컨벌루션 신경망 딥러닝을 활용한 미세먼지 예측

이동현[1], 권성준[2], 채상원[3], 강성원[4]

초 록

미세먼지는 인간의 건강에 악영향을 미칠 수 있는 주요 환경 리스크로 신속하고 정확하게 미세먼지를 예측할 수 있는 국가적 예측 관리 감시 시스템이 필요하다. 하지만, 예보관의 판단을 제외한 미세먼지 컴퓨터 예측모델 만의 예보 정확도는 한국의 경우 2016년 기준 52%에 불과하다. 한편, 딥 러닝 기술은 신속하고 정확한 예측이 필요한 분야에서 새로운 해법으로 제시되고 있다. 특히, 공간 정보를 포함하는 대용량의 환경 데이터는 이미지 데이터와 유사한 점이 있고, 이에 이미지 처리 및 예측에 높은 성능을 보이는 딥러닝 모형 중 컨벌루션 신경망(Convolutional Neural Network: CNN)을 활용하여 미세먼지를 예측하였다. 예측 결과 평균 제곱근 오차 기준 2.21 $\mu g/m^3$의 높은 정확도의 예측율을 보였다. 세부적으로, 동일 간격으로 측정되지 않은 환경 데이터인 미세먼지 데이터를 가상의 동일 간격 측정소의 데이터로 격자 보간 하였으며 이를 통해 미세먼지 예측에 CNN 기법을 적용할 수 있게 한 최초의 연구로 높은 미세먼지 예측뿐만 아니라, 향후 CNN딥러닝 기술의 환경 연구에 확장 가능성을 높였다.

주제어: 미세먼지, 딥러닝, 컨벌루션 신경망, CNN, PM10

[1] 한국산업기술대학교 경영학부 조교수

[2] 한국산업기술대학교 산업경영학과

[3] 한국산업기술대학교 IT경영학과

[4] 한국환경정책•평가연구원 선임연구위원

1. 서론

미세먼지 (PM10)는 지름이 10나노미터 보다 작은 입자물질로 미세먼지가 야기하는 대기오염은 유아뿐만 아니라 성인의 심폐질환과 같이 인체에 직접적인 영향을 미칠 수 있다. 세계보건기구 산하 국제 암 연구소(IARC)에 따르면, 미세먼지를 폐암과 방광암의 원인으로 지목, 1급 발암물질로 지정하였다 (IARC, 2013). 관련 연구를 살펴보았을 때, Šrám 외 (2005)는 대기오염이 초기 유아 사망률 특히 유아 호흡기 사망률 간에 유의한 상관관계가 있다고 밝혔으며 [1]. Dockery 외 (1993)은 대기오염이 심폐질환으로 인한 사망과 유의한 관련이 있다고 보고하였다 [2].

즉, 미세먼지는 천식·기관지염·폐암 등의 호흡기 질환, 심혈관 및 뇌혈관 질환, 피부 및 안구질환, 우울증 치매 등의 정신적 질환, 조기 사망 등 인간의 건강에 악영향을 미칠 수 있는 주요 환경 리스크로 신속하고 정확하게 미세먼지를 예측할 수 있는 높은 수준의 국가적 예측 관리 감시 시스템이 필요하고, 예측을 적절히 활용한다면 사회적, 개인적으로 미세먼지에 대응할 수 있는 하나의 길잡이가 될 수 있다. 하지만 국회 환경노동위원회 소속 김삼화 의원이 국립환경과학원으로부터 제출받은 '미세먼지(PM10) 예보 정확도 현황'에 따르면 미세먼지 예보 정확도가 평균 80% 후반 대를 유지하지만, 여기에는 예보관들의 판단이 평균 30%정도 개입된 수치이며 실제 컴퓨터 수치모델에서 산출한 예측 값의 정확도는 평균 50%대라고 설명하였다. 때문에 기존의 예보시스템보다 정확한 시스템이 필요하다.

한편, 딥 러닝 기술은 신속하고 정확한 예측이 필요한 분야에서 새로운 해법으로 제시되고 있다. 딥 러닝 기술은 방대한 양의 데이터를 스스로 학습한다는 특징을 바탕으로 많은 분야에서 실시간으로 다량의 데이터가 축적 되는 빅데이터 환경과 결합한다면 발전가능성이 높은 기술이다. 이러한 딥 러닝을 활용하여 다양한 방면에서 기존의 예측 성능을 발전시켰다는 보고가 있다 [3-4]. Lv 외 (2014)는 Stacked Autoencoder 기반의 딥 러닝 모델을 교통량예측에 사용하였다. 그 결과 이전의 방법들과 비교했을 때 딥 러닝을 사용한 모델이 보다 우수한 성능을 보였음을 밝혔다 [3]. Prasis Poudel 외 (2017)는 long short-term memory (LSTM)을 활용한 태양광 출력 예측 모델을 제시하였다. 그 결과 LSTM을 활용한 모델이 전통적인 분석 방법인 moving average 보다

성능이 월등함을 밝혔다 [4].

