<div align="center">

|  |  |  |
| --- | --- | --- |
| **2019** | **Report – 2019/ 1월 5주** | **Seong-gu** |

</div>

**I. Plan for this week**

***Personal Research Subject***

* 미세먼지 예측 머신러닝 기반 모델링 최신 논문 계속 검토할 것.
* 바람의 영향에 의한 미세먼지 유입을 최적 고려한 모델링 찾아볼 것.

**II. Plan for next week**

* 미세먼지 예측 머신러닝 기반 모델링 최신 논문 계속 검토할 것.
* 바람의 영향에 의한 미세먼지 유입을 최적 고려한 기초 모델링 고안.

(locally-connected 2d lstm)

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

* 현재는 시간에 따른 예측이지만, 누락된 공간적인 데이터도 잘 예측하는 방법 검토. (Unsupervised learning이 될텐데 방법을 찾아보자.)
* 랜드마크 등 지형, 지역적 특성 수기 추가
* Feature를 잘 선택 (데이터 프로세싱으로 정확도 개선)
* 지역은 한국을 전범위로 하지 말고 적당한 수준으로 결정함.
* 시간대 설정
* 효율성 및 정확도 면에서 트레이드 오프 분석 설계 방안을 제기
* 온도, 습도, 대기압, 바람세기, 풍향 데이터 외에 날씨(눈/비/안개 등)도 포함하면 좋을듯함

**IV. Milestone**

* 시공간 데이터 보간 기법 활용 (forward + IDW)
* Locally-connected 2D LSTM 기초 모델 개발
* Locally-connected 2D LSTM 다층 구조 설계
* Attention 구조를 결합한 모델링도 고려.
* 연관관계가 검증된 추가 input feature (2차생성을 일으키는 SO2나 NOx) 활용

