Journal of Korean Society for Atmospheric Environment
Vol. 35, No. 2, April 2019, pp. 214-225
https://doi.org/10.5572/KOSAE.2019.35.2.214
p-ISSN 1598-7132, e-ISSN 2383-5346

논 문

심층 신경망을 이용한 대기질 예측
Air Quality Prediction Using a Deep Neural Network Model

조경학, 이병영, 권명흠, 김석철*
㈜볼트시뮬레이션

Kyunghak Cho, Byoung-young Lee, Myeongheum Kwon, Seogcheol Kim*
BOOLT Simulation, Inc

접수일 2019년 2월 10일
수정일 2019년 3월 18일
채택일 2019년 3월 22일

Received 10 February 2019
Revised 18 March 2019
Accepted 22 March 2019

***Corresponding author**
Tel : +82-(0)2-3477-1963
E-mail : sckim@boolt.co.kr

Abstract A deep neural network (DNN) model of multi-layer perceptron with 3 or 4 hidden layers is developed to predict the air qualities. The DNN model takes the past 3 days of the hourly concentration measurements of the pollutants (CO, SO_2, NO_2, O_3, PM_{10}, $PM_{2.5}$) and the meteorology data (wind speed, wind direction, air temperature, air humidity), and then predicts the hourly concentration of the pollutants for the next 24 hours. The DNN model was compared against the observations from all nationwide air quality monitoring stations which includes 115 sites in 7 metropolitan cities in South Korea. The index of agreement (IOA) was found to be 0.7∼0.8, based upon the 6,505 comparison data sets from January 1, 2017 to September 30, 2017. In the unit of air quality grade, which can be evaluated from the pollutant concentration level, 60%∼80% cases of the DNN predictions agree with those of the observations. For the region-wide PM_{10} grade, the DNN predicts exactly the 75%∼85% cases of the observations, which is in about the same accuracy range of the numerical air quality models of the current operative use. Yet, for the region-wide $PM_{2.5}$ grade, the cases of the accurate predictions of DNN is about twice of those of the numerical model. In the metropolitan Gwangju, for an example, the DNN predicts exactly the 211 next days of the $PM_{2.5}$ grade, while the numerical model forecasts just 120 days correctly.

Key words: Deep neural network model, Machine learning, TensorFlow, Air quality model, Numerical model, National air quality monitoring station

1. 서 론

대기질 예측을 위한 통계적 모델의 연구는 꾸준히 진행되어 왔다 (Goyal *et al.*, 2006; Hooyberghs *et al.*, 2005). 수치모델에 비해서 연산속도가 매우 빠르다는 점이 통계적 모델의 장점인데, 적절히 구현될 경우 통계적 모델의 예측 정확도 또한 수치모델보다 높은 것으로 보고되어 있다 (Hrust *et al.*, 2009).

통계적 모델에 대한 연구사례에는 얕은 신경망 모형 (Shallow Neural Network, 이하 SNN), 입출력 층과 1개의 은닉층 (hidden layer)으로 구성된 인공 신경망 모형 (Artificial Neural Network, 이하 ANN), Support Vector Machine (SVM), 다항 로지스틱 회귀모형 (Multiple Logistic Regression), Random Forest (RF) 기법 등이 포함되어 있다.

심층 신경망 (Deep Neural Network, 이하 DNN) 모형은 ANN의 일종인데 여러 개의 은닉층을 지닌다는 점이 특징이다 (Schmidhuber, 2015; Bengio *et al.*, 2013). 은닉층이 많을수록 학습 시간과 연산량이 증가하고 과적합 문제가 발생한다. 여기에 대한 하나의 해결방안이 Hinton (2007)에 의해서 제시된 이후 여러 분야에서 DNN 사용이 급격히 확대되고 있다. 대

기질 예측에도 DNN을 활용한 사례가 증가하고 있다 (Shahraiyni and Sodoudi, 2016).

Perez and Reyes (2002)는 칠레의 산티아고에서 1998년부터 2000년까지 8개 측정소의 미세먼지 (PM_{10}) 관측농도와 기상정보 (온도, 상대 습도 및 풍속)를 토대로 SNN을 구축하였는데 이를 통해 대기질 예측 정확도가 개선되는 것을 확인하였다.

McKendry (2002)는 캐나다 Chilliwack 지역에서 시간당 대기오염물질 관측 농도 (NO, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$)와 관측 기상 (기온, 풍속, 풍향)을 활용하여 오존 (O_3)과 미세먼지 (PM_{10}), 그리고 초미세먼지 ($PM_{2.5}$)의 일최고 및 일평균 농도를 예측하는 통계적 기법을 연구하였다. 심층 신경망 모형 (Multi-Layer Perceptron, MLP)과 다중 선형 회귀 모형 (Multiple Linear Regression, MLR)을 비교한 결과, 오존에 대해서 MLP가 유의미하게 더 높은 정확도를 가지는 반면 미세먼지 (PM_{10})와 초미세먼지 ($PM_{2.5}$)에 대해서는 두 모형의 정확도가 비슷한 것으로 나타났다.

Jeon and Son (2018)은 2010년부터 2015년까지 국내 6개 대도시의 일별 미세먼지 관측데이터를 토대로 여러 가지 통계모델을 실험하였는데, 심층 신경망 모형에 의한 등급 예측이 다른 기법 (SNN, 다항 로지스틱 회귀모형, SVM, RF)보다 더 정확한 것으로 확인되었다.

Shahraiyni and Sodoudi (2016)는 도시지역 미세먼지 예측을 위해서 통계모델이 사용된 기존 연구들을 광범위하게 비교하였다. 그 결과 MLR에 비해 ANN이 우수한 것으로 나타났다. ANN 구조 중에서 MLP가 가장 빈번하게 사용된 것으로 조사되었으나 기존 연구결과에 대한 비교를 통해서 대기질 예측을 위한 최적의 ANN 구조는 판정할 수 없었다.

본 연구에서는 대전광역시의 측정망 자료를 토대로 DNN 모델을 구축하였다. 구축된 DNN 모델을 전국 7개 광역시 (서울, 부산, 대구, 인천, 대전, 광주, 울산)의 115개 측정소에 확대 적용하여 시간대별로 대기질을 예측한 후 관측값과 비교하였다. 또한 대기질 예측 농도를 등급으로 변환하여 얼마나 관측등급과 일치하는지를 비교하였다.

2. DNN 모델구현

본 연구에서는 관측결과를 토대로 관측지점의 대기오염농도 변화를 예측하도록 DNN 모델을 구성하였다. Python 기반의 Keras (2.1.1) 프레임워크를 사용하여 DNN 모델코드를 구현하였고, 인공지능 엔진 (AI engine)은 Google의 오픈소스 라이브러리 TensorFlow (1.4)를 적용하였다.