딥 러닝 기술을 통한 미세먼지 예측 선행 연구로는 시계열을 고려하는 RNN이나 전통적인 회귀모형을 활용하는 연구가 이루어졌다 [5-6]. Park et al.(2018)은 스페인 Oviedo에서 측정한 대기 오염 요소 데이터를 활용하여 MLP, RBF커널 기반의 SVM, ARIMA, VARIMA를 이용하여 PM10을 예측, 비교하였다. 분석결과 RBF커널 기반의 SVM의 예측성능이 가장 높았다고 보고하였다 [5]. Li et al.(2017)는 PM2.5를 예측하기위해 LSTM을 확장한 LSTME를 새로운 예측방식으로 제안하였다. LSTME는 기본적으로 PM2.5를 설명변수로 사용하는 LSTM에 기온, 습도, 풍속, 가시거리와 같은 기상요인, 발생월, 발생시간 등을 설명변수로 추가한 모델이다. LSTME모델의 비교모델로 STDL, TDNN, ARIMA, SVR, LSTM을 사용되었다. 보조 데이터의 중요성을 평가하기 위해 보조 데이터가 없는 LSTM NN 모델을 사용하여 추가실험을 수행하였다. 실험결과 LSTME모델이 다른 알고리즘보다 성능이 우수하다는 것을 확인하였다 [6].

하지만, 딥 러닝을 사용하여 미세먼지를 예측한 선행 연구는 대부분 공간정보를 활용하지 않았다. 특히 주변의 영향을 받게 되는 환경 및 대기 데이터의 경우, 공간정보를 활용할 수 있다면 더 좋은 성능을 낼 수 있다. 공간정보를 활용하는 방법으로 CNN을 제시하는 선행 연구가 있다. Xiaolei Ma et al.(2017)은 RNN과 같은 딥 러닝 방법은 시간적 상관관계만 고려할 뿐 공간관계를 고려하지 않는다고 보았다 [7]. 이러한 차이를 보완하기 위해 시공간 정보를 포함하는 베이징의 약 1만대의 택시에 장착된 GPS 데이터를 평균치를 계산하여 가로축은 시간간격을 세로축은 각 공간섹션으로 행렬을 만들고 이를 이미지로 변환하였다. 변환한 네트워크 트래픽 이미지를 CNN아키텍처를 사용하여 다음 시간의 속도를 예측하고자 하였다. 비교를 위한 예측모델로 OLS(Ordinary Least Square), KNN(K –Nearest Neighbors), RF(Random Forest), ANN(Artificial Neural Nerworks), SAE(Stack Autoencoder), RNN(Recurrent Neural Network), LSTMNN(Long Short-Term Memory Neural Network)를 사용하였는데, 다른 예측모델과 비교했을 때 CNN이 가장 좋은 성능을 보였다고 보고하였다 [7]. 특히, 공간 정보를 포함하는 대용량의 환경 데이터는 이미지 데이터와 유사한 점이 있고, 이에 이미지 처리 및 예측에 높은 성능을 보이는 컨벌루션 신경망(Convolutional Neural Network: CNN)을 활용 한다면 높은 정확도의 예측이 가능할 수 있다. 허나,

기본적으로 CNN 기술은 인접 픽셀의 특징들을 공유된 가중치를 가지는 Convolution 필터와 Pooling 필터로 학습하는 딥러닝 기술로, 일반적으로 개별 측정소가 있는 환경 데이터는 정확히 동일 위치만큼 떨어져서 측정소를 설치하는 것이 물리적으로 불가능하며, 특정 섹터에 센서가 몰려있는 경우도 많다. 이러한 이유로 일반적인 환경 데이터에 바로 CNN을 적용하기에는 무리가 있다. 이에 본 연구는 환경 데이터의 CNN 적용이라는 큰 맥락 아래 본 연구를 수행하였다. 세부적으로, 본 연구는 동일 간격으로 측정되지 않은 환경 데이터인 미세먼지 데이터를 가상의 동일 간격 측정소의 데이터로 격자 보간 하였으며 이를 통해 CNN 기법을 적용한 최초의 연구이다.

미세먼지를 예측하기 위해 관련문헌검토를 통해 예측에 사용할 데이터를 검토하고, 모형을 설계하였다. 데이터는 기상자료개방포털(https://data.kma.go.kr) 및 에어코리아(https://www.airkorea.or.kr/realSearch)에서 수집하고, 위·경도를 10개의 구간으로 나누어 결측값 대치와 보간 작업 등 데이터 전처리 과정을 진행하였다. 분석에 사용된 예측 모형은 지역별 공간정보를 활용할 수 있는 CNN을 기반한 모형을 구축하였다. 예측 모델의 최적화를 위해 4가지 CNN 아키텍처를 구성하였고, 각 아키텍처에서 사용되는 최적화 파라미터인 optimizer를 adam과 stochastic gradient descent 두 가지로 변화를 주어 예측 후 비교 하였다. 최종적으로, 미세먼지의 공간적 특성을 반영하여 보다 정확한 예측을 할 수 있는 분석방법을 제안하고, 환경 연구에서 CNN기술 기반의 딥 러닝의 확장가능성을 파악하고자 하였다.

