학사학위논문

딥러닝 기법을 이용한 코로나 확진자수 예측 프로그램

- 외부변수 및 해외데이터를 활용한 LSTM 모델의 성능 향상을 중심으로 -

지도교수 이수열

경희대학교
생체의공학과

김 지 후2018103906

2023년 12월 1일

딥러닝 기법을 이용한 코로나 확진자수 예측 프로그램

- 외부변수 및 해외데이터를 활용한 LSTM 모델의 성능 향상을 중심으로 -

지도교수 이수열

이 논문을 학사 학위논문으로 제출함

경희대학교

생체의공학과

김 지 후

2023년 12월 01일

김지후의 학사학위 논문을 인준함

지도교수 이수열

경희대학교

2023년 12월 01일

딥러닝 기법을 이용한 코로나 확진자수 예측 프로그램

List of Figure

List of Table

Abstract

2019년 말, 중국 우한에서 발생했다고 알려진 폐렴의 일종인 코로나 19(COVID-19)가 전 세계적으로 유행하며, 전세계적으로 여러 분야에서 큰 영향을 미쳤다. 이러한 전염병이 발생하게 되면 국가에서는 사회적 거리두기나 해외 입출국 통제 등의 여러가지 정책을 마련한다. 해당 정책은 감염자 수에 근거하여 시행하는 경우가 많은데, 확진자수를 예측 할 수 있다면 그를 대비하기 한결 수월해 질 것이다. 본 논문에서는 딥러닝 기법 중 하나인 LSTM을 통해 코로나 확진자수를 예측하는 모델을 제작하여, 해당 모델의 성능을 높일 수 있는 두가지 방안을 시험하는 것을 목표로 한다. 첫번째 방안은 확진자수 뿐만 아니라 외부적 변수를 추가하는 방안이며, 두번째 방안은 해외 확진자 데이터를 통해 Training data를 늘리는 방안이다. 시험 결과 외부적 변수를 추가하는 방식은 의미있는 결과를 냈지만, 해외 확진자 데이터를 통해 Training data를 늘리는 방안은 부적절한 방식이 라는 것을 알 수 있었다.

1. Introduction

인류는 살아오면서 많은 전염병을 겪어 왔다. 특히 6세기 중역, 로마제국의 도시 인구 40%를 죽음으로 인도했으며, 1300년대 유럽인구의 1/3이상의 목숨을 앗아갔다는 페스트나, 수백 년간 인류를 괴롭힌 천연두, 제 1차 세계대전 이후(1918년)에 세계인구의 약 1/50의 목숨을 앗아간 스페인 독감과 같이 전염병은 인류의 역사와 함께 하고 있음을 알 수 있다. 2019년 말 중국 우한에서 발생했다고 알려진 폐렴의 일종인 코로나-19(COVID-19)이 전 세계적으로 유행하며 전세계적으로 큰 영향을 미쳤다.

코로나는 사회에 큰 영향을 미쳤다. 특히 경제와 관련해 무척이나 많은 영향을 주었는데 코로나로 인해 관광객의 감소와 재택근무의 활성화로 인해 다수의 자영업자들이 힘들어 하는 기사가 다수 작성되었다. 또한 확진자수가 늘어나거나 줄어듬에 따라 정부에서 취하는 정책에 여러 변화가 생겼다. 예를들어 확진자수가 크게 증가할 경우, 사회적 거리두기 정도를 늘린다거나, 해외입출국을 막는 등의 정책을 펼쳤다. 만약 확진자수를 예측 할 수 있다면, 그에 따른 정책이나 대책을 마련하는데 있어 유용하게 사용할 수 있을 것이다.

해당 실험에서는 딥러닝의 기법인 RNN, 그중에서도 LSTM 을 이용하여, 코로나 확진자수를 예측하는 코드를 만드는 것을 목표로 한다. 해당 모델의 성능을 높이기 위해, 추가적인 사회적 모델을 추가하거나, training data의 수를 늘림에 따라 모델의 정확성을 판단하고 모델을 업그레이드 시키는 방법을 연구하는 것을 목적으로 한다.

2. Methods

A. RNN

RNN이란 Recurrent Neural Network 에 대한 약자로서, 어떤 특정 부분이 반복되는 구조를 통해 순서를 학습하기에 효과적인 딥러닝 기법이다. 해당 방식은 데이터에서 규칙적인 패턴을 인식하면서 가중치(W)를 통해 과거의 정보를 통해 현재의 정보 파악에 도움을 받는 구조로 만들어져 있다. 또한 각 time step 마다 가중치가 공유되어 지기 때문에 Backpropagation Trough Time(BIPT)를 통해 학습한다. 이러한 RNN은 각 출력 부분의 가중치는 현재 time step 이외에 이전 time steps에 매우 의존적이다. 많은 수의 뉴런 유신이나 많은 수의 입력 유닛이 있는 경우, 과거 학습 기능을 통해 반복적으로 곱해지는 가중치에 의해 에러값이 1보다 클 경우 누적에러가 기하 급수적으로 증가하거나, 1보다 작을 경우 누적에러가 감소하여 빠르게 0으로 수렴하는 문제가 발생할 수 있다. 이러한 문제 해결을 위해 LSTM모델이 제안 되었다.

B. LSTM

LSTM은 RNN의 한 종류로, 직전 data뿐만 아니라 과거의 data 또한 고려하여 미래 data를 예측하기 위해 나온 모델이다. 따라서 LSTM은 RNN의 장기 의존성 문제를 해결할 수 있을 것을 염두에 두고 만들어낸 모델이다.

LSTM은 RNN과 마찬가지로 체인 구조를 가지고 있지만, 4개의 Layer가 특별한 방식으로 서로 정보를 주고 받는다. RNN과 달리 Cell State라는 구조를 가지고 있는데 이는 이전 상태에서 현재 상태까지 유지되는 정보의 흐름을 나타내는 역할을 한다. 이를 통해 LSTM은 오래된 정보를 기억하고 새로운 정보를 적절하게 갱신할 수 있게 된다. 해당 구조가 RNN과 LSTM의 장기의존성 문제를 해결 할 수 있도록 만든 결정적 구조이다. LSTM은 forget gate, input gate, output gate 3개의 gate로 구성되어 있으며, forget gate 와 input gate를 통해 출력값을 조정하는 방식으로 학습하게 된다.

Forget gate는 과거의 정보의 유용성을 판단한다. Sigmoid function을 통해 유용성의 정도에 따라 0~1까지의 숫자를 이전 module에서 넘어온 output 값과 새로 들어온 input 값에 곱하여 중요한 정보는 1을 곱하여 대부분 정보를 보존하여 그대로 cell state에 전달되고, 중요하지 않은 정보는 0을 곱하여 정보를 보존하지 않는다. 이후 그 값을 cell state 값에 곱해준다.

이후 input gate를 통해 새로운 정보를 어떻게 반영할 것인지를 결정한다. 이를 통해 기존 정보와 새로운 정보를 적절하게 조합하여 정확성을 높일 수 있다. 이때 사용하는 함수는 sigmoid function 과 tanh function 두가지인데 tanh 함수는 RNN에서 사용되는 출력 계산 방법과 동일하게 진행 되며, sigmoid함수는 forget gate와 같이 후보값을 얼마나 전달할지 결정을 내리는 값으로, 이 두 값을 곱해 cell state에 더한다.

