<div align="center">

|  |  |  |
| --- | --- | --- |
| **2018** | **Report – 2018/ 12월 1주** | **Seong-gu** |

</div>

**I. Plan for this week**

***Personal Research Subject***

* 현재 모델링이 어떻게 프로세싱되는지 구체적으로 알아볼 것.
* 현재 결과값을 잘 정리하는 방법을 기존 논문에서 조사할 것.

**II. Plan for next week**

* 모델링 방법 구체적으로 학습 후 개선한 모델링으로 결과값 도출 및 정리.

**III. Response on Prof. Shin's feedback last week**

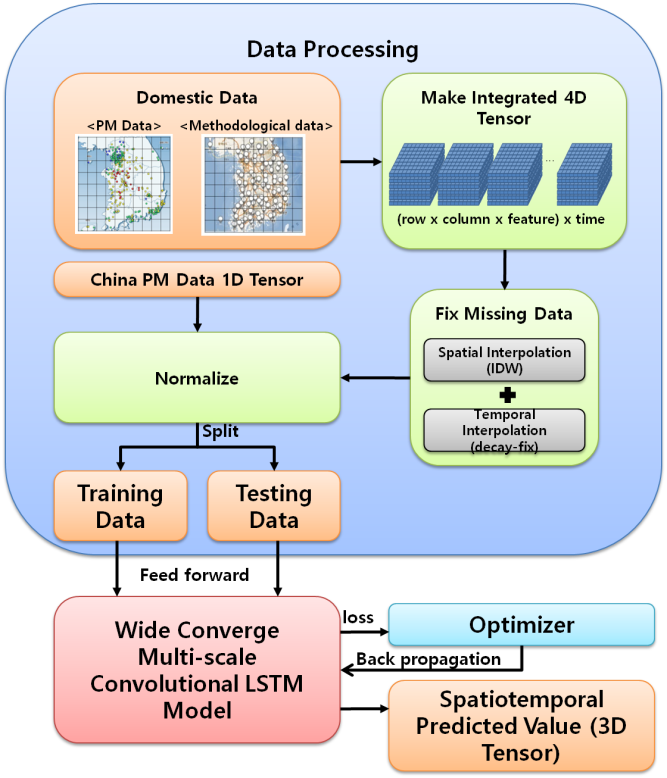
* 모델별로 구조 검토해보고, 가장 미세먼지 농도 예측에 나은 모델링 선정
* 결과로 보여준 median, min, max는 의미가 없어 보임.
* 결과를 잘 보여주는 방법에 대해 검토 (소수점 둘째자리까지)
* 모델링에 대해 자세히 검토할 것