2. 자료와 연구방법

2-1. 자료

미세먼지는 기상요인 및 대기오염 요소에 영향을 받는다는 다수의 연구들이 존재한다 [8-9]. Xie et al.(2015)는 중국 31개 지역의 2013년 3월 22일부터 2014년 3월 31일 사이에 수집된 286개의 데이터를 기반으로 SO2와 NO와 같은 대기오염물질이 PM10에 상관관계가 있는지 분석하였다. 그 결과 대기오염물질에 따라 PM이 지역별로 약간의 차이는 있지만 평균적으로 중간 정도의 상관관계를 가지는 것으로 보고하였다 [8]. Li et al. (2015)는 중국 사천성의 도시지역에서 발

생하는 대기오염물질을 방지하기 위한 대비책을 제안하기 위해 PM10, PM2.5, PM1.0의 시간적 분포 특성과 PM과 기상요인간의 관계를 분석하였다. 그 결과 상대습도 제외한 다른 기상요인과 PM사이에 유의한 상관관계가 존재한다고 보고하였다 [9].

이에 본 연구에서는 한국환경공단 에어코리아에서 제공하는 전국 17개 시·도 369개 측정소의 대기오염 측정소 데이터와 기상청 기상자료 개방포털에서 제공하는 전국 96개소의 종관기상관측장비(ASOS) 데이터를 활용하였다. 자료수집기간은 2014년부터 2016년까지 3년이다.

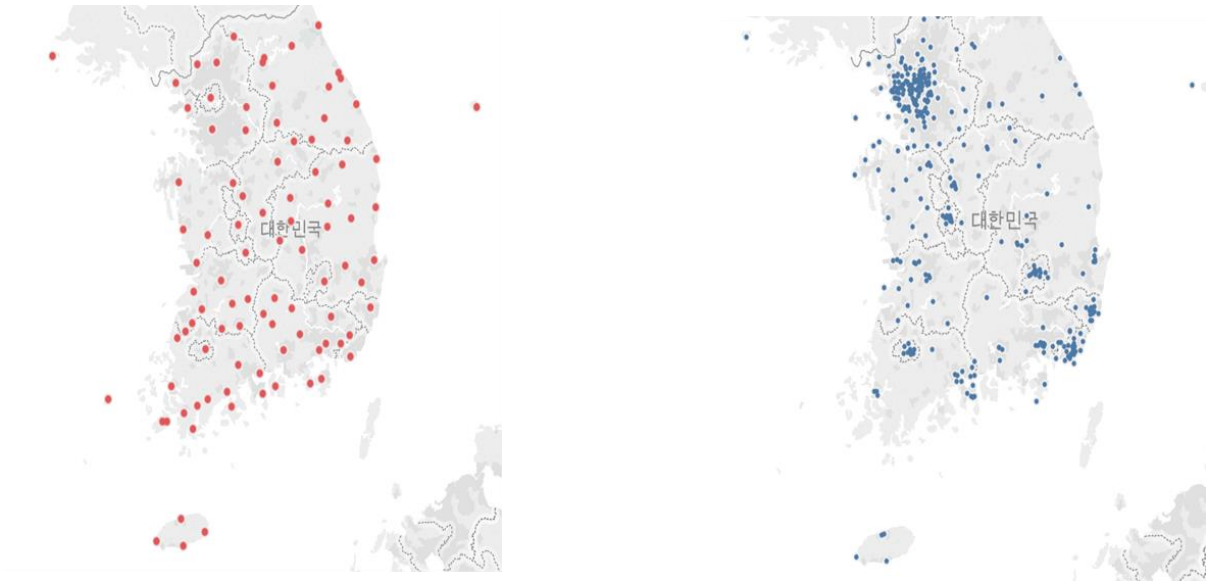

그림 1. 관측소 지점 (좌: 종관기상관측장비, 우: 대기오염측정소)

에어코리아에서 제공하는 대기오염 측정소 데이터는 아황산가스(*SO2*), 일산화탄소(*CO*), 오존(*O3*), 이산화질소(*NO2*)이고, 기상청 기상자료 개방포털에서 제공하는 데이터는 시간당 기온(℃), 강수량(*mm*), 풍속(*m/s*), 풍향(*deg*)이다.

2-2. 분석 방법론

먼저 종관기상관측장비 96개소와 대기오염 측정소 369개의 위치가 다르며 CNN에 사용할 동일한 구간의 이미지로 만들기 위해 GIS 보간을 실시하였다.

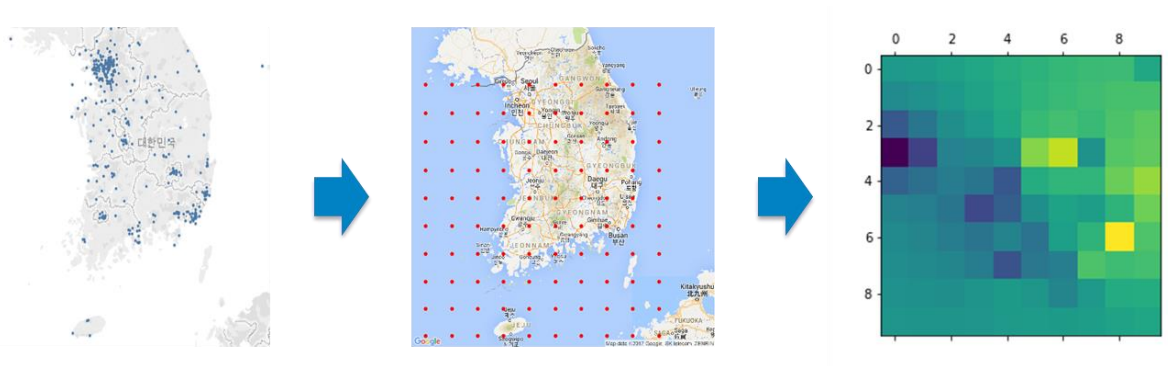

그림 2. 공간 환경 데이터의 컨벌루션 신경망(CNN) 적용을 위한 데이터 변환 (좌-미세먼지 측정소, 중간- 10x10 가상 관측소, 우-10x10 격자 변환 데이터)