Output gate에서 최종 cell state 값에 tanh 함수 곱한 값과 해당 출력이 얼마나 중요한지를 조

절하기 위해 입력된 정보와 hidden state의 sigmoid 함수를 곱한 값을 출력한다. 해당 값은 다음 모듈로 넘어가며, 위의 과정이 반복된다.

해당 실험에서 사용할 LSTM 모델을 사용하면, 예측할 시점의 14일전의 data를 이용해 다음날의 data를 예측 할 수 있다. 즉, t-14시점부터 t-1시점까지의 data를 제공 받아 t시점의 data를 예측하는 모델이다. 마찬가지로 t+1의 data는 제공된 t-13시점부터 t시점까지의 data를 가지고 t+1 시점의 data를 예측 하는 것을 뜻한다. 이를 one-step 예측이라고 부른다. 그러나 해당 방식은 바로 다음날의 데이터를 이전에 예측한 data가 아닌 실제로 측정한 data를 집어 넣어 주어야 하기 때문에 새로운 data가 측정되는 즉시 그 값을 입력해 주어야 하며, 먼 미래를 예측할 수 없다.

이를 보안하기 위해 multi-step 예측을 통해 몇 단위 앞을 예측하는 방식을 사용할 수 있다. Multi-step의 경우, test data의 첫번째 샘플을 사용해 나온 값을 입력 sequence에 포함 시켜 다음 값을 예측하고, 또 해당 값을 다음 입력 sequence에 포함시켜 다음 값을 예측하는 과정을 반복하여 미래를 예측한다. 따라서 '실제로 측정된 t시점의 data' 가 아닌, '이전 one-step을 통해 예측한 t시점의 data' 를 가지고 t+1시점의 값을 예측 하기 때문에, 이전 시점에서 예측한 data의 오차가 지속해서 누적된다. 따라서 더 멀리 예측을 할수록 그 오차가 더욱더 커진다는 문제점을 가지고 있다. 따라서, 오차율이 크게 나오는 것을 생각하여 해당 시험에서는 Multi step을 이용한 시험은 시행하지 않는다.

C. data 전처리

i. data 출처

확진자수 data는 **'Johns Hopkins 대학의 저장소[1]'** 에서 제공한 자료를 사용한다. 해당 data에는 한국(Korea South) 뿐 아니라 여러 나라의 코로나 확진자수를 포함하고 있지만, 여러 나라의 data를 training data에 포함 시키는 것은 각 나라별로 확진자가 늘어나는 변수가 다르기 때문에 포함시키지 않는 것이 나을 것이라 판단하여 오로지 한국(Korea South) data만 사용하기로 결정하였다.

사회적 거리두기 정도 data는 '**KRIHS 인터랙티브 리포트**[2]' 에서 제공한 자료를 사용한다. 거리두기 강도는 1, 1.5, 2, 2.5, 3 으로 총 5개의 단계로 나뉘어있다. Training data를 만들 때 필요한 data는 잠복기를 감안하여 예측하는 날의 일주일 전의 사회적 거리두기 강도를 제공하였다.

하루 평균 온도 data는 '**기상청 기상자료개방 포털**[3]' 에서 제공한 자료를 사용한다. 코로나 측정 기간이 아닌 잠복기를 생각해 실제 코로나가 걸렸을 날짜의 평균기온을 입력해 주기 위해, 예측 날의 일주일 전 평균기온을 제공하였다.

---

[1] 존스 홉킨슨 대학 CSSE 데이터 (https://coronavirus.jhu.edu/)

[2] KRIHS 인터랙티브 리포트 (https://interactive.krihs.re.kr/interactive/covid19/index.html)

[3] 기상청 기상자료개방 포털 (https://data.kma.go.kr/stcs/grnd/grndTaList.do?pgmNo=70)

ii. 최종 traning data

데이터 전처리를 할 경우 sequence data 형태로 데이터를 만들어 주어야 한다. 이때 sequence data의 길이는 14로 설정하여 14일간의 데이터를 한번에 넣어주는 형식으로 데이터를 처리해 주었다. 이를 통해 2020년 1월 22일부터 2020년 12월 18일 까지 총 332일의 확진자수 데이터를 14일씩 묶어서 행렬로 만들어 주었다. 이에 따라 총 318개의 sequence data가 생성되었다. 이때 다음날 확진자수의 경우 예측값과 비교할 값으로 남겨야 하기 때문에 마지막 sequence data는 12월 04일부터 12월 17일까지의 확진자수가 된다. 수집한 데이터는 총 8:1:1의 비율로 나누어 training data, validation data, test data로 활용한다. 아래 표는 data의 일 수 이다.

Table 1 Training data 날짜 및 형태

	Training data	Validation data	Test data
첫 sequence data	20.01.22-20.02.04	20.10.02-20.10.15	20.11.03-20.11.16
마지막 sequence data	20.10.01-20.10.14	20.11.02-20.11.15	20.12.04-20.12.17
Data 형태	(254,14,1)	(32,14,1)	(32,14,1)

평균 기온과 사회적 거리두기 정도의 경우 각 시계열 데이터 앞 부분에 배치하였다. 코로나 확진자수의 측정은 코로나에 실질적으로 걸린날에 잠복기를 더한 날짜라고 판단하여, 거리두기와 평균기온은 예측값의 일주일 전 data 를 입력해주었다. 따라서 최종적인 데이터의 형태는 아래 표와 같다

Table 2 실험A-C 변수 및 training data 형태

	평균 기온	사회적 거리두기 정도	한국 확진자수	Training Data 형태
실험 A	X	X	O	(254,14,1)
실험 B	X	O	O	(254,15,1)
실험 C	O	O	O	(254,16,1)

LSTM 은 그래프의 형태를 분석하는 방식의 딥러닝 기술이다. 따라서 하나의 시계열 데이터만을 분석해서 다음 데이터를 예측하는 것이 일반적이다. 그러나, 해당 실험에서 training data 의 개수는 254 로 그 수가 너무 적어 학습을 하기에는 무리가 있다고 판단하였다. 이를 통해 해외의 확진자수를 이용하여 training data 의 개수를 늘렸을 경우, 예측값에 어떠한 변화가 있는지를 확인해 보았다. 이때 해외의 사회적 거리두기라는 제도를 시행하지 않은 국가도 존재했기 때문에 추가적인 사회적 변수는 포함하지 않았다. 또한 training data 의 순서를 바꾸어 해당 데이터가 학습에 유효한지를 확인하였다. 해외의 데이터는 일본과 러시아의 data 를 사용하여, 총 4 개의 training data 를 통해 실험을 진행하였다. 아래의 표는 training data 의 형태이다.

