Airlyze

미세먼지 농도 예측 모델

김태영  
 Department of Software  
 성균관대학교  
 수원, 대한민국  
219tyk1234@gmail.com

하상수  
 Department of Software  
 성균관대학교  
 수원, 대한민국  
 [hsang325@gmail.com](mailto:email@email.com)

신재환  
 Department of Software  
 성균관대학교  
 수원, 대한민국  
 terry090222@gmail.com

한준호  
 Department of Software  
 성균관대학교  
 수원, 대한민국  
 hjunho2357@gmail.com

유재룡  
 Department of Software  
성균관대학교  
 수원, 대한민국  
 ujh1013@gmail.com

ABSTRACT

미세먼지(PM10)는 환경적,보건적 문제로 대두되고 있으며, 특히 국외 기상 조건이 미세먼지 농도에 미치는 영향에 대한 관심이 높아지고 있다. 이 프로젝트에서는 한국의 미세먼지 농도 예측 정확도 향상을 목표로, 기상청에서 제공하는 5개년치 기상 데이터를 활용하여 예측 모델을 개발하였다. 특히 중국발 미세먼지의 영향이라는 가설에 주목하여 공간 정보 활용의 중요성을 인지하고, 이를 반영하기 위해 1D CNN 과 LSTM 을 결합한 1DCNN-LSTM 모델을 설계하였다.

MLP 모델을 baseline model 로 설정하여 , 다양한 기상요인들의 중요도를 평가하였으며, LSTM 모델과 1DCNN-LSTM 모델을 설계하여 예측을 진행, MAE를 통해 그 정확도를 측정하였다.

그 결과 1DCNN-LSTM 모델은 기존의 시계열 분석 모델에 비해 미세먼지 농도 예측 성능 향상을 일부 보였으나, 당초 기대했던 만큼의 극적인 성능 향상은 달성하지 못하였다. 이는 모델 구조의 최적화 부족, 추가적인 공간 정보 변수의 부재, 또는 중국발 미세먼지 영향의 복잡성 등 다양한 요인에 기인 할 수 있다.

본 프로젝트는 미세먼지 예측에 있어 공간 정보 활용의 가능성을 탐색했다는 점에서 의의를 가지며, 향후 진행에서는 모델 구조 개선, 추가적인 공간 정보 변수의 활용 , 그리고 더 나은 데이터를 활용하여 예측 성능을 더욱 향상시킬 수 있을 것으로 기대한다.

1 Introduction

미세먼지(PM10)는 세계적으로 중요한 환경 문제이자 인류 건강에 심각한 영향을 미치는 요인 중 하나로 주목받고 있다. 특히, 대한민국은 계절적 요인 뿐 아니라 국외에서 유입되는 미세먼지의 영향이 복합적으로 작용하여 그 문제가 더욱 심화 되고 있다. 중국의 산업 활동과 기상 변화는 국경을 넘어 국내 대기 질에 직접적인 영향을 끼치는 주요 원인으로 지적되며, 이에 대한 정확한 예측은 정책적 대응과 국민 건강 보호에 필수적이다.

기존의 시간적 패턴만을 고려한 통계적 기법 또는 단일 딥러닝 모델은 국외 기상 조건과 같은 공간적 특성을 효과적으로 반영하지 못하는 한계를 가진다. 따라서 시공간적 특성을 동시에 고려할 수 있는 새로운 접근법이 필요하다.

본 과제에서는 한국 기상청에서 제공하는 기상 데이터와 중국의 기상 데이터를 결합하여 PM10 농도를 예측하는 모델을 설계하였다. 시간적 특성만을 학습하는 LSTM 모델과 더불어, 1차원 합성곱 신경망(1DCNN)을 통해 공간적 특성을 추출한 후 LSTM 과 결합하는1DCNN-LSTM 모델로 예측을 진행하였다. 이를 통해 국외 기상 조건이 미세먼지 농도에 미치는 영향을 분석하고, 기존 예측 모델과의 성능 비교를 수행한다.

본 과제의 기여점은 다음과 같다. 1. 국외 기상 데이터를 결합한 미세먼지 예측 모델 설계 및 시도. 2. 시간적, 공간적 특성을 동시에 학습하는 1DCNN-LSTM 하이브리드 모델 개발 . 3. 기존 MLP 및 LSTM 모델과의 비교를 통해 새로운 모델의 성능과 한계를 평가..