DNN 모델의 입력항목으로는 직전 3일간의 시간대별 연속 관측자료, 곧 72시간 동안의 데이터를 적용하였다. 입력 관측자료는 대기질 관측자료 (CO, SO_2, NO_2, O_3, PM_{10}, $PM_{2.5}$)와 기상 관측자료 (풍속, 풍향, 기온, 습도)로 구성된다. DNN 모델의 출력은 시간별 대기오염농도 예측치인데, 본 연구에서는 1시간 후부터 24시간 이후까지 각 시간대별 농도를 산출하도록 구성하였다.

2.1 사용자료 개요

본 연구에서 DNN 모델의 학습 및 검증에 사용된 자료는 10개의 항목 (feature)으로 구성된다. 대기오염물질 농도정보 6개 (CO, SO_2, NO_2, O_3, PM_{10}, $PM_{2.5}$)와 기상정보 4개 (기온, 풍향, 풍속, 습도)이다.

이중 대기오염물질 농도정보는 한국환경공단의 에어코리아 (www.airkorea.or.kr)에서 제공하는 국가측정망 (도시대기측정망, 도로변대기측정망) 농도자료를 이용하였다. 기상정보는 기상자료개방포털 (data.kma.go.kr)에서 제공하는 종관기상관측 (ASOS) 자료를 이용하였다.

대기오염물질 측정소별 관측정보에는 대기오염물질 농도만 제공되고 기상정보는 없다. 그래서 본 연구에서는 해당 광역시의 종관기상관측지점의 동일시간대 기상관측자료를 각 측정소별 대기오염물질 측정자료와 함께 이용하였다. 기상관측 항목 중 풍향과 풍속은 동서 및 남북 방향의 속도 벡터 성분으로 변

환하여 사용하였다.

농도나 기상 입력자료의 결측치는 k-nearest neighbor (kNN) 알고리즘을 적용하여 추정하였다. 본 연구에서는 R studio에서 제공하는 DMwR 패키지의 knnImputation () 함수를 이용하였다 (Torgo, 2010): k = 10, Scale = TRUE, Meth = weightAvg.

2.2 DNN 모델구성

DNN 모델의 최적 구조, 곧 은닉층의 개수와 각 층별 노드 분포를 결정하기 위해서 본 연구에서는 대전광역시의 관측자료를 활용하였다. 대기질 입력자료는 대전시 읍내동 관측소 (대기질 관측소번호: 525111) 데이터를 사용하였다 (NIE, 2017). 기상입력자료는 대전광역시 (기상 관측소번호: 133)의 데이터를 사용하였다 (KMA, 2017). 2015년 01월 03일부터 2017년 03월 28일까지 약 3년의 연속 시간별 데이터 (19,584개)를 사용하였다. 그 중에서 검증을 위해서 2017년 03월 29일부터 2017년 06월 29일까지 약 3개월 분량의 시간별 연속 데이터 (2,136개)를 사용하였다. 학습을 위한 데이터는 충분히 길어야 하는데, 최적의 학습데이터 기간은 DNN 모델의 계층구조와 문제의 성격에 따라 달라진다. 학습데이터를 늘려가면서 DNN 모델의 예측정확도를 확인해본 결과 학습기간이 3년 이상일 경우에 예측정확도는 더 이상 향상되지 않았다. 이때 에포크 (epoch)는 50회, 배치 (batch) 크기는 512로 설정하였다.

각 계층의 초기 가중치는 절삭된 정규분포 (TruncatedNormal: 표준편차의 2배보다 큰 값은 제외)에 의거하여 할당하였다. DNN 학습시 손실함수 (Loss function)를 최소화하는 방법으로는 Adam (Kingma, and Ba, 2015)을 적용하였다. 손실함수는 Mean Absolute Error (MAE)로써 정의하였다.

최적의 DNN 모델을 결정하기 위하여 모델의 구조 그리고 계산함수와 데이터 처리절차 등을 변경해가면서 학습과 검증을 반복하였다. 이 과정은 시행착오의 과정이다. 최종적으로 결정된 DNN 모델구조는 표 1과 같다. 표 1에서 은닉노드 개수는 은닉노드 수를 변경하면서 찾아낸 최적 결과이다. 이산화질소 (NO_2), 입자상 물질 (PM_{10}, $PM_{2.5}$), 일산화탄소 (CO)의 경우 각 계층 당 100개의 은닉 노드를 할당하였다. 오존 (O_3)의 경우 입력계층에서 가까운 순서대로 100개, 400개, 200개의 노드를 은닉계층별로 구성하였다. 이산화황 (SO_2)의 경우 각각 100개, 200개, 100개의 노드를 할당하였다.

최초 시도에서는 모든 오염물질에 대해서 신경망의 은닉층은 3개로 동일하게 설정하여 실험하였다. 그런데 이산화질소 (NO_2)와 오존 (O_3)의 경우 시간의 존성을 고려하기 위하여 시계열에 대한 1차원 convolution layer를 추가한 결과 예측정확도가 향상되는 것을 발견하였다. 이에 이 두 가지 물질에 대해서는 4개의 은닉층으로 신경망을 구성하였다.

표 1에서 dense layer는 입력과 출력이 완전히 연결되어있는 층을 의미하며, units은 층에 포함된 노드의 개수를 의미한다. Convolution layer는 인근의 노드끼리 묶어서 계산하는 convolution 계산방식을 지원하는 층인데, filters는 convolution layer의 커널 (kernel)

Table 1. Configuration of the DNN model.

Pollutants	Layer configurations
NO_2	#1 convolution layer – filters:48, dropout: 0.2 #2 dense layer – units:100, dropout: 0.2 #3 dense layer – units:100, dropout: 0.2 #4 dense layer – units:100, dropout: 0.2
O_3	#1 convolution layer – filters:48, dropout: 0.2 #2 dense layer – units:100, dropout: 0.5 #3 dense layer – units:400, dropout: 0.5 #4 dense layer – units:200, dropout: 0.5
PM_{10}	#1 dense layer – units:100, dropout: 0.2 #2 dense layer – units:100, dropout: 0.2 #3 dense layer – units:100, dropout: 0.2
$PM_{2.5}$	#1 dense layer – units:100, dropout: 0.2 #2 dense layer – units:100, dropout: 0.2 #3 dense layer – units:100, dropout: 0.2
SO_2	#1 dense layer – units:100, dropout: 0.0 #2 dense layer – units:200, dropout: 0.0 #3 dense layer – units:100, dropout: 0.0
CO	#1 dense layer – units:100, dropout: 0.2 #2 dense layer – units:100, dropout: 0.2 #3 dense layer – units:100, dropout: 0.2

의 수를 의미한다. Dropout은 각 은닉계층의 노드로부터 일괄적으로 삭제하는 노드의 비율이다 (Dahl *et al.*, 2013). 학습과 검증과정을 반복하면서 dropout의 최적치를 결정하였다. 대부분의 경우 dropout은 0.2가 적합하였는데 일부 경우 0.0 혹은 0.5가 적합하였다.

