Research Paper

Comparison Between Random Forest and Recurrent Neural Network for Photovoltaic Power Forecasting

Ramek Kim¹ · Kyungmin Kim² · Johng-Hwa Ahn^{1,2,*}

¹Department of Environmental Engineering, Kangwon National University

²Department of Integrated Energy and Infra System, Kangwon National University

(Received January 21, 2021; Revised March 24, 2021; Accepted May 10, 2021)

Objectives : Photovoltaic power generation which significantly depends on meteorological conditions is intermittent and unstable. Therefore, accurate forecasting of photovoltaic power generation is a challenging task. In this research, random forest (RF), recurrent neural network (RNN), long short term memory (LSTM), and gated recurrent unit (GRU) are proposed and we will find an efficient model for forecasting photovoltaic power generation of photovoltaic power plants.

Methods : We used photovoltaic power generation data from photovoltaic power plants at Gamcheonhang-ro, Saha-gu, Busan, and meteorological data from Busan Regional Meteorological Administration. We used solar irradiance, temperature, atmospheric pressure, humidity, wind speed, wind direction, duration of sunshine, and cloud amount as input variables. By applying the trial and error method, we optimized hyperparameters such as estimators in RF, and number of hidden layers, number of nodes, epochs, and validation split in RNN, LSTM, and GRU. We compared proposed models by evaluation indexes such as coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE).

Results and Discussion : The best RF at 1,000 of number of decision tree achieved test $R^2=0.865$, test RMSE=16.013, and test MAE=9.656. The best choice of RNN was 6 hidden layers and the number of nodes in each layer was 90. We set the epochs at 450. RNN achieved test $R^2=0.942$, test RMSE=10.530, and test MAE=6.390. To find the best result of LSTM, we used 3 hidden layers, and the number of nodes was 600. The epochs were set to 200. LSTM achieved test $R^2=0.944$, test RMSE=10.29, and test MAE=6.360. GRU was set to 3 hidden layer and the number of nodes was 450. The epochs were set to 500. GRU achieved test $R^2=0.945$, test RMSE=10.189, and test MAE=5.968.

Conclusions : We found RNN, LSTM, and GRU performed better than RF, and GRU model showed the best performance. Therefore, GRU is the most efficient model to predict photovoltaic power generation in Busan, Korea.

Keywords : Gated Recurrent Unit, Long Short-Term Memory, Photovoltaic Power Forecasting, Random Forest, Recurrent Neural Network

The Korean text of this paper can be translated into multiple languages on the website of <http://jksee.or.kr> through Google Translator.

* Corresponding author

E-mail: johngghwa@kangwon.ac.kr

Tel: 033-250-6357 Fax: 033-259-5550

© 2021, Korean Society of Environmental Engineers

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

연구논문

태양광 발전량 예측을 위한 랜덤포레스트와 순환신경망 비교

김라맥¹ · 김경민² · 안종화^{1,2,*}

¹강원대학교 공과대학 환경공학과

²강원대학교 신산업개발 T-EMS 융합학과

목적: 태양광 발전량은 기상상태에 따라 변동성이 크기 때문에 정확한 예측은 중요하다. 이에 본 연구에서는 랜덤포레스트(random forest, RF), 순환신경망(recurrent neural network, RNN), 장단기 메모리(long-short term memory, LSTM), 게이트 순환 유닛(gated recurrent unit, GRU) 등을 이용해 태양광 발전량 예측을 위한 효과적인 모델을 도출하고자 한다.

방법: 본 연구에서는 부산광역시 사하구 감천항로 태양광발전단지의 태양광 발전량 자료와 부산지방기상청의 기상자료를 이용하였다. 입력 인자로는 일사량, 온도, 기압, 습도, 풍속, 풍향, 일조시간, 운량 등을 사용하였다. 시행착오법을 이용하여 RF의 경우 트리 개수, RNN, LSTM, GRU의 경우 은닉층의 수, 노드의 수, 학습 횟수, 학습 자료 내 검증 자료 비율 등 모델별 하이퍼파라미터(hyperparameter)를 조절하여 최적화하였다. 모델의 예측 성능을 비교, 평가하기 위해 테스트 결정계수(coefficient of determination, R^2), 평균제곱근오차(root mean square error, RMSE), 평균절대오차(mean absolute error, MAE)를 사용하였다.

결과 및 토의: RF의 경우 결정 트리를 1,000개로 고정하였을 때 성능이 가장 좋았으며 $R^2=0.865$, RMSE=16.013, MAE=9.656이 나왔다. RNN의 경우 은닉층 6개, 노드의 수 90개, 학습 횟수 450회에서 가장 좋은 결과를 가져왔다. 그 결과 $R^2=0.942$, RMSE=10.530, MAE=6.390으로 나타났다. LSTM은 은닉층 3개, 노드의 수 600개, 학습 횟수 200회로 최적화할 수 있었다. 결과 값은 $R^2=0.944$, RMSE=10.29, MAE=6.360으로 도출되었다. GRU는 은닉층 3개, 노드의 수 450, 학습 횟수 500으로 고정하였다. 이때 성능평가 결과는 $R^2=0.945$, RMSE=10.189, MAE=5.968이었다.