2 Related Work

[[1]](#endnote-2)시공간 데이터를 활용한 연구인 Air quality prediction using spatio-temporal deep learning 을 참고 하였다. 해당 논문에서는 공간적 데이터와 시간적 데이터의 결합을 위해

GAT를 통해 그래프 데이터에서 공간적 데이터를 학습, BiLSTM 을 통해 시간적 데이터를 학습하였다.

3 Methodology

본 과제에서 사용한 데이터셋과 모델 설계 과정을 다음과 같이 설명한다.

3.1 Dataset

한국 기상청 데이터:

2018~2022년 총 5개년의 데이터를 수집하였으며, 서울, 안동, 대구 ,광주, 전주 5개 지역의 데이터를 사용하였다. 수집한 변수는 다음과 같다.

|  |  |
| --- | --- |
| 시간 변수 | TM(시간 정보) |
| 지역 변수 | STN(관측소 번호) |
| 기상 변수 | WD(풍향), WS(풍속), PA(현지 기압), VV(시정 거리)  PS(해면 기압), TA(기온), TD(이슬점 온도), etc.. |
| 환경 변수 | CA\_TOP(전체 운량), CA\_MID(중층 운량),  CT(운형), VS(시정), TS(지면 온도),  TE\_005, 01, 02, 03(지하 온도, cm단위),  IR(지면 적외선 온도), etc.. |
| 목표 변수 | PM (미세먼지 농도, μg/m³ 단위) |

중국 기상 데이터:

국외 기상 조건이 국내 미세먼지에 미치는 영향을 반영하기 위해 중국의 기온 , 기압 , 풍속 등 관련 기상데이터를 추가하였다. 2018~2022 5개년의 데이터로 Yanan, Tongliao, Qingdao, Chifeng, Dalian 5개 지역의 데이터를 사용하였다. 변수는 아래와 같다.

|  |  |
| --- | --- |
| 시간 변수 | TM(시간 정보) |
| 지역 변수 | STN(관측소 번호) |
| 기상 변수 | WD(풍향), WS(풍속), PA(현지 기압), VV(시정 거리)  PS(해면 기압), TA(기온), TD(이슬점 온도), etc... |
| 환경 변수 | CA\_TOP(전체 운량), CA\_MID(중층 운량),  CT(운형), VS(시정), TS(지면 온도),  IR(지면 적외선 온도), etc.. |
| 목표 변수 | PM (미세먼지 농도, μg/m³ 단위) |

데이터 전처리: 한국 데이터는 1시간 간격의 데이터, 중국 데이터는 3시간 간격에 데이터인 관계로, 1시간 간격의 데이터로 정리하기 위해 선형보간법으로 결측치를 처리 하였다.

3.2 Model design

모델은 MLP, LSTM, 1DCNN-LSTM 모델을 설계하였다.

다양한 환경 변수가 PM10에 영향을 미친다고 예측하였기 때문에 다양한 변수를 반영가능한 MLP 모델을 사용하였으며 , 이를 활용해, Permutation Importance를 측정하였다.

시계열 데이터 처리를 위해 LSTM 모델을 사용해 과거 정보를 학습하여 예측을 수행하였으며, 공간정보 활용을 위해 1DCNN-LSTM 을 설계하였다.

**3.2.1 MLP(Baseline 모델).**

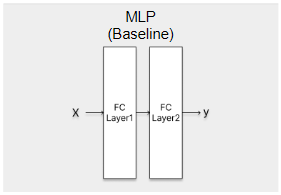


Figure 1: MLP(Baseline Model)

* **구조**: 입력 데이터를 완전 연결층(FC Layer)으로 전달.
  + Input → FC Layer(64) → FC Layer(32) → Output (PM10)
* 하이퍼파라미터:
  + 옵티마이저: Adam
  + 손실 함수: MSE (Mean Squared Error)
  + 성능 평가 지표: MAE (Mean Absolute Error)

MLP 모델은 단순한 구조로 미세먼지 예측의 기준 성능을 제공하며, 복잡한 시간 의존성을 반영하지 않는다.

**3.2.2 LSTM(Proposed 모델1).**

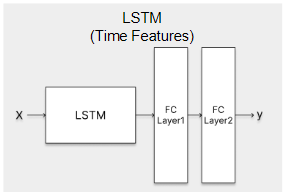


Figure 2: LSTM(Proposed Model)

* **구조**: 시간 순서에 따라 데이터의 과거 정보를 학습하는 시계열 모델.
  + Input → LSTM Layer(128 units) → Dropout(0.3) → FC Layer(32) → Output (PM10)
* 하이퍼파라미터:
  + 옵티마이저: Adam
  + 손실 함수: MSE
  + 드롭아웃: 0.3 (과적합 방지)

LSTM 모델은 시간 순서의 기상 데이터를 기반으로 미세먼지 농도를 예측하는 데 강점을 가지며, 기압, 풍향 등 시간적 의존성을 학습한다.

