碩士學位 請求論文 (표지)

指導敎授 辛 志 太

딥러닝 기반 국내 미세먼지 예측 모델링 연구

成均館大學校 一般大學院

휴먼 ICT 融合學科

李 聖 求

碩士學位 請求論文 (내표지)

指導敎授 辛 志 太

딥러닝 기반 국내 미세먼지 예측 모델링 연구

Research of Particulate Matter Prediction Modeling Based on Deep Learning

成均館大學校 一般大學院

휴먼 ICT 融合學科

李 聖 求

碩士學位 請求論文 (심사청구서)

指導敎授 辛 志 太

딥러닝 기반 국내 미세먼지 예측 모델링 연구

Research of Particulate Matter Prediction Modeling Based on Deep Learning

이 論文을 工學 碩士學位請求論文으로 提出합니다.

2019 年 4 月 日

成均館大學校一般大學院

휴먼ICT 融合學科

李聖求

(인정서)

이 論文을 李聖求의 工學
碩士學位 論文으로 認定함.

2019 年 6 月 日

審査委員長 _____

審査委員 _____

審査委員 _____

목차

표목차

그림목차

논문요약

딥러닝 기반 국내 미세먼지 예측 모델링 연구

미세먼지란 환경정책기본법에 따르면, 입자의 크기가 10um 이하인 먼지(PM-10)과 2.5um 이하인 먼지(PM-2.5)를 통칭하며, 사람의 생활 영역에 다양한 피해를 끼치고 있다.

이러한 미세먼지의 위험성에 대해 최근 시민과 정부의 인식이 높아지고 있으며, 미세먼지로 인한 피해를 예방하기 위한 예측 시스템의 중요성이 대두되고 있다. 그러나 환경부에서 국내 미세먼지 예측을 위해 쓰고 있는 모델은 결정론적 기법을 사용하고 있으며, 이 기법은 최근 딥러닝 기반 기법보다 성능이 떨어진다고 알려져 있다. 그럼에도 불구하고 국내환경에 맞추어 연구된 딥러닝 기반 미세먼지 예측 모델링의 연구는 국제 연구에 비해 아직 미흡한 실정이다.

따라서 본 연구에서는 국내 환경을 고려하여 중국 미세먼지 데이터와 국내 기상데이터 및 미세먼지 데이터를 활용한 딥러닝 모델 개발을 목표로 한다. 또한 기존에 연구된 딥러닝 기반 예측 모델보다 더 좋은 성능을 얻기 위해, 시공간 데이터를 동시에 분석하여 미세먼지 확산현상을 더 잘 고려할 수 있는 ConvGRU와, CNN과 달리 weight sharing을 하지 않아 중국에서 유입되는 미세먼지가 각 지역에 주는 영향을 더 잘 반영할 수 있는 locally-connected layer를 활용한 딥러닝 기반 국내 미세먼지 예측 모델링을 제안한다.

실험은 국내 전역을 8x10의 grid로 나눈 구역을 공간해상도로 하고, 1시간을 시간해상도로 하여 1시간, 4시간, 12시간, 24시간 뒤 미세먼지(PM10)농도를 예측하도록 설계한다. 제안한 모델에서 성능을 높이기 위해 가정한 부분들을 검증하고, 기존에 연구된 딥러닝 기반 미세먼지 예측 모델보다 더 나은 성능을 보이는지 확인하기 위해 5가지의 실험 가설을 세우고 이에 맞는 다양한 비교 모델을 만들어 결과를 분석하였다.

실험결과에 따른 검증된 가설을 종합하면 미세먼지를 예측하기 위한

모델링은 1) 시간정보뿐만 아니라 시공간정보를 동시에 고려할 때, 2) 단시간 예측에 대해서는 계산 복잡성이 낮은 기법을, 장시간 예측에 대해서는 복잡성이 높은 기법을 사용할 때, 3) 다음 시간 T를 예측하기 위해 다음 시간 T−1까지의 중간 과정을 고려할 때, 4) 국내의 미세먼지 확산 요인을 잘 반영하도록 설계할 때, 5) 국외 미세먼지 유입요인 또한 잘 반영하도록 설계할 때 좋은 미세먼지 예측 성능을 얻을 수 있다는 결론을 얻을 수 있으며, 제안한 모델이 이를 만족함으로써 기존에 연구된 모델보다 더 나은 예측 성능을 가짐을 보였다.

또한 단시간 예측 시 주로 이전 참값을 따라가는 delay shift 현상을, 장시간 예측 시 중간 값을 따라가려 하는 moving average 현상이 일어남을 예측 값과 참값의 비교 그래프를 통해 확인하였고, 이를 해결한다면 예측 성능을 더 개선할 수 있겠다는 결론을 얻었다.

주제어 : 미세먼지, 딥러닝, 시계열 예측, ConvGRU, Locally−connected layer

제1장 서 론

미세먼지란 환경정책기본법에 따르면, 입자의 크기가 10um 이하인 먼지(PM-10)과 2.5um 이하인 먼지(PM-2.5)를 통칭한다. 이 미세먼지는 눈에 보이지 않을 정도로 매우 작기 때문에 공기 중에 머물러 있다가 우리가 호흡할 때 호흡기를 거쳐 폐 등에 침투하거나 혈관을 따라 체내로 이동해 들어감으로써 건강에 나쁜 영향을 미친다. 세계보건기구(WHO)는 미세먼지에 대한 대기질 가이드라인을 1987년부터 제시해 왔고, 2013년에는 WHO 산하 국제암연구소(IARC)에서 미세먼지를 사람에게 발암이 확인된 1군 발암물질로 지정해왔다[1]. 또한, 미세먼지는 사람의 건강뿐만 아니라 인간의 다른 생활 영역에서도 피해를 미친다. 미세먼지 농도가 높아지면 시정이 좁아져 교통사고가 평소보다 더 자주 일어날 수 있다. 미세먼지는 가볍기 때문에 매우 먼 거리를 이동할 수 있으며, 이 미세먼지에 포함된 화학성분은 토양을 산성으로 오염시키며 농작물에 피해를 입히기도 한다[2].

이러한 미세먼지의 위험성에 대해 최근 시민과 정부의 인식이 높아지고 있으며, 미세먼지로 인한 피해를 예방하기 위한 미세먼지 예측 시스템의 중요성이 대두되고 있다. 현재까지 연구되고 있는 미세먼지 예측 기법은 크게 오염원의 물리·화학적 특성을 기반으로 오염원의 농도를 인과관계 식으로 결정하는 결정론적 기법(Deterministic Method)과 기존에 측정된 자료에 통계기법을 적용한 통계적 기법(Statistic Method)이 있으며, 통계적 기법은 다시 딥러닝 기반 기법과 non-딥러닝 기반기법으로 나뉜다. 환경부에서 국내 미세먼지 예측을 위해 쓰고 있는 모델은 미국 환경청(EPA)이 개발한 미세먼지 예측모델(CMAQ)[3]로 이는 결정론적 기법에 해당한다. 그런데, 결정론적 기법의 인과관계 식이 미세먼지가 발생하는 물리화학적 현상을 모두 설명할 수는 없다. 또한, 최근에는 딥러닝 기반 미세먼지 예측 기법이 다른 기법보다 좋은 성능을 낼 수 있다는 연구 결과[4]가

나왔으며, 딥러닝 기반의 미세먼지 예측 기법이 세계적으로 연구되고 있는 추세이다. 그러나 이러한 상황에도 불구하고 국내환경에 맞추어 연구된 딥러닝 기반 미세먼지 예측 모델링의 연구는 아직 미흡한 실정이다.

따라서 본 연구에서는 중국 미세먼지 데이터와 국내 기상데이터(온도, 습도, 바람, 날씨 등) 및 미세먼지 데이터를 이용하여 국내 환경에 맞는 딥러닝 기반 미세먼지 예측 모델링 개발을 목표로 한다. 이를 위해 시공간 데이터를 분석할 수 있는 ConvGRU와 개별 공간의 feature를 더 잘 뽑을 수 있는 Locally-Connected Layer를 활용한다. 제안한 모델의 구조 및 자세한 원리는 본문에서 소개하기로 한다.

제2장 관련연구

2-1. 미세먼지 농도 요인

미세먼지 농도를 예측하기 위해서는 먼저 어떤 요인들이 미세먼지 농도에 영향을 주는지 알아야 할 것이다. 기존 국제 연구의 경우에는 주로 세계적으로 미세먼지 수치가 가장 높은 도시인 중국의 베이징과 상하이를 대상으로 미세먼지 요인을 분석하였다. 분석결과로 다른 지역에서의 미세먼지 유입(베이징 25%, 상하이 20%)과 교통량에 의한 먼지 발생(베이징 22%, 상하이 25%)을 들었으며, 그 외 공장에서 화석연료를 태우거나, 휘발성유기화합물의 화학반응에 의한 2차 생성 등에 의해 미세먼지가 발생하는 것으로 조사되었다.

국내 미세먼지 농도의 영향 요인을 분석한 연구 논문에 따르면, 미세먼지 농도는 이러한 배출원 요인 이외에도 기상과 지형조건도 중요하게 작용한다고 한다[5]. 특히 한국환경정책평가연구원에서 간행한 학술지에 따르면 한국의 지리학적·계절적인 특성으로 인하여 중국에서의 미세먼지 유입 추정 기여율이 39%~53%정도로 다른 요인보다 압도적인 우위를 차지하였다[6].

또한 고농도 미세먼지 발생 시의 기상학적 요인을 분석한 연구에 따르면, 낮은 최저기온을 포함한 큰 일교차와 낮은 풍속이 고농도 발생과 높은 상관관계를 가지는 것으로 조사되었다[7]. 연구에서 이는 큰 일교차로부터 야간에 대기층이 안정화되면서 대기확산이 저하된 점과 낮은 풍속으로 인한 대기정체 때문인 것으로 판단하였다. 그러나, 봄철의 일부 지역에서는 편서풍과 중국 배출원의 영향이 함께 반영되어, 다소 강한 풍속에서도 고농도가 빈번하게 발생한 것으로 조사되었다. 반면에 고농도 현상이 대체로 발생하지 않은 경우는 잦은 강수에 의한 대기세정 효과와 계절적으로 남서풍이 주로 형성되어 중국의 영향이 낮기 때문인 것으로 분석하였다.

따라서 미세먼지 예측 모델링을 만들기 위해서는 일교차, 온도, 풍속과 같은 기상학적 요인과 풍향, 중국의 미세먼지와 같은 외부유입 요인 및 SO_2, NO_2와 같은 화합물에 의한 2차생성 요인 등 다양한 요인들을 모두 반영하여야 좋은 성능을 얻을 수 있다고 판단할 수 있다.

2-2. 미세먼지 예측을 위한 결정론적 기법

결정론적 기법은 역학적 기법(Mechanistic Method)이라고도 하며, 역 오염원의 물리·화학적 특성을 기반으로 미세먼지의 발생과 이동을 모델링 함으로써 미세먼지의 패턴과 오염도를 예측한다[8]. 미세먼지 예측에 쓰이는 결정론적 기법에는 기상데이터와 화학입자상 및 가스상 물질의 상호작용을 고려한 Weather Research and Forecasting(WRF)-Chem모델[9]과 대기 오염 물질의 배출, 확산 및 퇴적과 2차 대기 오염 물질의 생성 등 발생하는 광범위한 화학 반응 또한 시뮬레이션할 수 있는 Community Multi-scale Air Quality(CMAQ)모델[10] 등이 있다. 최근에는 이러한 기법들을 결합한 앙상블형태의 모델링을 구현하여 미세먼지를 예측하기도 한다[11].

이러한 방법은 미세먼지의 확산 메커니즘을 자세히 분석할 수 있다는 장점이 있다. 그러나 모델링에 쓰이는 변수의 인과관계를 직접 수식으로 표현해주어야 하기 때문에 수준 높은 사전지식이 요구된다. 또한 변수간 인과관계에 기반하여 만들어진 수식이 물리화학적 현상을 모두 설명하지는 못하기 때문에, 최근에 만들어진 통계적인 방법들보다 정확도가 떨어질 수 있다.

따라서 본 논문에서 사전지식이 없으면 구현이 어려운 점, 통계적 기법이 더 좋은 결과를 보였다는 점을 고려하여 미세먼지 예측을 위한 결정론적 기법은 다루지 않기로 하였다.

2-3. 미세먼지 예측을 위한 통계론적 기법

통계론적인 방법은 물리·화학적 지식을 활용 하지 않고 기존에 측정된 자료에 통계기법을 적용하여 미래 혹은 측정되지 않은 공간의 오염도를 예측한다[8]. 기존 미세먼지 예측 관련 논문들은 통계론적인 방법에서도 딥러닝을 적용하였는지 여부에 따라 다시 분류하기도 하며, 딥러닝을 적용한 기법은 따로 2-4에서 다루기로 한다.

딥러닝을 적용하지 않은 방법에는 Autoregressive Integrated Moving Average(ARIMA), Multiple Linear Regressions(MLR), Support Vector Regression(SVR) 등이 있다. ARIMA는 미세먼지 데이터의 이전 시점 값이 이후 시점 값에 영향을 미친다는 자기상관성과, 시간이 지나면서 변하는 평균값을 고려한 이동평균성을 합친 ARMA를 일반화한 기법이다. 이 기법은 ARMA와 달리 분석대상이 다소 비안정적인 시계열에도 적용될 수 있으며 초기에 미세먼지 예측의 통계론적 방법으로 사용되었다[12]. MLR은 예측을 요구하는 종속 변수가 하나의 설명 변수가 아닌 둘 이상의 설명 변수의 선형 조합으로 표현이 가능하다고 가정한 통계론적 방법이다. 이 기법을 기반으로 습도와 같은 기상요소들을 활용해 미세먼지를 예측하기도 하였다[13]. SVR은 분류문제 예측에 사용되는 Support Vector Machine(SVM)를 임의의 실수값을 예측할 수 있도록 ε-무감도 손실함수를 도입하여 일반화한 방법이다[14]. SVR은 MLR과 달리 비선형 커널 함수를 사용하여 비선형 관계를 가진 변수들을 설명할 수 있는 장점이 있으며, 이를 이용해 미세먼지 예측을 한 연구가 있다[15].