입력신호를 출력신호로 변환하는 활성화함수는 이산화질소 (NO_2)와 미세먼지 (PM_{10}, $PM_{2.5}$)의 경우에는 ReLU (Rectified Linear Unit)를 적용하였다. 나머지 물질에 대해서는 LeakyReLU (Leaky ReLU)가 적용하였다. 각 함수의 정의는 다음과 같다.

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

$$\mathrm{LeakyReLU}(x) = \max(0.1x, x) \tag{2}$$

DNN 모델의 예측결과를 관측자료와 비교하기 위하여 IOA (index of agreement, Willmott *et al.*, 1985), ME (bias), NRMSE를 산출하여 표 2에 제시하였다. 각 통계량은 다음 공식에 의해서 산출하였다.

$$\mathrm{IOA} = 1 - \frac{\sum_{i=1}^{n}(O_i - M_i)^2}{\sum_{i=1}^{n}(|M_i - \overline{M}| + |O_i - \overline{O}|)^2} \tag{3}$$

$$\mathrm{ME(bias)} = \frac{1}{n}\sum_{i=1}^{n}(O_i - M_i) \tag{4}$$

$$\mathrm{NRMSE} = \frac{100}{\overline{O}} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)^2} \tag{5}$$

여기서 O_i는 관측 값 (1시간 평균농도)이며 M_i은 DNN 모델 값 (1시간 평균농도에 대한 24시간 예측)이다. $\overline{O}$와 $\overline{M}$는 각각 관측과 모델 평균이다. n은 데이터 개수로 2,136개이다.

IOA는 두 결과의 시계열의 유사성을 나타내는 척도이다. IOA는 오염물질에 따라 0.5~0.8로 나타났다. IOA가 가장 높은 경우는 오존으로 약 0.8이다. IOA가 가장 낮은 경우는 이산화황으로 약 0.5이다. 입자상 물질의 경우 PM_{10} (0.5)보다 $PM_{2.5}$ (0.7)에 대해서 IOA가 더 높다.

ME (bias)는 관측과 모델간의 평균 편향을 나타내는 지표로, 단위는 각 오염물질의 농도와 동일하다. DNN 모델의 ME (bias)는 매우 낮다. 즉 평균량에 대한 예측 정확도가 매우 높은 편인데 이는 학습에 의존하는 DNN 모델의 특성으로 생각된다. 다만 입자상 물질인 PM_{10}에 대해서는 ME (bias)가 11 μg/m^3 정도인데, 이는 관측평균의 20%에 해당하는 상대적으로 높은 수치이다.

NRMSE (Normalized RMSE)는 RMSE를 관측평균으로 나누어 백분율로 나타낸 것이다. 일산화탄소 (CO)의 경우 34%로 가장 낮고, 이산화황 (SO_2)의 경우 71%로 가장 높다.

모델과 관측의 산점도 (scatter plot)는 그림 1과 같다. 수평축은 관측결과이며 수직축은 DNN 모델예측이다. 입자상 물질 (PM_{10}, $PM_{2.5}$)의 농도 단위는 μg/m^{3}이며 기체상 물질의 농도 단위는 ppb이다. 산점도 데이터의 개수와 모델 및 관측의 평균은 산점도의 좌측 상단에 표시하였다. 모델과 관측간의 일치도를 나타내는 통계량도 같이 표시하였다.

그림 2는 9개월 (2017.1.4.~2017.9.29.)의 DNN 모델 예측과 관측을 비교한 시계열 그래프이다. 회색으로 채워진 그래프는 1시간 평균 관측농도이며 흑색 굵은 실선은 DNN 모델 예측이다. 그래프의 수평축은 1시간 간격의 관측순서를 표시한 것이다. 수평축의 단위는 하루이다. 수직축은 ppb (기체상 물질) 혹은 μg/m^3 (입자상 물질) 단위의 농도이다. 시계열 그래프에서 주기성이 상대적으로 두드러지게 나타나는 오존의 경우, 며칠 단위의 농도 변화뿐 아니라 하루 혹은 이내의 급격한 농도 변화 또한 DNN 모델에 의

Table 2. Statistical Indices of the hourly DNN predictions and the national air quality monitoring station, #525111 (Daejeon), during 2017.03.29.~2017.06.29.

Pollutants	IOA	ME	NRMSE
NO_2	0.654	−0.103	53.612
O_3	0.793	3.925	42.249
PM_{10}	0.549	10.944	62.787
$PM_{2.5}$	0.660	−0.426	43.693
SO_2	0.547	0.115	71.199
CO	0.625	2.457	33.829