**3.2.3 1DCNN-LSTM(Proposed 모델2)**

텍스트, 도표, 스크린샷, 평행이(가) 표시된 사진

자동 생성된 설명

Figure 3: 1DCNN-LSTM(Proposed Model)

1DCNN-LSTM 모델은 기상 데이터의 시간적 특성과 공간적 특성을 동시에 반영하기 위해 설계되었다. 이 모델은 1차원 합성곱 신경망(1DCNN)을 활용하여 데이터의 공간적 특징을 추출하고, 이를 LSTM 레이어를 통해 시간적 의존성을 학습한다.

모델구조:

공간 정보 활용을 위해 입력 데이터로 중국의 지역별 PM10 데이터를 추가 하였다. 이후 K개의 Conv1D 연산을 통해 각 도시의 PM10 데이터로부터 공간적 특성을 추출 하였으며, 이러한 특성은 Concatenate Layer 에서 공간적 특성(SpatialFeatures)과 시간적 특성(Weather Features)을 결합하여 하나의 통합 특성 벡터로 변환된다. 이렇게 결합된 특성 벡터를 LSTM Layer 에 입력하여 시간 순서에 따른 기상 데이터의 의존성을 학습한다. 최종 출력값은 완전연결층(FC Layer)을 통과하여 PM10 예측값으로 반환 된다.

* 하이퍼파라미터:
  + 옵티마이저: Adam
  + 손실 함수: MSE
  + 드롭아웃: 0.2, 0.3, 0.4 (과적합 방지)(지역에 따라 다름)

4 Results

4.1 Permutation Importance

MLP 모델을 활용해 Important Feature 를 확인한 결과 상위 5 개의 Feature가 다음과 같이 선정되었다.

**안동**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 248.7513 |
| PS | 209.1231 |
| HM | 25.3198 |
| PA | 20.6332 |
| TA | 6.1239 |

**안동(중국 데이터 포함)**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 0.00928 |
| WD\_chifeng | 5.13473 |
| PS | 5.08325 |
| WS\_yanan | 4.91112 |
| TD\_qingdao | 3.78533 |

**대구**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 253.05799 |
| PA | 144.3779 |
| PS | 104.3526 |
| TA | 8.99368 |
| STN | 0.0 |

**대구(중국 데이터 포함)**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 20.5175 |
| PA | 0.67477 |
| TA | 0.14396 |
| WD\_chifeng | 0.13376 |
| TA\_yanan | 0.06397 |

**광주**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 276.8868 |
| PA | 124.6233 |
| PS | 122.3693 |
| HM | 42.3397 |
| TS | 3.5418 |

**광주(중국 데이터 포함)**

|  |  |
| --- | --- |
| Feature | Importance |
| PR\_qingdao | 0.000140 |
| TS | 0.000135 |
| PS\_qingdao | 0.000134 |
| PS\_tongliao | 0.0001332 |
| HM\_yanan | 0.0001330 |

**전주**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 424.6587 |
| PA | 210.5079 |
| PS | 171.9294 |
| HM | 64.1256 |
| WD | 1.15517 |

**전주(중국 데이터 포함)**

|  |  |
| --- | --- |
| Feature | Importance |
| IX | 0.00012 |
| VV\_chifeng | 0.00010 |
| VV\_dalian | 6.60388 |
| PT\_dalian | 6.01990 |
| PM10\_yanan | 5.97745 |

**서울**

|  |  |
| --- | --- |
| Feature | Importance |
| VS | 393.0197 |
| PS | 358.2524 |
| PA | 32.47354 |
| HM | 24.64650 |
| TE\_005 | 7.94076 |

**서울(중국 데이터 포함)**

|  |  |
| --- | --- |
| Feature | Importance |
| TA | 0.0003276 |
| TS | 0.0002916 |
| TD\_chifeng | 0.0002908 |
| PS\_qingdao | 0.0002792 |
| TA\_chifeng | 0.0002788 |

전체적으로 풍향과 기압이 중요한 변수로 나타났음을 확인 할 수 있다. 풍향과 기압의 영향은 시간적 변화에 따라 달라질 가능성이 크기 때문에 시계열 모델이 중요할 것이라 예측 가능하였다.