그러나 최근 연구의 통계론적 모델들간의 미세먼지 예측 결과에 따르면, 딥러닝을 이용한 기법이 딥러닝을 이용하지 않은 통계론적 기법보다 좋은 연구결과를 보이는 것으로 나타났다[16]. 따라서 본 논문에서는 미세먼지 예측을 위한 통계론적 기법 중에서 딥러닝을 이용한 기법에 중점을 두었다.

2-4. 미세먼지 예측을 위한 딥러닝 기법

미세먼지농도 예측 문제는 일정시간 간격으로 측정된 이전 미세먼지 농도로부터 다음 시간의 미세먼지 농도를 구한다는 점에서 시계열 예측 문제라고 할 수 있다. 따라서 Recurrent Neural Network(RNN)기반의 딥러닝 모델링이 많이 쓰이고 있으며, 여기에 주변 미세먼지 농도와 풍향·풍속과 같은 공간적인 데이터 특성을 입력 데이터에 추가하거나 공간 데이터를 더 잘 분석할 수 있도록 모델 구조를 변형하기도 한다.

2-4-1. Long Short Term Memory (LSTM) [17]

시계열 데이터 예측에 쓰이는 기본모델인 RNN의 경우, 예측 시계열이 길어질수록 초기 정보가 최근 정보에 전달이 잘 되지 않는 vanishing gradient 및 exploding gradient 문제가 있다. LSTM은 이 문제를 해결하기 위해 식(1)-(5)와 같이 RNN의 각 cell마다 장시간 정보를 저장하는 memory cell인 c_t와 정보의 흐름을 조절하는 input gate i_t, forget gate f_t 그리고 output gate o_t를 추가한 기법이다. x는 입력데이터, h는 final state, 연산은 요소별 곱셈을 뜻하는 Hadamard product 연산자; W와 b는 각각 weight matrix와 bias vector를 의미한다. σ와 tanh는 각각 sigmoid activation과 hyperbolic tangent activation을 의미한다.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (3)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co} \circ \mathbf{c}_t + \mathbf{b}_o) \quad (4)$$

$$h_t = o_t \circ \tanh(c_t) \quad (5)$$

LSTM모델은 매 시간마다 이전 시간의 상태값인 c_{t-1}와 h_{t-1} 및 현재 입력값인 x_t를 입력으로 받는다. 식(1)과 (3)에 의해 현재 입력값인 x_t는 input gate를 거쳐 현재 상태 c_t에 기억된다. 그리고 식(2)과 식(3)에 의해 과거 상태 c_{t-1}는 forget gate를 거쳐 잊혀진다. 마지막으로 식(4)와 (5)에 의해 현재 입력값 및 상태값 x_t, c_t와 과거 출력값 h_{t-1}이 output gate를 거쳐 현재 출력값 h_t을 결정한다.

그림 1은 LSTM을 이용해 이전 r시간부터 1시간 전까지의 1시간 간격의 연속된 미세먼지 데이터에서 temporal feature를 뽑아낸 뒤, 이를 기상데이터와 Hour Of day, Month Of Year과 같은 temporal predictor를 합쳐서 Fully Connected layers(FCs)에서 처리하여 다음 미세먼지 농도를 예측한 연구에서 사용된 모델이다[16]. 이 모델은 딥러닝을 쓰지 않은 통계적 기법인 ARMA, SVR보다 더 좋은 결과를 나타냈다.

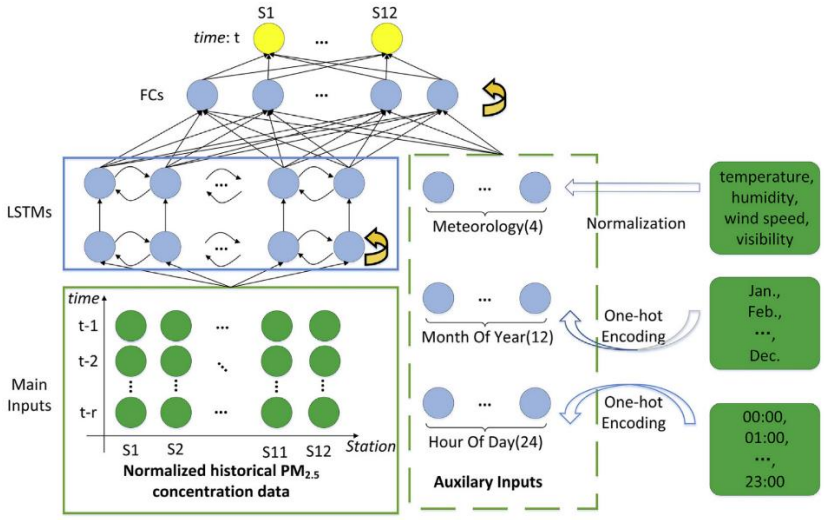
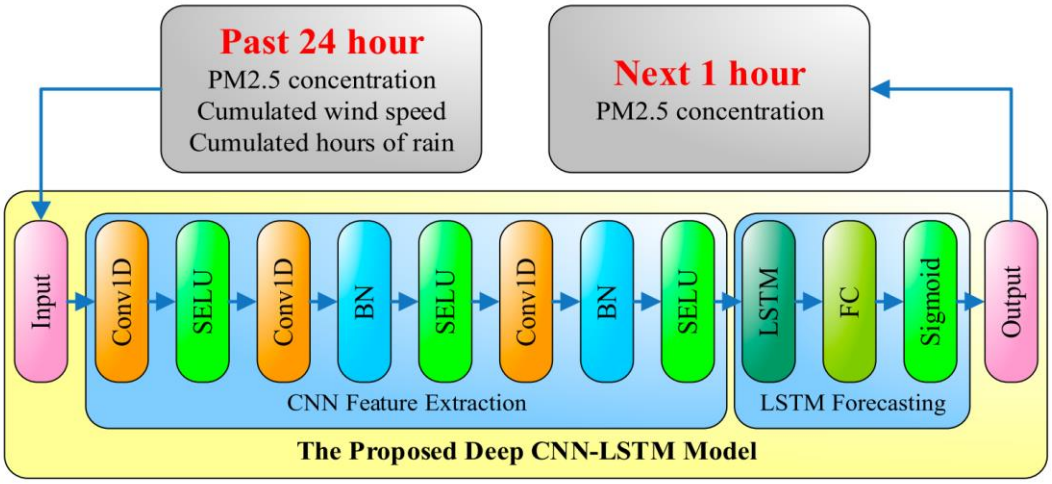


그림 1 : LSTM 기반 Model 구조 예시 [16]

2-4-2. CNN-LSTM 모델

CNN-LSTM은 시간 또는 공간축에서 인접한 데이터에 convolution 연산을 적용해 feature를 뽑아내는 CNN(Convolutional Neural Network) layer를 LSTM layer에 적층한 구조이다. 그림 2는 베이징 미세먼지를 예측하기 위한 연구에서 사용된 CNN-LSTM기반 모델링이며, 24시간 전부터 1시간 전까지의 PM2.5 데이터와 바람속도, 시간당 강수량을 이용해 1시간 뒤를 예측한다[18]. 중첩된 3개의 CNN layer는 시간축에 대해 인접한 데이터간의 feature를 압축하는 역할을 하며, 이 압축된 feature는 LSTM의 입력으로 들어간다. 해당 연구에서 제안한 모델은 LSTM만 이용한 모델보다 조금 더 나은 예측 결과를 보여주었다.

그러나 LSTM layer과 CNN layer를 단순히 이어서 만든 이 CNN-LSTM 모델은 LSTM layer 입력이 각 feature마다 unfold된 1차원으로 주어져야 하므로, 시공간 특성을 동시에 분석할 수 없다는 단점이 있다. 즉, CNN-LSTM 모델은 미세먼지가 공간으로 확산하는 현상을 고려하기 어렵다.



Conv1D: One Dimensional Convolution Layer
LSTM: Long Short Term Memory Neural Network
FC: Fully Connected Neural Network
Sigmoid: Sigmoid function
SELU: Scaled Exponential Linear Unit
BN: Batch Normalization

그림 2: CNN-LSTM 기반 Model 구조 예시 [18]

2-4-3. Convolutional LSTM 모델

LSTM layer에 convolution 연산을 결합한 Convolutional LSTM 모델은 2차원 공간정보가 데이터에 포함되며 시공간 특성을 따로 분석하는 것이 아니라 동시에 분석한다는 점에서 CNN-LSTM의 단점을 해결한다[19]. LSTM의 모든 행렬 곱은 식(6)-(9)와 같이 convolution operation으로 대체되었다. 또한, 2차원 공간정보를 포함하기 위해, input gate i_t, forget gate f_t, output gate o_t와 입력값 X, final state H, memory cell C 그리고 가중치 W가 feature를 포함한 3차원 tensor로 쓰인다.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \quad (6)$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \quad (7)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * H_{t-1} + b_f) \quad (8)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \quad (9)$$

$$H_t = o_t \circ \tanh(C_t) \quad (10)$$

Convolutional LSTM의 시공간정보의 전달은 그림 3과 같이 이루어진다. 각 time-step에서는 convolution kernel이 이전 상태인 H_{t-1}, C_{t-1}의 3차원 tensor를 sliding window하여 kernel size 범위(receptive field)에 포함된 인접 데이터의 정보를 뽑아내 다음 상태(H_t, C_t)로 전달하는 state-to-state transition이 발생한다. 또한 현재 입력값 X_t의 3차원 tensor를 convolution kernel이 sliding window하여 다음 상태에 전달하는 input-to-state transition이 발생한다. kernel size를 적절히 설정한다면 이 transition들에 의해 각 time-step마다 상관관계가 상대적으로 큰 인접한 데이터에서만 feature를 extraction할 수 있으며, time-step이 누적되면서 원거리에 있는 데이터 또한 전파되어 같이 feature

extraction할 수 있는 장점이 있다.

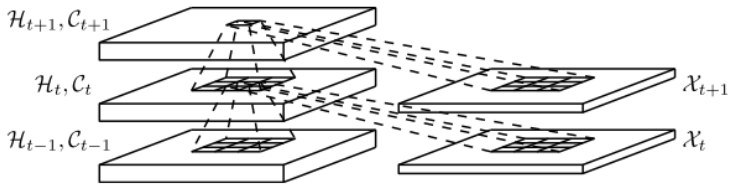


그림 3: Convolutional LSTM의 데이터 transition [19]

제3장 ConvGRU 기반 미세먼지 예측 모델링

3-1. 관련 이론

3-1-1. ConvGRU

ConvLSTM은 시공간을 동시에 분석할 수 있다는 장점이 있지만, 많은 연산과정이 추가되면서 계산복잡성이 높다. 컴퓨팅 자원이 충분하다면 ConvLSTM을 사용해도 큰 무리가 없으나 컴퓨팅 자원이 부족할 경우 계산 복잡성을 낮출 필요가 있다. Convolutional Gate Recurrent Unit(ConvGRU)는 ConvLSTM와 비슷하나 LSTM가 아닌 GRU에 Convolution 연산을 추가한 기법으로, training parameter를 줄이고 계산 복잡성을 낮추도록 하는데 해결 방법이 될 수 있다. 또한 계산복잡성을 낮춤으로써 일부 성능향상도 기대할 수 있다.

GRU는 LSTM에서 게이트 일부를 생략하여 연산을 좀더 간소화시킨 셀 구조이다[20]. 식(11)-(12)처럼 update gate인 z와 reset gate인 r이 있으며, 모두 현 시점의 입력 값 x_t와 이전 시점 출력 값 h_{t-1}을 반영해 구해진다. 활성화 함수로는 시그모이드 함수 σ를 사용하며 W와 U는 각각 input weight matrix, output weight matrix를 나타낸다. 식(13)은 현 시점에서 기억할 정보를 정의하며, 현 시점 입력 정보 Wx_t에 과거 정보인 Uh_{t-1}를 모두 반영하나, 과거 정보는 reset gate의 값에 따라 반영되는 비가 달라질 수 있다. 현재 출력 값 h_t는 식(14)에 의해 정의되며, 이전 출력 값 ht-1과 식(13)에서 구한 값에 대해 update gate인 z_t를 가중치로 둔 가중 평균값을 계산한다. 즉, zt가 0면 과거 출력 값을 그대로 현재 출력 값으로 사용하게 된다.

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (11)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (12)$$

$$\tilde{h}_t = \tanh(W x_t + U(r_t \circ h_{t-1})) \quad (13)$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad (14)$$

ConvGRU는 GRU의 식(11)-(13)에서 weight matrix와 x_t및 h_{t-1}사이에 convolution 연산이 식(15)-(17)과 같이 추가되면서 ConvLSTM와 동일하게 3차원 시공간정보를 동시에 분석할 수 있다[21].

$$z_t = \sigma(W_z * x_t + U_z * h_{t-1}) \quad (15)$$

$$r_t = \sigma(W_r * x_t + U_r * h_{t-1}) \quad (16)$$

$$\tilde{h}_t = \tanh(W * x_t + U * (r_t \circ h_{t-1})) \quad (17)$$

3-1-2. Locally-Connected Layer

Locally-Connected layer는 CNN과 그림 4와 같이 동일하게 현재 layer의 각 뉴런이 kernel size(receptive field) 범위 내에 있는 뉴런에만 연결되는 local connectivity 특성을 가지나, receptive field마다 가중치 파라미터를 독립적으로 두기 때문에 weight sharing 특성을 가지지 않는다.