: forward-fixed 기법과 IDW 기법에 대해 가중치를

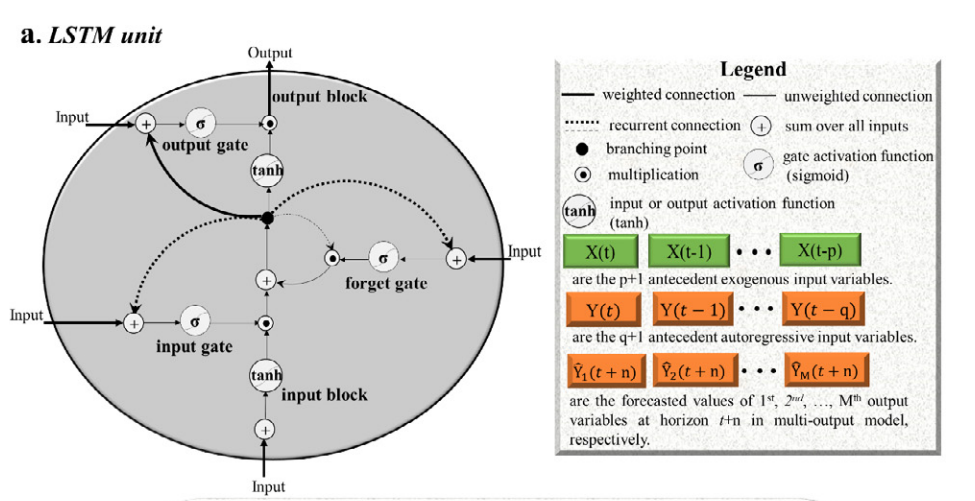
**V. Reports**

* 개인연구 진행사항
* **Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts 리뷰**

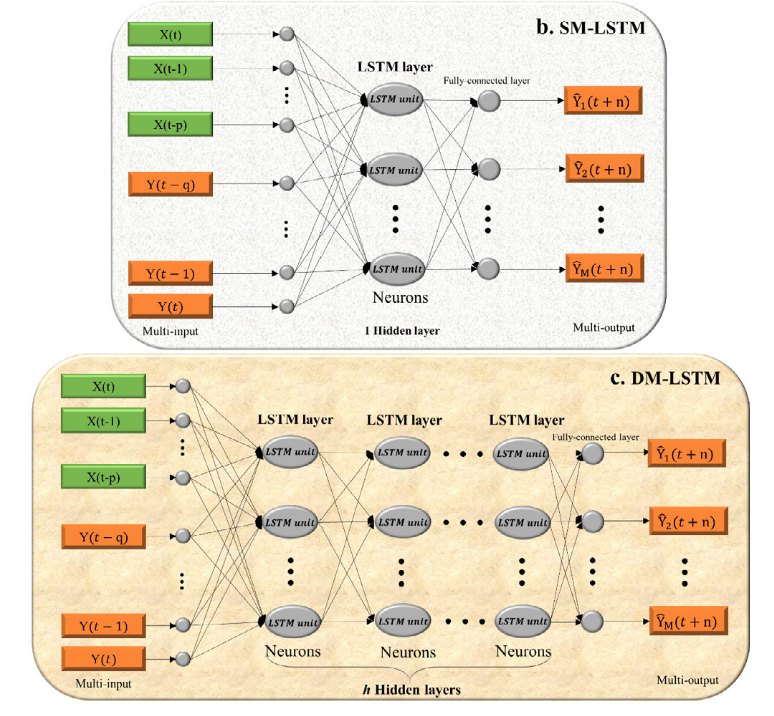
본 논문은 다변수 시계열 공기질 데이터 예측을 위해 a Deep Multi-output LSTM (DM-LSTM)를 사용하였습니다. LSTM층을 단층이 아닌 다층으로 구현하고, overfitting 및 결과값에 대한 instability를 줄이기 위해 머신러닝에서 사용되는 3가지 기법을 추가로 사용하였습니다.

* **모델링 및 트레이닝 방법**

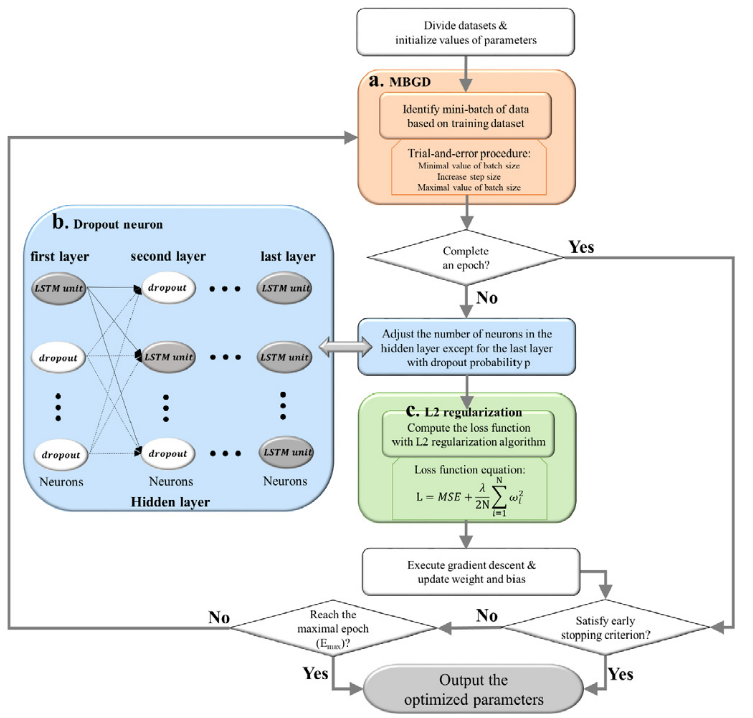
사용된 LSTM unit은 아래와 같이 일반적인 unit과 같습니다.



비교 모델로 단층 LSTM 모델인 Shallow Multi-output LSTM과 다층모델인 DM-LSTM은 아래와 같습니다.



overfitting 및 결과값에 대한 instability를 줄이기 위해 머신러닝에서 사용되는 3가지 기법을 사용한 트레이닝 프로세스는 아래와 같습니다.



