<별지 제3호 서식>

PBL 결과 보고서

<1. PBL결과 개요보고서>

PBL책임자 (소속, 서명)	팀장	경북대학교 컴퓨터학부	배홍직 (서명 또는 인)
	멘토	(주)산들정보통신	최상식 (서명 또는 인)
PBL과제명	딥러닝(텐서플로우) 미세먼지 예측		
PBL결과 요약	경북대학교 근처에 위치한 신암동 대기측정소의 공공데이터를 기반으로 딥러닝 DNN 모델을 만들어 미세먼지 PM_{10}, $PM_{2.5}$의 한 시간 후 농도를 예측하였다. 이는 각각 약 78%, 89%의 정확도를 나타냈다. 저농도일 때의 정확도는 높았으나, 고농도로 급변하는 경우, 정확도가 떨어졌다.		
PBL결과 및 활용방안	계속해서 개발·개선해나가며 높은 정확도의 모델을 만든 후, 웹 또는 앱 어플리케이션을 활용해서 사용자들에게 알려주는 서비스를 제공하고자 한다.		
참 고 및 건의사항	프로젝트 GitHub 주소 https://github.com/SNMHZ/finedust_predict		
PBL결과 발표계획	포스터를 이용한 게시판 발표		

<2. PBL결과 요약문>

PBL과제명	**딥러닝(텐서플로우) 미세먼지 예측**	PBL책임자	최상식 (서명 또는 인)

연 구 내 용

산업이 발전하며 미세먼지의 농도가 계속해서 증가하고 있다. 미세먼지는 우리에게 심각한 악영향을 끼쳐 우리나라에 커다란 환경 문제 중 하나로 언급되고 있다. 따라서 우리는 딥러닝(텐서플로우)를 활용하여 다양한 환경인자들을 통해 미세먼지 농도를 예측해보고자 하였다.

우선 연구는 대구 신암동에 위치한 신암동 대기측정소의 데이터만을 사용하여 진행되었다. 이는 미세먼지가 환경인자들의 영향을 받을 뿐만이 아니라, 측정소가 위치한 지형·지리적 정보와 인위적 요인(자동차, 공장, 건설현장 등)에도 영향을 받아 측정소 별로 같은 환경에서도 다른 결과가 나오기 때문이다.

미세먼지 농도는 한 시간 후의 농도만을 예측하였다. 이는 풍속, 강수와 같은 환경인자들이 급변할 수 있기 때문에 그 이상의 미세먼지 농도를 예측하는 것은 정확도가 매우 떨어지기 때문이다.

데이터를 수집하고 전처리, 정규화 과정을 거쳐 데이터를 가공해주었다. 그 후, 인자를 각각 PM_{10}, PM_{10}+대기 데이터, PM_{10}+대기 데이터+기상 데이터를 사용하여 심층 신경망 모델을 만들어주었다. 각각의 정확도는 약 78%, 74%, 78%로 확인되었다. $PM_{2.5}$+대기 데이터+기상 데이터를 사용한 심층 신경망 모델의 정확도는 약 89%로 확인되었다.

<3. 결과보고서>

<차 례>

Ⅰ. PBL 목적 및 방법

1. PBL의 목적, 필요성 및 PBL목표

• 미세먼지 정의

미세먼지란 사람 눈에는 보이지 않는 아주 작은 미세한 입자로 크기에 따라 PM_{10}(미세먼지), $PM_{2.5}$(초미세먼지)로 구분된다. 이는 자동차 배기가스, 화학 발전소, 공장 등에서 발생하거나 황산화물(SOx), 질소산화물(NOx), 암모니아(NH3), 휘발성 유기화학물(VOCs) 등의 전구물질이 대기 중에서 반응을 일으켜 발생한다.

• 미세먼지 위험성

미세먼지는 호흡기를 통해 신체 내부에 들어가 단기적으로는 천식 발작, 급성 기관지염, 부정맥과 같은 증상을 악화시키고 장기적으로는 심혈관질환, 호흡기질환, 폐암과 같은 증상을 일으킬 수 있다. 이처럼 미세먼지에 장기간 노출되면 건강을 해칠 뿐만 아니라 심하면 사망에 이르기까지 악영향을 끼치게 된다. 이에 따라 세계보건기구(WHO)에서는 2013년, 대기오염과 함께 미세먼지를 1군 발암물질로 분류했다.

• 미세먼지 예측 목적

우리나라의 많은 공장, 발전소와 늘어나는 자동차 보급률에 따라 미세먼지양은 계속해서 늘어나고 있다. 이에 따라 미세먼지를 예측하고 대비하고자 다양한 연구들이 진행되었지만 미세먼지 농도는 대상지역에 따라 기상 · 지리적 특성, 주변 도시의 영향, 내부배출원 등이 달라지기 때문에 각각의 지역에 따라 적합한 모델을 찾아야한다(안지민, 2018). 따라서 우리 팀은 우리 경북대학교 지역에 맞는 예측 모델을 만듦으로써 학생들의 외출 여부 및 마스크 착용 결정에 도움을 주고자 한다.