종관기상관측장비의 위경도를 대기오염 측정소에 맞추어 계산하기 위해 역거리가중법(IDW)을 사용하였다. 역거리가중법은 가까운 거리일수록 높은 가중치를 두어 보간점의 값을 계산하는 보간 방법이다. 위 보간 방법으로 두 데이터를 통합하였다. 마지막으로 동일 거리의 데이터로 변환하여 CNN을 적용하기 위해 최서북단 위경도와 최동남단 위, 경도 사이를 10 X 10 의 동일 구간 격자로 나누어 가상의 측정소를 구성한 뒤 총 100개의 지역의 동일 구간으로 떨어진 가상의 측정소 위치로 데이터를 격자 변환하였다.

2-3. CNN

우리는 미세먼지 공간정보를 반영한 예측을 위해 CNN을 사용하였다. CNN은 Input 이미지에 필터를 통해 output 이미지를 출력한다. 이미지 내 각 위치는 고유한 값을 가지며, 필터는 학습을 통해 변하는 값을 가진다. padding = valid를 사용하였을 때, Input 이미지의 특정 위치의 값이 필터의 값과 곱해지고 그 값들을 모두 더해 output 이미지를 출력한다. 이런 과정을 거치게 되면 필터의 크기에 따라 output 이미지의 크기가 변하게 된다. padding = same을 사용하면 output 이미지의 크기를 유지하기 위해 input 이미지 테두리에 임시로 값을 추가하기도 한다. 그림 3과 그림 4는 padding에 따른 CNN모형이다.

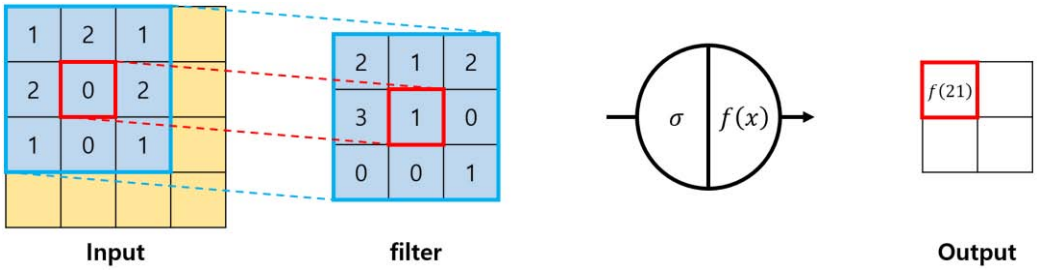

그림 3. CNN 모형 (padding = valid)

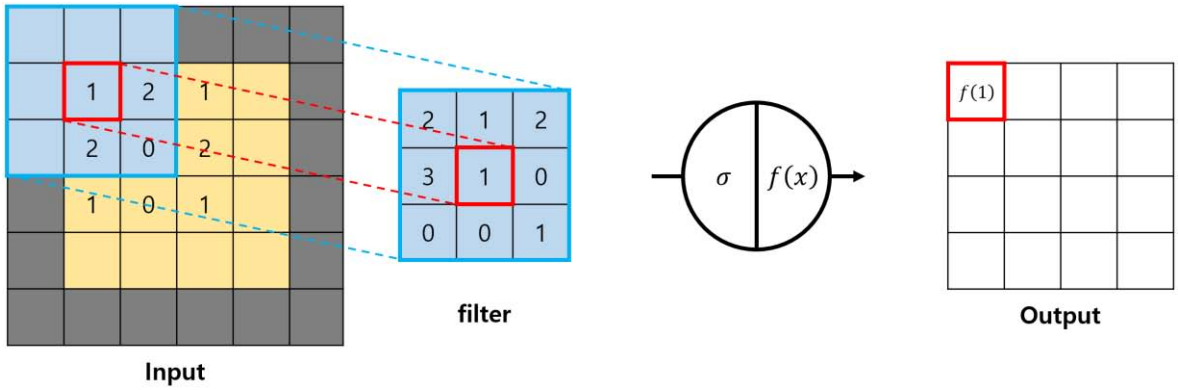

그림 4. CNN 모형 (padding = same)

CNN은 컨벌루션 필터의 크기와 개수 및 파라미터에 따라 다양한 아키텍처를 만들 수 있다. 본 연구에서는 표1과 같은 4가지의 CNN 기반의 미세먼지 예측 특화 모델을 개발하여 예측에 사용하였다.

