LSTM을 이용한 국내 태양광 발전량 예측 연구

지도교수 Aziz Nasridinov

이 논문을 학사학위 논문으로 제출함

2022年 06月

충 북 대 학 교

빅데이터 융합전공

정 연 휘

정연휘의 학사학위논문을 인준함.

지도교수 Aziz Nasridinov (인)

2022年 06月

충북대학교 빅데이터 융합전공

차 례

표 차례

그림 차례

Ⅰ. 서론

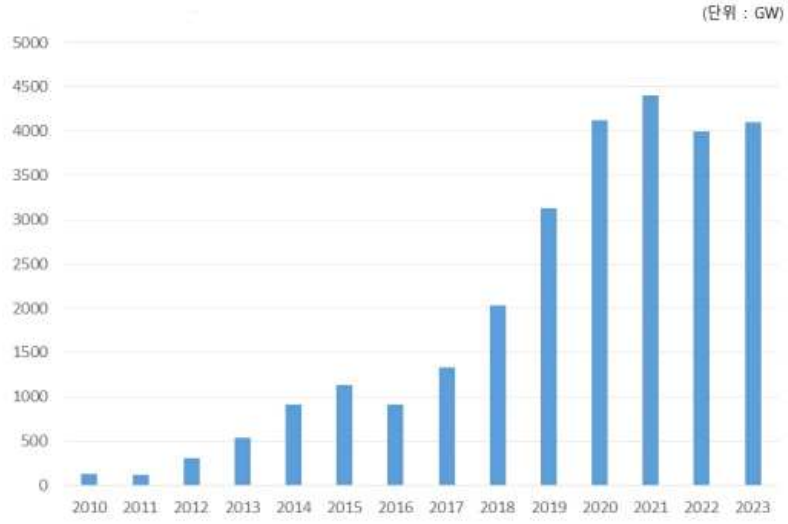
1. 연구 목적

2021년 8월 제54차 IPCC(Intergovernmental Panel on Climate Change) 총회에서 「IPCC제6차 평가보고서(AR6) 제1실무그룹 보고서」를 승인하였다. 보고서에 따르면 2001-2020년의 지구 표면 온도는 1850-1900년보다 0.99℃ 높았으며, 2011-2020년 지구 표면 온도는 1980-1990년 보다 1.09℃ 높았다고 하였다. 또한 향후 몇 십 년 내 CO_2와 기타 온실가스에 대한 심층 감축이 진행되지 않는다면 21세기 중 지구 온난화는 1.5℃ 및 2℃를 넘어설 것이라고 예측하였다. 2℃ 이상 상승하는 경우 가뭄과 폭우 발생 빈도와 강도는 1.5℃ 대비 더 높아지고, 예측의 정확도도 함께 높아질 것이라고 하였다. 인류가 배출한 CO_2의 순배출 제로를 달성하는 것은 인류가 초래한 지구 온도 상승을 안정화시키는 데 필수적인 요소라고 발표하였다.

이처럼 전 세계적으로 환경 규제에 따라 청정 에너지에 대한 중요성이 강조되었고, 국내에서도 화석 연료의 사용을 축소하며, 이를 대체할 에너지를 위한 다양한 정책이 나타나고 있다. 그 중, 2020년 12월에 발표한 산업통상자원부에서 발표한 「제5차 신재생에너지 기술개발 및 이용·보급 기본계획(2020-2034)」에 따르면 2030년까지 1차 에너지 14.3%, 발전량 중 21.6%를 신재생 에너지로 공급하고자 한다. 그로 인해, 14~30년간 CO_2는 주적 9.9억톡 감축 전망을 예측하고 있다. 또한 태양광 등 청정에너지 중심의 에너지 확산 기반을 마련해야 한다.

해외경제소에서 발표한 「2021년 하반기 태양광산업 동향」에 따르면 2021년 글로벌 태양광 신규 설치량은 180~190GW로 추정되며, 2022년 전망치는 230GW로 예상된다고 발표하였다. <그림 1>은 글로벌 태양광 설치량 현황 및 전망을 나타낸다. 중국 및 미국 시장의 양호한 태양광 수요는 가스 가격 상승에 따른 발전단가 급등과 에너지안보 측면에서 태양광 발전의 중요성 부각 등으로 태양광 수요는 당분간 지속적으로 늘어날 전망이라고 하였다. 또한 중동 및 아시아 등을 중심으로 한 개발도상국 또한 태양광 수요는 선진 태양광 시장대비 규모는 작으나, 성장률은 선진 시장 대비 두 배 이상 빠르게 증가할 전망이라고 하였다.

따라서 본 논문에서는 LSTM(Long Short-Tem Memory)을 이용하여 국내의 태양광 발전량 예측 연구를 하고자 한다. 데이터 학습 모델을 구축하여 데이터를 5시간, 10시간, 20시간, 23시간으로 나누어서 진행해야 한다.



<그림 1> 글로벌 태양광 설치량 현황 및 전망
(출처 : 한국수출입은행 해외경제연구소)

<표 1>는 한국에너지공단 신재생에너지센터 「신재생에너지보급통계」의 우리나라의 신재생에너지별 생산량 현황이다. 이처럼 국내에서도 2012년경부터 태양광 발전량은 꾸준히 증가하는 추세이다.