1. MBGD (Mini-Batch Gradient Decent)

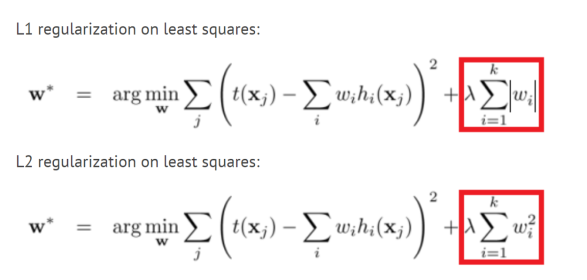
: 한 epoch에 대해 training dataset을 mini-batch size로 나누어서 학습을 시키는 것으로, 파라미터를 training sample 하나마다 업데이트를 하는 Stochastic Gradient Descent(SGD)의 연산이 빠른 장점과 전체 training sample를 이용해 한번에 파라미터를 업데이트하는 Batch Gradient Descent의 안정적으로 비용함수를 수렴시킬 수 있는 장점을 결합한 파라미터 최적화 기법입니다.

2. Dropout

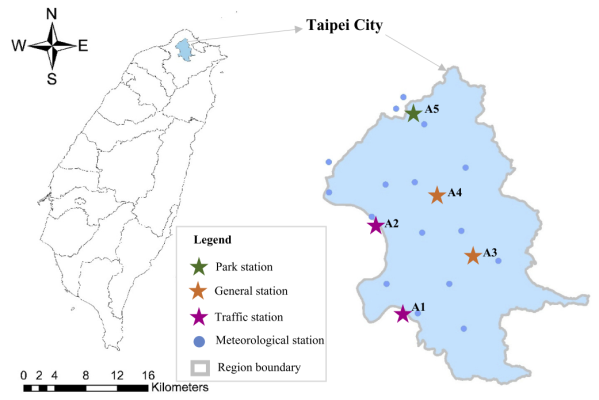
: 확률적으로 일부의 neuron들을 사용하지 않는 것으로, 마치 다수의 모델을 학습시키는 것과 비슷한 효과를 내면서overfitting을 막는 기법입니다.