2. PBL 내용, 범위 및 방법

• 수집 데이터

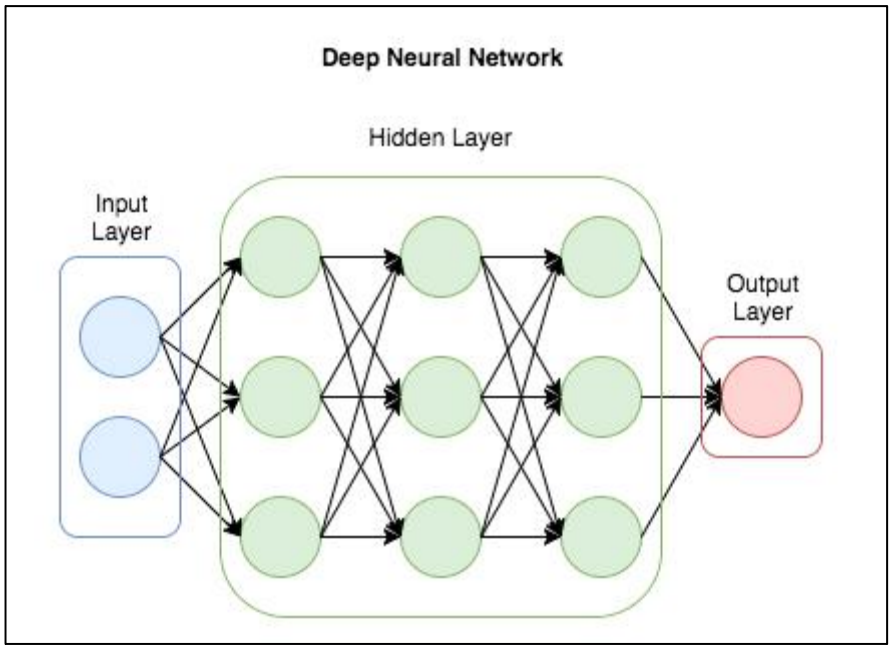
다음 [표1]과 같이 각각 데이터를 에어코리아, 기상청에서 CSV 파일 형식의 데이터를 수집하였다. 에어코리아에서는 대구에 위치하고 있는 각 측정소별로 시간별 데이터를 얻을 수 있었다. 반면에 기상청에서는 대구시 전체에 대한 시간별 데이터를 얻을 수 있었다. 따라서 대기요인에 관한 데이터를 동으로 구분하고 대구시 전체에 대한 기상요인 데이터를 그대로 적용하였다.

구 분	인 자	데이터 단위	데이터 범위	데이터 형태
미세먼지 및 대기요인	PM_{10}, $PM_{2.5}$, O_3, NO_2, CO, SO_2	시간별	2003~2017	csv
기상 데이터	기온, 강수량, 풍속, 습도	시간별	2003~2017	csv

[표1] 수집 데이터

• 예측 방법

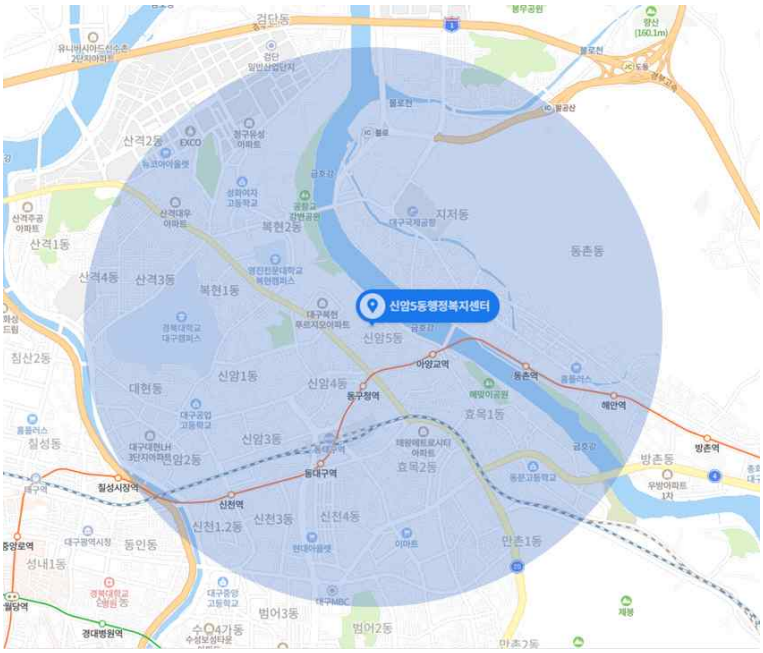
파이썬 텐서플로우(TensorFlow)를 활용하여 다양한 인자들을 추가 혹은 제거하면서 심층 신경망(Deep Neural Network) 모델을 만듦으로써 정확도 높은 예측 모델을 찾고자 하였다.



[그림1] 심층 신경망 모델

• 예측 범위

미세먼지 농도는 한 시간 후만의 농도를 예측하였다. 이는 미세먼지의 농도는 대기 인자, 기상 인자뿐만 아니라 미세먼지를 발생시키는 1차 발생원인 자동차, 공장, 발전소, 건설현장 등의 영향도 받기 때문에 예측이 어렵기 때문이다.
경북대학교에서 가장 가까운 대기 측정소 위치는 신암동으로, 신암동 대기 측정소는 신암5동행정복지센터에 위치하고 있다. 거리는 반경 3km 이내로, 해당 측정소에 대한 예측 모델의 결과가 유의미할 것이라고 판단하였다.