Table 3 실험 D-G 변수 및 Data 형태

	Training data	사회적 거리두기 및 평균 기온	Data 형태
실험 D	일본 확진자수 + 한국 확진자수	X	(508,14,1)
실험 E	한국 확진자수 + 일본 확진자수	X	(508,14,1)
실험 F	러시아 확진자수 + 한국 확진자수	X	(508,14,1)
실험 G	한국 확진자수 + 러시아 확진자수	X	(508,14,1)

LSTM 모델은 제공되는 14 일 동안의 확진자수 데이터의 분석을 통해 다음 날의 데이터를 예측한다. 따라서 하나의 시계열에서 training data 를 추출하는게 아닌, 2 개의 시계열 data 에서 training data 를 추출하여 모델을 시험하는 해당 실험이 시행되기 위해서는 '한국과 해외 어떤 나라에서든 14 일 동안의 데이터에는 동일한 규칙이 적용되어야 한다.'라는 전제조건이 따른다. 해당 전제 조건을 충족하기 위하여, 일본과 러시아 모두 '주중에는 외출의 양이 증가하여 확진자수가 증가하고, 주말에는 외출의 양이 감소하여 확진자수 또한 감소할 것이다.'라는 규칙이 한국과 공유 된다고 판단하여 해당 실험을 진행하였다.

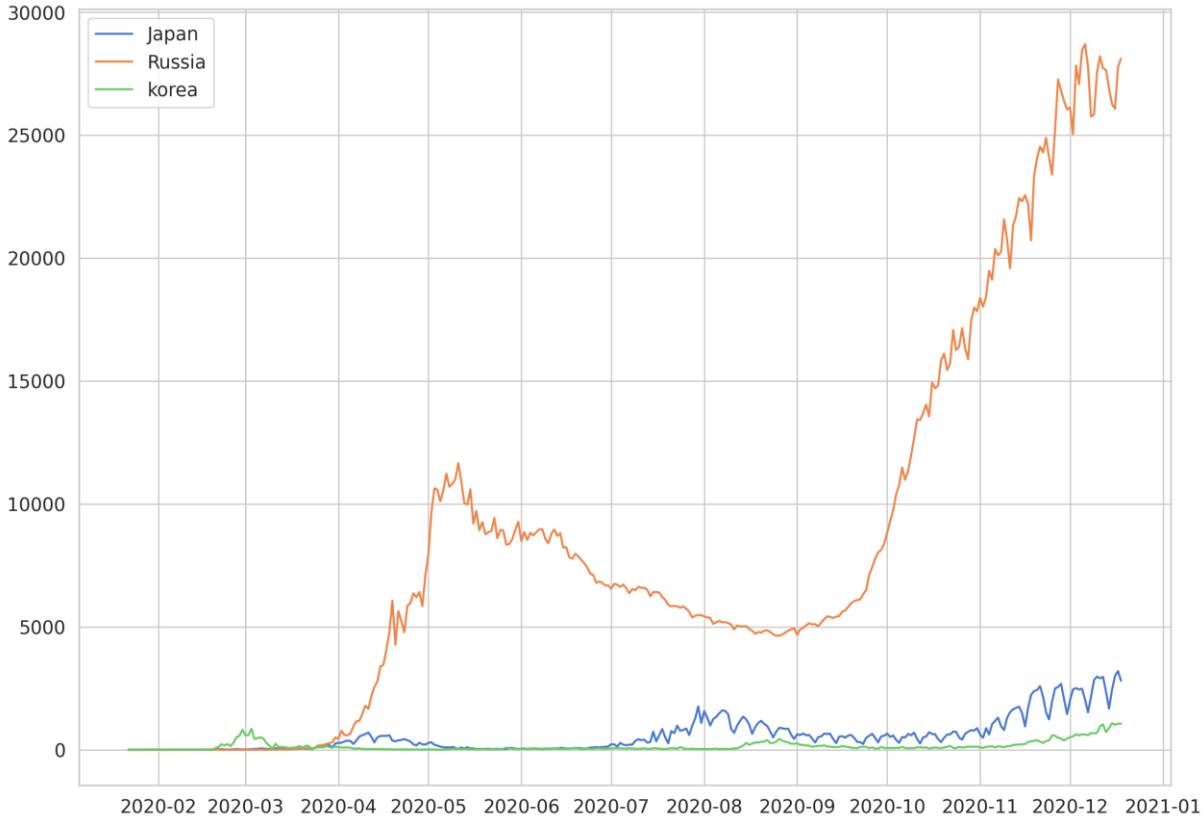

Figure 1일본, 러시아, 한국 확진자수 data

위의 그래프는 데이터 전처리를 하지 않은 확진자수의 그래프이다. 이때 러시아와 일본의 확진자수가 한국에 비해 확실히 그 수가 많다는 것을 확인 할 수 있다.

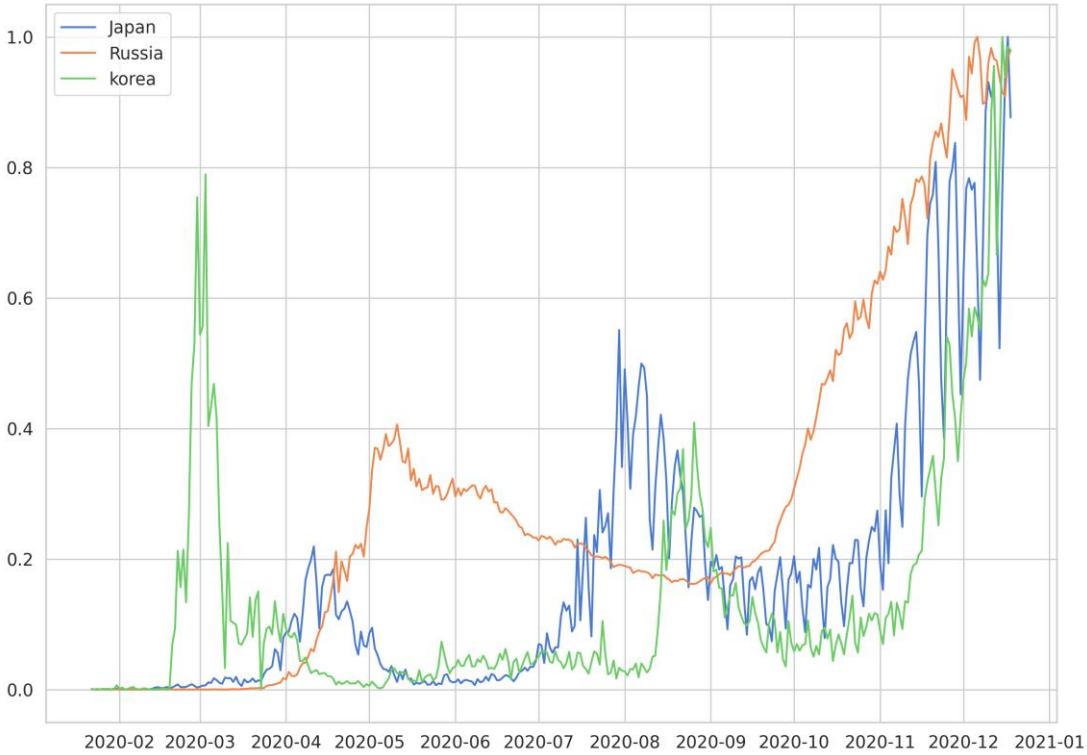

Figure 2 Regulation 이후 일본, 러시아, 한국 확진자수 data

이를 MinMaxScale을 통해 전처리를 하였을 경우 위와 같은 모습이 나오게 된다.