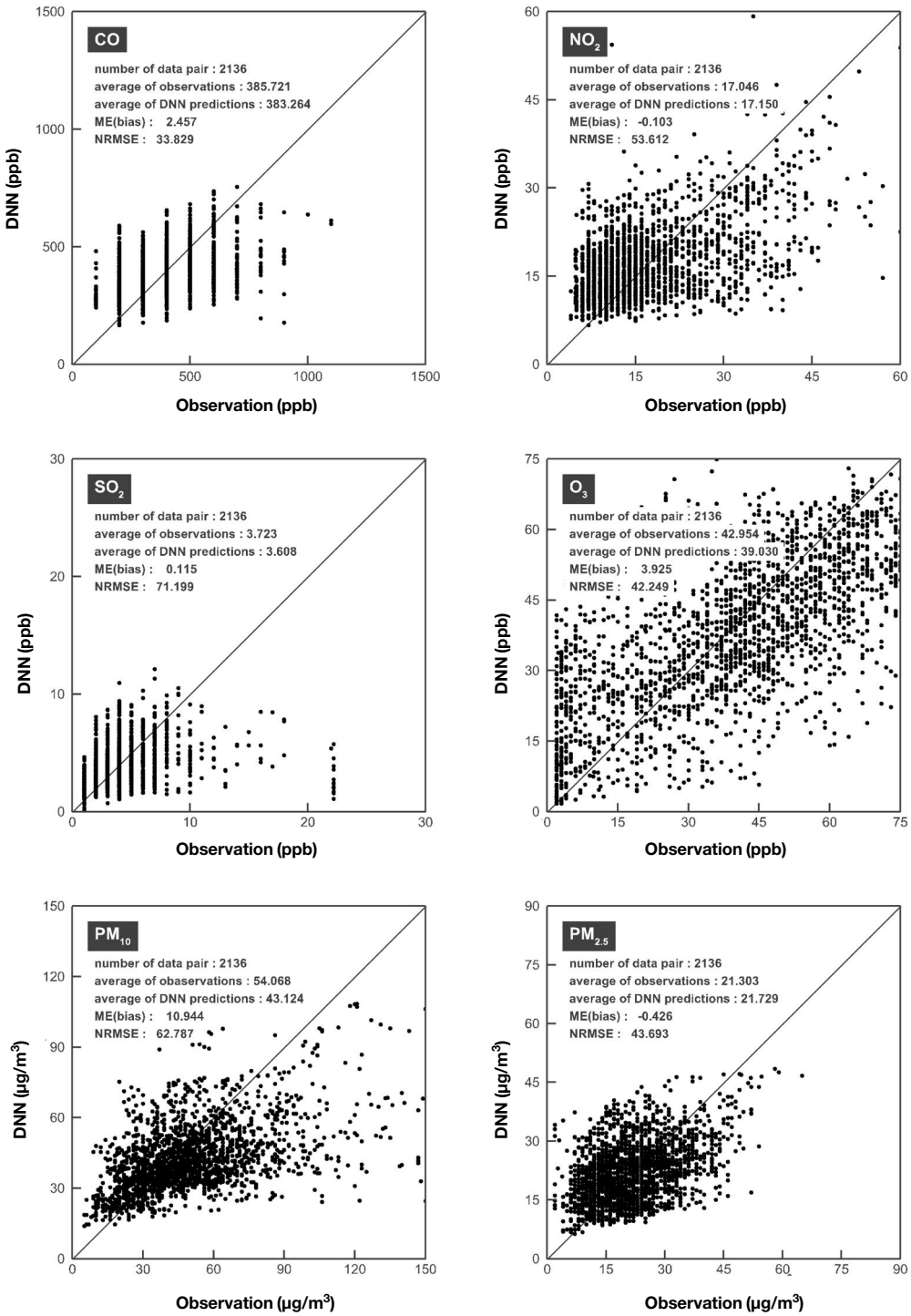

Fig. 1. Scatter plots of the DNN predictions and the hourly averaged measurements at station #525111 during 2017.03.29.~ 2017.06.29.

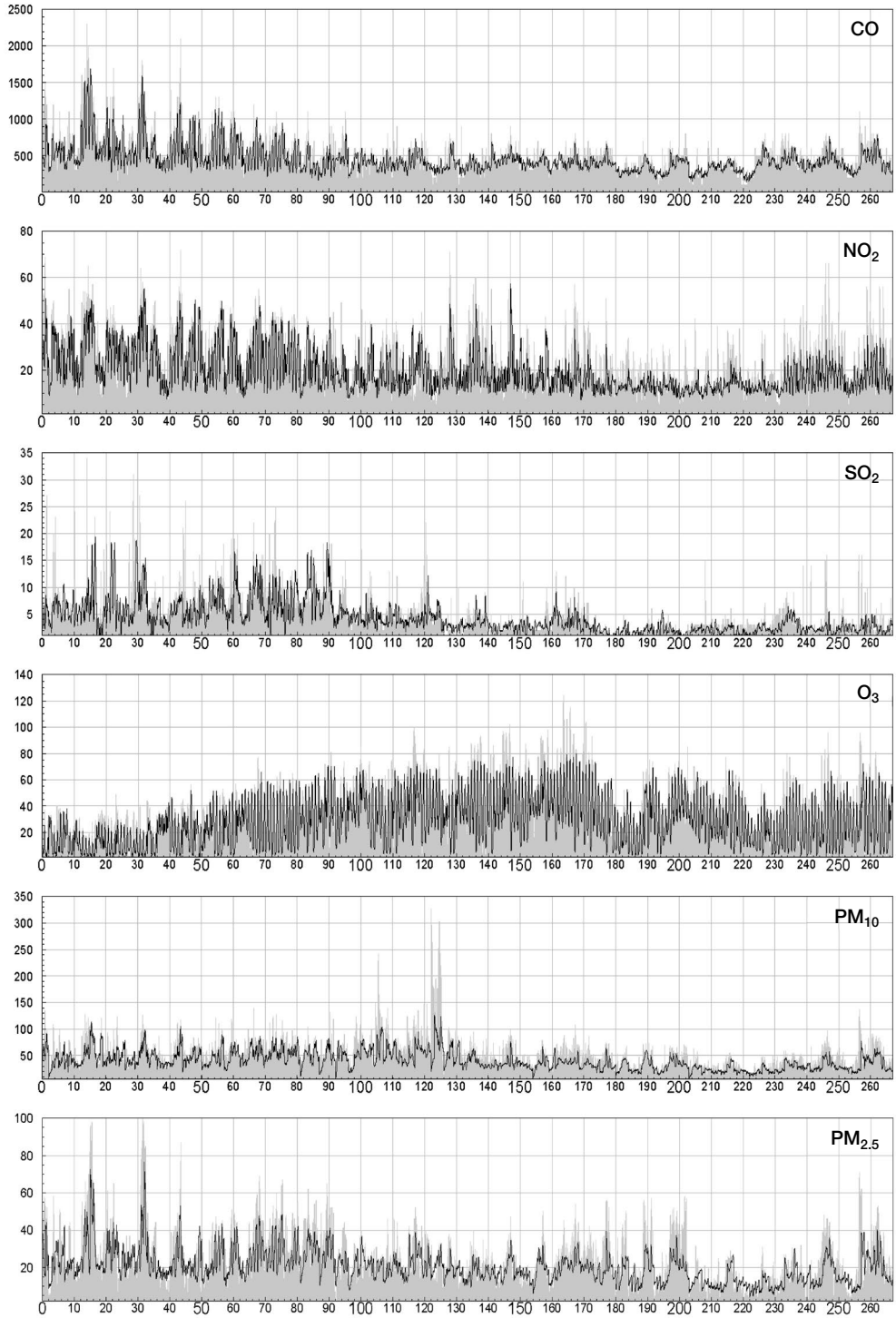

Fig. 2. Time series plot of the DNN model predictions (black solid lines) against the hourly averaged measurements (gray fill) during 2017.01.04.～2017.09.29.

해서 비교적 잘 예측된다. 오존 농도의 일 최고치는 비교 기간 내에서 상대적으로 균일하다. 오존 농도의 일 최저치 또한 비교기간 동안 큰 변동이 없었다.

반면에 하루 단위로 농도변화가 급격하게 나타났던 다른 오염물질에 대해서는 DNN 모델 예측결과가 상대적으로 더 부정확하다. 급격히 상승하는 경우에 DNN 모델은 실제보다 과소 예측하는 경향이 있다. 이는 PM_{10}에 대한 비교 결과에서 잘 관찰된다. 익일의 PM_{10} 농도변화가 급격히 상승하는 경우는 DNN 모델에 의해서는 잘 예측되지 않는다. 반면 $PM_{2.5}$에 대해서는 경향이 매우 다르다. 실제로 발생한 급격한 $PM_{2.5}$ 농도변화가 DNN 모델예측에 의해서도 비교적 잘 포착된다.

각 오염물질에 대해서 표 2의 IOA는 그림 2의 시계열 그래프 비교특성과 일치한다. IOA는 또한 그림 1의 산점도 비교특성 및 관련 통계량과도 대체로 일치한다. 본 연구에서 IOA는 DNN 모델과 관측간의 일치도를 측정하는 통계량으로 적합한 것으로 판단되므로 이하 전국 데이터에 대한 비교에서는 IOA를 기준으로 삼았다.