<표 1> 우리나라의 신재생에너지별 생산량 현황 (단위 : 천 toe)
(출처 : 한국에너지공단 신재생에너지센터)

구분	2012	2013	2014	2015	2016	2017	2018	2019	2020
태양열	26.3	27.8	28.5	28.5	28.5	28.1	27.4	26.9	26.4
태양광	243.0	358.7	579.1	907.0	1,183.3	1,672.4	2,194.0	3,055.2	4,156.0
풍력	192.7	242.4	241.8	283.5	355.3	462.2	525.2	570.8	671.1
수력	814.9	892.2	581.2	453.8	603.2	600.7	718.8	594.5	826.3
바이오	1,334.7	1,558.5	2,822.0	2,765.7	2,765.5	3,598.8	4,442.4	4,162.4	3,899.2
폐기물	5,998.5	6,502.4	6,904.7	8,436.2	8,742.7	9,359.0	9,084.2	1,119.8	1,166.0
해양	98.3	102.1	103.8	104.7	104.6	104.3	103.4	101.0	97.4
지열	65.3	87.0	108.5	135.0	162.0	183.9	205.5	224.7	241.0
수열	0.0	0.0	0.0	4.8	6.0	7.9	14.7	21.2	21.3
연료전지	82.5	122.4	199.4	230.2	241.6	313.3	376.3	487.2	750.8

2. 연구의 필요성

태양광 발전은 에너지원이 무한하고, 청정하며, 유지 보수가 유지보수가 용이하고, 수명이 길다는 장점이 있다. 그러나 초기 투자비와 발전단가가 높고, 설치장소가 한정적이며, 에너지 밀도가 낮아 큰 설치면적이 필요하다는 단점이 있다. 특히, 기상상태에 의존하기 때문에 발전량이 매우 간헐적인 특성이 있다. 그렇기에 태양광 발전량의 불확실성을 줄이기 위해서 태양광 발전량 예측은 필수적이다.

IEA(International Enegy Agency)의 「재생에너지 3020이행계획」에 따르면 2030년까지 재생에너지 비중을 20%로 높이겠다는 목표를 달성하기 위해서는 재생에너지를 수용할 수 있는 탄력적이고 유연한 전력 시스템이 필요하다고 하였다. 이런 시스템 관리를 위해 정확한 재생 에너지 발전량 예측의 중요성은 커지고 있다. 산업통상지원부와 한국전력거래소는 재생에너지 확대에 따른 출력 변동성 대응을 위해 재생에너지 발전량 예측제도를 도입하였다. 이 제도는 20MW 이상 태양광 및 풍력 발전사업자 등이 재생에너지 발전량을 하루 전에 미리 예측하여 제출하고, 당일 날 일정 오차율 이내로 이를 이행할 경우 정산금을 지급하는 제도이다. 예측 제도를 통해 재생 에너지 발전량을 예측할 수 있고 효율적인 전력 시스템이 운영 될 수 있을 것이다. 이로 인해 에너지의 이용률을 극대화 할 수 있다.

한전 데이터사이언스연구소는 재생에너지 발전량 예측제도 차명를 지원하기 위해서 '태양광 발전량 예측 기술'을 자체 개발했다. AI(인공지능) 기반으로 태양광발전소의 발전실적과 기상 관측 데이터를 딥러닝 기법으로 분석해 알고리즘을 도출하고, 기상예보 데이터가 입력되면 발전량을 예측하는 기술이다. 이처럼 국내에서는 다양하게 태양광 발전량 예측에 대해서 큰 관심을 보인다.

따라서 국내의 태양광 발전량 예측의 연구가 중요함을 인지하였고, 이에 따라 연구를 진행할 예정이다. 본 논문은 장기간의 시계열 데이터에 유리한 LSTM(Long Short-Term Memory)을 기반으로 하여 태양광 발전량 예측을 할 예정이다. 미래의 시간별 태양광 발전량 데이터를 구하기 위해서는 미래의 시간별 태양광 발전량에 영향을 미치는 기후 데이터들이 필요하다. 그러나 미래의 기후 데이터들을 수집하기 어려운 상황이다. 이이 대해 본 논문에서는 과거의 시간별 태양광 발전량에 영향을 미치는 기후 데이터들을 에측 정확도를 향상하기 위한 모델을 구축하여 활용할 예정이다.

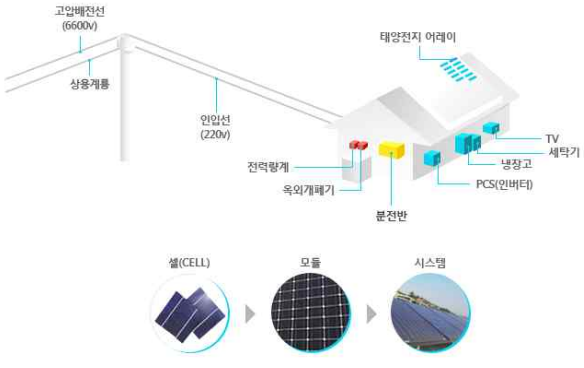
Ⅱ. 이론적 배경

본 논문에서 태양광 발전은 햇빛을 받아 전기를 발생하는 태양전지를 이용하여 직접 전기에너지로 변환시키는 기술이다. 이런 태양광 발전은 기후 데이터들에 의해서 발전량이 달라진다. 그렇기에 태양광 발전량 예측에 대한 관한 연구는 끊임없이 증가하고 있다.

1. 태양광 발전

태양광 발전은 무한정, 무공해의 태양 에너지를 직접 전기에너지로 변환시키는 발전 기술이다. 즉, 햇빛을 받으면 광전효과에 의해 전기를 발생하는 태양전지를 이용한 발전방식이다. 태양광 발전시스템은 태양전지로 구성된 모듈과 고촉전지 및 전력변환장치로 구성되어 있다.

태양광 발전은 <그림 2>와 같이 구성되어 있다. 태양전지는 태양에너지를 전기에너지로 변환할 목적으로 제작된 광전지로서 금속과 반도체의 접촉면 또는 반도체의 pn접합에 빛을 조사하면 광전효과에 의해 광기전력이 일어나는 것을 이용한 것이다. 일반적으로 태양광 모듈은 60개 또는 72개의 태양광 셀로 이루어져 있으며 이 셀들을 전기적으로 연결하여 내구성이 강한 유리 패널과 프레임으로 안전하게 만든 것이 바로 태양광 모듈이다. 여러 개의 모듈이 모여 강력하고 효율적인 태양광 발전 솔루션을 구성하게 되며, 태양광 모듈의 단자에서 생성된 전류는 연결된 케이블을 통해 인버터로 이동하며, 인버터는 전기 제품에 사용될 수 있도록 교류로 변환한다. 최근에는 발전량이 변동적인 태양광 에너지를 안정적으로 공급하기 위해 대용량 에너지 저장장치의 사용이 증가하는 추세이다.