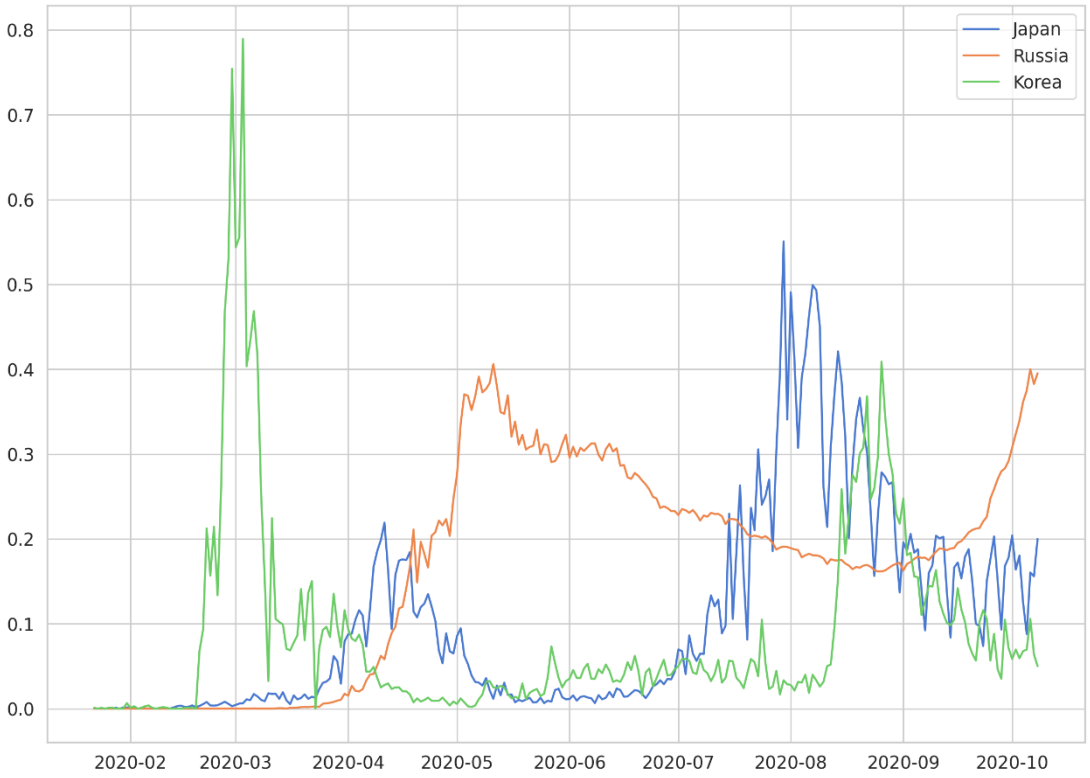

Figure 3 Regulation 이후 일본, 러시아, 한국 확진자수 training data

해당 그래프는 한국과 일본 러시아의 확진자수를 training data에 포함된 일 수만을 보인 그래프이다. 해당 그래프를 통해 각 나라별 확진자수의 특징을 정리해 본다면, 한국의 경우 2월 중순부터 3월 중순까지 한달 사이에 확진자수가 급격하게 증가하였다가 감소하는 특징과 8월 중순에 다시 한번 확진자수가 급증했다가 9월 중순까지 확진자수가 급감하는 모습을 보이고 있다. 이후 10월 중순까지는 확진자수가 크게 변화하는 구간 없이 유지되는 듯한 형태를 보이고 있다.

일본의 확진자수의 경우 7월 중순부터 8월 초까지 급격히 증가하였다가 9월 초까지 급감하는 모습을 확인 할 수 있다. 또한 마지막 10월 중순까지는 4일정도의 주기로 매일 확진자수가 증가하였다가 감소하는 모습이 확인되지만 약 0.1에서 0.2 사이에서 확진자수가 유지되는 모습을 확인 할 수 있다.

러시아의 경우 4월 중순부터 5월 중순까지 확진자수가 급격하게 증가하지만 이후 9월 초까지 그 수가 완만하게 감소하는 형태의 그래프를 그리는 것을 확인 할 수 있다. 또한 9월 중순부터 10월 2중순까지 다시 한번 급격하게 확진자수가 급증하는 모습을 보이며 그래프가 끝이 난다.

D. MAE

오차의 정도를 알아보는 여러가지 방법 중에 해당 실험에서 사용하는 방식은 MAE 방식이다. MAE 방식이란 '평균 절대 오차(Mean Absolute Error)'의 약자로서, 실제 값과 측정(예측)값 사이의 차이값의 평균을 의미한다. 이에 대한 정확한 수식은 아래와 같다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

해당 수식에서 n은 오차의 개수, $|x_i - x|$는 절대 오차를 의미한다. 따라서 MAE를 측정하기 위해서는 각 측정값과 실제값 사이의 절대 오차를 구하고 구한 절대 오차들을 모두 더해 이를 평균을 메기면 MAE의 값을 얻을 수 있다.

MAE는 loss function이 오차와 비례하여 일정하게 증가하는 특징이 있다. MSE는 오차 제곱의 평균값이므로 오차가 커질수록 손실의 함수의 값이 빠르게 증가한다. 이러한 이유로 MAE는 outlier에 강건하다. 이는 오차가 유난히 큰 값은 outlier로서 간주하고 해당 값을 무시하고 학습한다는 의미다. 예를 들어, 모델 학습이 잘 되었다면 대부분의 오차가 작기 때문에 밀집되어 나타난다. 그 중 오차가 유난히 큰 Outlier가 발생 할 겨우, 이 오류의 오차를 줄이기 위해 오차가 작게 나온 밀집된 데이터의 값을 모두 변화 시켜 얻는 loss 이득이나 outlier의 큰 오차를 무시하고 현재 상태에서 학습을 진행할 때의 loss이득이 동일하다. 따라서 MAE는 이러한 outlier을 무시하고 학습을 한다. 즉 MAE는 outlier가 있어도 최대한 잘 추정된 데이터들의 특성을 반영할 수 있기 때문에 통계적으로는 중앙값과 연관이 깊다. 반면 MSE의 경우 error 값이 증가함에 따라 loss function이 제곱배만큼 커지기 때문에 outlier의 오차를 줄이면 얻을 수 있는 loss이득이 훨씬 큽니다. 즉 MSE는 outlier에 민감하다. 이러한 이유로 회귀(regulation) 문제에서는 MAE를 더 자주 사용하며, 해당 실험에서도 Loss Function을 MAE로 설정하였다.

