<div align="center">

|  |  |  |
| --- | --- | --- |
| **2019** | **Report – 2019/ 4월2주** | **Seong-gu** |

</div>

**I. Plan for this week**

***Personal Research Subject***

* GCN-LSTM 모델링 개선 방법 계속 진행

**II. Plan for next week**

* 각 측정소별Terrain 정보를 뽑아서 feature로 활용해보기
* 트레이닝 시간을 고려하여 모델에Attention을 적용하는 대신 CNN을 추가해도 좋을 것. 시간대별 feature를 CNN 방식으로 뽑을 것.
* 현재까지 방식은 다수의 측정소의 다수의 시간대별 미세먼지 농도를 예측하는 것이었으나, 성능이 잘 나오지 않았기 때문에 개별 Station에 대해 특정 시간 뒤 미세먼지 농도를 예측하는 방식으로 수정하여 성능이 얼마나 증가하는지 검토. – small scale 테스트

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

* 랜드마크 등 지형, 지역적 특성 수기 추가
* Feature를 잘 선택 (데이터 프로세싱으로 정확도 개선)
* 지역은 한국을 전범위로 하지 말고 적당한 수준으로 결정함.
* 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
* JEET에 내는 것을 목표로 함.

**IV. Milestone**

* 시공간 데이터 보간 기법 활용 (forward + IDW)
* 추가 feature에 대한 성능 개선 효과 확인 (location predictor, weather, china PM ..)
* feature extraction 방법
* : forward-fixed 기법과 IDW 기법에 대해 가중치를
* . GCN-LSTM 또는 ConvLSTM내부 연산 방법 개선

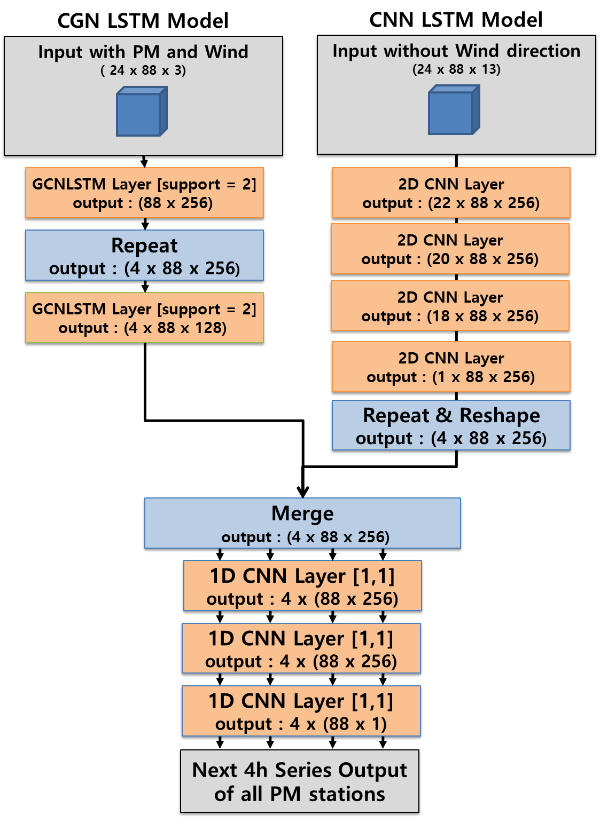
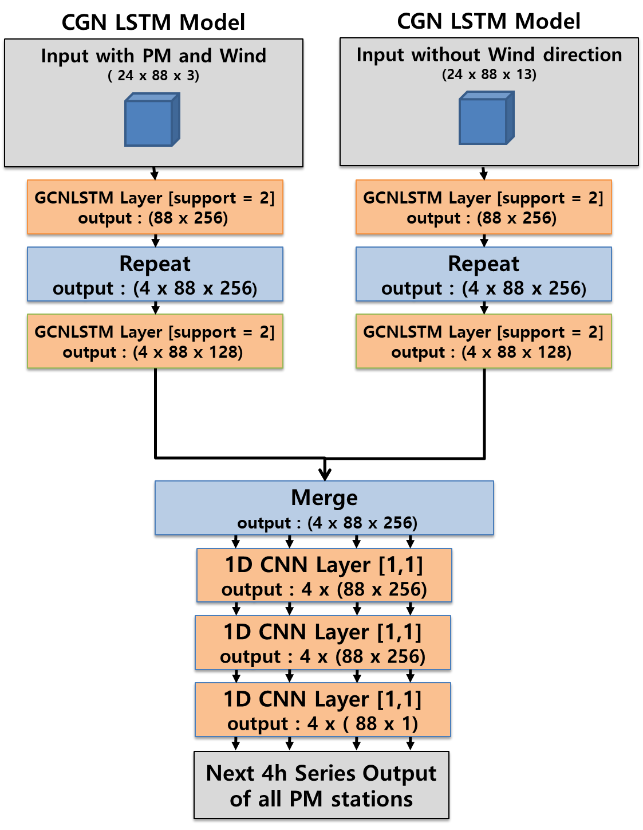
**V. Reports**