결론: 본 연구에서 태양광 발전량 예측 성능은 RF보다 순환신경망 계열 모델이 뛰어났다. RNN, LSTM, GRU 중에서는 GRU의 성능이 가장 우수하였다. 따라서 본 연구에서는 GRU 모델이 부산지역의 태양광 발전량을 예측하는데 적절한 모델임을 알 수 있다.

주제어: 게이트 순환 유닛, 랜덤포레스트, 순환신경망, 장단기 메모리, 태양광 발전량 예측

1. 서론

전 세계적으로 에너지 수요가 증가함에 따라 화석연료에 대한 의존도가 높아지고 있다. 화석연료는 온실가스를 배출해 지구온난화와 같은 기후변화를 일으켜 심각한 문제를 불러일으키고 있다.¹⁾ 이러한 화석연료를 대체하기 위하여 최근 오염물질이 없는 청정한 신재생 에너지가 주목받고 있다. 국내에선 ‘재생에너지 3020’ 정책하에 2030년까지 전체 전력 생산 중 신재생 에너지 생산량을 20%까지 상승시키는 계획을 발표했다.²⁾ 이중 태양광 발전은 무한한 태양에너지 자원을 쓴다는 장점이 있다.³⁾ 2018년 국내 신재생에너지 발전량은 전체 발전량의 8.88%에 해당하여 전년 대비 13.07%가 증가하였다.⁴⁾ 이중 태양광 발전량은 신재생에너지의 17.5%를 차지하여 전년도 대비 30.5%가

증가하여 앞으로도 지속적인 수요 및 성장이 예상된다. 태양광 발전량은 기상 상태에 의존하기 때문에 매우 간헐적이며 계절 변화에 큰 영향을 받는다.⁵⁾ 이는 전력량과 수요 사이의 불균형 문제를 초래하여 지역에 불안정한 전력공급 등 다양한 문제를 일으킬 수 있다.⁶⁾ 따라서 정확한 태양광 발전량 예측을 위한 모델 개발은 기존 태양광 발전 시스템의 효과적인 운영 및 추후 발전 사업계획의 정확성 향상 등을 이끌어 발전량의 과잉 공급, 부족 등의 불확실성을 해결할 수 있다.⁷⁾

태양광 발전량 예측모델 개발 현황은 현재까지 개발된 모델의 실제 구현을 위한 물리적 모델 설계(physical modeling), 자료기반 예측 등을 통한 통계적 모델 설계(statistical modeling), 기존 모델과 통계적 모델의 결합 등을 시도하는 앙상블 모델 설계(ensemble modeling) 등 크게 3가지로 나눌 수 있다.⁸⁾ 물

리적 모델 설계 사례로는 수치기상예보(numerical weather prediction, NWP)가 있으며, 주어진 시간 척도의 대기조건 정보를 받아 모델의 예측 성능을 향상시키는 모델이다. 통계적 모델 설계는 설계자가 지정한 수식 등을 이용하여 수동적으로 계산하는 기존 수학적 모델(numerical model)과 다르게 무작위, 시계열 자료 등 과거 자료를 기반으로 스스로 학습하여 모델을 구축한 다음 이를 이용하여 발전량을 예측하게 설계한다.⁹⁾ 대표적으로 기계학습(machine learning) 및 심층학습(deep learning) 등이 있으며 이러한 학습 모델로 서포트 벡터 머신(support vector machine, SVM), 인공신경망(artificial neural network, ANN) 등이 있다. 앙상블 기법은 개발된 다양한 모델을 합쳐 더 효율적인 예측을 수행하는 모델을 개발하는 것으로, 개별 모델이 가지는 한계를 극복하기 위해 다른 모델들을 결합시킨다. 대표적인 앙상블 모델로는 다수의 의사결정나무(decision tree, DT)를 이용하는 랜덤 포레스트(random forest, RF)가 있다.

기계학습은 기존 수학적 모델이 가지고 있던 한계인 높은 차원의 복잡한 비선형적 특징에 대한 해석을 보다 효과적으로 수행하여 정확한 예측값을 도출할 수 있다는 장점이 있다.¹⁰⁾ 따라서 기계학습은 현재 태양광발전량을 예측하는 데 다양한 모델로 시도되고 있다. 선행 연구에선 ANN과 SVM을 이용하여 기계학습 모델 설계를 시도하였다.^{8,11)} 하지만, 기계학습 중 RF^{12,13)}, 순환신경망(recurrent neural network, RNN)¹⁴⁾, 장단기 메모리(long-short term memory, LSTM)¹⁵⁻¹⁷⁾, 게이트 순환 유닛(gated recurrent unit, GRU)¹⁸⁾ 등을 태양광 발전량 예측에 사용한 문헌은 모델별 1~3편으로 제한적이었으며, RF와 순환신경망 계열을 비교한 논문은 찾아볼 수 없었다. 따라서 정확한 예측모델 선정을 위하여 다양한 기계학습 알고리즘을 이용한 태양광 발전량 예측 비교연구가 필요하다.