3. DNN 모델적용

3.1 전국 광역시 측정소 예측결과

대전광역시 데이터를 토대로 구조를 확정한 DNN 모델을 전국에 대해서 확대 적용하였다. 7개 광역시(서울, 부산, 대구, 인천, 대전, 광주, 울산)의 115개 대기오염물질 측정소에 대하여 DNN 예측 결과와 관측자료를 비교하였다. 전체 측정소의 구성을 살펴보면 도시대기측정망이 93곳으로 대부분을 차지하며 나머지는 도로변대기측정망으로 22곳이다. 광역시별 측정소 분포 현황을 살펴보면 서울시가 34곳으로 측정소가 가장 많다. 측정소 현황은 표 3에 나타냈다.

2014년 01월 01일부터 2016년 12월 31일까지 전국 측정소의 시간별 데이터를 DNN 모델 학습 자료로 적용하였다. 학습용 데이터 수는 26,305개이며, 학습과정은 전술한 모델구축용 과정과 동일하다. 2017년 01월 04일부터 2017년 09월 29일까지 전국 측정소의 시간별 데이터를 통해 각 측정소별 신경망이 예측한 값과 실제 관측값을 비교하였다. 서울과 울산의 일부 측정소에서는 $PM_{2.5}$를 측정하지 않는다. 부산의 일부 측정소에서는 일산화탄소(CO)와 이산화황(SO_2)을 측정하지 않는다. 측정되지 않는 대기오염물질은 해당 지점의 예측항목에서 제외하였다.

Table 3. The national air quality monitoring stations in Korea.

City	Number of stations		
	Total	Urban	Roadside
Seoul	34	23	11
Busan	20	18	2
Daegu	10	9	1
Incheon	17	14	3
Gwangju	9	7	2
Daejeon	10	8	2
Ulsan	15	14	1

표 4에 광역시별 도시대기측정망과 도로변대기측정망의 평균 IOA, 그리고 광역시 전체의 평균 IOA를 나타냈다. 또한 광역시 전체의 도시대기측정망 평균과 도로변대기측정망 평균, 그리고 115개 측정소 전체에 대한 평균 IOA도 함께 나타냈다. IOA 값을 기준으로 색을 달리하여 청색(0.0~0.6), 녹색(0.6~0.7), 황색(0.7~0.8), 적색(0.8~1.0)으로 표 바탕을 표시하였다.

표 4에서 STEP01과 STEP24는 DNN 모델의 예측작업 반복 주기이다. STEP01은 DNN 예측작업을 1시간 주기로 반복한 경우로, 매 시간마다 예측시점으로부터 24시간 이후의 값을 예측한 결과이다. STEP24는 예측작업의 반복 주기가 24시간인 경우로, 24시간마다 예측시점으로부터 24시간 이후의 값을 예측한 결과이다. 표 4에서 예측 반복 주기에 따른 IOA의 차이는 경미하다. 이하 비교분석에서는 24시간 단위로 예측한 결과(STEP24)를 사용하였다.

대부분의 관측소에서 이산화황(SO_2)의 예측 정확

Table 4. IOA averaged over each metropolitan area and all national air quality monitoring stations during 2017.01.04.~2017.09.29. (blue for IOA value of 0.0~0.6, green 0.6~0.7, yellow 0.7~0.8, red 0.8~1.0).

		STEP01						STEP24					
		CO	NO_2	SO_2	O_3	PM_{10}	$PM_{2.5}$	CO	NO_2	SO_2	O_3	PM_{10}	$PM_{2.5}$
Seoul	Urban	0.756	0.747	0.613	0.839	0.704	0.738	0.767	0.763	0.608	0.842	0.699	0.743
	Roadside	0.737	0.741	0.691	0.738	0.705	–	0.745	0.746	0.685	0.741	0.729	–
	All	0.750	0.745	0.638	0.807	0.705	0.738	0.760	0.758	0.633	0.809	0.709	0.743
Busan	Urban	0.745	0.697	0.507	0.761	0.699	0.698	0.755	0.708	0.509	0.768	0.681	0.707
	Roadside	0.712	0.794	0.559	0.695	0.699	0.700	0.703	0.791	0.532	0.726	0.668	0.696
	All	0.742	0.706	0.514	0.754	0.699	0.698	0.749	0.716	0.511	0.764	0.680	0.706
Daegu	Urban	0.762	0.772	0.655	0.859	0.694	0.729	0.783	0.783	0.659	0.859	0.702	0.752
	Roadside	0.843	0.764	0.719	0.850	0.788	0.805	0.859	0.781	0.721	0.854	0.809	0.828
	All	0.770	0.771	0.661	0.858	0.703	0.737	0.791	0.783	0.665	0.859	0.713	0.760
Incheon	Urban	0.751	0.704	0.541	0.798	0.653	0.684	0.763	0.715	0.532	0.804	0.641	0.683
	Roadside	0.720	0.750	0.559	0.765	0.675	0.653	0.731	0.757	0.549	0.777	0.645	0.656
	All	0.745	0.713	0.544	0.792	0.657	0.679	0.757	0.722	0.535	0.799	0.642	0.678
Gwangju	Urban	0.761	0.754	0.642	0.838	0.707	0.694	0.773	0.760	0.635	0.847	0.681	0.719
	Roadside	0.681	0.762	0.710	0.820	0.738	0.704	0.688	0.782	0.713	0.830	0.740	0.740
	All	0.743	0.756	0.657	0.834	0.714	0.696	0.754	0.765	0.652	0.844	0.695	0.724
Daejeon	Urban	0.791	0.785	0.731	0.859	0.718	0.743	0.818	0.801	0.736	0.862	0.693	0.760
	Roadside	0.771	0.817	0.731	0.840	0.733	0.726	0.787	0.812	0.721	0.844	0.704	0.736
	All	0.787	0.792	0.731	0.856	0.721	0.740	0.812	0.803	0.733	0.859	0.695	0.755
Ulsan	Urban	0.648	0.700	0.461	0.776	0.693	0.685	0.667	0.706	0.430	0.781	0.680	0.708
	Roadside	0.631	0.746	0.340	0.768	0.695	–	0.621	0.739	0.329	0.756	0.666	–
	All	0.647	0.703	0.453	0.775	0.693	0.685	0.664	0.708	0.424	0.780	0.679	0.708
Averaged urban		0.741	0.730	0.578	0.812	0.694	0.713	0.755	0.742	0.570	0.817	0.682	0.724
Averaged roadside		0.730	0.757	0.651	0.761	0.709	0.702	0.737	0.762	0.644	0.768	0.711	0.714
Averaged all		0.739	0.735	0.592	0.802	0.697	0.712	0.752	0.745	0.585	0.807	0.688	0.723

도가 가장 낮고 오존(O_3)의 예측 정확도가 가장 높다. IOA를 살펴보면 일산화탄소(CO)는 0.621~0.859, 이산화질소(NO_2)는 0.697~0.817, 이산화황(SO_2)은 0.329~0.736, 오존(O_3)은 0.695~0.862, 미세먼지(PM_{10})는 0.641~0.809, 초미세먼지($PM_{2.5}$)는 0.653~0.828로 나타났다. IOA 결과에서 측정소별 편차는 크지 않다. 대전의 한 개 측정소 자료를 토대로 구성된 DNN 모델 구조를 전국 측정소에 대해서 동일하게 적용하여도 무난한 것으로 생각된다. 빈 칸은 관측 자료가 없는 경우이다.