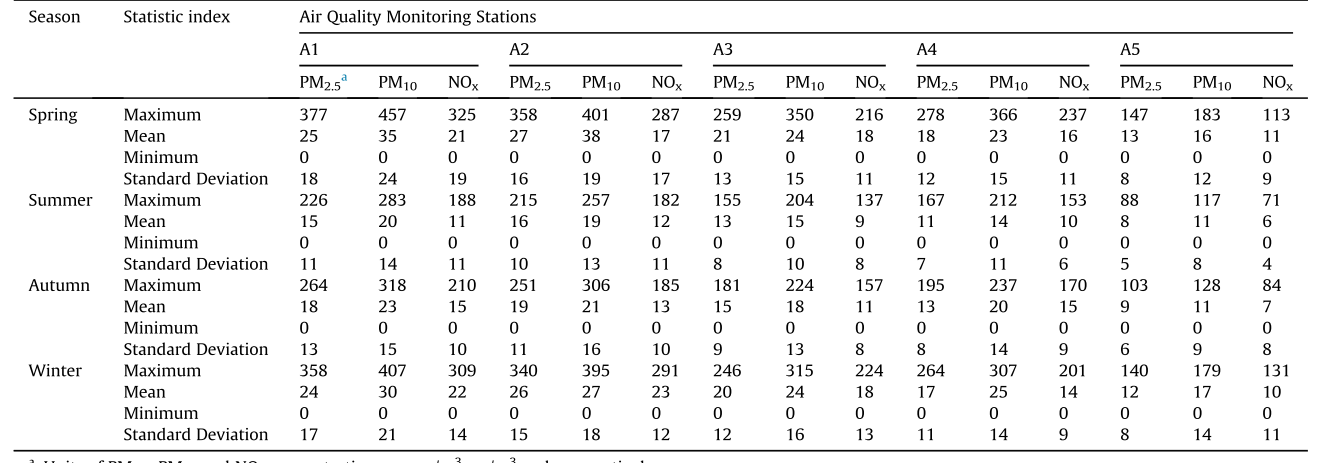
3. L2 regularization

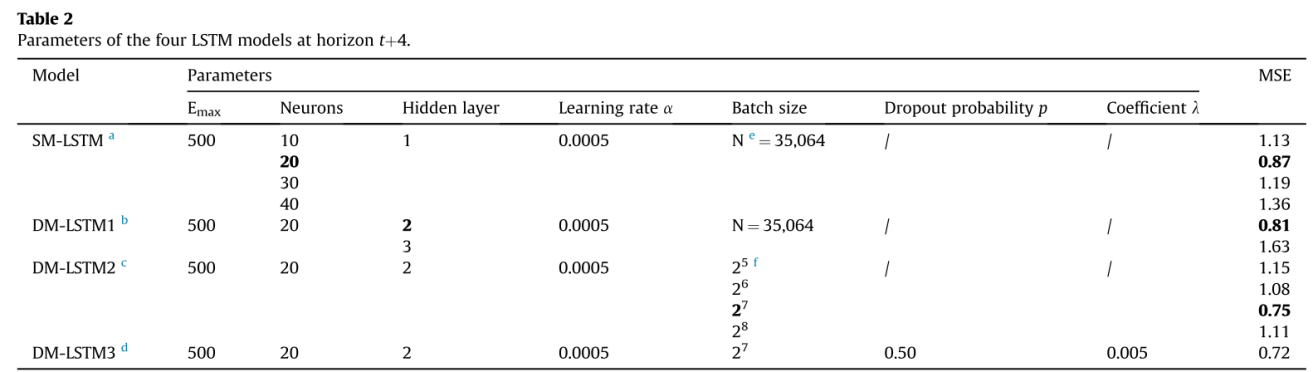
: 손실함수를 줄이는 과정에서 weight 또한 감소시키는 방향으로 줄여가는 기법으로, 특정 weight가 학습 시에 비이상적으로 커지는 것을 방지하여 overfitting을 막는 기법입니다.



* **사용된 데이터 & 결과**

2010 년부터 2016 년까지의 7개년 데이터에서 8개의 공기질 데이터 (PM2.5, PM10, O3, NOx, NO2, NO, SO2, CO)와 5개의 기상청데이터(rainfall, temperature, wind speed, wind direction, and relative humidity)를 5개의 측정소로부터 수집하였습니다. 측정소는 특성이 좀 다른데, A5는 인적, 교통량이 없는 공원에 있는 측정소, A3, A4는 일반적인 측정소, A1, A2는 도로변에 있는 측정소입니다.





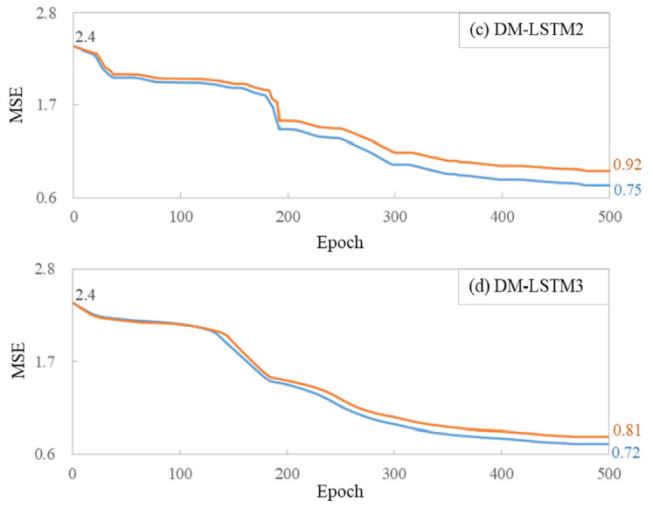
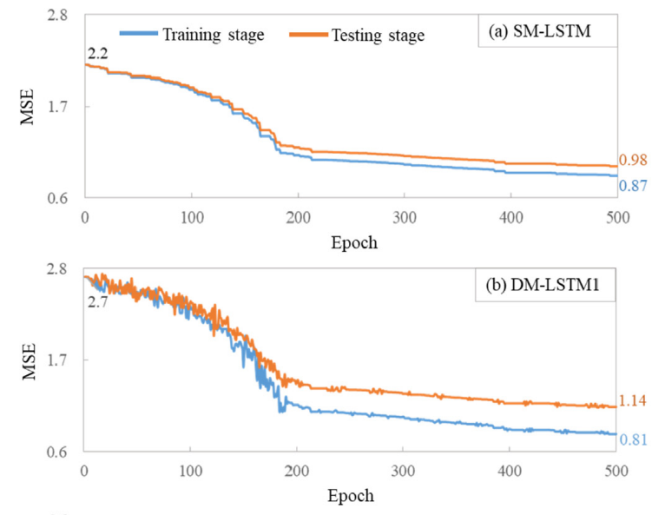
결과는 위 표와 같으며, MSE는 에 의해 표준화된 값에 대해 계산되었습니다.