<그림 2> 태양광 발전 시스템 (출처 : 한국전력공사)

2. 태양광 발전량 예측에 관한 선행 연구

태양광 발전은 태양의 위치에 따라 출력 변동이 심하고 출력 예측이 어려운 불규칙한 재생에너지로서 전체 신재생 에너지 공급에 차지하는 비중이 커질수록 전력계통에 안정적으로 연계할 수 있는 기술이 필요하다. 이에 따라 태양광 에너지 시설 최적 입지 연구, 태양광 설비 고장 예측 연구 등 태양광 발전에 관련된 다양한 연구가 진행되고 있다. 특히 태양광 발전량 예측에 관한 연구가 활발하게 되고 있다. 본 논문도 태양광 발전량 예측 연구를 진행하려 한다.

딥러닝 알고리즘 중에서 시계열 데이터 분석과 예측에 효과적인 RNN, LSTM을 이용한 모델들의 태양광 발전량 예측연구가 많이 있다. 시계열 데이터는 데이터의 input sequence의 차이에 의해 데이터 학습 모델이 달라진다. input sequence는 2년 이하의 단기부터 10년 이상의 장기로 구분할 수 있다. (신동하 et al.) RNN-LSTM을 이용한 태양광 발전량 단기 예측에서는 시계열 데이터 학습에 적합한 RNN과 LSTM에 비해 예측율이 많이 떨어지는 것을 확인할 수 있었으며, 다양한 입력 요소의 결합으로 보다 향상된 예측 결과를 도출할 수 있을 것으로 기대된다라고 언급하였다. 이처럼 1년 미만의 대부분의 연구에서는 태양광 발전량 데이터의 특성을 파악하기 어려워 5년 미만의 중기 데이터를 활용하였다. (김백천 et al.) 계절별 기상조건에 기반한 태양광 발전량 예측 연구에서는 계절적 특성을 고려하기 위해 총 4년의 데이터를 계절별로 분류하였고, RMSE(Root Mean Square Error)가 최대 54.05에서 최저 10.31의 값을 얻었다. (손혜숙 et al.) 기상 예보를 활용한 LSTM 기반 24시간 태양광 발전량 예측모델 연구에서는 .1시간 단위의 3년의 태양광 데이터를 사용하여 24시간 태양광 발전량 예측 모델을 제안하였다. 초기의 11개의 기상 변수에서 7개의 기상 변수를 활용하여 발전량 모델링 연구에서 변수 선택에 사용될 수 있음을 언급하였다. 위의 선행 연구를 <표 2>와 같이 input sequence 별로 연구를 정리 하였다..

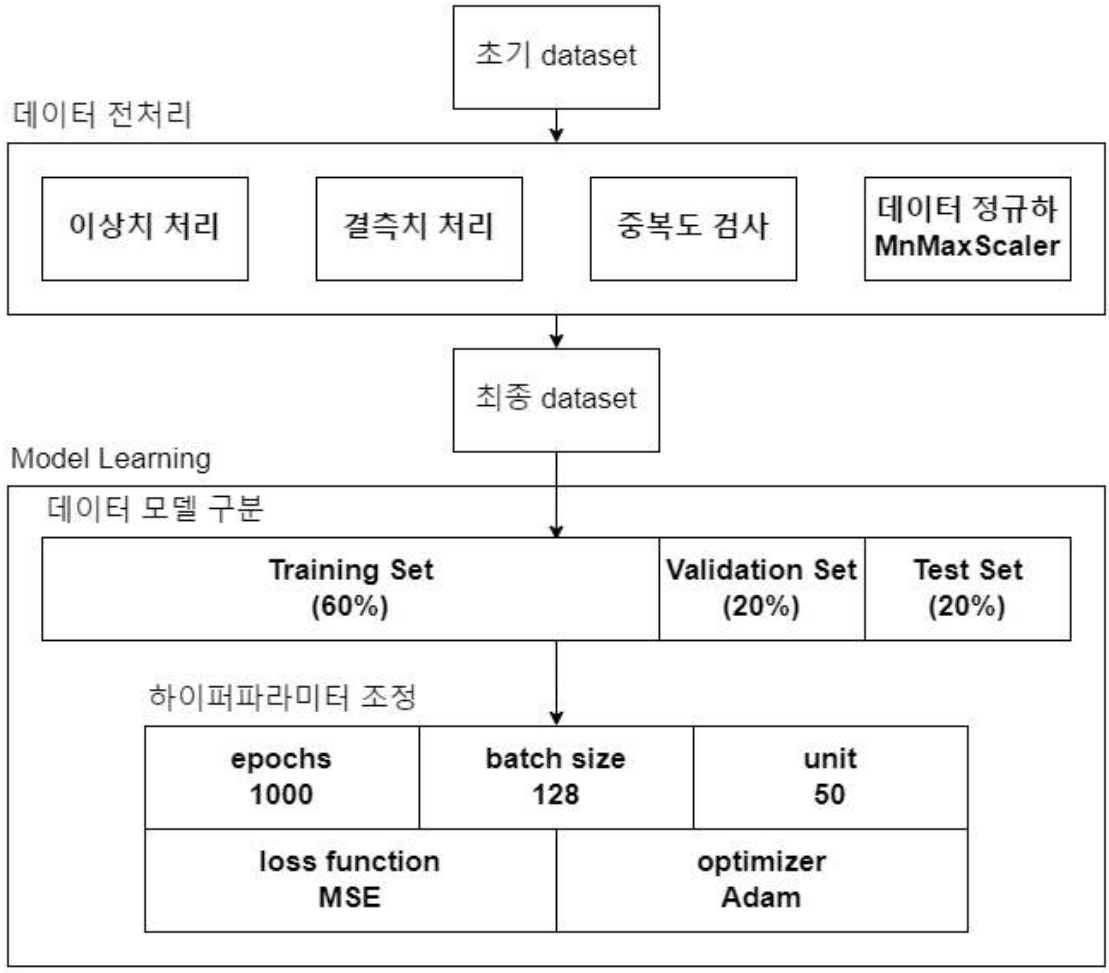
<표 2> input sequence별 선행 연구 표

구분	input sequence	방법론	저자
단기	2013 ~ 2015 (2년)	RNN, LSTM	신동하 et al.
중기	2016 ~ 2019 (4년)	LSTM	김백천 et al.
	2016.01 ~ 2018.12 (3년)	LSTM	손혜숙 et al.