**IV. Milestone**

1. 실험실 환경에서 미세먼지 농도 예측 시스템 구축   
   1. 머신러닝 기법을 활용한 미세먼지 농도 예측  
        
      **- 한국의 지리학적 특성 및 계절풍의 영향을 고려하여, 활용할 측정소 범위를 국내 전역으로 하고, 중국의 미세먼지 농도를 모델링 데이터로 사용함.**(중국데이터를 확보하는 방법 조사할 것)

**- 결측치는 공간적 보간법인 IDW 기법과 시간적 보간법인 decay-fix 기법을 결합한 시공간 보간법을 사용하여 추정함.**

**- 모델링은 미세먼지 유입의 공간적인 특성을 적극 반영하기 위해서 최근에 연구되고 있는 Convolutional LSTM 모델을 사용.**(모델링 최적화 필요)



**V. Reports**

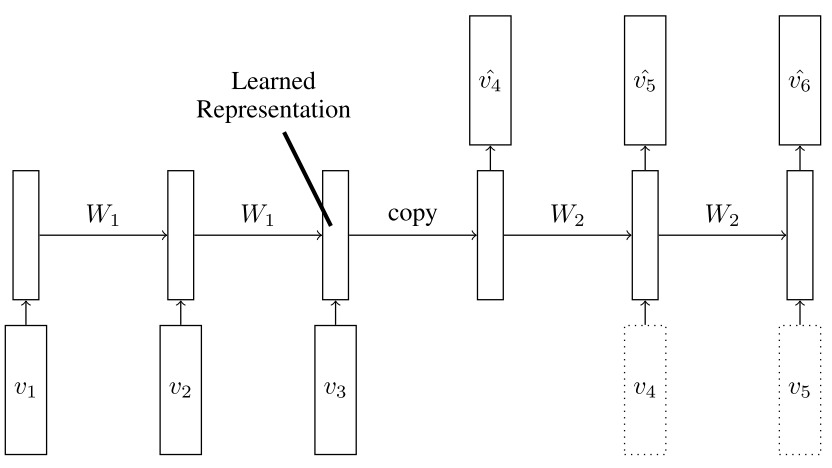
* 개인연구 진행사항
* **모델링 관련**

**1. 모델 encoder – decoder 학습 관련  
:** series input을 받아, series output을 낼 수 있는 모델로 input 길이와 output 길이는 다를 수 있음.

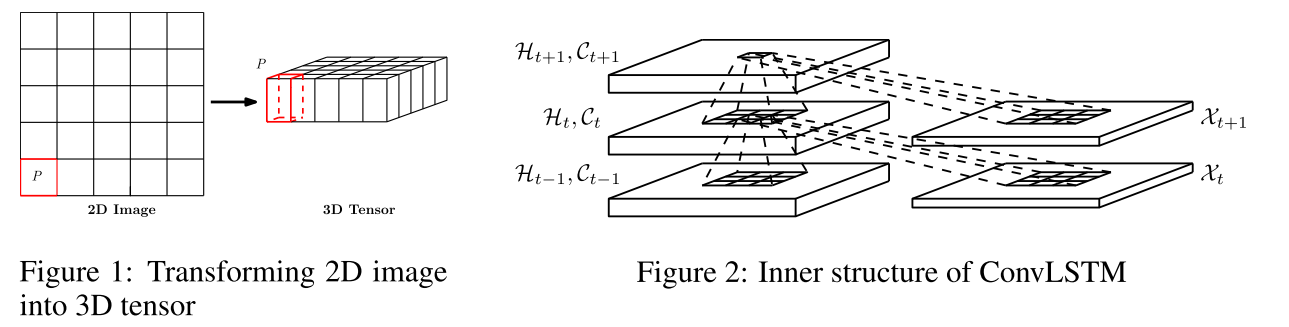
<https://chunml.github.io/ChunML.github.io/project/Sequence-To-Sequence/>

<https://www.google.co.kr/url?sa=i&source=images&cd=&ved=2ahUKEwjq-Lm3tPbeAhXCiLwKHUvXD1wQjxx6BAgBEAI&url=https%3A%2F%2Fstackoverflow.com%2Fquestions%2F51749404%2Fhow-to-connect-lstm-layers-in-keras-repeatvector-or-return-sequence-true&psig=AOvVaw1M8Bk6jvHIDzfJaXY5wgv1&ust=1543471187265562>