4.2 LSTM

|  |  |  |
| --- | --- | --- |
| 지역 | 한국 | 한국 + 중국 |
| 서울 | 18.8039 | 25.3115 |
| 전주 | 16.6403 | 17.1193 |
| 광주 | 19.2508 | 47.2586 |
| 대구 | 10.8149 | 15.4643 |
| 안동 | 8.9762 | 16.7488 |

**서울**

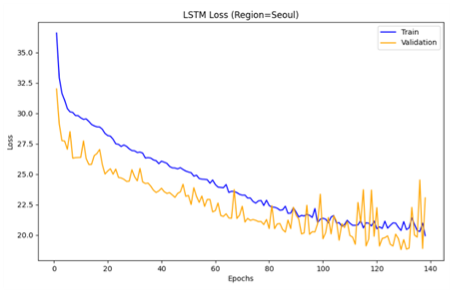
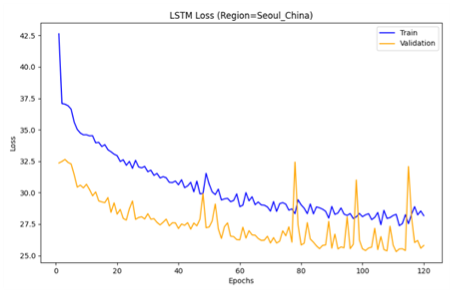


Figure 4: Loss Graph ( Left = 서울, Right = 서울+중국)

**안동**

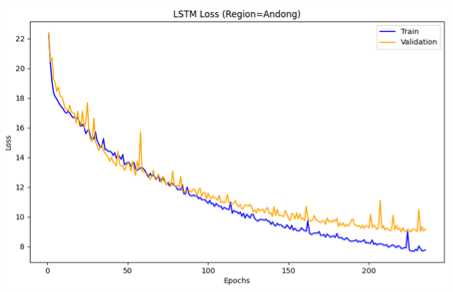
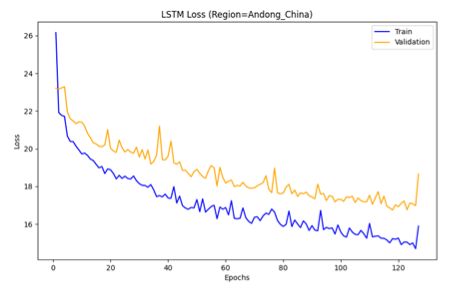


Figure 5: Loss Graph ( Left = 안동, Right = 안동+중국)

**광주**

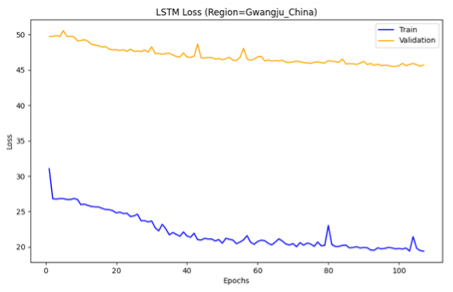
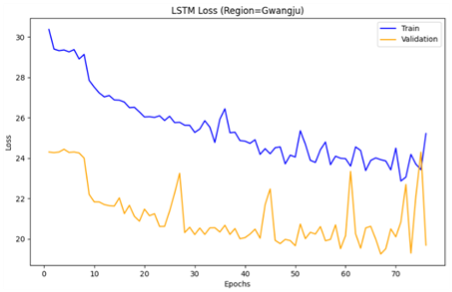


Figure 6: Loss Graph ( Left =광주, Right = 광주 +중국)]

**전주**

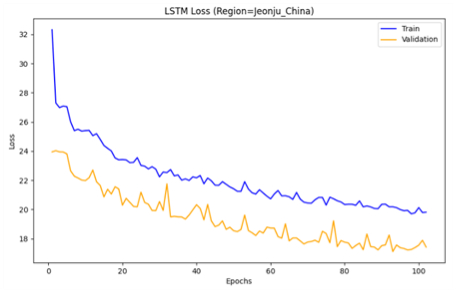
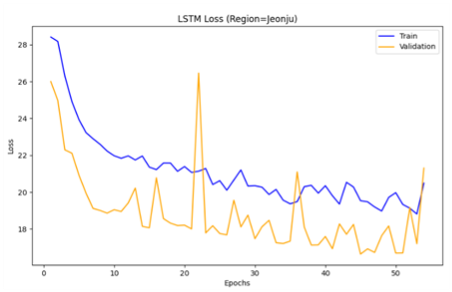