[그림2] 신암동 측정소 위치 및 3km 반경 표시

Ⅱ. PBL수행 내용 및 결과

• 데이터 수집

미세먼지 예측을 위한 모델 설계에 앞서 학습에 필요한 대구시의 기상 및 대기 인자들에 대한 데이터 셋을 수집하였다. 에어코리아(airkorea.or.kr)에서 대구시 내에 있는 측정소들의 PM_{10}, $PM_{2.5}$, SO_2, O_3, NO_2, CO 인자들과 기상자료개방포털(data.kma.go.kr)에서 기온, 강수, 풍속, 습도 인자들에 대한 매 시간 데이터를 2003년부터 2017년까지 .csv 파일 형식으로 제공받았다.

• EDA(탐색적 데이터 분석) 과정

EDA란 수집한 데이터를 도표, 그래프, 통계 등을 활용하여 다양한 각도에서 관찰하고 이해하는 과정을 뜻한다. 따라서 우리 팀도 수집한 데이터를 딥러닝 실행 전에 다양한 통계적 방법을 통해 관찰하고 이해하기 위해 아래의 EDA 과정을 거쳐주었다.

- 데이터 전처리

데이터 전처리를 위해 N/A값들을 제거해주었다. N/A값들을 추적해본 결과, 기기 센서 고장 또는 점검 등에 의한 미측정으로 확인되었다. 데이터는 2017년 신암동의 수집된 데이터 8,759개 중 결측 데이터 1,449개를 제거한 7,310개를 통해 학습하였다.

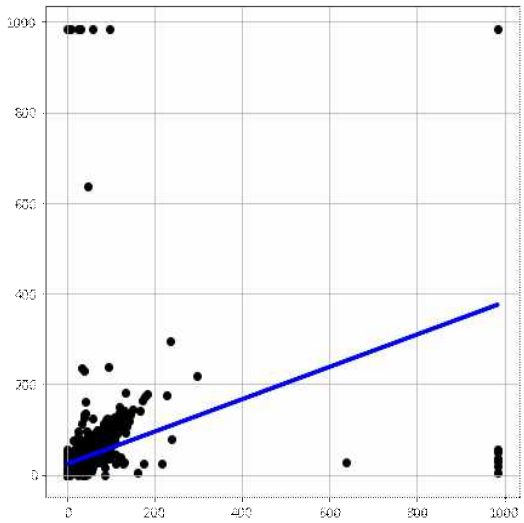
구분	데이터 수	결측 수
대기 데이터	8,759	1,449
기상 데이터	8,759	0

[표2] 수집 데이터 수

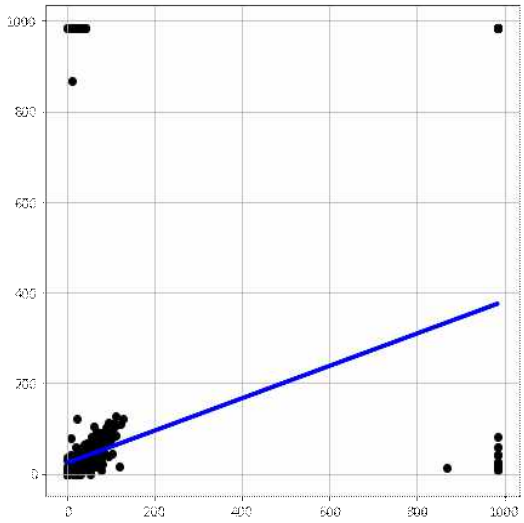
- **선형회귀분석**

파이썬의 사이킷런(Scikit-learn)을 활용해서 PM_{10}, $PM_{2.5}$만을 사용하여 미세먼지 단순선형회귀분석을 실행하였다. 시각화 결과 이상값이 다수 존재한다는 것을 발견할 수 있었다.