표 1. CNN기반 미세먼지 예측 모형 아키텍처

아키텍처	Arc 1	Arc 2	Arc 3	Arc 4
input shape	7, 10, 10, 1	1, 10, 10, 9	7, 10, 10, 9	7, 10, 10, 9
filter 1	conv3D(7, 1, 1) * 1 padding = valid	conv3D(1, 2, 2) * 9 padding = same	conv3D(7, 1, 1) * 9 padding = valid	conv3D(7, 1, 1) * 9 padding = valid
filter 2	conv3D(1, 2, 2) * 1 padding = same	conv3D(1, 2, 2) * 1 padding = same	conv3D(1, 2, 2) * 9 padding = same	conv3D(1, 2, 2) * 9 padding = same

filter 3	-	-	conv3D(1, 1, 1) * 1 padding = same	reshape
output shape	1, 10, 10, 1	1, 10, 10, 1	1, 10, 10, 1	100,1

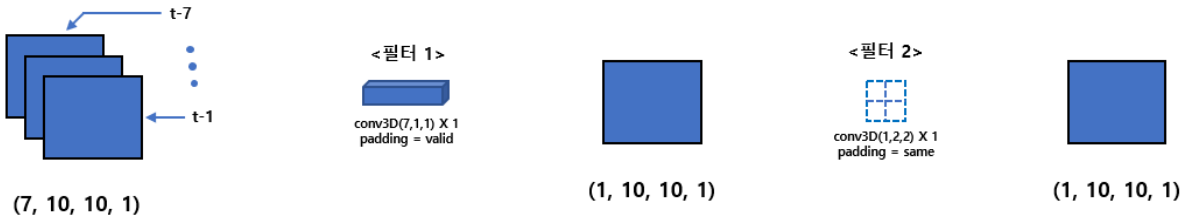

그림 4. Arc 1 모형

그림 4은 Arc 1을 도식화한 것이다. Arc 1은 PM10의 이전 7시간으로 이후 1시간을 예측하는 모형이다. Arc 1의 input 데이터는 PM10의 10X10 이미지를 이전 7시간을 겹쳐 만든 3차원 이미지다. 필터 1은 지역별로 이전시간을 고려해 다음시간을 예측하는데 사용하고 padding을 valid로 설정하여 필터 1을 통해 나온 이미지는 10X10 이미지 1개가 된다. 필터 2는 필터 1에서 만들어진 10X10 이미지에 주변공간을 고려하여 이미지를 생성한다.

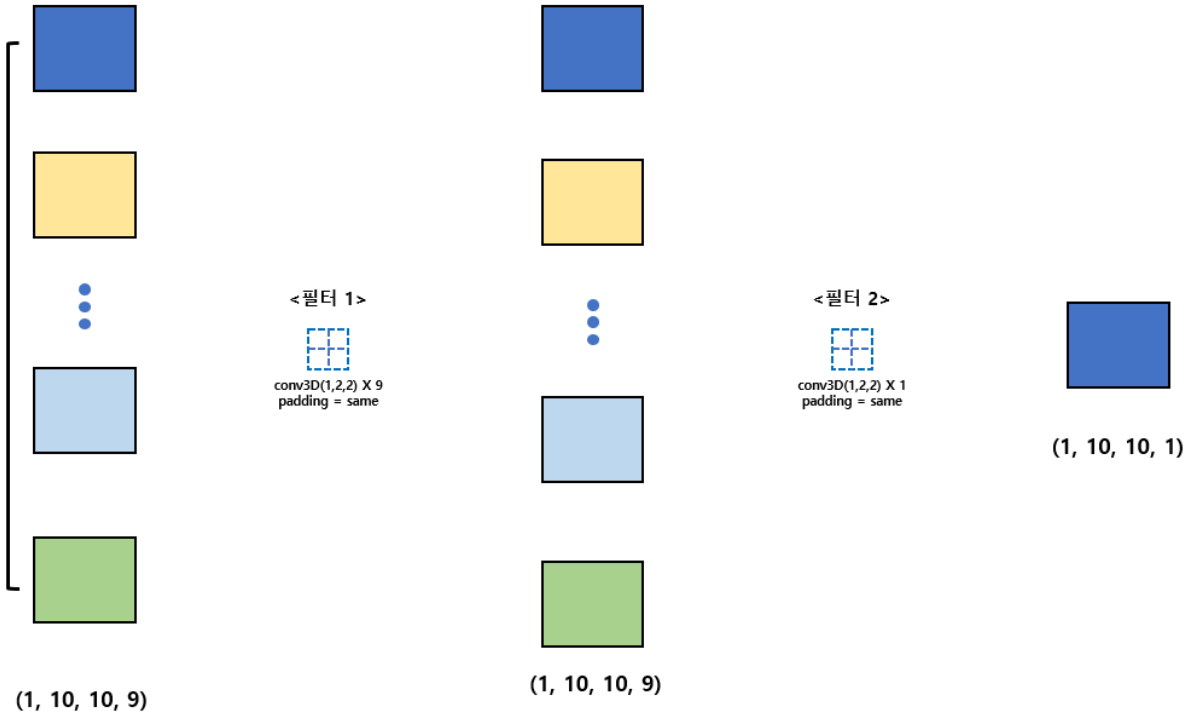

그림 5. Arc 2 모형

그림 5은 Arc 2을 도식화한 것이다. Arc 2는 PM10에 다른 요인들을 추가한 모형이며, 각 요인들의 이전 1시간을 고려해 이후 PM10을 예측한다. Arc 2의 input 데이터는 각 요인들의 10X10 이미지를 9개의 채널로 구성한 이미지다. 필터 1은 각 변수들의 주변 공간정보를 고려하여 변수별로 10X10 이미지를 생성한다. 필터 1을 통해 나온 변수별로 이미지는 필터 2를 통해 종합하여 1개의 10X10 이미지를 생성한다.