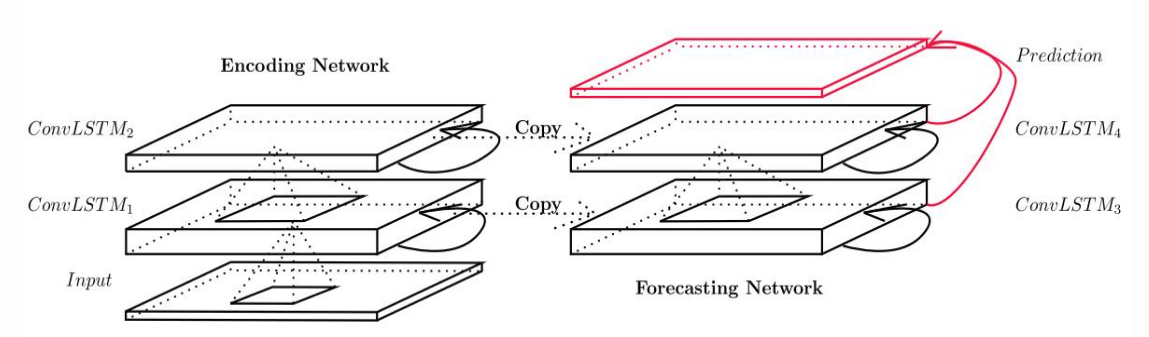
repeat vector를 사용한 예시



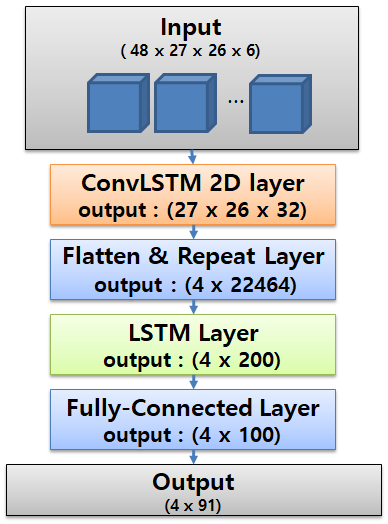
**Convolutional LSTM 기본 구조**



**Encoder-decoder Convolutional LSTM 모델 구조.**



**현재 모델**



차후 모델은 Decoder로 쓰인 LSTM 대신에 ConvLSTM 을 넣어서 결과값을 확인할 예정입니다.

**2. 논문 리뷰 (결과값 정리 및 모델링 방법에 대한 survey)**Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., & Chi, T. (2019). A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Science of The Total Environment, 654, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086

의 논문 리뷰

**논문의 목적**

convolutional long short-term neural network을 이용해 spatiotemporal한 feature를 뽑아내서 미세먼지 농도를 더 정확하게 예측하는 것을 목적으로 합니다.

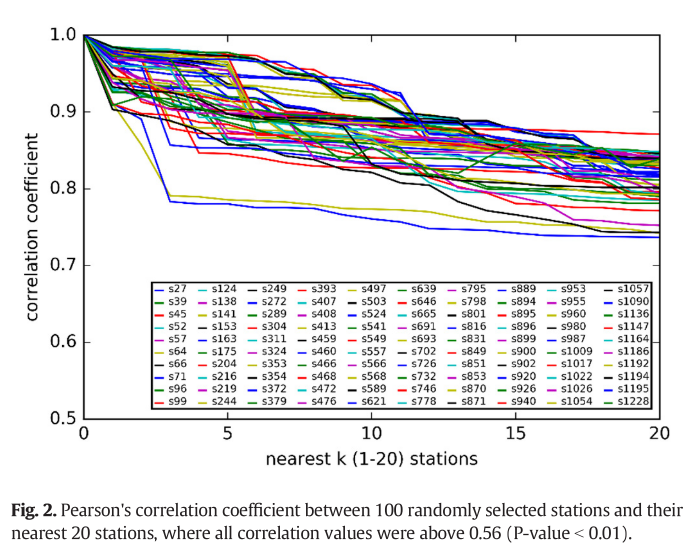
단, 본 논문에서 쓰인 모델은 3D CNN과 LSTM 레이어가 하나의 레이어로 결합된 것이 아닌, 개별적으로 쌓여있는 모델이라 일반적으로 얘기되는Convolutional LSTM 모델과는 다릅니다.

**데이터 셋**

미세먼지 데이터 셋으로는 두 번의 실험을 진행하였는데, 첫 번째는 large scale로 1233개의 중국 전역 측정소 데이터를, 두 번째로는 small scale로 12개의 베이징 측정소를 사용했습니다.

추가적으로 기상데이터 (hourly humidity, temperature and wind speed), 및 planetary boundary layer height (PBLH), and aerosol optical depth (AOD)를 사용했습니다. 이 추가 데이터가 미세먼지 농도와 관련있다는 것은 reference논문들((Chu et al., 2003; Engel-Cox et al., 2004; P. Gupta et al., 2006; Koelemeijer et al., 2006; Díaz-Robles et al., 2008; Saide et al., 2011). These)을 통해 제시하였습니다.

본 논문에서 가장 중요하게 생각하고 있는, 공간적인 특성을 뽑아내기 위해서 먼저 각각의 측정소에 대해 k번째로 가까운 측정소에서 측정한 pearson’s corrleation값을 그래프로 나타내었습니다. 이 그래프는 주변 측정소 데이터를 타겟 측정소 데이터를 예측하는 데 쓰이는 것이 타당함을 증명하는데 쓰였습니다..