PM_{10}



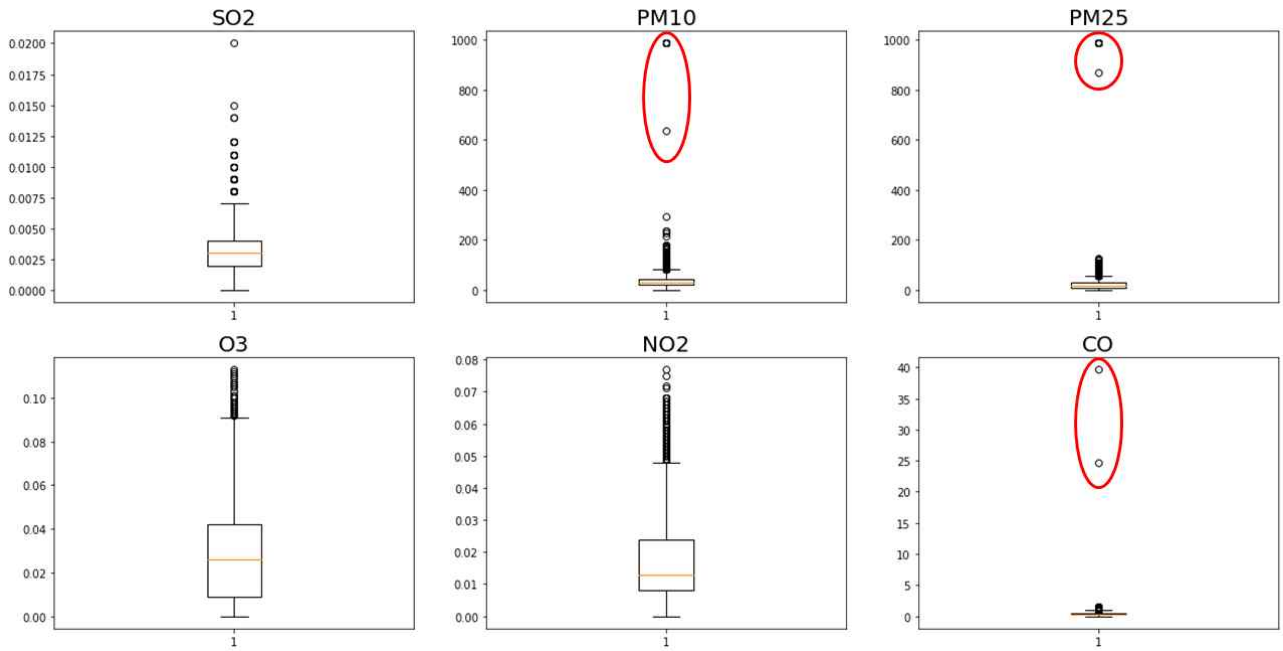
$PM_{2.5}$

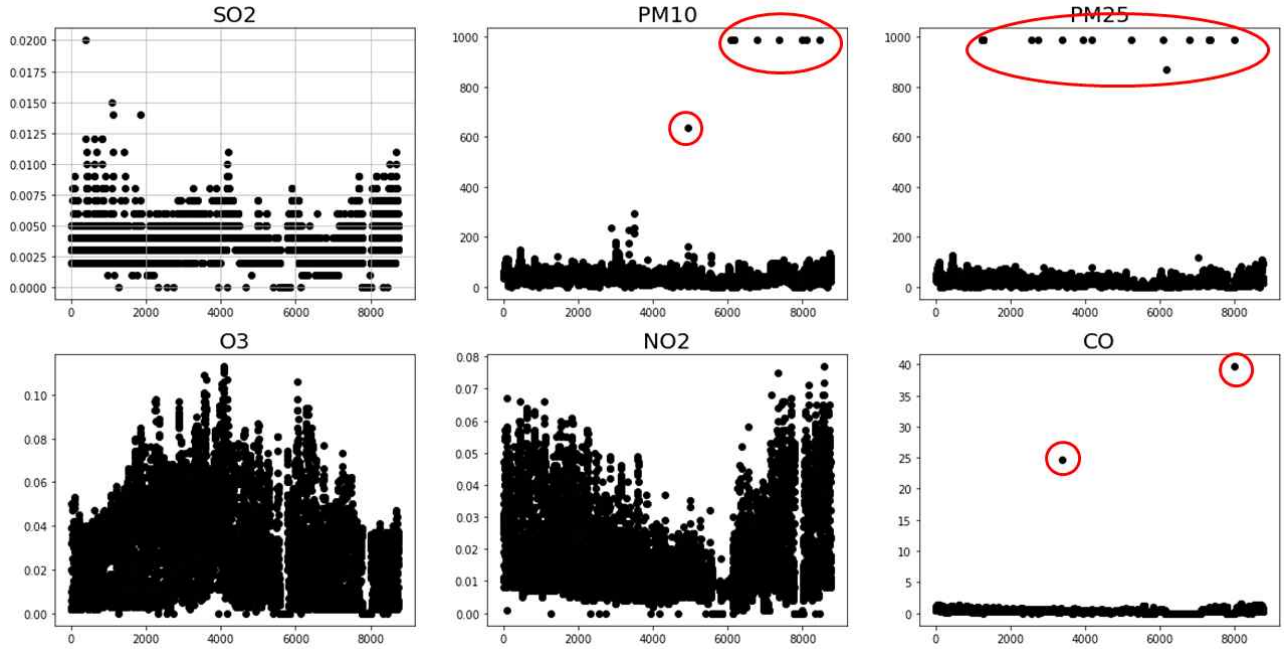


[그림3] PM_{10}, $PM_{2.5}$에 대한 단순선형회귀분석 결과

- **이상값(Outlier) 제거**

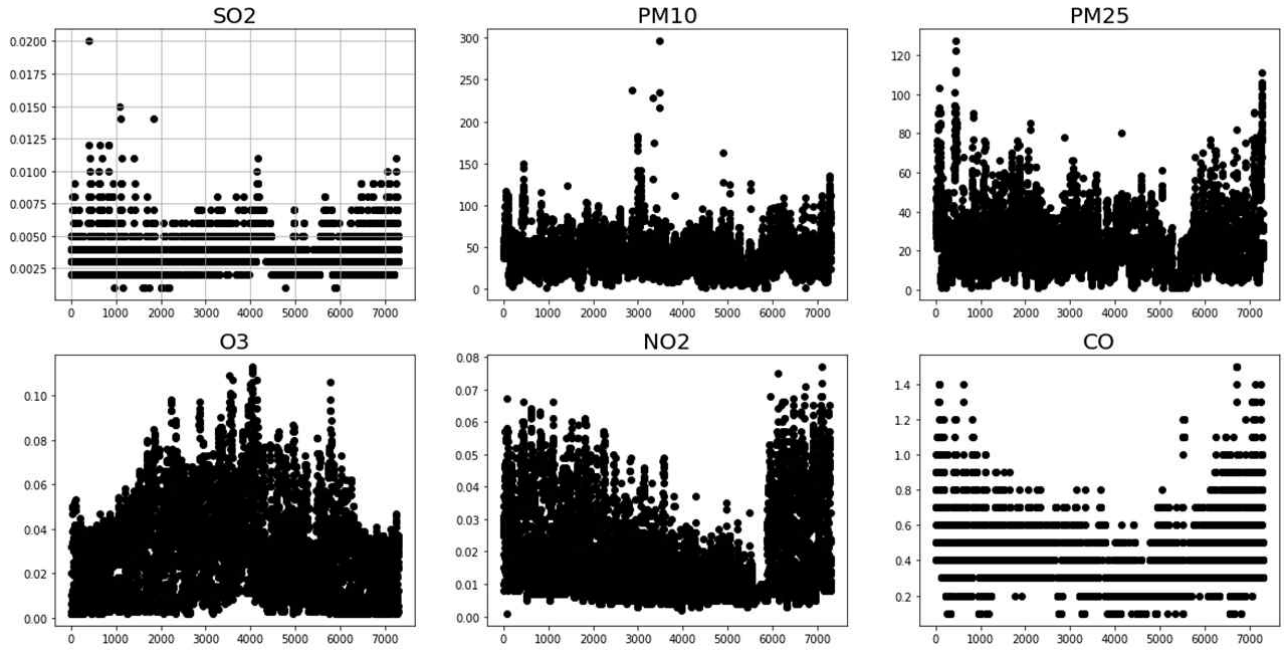
기상 데이터는 결측값, 이상값이 없기 때문에 대기 데이터에 대해서만 outlier 제거를 해주었다. 일반적으로 outlier 제거에 1.5 IQR Rule 등을 활용한다. 하지만 미세먼지의 경우 대부분 저농도 구간에 편향되어있고 고농도 구간은 극히 적기 때문에 1.5 IQR Rule과 같은 통계론적 방법을 사용할 경우, 고농도 데이터들이 전부 outlier로 제거된다(전영태, 2020). 따라서 대기인자들의 기계 센서 오류로 인한 측정값으로 예상되는 데이터 구간을 아래 [그림2]와 같이 구간을 선정하고 이를 outlier로 제거하였다.





[그림4] 이상치 구간 설정

그 결과, 아래 [그림5]와 같이 outlier가 제거된 데이터 셋을 얻을 수 있었다.



[그림5] 이상치 제거 후 모습

- **데이터 스케일링**

각 데이터들의 스케일을 0과 1 사이의 값으로 변환하기 위해 min-max scaling을 사용하였다.