3. Results & Discussion

A. 한국의 확진자수 만을 가지고 예측한 경우

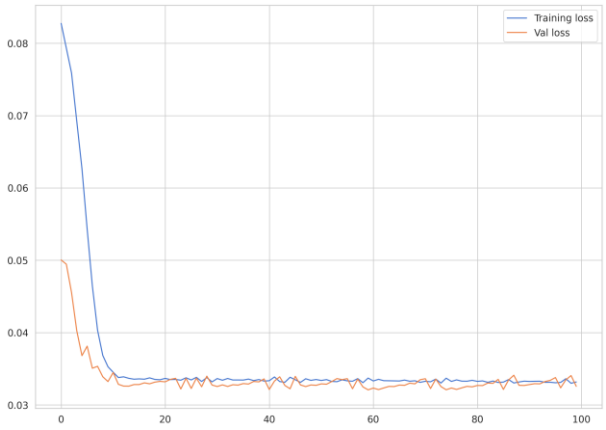

Figure 4 Loss 그래프 (training data: 한국 확진자)

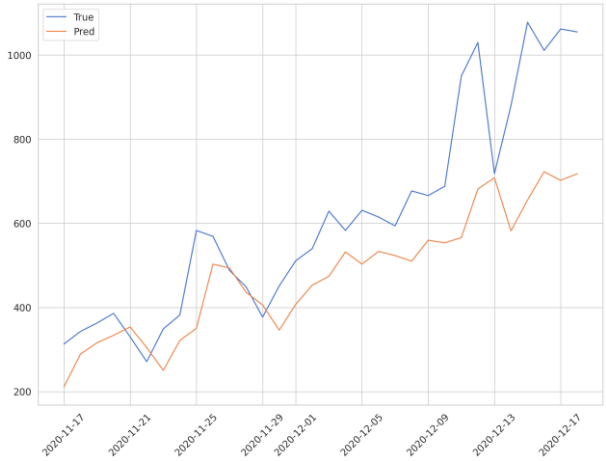

Figure 5 Predict data 그래프 (training data: 한국 확진자)

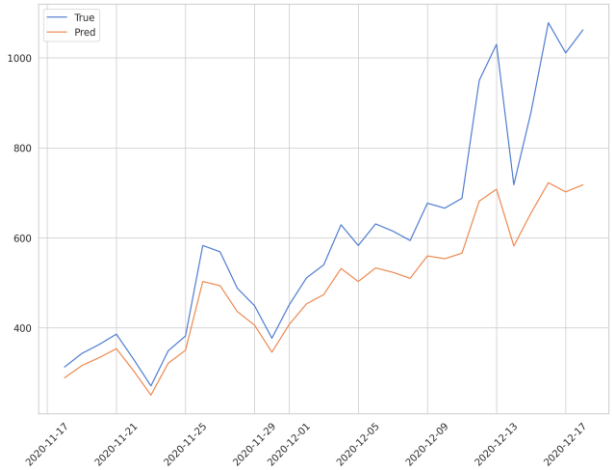

Figure 6 하루 앞당긴 Predict data 그래프 (training data: 한국 확진자)

B. 한국의 확진자수와 사회적 거리두기 정도를 통해 예측한 경우

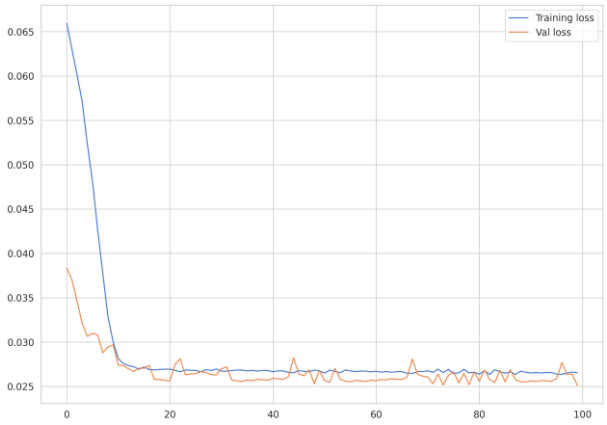

Figure 7 Loss 그래프 (training data: 한국 확진자, 사회적 거리두기 정도)

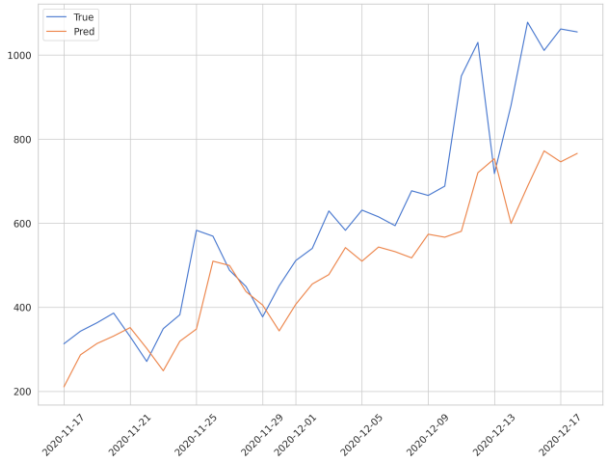

Figure 8 Predict data 그래프 (training data: 한국 확진자, 사회적 거리두기 정도)

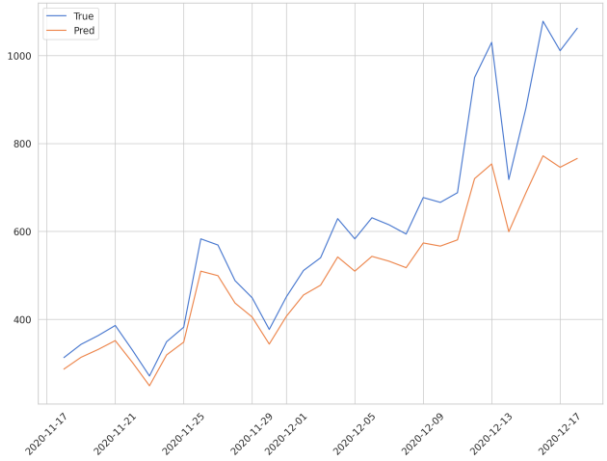

Figure 9 하루 앞당긴 Predict data 그래프 (training data: 한국 확진자, 사회적 거리두기 정도)

C. 한국의 확진자수와 사회적 거리두기, 평균 기온을 통해 예측한 경우

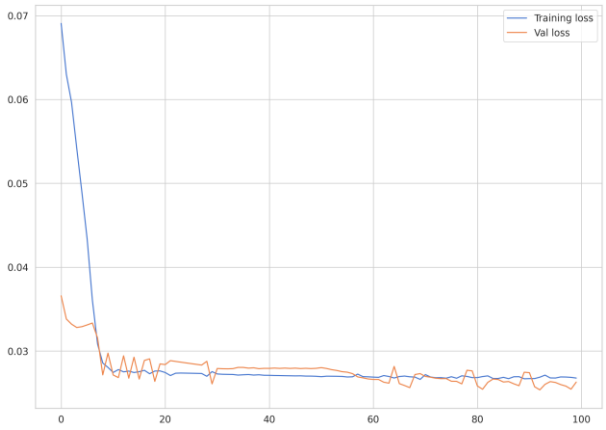

Figure 10 Loss 그래프 (training data: 한국 확진자, 사회적 거리두기 정도, 평균기온)

Figure 11 Predict data 그래프 (training data: 한국 확진자, 사회적 거리두기 정도, 평균기온)