**모델**

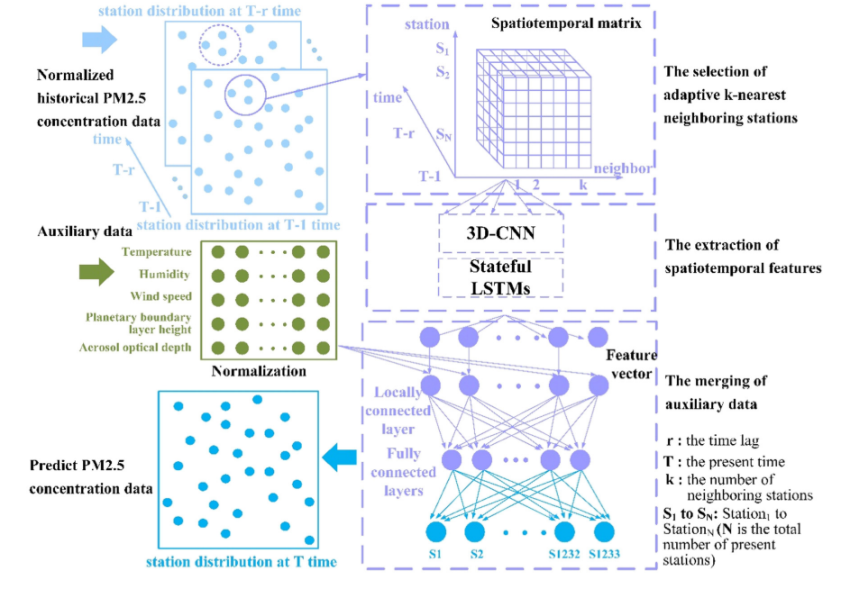
모델은 아래와 같으며 현재시간 T일 때 모든 station의 미세먼지 농도를 예측합니다. 크게 input part, modeling part, output part로 나뉘며 세부적으로는 다음과 같이 나뉩니다.

Input : 과거 T-r부터 T-1시간 까지 normalized된 미세먼지 농도 데이터 와 기상데이터 등 추가 데이터 두 가지

Model :

The selection of adaptive k-nearest neighboring stations - 미세먼지 농도 데이터를 3차원 matrix (input time-step, target station , neighbor station data)에 mapping 함. k값은 타겟 측정소마다 다르며, 일정 correlation threshold 값을 정하고, 이 값 이상만을 가지는 주변측정소 (=k개) 만을 모델링 데이터로 활용하는데 씀.  
The extraction of spatiotemporal features – basic CNN(k\*1\*r kernel size CNN layer + ReLU)뒤에 Stateful LSTM(1000개의 노드를 가진 LSTM layer 2개)을 적층, spatiotemporal한 feature를 뽑아냄. (stateful LSTM이란, 각 batch에서 cell state가 다음 batch의 cell state로 유지되는 LSTM으로 long sequence에 적합함)

The meriging of auxiliary data : CNN+LSTM 모델에서 뽑아낸 1D Feature vector와 추가 데이터를 Locally connected layer의 input으로 활용하여 미세먼지 농도를 예측



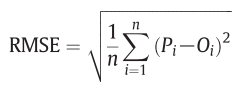
**결과값 정리**

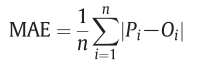
결과값은 다음 3개의 indicator를 주로 사용합니다.

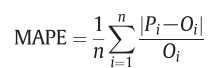
RMSE(Root Mean Square Error)

MAE(Mean Absolute Error)

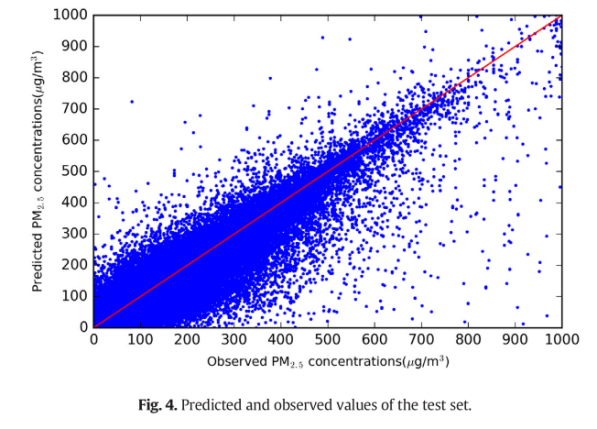
MAPE(Mean Absolute Percentage Error)