본 논문에서도 시계열 예측에 효율적인 LSTM 알고리즘을 통하여 태양광 발전량 예측 모델을 생성한다. 본 논문의 데이터는 보다 태양광 발전량 데이터의 특성을 파악하기 쉬운 중기 데이터를 이용하고자 한다. 생성된 모델을 통하여 정확한 예측 값이 출력 되고, 안정적인 발전소 운영에 도모하고자 한다.

Ⅲ. 연구 방법

본 논문에서 사용한 연구 방법 모형도는 아래의 <그림 >와 같다. 초기의 dataset을 이상치 처리, 결측치 처리, 중복도 검사, MinMaxScaler를 이용한 데이터 정규화를 통해 데이터를 전처리 하였다. 이후 최종 dataset을 Training Set 60%, Validation 20%, Test Set 20%으로 나누었다. 이후 LSTM 알고리즘을 적용하기 위해 하이퍼파라미터를 조정하였다.



<그림 3> 연구 방법 모형도

1. 데이터 수집

본 논문에서는 사용한 태양광 발전량 데이터는 <표 3>와 같은 형태이다. 해당 데이터는 신재생 에너지 발전 통합 운영 관리 전문 기업인 (주)대연씨앤아이 제공되는 데이터이다. 2016년 12월 13일 14시부터 2020년 12월 31일 23시까지의 약 4년간 수집된 발전량, 기온, 습도, 이슬점 온도, 일사량, 구름양 데이터이다.

<표 3> 초기 Dataset 일부

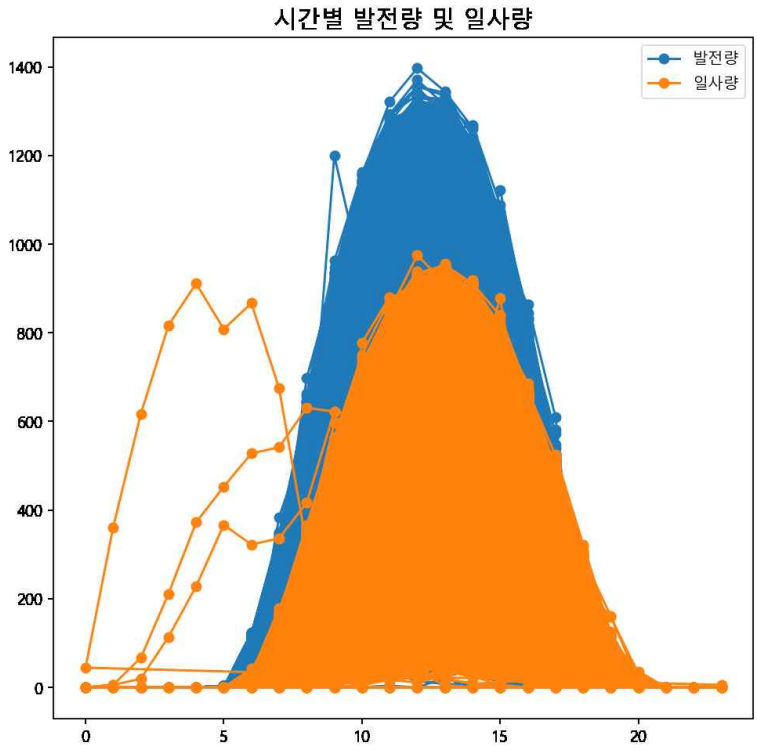
일자	발전량	월	시	기온	습도	이슬점 온도	일사량	구름양
2016-12-13-14	140.631	12	14	4.3	77	0.6	100	0
2016-12-13-15	85.793	12	15	4.3	80	1.1	72.222	0
2016-12-13-16	32.59	12	16	3.8	80	0.6	55.556	0
2016-12-13-17	0.154	12	17	3.4	80	0.2	13.889	0
2016-12-13-18	0	12	18	2.6	86	0.4	2.778	0
2020-12-31-19	0	12	19	-0.9	65	-6.6	0	8
2020-12-31-20	0	12	20	-1.5	66	-7	0	8
2020-12-31-21	0	12	21	-1.9	75	-5.7	0	8
2020-12-31-22	0	12	22	-2.7	73	-6.8	0	8
2020-12-31-23	0	12	23	-4	84	-6.3	0	5

<표 4>은 각 데이터 변수를 요약한 표이다. 스피어만 상관계수 분석에 따라 종속변수와 독립변수 간의 상관관계 분석을 진행하고, 상관계수를 나타내었다. 이 때, 독립변수는 기온, 습도, 이슬점 온도, 일사량, 구름양이며, 종속 변수는 발전량으로 고정하였다. 스피어만 분석에 의하면 상관계수의 값은 1부터 -1 사이의 값을 가지며, 상관계수의 절대 값이 커질수록 상관성이 높다고 한다. 상관계수가 양수일 경우 선형적으로 커지며, 음수일 경우 선형적으로 감소한다. 그 결과, 발전량과 일사량은 0.849의 큰 상관관계를 보였다. 이러한 수치처럼 일사량이 높아지면 발전량이 높아짐을 의미한다.

<표 3> 변수 요약 표

구분	최소	최대	평균	표준편차	상관계수
발전량	0	1396.853	238.973	361.765	-
기온	-12.9	39.2	13.678	10.0795	0.317
습도	0	100.0	67.674	23.528	-0.368
이슬점 온도	-26.9	28.0	6.833	11.880	0.067
일사량	0	975.0	144.024	225.609	0.849
구름양	0	10.00	3.040	3.948	0.009

<그림 4>을 보면 시간에 따라 발전량과 일사량의 그래프가 유사하며, 두 변수는 비슷하게 증가하는 것을 확인할 수 있다. 반대로 발전량과 습도의 상관계수는 -0.368로, 습도가 높아지면 발전량이 낮아짐을 의미한다.