Figure 7: Loss Graph ( Left = 전주, Right = 전주 +중국)

**대구**

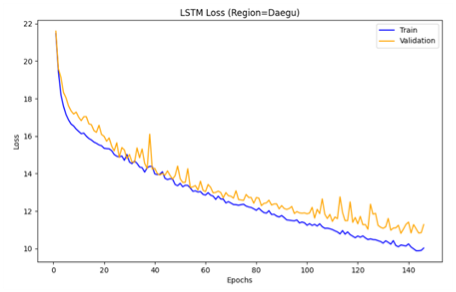
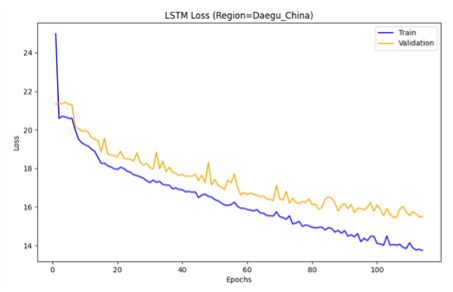


Figure 8: Loss Graph ( Left = 대구, Right = 대구+중국)

중국 데이터를 포함한 결과, 모든 지역에서 loss 가 증가하는 현상을 보였다.이를 통해 단순한 시점 정보 뿐 아니라 추가적인 feature가 필요함을 확인하였으며, 공간 정보를 활용하기로 결정하였다.

4.3 1DCNN-LSTM

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K | 서울 | 전주 | 광주 | 대구 | 안동 |
| 1 | 25.8446 | 16.6978 | 26.3773 | 14.6803 | 13.5553 |
| 2 | 25.6843 | 15.4528 | 49.6924 | 14.4918 | 14.8014 |
| 3 | 32.3613 | 23.9469 | 49.6911 | 18.5006 | 23.1646 |

K의 값을 1,2,3 으로 설정하여 진행하였으며, 아래 그래프는 k=2의 결과이다.