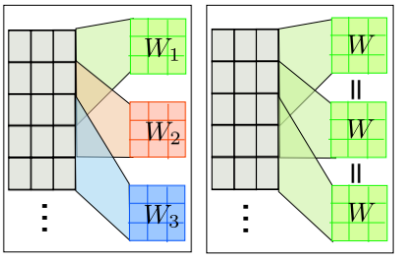


그림 4: CNN (왼쪽)과 locally-connected layer (오른쪽)의 방식 비교[22]

이러한 성질 때문에 locally-connected layer가 학습시켜야 할 가중치 파라미터 개수는 같은 입력데이터가 주어진 CNN의 파라미터의 개수와 현재 layer의 뉴런 개수의 곱으로 결정되며 상당히 많은 메모리와 연산을 요구하게 된다. 또한, 주어진 입력 feature에 대한 일반화 능력이 떨어지는 단점이 있다[22]. 그러나 Locally-connected layer는 receptive field에 독립적인 feature를 더 잘 뽑을 수 있다는 장점이 있다.

3-2. 제안 모델링

기존 예측 시스템에서는 2차원 공간정보를 분석할 수 없어 미세먼지 확산 현상을 고려하기 어려운 CNN-LSTM이나 계산복잡성이 높은 ConvLSTM을 사용한다. 이러한 단점을 보완하기 위해 제안 모델링에서는 2차원 공간정보 또한 분석할 수 있으면서 계산복잡성을 줄인 ConvGRU를 사용한다. 또한 CNN과 달리 weight sharing을 하지 않아 외부 유입요인이 국내의 개별 지역에 어떠한 영향을 미치는지 더 잘 분석할 수 있는 locally-connected layer를 사용해 예측성능을 높인다.

또한 기존의 예측모델이 모든 input variable들을 한번에 모두 분석하여, 2-2의 미세먼지에 영향을 주는 다양한 요인들을 동시에 반영할 수는 있으나, variable수가 많은 경우 연산과정에서 각 요인에 상관없는 input variable까지도 동시에 고려할 수 있다. 따라서 모델이 개별 요인에 의한 영향을 정확히 반영하기 어렵다. 곧, 모델 예측 성능을 악화시키거나 최적해 수렴에 많은 시간이 소요될 수 있다.

따라서 제안한 모델은 그림 5와 같이 세 파트로 이루어진 입력으로 이전 24시간의 데이터를 이용해 다음 T시간 뒤의 국내 전역의 미세먼지를 예측한다. 각 파트는 각각 지역 의존적인 데이터에 기반해 feature를 뽑는 Terrain(T)파트, 바람의 영향을 고려해 feature를 뽑는 Wind(W)파트 그리고 중국의 영향을 고려해 feature를 뽑는 China(C)파트로 나뉜다. 분리된 각 파트들은 목적에 관련된 input variable만을 이용해 1차적인 feature를 뽑아내며, 뽑힌 feature들은 뒷부분에서

결합되어 다른 layer에 의해 분석되면서 미세먼지의 영향을 주는 요인들을 모두 반영할 수 있게 된다.

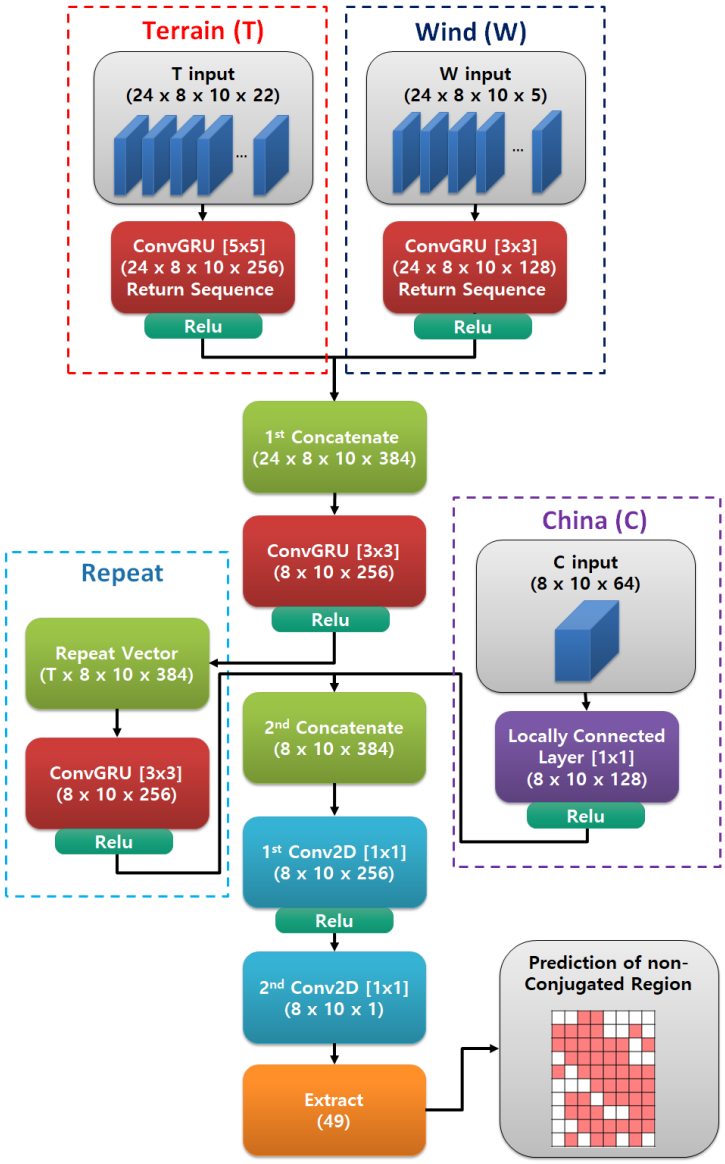


그림 5: 제안한 미세먼지 예측 모델링

각 파트 input의 소괄호()는 입력 dimension을 나타낸다. T, W파트는 시공간차원을 포함한 4차원 텐서 데이터를 입력으로 사용하며 (time-step, X좌표, Y좌표, input feature)의 차원을 가진다. C파트는 공간차원만을 포함한 3차원 tensor 데이터를 사용하며, (X좌표, Y좌표, input feature)의 차원을 가진다. 여기에서 X좌표, Y좌표는 국내 전역을 8x10의 grid로 나누었을 때의 공간좌표이며, 이에 대한 자세한 설명은 제 4장에 서술하였다. layer의 소괄호는 출력 dimension을 나타내며 dimension의 마지막 차원은 해당 layer의 filter개수를 의미한다. Convolution연산이 포함된 layer의 대괄호는 공간(XY)에 대한 kernel size를 나타낸다.

각 파트의 입력 dimension은 파트에 해당하는 layer의 역할과 관련이 있다. T, W파트는 각 지역별로 독립적인 데이터가 주어지며, 이 데이터의 시공간 특성을 동시에 분석하기 위해 4차원 데이터를 필요로 하는 ConvGRU를 사용하였다. C파트는 중국 미세먼지가 국내의 각 지역에 어떻게 독립적으로 영향을 미치는지 분석하기 위해서 locally-connected layer를 사용하였고, 모든 지역에 대해 X, Y좌표를 제외하고 중국의 영향과 관련된 데이터가 공통으로 주어진다. locally-connected layer는 3차원 데이터를 필요로 하기 때문에 시간차원을 생략한 대신, input variable에 과거 데이터를 일부 포함한다. 각 파트에 해당하는 데이터에 대한 설명은 뒷부분에 서술한다.

kernel size는 2-4-3에서 서술한 것처럼, time-step마다 convolution 연산을 할 때 어느 범위까지의 인접한 데이터를 활용할지 결정하는 receptive field 크기를 의미한다. 모델링에 쓰인 ConvGRU와 locally-connected layer의 kernel size는 2차원 공간상에서 잡는 receptive field의 크기를 의미한다. T파트의 ConvGRU는 조금 넓은 범위의 주변 데이터도 고려하기 위해 5x5의 kernel size를 가진다. W파트의 ConvGRU는 본 연구의 시간해상도(time-step)인 1시간안에 바람에 의해 확산이 가능한 범위를 고려하여 3x3의 kernel size를 가진다. C파트의 locally-connected layer는 중국데이터가 국내에 어떻게 영향을 미치는지 각

지역별로 독립적으로 분석하도록 1x1의 kernel size를 가진다.

T와 W파트에서 뽑힌 1차 feature들은 합쳐진 뒤, 다음 ConvGRU의 input으로 들어가며 이 layer에서 이전 24시간 동안의 국내 데이터에 의한 요인이 결합된 2차 feature를 얻게 된다. 이 결합된 feature들을 다음 T시간의 미세먼지 예측을 위한 마지막 feature로 쓸 수도 있다.

그러나 이 feature들은 이전 24시간에서 뽑힌 feature이며 다음 1시간부터 T-1시간의 feature도 순차적으로 고려하는 중간과정을 넣는다면, 예측성능이 향상될 것을 기대할 수 있다. 따라서 본 연구에서는 이를 반영하기 위해 Repeat파트를 추가하여, 이 2차 feature들을 T번 복사한 뒤 ConvGRU의 입력으로 사용한다.

Repeat파트에서 ConvGRU에서 뽑힌 3차 feature들은 C파트에서 뽑아낸 feature과 다시 합쳐져 1^{st} Conv2D에 의해 국내와 중국의 요인을 모두 고려한 256개의 final feature를 만들어내고, 마지막 2^{nd} Conv2D를 통해 모든 지역의 미세먼지 예측값을 얻게 된다. 단, 몇몇 지역은 측정소가 포함되어 있지 않으며, 보간법을 통해 주변 지역에서 보간한 데이터이기 때문에 실제로 측정소가 포함된 49곳의 지역만 예측결과로 사용한다.

예측값을 최종적으로 뽑기 위한 2^{nd} Conv2D를 제외하고 training parameter를 포함한 모든 layer에는 비선형 연산을 추가하기 위해 최근 딥러닝 연구에서 주로 사용되고 있는 Relu를 활성화 함수로 추가한다.

각 파트에 사용된 입력 variable들을 정리하면 표 1과 같다. T파트는 지역별 특성에 의한 feature extraction을 목적으로 하기 때문에 기상요인 오염원 요인 및 측정한 시간대와 위치좌표 등을 포함하는 spatial predictor & temporal predictor를 모두 포함한다. 반면, W파트는 바람에 의한 미세먼지 확산을 고려하기 때문에 이와 상관관계가 큰 PM10, 풍속, 풍속_u성분, 풍속_v성분, 습도만을 고려한다. 중국의 영향을 고려한 C파트는 중국의 미세먼지가 국내로 유입되는 영향을 고려하여, 중국의 데이터를 포함한 기타 데이터를 모두 포함한다. 다만,

사용된 기상데이터 및 오염원 데이터(#0~#11)은 중국데이터를 얻지 못했기 때문에 중국의 영향을 가장 직접적으로 받는 백령도 데이터를 사용하였다. 또한, C파트의 #0, #7, #8, #12에 대해서는 중국 및 백령도 미세먼지 농도가 거리차이 때문에 국내에 전달되는 지연을 고려하여 scalar 데이터가 아닌 각각 6시간전, 23시간전 데이터부터 현재까지의 vector 데이터를 사용한다. variable에 대한 자세한 설명은 제 4장에서 진행한다.

표 1: 파트 별 입력 variable

#	Variable	Dimension (for C input)	T input	W input	C input
0	PM10	scalar (7)	O	O	O
1	SO2	scalar	O	X	O
2	NO2	scalar	O	X	O
3	기온	scalar	O	X	O
4	풍속	scalar	O	O	O
5	상대습도	scalar	O	O	O
6	증기압	scalar	O	X	O
7	풍속_u성분	scalar (7)	X	O	O
8	풍속_v성분	scalar (7)	X	O	O
9	강수량	scalar	O	X	O
10	일교차	scalar	O	X	O
11	일최저기온	scalar	O	X	O
12	Beijing PM2.5 [t-23,t]	24	X	X	O
13	계절	4	O	X	O
14	시간대	6	O	X	O
15	X좌표	scalar	O	X	O
16	Y좌표	scalar	O	X	O

제4장 실험 설계

4-1. 실험 개요

4-1-1. 실험 가설

본 연구에서 제안한 모델은 다양한 가정을 통해 만들어졌기 때문에, 해당 가정을 검증하는 것이 필요하다. 또한, 기존의 딥러닝 기반 예측 모델과 성능비교를 통해 제안하는 모델이 이전에 연구된 모델보다 더 나은 성능을 가짐으로써 본 연구가 의미를 가지는지 검증할 필요가 있다. 따라서 다음과 같은 5가지 가설을 세우고 실험을 진행한다.