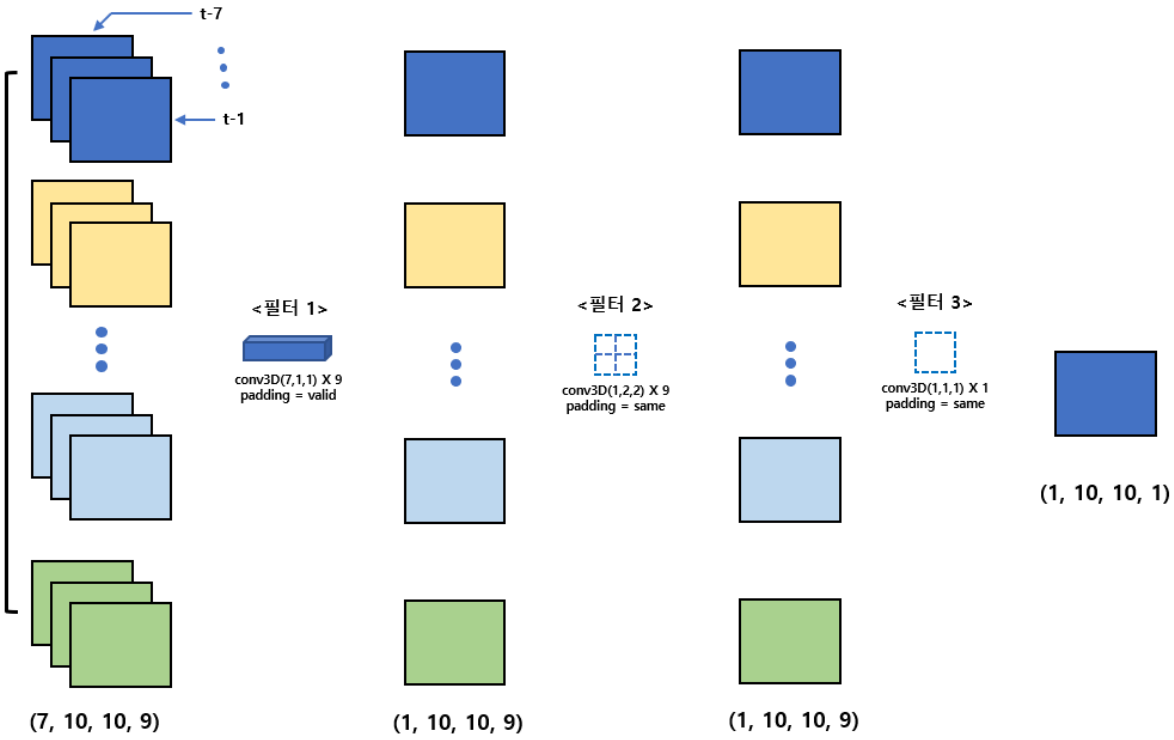

그림 6. Arc 3 모형

그림 6은 Arc 3을 도식화한 것이다. Arc 3은 모든 변수들의 이전 7시간의 이미지로 PM10을 예측하는 모형이다. Arc 3의 input 이미지는 모든 변수들의 이전 7시간의 10X10 이미지를 겹쳐 만든 3차원 이미지를 9개의 채널로 변환한 이미지다. 필터 1은 변수별로 이전 7시간을 고려하여 10X10 이미지를 생성한다. 필터 2는 생성된 10X10 이미지에 주변 공간정보를 고려하여 다시한번 변수별로 10X10 이미지를 생성한다. 필터 3은 필터 2를 통해 생성된 변수 별 10X10 이미지를 고려하여 PM10 이미지를 생성한다.