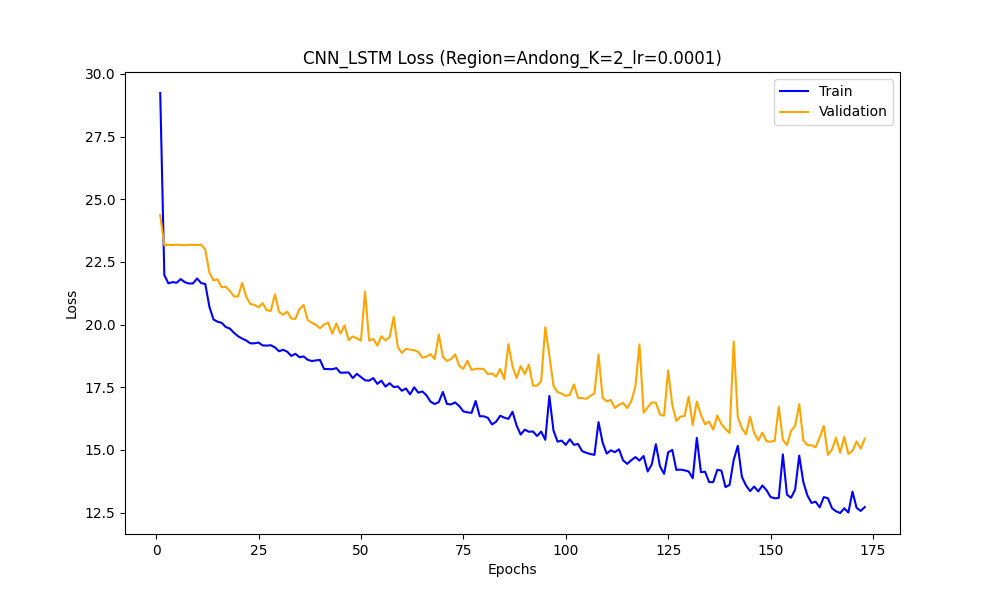


Figure 9: Loss 안동 for k =2

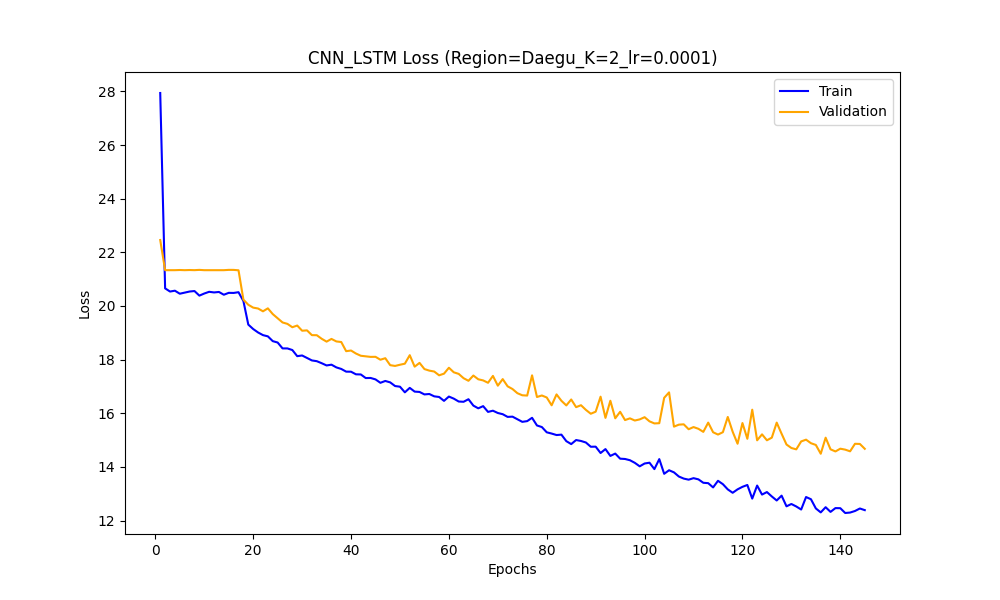


Figure 10: Loss 대구 for k =2

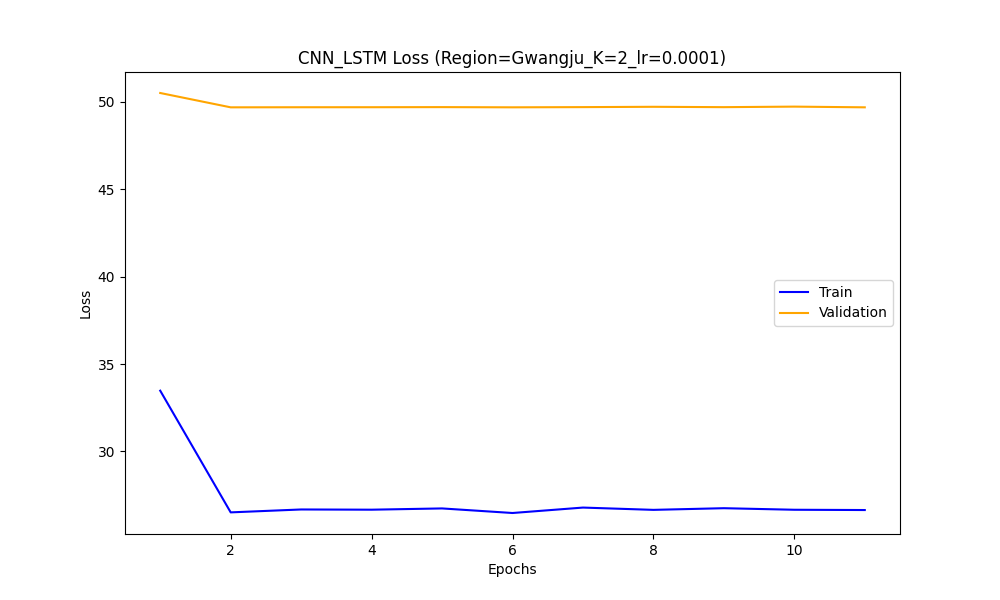


Figure 11: Loss 광주 for k =2