- 가설 1: 제안한 모델은 시공간정보까지 동시에 고려할 수 있기 때문에, 시간정보만 고려한 CNN-LSTM모델보다 더 나은 성능을 보여준다.
- 가설 2: ConvGRU는 기존에 사용되는 ConvLSTM보다 계산복잡성을 낮출 수 있으므로, 적은 training parameter로도 비슷하거나 향상된 예측성능을 가질 것이다.
- 가설 3: 제안된 모델에 포함된 Repeat 파트는 다음 1시간부터 T-1시간의 중간 예측과정을 고려할 수 있기 때문에 성능을 향상시킬 것이다.
- 가설 4: 풍향, 풍속, 습도 및 주변 미세먼지 농도를 이용해 feature extraction하면 국내의 미세먼지 확산요인을 잘 고려할 수 있기 때문에 예측 성능을 높일 수 있을 것이다.
- 가설 5: 중국 미세먼지 데이터와 풍향 풍속 정보를 적절히 feature extraction하면 국외 미세먼지 유입요인도 고려할 수 있기 때문에, 중-장기간의 예측 성능을 높일 수 있을 것이다

4-1-2. 비교 모델링

실험 가설을 검증하기 위해서 아래와 같이 제안 모델의 일부를 변경된 모델 및 기존 연구에서 사용된 방식의 비교 모델을 설계하였다.

- T+W+C모델: 본 연구에서 제안하는 모델의 모든 입력 파트를(T, W, C) 포함한다.
- T+W모델: 본 연구에서 제안하는 모델에서 C파트를 생략한다. 가설 5를 확인하기 위해 T+W+C모델과 비교될 수 있다.
- T모델: 본 연구에서 제안하는 모델에서 C, W파트를 생략하고 T파트만 남긴다. 가설 4를 확인하기 위해 T+W모델과 비교될 수 있다.
- T(simple)모델: T모델에서 Refeat파트를 생략한다. 가설 3을 확인하기 위해 T모델과 비교될 수 있다.
- ConvLSTM모델: T모델에서 모든 ConvGRU를 ConvLSTM으로 바꾸되, filter size와 같은 hyper-parameter는 유지한다. 가설 2를 확인하기 위해 T모델과 비교될 수 있다.
- CNN+ConvLSTM(1x1)모델: 그림 6과 같이 기존 미세먼지 예측을 위해 구현된 CNN-LSTM[18]모델과 유사한 구조를 가지는 모델을 설계한다. 이 모델은 T모델의 입력을 그대로 가지며, 가설 1을 확인하기 위해 T모델과 비교될 수 있다. 비교를 위해 Conv1D가 아닌 Conv3D, LSTM이 아닌 ConvLSTM을 쓴 이유는 해당 논문에서 제시한 모델이 하나의 측정소에 예측하는 문제이며, 다수의 지역을 예측하는 본 연구의 실험 환경에 맞게 다시 디자인할 필요가 있기 때문이다. 대신, ConvLSTM의 kernel size는 1x1로 맞추고, Conv3D의 kernel size 또한 3x1x1로 맞춰 시간적 feature는 뽑되, 2차원 공간적으로 인접한 지역의 데이터에서는 뽑지 않고 개별 지역의 데이터만 사용하게 함으로써 기존 논문의 실험

조건과 유사하게 맞추었다.

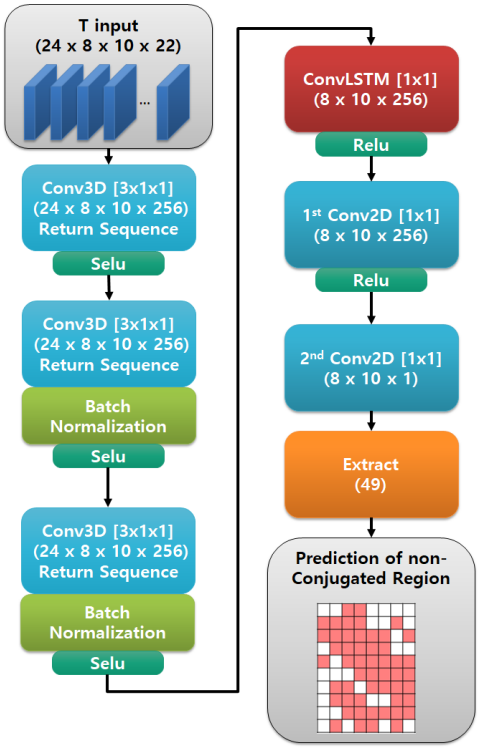


그림 6: CNN+ConvLSTM(1x1)모델

4-1-3. 실험 시스템 설계

가설을 검증하기 위한 실험 시스템은 그림 7과 같이 구성된다. 모델에 필요한 데이터인 input variable은 실험환경에 맞게 전처리된 후, training data, validation data, test data로 나뉜다. Input variable과 전처리 구성요소는 뒷장부터 자세히 설명한다. Training data는 모델의 가중치를 업데이트하기 위한 학습데이터이며, 이전에 설정된 가중치를 기반해 계산한 예측값과 참값간의 training loss를 구하게 되면 optimizer의 back propagation을 통해 가중치가 업데이트 된다. Validation data는 early stopping 기법을 통해 overfitting을

방지하거나 hyper-parameter를 개선하기 위해 testing data대신 사용되었다. Early stopping이란 모든 학습 데이터를 한번 학습하는 과정인 1 epoch마다 validation loss를 계산해서 이전보다 loss의 개선이 진행되지 않으면 학습을 중단하는 기법이다. Testing data는 모델을 학습하는 과정에는 사용되지 않는 데이터이며 모델의 성능을 최종적으로 평가하기 위한 예측값을 얻기 위해서 사용된다.

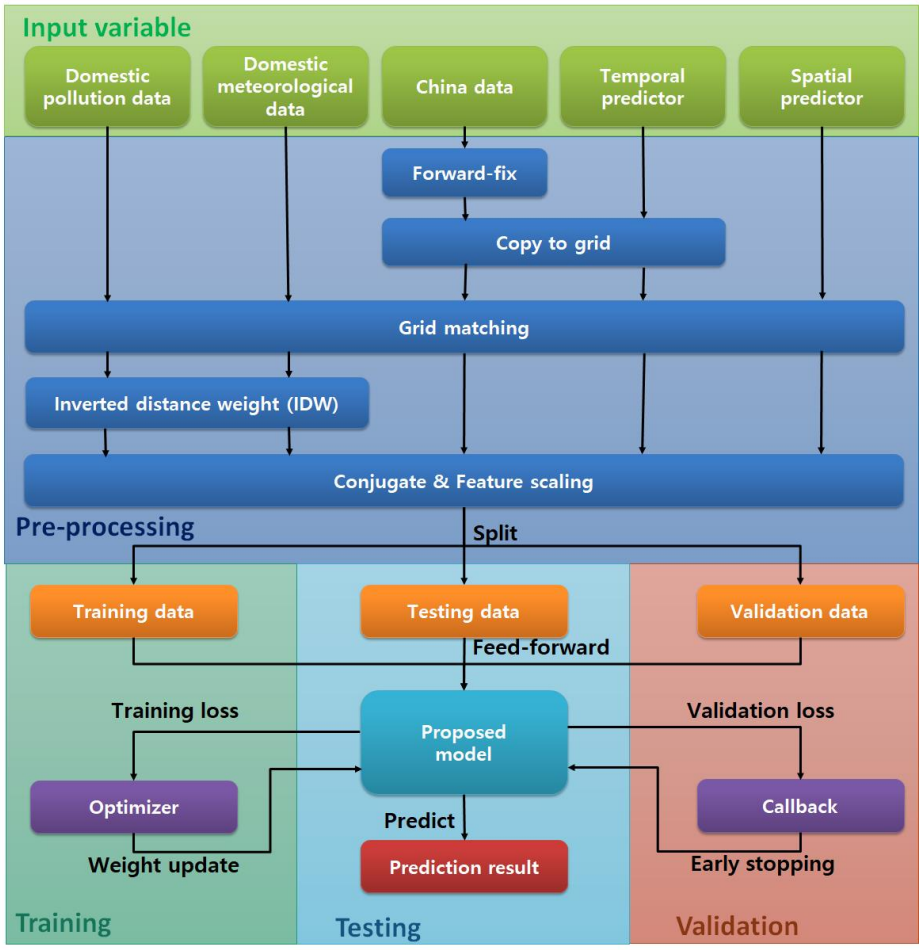


그림 7: 실험시스템 흐름도

4-2. 데이터 수집 및 전처리

4-2-1. 보간법

미세먼지 예측에 필요한 측정 데이터들은 통신 장애, 측정소 자체 문제 등으로 측정되지 못한 경우가 있다. 이 결측치들이 어떻게 채워지는지에 따라 잘못된 모델 training이 발생할 수 있으며 이는 예측 성능에 직접적인 영향을 미친다. 결측치를 채우는 방법을 보간법이라고 하며, 공간적으로 주변의 측정값을 이용하여 결측치를 채우는 공간 보간법 및 시간적으로 가까운 데이터에서 보간을 하는 시간 보간법이 있다.

가. Inverse Distance Weighting (IDW)

대기환경 분야에서는 공간상에 존재하는 데이터를 보간하는 방법으로 Inverse Distance Weighting (IDW) 기법, Kriging 기법 등 다양한 방법이 있으나, 본 연구에서는 필요한 연산이 간단한 IDW를 사용한다. IDW 기법은 공간적으로 인접한 지점 사이에 값은 비슷하나, 지점간 거리가 멀수록 그 유사성이 감소한다는 가정하에, 거리에 따른 가중치를 두어 결측치를 추정하는 방법이다 [23]. 계산식은 식(18)과 같다.

$$R_t = \sum_{i=1}^{n}\left(P_i / D_i^b\right) / \sum_{i=1}^{n}\left(1 / D_i^b\right) \qquad (18)$$

R_t는 보간을 통한 추정값이며, n은 보간에 쓰이는 인접한 지점의 수이다. P_i는 인접 지점의 측정값, D_i는 인접한 지점과 추정 지점 사이의 거리이며, b는 역거리가중치이다.

나. Forward-fix

시간 보간법은 시계열로 측정되는 모든 누락된 데이터에 쓰이며, forward-

fix 기법은 시간 보간법에 쓰는 일반적인 기법 중에 하나로 가장 최근에 측정된 값으로 결측치를 채운다[24]. 식(19)-(20)은 이 기법을 수식으로 나타낸 것이며, 현재 측정값 x_t가 측정되었으면 m_t는 1, 결측되었으면 m_t는 0을 가진다. $x_{t'}$는 측정된 최신 값을 나타내며, 식(20)에 의해 현재 측정값은 결측된 경우에만 측정된 최신 값을 그대로 가지게 된다.

$$m_t = \begin{cases} 1, & x_t \; valid \\ 0, & x_t \; missed \end{cases} \tag{19}$$

$$x_t \leftarrow m_t x_t + (1 - m_t) x_{t'} \tag{20}$$

4-2-2. 실험 데이터

실험을 위해 수집된 데이터는 크게 표 2와 같이 국내 대기 오염 측정소에서 수집된 오염원 데이터[25], 국내 기상 관측소에서 수집된 기상 데이터[26], 그리고 중국 PM2.5 데이터[27]를 포함한다. 측정소는 국내 전역을 평면좌표계에 쓰이는 Transverse Mercator (TM)좌표를 기준으로 X좌표(45.85~97.92) 및 Y좌표(65.81~133.75)구간 범위에 있는 것만을 포함한다. 데이터를 수집할 필요가 없는 temporal predictors, spatial predictors에 대해서는 이 문단의 '라.'에서 서술한다.

표 2: 수집된 데이터 개요

	국내 대기 오염	국내 기상 관측	중국 PM2.5
측정 기간	2014.1.1 1:00 ~ 2018.12.31 23:00 (1시간 간격)		
측정 시간	43823 시간		
측정 지역	일부 섬을 제외한 전 지역		Beijing
사용된 측정소 수	207	54	1
사용된 Variable 수	3	7	1

가. 국내 오염원 데이터

국내 오염원 데이터는 표 3과 같이 본 연구의 예측 대상인 PM10을 포함하며, 미세먼지 2차 생성에 관여하는 SO_2와 NO_2도 포함한다. PM2.5의 경우 2015년부터 관측이 시작되었으며, 관측 가능한 측정소도 한정이 되어있기 때문에 예측 대상 및 input variable에서 제외하였다.

표 3: 국내 오염원 데이터

Variable	Unit	Mean	St. dev.	Range
PM10	$\mu g/m^3$	45.83	31.50	[0, 1484]
SO2	ppm	0.0044	0.0037	[0, 0.3760]
NO2	ppm	0.0198	0.0148	[0, 0.364]

국내 대기 오염 예측을 할 때, 측정소 단위로 공간해상도를 맞출 경우 미세먼지의 농도가 국소적 범위의 주변 환경에서 생기는 변수의 영향을 매우 강하게 받기 때문에 국내 전역에 대한 장기 예측이 어려워진다. 따라서 본 연구에는 그림 8과 같이 측정소가 포함한 TM 좌표 구간에 대해 8 x 10의 grid로 나누고, 공간해상도를 이 grid로 나누어진 구역으로 정의한다.

공간해상도를 맞추기 위해서, 거리가 상대적으로 다른 측정소와 많이 떨어져 있고 중국 미세먼지의 영향을 강하게 받는 백령도 데이터를 제외한 206곳의 측정소를 이 grid에 mapping한다. 각 구역에 Mapping된 측정소의 개수를 그림 8의 오른쪽에 표시하였다. grid로 나누어진 구역의 데이터는 구역에 포함된 미세먼지 측정소의 측정값을 평균하여 구한다.