<그림 4> 시간별 발전량 및 일사량

2. 데이터 전처리

데이터 전처리 작업은 데이터 분석의 품질에 큰 영향을 미친다. 분석하기 좋게 데이터를 고치는 작업을 해야 한다. 태양광 발전량 데이터의 이상치, 중복 데이터, 결측치 데이터를 처리하고 끝으로 데이터 정규화를 한다. 데이터 전처리가 완료된 Dataset을 별도로 저장한 것을 최종 Dataset으로 사용한다.

(1) 데이터 이상치 처리

각 변수들은 눈으로는 확인되지는 않지만 비정상적인 값을 가지는 경우가 있다. 이런 이상치를 확인하기 위해서 각 데이터를 IQR(InterQuartile range, 사분범위)를 계산하였다. IQR은 데이터를 4개의 동일한 부분으로 나누어서, $Q_3 - Q_1$의 값을 의미한다. 이 때, Q_1은 데이터의 25%가 이 값보다 작거나 같은 값이며, Q_3은 데이터의 75%가 이 값보다 작거나 같은 값을 의미한다. IQR을 계산하여 상한 제한선 혹은 하한 제한선을 넘어가는 값을 이상치로 판단하였다. 상한 제한선은 $Q_3 + 1.5 \times IQR$, 하한 제한선은 $Q_1 - 1.5 \times IQR$로 설정하였다. <표 5>를 통하여 각 변수들의 이상치를 확인할 수 있다. 제한선을 넘어가는 값들은 결측치 데이터로 분류하였고, 해당 데이터는 삭제 처리 하였다.

<표 5> 이상치 확인

구분	Q_1	Q_3	IQR	하한 제한선	상한 제한선
발전량	0.0	411.735	411.735	-617.602	1029.336
기온	5.0	21.3	16.3	-19.45	45.75
습도	51.0	90.0	39.0	-7.5	148.5
이슬점 온도	-2.2	17.1	19.3	-31.15	46.05
일사량	0.0	177.778	177.778	-266.667	444.445
구름양	0.0	7.0	7.0	-10.5	17.5

(2) 중복 데이터 처리

변수들은 눈으로는 확인되지는 않지만 중복된 값을 가지는 경우가 있다. 이런 중복된 값들은 데이터의 품질이 떨어지고, 왜곡된 값으로 치우칠 수도 있다. 태양광 데이터를 확인할 결과 중복된 데이터는 존재하지 않았다.

(3) 결측치 데이터 처리

데이터를 파일로 입력할 때 빠트리는 등 데이터가 소실되는 경우가 존재한다. 일반적으로 유효한 데이터 값이 존재하지 않는 결측치 데이터를 NaN으로 표시한다. 결측치 데이터가 많아지면 데이터의 품질이 떨어지고, 머신러닝 분석 알고리즘을 왜곡할 수도 있다. 결측치 데이터를 확인 한 결과 결측치 데이터는 존재하지 않았다.

(4) 데이터 정규화

머신러닝 모델은 학습 데이터를 기반으로 학습되기 때문에 반드시 테스트 데이터는 학습 데이터의 정규화 기준에 따라야 한다. 정규화를 하지 않는다면 수치가 다른 값에 매우 크거나 매우 작은 데이터에 영향을 줘서 정확한 결과를 얻어내기 어렵다. 데이터 정규화에는 다양한 방법이 있다. standardScaler 방법은 각 열의 값을 평균 0, 표준 편차 1로 하는 방법으로 데이터의 특징을 모르는 경우에 사용에 유용하다. RobustScaler 방법은 중앙값 0, IOR을 이용하여 정규화하는 방법이며 데이터 이상치가 많은 경우에 유용하다. 사용할 정규화 방법은 MinMaxScaler이다. MinMaxScaler는 각 변수의 최솟값과 최댓값을 기준으로 0에서 1 구간 내에 균등하게 값을 배정하는 정규화 방법이다. 태양광 데이터는 단위부터 최댓값과 최솟값의 범위가 차이가 심하기 때문에 MinMaxScaler 방법을 이용하였다. <표 6>은 MinMaxScaler 방법을 이용한 후 최종적인 변수 요약 표이다.

<표 6> MinMaxScaler 방법을 이용한 후 변수 요약 표

구분	최소	최대	평균	표준편차	상관계수
발전량	0	1	0.179	0.291	-
기온	0	1	0.407	0.165	0.338
습도	0	1	0.685	0.215	-0.384
이슬점 온도	0	1	0.536	0.282	0.058
일사량	0	1	0.119	0.205	0.848
구름양	0	1	0.287	0.388	0.016

최종 Dataset의 일부는 <표 7>과 같다. 전처리가 완료된 발전량 데이터를 정리해야 한다. 초기의 Dataset에서는 일자의 형태가 'YYYY-MM-DD-HH'의 Date 타입이 string이었다. 이러한 형태의 일자를 보다 쉽게 Date 타입하여 시계열 데이터 분석에 보다 쉽게 접근하기 위해 형태를 수정하였다. 바꾼 일자의 형태를 'YYYY-MM-DD HH:MM:SS' 이다. 현재의 Dataset에서는 기존 Dataset의 정보를 받았기 때문에 분(MM)과 초(SS)는 값이 모두 '00:00'으로 통합되어 있다.