$$\frac{value - \min}{\max - \min}$$

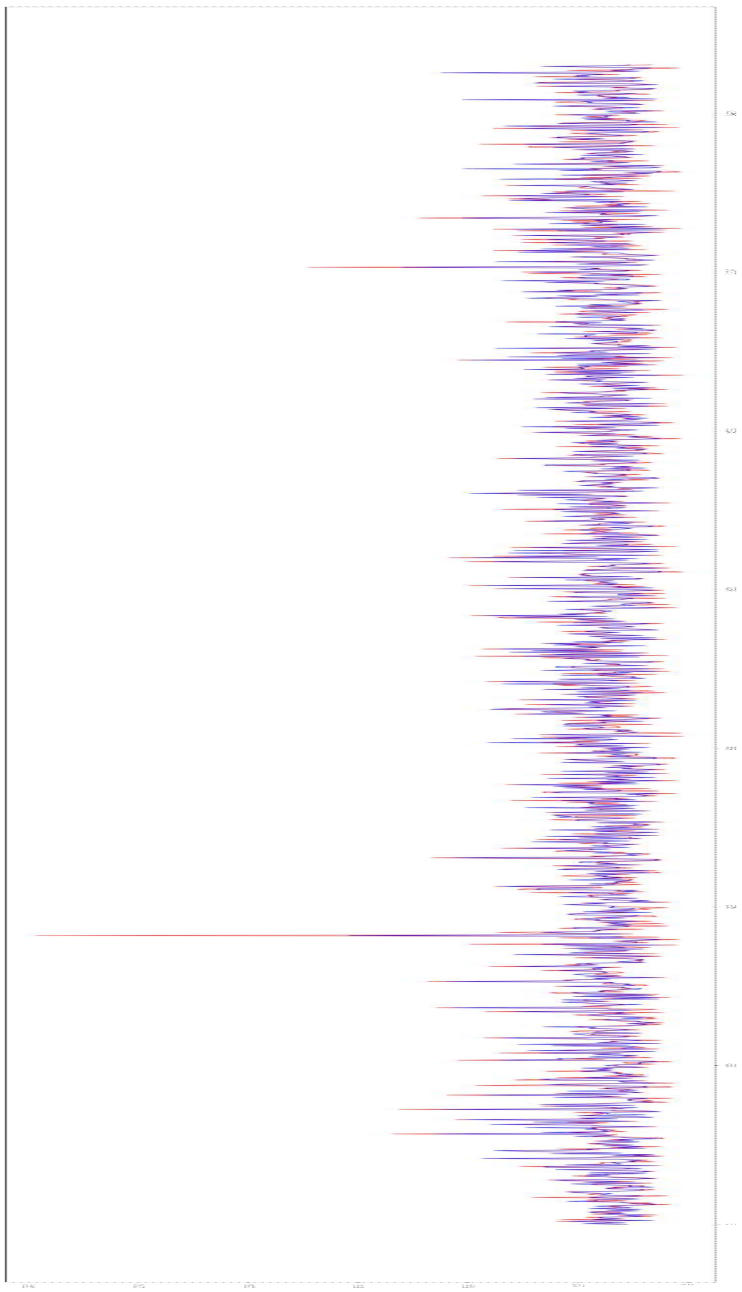
• PM_{10} 인자만 사용한 심층 신경망 모델

우선, 대기와 기상 인자를 사용하지 않고 PM_{10} 데이터만을 사용해서 한 시간 후의 PM_{10}을 예측하는 모델을 만들었다. Train, Validation, Test Set은 6:2:2의 비율로 나누었다. 하이퍼 파라미터는 우리 팀에서 임의의 값들을 선정하여 모델 테스트 셋에 대해 결정계수(R^2) 값을 비교하여 가장 높은 값을 선정하였다.

구분	값
Training Data	60%
Validation Data	20%
Testing Data	20%
Optimizer	Adam
Activation	relu
Learning rate	0.001
Epochs	100
Loss function	mse
Callback method	Ealry Stopping(patience=10)

[표3] 하이퍼 파라미터 설정

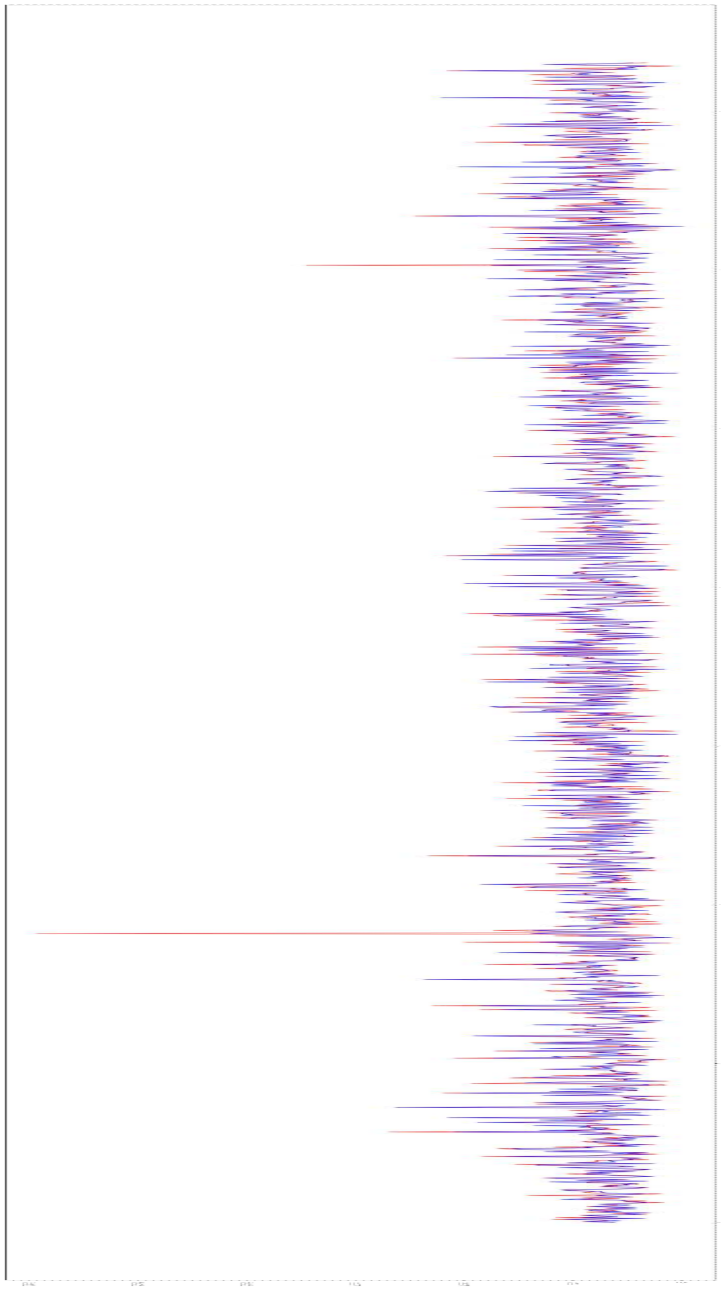
테스트 셋에 대해 예측한 결과, 약 76%의 정확도를 나타내었다.