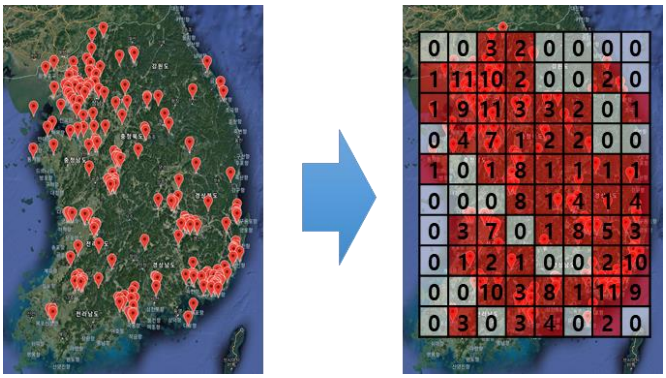


그림 8: 대기 오염 측정소 분포와 grid mapping

이때 측정소가 포함되어 있지 않은 구역이나, 측정소가 있으나 측정소의 통신 장애, 측정소 자체 문제로 특정 기간 동안 데이터가 결측된 구역이 생길 수 있다. 이러한 구역의 데이터를 그대로 비워두게 되면, ConvGRU의 Convolution 연산과정에서 인접지역의 데이터가 모두 채워진 지역과 인접지역 일부 데이터가 없는 지역간의 연산 불균형이 발생하기 때문에 예측 성능이 낮아질 수 있다.

따라서 데이터가 없는 구역들에 대해서는 인접한 구역에 포함된 측정소들의 TM좌표와 데이터 없는 구역의 중심부 TM좌표간 거리를 기준으로 IDW를 적용하여 데이터 보간을 한다. 본 연구에서는 더 가까운 지점에 가중치를 더 많이 두기 위해 IDW를 사용할 때 역거리가중치 b를 4로 정의한다.

나. 국내 기상 데이터

국내 기상 데이터는 2-1의 미세먼지 농도 요인에 근거하여, 표 4와 같이 기온, 풍속, 풍향, 상대습도, 증기압, 강우량, 적설량을 포함한다. 표에서의 강우량, 적설량은 비 또는 눈이 내린 날에 대해서만 집계되었다.

표 4: 국내 기상 데이터

Variable	Unit	Mean	St. dev.	Range
기온	℃	13.02	10.44	[−22.7,41]
풍속	m/s	1.84	1.54	[0,27.4]
풍향	(°)	175.6	123.15	[0,360]
상대습도	%	67.67	22.46	[1,100]
증기압	hPa	12.20	8.34	[0.1,43.4]
강우량	mm	1.525	3.332	[0,107.5]
적설량	cm	5.1478	9.841	[0,110]

그런데, 국내 기상 관측소에서 수집된 이 기상 데이터를 그대로 딥러닝 모델링 input으로 쓸 경우 몇 가지 문제가 발생한다. 첫 번째로 풍향이 특정 두 시점에서 각각 0°과 350°일 때 실제로는 10°의 차이밖에 없으나, 모델이 받아들이는 데이터상으로는 350의 차이를 가지게 된다. 이는 편차가 똑같은 상황인, 특정 두 시점에서 풍향이 각각 0°과 10°일 때와 완전히 다른 결과를 가져올 수 있다. 두 번째로 2−1에서 언급된 중요 기상인자인 일교차와 최저기온은 모델이 주어진 기온데이터로만 알기 어렵다. 또한 모델에 강우량과 적설량을 분리하여 feature로 제공할 필요 없이, 합쳐서 강수량이라는 feature로 제공하면 모델이 강수에 의한 대기 세정효과를 좀 더 잘 학습할 수 있다.

따라서 본 연구에서는 식(21)−(22)을 사용하여 풍속과 풍향으로부터 풍속_u, 풍속_v 성분을 얻어 이를 추가 feature로 사용한다. 식(21)−(22)에서 $wind_u$, $wind_v$는 각각 풍속_u, 풍속_v 성분이며, W는 풍속, D는 degree로 나타낸 풍향이다. 풍향은 0°일 때 북쪽에서 불어오고 값이 커질수록 불어오는 방향이 시계방향으로 바뀌기 때문에, u 성분 및 v 성분으로 변환 시 270°에서 풍향각도를 뺀 각도를 사용하여 변환한다.

$$wind_u = W\cos\frac{(270-D)\pi}{180} \qquad (21)$$

$$wind_v = W \sin \frac{(270 - D)\pi}{180} \qquad (22)$$

그리고 강우량과 적설량을 더해 강수량 feature를 만들고, 이전 24시간부터 현재까지의 최저기온을 일최저기온 feature로, 최고기온과 최저기온의 차를 일교차 feature로 만든다. 이런 전처리를 통해 가공된 기상 정보를 표 5에 제시하였다.

표 5: 전처리를 통해 가공된 추가 기상 정보

Variable	Unit	Mean	St. dev.	Range
풍속_u 성분	m/s	−0.3987	1.699	[−18.9,20.98]
풍속_v 성분	m/s	0.2461	1.631	[−21.14,22.27]
강수량	mm	0.1813	1.5247	[0,115.9]
일교차	℃	10.05	4.06	[0.3, 26.3]
일최저기온	℃	8.39	10.30	[−22.7,31.1]

이 기상 데이터도 마찬가지로 국내 오염원 데이터와 공간 해상도를 맞추고 백령도의 기상데이터를 제외한 53곳의 측정소를 grid의 각 지역에 mapping한다. 측정소의 분포와 각 구역에 mapping된 측정소의 개수를 그림 9에 표시하였다. 측정소가 포함되어 있지 않은 구역과, 결측치가 있는 지역에 대한 보간은 국내 오염원 데이터와 동일하게 IDW를 적용한다.

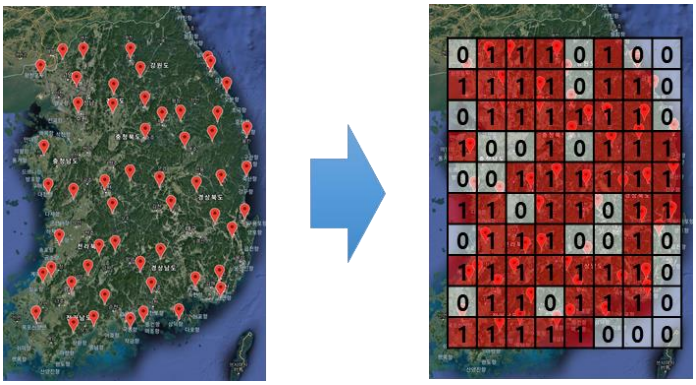


그림 9: 기상 측정소 분포와 grid mapping

다. 중국 데이터

본 연구에서 사용하는 중국 데이터는 베이징의 대기오염 측정소에서 측정한 PM2.5 데이터이며, 표 6에 제시하였다. 중국의 기타 오염원 데이터와 기상 데이터를 포함시키려 하였으나 자료를 구하지 못하였기 때문에, 기상데이터는 지리적으로 가장 가까워 베이징의 영향을 가장 많이 받는 백령도의 데이터로 대체한다. 백령도 데이터는 4-2-2의 '가.', '나.'에서 구하였다.

중국데이터 또한 결측치를 포함하나, 주변 측정소가 없기 때문에 IDW를 적용할 수 없다. 따라서 중국 데이터에 대해서는 시간적 보간 기법인 forward-fix를 사용하여 보간을 하였다.

표 6: 중국 데이터

Variable	Unit	Mean	St. dev.	Range
Beijing PM2.5	$\mu g/m^3$	72.59	78.95	[0,722]

라. Temporal predictors, spatial predictors

데이터가 어느 기간 또는 시간에 측정되었는지 알 수 있는 temporal predictors와 어느 위치에서 측정되었는지 알 수 있는 spatial predictors가 추가적인 variable로 주어진다면 모델이 이 variable들을 기준으로 세분화된 feature extraction이 가능하다. 본 연구에서 사용된 temporal predictors와 spatial predictors를 표 7에 제시하였다.

Variable의 계절은 해당 데이터가 측정된 계절을 의미하며, 시간대는 해당 데이터가 측정된 시간을 0시부터 4시간 간격으로 묶었을 때 해당되는 시간대를 의미한다. 두 variable은 temporal predictors에 해당되며 one-hot-encoding 처리되어 각각 4, 6의 variable dimension을 가진다.

Variable의 X좌표 및 Y좌표는 국내 전역을 8 x 10의 grid로 나누었을 때 각각 열 번호, 행 번호에 대응되는 좌표로 spatial predictors에 해당되며 scalar format을 가진다.

표 7: temporal predictors, spatial predictors

Variable	Range	Format
계절	봄, 여름, 가을, 겨울	4-dimension Vector (one-hot-encoding)
시간대	0~3 시, 4~7 시, 8~11 시, 12~15 시, 16~19 시, 20~23 시	6-dimension Vector (one-hot-encoding)
X 좌표	[0,7]	scalar
Y 좌표	[0,9]	scalar

4-2-3. Feature scaling

Variable마다 값의 범위가 다르기 때문에 variable의 값을 그대로 모델의 입력에 사용한다면, 모델을 트레이닝할 때 값이 상대적으로 큰 variable에 영향을 더 많이 받아 최적해를 찾기 어려워지며, 트레이닝에 더 많은 시간을 소요하게 된다. 따라서 입력 variable을 그대로 모델의 입력 feature로 사용하지 않고, 모든 variable의 값의 범위를 비슷하게 맞춰주는 feature scaling과정을 거친다. feature scaling에는 다양한 방법이 있으며, 일반적으로 원래 variable값에 최소값을 뺀 뒤 (최대값-최소값)으로 나누는 min-max scaling이 주로 사용된다. 하지만 이 방식을 사용하게 되면 극성(+,-)을 가지고 있는 variable 데이터가 모두 양수 또는 0으로 scaling되면서 왜곡이 생길 수 있다. 예를 들어 본 연구에서 사용되는 variable인 풍속_u성분, 풍속_v성분은 값의 극성에 의한 풍향의 특성을 포함하여, 양수일 때와 음수일 때의 풍향은 서로 반대가 된다. 그러나 min-max scaling이 적용되면 이러한 값의 극성이 없어지면서 데이터에 포함된 풍향 특성이 왜곡될 수 있다. 따라서 본 연구에서는 식(23)의 feature scaling 방법을 사용하여 scaling된 input variable의 극성을 그대로 유지하면서 값의 범위를 [-1,1] 이내로 유지한다. 식(23)에서 x_{txyv} 는 시점 t일 때, grid (x,y)좌표에 해당하는 구역의 v번째 variable이며, $|X_v|$ 는 모든 시점의 모든 구역에서 v번째 variable의 절대값 집합이며, max는 집합의 최대값을 반환하는 함수이다.

$$x'_{txyv} = \frac{x_{txyv}}{\max(|X_v|)} \qquad (23)$$

4-3. Hyper-parameter 설정

Hyper-parameter 란 모델이 훈련하면서 업데이트하는 가중치 파라미터나 입력 파라미터를 제외하고 모델을 설계할 때 직접 설정해줘야 하는 파라미터를 말한다. 설정한 hyper-parameter 를 표 8 에 제시하였다. filter 개수나 kernel size 와 같은 layer 에 독립적인 hyper-parameter 는 이전 장에서 제시하였으므로 제외한다. 표에서 prediction length 는 예측을 할 다음 T 시간을 의미하며, 본 연구에서는 여러 시간대를 동시에 예측하도록 모델을 설계하지 않았기 때문에 prediction length 마다 따로 모델을 만들고 결과를 정리하였다. Callback method 에서 early stopping 의 patience 는 앞으로 몇 epoch 까지 validation loss 가 개선되지 않아도 계속 학습을 진행할 것인지에 대한 문턱값을 의미한다.

표 8: Hyper-parameter 설정

Parameter	Value
Training data	60% (2014~2016)
Validation data	20% (2017)
Testing data	20% (2018)
Prediction length (T, hour)	[1, 4, 12, 24]
History length (hour)	24
Time interval (hour)	1
Optimizer	Amsgrad
Learning rate	0.00075
Max training epochs	100
Loss function	Mean square error
Callback method	Early stopping with patience = 10

제 5 장 실험 결과

실험결과를 내기 위해, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (IA), Accuracy 가 indicators 로 쓰였다. 각 indicator 들의 계산식은 식(24)－(27)와 같으며, N 은 테스트 데이터 개수, y_i 는 예측값, y_i^* 는 참값, T 는 전체 예측 수(N)에서 예측값의 labeling 과 참값의 labeling 이 일치한 예측 수를 의미한다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2} \tag{24}$$

$$\mathrm{MAE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|} \tag{25}$$

$$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{N}(|y_i - y_i^*|)^2}{\sum_{i=1}^{N}(|y_i - < y_i^* >| + |y_i^* - < y_i^* >|)^2} \tag{26}$$

$$\mathrm{Accuracy} = \frac{100T}{N} \tag{27}$$

IA 는 예측값이 참값과 얼마나 유사한지를 나타내는 지표로 1 에 가까울수록 참값과 예측값이 유사하다고 할 수 있다. Accuracy 는 실수값에 대한 예측 정확도를 나타내는 다른 regression 지표와 달리 classification 지표를 나타낸다. 본 연구에서는 환경부 기준에 따라 미세먼지 농도를 4 단계 분류했을 때의 분류 성능을 나타내며 2019 년 5 월 1 일 기준으로 미세먼지 농도는

좋음(31 미만), 보통(31 이상 81 미만), 나쁨(81 이상 151 미만), 매우나쁨(151 이상)으로 분류된다.

먼저, 제안한 모델을 포함한 비교 모델들의 학습 파라미터 수(가중치)는 표 9 와 같다. CNN+ConvLSTM[1x1]모델이 가장 적고, ConvLSTM 모델이 가장 많다. 같은 구조이나 ConvLSTM 대신 ConvGRU 를 사용한 T 모델이 ConvLSTM 모델보다 3/4 의 가중치 개수만을 가지며 계산 복잡도가 더 작음을 알 수 있다.