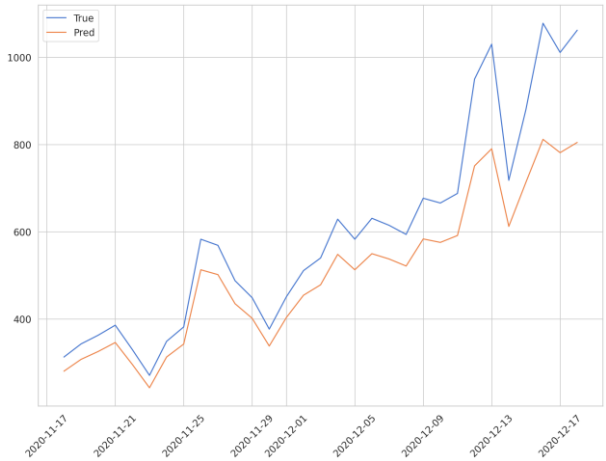

Figure 12 하루 앞당긴 Predict data 그래프 (training data: 한국 확진자, 사회적 거리두기 정도, 평균기온)

D. 한국 확진자수 예측 (training data: 일본 확진자 + 한국 확진자)

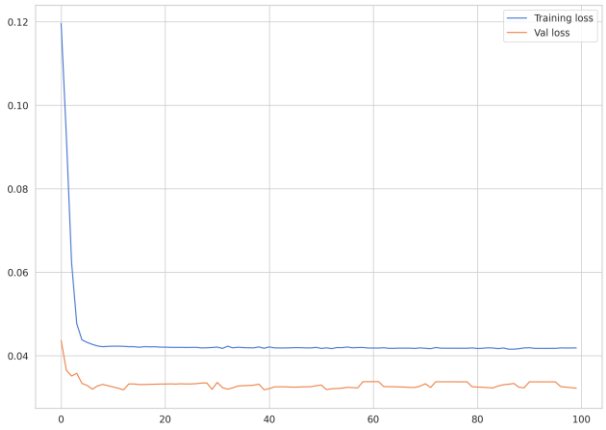

Figure 13 Loss 그래프 (training data: 일본 확진자 + 한국 확진자)

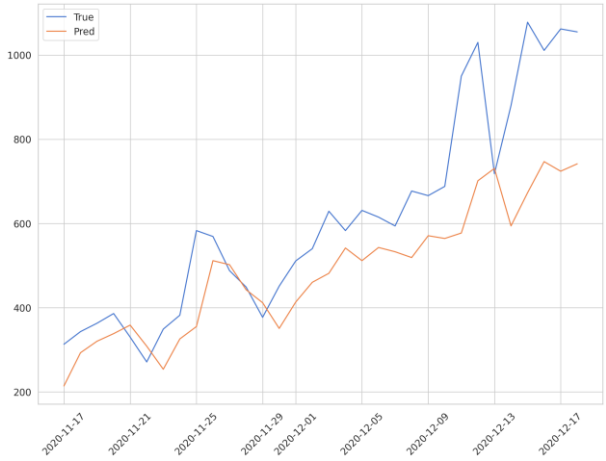

Figure 14 Predict data 그래프 (training data: 일본 확진자 + 한국 확진자)

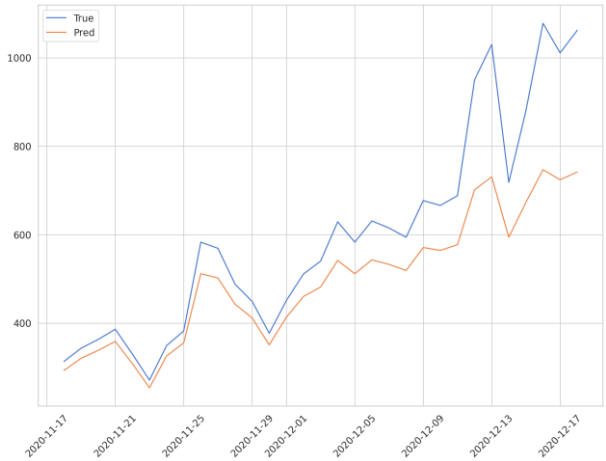

Figure 15 하루 앞당긴 Predict data 그래프 (training data: 일본 확진자 + 한국 확진자)

E. 한국 확진자수 예측 (training data: 한국 확진자 + 일본 확진자)

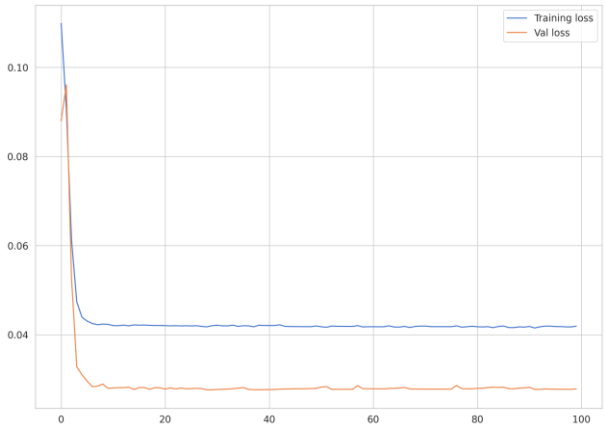

Figure 16 Loss 그래프 (training data: 한국 확진자 + 일본 확진자)

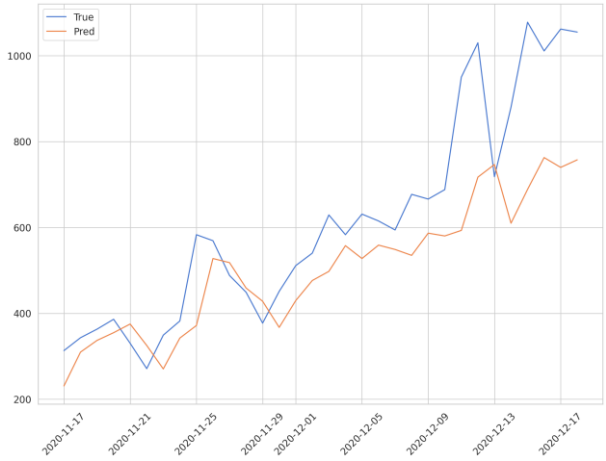

Figure 17 Predict data 그래프 (training data: 한국 확진자 + 일본 확진자)

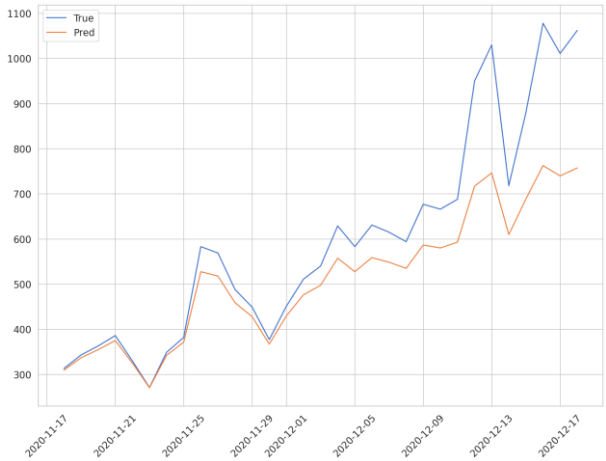

Figure 18 하루 앞당긴 Predict data 그래프 (training data: 한국 확진자 + 일본 확진자)