제안한 DM-LSTM은 3가지 모델로 나뉘었으며,

DM-LSTM1 : LSTM 층만 다층인 모델

DM-LSTM2 : DM-LSTM1에 MBGD를 적용한 모델

DM-LSTM3: DM-LSTM2에 dropout과 L2 regularization 을 적용한 모델 입니다.

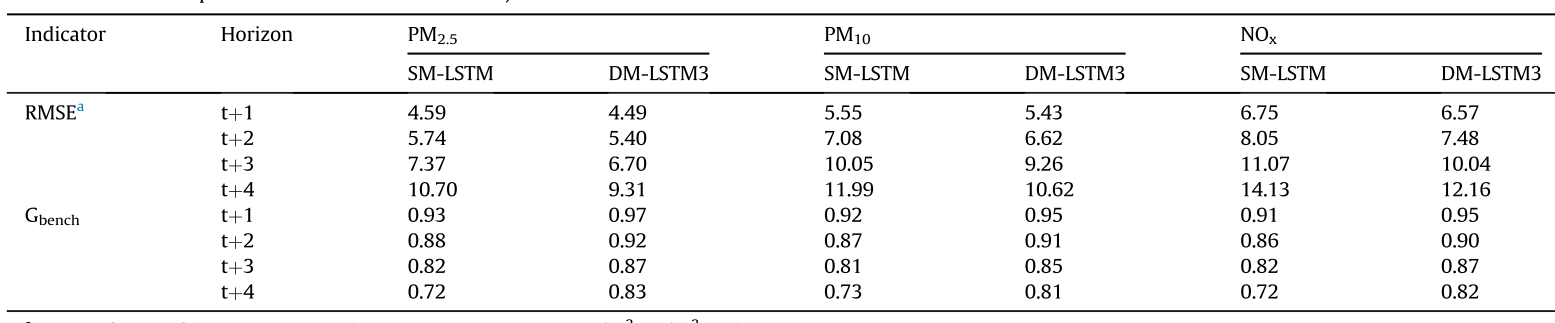


위 그래프는 epoch에 따른 training, testing MSE값입니다. SM-LSTM은 baseline이 됩니다.

LSTM layer만 다층으로 쌓은 DM-LSTM1은 SM-LSTM과 비교하여 training error는 적지만, testing error는 오히려 높은 것을 보아 over fitting이 발생했음을 알 수 있습니다. 또한 200epoch까지 error값이 진동하는데 이는 모델의 stablity가 떨어짐을 나타냅니다.

이 결과는 LSTM 층을 깊게만 만든다고 해서 성능개선으로 이어지지는 않음을 보여줍니다.

MBGD를 적용한 DM-LSTM2은 DM-LSTM1과 비교하여, training, testing error 모두 월등히 나은 결과를 보여주었고, error값이 진동하는 부분이 없어져서 stability를 개선했습니다.

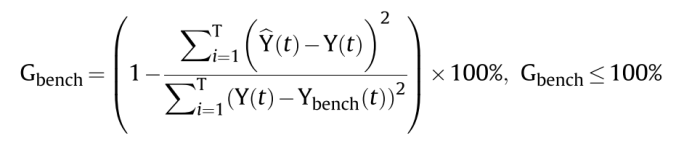
MBGD, dropout, L2 regularization 를 모두 적용한 DM-LSTM3은 DM-LSTM2과 비교하여 training, testing error 모두 좀 더 나은 결과를 보여주었으며, training error와 testing error의 차이도 좁혀줌으로써 overfitting을 개선했습니다.  