<표 7> 최종 Dataset 일부

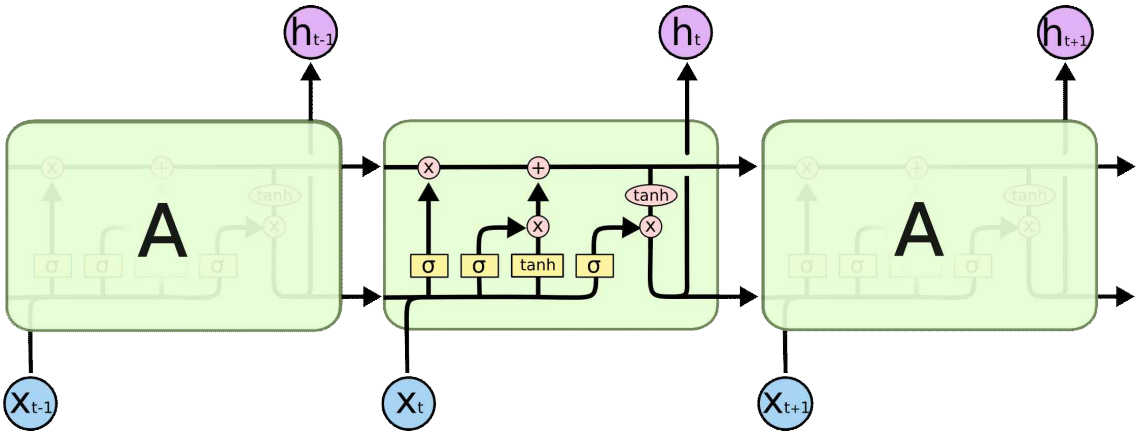
일자	발전량	월	시	기온	습도	이슬점 온도	일사량	구름양
2016-12-13 14:00:00	0.136635	12	14	0.267	0.77	0.396	0.105	0
2016-12-13 15:00:00	0.083355	12	15	0.267	0.80	0.414	0.076	0
2016-12-13 16:00:00	0.031664	12	16	0.256	0.80	0.396	0.058	0
2016-12-13 17:00:00	0.000150	12	17	0.248	0.80	0.381	0.015	0
2016-12-13 18:00:00	0	12	18	0.231	0.86	0.388	0.003	0
2020-12-31 19:00:00	0	12	18	0.158	0.65	0.137	2.778	0.8
2020-12-31 20:00:00	0	12	19	0.145	0.66	0.122	0	0.8
2020-12-31 21:00:00	0	12	20	0.137	0.75	0.169	0	0.8
2020-12-31 22:00:00	0	12	21	0.120	0.73	0.129	0	0.8
2020-12-31 23:00:00	0	12	22	0.092	0.84	0.148	0	0.5

3. LSTM 알고리즘

LSTM(Long Short-Tem Memory)은 RNN(Recurrent Neural Network)의 한 종류로, 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완한 신경망 구조이다.

LSTM은 <그림 5>와 같이 반복 모듈은 단순한 한 개의 tanh layer가 아닌 4개의 layer가 서로 정보를 주고받는 구조로 되어 있다. LSTM 셀에서는 상태(state)가 크게 두 개의 벡터인 ht(단기 상태)와 ct(장기 상태)로 나눌 수 있다. LSTM의 반복 모듈에는 4개의 상호작용하는 layer가 들어있다. Cell state는 정보가 바뀌지 않고 그대로 흐르도록 하는 역할을 한다. LSTM은 cell state에 뭔가를 더하거나 없앨 수 있는 능력이 있는데, 이 능력은 gate라고 불리는 구조에 의해서 된다. Gate는 정보가 전달될 수 있는 추가적인 방법으로, sigmoid layer와 pointwise 곱셈으로 이루어져 있다.

Forget gate는 cell state에서 sigmoid layer를 거쳐 어떤 정보를 버릴 것인지 정한다. Input gate에서 앞으로 들어오는 새로운 정보 중 어떤 것을 cell state에 저장할 것인지를 정한다. 먼저 sigmoid layer를 거쳐 어떤 값을 업데이트할 것인지를 정한 후, tanh layer에서 새로운 후보 Vector를 생성한다. 이전 gate에서 버릴 정보들과 업데이트할 정보들을 정했다면, Cell state update 과정에서 업데이트를 진행한다. Output gate는 어떤 정보를 output으로 내보낼지 정한다. 먼저 sigmoid layer에 input data를 넣어 output 정보를 정한 후 Cell state를 tanh layer에 넣어 sigmoid layer의 output과 곱하여 output으로 내보낸다.



<그림 5> LSTM 구조 (출처 : Christopher Olah)

(1) 데이터 모델 구분

LSTM 모델은 출력 관찰에 대한 입력으로 과거 관찰 시퀀스를 매핑하는 기능을 학습한다. 모델 학습을 위하여 2016년 12월 ~ 2020년 12월의 데이터에서 <표 8>과 같이 Dataset을 구분하였다. 60%는 Training Set으로, 20%는 Validation Set으로, 나머지 20%는 Test Set으로 하였다.

<표 8> Dataset 구분 표

구분	비율	기간
Training Set	60%	2016.12.13. ~ 2019.12.05
Validation Set	20%	2019.12.06. ~ 2020.03.05
Test Set	20%	2020.03.06. ~ 2020.12.31

(2) 하이퍼파라미터 조정

네트워크를 구성할 때 조정해야 하는 하이퍼파라미터(Hyperparameter)는 epochs, batch size, unit, 활성화 함수, 손실함수이 있다. epoch는 전체 Training Set에 대해 한 번 학습을 완료한 횟수를 의미하며, batch size는 모델의 가중치를 한 번에 업데이트시킬 때 사용되는 묶음을 의미한다. unit은 LSTM의 hidden layer의 수를 의미한다. 활성화 함수는 모델에 비선형성을 갖게 하여 선형 분리 불가능한 데이터에 대응하기 위해 사용한다. 모델 학습에서는 최대한 틀리지 않는 방향으로 학습해 나가야 하는데, 여기서 얼마나 틀리는지 알게 하는 함수가 손실함수이다. 이러한 손실함수의 최솟값을 찾는 것을 최적화라고 하고, 수행하는 알고리즘을 최적화 알고리즘이라고 한다. 하이퍼 파라미터의 값에 따라 모델의 성능이 달라지기 때문에 적절하게 값을 주어야 한다.