본 연구에선 RF, RNN, LSTM, GRU 등을 이용하여 부산 사하구 감천항로 지역에 있는 남부발전의 부산 태양광발전단지 태양광 발전량 예측모델을 개발하여 성능을 비교하고자 한다. 특히, 기존 순환신경망 계열 모델의 성능 비교연구에선 GRU에 최적화한 하이퍼파라미터(hyperparameter)를 RNN과 LSTM에 적용하여 비교했으나¹⁸⁻²⁰⁾, 본 연구에선 하이퍼파라미터를 모델별로 최적화하여 비교하고자 한다.

2. 실험 방법

2.1. 모델

2.1.1. 랜덤포레스트

RF는 앙상블을 기반으로 한 기계학습 모델로 DT가 가진 학습 자료에 대한 성능과 테스트 자료에 대한 성능 차이가 큰 과대 적합(overfitting), 낮은 정확도 등의 문제를 개선하기 위해 개발되었다.²¹⁾ RF는 주어진 자료에서 무작위로 자료를 추출하고 다시 복원하기를 반복하여 자료를 분배하는 bagging (bootstrap aggregating)을 통하여 과대 적합을 효과적으로 예

방한다(Fig. 1(a)). 이후 다수의 DT에 자료를 대입하여 얻어진 결과를 다수결 투표(majority voting)를 통하여 최적 DT를 선정하여 예측값을 도출해낸다.

2.1.2. 순환신경망

RNN은 시간적 관계에 있는 자료를 인식하는데 비효율적인 인공신경망을 보완하기 위해 개발되었다.²²⁾ RNN은 이전 연산의 결과인 파라미터(parameter)의 가중치를 이용하여, 은닉층(hidden layer)에서 같은 작업을 반복하기 때문에 순환이라고 불린다(Fig. 1(b)).²³⁾ 따라서 RNN은 필기 인식이나 음성 인식과 같은 자료나 시간의 흐름에 따라 기록된 시계열 자료 등과 같은 연속성 자료(sequential data)에 좋은 성능을 보인다고 알려져 있다.²⁴⁾ RNN은 자료의 양이 많아질수록 학습량이 증가하여 학습 시간이 증가한다. 따라서 RNN은 비교적 적은 양의 시계열 자료 처리에 효과적이다.

2.1.3 장단기 메모리

RNN은 모델 내 계산 과정의 특성상 자료의 계산이 반복될수록 가중치가 점점 줄어들어 0에 수렴하는 경사 하강 소실 (gradient vanishing) 현상이 발생하여 장기 기억에 취약하다는 단점이 있다. LSTM은 RNN에서 발생한 한계에 대하여 계산 과정에서 다른 파라미터를 이용하여 이를 보완하여 효과적인 시계열 자료 예측을 위하여 개발되었다.²⁵⁾ LSTM은 정보를 메모리 셀에 들여보낼 것인지 내보낼 것인지를 통제하는 forget gate, input gate, output gate를 통하여 연산을 수행한다 (Fig. 1(c)).²⁶⁾ Forget gate는 sigmoid 함수를 이용하여 기존 파라미터를 계산한 다음 이를 통해 기존 자료에 대한 활용 여부를 결정한다. Input gate에선 파라미터에 줄 정보를 sigmoid, hyperbolic tangent 등의 함수를 이용하여 연산한 다음 적용한다. 마지막으로 output gate는 계산 결과를 출력하여 다음 층의 노드로 보내게 된다.²⁷⁾ 이를 통해 기존 RNN 모델의 단점을 효과적으로 해결하였다.

2.1.4. 게이트 순환 유닛

GRU는 LSTM을 개량한 모델로 LSTM보다 적은 게이트(gate)를 이용하기 때문에 연산 과정이 단순하다는 장점이 있다. 세 가지 게이트를 이용한 LSTM과 다르게 GRU는 두 가지 게이트인 reset gate, update gate로 구성된다(Fig. 1(d)).²⁸⁾ Reset gate는 이전 파라미터의 반영 여부를 결정한다. 이를 이용하여 update gate에선 이전 파라미터와 현재 자료의 비율을 결정하여 연산을 수행한 다음 계산 결과를 전달한다. 이를 통하여 기존 3개의 파라미터와 3개의 게이트를 거치며 복잡한 연산을 수행하던 LSTM과 다르게 GRU는 2개의 파라미터, 2개의 게이트만을 사용하여 모델을 단순화가 가능하였으며, 더 빠른 연산을 수행할 수 있다.

Fig. 1. Structure of (a) random forest, (b) recurrent neural network, (c) long short-term memory, and (d) gated recurrent unit.

