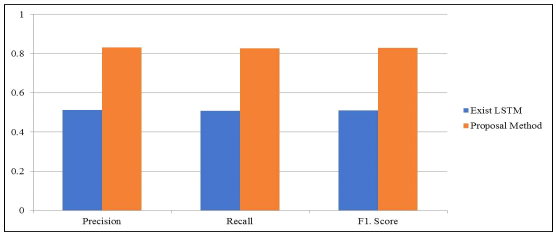

Fig. 8 Comparison of Exist and Proposal Method's F1 Score

그림 9에서는 극단적인 데이터 예측에 대한 구체적인 사례를 제시한다.

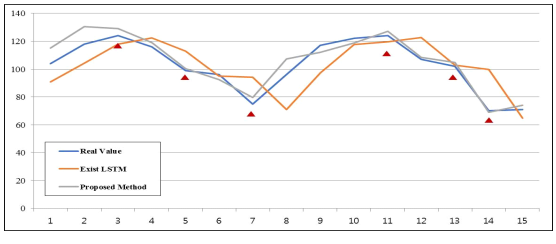

Fig. 9 Comparison of predicted values for the first 15 data occurrences

사례에서 기존 LSTM은 실제 값에서 데이터 변화가 발생하고 이를 look_back을 통해 입력된 이후 예측 값이 생성되는데 반해, 제안 방법은 변화 시점을 예측하여 예측 값을 생성한다.

Ⅴ. 결 론

본 논문에서는 환경 데이터를 활용하여 미세먼지 오염도에 대한 시계열 예측 성능을 개선하기 위해 CNN-LSTM 결합 방법을 제안하였다.

제안 방법의 성능을 검증하기 위해 Beijing $PM_{2.5}$ 데이터를 활용하여 5개 시계열 구간에 대한 RMSE를 비교한 결과 모든 시험 구간에서 제안 방법의 예측 성능이 우수함을 알 수 있다. 또한 극단적인 데이터가 발생하는 시점에 대한 예측 성능을 F1 Score로 살펴본 결과 역시 제안방법이 우수하게 나타나 미세먼지 오염도 예측에서 제안 방법이 효율적인 것을 검증하였다.

본 논문에서는 여러 제약 조건으로 인해 제안 방법과 비교 대상을 결합 모형인 APNet이나 ConvLSTM과 성능 비교하지 못하고 기존 LSTM 단독 모델과 비교함으로써 타 결합 방법과 성능 비교를 수행하지 못한 제약을 가지고 있다. 이는 향후에 추가적으로 연구가 되어야 할 분야이다.

본 연구를 통해 확보된 제안 방법과 실험 결과를 활용할 경우 미세먼지 오염도 데이터뿐만 아니라 급격히 변화하는 시계열 데이터가 많이 발생하는 환경, 사회과학 등 다양한 영역에서 응용될 수 있다.

REFERENCES

[1] D. Y. Jin, K. J. Han, and J. H. Kim, “Estimation of Fine Dust Pollution Using The Atmospheric Images,” *Korea Environment Institute Working Paper*, 2018.

[2] C. J. Huang, and P. H. Kuo, “A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cites,” *Sensors* 18, Jul. 2018.

[3] Y. Qi, Q. Li, H. Karimian, and D. Liu, “A hybrid model for spatiotemporal forecasting of ($PM_{2.5}$) based on graph convolutional neural network and long short-term memory,” *Sci. Total Environ.*, vol. 664, pp. 1-10, May. 2019.

[4] S. G. Lee, and J. T. Shin, “Hybrid Model of Convolutional LSTM and CNN to Predict Particulate Matter,” *IJIEE*., vol. 9, no. 1, pp. 34-38, Mar. 2019.
[5] C. Liu, F. Tsow, Y. Zou, and N Tao, “Particle Pollution Estimation Based on Image Analysis,” *PLoS ONE* 11(2) :doi: 10.1371/journal.pone.0145955.
[6] C. H. Hwang, H. S. Kim, and H. K. Jung, “Detection and Correction Method of Erroneous Data Using Quantile Pattern and LSTM,” *JICCE*., vol. 16, no. 4, pp. 242-247, Dec. 2018.
[7] V. Q. Nguyen, L. V. Ma, and J. Kim, “LSTM-based anomaly detection on big data for smart factory monitoring,” *Journal of Digital Contents Society*, vol. 19, no. 4, pp. 789-799, 2018. DOI:10.9728/dcs.2018.19.4.789.

황철현(Chul-Hyun Hwang)

1991년 금오공과대학교 전자공학과(공학사)
1995년 경남대학교 컴퓨터공학과(공학석사)
2015년 배재대학교 컴퓨터공학과(공학박사)
2019~현재 경복대학교 스마트IT학과 부교수
※관심분야 : 빅데이터, 인공지능, 기계학습, 딥러닝, IoT, Data Architecture

신강욱(Gang-Wook Shin)

1987년 동국대학교 전자공학과 학사
1993년 홍익대학교 전자공학과 석사
2005년 홍익대학교 전기공학과 박사
1993년~현재 : 한국수자원공사 수석연구원
※관심분야 : 플랜트제어 및 응용, 모델링, 지능제어, 원격감시제어, 센서응용