[그림6] PM_{10} 모델의 PM10 예측

• **PM_{10} 인자 + 대기 데이터를 사용한 심층 신경망 모델**

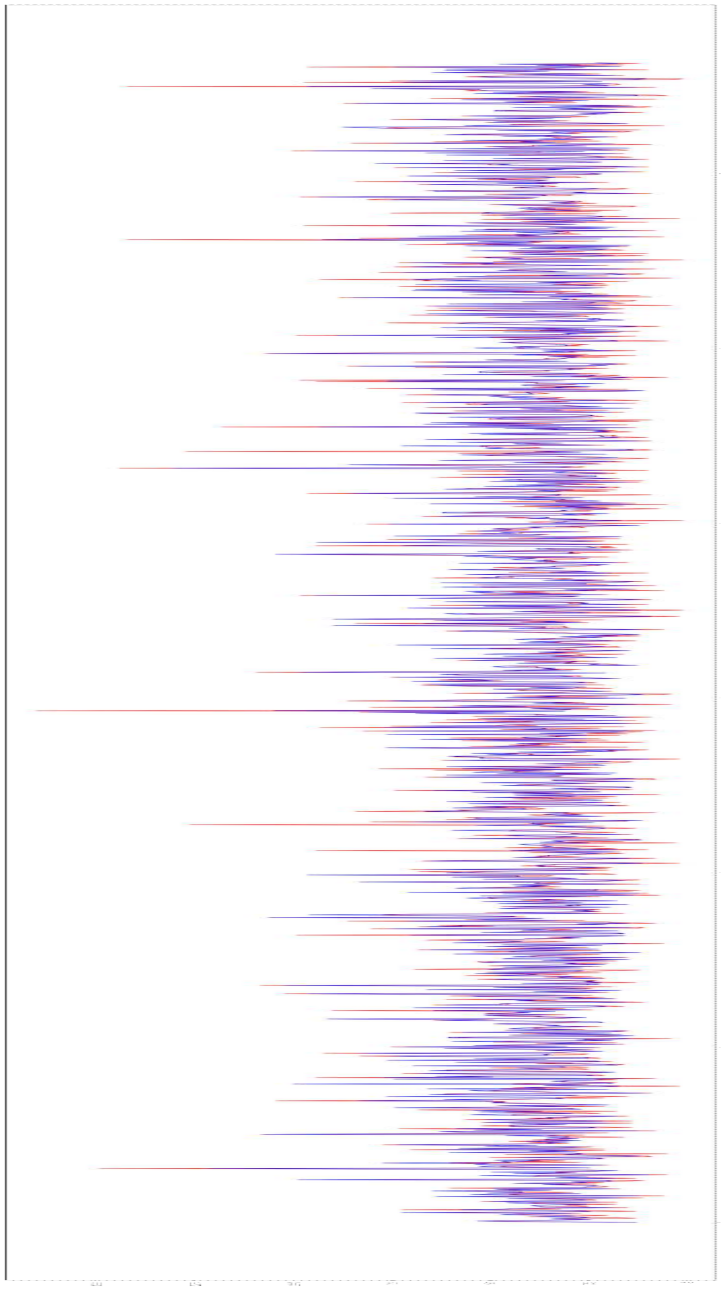
위의 모델에 대기 데이터 SO2, O3, NO2, CO 인자를 추가하여 한 시간 후의 PM_{10}을 예측하는 모델을 만들었다. 동일한 하이퍼 파라미터를 사용하여 테스트 셋에 대해 예측한 결과, 정확도가 약 74%로 PM_{10} 인자만을 사용했을 때에 비해 오히려 정확도가 떨어졌다.



[그림7] PM_{10} + 대기 데이터 모델의 PM_{10} 예측

• **PM_{10} 인자 + 대기 데이터 + 기상 데이터를 사용한 심층 신경망 모델**

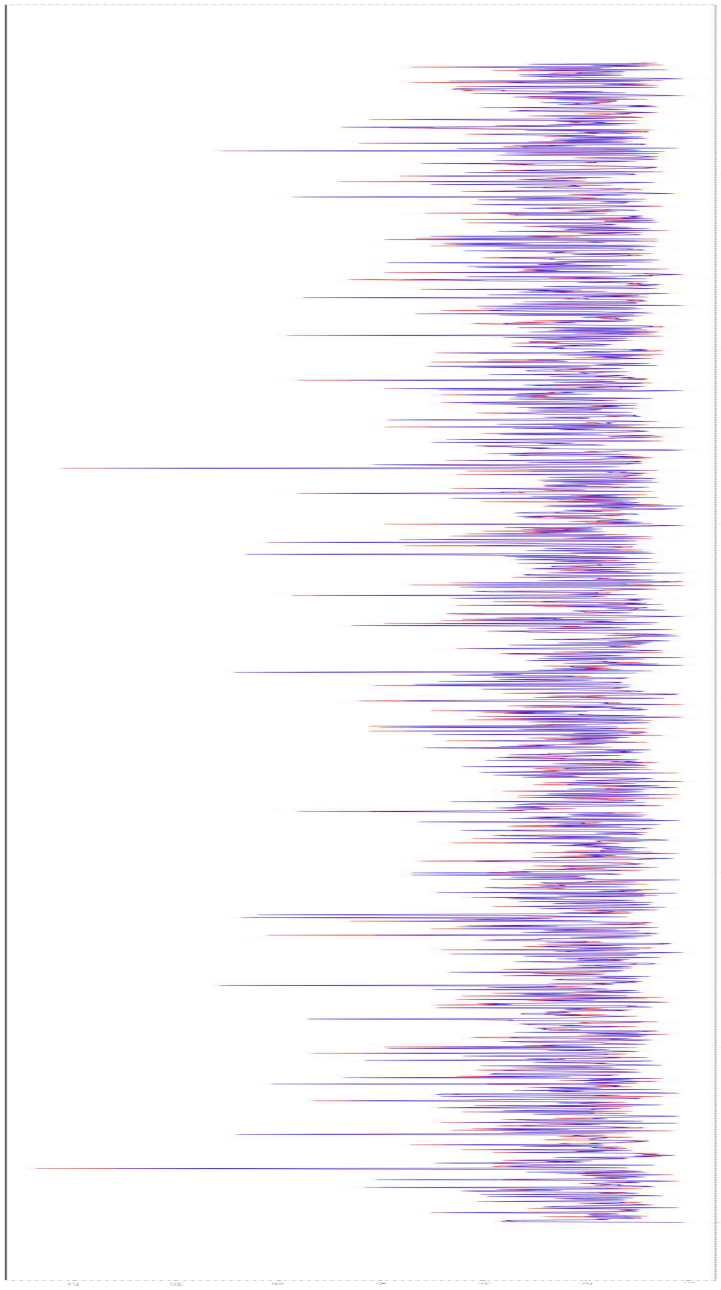
대기 데이터를 포함한 기상 데이터 기온, 강수량, 풍속, 습도 인자를 추가하여 한 시간 후의 PM_{10}을 예측하는 모델을 만들었다. 동일한 하이퍼 파라미터를 사용하여 테스트 셋에 대해 예측한 결과, 정확도가 약 78%로 PM10 인자만을 사용하였을 때와 비슷한 정확도가 나왔다.