위 표는 공기질(PM2.5, PM10, NOx)에 대한 실질적인 RMSE값 및 Gbench값을 제시하였으며

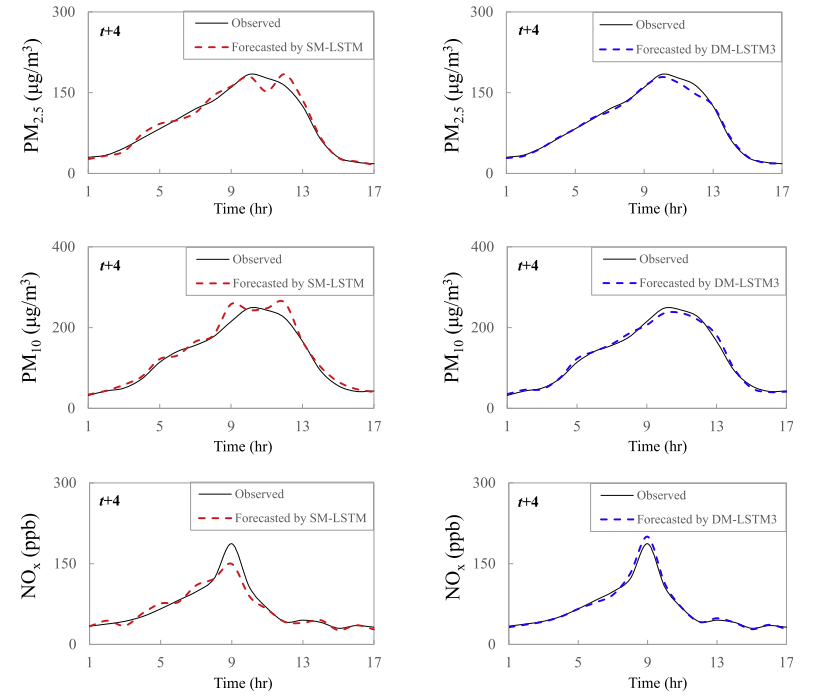
예측 시간이 길어질수록 SM-LSTM과 DM-LSTM3의 error값 격차가 늘어남을 알 수 있습니다.

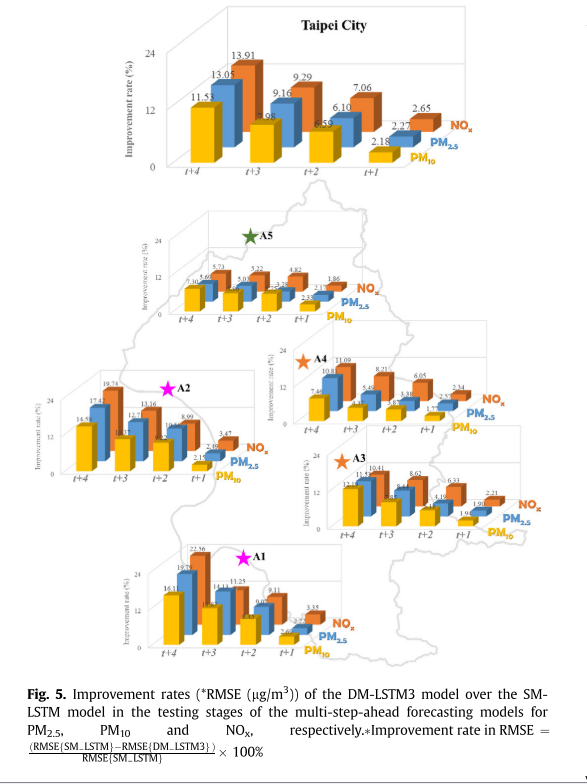
즉, 제시된 결과로써 제안한 모델은 장기 예측에 도움이 될 수 있음을 알 수 있습니다.

Gbench(the goodness-of-fit with respect to the benchmark)값은 아래와 같이 계산됩니다.   
() = Y(t-n))



아래는 비교적 예측이 어려운 peak값에서의 미세먼지농도 예측값과 실제값의 비교 곡선으로, DM-LSTM3의 예측력이 SM-LSTM보다 좋음을 보여주고 있습니다.





위의 그림은 주변 환경이 다른 측정소별로 SM-LSTM에 비해 DM-LSTM3를 썼을 때 얼마나 예측성능이 개선되었는지를 보여주는 그림입니다.

가장 공기질이 좋고 변화가 적은, 공원에 위치하는A5보다 공기질이 나쁘고 변화가 심한, 도로변에 있는 A1,A2에서 예측성능이 높게 나왔음을 알 수 있습니다.