2.2. 자료 설정

본 연구에선 부산지역의 태양광 발전량을 예측하기 위해 태양광 발전량 자료와 기상 자료를 이용하였다(**Fig. 2**). 태양광 발전량 자료는 부산광역시 사하구 감천항로에 있는 남부발전의 부산 태양광발전단지에서 수집하였으며, 기상 자료는 부산지방기상청에서 수집하였다. 수집한 자료는 시간별 자료로 2015년 1월부터 2019년 11월까지 총 23,095개였다. 2015년 1월부터 2018년 11월까지 학습(Training)으로 사용하였고 2018년 12월부터 2019년 1월까지 확인(Test)으로 사용하였다(**Fig. 3**). 자료의 전처리는 먼저 계측의 운영에 있어 발생한 결측치를 제거했으며, 그다음 Scikit-learn 라이브러리

Fig. 2. Map of the power plants located in Busan, South Korea.

Fig. 3. Photovoltaic power output data from 2015/01/01 until 2019/11/30.

의 `standardscaler` 함수를 이용하여 평균을 0, 표준편차를 1로 하는 표준화 작업을 수행하였다. 사용한 기상 자료와 발전량 자료 모두 오전 6시부터 오후 6시로 설정하였고 나머지 시간 자료는 모두 제거하였다.

2.3. 입력 자료

결측치를 제거하고 최종적으로 일사량, 온도, 기압, 습도, 풍속, 풍향, 일조시간, 운량 등의 기상 자료들을 입력 인자로 적용하였다. 일사량은 표면에 도달하는 태양 복사에너지의 양으로, 태양광 발전량을 예측하는 데 가장 중요한 요소이다.^{29,30)} 온도는 상승할 때 시간당 도선에 흐르는 전하의 양을 증가시켜 전압이 감소하게 된다. 이로 인해 온도가 상승하면 태양광 발전량은 감소하게 되는 경향성을 보여준다.³¹⁾ 기압은 온도와 반비례 관계에 있어 태양광 발전량 예측을 위해 고려되어야 한다.³²⁾ 습도는 태양 빛이 공기 중의 물방울을 통해 굴절, 반사, 확산할 수 있어 일사량을 감소시키는 영향을 준

다.³³⁾ 따라서 습도가 증가하면 태양광 발전량이 감소하는 경향이 있다.³⁴⁾ 풍속과 풍향은 열을 분산시키는 데 기여해 태양광 발전 패널의 열을 낮추어 발전량에 영향을 준다.³⁴⁾ 일조시간의 경우 일사량 측정에 가장 우선하여 고려되는 인자로서 일사량과 태양광 발전량을 예측하는데 고려되는 요소이다.³⁵⁾ 운량은 햇빛을 가려 태양광 발전량의 변동성에 가장 주된 영향을 미치는 요소이다. 운량은 일사량을 감소시켜 발전량을 빠르게 감소시킨다.^{36,37)}

2.4. 모델 최적화 및 성능 지표

모델 최적화를 위하여 모든 모델의 하이퍼파라미터를 조절하였다. RF의 경우 트리 개수(estimators)를 통해 DT의 개수를 조절할 수 있으며, 각 트리의 최대 깊이(depth)에 대한 제어, 최소 분할 자료 수 조정 등을 통하여 최적화가 가능하다. 본 연구에선 트리의 최대 깊이는 출력값인 발전량을 얻을 때까지 구성하는 것으로 하였으며, 최소 분할 자료 수는 과적합 제어를 위하여 2로 하였다. 트리 개수의 경우 시행 착오법(trial & error method)을 통하여 최적 개수를 결정하였다. 심층학습 모델인 RNN, LSTM, GRU의 경우 은닉층 수, 은닉층 내 노드 개수, 학습 반복 횟수(epoch), 학습 자료 내 검증자료 비율 등을 조절 가능하며 본 연구에선 이를 시행 착오법으로 모델별 최적화를 하였다. 활성화 함수는 모든 심층학습 모델에 rectifier linear unit (ReLU)를 적용하였으며 층별 노드는 동일하게 구성하였다. 모든 모델은 Python 환경에서 설계되었으며, RF 모델의 경우 Scikit-learn 라이브러리의 Random Forest Regressor 함수를 이용하여 설계하였다. RNN, LSTM, GRU는 Keras 라이브러리의 Simple RNN, LSTM, GRU 함수를 이용하였다.

태양광 발전량 예측모델의 성능 평가는 예측과정에서 중요하다.³⁸⁾ 본 연구에서는 관측값과 예측값의 관계 등을 확인 가능한 결정계수(coefficient of determination, R^2)(식 1)와 태양광 발전량 예측 평가 지표로 주로 사용되는 평균제곱근오차(root mean square error, RMSE)(식 2), 평균절대오차(mean absolute error, MAE)(식 3)를 이용하였다.

$$R^2 = 1 - \frac{\sum_{i=1}^n (X_i - Y_i)^2}{\sum_{i=1}^n (X_i - X_{avg})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - Y_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i| \quad (3)$$

여기서, n 은 자료의 개수, X 는 발전량의 관측된 값, Y 는 모델을 통한 예측값, X_{avg} 는 관측된 값의 평균이다. 모든 성능 지표는 Scikit-learn 라이브러리에서 사용하였다.