활성화 함수는 종류가 tanh, Sigmoid, ReLU, Swish 등이 있다. tanh는 결과 데이터의 평균이 0이며 Vanishing gradient 현상(input값이 어느정도 크거나 작으면 기울기)이 발생하는 문제가 있다. Sigmoid는 tanh와 같이 Vanishing gradient 현상을 발생시키지만, 결과 값을 0에서 1 사이로 만들어 준다. ReLU 활성화 함수는 가중치를 곱하여 결과값을 0으로 나타나게 하는 방법이다. Swish 활성화 함수는 ReLU를 대체하기 위해 고안된 함수로서, CNN 아키텍쳐 가운데 하나인 mobilenet을 학습시키는데 보통 사용된다.

손실함수도 binary crossentropy (이항 교차 엔트로피), categorical crossentropy (범주형 교차 엔트로피), sparse categorical crossentropy, means squared error(MSE, 평균 제곱 오차 손실) 등이 존재한다. binary crossentropy는 이진 분류

기를 훈련할 때 자주 사용된다. categorical crossentropy에서는 출력을 클래스 소속 확률에 대한 예측 문제에서 자주 사용된다. sparse categorical crossentropy는 ategorical crossentropy와 유사하게 여러 개의 클래스 분류에서 사용된다. MSE는 연속값인 dataset에서 사용되는 함수이기 때문에 국내 태양광 발전량 dataset에 적합하다.

최적화 함수는 Adam, RMSprop, SGD, Adadelta 등이 있다. SGD(Stochastic Gradient Descent, 확률적 경사 하강법)는 최적의 매개변수의 기울기를 구해, 기울어진 방향으로 매개변수의 값을 반복해서 갱신하는 것을 의미한다. RMSprop 함수는 학습이 진행될수록 학습률이 감소되는 문제점을 해결하기 위해 나타났다. Adam 함수는 현재까지 계산해온 기울기의 지수 평균을 저장하고, 기울기의 제곱값의 지수 평균을 저장하는 방식으로 사용한다. Adaadelta 함수는 기울기의 이동 평균을 구하는 방법으로 한 번의 업데이트를 위해 모든 데이터가 계산에 포함되기 때문에 속도가 매우 느리다.

<표 9> 하이퍼파라이터 조정 표

구분	
epochs	**100** (10, 50, 100, 1000)
batch size	**128** (32, 64, 128, 256)
unit	**50** (10, 50, 100)
활성화 함수	**Sigmoid**(tanh, Sigmoid, ReLU, Swish)
손실함수	**MSE**
최적화 함수	**Adam** (Adam, RMSprop, SGD, Adadelta)

(3) 시간 조절

본 논문은 딥러닝 기반의 LSTM 모델을 구축한 이유는 시간별 태양광 발전량 예측을 위해서이다. 시계열 데이터는 데이터의 input sequence의 차이에 의해 데이터 학습 모델이 달라진다. inut sequence 별로 차이를 두어 LSTM 모델을 각각 이용해 보았다. 5시간, 10시간, 20시간, 22시간, 23시간, 24시간을 기준으로 이용하였다.

Ⅳ. 연구 결과

본 논문에서는 데이터 모델 성능 평가 방법으로 RMSE와 R-squared를 통해 확인하였다. 시간별로 데이터 모델을 평가한 결과 23시간을 기준으로 한 것이 가장 성능이 좋은 데이터 모델이었다.

1. 데이터 모델 성능 평가 방법

본 논문에서는 시간별 태양광 발전량 예측을 위해 딥러닝 기반의 LSTM으로 모델을 구축하였다. 본 논문의 데이터 모델 성능 평가는 RMSE와 R-squared를 통해 확인해 볼 것이다.

(1) RMSE(Root Mean Squared Error)

MSE(Mean Squared Error)는 모델의 예측값과 실제값 오차의 제곱을 하는 방법이다. 이를 아래의 식과 같이 루트를 씌워 사용한 값이 RMSE이다. RMSE를 사용하면 오류 지표를 실제값과 유사한 단위로 다시 변환하여 해석을 쉽게 한다. 또한 예측 대상의 크기에 영향을 바로 받는다. 이 때 x는 input, y는 output, $\hat{y}$는 예측값, n은 dataset의 개수를 의미한다.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum(\hat{y} - y)^2}{n}}$$

(2) R-squared(Coefficient of determination, 결정계수)

R-squared는 현재 사용하고 있는 x가 y의 분산을 얼마만큼 줄였는지 확인 할 수 있는 평가 기준이다. 즉 y의 평균값 모델을 사용했을 때 대비 x 변수를 사용함으로써 얻는 성능 향상의 정도이다. R-squared는 아래의 식으로 구성되어 있다. SSE는 관측값과 예측값의 차이, SSR은 예측값과 평균의 차이, SST는 관측값과 평균의 차이이다. R-squared의 값은 1에 가까우면 데이터 모델의 성능이 우수하다고 평가할 수 있다. 이 때 y_i : output, $\bar{y}$: output 평균, $\hat{y_i}$: 예측값, n: dataset 개수을 의미한다.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