표 9: 모델의 학습 파라미터 수

	T	T+W	T+W+C
Training parameters	12,483,841	13,828,609	14,526,977
	T(simple)	ConvLSTM	CNN+ConvLSTM [1x1]
Training parameters	8,944,129	16,623,105	1,003,265

본 연구의 예측 결과는 4-3 에서 설정한 것처럼 1 시간, 4 시간, 12 시간, 24 시간 뒤의 미세먼지(PM10) 예측을 대상으로 하며, 1 시간과 4 시간을 단기간 예측, 12 시간과 24 시간을 중장기간 예측으로 각각 묶어서 결과를 설명한다. 예측은 다수의 구역을 대상으로 하기 때문에, 결과에 해당하는 모든 지표는 모든 지역에 대해 계산된 지표의 평균값으로 산정한다.

비교적 단기간 예측결과에 해당하는 표 10 과 표 11 은 이전 24 시간 데이터로부터 각각 1 시간 뒤와 4 시간 뒤 미세먼지 농도를 예측한 결과를 나타낸다. 제안된 T, T+W, T+W+C 모델에서 비교를 했을 때는 T+W 이 모든 지표에서 가장 좋은 성능을 나타냈다. 이 결과는 단기간 예측에 대해서 바람에 의한 국내 미세먼지 확산 요인과 지역특성에 의존적인 요인들을 같이 고려했을 때 좋은 성능을 보인다는 것을 나타낸다. T+W+C 는 T+W 보다 낮은 성능을 보인 이유는

단기간 예측에서 국내 데이터가 중국 데이터가 보다 예측값간의 상관관계가 월등히 높아 중국 데이터는 국내 데이터와 예측값의 상관관계를 해석하는 데 방해되는 feature 로 작용하기 때문인 것으로 추측된다.

1 시간 뒤, 4 시간 뒤 모두 CNN+ConvLSTM[1x1]이 모든 지표에서 가장 나쁜 예측성능을 보였으며, 단기간 예측에 대해서는 제안된 T, T+W, T+W+C 모델이 기존 연구에서 쓰이는 예측모델보다 더 나은 성능을 보여줌을 알 수 있다. T 모델에서 ConvGRU 를 ConvLSTM 으로 바꾼 ConvLSTM 모델과 T 모델을 비교했을 때는, 1 시간 뒤의 성능은 비슷하나, 4 시간 뒤 예측결과에 대한 모든 지표에서 더 나은 성능을 보여주고 있으며 GRU 가 LSTM 보다 더 적은 계산복잡도를 가지기 때문에 단기간 예측에 대해서는 적은 트레이닝 파라미터를 사용하면서도 비슷하거나 더 나은 결과를 보일 수 있음을 알 수 있다.

T 모델에서 repeat 파트를 생략한 T(simple)모델보다 T 모델이 1 시간 뒤, 4 시간 뒤 예측 결과에 대한 대부분의 지표에서 조금 더 나은 성능을 보여주고 있으며 단기간 예측에 대해서는 repeat 파트를 모델에 포함시키는 것이 타당함을 알 수 있다.

표 10: 1 시간 뒤 예측결과 비교

	T	T+W	T+W+C	T (simple)	Conv LSTM	CNN+ ConvLSTM [1x1]
RMSE	6.33	**6.25**	6.32	6.35	6.30	7.41
MAE	4.08	**4.04**	4.11	4.11	4.08	4.82
IA	0.9857	**0.9859**	0.9857	0.9856	0.9858	0.9804
Accuracy	90.93	**91.06**	90.87	90.93	90.95	89.43

표 11: 4 시간 뒤 예측결과 비교

	T	T+W	T+W+C	T (simple)	Conv LSTM	CNN+ ConvLSTM [1x1]
RMSE	12.84	**12.29**	12.60	13.53	13.48	15.07
MAE	8.18	**7.74**	8.13	8.74	8.51	9.45
IA	0.9318	**0.9406**	0.9360	0.9242	0.9252	0.9049
Accuracy	81.79	**83.02**	81.98	80.73	81.18	79.04

비교적 장기간 예측결과에 해당하는 표 12 와 표 13 은 이전 24 시간 데이터로부터 각각 12 시간 뒤와 24 시간 뒤 미세먼지 농도를 예측한 결과를 나타낸다. 이 결과에서는 단기간 예측결과와 달리 T+W 모델보다 T+W+C 모델이 더 좋은 예측 성능을 보였다. 이는 장기간 예측에 영향을 주는 요인이 국내의 데이터만으로는 모두 설명이 되지 않으며, 바람에 의한 중국 미세먼지의 국외 유입 요인을 같이 고려할 필요가 있음을 보여준다. T 모델과 T+W 모델을 비교했을 때는 T+W 모델이 더 나은 성능을 보임으로써 바람에 의한 국내 미세먼지 확산 요인을 고려하는 것이 장기간 예측에도 여전히 유효함을 확인할 수 있었다.

제안한 모델과 다른 모델들을 비교했을 때는 제안한 모델인 T+W+C 가 다른 모델보다 월등한 성능을 보였으나, 24 시간 뒤 예측을 기준으로 T 모델이 같은 입력값을 가지는 ConvLSTM 모델과 CNN+ConvLSTM[1x1]모델에 비해 특정 지표에서 더 낮은 결과를 보였음을 알 수 있다. 이는 상대적으로 예측이 쉬운 단기간 예측과 달리, 24 시간 뒤 예측 문제를 T 모델이 만들어진 구조와 주어진 국내 데이터만으로는 해결이 난해하기 때문에 비교 모델들과 비슷한 예측 성능을 가지고 있어 나타난 현상이라고도 볼 수 있다.

단기간 예측과 마찬가지로 T 모델과 T(simple)을 비교했을 때는 T 모델이 12 시간 뒤, 24 시간 뒤 예측 성능이 모든 지표에서 더 좋게 나옴으로써 비교적

장기간 예측에 대해서도 Repeat 파트를 포함시키는 것이 성능을 향상시키는 것을 보여주고 있다.

표 12: 12 시간 뒤 예측결과 비교

	T	T+W	T+W+C	T (simple)	Conv LSTM	CNN+ ConvLSTM [1x1]
RMSE	20.03	19.39	**19.38**	21.62	20.29	22.02
MAE	12.86	12.79	**12.29**	13.40	13.46	14.25
IA	0.7768	0.8006	**0.8071**	0.7364	0.7600	0.7399
Accuracy	70.89	70.96	**73.14**	69.70	68.90	68.43

표 13: 24 시간 뒤 예측결과 비교

	T	T+W	T+W+C	T (simple)	Conv LSTM	CNN+ ConvLSTM [1x1]
RMSE	23.51	22.91	**22.41**	23.91	23.66	23.66
MAE	16.47	15.57	**15.13**	16.82	16.34	16.67
IA	0.6236	0.6269	**0.7093**	0.6087	0.6338	0.6452
Accuracy	62.06	63.58	**66.15**	60.15	63.63	64.08

주어진 표 10~13 으로는 개별 지역의 실험 결과값을 알 수 없기 때문에, 49 개 지역의 MAE 값을 각 모델 별로 모두 비교할 수 있도록 그림 10~13 와 같이 그래프를 제시하였다. 그래프의 세로축은 MAE 이며, 가로축은 각 지역을 의미한다. 이 그래프에서 Proposed model 은 제안된 T, T+W, T+W+C 모델 중 각 예측 시간대별로 가장 좋은 성능이 나온 모델을 의미하며 1 시간, 4 시간 뒤 예측에는 T+W 모델이, 12 시간, 24 시간 뒤 예측에는 T+W+C 모델이 사용되었다. 결과에서 소수의 지역을 제외하고 전반적으로 Proposed model 이 비교모델보다 더 나은 성능을 보임을 알 수 있다.