[그림8] PM_{10} + 대기 데이터 + 기상 데이터 모델의 PM_{10} 예측

• $PM_{2.5}$ 인자 + 대기 데이터 + 기상 데이터를 사용한 심층 신경망 모델

PM_{10} 예측 모델뿐만 아니라 $PM_{2.5}$(초미세먼지)의 한 시간 후 농도를 예측하는 모델을 위와 동일한 과정을 통해 만들었다. 테스트 셋에 대해 예측한 결과, 정확도가 약 89%로 PM_{10} 예측에 비해 매우 높게 나왔다. 이는 PM_{10}에 비해 $PM_{2.5}$의 고농도가 적고, 저농도 → 고농도, 고농도 → 저농도로 변화가 크게 일어나는 경우가 적어 나타난 결과로 보인다.



[그림9] $PM_{2.5}$ + 대기 데이터 + 기상 데이터 모델의 $PM_{2.5}$ 예측

• **결과**

딥러닝 DNN 모델을 활용하여 한 시간 후의 PM_{10}, $PM_{2.5}$ 미세먼지 농도를 예측하는 것은 전체적으로 정확도 약 75% 이상을 나타내었다. 대체로 미세먼지 농도가 낮은 경우 오차가 적은 값을 예측해냈으나, 고농도로 급변하는 경우에는 오차가 크게 나타났다.

Ⅲ. PBL결과 활용계획

1. PBL결과 활용계획

• **추후 개발의 기초 자료로 활용**

(1) 현재 한 시간 후만을 대상으로 미세먼지 예측을 했지만 이는 실용성이 크게 없을 것으로 예상된다. 따라서 1시간 후뿐만이 아니라, 2시간·3시간, 나아가 최대 6시간 정도의 미세먼지 농도를 예측할 때 기초 자료로 사용하고자 한다.

(2) 현재 신암동 측정소를 대상으로만 모델을 개발하였다. 따라서 신암동뿐만이 아니라 다른 측정소에도 이를 기반으로 모델을 개발하여 각각의 측정소 위치에 맞는 모델을 만드는 것에 대한 기초 자료로 사용하고자 한다.

(3) 미세먼지 농도에 영향을 끼치는 인자로 대기, 기상인자뿐만 아니라 자동차 통행량, 발전소 혹은 공장 가동 여부, 화재 등이 있다. 이러한 인자들을 추가하여 좀 더 넓은 범위(시간적·공간적)를 예측할 수 있는 모델을 개발하는데 기초 자료로 사용하고자 한다.

• **웹 / 앱 어플리케이션 개발**

OpenAPI를 활용해서 실시간으로 인자들에 대한 정보를 수집하고, 모델을 통해 계속해서 학습하며 미세먼지 농도 예측에 대한 결과를 사용자들에게 알려주는 웹 또는 앱 어플리케이션을 개발하는 것에 사용하고자 한다.

Ⅳ. 참고문헌

[학위 논문]

안지민. "대구지역 PM10과 PM2.5 농도 특성 및 기여인자 분석 연구." 국내석사학위논문 계명대학교 대학원, 2018. 대구

이기혁. "복합 신경망 구조를 이용한 미세먼지 위험 단계 예측 모델 설계 및 분석." 국내석사학위논문 한양대학교 대학원, 2020. 서울

조경우. "미세먼지 예측을 위한 딥러닝 기반 농도별 분리 예측 모델." 국내박사학위논문 한국기술교육대학교 일반대학원, 2020. 충청남도

채상미. "학교 미세먼지 대응 위한 RNN-LSTM 순환 신경망 미세먼지 예측." 국내석사학위논문 아주대학교 대학원, 2020. 경기도

[학술지]

성상하,김상진,and 류민호. "미세먼지 예측을 위한 기계학습 모델 간 성능 비교 연구:국내 발생 데이터를 중심으로." 한국혁신학회지 15.4 (2020): 339-357.

전영태,유숙현,and 권희용. "Outlier 데이터 제거를 통한 미세먼지 예보성능의 향상." 멀티미디어학회논문지 23.6 (2020): 747-755.

조경우,정용진,강철규,오창헌,Cho Kyoung-woo,Jung Yong-jin,Kang Chul-gyu,and Oh Chang-heon. "미세먼지 예측을 위한 기계 학습 알고리즘의 적합성 평가." 한국정보통신학회논문지 23.1 (2019): 20-26.