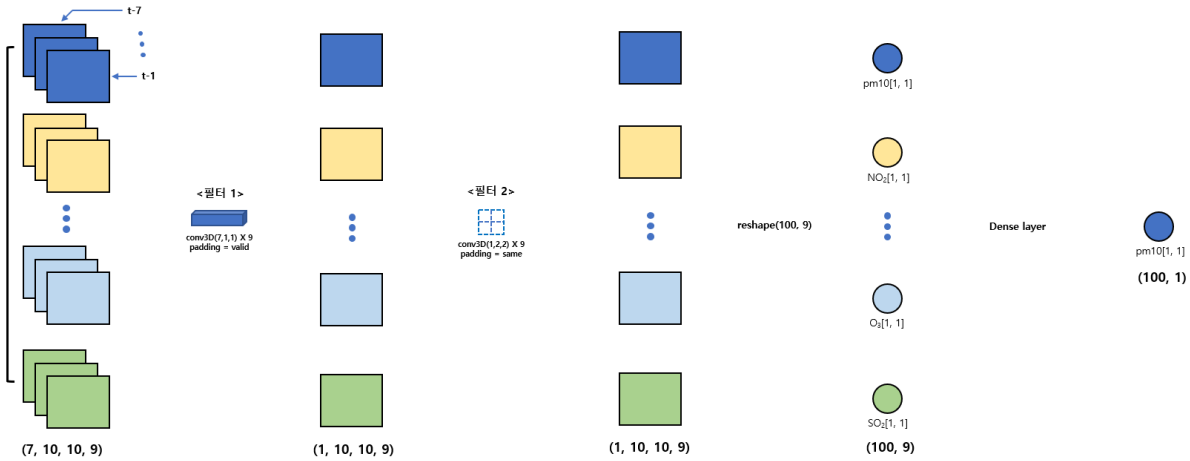

그림 7. Arc 4 모형

그림 7는 Arc 4을 도식화한 것이다. Arc 4는 Arc 3을 확장한 것으로 convolutional filter로 전체지역을 한번에 예측하는 것이 아니라 지점별로 나누어 예측하는 모형이다. Arc 4의 input 데이터는 Arc 3의 input 데이터와 같다. 필터 1과 필터 2의 역할도 Arc 3의 필터 1, 필터 2와 같다. 필터 2를 통해 만들어진 이미지는 reshape layer를 통해 이미지의 지점별로 나누어진다. 지점별로 나누어진 데이터는 전결합층을 통해 최종적으로 PM10을 지점별로 예측하게 된다.

분석에 사용한 모든 데이터는 전체 기간을 시계열 순으로 60%의 훈련데이터와 20%의 검증데이터, 20%의 예측데이터로 나누어 분석을 진행하였고, 아키텍처 구현은 Python 프로그래밍 언어의 Keras 패키지를 사용하였다.

3. 결과

CNN을 활용한 4가지 아키텍처에서 Optimizer를 SGD(Stochastic Gradient Descent) Adam으로 나누어 예측, 비교하였다.

표 2. 예측 결과

Optimizer	Arc 1	Arc 2	Arc 3	Arc 4
SGD	17.54	18.17	17.55	NaN

Adam	2.36	2.76	2.21	2.54

표 2는 예측결과이다. Optimizer을 SGD로 설정했을 때보다 Adam으로 설정했을 때 예측 성능이 월등히 뛰어남을 알 수 있었다. 아키텍처별로 비교하였을 경우 이전 시간만을 고려한 Arc 2 보다 PM10의 이전 7시간으로 예측한 Arc 1과 이전 7시간의 다변수를 함께 고려한 Arc 3가 비교적 오차가 적음을 확인할 수 있었다. Arc 3를 지역별로 예측한 Arc 4는 오차가 Arc 3보다 크게 나왔다. 모든 모형 중 가장 높은 성능을 나타낸 모형은 Optimizer을 Adam으로 설정한 Arc 3가 가장 오차가 적어 예측성능이 가장 높음을 알 수 있었다.

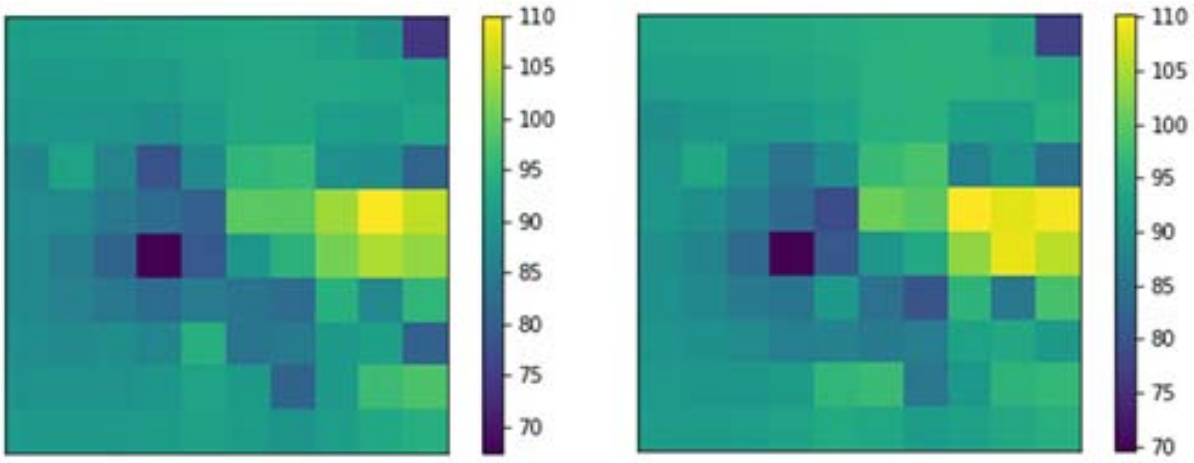

그림 8. Arc 3(Adam)의 예측 이미지 예시 (좌-예측, 우-실제)