3. 결과 및 고찰

3.1. 랜덤포레스트

RF의 하이퍼파라미터(hyperparameter)인 트리 개수를 10부터 10,000까지의 범위 안에서 조절하였다. 선행 연구에선 실험마다 최적의 트리 개수의 차이는 있지만 트리 개수가 증가할수록 예측 정확도가 증가하는 경향성을 볼 수 있다고 하였다.³⁹⁾ 본 실험에서도 트리 개수를 10에서부터 증가시켰을 때 결과도 좋아지다가 400개부터 test $R^2=0.865$ 로 고정됨을 확인하였다. test RMSE의 경우 트리 개수가 10,000일 때를 제외하고 트리 개수 500에서 가장 낮은 값인 16.013이 나왔다. Test MAE의 경우도 트리 개수 10,000일 때를 제외하고 트리 개수 500일 때 9.656로 가장 낮았다. 트리 개수는 모델을 운용하는 데 있어 계산하는 시간과 비용에 직접적인 영향을 준다. 트리 개수가 증가할수록 계산 시간과 비용이 증가하기 때문에 모델 성능 및 트리 개수 사이의 절충을 해야 한다.⁴⁰⁾ 본 연구에선 트리 개수가 10,000일 때 test RMSE=16.010, test MAE=9.655라는 값을 도출하였지만 트리 개수 500일 때보다 훨씬 뛰어난 성능 차이를 내지 않았다. 따라서 최적 트리 개수를 500개로 선정하였다(Table 1). 완성된 모델은 RF를 사용하여 태양광 발전량 예측 선행 연구보다 1/8-1/2 규모로 더 간결한 설계가 되었다.^{12,13)} 최적화된 RF 모델의 결과는 train $R^2=0.985$, train RMSE=5.195, train MAE=3.037, test $R^2=0.865$, test RMSE=16.013, test MAE=9.656로 나왔다(Table 2).

RF의 입력 인자 중요도 분석(feature importance analysis)을 통하여 인자별 중요도를 분석한 결과 일사량(78.20%), 온도(10.88%), 기압(3.19%), 습도(2.22%), 풍속(2.07%), 풍향(1.18%), 일조시간(1.14%), 운량(1.13%) 순으로 나타났다(Fig. 4). 이를

Table 1. The specific hyperparameter settings of random forest, recurrent neural network, long short-term memory, and gated recurrent unit.

Model	Number of trees	Hidden layers	Nodes	Epochs	Validation split	Activation function
RF	500	-	-	-	-	-
RNN	-	6	90	450	0.1	ReLU
LSTM	-	3	600	200	0.1	ReLU
GRU	-	3	450	500	0.1	ReLU

Table 2. Performance comparison of random forest, recurrent neural network, long short-term memory, and gated recurrent unit.

	R^2		RMSE		MAE	
	Train	Test	Train	Test	Train	Test
RF	0.985	0.865	5.195	16.013	3.037	9.656
RNN	0.991	0.942	4.01	10.53	0.946	6.39
LSTM	0.991	0.944	4.016	10.29	1.122	6.36
GRU	0.992	0.945	3.899	10.189	0.923	5.968

Fig. 4. Variable importance for photovoltaic power generation.

통해 기존 문헌에서 그렇듯 일사량이 태양광 발전량을 예측할 때 가장 중요한 인자임을 알 수 있었다.

RF를 이용한 태양광 발전량 예측 사례는 2개 정도를 찾을 수 있었다.^{12,13)} 한 선행 연구는 영국 카디프(Cardiff) 지역의 2015년 1월부터 2015년 12월까지의 기상 자료를 이용해 태양광 발전량 예측 모델을 구축하였다.¹²⁾ 해당 연구에선 트리 개수를 1,000으로 정할 때 가장 뛰어난 성능을 보였으며, 이는

본 연구보다 2배 많은 트리를 사용했다는 차이가 있었다. 입력 인자로는 일사량, 온도, 습도, 풍속 등을 사용하였다. 최적 모델의 성능은 $R^2=0.923$, $RMSE=2.247$, $MAE=1.083$ 이 나왔다(Table 3). 선행 연구에서 사용한 인자 4개는 본 연구에서의 인자 중요도 분석의 상위 5개(일사량, 온도, 기압, 습도, 풍속)에 해당함을 알 수 있었다. 이를 통해 일사량이 태양광 발전량을 예측할 때 가장 중요한 인자이며 온도, 습도, 풍속 등도 주요 인자임을 알 수 있었다. 다른 연구에선 이탈리아 지역의 2014년 3월부터 2015년 9월까지의 기상 자료를 이용해 태양광 발전량을 예측하는 실험을 하여, 결정트리를 4,000으로 설정해 $RMSE=1.949$, $MAE=3.800$ 이라는 값을 얻었다(Table 3).¹³⁾ 본 연구에서 사용한 인자와 달리 태양광 패널의 온도를 추가했다는 점과 더 많은 트리 개수를 적용하였다는 점의 차이가 있었다.