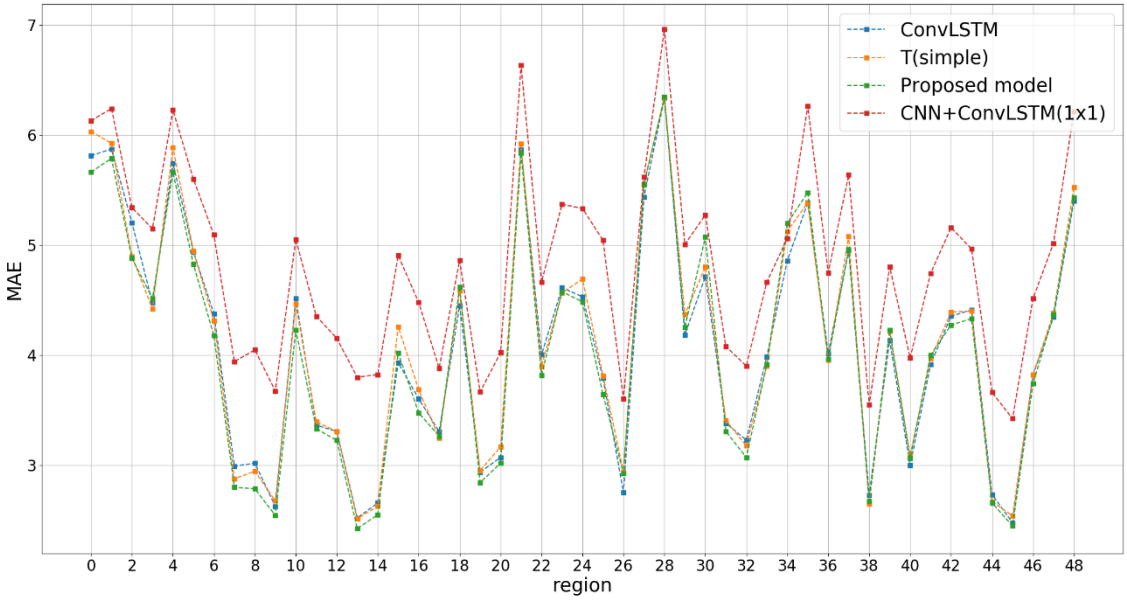


그림 10: 모든 지역에 대한 1 시간 뒤 미세먼지 예측 비교 그래프

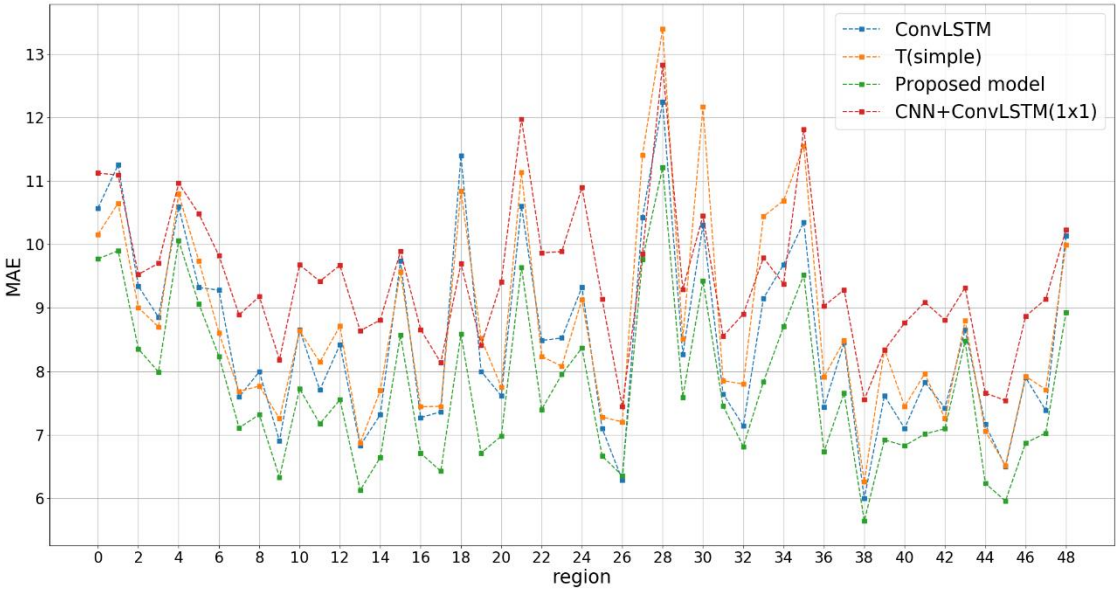


그림 11: 모든 지역에 대한 4 시간 뒤 미세먼지 예측 비교 그래프

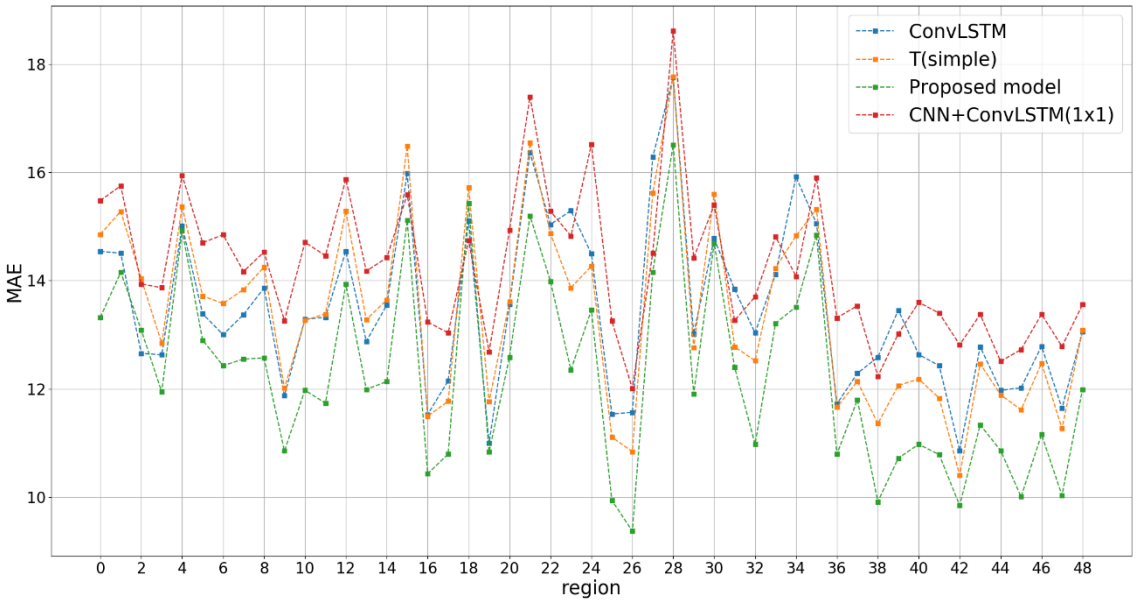


그림 12: 모든 지역에 대한 12 시간 뒤 미세먼지 예측 비교 그래프

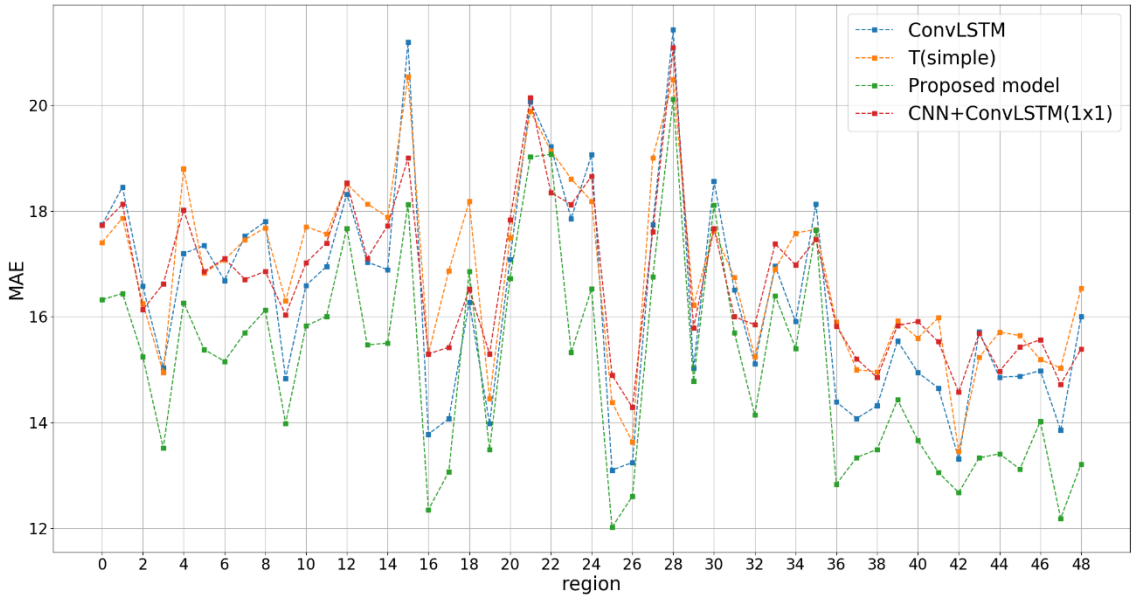


그림 13: 모든 지역에 대한 24 시간 뒤 미세먼지 예측 비교 그래프

또한, 제안한 모델의 예측 성능이 어떻게 되는지 가시적으로 확인하기 위하여, 특정 지역(그림 10~13 에서 22 번)의 예측결과와 참값과 비교한 그래프들을 그림 14~17 와 같이 제시하였다. 각 그림은 4 개의 그래프로 이루어져 있는데, 가장 위의 그래프는 test 전 기간인 1 년(2018.1.1~2018.12.31)에 대한 비교이며, 그 다음 위쪽부터 각각 처음 한달 간의 비교, 중간 한 달간의 비교, 마지막 한 달간의 비교를 의미한다. 그래프의 세로축은 PM10 농도를 의미하며, 가로축은 시간을 의미한다.

1 시간 뒤 예측 결과는 미세먼지 농도의 추이 및 peak 와 valley 들을 잘 잡는 것을 확인할 수 있다. 4 시간 뒤 예측 결과 또한 전반적으로 추이 및 peak 와 valley 들을 잘 잡는 것을 확인할 수 있으나, 몇몇 구간에 대해서는 이전 시간의 참값을 따라가는 delay shift 현상과 및 peak 와 valley 의 중간 값을 가져가는 moving average 현상을 관찰할 수 있다. 이는 예측이 상대적으로 어려운 peak 와 valley 값을 피하고 loss 를 최대한 줄이기 위해 모델이 만들어낸 결과라고 볼 수 있다.

12 시간 뒤와 24 시간 뒤 예측 결과는 1 시간 뒤, 4 시간 뒤 예측 결과와 비교했을 때 상대적으로 예측이 잘 안되고 있음을 알 수 있다. delay shift 현상은 잘 보이지 않으나, moving average 현상이 눈에 띄게 증가했음을 알 수 있으며, 예측 문제가 어렵기 때문에 모델이 loss 를 안전하게 줄일 수 있는 방법인 중간 값을 택하는 방향으로 학습이 되었다는 것을 알 수 있다. 그러나 그래프를 자세히 확인해보면 몇몇 구간에 대해 peak 나 valley 를 잡는 것을 볼 수 있으며, 대부분의 구간에 대해서는 peak 나 valley 가 발생한 시점에서 상대적으로 작은 peak 나 valley 를 발견할 수 있다. 이는 제안한 모델이 장기간 예측에서 절대치를 잡기는 어렵지만 미세먼지가 가장 높아지거나 낮아지는 시점은 어느 정도 잡아낼 수 있음을 알 수 있다.