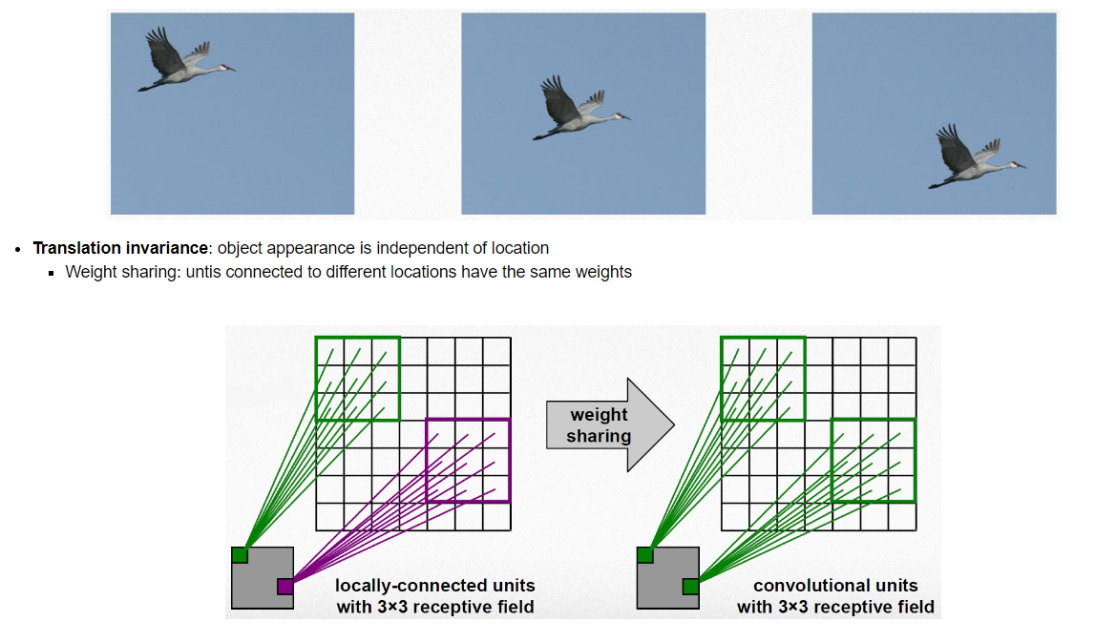
즉, 제안한 모델(DM-LSTM3)은 baseline으로 쓰인 SM-LSTM에 비해 예측이 어려운 측정소의 데이터를 더 잘 예측한다고 결론지을 수 있습니다.

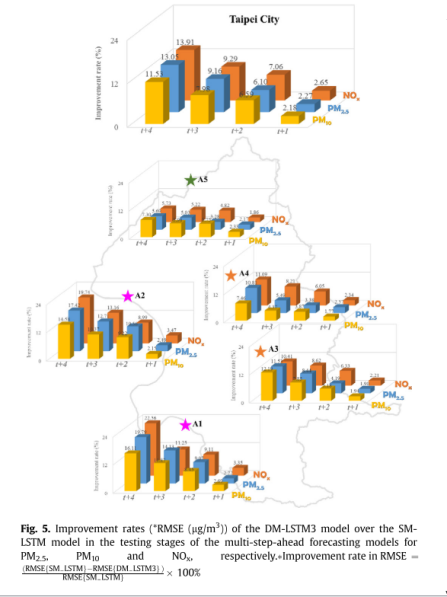
* **결론**

본 논문은 LSTM layer를 다층구조로 하고, overfitting과 instability를 방지할 수 있는 기초적인 3가지 기법(MBGD, dropout, L2 regularization)을 활용해 데이터를 training하였습니다. 그 결과로 단층 LSTM 구조만을 가진 기본 모델보다 개선된 예측성능을 보였고 특히 예측이 어려운 peak값이나, 예측 난이도가 높은 측정소에 대해서 월등히 좋은 성능을 보였습니다. 또한 layer를 깊게한다고 성능이 개선되는 것은 아니며, overfitting문제가 생길 수 있기 때문에 이를 보완할 방법을 결합해야 비로소 성능 개선이 됨을 알 수 있었습니다.