**1. 측정값-실제값 비교**

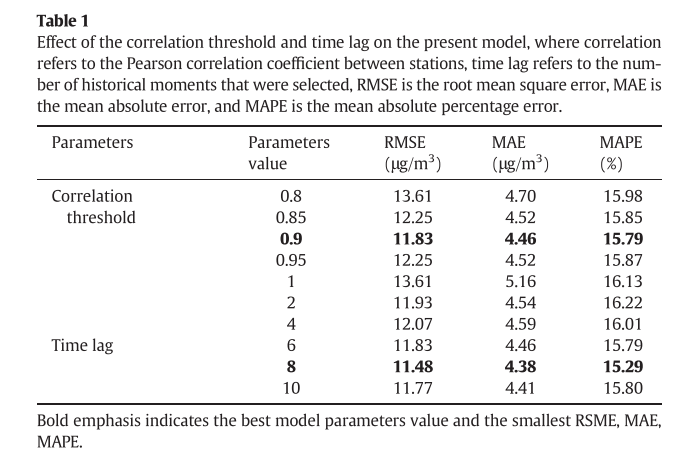


R^2 (coefficient of determination) = 0.92

**2. 라벨링(classification) 정확도**

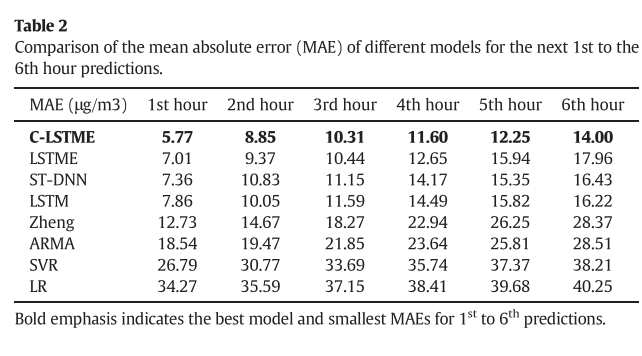
6개의 rank로 이루어져 있으며, 정확도는 87.6%, 다른 모델과 비교는 이루어지지 않음.

**3. correlation threshold와 Time lag 최적값을 정하기 위한 비교표**



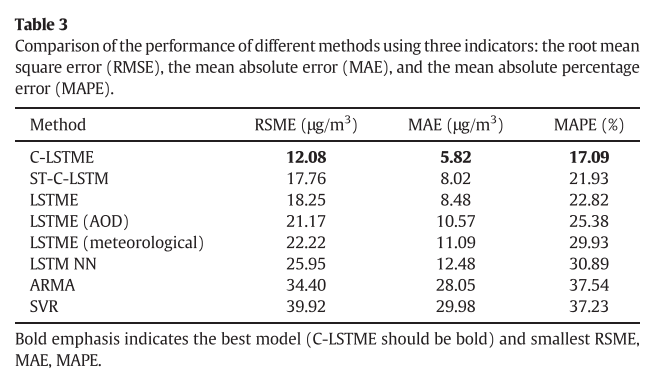
**4. 다른 모델과의 비교 (large scale - 중국 전역의 1233개 측정소 데이터를 사용)**

가장 낮은MAE를 기록한 측정소의 결과값만을 보여주었습니다. 본 논문에서 제시한 모델이 가장 낮은 오차값을 보였습니다.  
(LR -Linear regression model, the support vector regression (SVR) model, the autoregressive moving average (ARMA) model, and the Zheng model (Zheng et al., 2015),  ST-DNN(Space-Time Delay Neural Network)