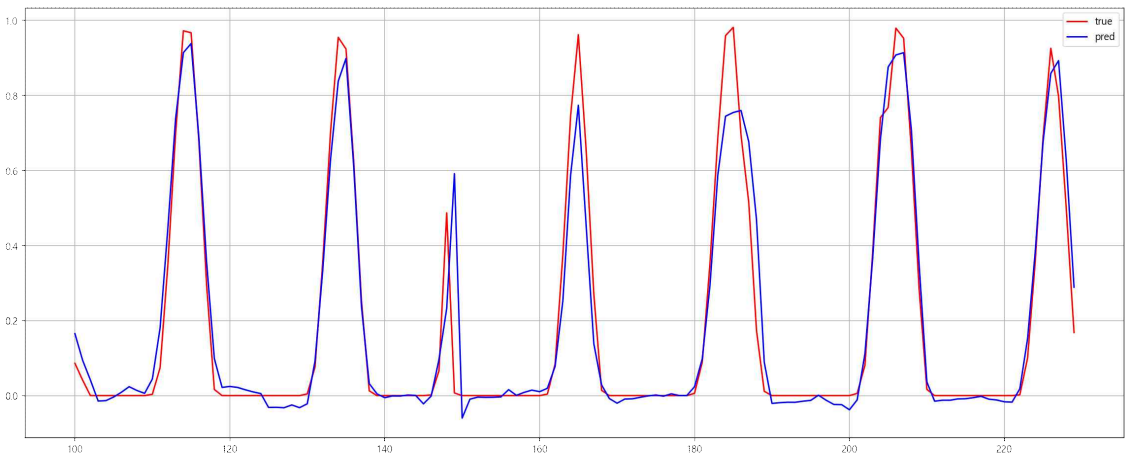
2. 데이터 모델 성능 평가

본 연구 모델에서의 시간별 예측에 대한 성능 평가 결과는 다음 <표 10>이다. 5시간, 10시간, 20시간, 22시간, 23시간, 24시간을 기준으로 LSTM 모델 성능 평가 하였다. 앞에서의 최적의 하이퍼파라미터를 동일하게 적용하였다. epochs는 1000, batch size는 128, unit은 50, 활성화 함수는 Sigmoid, 손실 함수는 MSE, 최적화 함수는 Adam으로 적용하였다. RMSE는 시간과 상관 없이 0.1의 값이 나왔지만, R-Squared에서는 23시간을 기준으로 하였을 때 최대의 성능이 나왔다.

<표 10> 모델 성능 평가 결과

구분	5시간	10시간	20시간	22시간	23시간	24시간
RMSE	0.1	0.1	0.1	0.1	0.1	0.1
R-Squared	78.7	78.7	85.2	86.9	87.1	85.4

아래의 <그림 6>은 Test Set에서 무작위로 추출한 날들의 실제 값과 모델에서의 예측 값을 비교하는 그래프이다. 실제 발전량 값을 빨간색으로, 모델에서 예측한 값을 파란색으로 표현하였다. 기간은 무작위로 2016.12.19.부터 2016.12.24.까지 선택하였다.



<그림 6> LSTM 알고리즘의 예측 결과

Ⅴ. 결론

본 논문은 LSTM을 이용한 국내 태양광 발전량 예측을 하는 것이 목적이었다. 태양광 데이터는 기온, 습도, 일슬점 온도, 구름양, 일사량 등 다양한 기후 조건에 따라 간헐적이고, 극심한 차이를 보이고 있다. 태양광 발전의 효율성과 정확한 발전량 예측이 요구된다. 이에 따라 기후 조건에 따른 시간별의 발전량을 통하여 예측하였다.

초기의 국내 태양광 발전량 dataset는 (주)대연씨앤아이 제공하는 데이터이다. 이러한 데이터를 이상치, 중복 데이터, 결측치 데이터, MinMaxScaler를 통하여 전처리를 하였다. 이후 데이터를 Training Set을 60%, Validation Set을 20%, Test Set을 20% 나누어서 LSTM 모델을 적용하였다. LSTM은 시계열 데이터에 적합한 알고리즘이다. LSTM의 하이퍼파라미터를 조정해서 보다 좋은 성능의 모델을 만들었다. 끝으로 RMSE와 R-Squared를 통하여 성능을 평가하였다.

5시간, 10시간, 20시간, 22시간, 23시간, 24시간을 기준으로 생성한 LSTM 모델을 적용하였고, 0.1의 RMSE로 동일하였지만, 23시간을 기준으로 하였을 때는 87.1의 R-Squared 수치로 최대의 성능을 보였다. 정확한 예측을 위해서 본 데이터의 기후 조건들을 중요 변수만을 판별할 필요성이 있다.

<표 11> 모델 최대 성능 평가 결과

RMSE	R-squared
118.5	85.1

위와 같은 연구가 추가적으로 이루어진다면 본 모델의 예측 성능을 더욱 향상 시킬 수 있고, 태양광 에너지에도 활성화 될 수 있다.

참고문헌

1. 국내 문헌

(1) 논문

김용수·이상현·김홍원(2019), "기상정보를 활용한 LSTM 기반 태양광 발전량 예측 기법", 한국통신학회논문지

손혜숙·김석연·작윤(2018) "기상 예보를 활용한 LSTM 기반 24시간 태양광 발전량 예측모델 연구"

신동하·김창복(2018), "RNN-LSTM을 이용한 태양광 발전량 단기 예측 모델", 한국항행학회논문지

김백천·정승환·김민석·김종근·김성신(2021) "계절별 기상조건에 기반한 태양광 발전량 예측에 관한 연구"

김정원(2019), "기상 예보를 이용한 머신러닝 알고리즘 기반 태양광 발전량 예측 기법", 한국정보화기술학회논문지

(2) 학위논문

최다빈 "딥러닝을 활용한 시간별 태양광 발전량 예측." 국내석사학위논문 충북대학교, 2021. 충청북도

유복종 "태양광발전시스템의 발전량 예측 정확도 향상 연구." 국내박사학위논문 한양대학교 대학원, 2018. 서울

이용규 "발전소 데이터를 이용한 태양광발전량의 선형회귀모델 예측분석." 국내석사학위논문 중남대학교 에너지과학기술대학원, 2022. 대전

안연주 "기상 매개 변수를 포함한 LSTM 기반 태양광 전력 시스템의 발전량 예측." 국내석사학위논문 한경대학교, 2021. 경기도

(3) 기타자료

산업통상자원부(2017), "재생에너지 3020 이행계획".

산업통상자원부(2020), "재생에너지 발전량 예측제도 도입".

안재균(2018), "신재생에너지 보급 확산을 대비한 전력계통 유연성 강화방안 연구", 에너지경제연구원.

한국과학기술정보연구원(2002), "태양광발전 시스템".

2. 국외 문헌

(1) 기타자료

IEA(2020a), *'Renewables 2020'*.

IEA(2020b), *'Korea 2020 Energy Policy Review'*.

Olah, Christopher(2015), *'Understanding LSTM Networks'*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/.