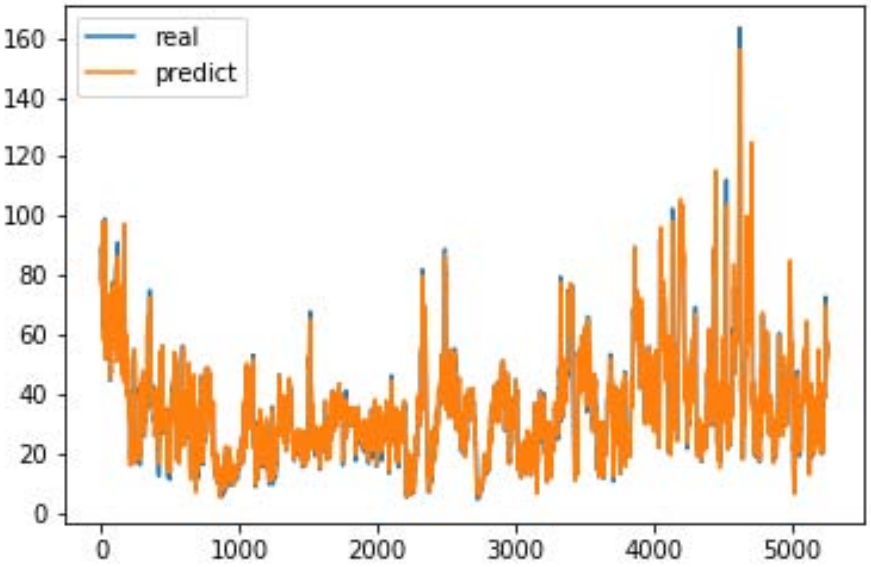

그림 9. Arc 3(Adam)의 지역(4, 4) 예측 그래프

그림 8은 예측 성능이 가장 높은 Arc 3로 예측된 이미지와 실제 이미지를 비교한 예시이다. 예측 이미지와 실제 이미지를 비교하였을 때 높게 예측한 지역, 다시 말해 그림 중앙 우측에서 약간의 차이만 있을 뿐 큰 차이가 없는 것으로 나타났다. 특히 평균적인 예측을 한 그림 좌측은 차이가 거의 없는 것으로 나타났다. 따라서 CNN을 활용한 예측모델의 성능이 매우 좋은 것으로 평가된다. 그림 9는 그림 8의 결과에서 중앙지점의 예측치를 그래프로 나타낸 예시이다. 모델의 예측 값이 실제 값과 큰 차이가 보이지 않음을 알 수 있다.

4. 결론

분석 결과에 따른 함의는 다음과 같다. 본 연구는 CNN을 활용하여 공간정보를 활용한 예측을 미세먼지 분야에서 최초로 제안하였고, 평균 제곱근 오차 기준 2.21$\mu g/m^3$ 의 높은 정확도의 예측 율을 나타냈다. 일반적인 환경 데이터가 동일 구간이 떨어져서 측정된 데이터가 아니기 때문에 CNN 적용이 힘들었다면, 위 연구의 방법을 활용한다면 환경 데이터를 CNN 적용이 가능하도록 변환할 수 있고, 이를 통해 공간정보를 충분히 예측에 활용하는 환경 CNN 적용을 가능하게 한 최초의 연구이다. 기존의 분석방식은 시계열만을 고려하거나 혹은 공간정보를 고려하더라

도 공간정보를 시계열 분석방식에 대입하는 방식이었다. 하지만 CNN을 활용할 경우 분석방법 자체가 공간정보를 활용하고 있기 때문에 높은 성능의 예측을 할 수 있게 된 것으로 보인다.

파라미터 조정은 성능에 영향을 미치는 만큼 시간과 노력을 요하는 작업이기에 향후, 파라미터를 다양하게 변경하며 최적화를 한다면 예측 성능이 더 높아질 가능성이 있다. 이 연구에서는 이미지 인식분야에서 사용하였던 CNN기법을 미세먼지예측에 활용해 보았다. 추후 미세먼지뿐만 아니라 이미지로 표현할 수 있는 그리고 정확하고 신속한 예측이 필요한 다양한 환경 분야에서의 확장이 가능할 것으로 기대된다.

참고문헌

1. Radim J. Šrám et al. (2005). Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. Environ Health Perspect, 2005 Apr; 113(4): 375–382.
2. Douglas W. Dockery et al. (1993). An Association between Air Pollution and Mortality in Six U.S. Cities. The New England Journal of Medicine, 1993; 329: 1753-1759.
3. Yisheng Lv et al. (2014). Traffic Flow Prediction With Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, Volume: 16 Issue: 2.
4. Prasis Poudel and Bongseog Jang. (2017). Solar Power Prediction Using Deep Learning Technique. Advanced Science and Technology Letters, Vol.146 (FGCN 2017), pp.148-151
5. P.J.García Nieto et al. (2018). PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study. Science of The Total Environment, Volume 621, 15 April 2018, Pages 753-761
6. Xiang Li et al. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution, Volume 231, Part 1, December 2017, Pages 997-1004
7. Xiaolei Ma et al. (2017). Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. Open Access, Sensors 2017, 17(4), 818
8. Yangyang Xie et al. (2015). Spatiotemporal variations of PM 2.5 and PM 10 concentrations between 31 Chinese cities and their relationships with SO2 , NO2 , CO and O3. Particuology, ISSN: 1674-2001, Vol: 20, Page: 141-149
9. Yang Li et al. (2015). Variations in PM10, PM2.5 and PM1.0 in an Urban Area of the Sichuan Basin and Their Relation to Meteorological Factors. Atmosphere 2015, 6(1), 150-163