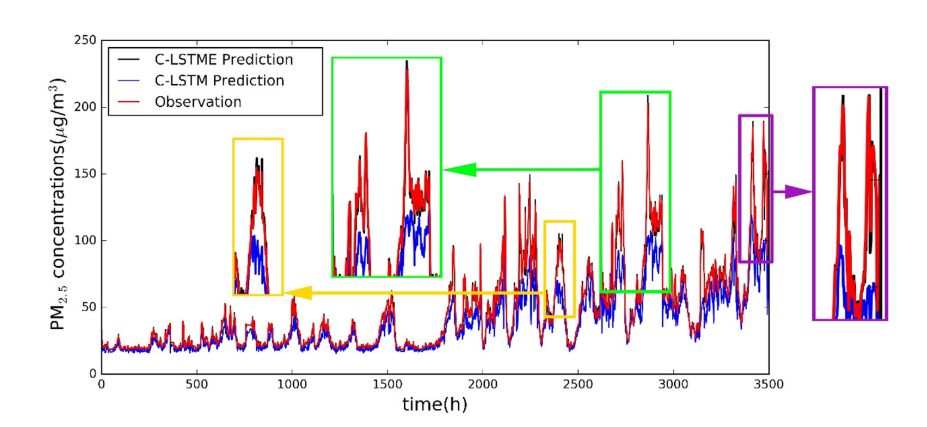
**5. 다른 모델과의 비교 2 (small scale - 베이징시의 12개 측정소 데이터를 사용, 1시간 뒤 예측만 해당)**

가장 낮은RSME, MAE, MAPE를 기록한 측정소의 결과값을 사용했습니다. 마찬가지로 본 논문에서 제시한 모델이 가장 낮은 오차값을 보였습니다.



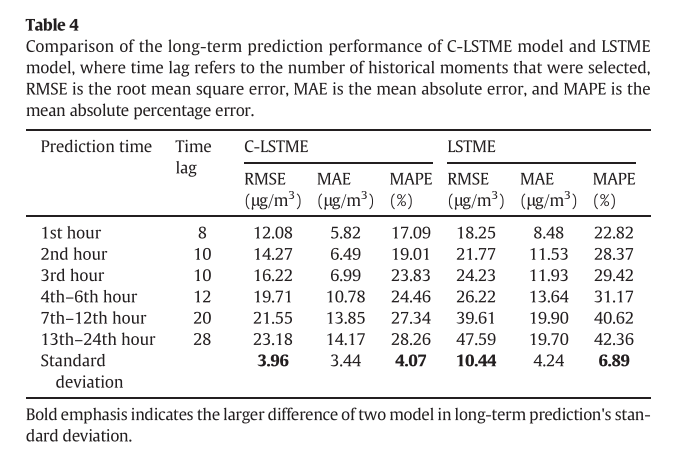
**6. 추가데이터를 사용한 모델(C-LSTME)과 사용하지 않은 모델(C-LSTM) 및 실제값과의 비교 그래프**

임의로 정한 측정소의 결과값을 활용했습니다. 추가데이터를 사용했을 때, 예측값이 실제값을 좀 더 잘 따라갑니다. (검은색 선)



**7.** long-term 예측에 따른 모델링 결과값 비교

추가데이터를 활용하되, LSTM만 사용한 모델과 본 논문에서 제시한 모델과 결과값을 비교했습니다. 모든 예측시간 간격에서 오차값이 본논문에서 제시된 모델에서 더 적게 나왔으며, 예측 시간 간격이 길어질수록 제시한 논문의 성능이 더 두드러지게 나타났습니다.



**결론**

* 본 논문에서는 타겟 측정소와 주변 측정소와의 미세먼지 측정값에 대한correlation을 활용해, 일정 수준 이상의 correlation을 가진 주변 측정소만을 선별하였고, 이 주변 측정소 데이터를 CNN을 통해 feature를 뽑아내어 모델 성능을 높이는 걸 main contribution으로 하였습니다.
* 본 모델은 시계열 데이터 예측에서 기존에 제시되는 non – deep learning 방식인 SVR, ARMA 및 deep-learning 방식인 LSTM , Zheng model (Zheng et al., 2015), DNN 보다 좋은 성능을 나타냈습니다**.**
* 본 모델에서 사용한 Convolutional LSTM은 저희쪽에서 사용하는 Convolutional LSTM 방식과 다르지만, 결과값 정리 (실제값과 예측값간의 R^2 (coefficient of determination) 도출, RMSE, MAE, MAPE를 evaluation index로 사용, 다른 모델과 비교)방법을 논문에 적용하면 좋을 것으로 보입니다.