3.2. 순환신경망

RNN 모델의 하이퍼파라미터 최적화 결과 은닉층 6층에 은닉층별 노드 90, 학습 횟수 450회, 학습 자료 내 검증자료 비

Table 3. Existing photovoltaic power research with random forest, recurrent neural network, long short-term memory, and gated recurrent unit.

Model	Hyperparameter	Outcome	Data period	Location	Literature
RF	Number of trees=4,000	MAE=3.8 RMSE=1.949	2014/03- 2015/09	Italy	Oneto et al., 2018 ¹³⁾
RF	Number of trees=1,000	$R^2=0.923$ RMSE=2.247 MAE=1.083	2015/01/01/- 2015/12/31	UK	Ahmad et al., 2018 ¹²⁾
RNN	N/A	MAE=0.1327	2005	Japan	Yona et al., 2013 ¹⁴⁾
LSTM	Hidden layer=1 Nodes=4 Epoch=50 Sigmoid	RMSE=82.15	a whole year	Egypt	Nasser et al., 2019 ¹⁵⁾
LSTM	Hidden layer=2 Nodes=64, 128 Epoch=100	MAE=0.327 RMSE=0.709	half year	Australia	Wang et al., 2019 ¹⁶⁾
LSTM	Hidden layer=2 Nodes=32	RMSE=2.11 MAE=1.48	2014-2018	China	Zhou et al., 2019 ¹⁷⁾
GRU	Hidden layer=1 Nodes=15 Epoch=100 Sigmoid	RMSE=0.0683	2012/04/01- 2014/07/01	61 countries	Wang et al., 2018 ¹⁸⁾

을 0.1로 설정한 경우 최적화가 되었다(Table 1). 은닉층의 경우 2~9층까지 변화를 주며 최적화한 결과 6층으로 적용했을 시 학습 결과 $R^2=0.991$, $RMSE=4.093$, $MAE=1.302$ 로 7층을 적용한 결과보다 $RMSE$, MAE 가 각 0.016, 0.071 높게 나왔으나 테스트 결과 $R^2=0.943$, $RMSE=10.395$, $MAE=6.596$ 으로 오차가 가장 낮았으며, R^2 는 가장 높았다. 따라서 은닉층은 6층에서 최적화가 되었음을 알 수 있었다. 은닉층별 노드의 경우 40~100에서 10씩 변화를 주며 최적화를 진행하였으며, 90에서 가장 좋은 결과를 얻었다($train R^2=0.991$, $train RMSE=4.081$, $train MAE=0.958$, $test R^2=0.944$, $test RMSE=10.311$, $test MAE=6.366$). 학습 반복 횟수는 300~800에서 최적화를 진행하였으며 450회에서 가장 좋은 성능을 얻었다($train R^2=0.991$, $train RMSE=3.915$, $train MAE=1.004$, $test R^2=0.944$, $test RMSE=10.308$, $test MAE=6.199$). 학습 자료 내 검증자료 비율 최적화 결과 학습 자료의 10%를 설정했을 경우보다 20%, 30%로 했을 경우 점점 모델 성능이 감소하여 10%로 하였다($train R^2=0.991$, $train RMSE=3.915$, $train MAE=1.004$, $test R^2=0.944$, $test RMSE=10.308$, $test MAE=6.199$). 최적화가 완료된 모델 성능은 $train R^2=0.991$, $train RMSE=4.010$, $train MAE=0.946$, $test R^2=0.942$, $test RMSE=10.530$, $test MAE=6.390$ 였다(Table 2).

RNN 모델을 이용한 태양광 발전량을 예측한 기존 논문은 1편을 찾을 수 있었다.¹⁴⁾ 인자로 운량과 습도, 온도를 인자로 설정하여 실험한 결과, $MAE=0.133$ 이 나왔다(Table 3). 이 논문에서는 일사량 자료를 확보하지 못해 먼저 일사량을 예측하였다. 이후 예측한 일사량과 확보된 인자들로 태양광 발전량을 예측했기 때문에 본 연구의 실험 과정과 차이가 있음을 확인할 수 있었다.