표 2에서 DNN 모델을 구축한 뒤 확인한 검증결과와 비슷하게 각 광역시마다 오존(O_3)에 대한 IOA가 가장 높으며, 이산화황(SO_2)에 대한 IOA가 가장 낮음을 확인할 수 있다.

4. DNN 모델의 성능평가

4.1 대기질 등급예측

한국환경공단에서는 대기질 등급 기준을 제시하고 있는데 표 5와 같이 '좋음(good)', '보통(moderate)', '나쁨(bad)', '매우나쁨(very bad)'의 4가지 등급으로 대기질을 구분한다. 대기질 등급을 구분하는 기준은 대기오염 농도로, 가스상 물질의 경우에는 1시간 평균 농도이며 입자상 물질은 24시간 이동평균 농도이다.

DNN 모델의 예측 농도를 대기질 등급으로 변환하여 관측등급과 비교하였다. DNN 모델의 예측등급이 관측등급과 일치하는 비율(%)을 표 6에 나타내었다. 모든 측정소와 모든 시간에 대해서 비교한 결과이다. 일산화탄소(CO)와 이산화황(SO_2)의 경우 일치도가

Table 5. Air quality grade (www.airkorea.or.kr/khaiInfo).

Pollutant	Pollution level grade			
	Good	Moderate	Bad	Very bad
NO_2 (ppm)	0∼0.030	0.031∼0.060	0.061∼0.200	0.201∼2
O_3 (ppm)	0∼0.030	0.031∼0.090	0.091∼0.150	0.151∼0.600
PM_{10} (μg/m^3)	0∼30	31∼80	81∼150	151∼600
$PM_{2.5}$ (μg/m^3)	0∼15	16∼35	36∼75	76∼500
SO_2 (ppm)	0∼0.020	0.021∼0.050	0.051∼0.150	0.151∼1
CO (ppm)	0∼2	2.01∼9	9.01∼15	15.01∼50

Table 6. Percentage of agreement between observed air quality grades and DNN predictions, summed over all national air quality monitoring stations during 2017.01.04.∼2017.09.29.

	Percentage of grade agreement					
	CO	NO_2	SO_2	O_3	PM_{10}	$PM_{2.5}$
Seoul	99.96	67.24	99.98	80.75	69.97	72.86
Busan	100.00	82.90	99.26	75.75	69.03	59.22
Daegu	99.91	84.11	99.82	78.40	68.89	60.20
Incheon	99.82	76.09	99.23	77.72	69.47	59.09
Gwangju	99.99	85.66	99.93	78.49	67.03	61.82
Daejeon	99.99	88.57	99.89	78.59	72.20	63.40
Ulsan	99.40	79.60	92.64	76.23	67.51	82.82

90% 이상으로 다른 대기오염물질에 비해 높다. 오존(O_3)의 경우 등급일치 비율이 75%∼80% 가량이며, 이산화질소(NO_2)는 서울 지역을 제외하고는 75%∼85% 가량이다. 입자상 물질(PM_{10}, $PM_{2.5}$)은 일치도가 가장 낮은데 예측등급의 일치비율은 60%∼70% 가량이다.

표 7은 관측과 예측값의 농도 등급을 positive (표 5에서의 대기오염물질 농도 등급에서 '좋음' 또는 '보통'에 해당하는 경우)와 negative (표 5에서의 대기오염물질 농도 등급에서 '나쁨' 또는 '매우 나쁨'에 해당하는 경우)로 구분하여 일치하는 정도를 나타낸 결과이다. 표 7에서 p는 관측과 예측이 전부 positive인 경우의 비율이며, n은 관측과 예측이 전부 negative인 경우의 비율이다.

대기질 농도등급에 대한 결과를 비교해 보면, 대기질 등급이 '좋음', '보통'에 해당하는 경우의 일치율이 '나쁨', '매우나쁨구간'에 해당하는 경우에 비해 상대적으로 높다. 이는 현재 구축된 DNN 모델이 가지는

Table 7. Percentage of agreement between observed air quality grades and DNN predictions, summed over all national air quality monitoring stations, just for bad or very bad grade of observation period only during 2017.01.04.∼2017.09.29.

		Percentage of agreement (bad/very bad grade events percentage)					
		CO	NO_2	SO_2	O_3	PM_{10}	$PM_{2.5}$
Seoul	p	100.0	96.75	99.99	100.0	97.63	95.82
	n	–	38.51	–	0.51	25.81	40.03
Busan	p	100.0	99.69	99.98	100.0	98.10	90.15
	n	–	13.36	0.00	0.00	19.15	40.36
Daegu	p	99.91	99.73	99.88	99.72	97.77	93.61
	n	–	9.52	0.00	0.16	30.40	45.26
Incheon	p	100.0	99.38	100.0	100.0	98.46	93.60
	n	–	16.17	0.00	0.00	19.90	35.88
Gwangju	p	100.0	99.91	99.93	99.99	98.26	95.62
	n	–	6.00	–	0.00	27.84	38.48
Daejeon	p	100.0	99.99	99.92	99.93	98.19	96.66
	n	–	2.81	–	1.61	24.40	34.58
Ulsan	p	100.0	99.93	99.80	100.0	97.59	97.82
	n	–	3.35	2.22	0.00	20.67	37.01

고농도 예측의 한계점으로 생각된다. 이 문제를 개선하기 위한 방안으로 여러 가지를 생각해볼 수가 있다. 본 연구에서는 DNN 모델 학습에 적용한 손실함수는 Mean Absolute Error (MAE)였는데 이를 변경하여 고농도 예측정확성에 더 가중치가 높은 기준을 시도해볼 수도 있을 것이다. 그 결과 MAE는 증가하겠지만 고농도 예측의 정확도는 향상시킬 수 있을 것으로 기대한다.