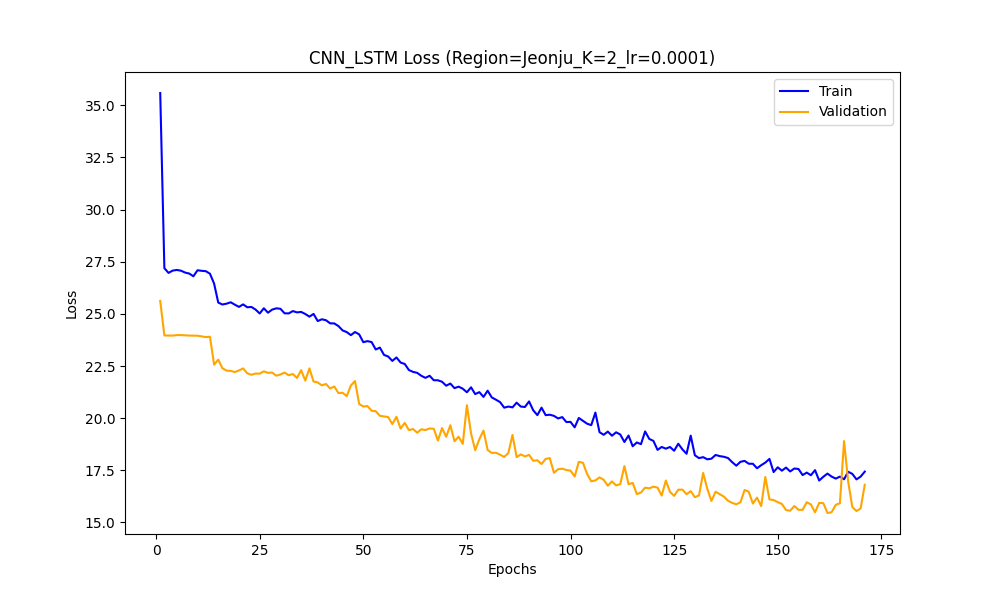


Figure 12: Loss 전주 for k =2

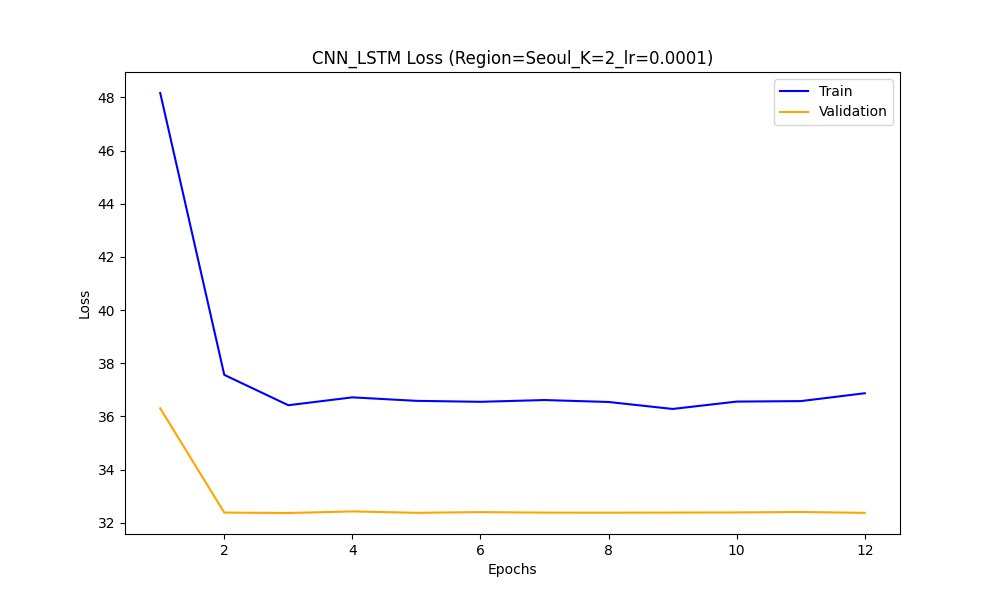
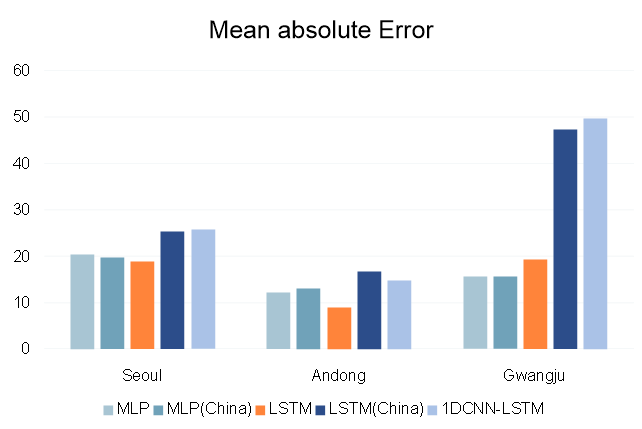


Figure 13: Loss 서울 for k =2

일반적으로 k=2 경우 성능향상에 도움이 되었으나 , 하이퍼 파라미터에 따라 성능에 큰 변동이 있었다.