F. 한국 확진자수 예측 (training data: 러시아 확진자 + 한국 확진자)

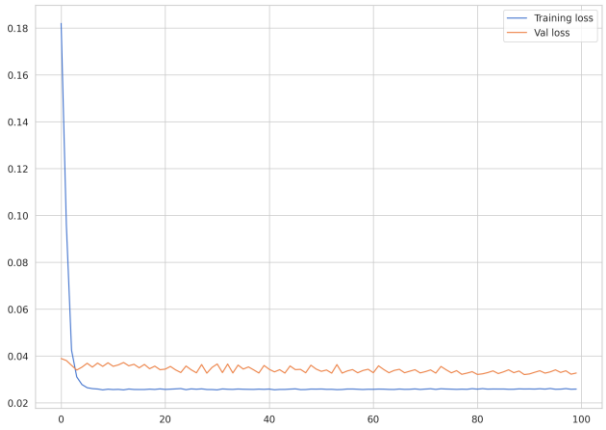

Figure 19 Loss 그래프 (training data: 러시아 확진자 + 한국 확진자)

Figure 20 Predict data 그래프 (training data: 러시아 확진자 + 한국 확진자)

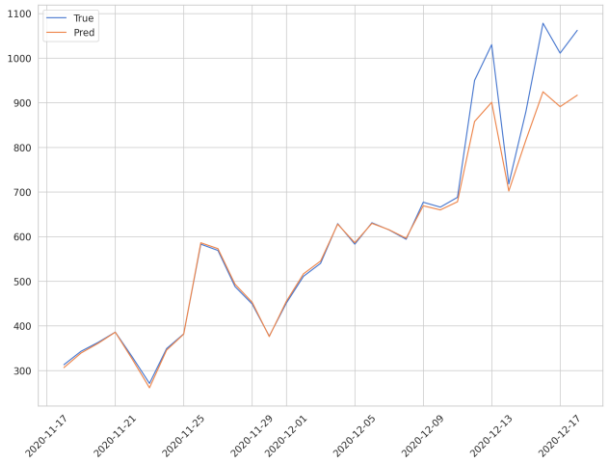

Figure 21 하루 앞당긴 Predict data 그래프 (training data: 러시아 확진자 + 한국 확진자)

G. 한국 확진자수 예측 (training data: 한국 확진자 + 러시아 확진자)

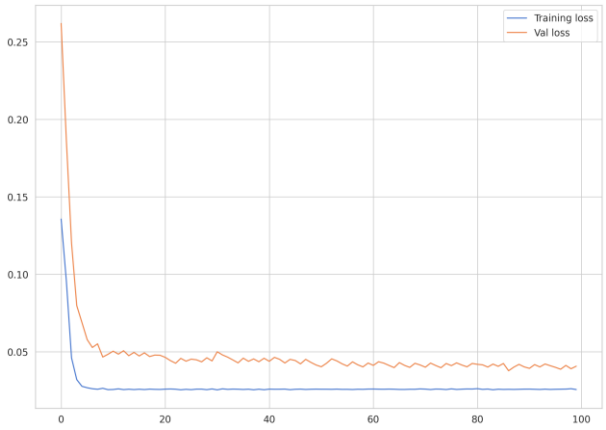

Figure 22 Loss 그래프 (training data: 한국 확진자 + 러시아 확진자)

Figure 23 Predict data 그래프 (training data: 한국 확진자 + 러시아 확진자)

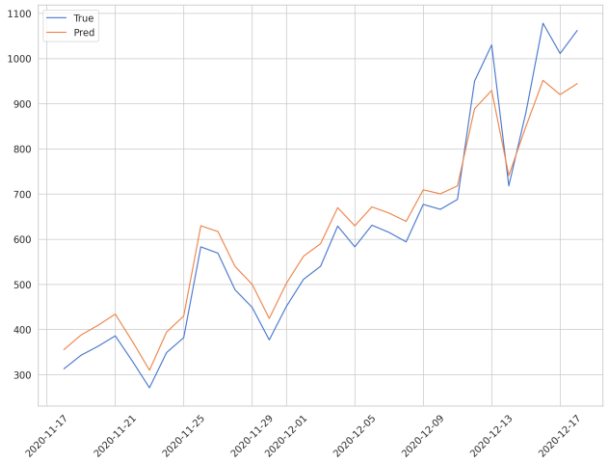

Figure 24 하루 앞당긴 Predict data 그래프 (training data: 한국 확진자 + 러시아 확진자)

4. Conclusion

A. 사회적 변수를 고려한 확진자수 예측

Table 4 실험 A-C 결과값 (Loss 및 MAE)

	실험 A	실험 B	실험 C
Training loss	0.0332	0.0265	0.0267
Validation loss	0.0328	0.0255	0.0259
Single-step 예측 MAE	246.706	242.444	251.890
전날 과의 유사성 MAE	58.543	99.119	91.896

딥러닝을 통해 원하는 결과는 training loss와 validation loss 값의 최소화다. 이때, loss가 가장 적은 모델은 위 실험의 그래프를 바탕으로, '내일의 확진자수의 추세는 오늘의 확진자수의 추세와 동일한 것이다'라는 전제를 학습하는 것으로 판단 된다. 따라서 하루 뒤의 data는 그 전날의 data와 비슷한 형식으로 따라가게 된다. 이를 확인 하기 위해 모든 predict data를 하루 앞으로 당긴 후 MAE를 비교해 보는 실험 또한 동행했다.

사회적 거리두기 변수를 추가하였을 경우, 확진자수 data만을 이용하였을 때 보다 training loss와 validation loss값이 확실히 낮아지며, 모델의 정확도가 올라갔다고 판단하였다. RNN모델을 사용하는 대표적인 문제인 '자연어 처리' 문제의 경우, 다음에 올 문자를 예측하기 위해 '주어인 명사 뒤에 서술어인 동사가 와야 한다.'와 같은 몇가지 규칙이 존재한다. 그러나 '이제까지의 확진자수'만을 가지고 미래의 확진자수를 예측할 때, 학습할 수 있는 것은 이전의 그래프의 모양 및 추세 정도일 뿐 정확한 규칙을 찾아내는 데에는 어려움이 있다. 따라서 사회적 거리두기 와 같은 외부 변수를 추가하여 '거리두기의 강도가 강해지면 확진자수는 줄어든다'와 같은 규칙을 학습 할 수 있도록 해주었기 때문에 validation loss와 training loss가 줄어든 것으로 판단된다. 추가적으로 평균 기온 변수 또한 추가해 주었을 때는 loss값이 사회적 거리두기 변수를 추가하였을 때 보다 급격하게 좋아지는 모습은 보이지 않는다.

모델의 성능인 MAE 값을 비교하였을 경우, 거리두기 변수를 추가하였을 경우는 그 성능이 증가하였음을 보이지만, 평균기온을 추가하였을 경우 성능이 오히려 떨어진 것을 확인 할 수 있다. 이를 통해 확진자수에 영향을 크게 미치는 변수인 사회적 거리두기 정도는 규칙을 학습하는데 있어 유효한 변수이며, 확진자수의 큰 영향을 미치지 않는 변수인 평균 기온은 오히려 규칙을 학습하는데 있어 방해가 되는 변수라 판단하였다. 예를 들어 사회적 거리두기 변수를 해당 모델에 추가적으로 제공하였을 경우, '사회적 거리두기가 증가하면 몇일 후 확진자수가 감소하는 추세를 보인다'라는 규칙을 학습한 것이라 판단할 수 있다.