아울러 시계열 데이터를 처리하는데 방식으로 본 연구에서는 시도해보지 않은 순환 인공 신경망

(RNN, Recurrent Neutral Networks)이 있다. RNN 기본 모델은 오래된 과거 데이터의 기울기 값이 소실되는 문제가 있는 것으로 알려져 있지만(Bengio *et al.*, 1994), 이러한 문제를 보완하기 위해 제시된 다양한 모델이 있다. 현재 구축된 DNN 모델의 신경망에 적합한 RNN 응용 모델을 찾아 추가하는 방안을 시도해볼 수 있다. 또한 현재 모델에서 사용되는 입력자료 외에 고농도 예측에 도움이 될 수 있는 입력자료, 예컨데, 중국의 시간별 기상 및 대기질 자료를 추가하는 것도 고려해볼 수 있을 것이다. 특별히 국내에서 발생하는 미세먼지 고농도 상황은 중국으로부터의 유입여부와 밀접하게 관련되어 있는 듯하기 때문이다. 고농도 사건에 대한 예측정확도를 개선하기 위해서 향후 연구가 필요하다.

4.2 수치예보모델과의 비교

국가 대기오염물질 예보는 미세먼지의 장거리 이동, 광화학반응에 의한 2차 생성 및 소멸 등의 복합적인 대기오염 기작을 모의할 수 있는 수치예보모델에 근간을 두고 있다. 수치예보모델은 기상 모델(WRF)과 배출량 모델, 대기질 모델(CMAQ) 등이 연결된 시스템으로 구성되어 있다(NIE, 2014). 미세먼지와 오존의 농도등급에 대한 예측자료는 매일 4회(05시, 11시, 17시, 23시) 공개되고 있다. 장기간에 걸친 예보자료가 누적되어 있기 때문에 DNN 모델예측 성능을 비교하기에 적합하다.

국가 대기오염물질 예보는 전국을 19개 권역으로 구분하여(airkorea.or.kr/dustForecast), 각 권역별로 미세먼지에 대한 대표 등급을 예보하고 있다. 19개 권역을 구체적으로 살펴보면, 서울, 인천, 경기북부, 경기남부, 강원영서, 강원영동, 대전, 세종, 충북, 충남, 광주, 전북, 전남, 부산, 대구, 울산, 경북, 경남, 제주로 구분되어 있다. 미세먼지에 대한 국가 등급예보와 비교하기 위해서 시간별, 측정소별로 생성되는 DNN 모델의 예측자료를 일별, 광역시별로 대수 평균하였다.

PM_{10}에 대하여 DNN 모델과 수치예보모델에 의한 대기질 예측 등급의 정확도를 표 8에 제시하였다. 총

Table 8. Number percentage of the correctly forecasted days for regional PM_{10} concentration grade during 2017.01.04.~2017.09.29.

	Percentage of the correctly forecasted days (the number of correctly forecasted days / the number of whole days)	
	DNN model	Numerical model
Seoul	80.15% (214/267)	82.40% (220/267)
Busan	81.65% (218/267)	79.40% (212/267)
Daegu	85.02% (227/267)	79.40% (212/267)
Incheon	85.77% (229/267)	82.77% (221/267)
Gwangju	75.66% (202/267)	79.78% (202/267)
Daejeon	88.01% (235/267)	85.02% (227/267)
Ulsan	80.15% (214/267)	80.15% (214/267)

비교일수는 267일(2017.1.4.~2017.9.29.)로 예측등급이 관측자료와 일치하는 비율과 일수를 표시했다. PM_{10}의 경우 DNN 모델에 의한 등급예측 정확도는 수치예보모델과 비슷하다. 예측정확도가 가장 높은 경우는 대전권역에 대해 구성된 DNN 모델인데 익일 등급을 정확하게 맞춘 비율이 88%이다. 267일 가운데 235일의 익일 PM_{10}등급을 정확하게 예측하였다. 대전권역에 대한 수치모델의 정확도는 85%로 DNN보다 낮았다. 예측정확도가 가장 낮은 경우도 DNN 모델로 나타났다. 광주권역에 대해서 DNN 모델의 익일 등급예측 정확도는 76%였다. 광주권역에 대한 수치모델의 정확도는 80%로 DNN보다 낮았다.

$PM_{2.5}$에 대하여 DNN 모델과 수치예보모델에 의한 대기질 예측 등급의 정확도를 표 9에 제시하였다. $PM_{2.5}$의 경우 DNN 모델의 익일 등급예측 정확도는 수치모델에 비교하여 현저히 높다. 수치모델의 익일 예측 정확도는 50% 이하이나 DNN 모델의 경우 대부분 권역에서 80%를 상회한다. 예측정확도가 가장 높은 경우는 광주권역에 대해 구성된 DNN 모델로 익일 등급을 정확하게 예측한 비율이 87%이다. 광주권역에 대한 DNN 모델은 267일 가운데 211일의 익일 $PM_{2.5}$등급을 정확하게 맞추었다. 반면에 광주권역에 대한 수치모델의 정확도는 45%로 DNN의 절반 정도로 나타났다. DNN 모델의 예측정확도가 가장 낮은 경우는 부산권역이었는데 익일 등급예측 정확도는

Table 9. Number percentage of the correctly forecasted days for regional $PM_{2.5}$ concentration grade during 2017.01.04.~2017.09.29.

	Percentage of the correctly forecasted days (the number of correctly forecasted days / the number of whole days)	
	DNN model	Numerical model
Seoul	82.02% (219/267)	44.94% (120/267)
Busan	78.65% (210/267)	50.19% (134/267)
Daegu	83.90% (224/267)	45.69% (122/267)
Incheon	79.03% (211/267)	39.33% (105/267)
Gwangju	87.27% (233/267)	44.94% (120/267)
Daejeon	83.15% (222/267)	42.70% (120/267)
Ulsan	80.52% (215/267)	49.06% (131/267)

79%였다. 부산권역에 대한 수치모델의 정확도는 50%로 DNN보다 현저히 낮았다.

PM_{10}과 비교하여 $PM_{2.5}$에 대한 수치모델링의 정확도가 급격히 감소하는 이유를 정확히 알 수는 없다. 다만 $PM_{2.5}$의 경우 PM_{10}보다 2차생성 메카니즘이 훨씬 더 중요하다는 점을 감안할 때 이 과정에 대한 수치모델링의 해석정확도가 상대적으로 떨어지는 것이 아닌가 추측이 된다. 수치모델과 달리 DNN 모델은 2차생성 메카니즘을 따로 구분하여 명시적으로 모델링하지 않는다. DNN 모델은 관측결과와 예측자료 간의 상관성을 분석하여 모형화한 것인데, $PM_{2.5}$와 PM_{10} 모두에 대해서 이러한 상관성의 불확도가 비슷한 것으로 추정된다. 그 결과 PM_{10}과 비교하여 $PM_{2.5}$에 대해서도 DNN 모델의 예측정확도가 유사하게 나타난 것으로 생각된다.

4.3 연산시간

DNN 모델의 예측정확도는 학습자료에 크게 의존한다. 사실상 학습자료가 제공되는 위치에 대해서만 본 연구의 DNN 모델구성이 가능하다. 본 연구에서는 대기질 측정망 자료를 활용하여 학습하였으므로 DNN 모델을 통해 대기질을 예측할 수 있는 위치는 측정소에 국한된다.