3.3. 장단기 메모리

LSTM의 하이퍼파라미터 최적화 결과 은닉층 3층에 층별 노드 600, 학습 횟수 200회, 학습 자료 내 검증자료 비율 0.1로 설정할 때 최적화되었다(Table 1). 은닉층의 경우 2~6층에서 변화를 주며 최적화한 결과 3층을 적용한 경우 2층보다 학습 성능은 낮았으나 테스트 성능이 높았다. 이후 층수를 올릴수록 모델의 성능이 감소하여 3층을 최적 층수로 하였다($train R^2=0.990$, $train RMSE=4.356$, $train MAE=1.888$, $test R^2=0.903$, $test RMSE=13.680$, $test MAE=11.056$). 층별 노드의 경우 400~800에서 최적화를 진행한 결과 600 이하에선 점차 성능이 향상되다 600 이후부터 감소하였다. 따라서 최적 노드는 600으로 하였다($train R^2=0.990$, $train RMSE=4.143$, $train MAE=1.695$, $test R^2=0.936$, $test RMSE=10.992$, $test MAE=7.750$). 학습 반복 횟수의 경우 100~800을 적용한 결과 200에서 가장 좋은 성능을 보였으며($train R^2=0.991$, $train RMSE=3.983$, $train MAE=1.113$, $test R^2=0.946$, $test RMSE=10.160$, $test MAE=6.227$) 그 이상으로 적용할 경우 테스트 성능이 감소하여 과대 적합이 발생

할 가능성이 있었다. 학습 자료 내 검증자료 비율 최적화 결과 학습 자료의 10%를 설정했을 경우 모델 성능이 향상되었으나($train R^2=0.992$, $train RMSE=3.967$, $train MAE=1.058$, $test R^2=0.948$, $test RMSE=10.144$, $test MAE=6.210$) 20% 이상으로 설정할 경우 학습, 테스트 성능이 점차 감소하여 적절한 학습 자료량을 유지하지 못하는 것으로 보였다. 최적 모델의 성능은 $train R^2=0.991$, $train RMSE=4.016$, $train MAE=1.122$, $test R^2=0.944$, $test RMSE=10.290$, $test MAE=6.360$ 이었다(Table 2).

LSTM을 이용하여 태양광 발전을 예측한 논문은 3편을 찾을 수 있었다.¹⁵⁻¹⁷⁾ 이집트 카이로(Cairo) 지역의 1년 치 자료를 이용하여 태양광 발전량을 예측한 논문에서는 은닉층 1개, 노드의 수 4개, 학습 반복 횟수 50으로 하이퍼파라미터를 설정할 때 $RMSE$ 가 82.15로 가장 뛰어난 성능의 모델이 도출되었다(Table 3).¹⁵⁾ 오스트레일리아 앨리스 스프링스(Alice Springs)의 반년 치 자료를 이용하였다. 최적화한 태양광 발전량 모델로 2개의 은닉층, 은닉층별 노드 수는 64, 128개, 학습 반복 횟수는 100으로 제시하였다.¹⁶⁾ 결과 값은 $RMSE=0.709$, $MAE=0.327$ 이다(Table 3). 중국 저지양주 샤오싱시(紹興市)의 2014년부터 2018년까지 총 5년의 기상 자료와 발전량 자료를 이용해 태양광 발전량을 예측하였다.¹⁷⁾ 최적화한 LSTM의 하이퍼파라미터는 은닉층 2개, 노드의 수 32개로 $RMSE=2.110$, $MAE=1.480$ 의 결과를 얻어냈다.

본 논문의 최적 노드 수는 앞에 언급한 세 논문과 차이를 보였다. 하이퍼파라미터는 모델의 성능, 운용 시간과 비용에 영향을 준다. 따라서 LSTM의 성능을 높이기 위해서는 하이퍼파라미터의 최적화 과정이 필요하다.²⁷⁾ 하이퍼파라미터 중 은닉층과 노드의 수로 LSTM 성능에 큰 영향을 미친다. 너무 적은 은닉층 수와 노드 수는 과소 적합(underfitting)을 일으키고 너무 많은 은닉층 수와 노드 수는 과대 적합을 일으킨다.⁴¹⁾ 그러므로 은닉층 수와 노드 수를 적절히 선택할 필요가 있다.

3.4. 게이트 순환 유닛

하이퍼파라미터 조정을 통한 GRU 최적화 결과 은닉층 3층, 층별 노드 450, 학습 횟수 500회, 학습 자료 내 검증자료 비율 0.1을 적용한 경우에서 가장 좋은 성능을 보였다(Table 1). 은닉층의 경우 2~8층까지 변화를 주며 최적화를 시도한 결과 2층에선 좋은 학습 성능을 볼 수 있었으나($train R^2=0.989$, $train RMSE=4.596$, $train MAE=1.810$, $test R^2=0.926$, $test RMSE=11.851$, $test MAE=8.915$) 테스트 결과 3층보다 낮은 성능을 보여 실제 자료에 적용 시 3층에서 더 높은 성능을 보였다($train R^2=0.988$, $train RMSE=4.597$, $train MAE=1.911$, $test R^2=0.929$, $test RMSE=11.638$, $test MAE=8.580$). 이후 은닉층 수를 증가시킬수록 테스트 성능의 감소를 확인하여 최적 은닉층을 3층으로 하였다. 은닉층별 노드를 10~1,000 범위에서 최적화를 시도한 결과 400 미만으로 적용했을 경우 400