'내일의 확진자수 추세는 오늘의 확진자수 추세와 동일할 것이다'라는 전제를 얼마나 학습하였는가를 측정하기 위해, 예측 그래프를 하루 앞당겨 MAE를 측정하여 비교해 보았을 때, 확진자수만을 추가하였을 때가 해당 전제를 가장 잘 학습 하였으며, 두번째로 거리두기 변수와 평균기온을 추가하였을 때 해당 전제를 잘 학습하였다. 마지막으로 '사회적 거리두기 정도'만을 추가하였을 때 해당 전제의 학습을 제일 크게 벗어난 것을 확인 할 수 있다.

또한 급격히 확진자수가 급격하게 변화하는 구간인 2020년 12월 11일부터 2020년 12월 17일까지는 모든 실험결과에서 동일하게 오류가 특히 크게 나타나는 것을 확인할 수 있다. 이는 확진자수 data만으로는 외부의 변수에 의해 급격히 그 수가 변하는 날짜를 예측 할 수 없기 때문에 급격한 변화 구간에서 오차가 더 크게 나타나는 것으로 판단된다.

B. 해외 확진자수를 포함한 training data를 통한 한국의 확진자수 예측

Table 5 실험 D-G 결과값 (Loss 및 MAE)

	실험 A	실험 D	실험 E	실험 F	실험 G
Training loss	0.0332	0.0419	0.0418	0.0259	0.0259
Validation loss	0.0328	0.0337	0.0281	0.0323	0.0393
Single-step 예측 MAE	246.706	242.548	238.375	254.722	250.741
전날 과의 유사성 MAE	58.543	100.632	84.550	26.314	52.179

해당 실험에서 눈에 띄게 특이한 점은, 러시아의 경우 모든 실험 중에서 유일하게 전날과의 유사성을 확인하기 위해 예상 그래프를 하루 앞당겼을 때 실제 data보다 높은 값으로 다음날의 data를 예측한 모습을 확인 할 수 있다는 점이다.

이러한 결과가 나온 이유로는 2가지의 가능성이 있다. 첫번째로, Russia data만이 가장 최근에 급증하는 추세를 가지고 있기 때문에 Russia data를 학습하였을 경우 predict data가 크게 측정되었다는 가설이다. 그러나 해당 가설은 한국과 외국의 training data순서를 바꾸어서 확인해 본 결과 한국의 training data를 가장 뒤에 둔 경우들을 확인해 보았을 때, Russia의 data를 학습하였을 때가 Predict data가 높게 나온 것이 확인되며, 틀린 가설 임을 확인 할 수 있었다.

두번째로, regulation 과정에서 Russia의 training data의 최댓값이 일본과 한국에 비해 작기 때문에 상대적으로 높은 data만을 가지고 있는 validation data와 test data에서 낮아지는 추세를 학습하지 못했기 때문이라는 가설이다. Validation data와 test data는 증가하는 추세에서의 data이기 때문에 대부분 training data 보다 큰 값을 가진 값들로 존재한다. 이에 따라 큰 값에서의 그래프 추세를 학습이 덜된 Russia의 predict data 값들이 다른 모델보다 크게 예측 되었을 것이다.

실험 D-G의 결과를 확인해 보면, 일본 data를 추가하였을 경우는 training loss가 오히려 늘어났으며, 러시아 data를 추가하였을 경우에는 training loss가 줄어드는 모습을 확인할 수 있다. 그러나 MAE의 경우는 일본의 data를 사용했을 때는 더 낮은 MAE를 보이고, 러시아의 data를 사용했을 때는 더 높은 MAE를 보이는 정반대의 결과를 보인다.

해당 실험은 부족한 training data를 보완하기 위해 '14일동안의 데이터에서는 동일한 규칙이 적용되어야 한다'라는 전제를 바탕으로 시행되었던 실험이다. 그러나 러시아의 data와 같이 성능을 높이는 것이 아닌 낮추는 결과가 발생하였다. 이는 무리한 전제조건을 사용하면서 training data를 늘리는 것에 비해, 눈에 띄는 성능 향상을 보여주지 못하는 것을 보여준다. 따라서 해외의 data를 이용하여 training data의 수를 늘리는 방식으로 모델의 성능을 높이는 것은 부적절한 방법임을 알 수 있다.

C. 결론

2가지의 실험을 통해, '코로나 확진자수 예측 모델'의 성능을 높이는 방법으로 해외의 확진자 data를 추가하여 training data의 수를 높이는 방식은 잘못된 방식이며, 적절한 사회적 변수를 추가하는 것이 더 적합한 방식이라는 것을 알 수 있다. 따라서, 모델의 성능을 높이기 위해서는 선행적으로 코로나의 확진에 영향을 미치는 요소에 대한 연구가 선행적으로 요구된다. 또한, 예측에 도움이 되는 변수 (ex. 구글 '코로나' 검색 횟수, 마스크 구매 수량, 등)를 추가적으로 제공해 주는 것 또한 모델의 성능을 높이는데 도움이 될 것이다.

또한 해당 실험 A-C 에서는 사회적 거리두기 정도와 평균기온을 예측날의 일주일 전 data 만을 제공해 주었다. 그러나 실제 사람들의 잠복기는 정확히 일주일이지 않으며 해당 변수의 효과가 얼마나 빠르게 적용되는지 확실하지 않기 때문에, 일정 기간의 변수를 추가적으로 제공하는 것이 더 효과적일 것으로 생각된다. 즉, 일주일 전의 data 만을 주는 것이 아닌 3 일전부터 10 일전까지의 사회적 거리두기 정도와 평균기온을 변수로 제공하는 것이 더 효과적인 모델을 만드는데 도움이 되었을 것이다.

5. Reference

[1]. 김대호, 김세일 and 고진환. (2023). 기상 데이터를 활용한 CNN-LSTM 기반 유량 예측 모델 연구. Journal of the Korea Academia-Industrial, Vol. 24, No. 8 pp. 22-31

[2]. 김재호, 김장영. (2022). 코로나 확진자수 예측을 위한 BI-LSTM과 GRU 알고리즘의 성능 비교 분석. 한국정보통신학회논문지, Vol. 26, No.2: 187-192

[3]. 노윤아, 정승원, 문재욱 and 황인준. (2022). 사회적 변수를 고려한 LSTM 기반 코로나19 일별 확진자수 예측 기법. 정보과학회 컴퓨팅의 실제 논문지, 28(2), 116-121.

[4]. 질병 관리청 "과거대유행경험" doi: https://www.kdca.go.kr/contents.es?mid=a21104010000

[5] Felix A. Gers, Jürgen Schmidhuber, Fred Cummins; Learning to Forget: Continual Prediction with LSTM. Neural Comput 2000; 12 (10): 2451–2471. doi: https://doi.org/10.1162/089976600300015015