* **Deep Locally-connected 2D LSTM 설계**

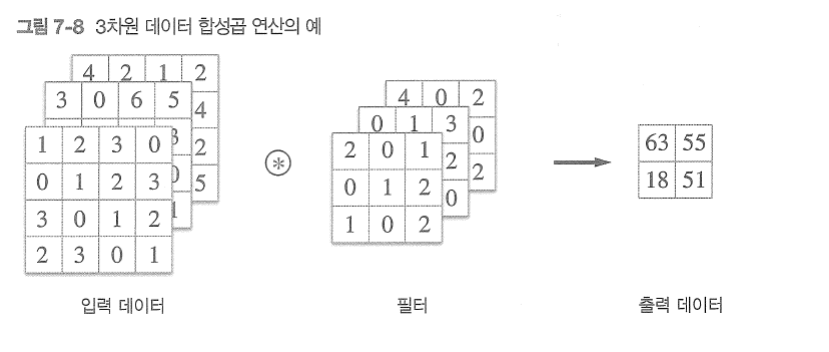
이미지프로세싱에 쓰이는 2D CNN은 kernel이 weight를 공유하기 때문에, 사물의 위치에 상관없이 사물을 검출하는데 문제가 없습니다.



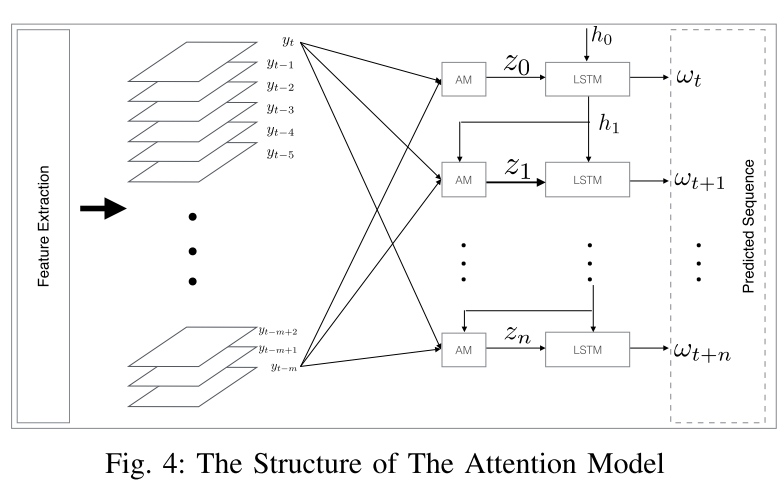
오히려, 미세먼지를 예측하기 위해 공간적 데이터를 분석할 경우에는, location에 따라 미세먼지의 유입특성이 달라질 수 있기 때문에 location을 같이 고려해야 합니다. 따라서 weight는 공유되어서는 안되며, 이를 위해 weight를 공유하지 않는 locally-connected 2D 기법을 LSTM과 결합하는 방법에 대해 연구하고 있습니다.

locally-connected 2D와 LSTM을 결합한 방법은 현재까지 라이브러리나 연구논문으로 제시된 케이스를 본적이 없기 때문에 새롭게 개발하려고 합니다.

또한 feature로 사용될 농도, 바람, 습도는 서로간 선형적 관계에 의해 영향을 받지 않으며 비선형적인 관계에 의해 영향을 받기 때문에 이를 적용하기 위해서는 다층구조 설계가 필요합니다.



(다채널을 포함한 합성곱 연산은 채널간 데이터는 단순히 합산하므로, 이를 보완하기 위해 layer를 다층으로 구현.)



또한 다른 논문에서는 attention 구조를 통해, 과거 데이터를 좀더 예측에 잘 반영할수록 적용한 사례가 있으며, 이를 모델링에 결합하는 방법도 고려할 예정입니다.

따라서 향후 필요한 연구 항목을 다음과 같이 잡으려고 합니다.

* Locally-connected 2D LSTM 기초 모델 개발
* Locally-connected 2D LSTM 다층 구조 설계
* Attention 구조를 결합한 모델링도 고려.
* 연관관계가 검증된 추가 input feature (2차생성을 일으키는 SO2나 NOx) 활용