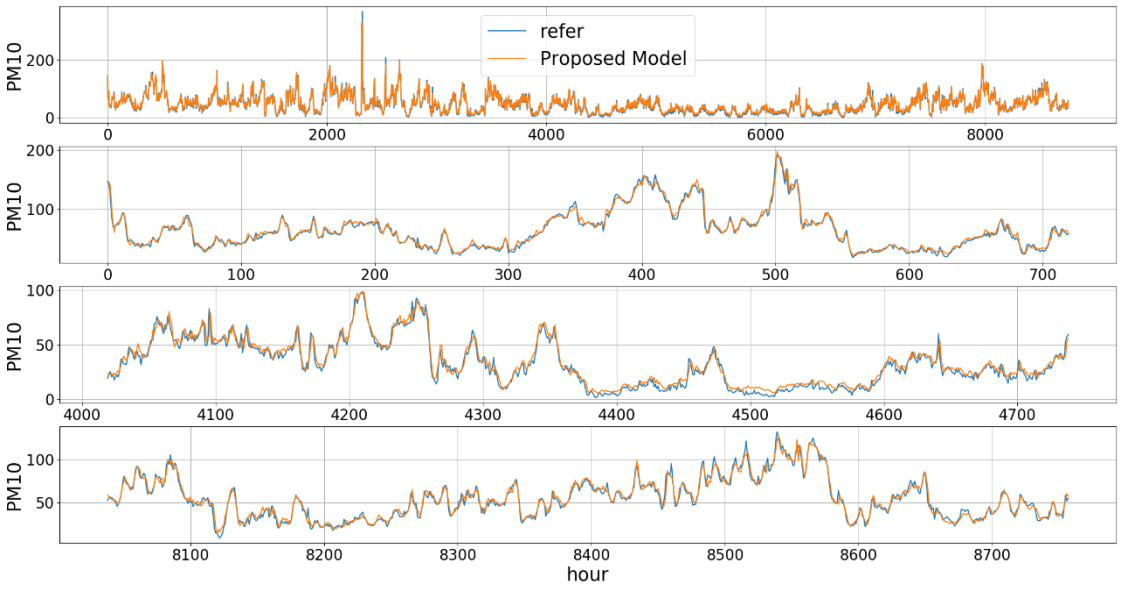


그림 14: 제안 모델의 1 시간 뒤 예측값과 참값의 비교 그래프

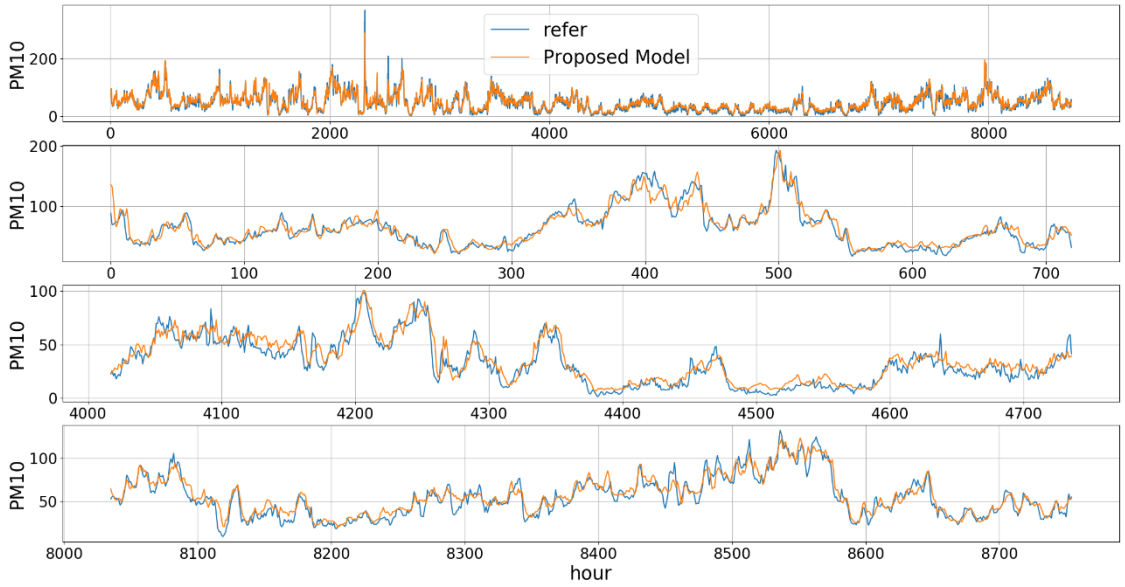


그림 15: 제안 모델의 4 시간 뒤 예측값과 참값의 비교 그래프

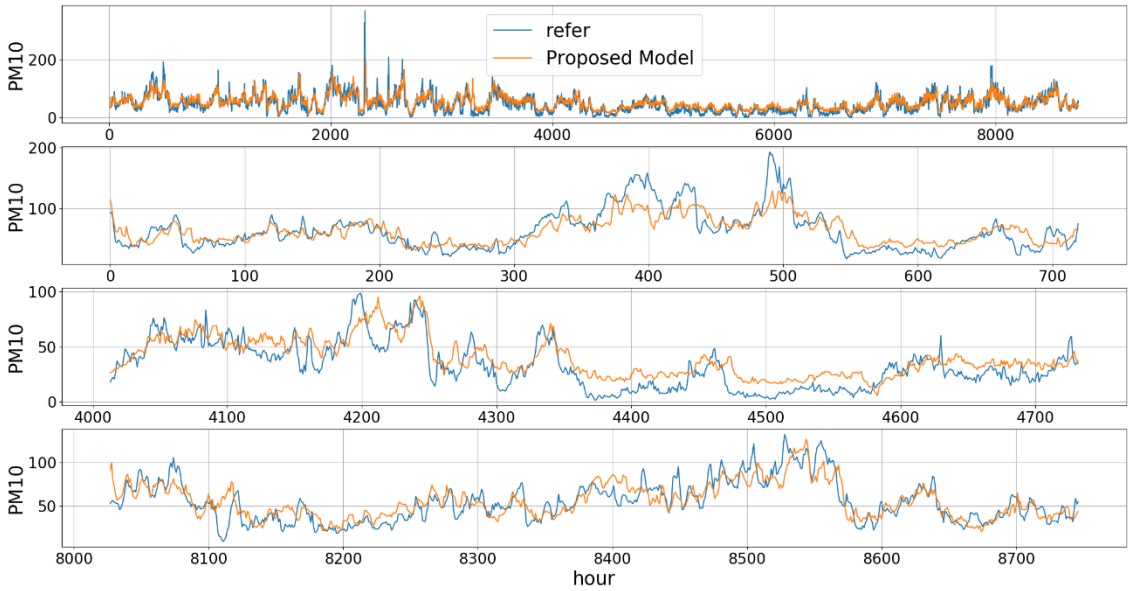


그림 16: 제안 모델의 12 시간 뒤 예측값과 참값의 비교 그래프

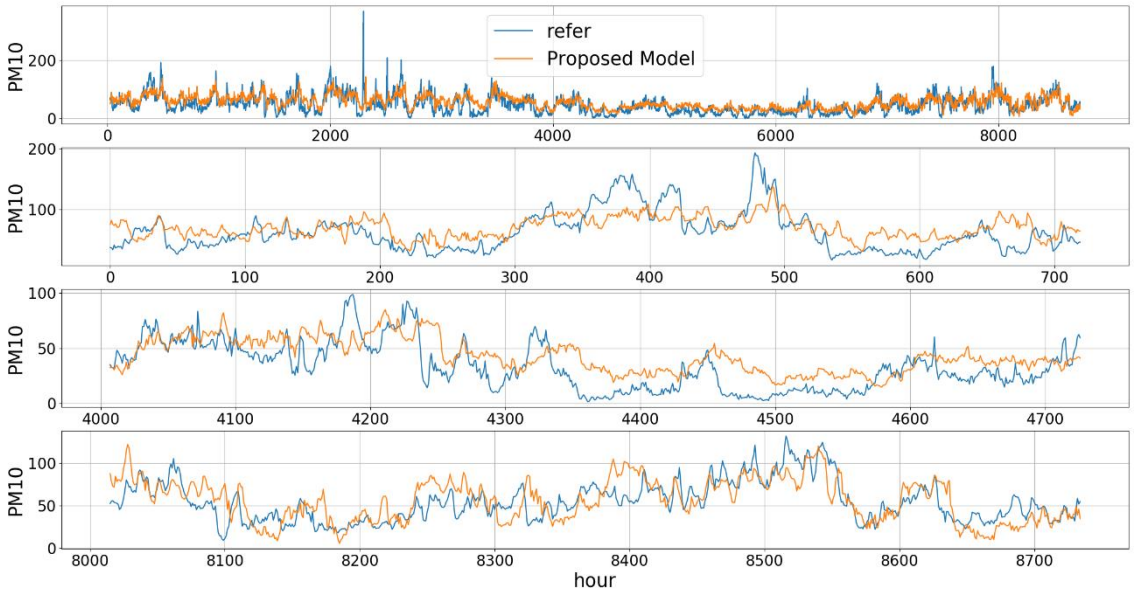


그림 17: 제안 모델의 24 시간 뒤 예측값과 참값의 비교 그래프

제6장 결론

본 연구에서는 국내 전역을 8x10 의 grid 로 나누고, 나뉘어진 구역을 공간해상도로 하고, 1 시간을 시간해상도로 하여 1 시간, 4 시간, 12 시간, 24 시간 뒤 미세먼지(PM10)농도를 ConvGRU 와 locally-connected layer 를 활용한 딥러닝 기반 모델을 통해 예측했다. 미세먼지 농도에 영향을 주는 다양한 요인을 고려하기 위해 입력 데이터로는 국내 오염원 데이터, 국내 기상데이터, 중국 미세먼지 데이터, temporal predictor & spatial predictor 를 사용하였다. 본 연구에서 제안한 모델에서 성능을 높이기 위해 가정하여 설계한 부분을 검증하고, 기존에 연구된 딥러닝 기반 미세먼지 예측 모델보다 더 나은 성능을 보이는지 확인하기 위해 5 가지의 실험 가설을 세우고 다양한 비교 모델을 만들어 결과를 분석하였다. 제 5 장의 실험결과를 토대로 가설을 검증하면 다음과 같다.

- 제안한 모델은 시공간정보까지 동시에 고려할 수 있기 때문에, 시간정보만 고려한 CNN-LSTM 모델보다 더 나은 성능을 보여준다는 가설 1 은 1 시간, 4 시간, 12 시간 뒤의 예측 결과를 기준으로 하면 모든 지표에서 T 모델이 같은 입력을 가진 CNN-ConvLSTM[1x1]보다 더 나은 결과를 보여주면서 만족한다. 단, 24 시간 뒤 예측 결과에 대해서는 일부 지표에 대해서는 더 낮은 결과를 보여줬기 때문에 만족한다고 보기 어렵다.
- ConvGRU 는 기존에 사용되는 ConvLSTM 보다 계산복잡성을 낮출 수 있으므로, 적은 training parameter 로도 비슷하거나 향상된 예측성능을 가질 것이라는 가설 2 는 가설 1 와 마찬가지로 1 시간, 4 시간, 12 시간 뒤의 예측결과를 기준으로 하면 T 모델이 ConvLSTM 모델보다 비슷하거나 나은 결과를 보여주기 때문에 만족하지만, 24 시간 뒤는 만족한다고 보기 어렵다.

- 제안된 모델에 포함된 Repeat 파트는 다음 1 시간부터 T−1 시간의 중간 예측과정을 고려할 수 있기 때문에 성능을 향상시킬 것이라는 가설 3 은 모든 예측 시간과 모든 결과 지표에서, 같은 입력을 가지는 T 모델이 T(simple)모델보다 더 나은 결과를 보여주었으므로 타당하다.
- 풍향, 풍속, 습도 및 주변 미세먼지 농도를 이용해 feature extraction 하면 국내의 미세먼지 확산요인을 잘 고려할 수 있기 때문에 예측 성능을 높일 수 있을 것이라는 가설 4 는 모든 예측 시간과 모든 지표에서, T+W 모델이 T 모델보다 더 나은 결과를 보여주었으므로 타당하다.
- 중국 미세먼지 데이터와 풍향 풍속 정보를 적절히 feature extraction 하면 국외 미세먼지 유입요인도 고려할 수 있기 때문에, 중장기간의 예측 성능을 높일 수 있을 것이라는 가설 5 는 12 시간 뒤와 24 시간 뒤를 중장기간의 예측이라고 생각했을 때 모든 지표에서 T+W+C 모델이 T+W 모델보다 더 나은 결과를 보여주었으므로 타당하다.

가설 1~5 를 종합하면, 미세먼지를 예측하기 위한 모델링은 1) 시간정보뿐만 아니라 시공간정보를 동시에 고려할 때, 2) 단시간 예측에 대해서는 계산 복잡성이 낮은 기법을, 장시간 예측에 대해서는 복잡성이 높은 기법을 사용할 때, 3) 다음 시간 T 를 예측하기 위해 다음 시간 T−1 까지의 중간 과정을 고려할 때, 4) 국내의 미세먼지 확산 요인을 잘 반영하도록 설계할 때, 5) 국외 미세먼지 유입요인 또한 잘 반영하도록 설계할 때 좋은 미세먼지 예측 성능을 얻을 수 있다는 결론을 얻을 수 있으며, 제안한 모델이 이를 만족함으로써 기존에 연구된 모델보다 더 나은 예측 성능을 가지고 있음을 보였다. 또한 단시간 예측 시 주로 이전 참값을 따라가는 delay shift 현상을, 장시간 예측 시 중간 값을 따라가려 하는 moving average 현상이 일어남을 예측값과 참값의 비교 그래프를 통해 확인하였고, 이를 해결한다면 예측 성능을 더 개선할 수 있다는 결론을 얻을 수 있다.

본 연구에서는 국내 전역의 미세먼지 예측을 목표로 했으므로 컴퓨팅 자원을 고려해 전역을 8x10 grid 로 나누었을 때의 cell 을 공간해상도로 했지만, 격자 cell 에 하나의 측정소만 포함되도록 하는 최소 공간해상도로 grid 를 나누어 측정소 단위 공간해상도로 예측이 가능하다.

또한, 본 연구에서 측정소를 포함하지 않는 cell 들을 보간하기 위해 IDW 를 사용하였으나, 보간을 위한 새로운 딥러닝 모델링을 만들어 좀 더 정확한 보간데이터를 확보함으로써 예측 성능을 높이는 것도 가능할 것으로 생각한다. 예를 들면, 먼저 측정소가 포함된 기준 cell 에 대해 인접한 cell 들의 데이터를 입력으로 활용하고 기준 cell 데이터를 라벨로 활용해 모델을 훈련한 뒤, 훈련한 모델로 측정소를 포함하지 않은 cell 의 데이터 값을 얻어내는 방법이 가능하다. 이에 관련된 후속 연구도 시공간적으로 정확한 미세먼지 예측을 위해 필요할 것으로 사료된다.

참고문헌

[1] "미세먼지상식." 서울시 미세먼지정보센터, 2019년 04월 01일 접속, https://bluesky.seoul.go.kr/finedust/common-sense/page/10?article=728.

[2] "Particulate Matter (PM) Basics." United States Environmental Protection Agency, 2019년 04월 01일 접속, https://www.epa.gov/pm-pollution/particulate-matter-pm-basics.

[3] "미세먼지 수치예보 정확도 고작 50%." 데일리환경, 2019년 04월 01일 접속, http://www.dailyt.co.kr/news/articleView.html?idxno=19356.

[4] Shahraiyni, H. T. and S. Sodoudi. (2016). Statistical Modeling Approaches for PM10 Prediction in Urban Areas. A Review of 21St-Century Studies, Atmosphere, 7(2), 15-38.

[5] 전병일. (2012), "부산지역 겨울철 고농도 미세먼지 발생일의 기상학적 특성. 한국환경과학회지", 21(7), 815-824

[6] 박순애, 신현재. (2017), "한국의 초미세먼지(PM2.5)의 영향요인 분석. 환경정책", 25(1), 227-248.

[7] 김동혁, 이상신, 권지수. 고농도 미세먼지 발생 시 기상학적 특성 파악에 대한 연구. 충남연구원 서해안기후환경연구소, 2017.

[8] 김진형, 강성원. (2018), "국내 미세먼지 오염도에 영향을 미치는 요인에 대한 분석. 한국환경경제학회 하계학술대회논문집 ", 2018s(0), 779-791.

[9] Saide, P.E., Carmichael, G.R., Spak, S.N., Gallardo, L., Osses, A.E., Mena-Carrasco, M.A., Pagowski, M. (2011). Forecasting urban PM10 and PM2. 5 pollution episodes in very stable nocturnal conditions and complex terrain using WRFeChem CO tracer model. Atmospheric Environment, 45, 2769-2780.

[10] Chen, J., Lu, J., Avise, J. C., Damassa, J. A., Kleeman, M. J., & Kaduwela,

A. P. (2014). Seasonal modeling of PM2.5 in California’s San Joaquin Valley. Atmospheric Environment, 92, 182-190.

[11] 김성태, 구윤서. (2015), “미세먼지 앙상블 예보기법 개발. 한국도시환경학회지”, 15(3), 251-260.

[12] 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

[13] Li, C., Hsu, N. C., & Tsay, S. C. (2011). A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmospheric Environment, 45(22), 3663-3675

[14] Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. (1997). Support vector regression machine. Advances in Neural Information Processing Systems, 9, 155-161.

[15] García Nieto, P. J., Combarro, E. F., Del Coz Díaz, J. J., & Montañés, E. (2013). A SVM-based regression model to study the air quality at local scale in. Oviedo urban area (Northern Spain): A case study. Applied Mathematics and Computation, 219(17), 8923-8937.

[16] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution, 231, 997-1004.

[17] Hochreiter, S., Schmidhuber, J., (1997). Long short-term memory. Neural Comput. 9, 1735-1780.

[18] Huang, C. J., & Kuo, P. H. (2018). A deep cnn-lstm model for particulate matter (Pm2.5) forecasting in smart cities. Sensors (Switzerland), 18(7).

[19] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, In arXiv preprint arXiv:1506.04214.

[20] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In arXiv preprint arXiv:1406.1078.

[21] Ballas, N., Yao, L., Pal, C., & Courville, A. (2015). Delving Deeper into Convolutional Networks for Learning Video Representations. In arXiv preprint arXiv :1511.06432.

[22] Pang, L., Lan, Y., Xu, J., Guo, J., & Cheng, X. (2017). Locally Smoothed Neural Networks, (1989), In arXiv preprint arXiv:1711.08132.

[23] Guocai, Z.. (2004). Progress of weather research and forecast (WRF) model and application in the United States. Meteorological Monthly. Mon. 12, 5

[24] Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., & Lin, S. (2013). A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Annals of the Photogrammetry. Remote Sensing and Spatial Information Sciences, 4(4W2), 15-22.

[25] "최종확정 측정자료 조회." 에어코리아, 2019년 05월 01일 접속, https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123.

[26] "종관기상관측." 기상자료개방포털, 2019년 05월 01일 접속, https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36.

[27] "Beijing - PM2.5." U.S. Department of State Air Quality Monitoring Program, 2019년 05월 01일 접속, http://www.stateair.net/web/post/1/1.html.

ABSTRACT

Research of Particulate Matter Prediction Modeling Based on Deep Learning

Lee, Seong Gu
Department of Human ICT Convergence
Sungkyunkwan University

Particulate matter (PM) is the term for a very small mixture of solid particles and liquid droplets found in the air. This term includes PM10 whose diameters are generally 10 micrometers and smaller; and PM2.5 whose diameters are generally 2.5 micrometers and smaller.

Recently, people and government get more awareness of the risk of PM. Also, the importance of PM prediction system to prevent the damage caused by PM is emerging. However, the prediction system in Korea is based on deterministic method, which is known to have lower performance than the recent deep learning based method. Nevertheless, the study of deep learning based PM prediction modeling in Korea is still insufficient compared to the international studies.

Therefore, this study aims at developing a deep learning model using China PM data, domestic meteorological data and domestic pollution data to consider the domestic environment. To do this, we propose the PM prediction modeling based on deep learning by using ConvGRU which can simultaneously analyze spatiotemporal information, and using a locally-connected layer which

can better extract features of individual fields.

Experiments are designed to predict the PM10 of next 1 hour, 4 hours, 12 hours, and 24 hours with the spatial resolution divided by the 8x10 grid of all regions in Korea. In order to verify the performance of the proposed model, we make five experimental hypotheses, which confirm that the proposed model is better than the other deep-learning based prediction model.

In the result, the prediction performance got better 1) when it analyzes spatiotemporal information simultaneously, 2) when it has low computational complexity for short-term prediction; and it has high complexity for long-term prediction, 3) when it considers the intermediate process up to the next 1 hour to predict the next T-1 hour, 4) when it considers the factors of PM diffusion in Korea 5) when it considers the factors of China PM. So, the proposed model showed the better prediction performance than the previously studied models.

Also, the result showed the delay shift phenomenon in the short-term prediction, and showed the moving average in the long-term prediction. So, we can conclude that the prediction performance can be improved if those phenomenon are solved.

Keywords: Particulate matter, Deep learning, Time-series prediction, ConvGRU, Locally-connected layer

碩士學位請求論文

딥러닝 기반 국내 미세먼지 예측 모델링 연구

2019

李聖求