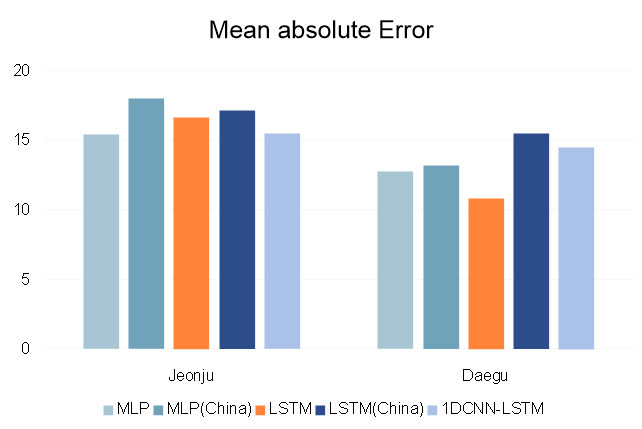


Figure 14: MAE Graph

전체적인 결과는 위와 같다. LSTM 이 대체적으로 가장 좋은 결과를 보여 주었다.

5 Conclusion and Future Work

본 과제에서는 미세먼지 농도(PM10) 에측을 위해 다양한 모델(MLP, LSTM, 1DCNN-LSTM)을 설계하고 이들의 성능을 비교하였다. 결과적으로, 풍향과 기압과 같은 특정 기상 변수가 예측 성능에 유의미한 영향을 미치는 것으로 나타났으며, 이를 반영한 학습에서 예측 정확도가 향상되었음을 확인하였다. 또한 시간적 패턴을 반영하는 LSTM 모델이 단순한 MLP 모델보다 더 나은 성능을 보였으나, 중국의 기상 데이터를 포함한 1DCNN-LSTM 모델은 성능 개선 효과를 충분히 달성하지 못했다. 특히 중국 측 미세먼지 데이터는 오히려 예측 성능 저하를 초래했으며, 이는 데이터 품질 및 모델 일반화 문제와 관련이 있을 가능성을 뜻한다.

이 과제는 국외 데이터를 통한 미세먼지 예측 시도의 의의를 가지지만 , 데이터의 품질 , 양, 그리고 특성 반영에 대한 한계를 드러내었다.

5.1 Limitations

본 과제에서 직면한 주요 한계점은 크게 데이터 품질, 모델의 기술적 제약, 그리고 정보의 부족이라는 세 가지 측면에서 나타났다. 우선 데이터 품질 측면에서, 중국에서 수집된 데이터의 신뢰성과 정확성이 충분히 확보되지 못했다는 점이 가장 큰 문제점으로 지적되었다. 또한 실제 활용 가능한 데이터의 양이 제한적이었으며, 시간적, 공간적 다양성을 충분히 반영하지 못하는 한계를 보였다.

모델 측면에서는 본 연구에서 활용된 1DCNN-LSTM 모델이 미세먼지 발생에 영향을 미치는 다양한 변수들의 중요도를 적절하게 반영하지 못했다는 기술적 한계가 존재했다. 이는 모델의 예측 정확도와 신뢰성에 직접적인 영향을 미치는 요소로 작용했다.

마지막으로, 미세먼지의 초기 발생 근원지와 같은 구체적인 원인 데이터가 부족했다는 점도 주요한 한계점으로 나타났다. 이러한 정보의 부재는 미세먼지 발생 메커니즘을 보다 정확하게 이해하고 예측하는 데 있어 중요한 제약 요소로 작용했다.

5.2 Future Work

중국 데이터를 포함하여 고품질의 일관된 데이터셋을 확보하며, Attention 메커니즘을 통해 중요한 기상변수를 더욱 효과적으로 학습한다. 다른 기계학습 모델(Transformer 기반 모델) 과의 비교 연구를 진행가능 하며, 결측값 처리에 있어 기상변수 각각의 특성을 고려하여 다양한 보완법을 적용가능 할 것이다.

5.3 Github Link

REFERENCES

[1] Zhang, Y., & Wang, S. (2022). A novel approach for PM2.5 concentration prediction based on VMD-GAT-BiLSTM model. \*Atmospheric Pollution Research\*, 13(4), 101224. https://doi.org/10.1016/j.apr.2022.101224

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00

1. 참조 [↑](#endnote-ref-2)