측정소당 6가지의 대기오염물질에 대한 DNN 모델을 구성하는데 소요되는 시간, 곧 학습 시간은 표 10에 제시된 컴퓨터 환경에서 약 9분이었다. 학습 자료가 제공되는 예측지점에 대해서 대기오염물질 당 DNN 모델구성에 1분 30초의 학습시간이 필요한 셈이다.

Table 10. Personal computer systems used to run the DNN model code.

ITEM	SPECIFICATIONS
OS	MS Windows 10
CPU	Intel i7 (2 cores, 2.50 GHz)
MEMORY	8 GB
GPU	Not used

모델구성이 완료된 이후 DNN 모델을 통하여 2017년 1월 4일부터 2017년 9월 29일까지 익일 농도예측에 소요되는 시간을 측정하였다. 여기서 익일 예측은 1시간 후부터 24시간 후까지 도합 24개의 오염농도를 6가지 물질에 대해서 산출하는 것을 의미한다. 즉 144개의 농도값이 익일 예측치로 산출된다. DNN 모델을 실행한 전산환경은 일반사양의 PC로 표 10과 같다. 익일 예측에 소요되는 시간은 PC에서 지점에 따라 27±3초로 확인되었다.

DNN 예측은 각 지점별로, 오염물질별로 독립적으로 진행되므로 효과적인 병렬연산이 가능하다. 병렬연산이 가능한 장비에서 운용할 경우 DNN 모델을 사용하여 전국 권역에 대해서 실시간 예측도 가능할 것으로 생각된다.

5. 결 론

입력층과 출력층 사이에 3~4개의 중간 은닉층을 지니는 DNN 모델을 구성하여 시간별 대기오염 농도를 24시간 이후까지 예측해 보았다.

전국 광역시의 115개 대기오염물질 측정소에 대해서 DNN 모델의 예측 값을 관측자료와 비교한 결과 IOA가 0.7~0.8로 확인되었다.

DNN 모델의 예측농도 등급을 전국광역시 측정소의 관측등급과 비교하였다. DNN 모델예측 결과는 관측등급과 60%~80%의 빈도로 일치하였다.

DNN 모델을 수치예보모델과 비교하였다. PM_{10}에

대해서 DNN 모델의 익일 대기질 등급 예보는 75%~85% 빈도로 관측등급과 일치하였는데, 이는 수치예보모델과 비슷한 수준이었다.

그러나 $PM_{2.5}$에 대해서는 수치예보모델의 정확도는 PM_{10}의 절반수준으로 낮아졌으나, DNN 모델의 정확도는 PM_{10}과 유사한 수준을 유지하였다. 즉, $PM_{2.5}$에 대해서는 수치모델에 비해서 DNN 모델의 예측결과가 빈도를 기준으로 2배가량 더 정확하였다.

계산소요시간의 측면에서 DNN 모델의 성능은 매우 우수한 것으로 나타났다. PC에서 운용하였음에도 불구하고, DNN 모델을 사용하여 익일 대기오염농도를 예측할 경우 DNN 처리 소요시간은 1개 지점에 대해서 30초 정도에 불과하였다.

References

Bengio, Y., Courville, A., Vincent. P. (2013) Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828, DOI: 10.1109/TPAMI.2013.50.

Bengio, Y., Simard, P., Frasconi, P. (1994) Learning Long-Term Dependencies with Gradient Descent is Difficult, IEEE Transactions on Neural Networks, 5(2), 157-166, DOI: 10.1109/72.279181.

Dahl, G.E., Sainath, T.N., Hinton, G.E. (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 8609-8613, DOI: 10.1109/ICASSP.2013.6639346.

Goyal, P., Chan, A.T., Jaiswal, N. (2006) Statistical models for the prediction of respirable suspended particulate matter in urban cities, Atmospheric Environment, 40(11), 2068-2077, DOI: 10.1016/j.atmosenv.2005.11.041.

Hinton, G.E. (2007) Learning multiple layers of representation, Trends in Cognitive Sciences, 11(10), 428-434, DOI: 10.1016/j.tics.2007.09.004.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O. (2005) A neural network forecast for daily average PM_{10} concentrations in Belgium, Atmospheric Environment, 39(18), 3279-3289, DOI: 10.1016/j.atmosenv.2005.01.050.

Hrust, L., Klaić, Z.B., Križan, J., Antonić, O., Hercog, P. (2009) Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations, Atmospheric Environment, 43(35), 5588-5596, DOI: 10.1016/j.atmosenv.2009.07.048.

Jeon, S.H., Son, Y.S. (2018) Prediction of fine dust PM_{10} using a deep neural network model, The Korean Journal of Applied Statistics, 31(2), 265-285. (in Korean with English abstract), DOI: 10.5351/KJAS.2018.31.2.265.

Kingma, D.P., Ba, J.L. (2015) Adam: A Method for Stochastic Optimization, the 3rd International Conference for Learning Representations (ICLR), San Diego.

Korea Meteorological Administration (KMA) (2017) ANNUAL CLIMATOLOGICAL REPORT, Publication Number: 11-1360000-000016-10.

McKendry, I.G. (2002) Evaluation of artificial neural networks for fine particulate pollution (PM_{10} and $PM_{2.5}$) forecasting, Journal of the Air & Waste Management Association, 52(9), 1096-1101, DOI: 10.1080/10473289.2002.10470836.

National Institute of Environmental (NIE) (2017) Monthly Report of Air Quality, December 2017, Publication Number: 11-1480083-000177-06.

Perez, P., Reyes, J. (2002) Prediction of maximum of 24-h average of PM_{10} concentrations 30h in advance in Santiago, Chile, Atmospheric Environment, 36(28), 4555-4561, DOI: 10.1016/S1352-2310(02)00419-3.

Schmidhuber, J. (2015) Deep Learning in Neural Networks: An Overview, Neural Networks, 61, 85-117, DOI: 10.1016/j.neunet.2014.09.003.

Shahraiyni, H.T., Sodoudi, S. (2016) Statistical Modeling Approaches for PM_{10} Prediction in Urban Areas; A Review of 21st-Century Studies, Atmosphere, 7(2), 15, DOI: 10.3390/atmos7020015.

Torgo, L. (2010) Data Mining using R: learning with case studies, CRC Press (ISBN: 9781439810187).

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M. (1985) Statistics for the evaluation of model performance, Journal of Geophysical Research, 90(C5), 8995-9005, DOI: 10.1029/jc090ic05p08995.

Authors Information

조경학 (주)볼트시뮬레이션 이사)
이병영 (주)볼트시뮬레이션 기술이사)
권명흠 (주)볼트시뮬레이션 주임)
김석철 (주)볼트시뮬레이션 대표이사)