* 개인연구 진행사항 – 결과 비교

앞으로 바람에 의한 미세먼지 데이터 유입을 고려하여 데이터 측정 범위를 늘리기 위해 사용할 측정소의 개수를 늘렸습니다. (53-> 88)

24시간 이전 데이터를 기반으로 4시간 뒤 데이터들을 예측했습니다. GCN-LSTM은 2D CNN을 바람 이외의 내부적 요인을 고려한 part(모델의 오른쪽)에 대체한 GCN-LSTM + 2D CNN를 추가로 실험했습니다. 이 모델은 논문에 제시했던 Hybrid 모델의 GCN버전이라고 할 수 있습니다.

Double GCNLSTM (왼쪽)과 GCN-LSTM + 2D CNN (오른쪽) 구조



Input: 24시간, Output: 4시간 Encoder-Decoder 구조

| Model | Prediction hour | | | | Average |
| --- | --- | --- | --- | --- | --- |
| 1h | 2h | 3h | 4h |
| **Follow** | 10.51 | 15.05 | 18.20 | 20.57 | 16.08 |
| **Double ConvLSTM 3x3** | **9.95** | **12.87** | **15.33** | **17.36** | 13.87 |
| **Double GCNLSTM** | 12.84 | 15.18 | 17.20 | 18.84 | 16.01 |
| **GCN-LSTM + 2D CNN** | 10.77 | 14.18 | 16.70 | 18.64 | 15.07 |

표에서 제시된 Follow 모델은 “앞으로의 미세먼지 예측량은 최근까지 업데이트 된 미세먼지량과 같다.”는 단순 가정에 의해 산출된 오차값이며, 시계열 데이터 예측에서 Baseline으로 쓰이는 매우 단순한 기법입니다.

다만 2D CNN를 결합한 GCN-LSTM + 2D CNN이 기존 Double GCNLSTM보다 상당히 좋은 결과를 보였습니다. 대체된CNN파트가 GCN보다는 모델링 목적에 맞게 이전 24시간동안의 내부요인(바람에 의한 주변 미세먼지 유입이 아닌 온도, 습도, 이전 미세먼지농도 등)만을 고려하여 시간에 parallel하게 feature를 추출하였기 때문으로 보입니다.

그렇지만 기존 Double ConvLSTM이 아직까지 가장 좋은 결과를 나타냈으며, GCN이 ConvLSTM보다 좋은 결과를 못 내는 이유로는 Station간 상대적 위치좌표에 대한 정보를 활용하지 못하며 그 정보에 dependent한 training parameter를 쓸 수 없기 때문으로 보입니다. (ConvLSTM의 경우 모든 station의 x,y좌표를 matrix의 row, col에 매핑시키고, 3x3 Kernel을 사용하여 전후좌우의 station에 대해 각각 다른 parameter값을 학습시킬 수 있습니다.)

이외에도 Double ConvLSTM를 다시 ConvLSTM + 3D CNN으로 바꾸어 (Hybrid 모델과 유사) 결과를 비교하였으나 더 나은 결과를 얻지 못했기 대문에 본 결과에서는 제외 했습니다.

다음주에는 현재까지 방식은 다수의 측정소의 다수의 시간대별 미세먼지 농도를 예측하는 것이었으나, 성능이 잘 나오지 않았기 때문에 개별 Station에 대해 특정 시간 뒤 미세먼지 농도를 예측하는 방식으로 수정하여 성능이 얼마나 증가하는지 검토해보도록 할 것이며, GCN의 위 문제점을 해결하는 방법도 고려해보도록 할 예정입니다.