으로 적용한 모델보다 낮은 성능을 보였으며(train $R^2=0.988$, train RMSE=4.817, train MAE=2.375, test $R^2=0.938$, test RMSE=10.844, test MAE=7.489), 500 이상으로 적용한 결과 학습, 테스트 성능 모두 감소하였다. 따라서 400, 450을 적용하여 성능을 비교한 결과 450에서 학습, 테스트 모두 높은 성능을 보여 450을 최적값으로 하였다(train $R^2=0.990$, train RMSE=4.348, train MAE=1.809, test $R^2=0.940$, test RMSE=10.691, test MAE=7.323). 학습 반복 횟수는 200~800 범위에서 최적화를 진행하였으며, 500에서 가장 좋은 성능을 얻었다(train $R^2=0.992$, train RMSE=3.823, train MAE=0.856, test $R^2=0.947$, test RMSE=10.040, test MAE=5.873). 학습 자료 내 검증자료 비율은 10%에서 가장 좋은 성능을 얻을 수 있었으며, 비율을 증가시킬 경우 모델 성능이 감소하였다(train $R^2=0.992$, train RMSE=3.824, train MAE=0.821, test $R^2=0.948$, test RMSE=10.033, test MAE=5.671). 최적화한 GRU 모델의 최종 결과는 train $R^2=0.992$, train RMSE=3.899, train MAE=0.923, test $R^2=0.945$, test RMSE=10.189, test MAE=5.968이 나왔다(**Table 2**).

태양광 발전량을 예측하는 GRU 모델은 61개국 자료를 수집한 global energy forecasting competition 2014 자료를 이용한 한 편을 찾을 수 있었다.¹⁸⁾ 해당 연구에선 하이퍼파라미터는 은닉층 1개, 노드의 수 15, 학습 횟수 100, 활성화 함수를 sigmoid로 적용하여 결과 값은 RMSE=0.068이 나왔다(**Table 3**). 선행 연구의 경우 1개의 은닉층에 노드 15로 최적화한 것과 달리 본 연구에선 은닉층을 3층으로 하여 은닉층별 노드를 450으로 하였다는 점에서 모델이 더욱 복잡하게 설계되었음을 알 수 있었다. 따라서 선행 연구¹⁸⁾에서의 은닉층 수와 노드의 수는 본 연구의 GRU 모델 실험에서의 최적의 값이 될 수 없음을 알 수 있었으며 사용하는 자료의 구성, 기간, 지역 등에 따라 달라지는 하이퍼파라미터를 시행 착오법 등을 통해 최적점을 찾아야 함을 알 수 있었다.²⁷⁾

3.5. 모델 비교

모델별 최적화를 완료한 RF, RNN, LSTM, GRU의 예측 성능을 비교한 결과, 학습 과정에서 RF가 가장 낮은 예측률(0.985)을 보였으며, 나머지 모델에서는 GRU (0.992), LSTM (0.991), RNN (0.991) 순서로 나타났다(**Fig. 5**). 테스트 과정에서도 역시 RF (0.865)가 4개 모델 중 가장 낮았고, GRU (0.945), LSTM (0.944), RNN (0.942) 순서로 예측률을 보였다. 학습 및 테스트 과정 모두 순환신경망 계열의 모델인 RNN, LSTM, GRU가 RF보다 뛰어난 성능을 보였다. 순환신경망 계열의 모델에서는 모델 간의 성능 차이가 두드러지게 나타나지 않았지만 RNN < LSTM < GRU 순으로 예측 결과가 점점 좋아지는 것을 확인할 수 있었다. 최종적으로 해당 지역 태양광 발전량 예측을 위해 상대적으로 우수한 모델은 GRU임을 확인할 수 있었다.

Fig. 5. Prediction results of photovoltaic power on training and test dataset using random forest, recurrent neural network, long short-term memory, and gated recurrent unit.

4. 결론

본 연구에서는 태양광 발전량 예측에 효과적인 모델을 찾기 위해 RF, RNN, LSTM, GRU를 사용하였다. 입력 인자 중 일사량이 가장 중요하였다(78.2%). 하이퍼파라미터 최적화 과정을 진행한 결과 시계열 자료에 효과적인 순환신경망 계열인 RNN, LSTM, GRU 모두 RF보다 성능이 뛰어났다. 순환신경망 계열에서는 GRU ($R^2=0.945$, RMSE=10.189, MAE=5.968)의 예측 정확도가 가장 뛰어났다. 따라서, 시계열 자료를 이용한 부산지역의 태양광 발전량 예측에 있어 순환신경망 기법이 상대적으로 우수하였으며 그중에서도 GRU가 가장 적합함을 알 수 있었다. 따라서 추후 다양한 지역의 태양광 발전 예측용 GRU 모델 개발을 통해 더욱 정확한 예측 가능한 모델을 개발할 수 있으리라 기대된다.

Acknowledgement

본 논문은 4단계 BK21 사업(강원대학교, 다학제 융합 에너지자원 신산업 핵심인력 양성사업단)으로 지원된 연구입니다.

References

1. H. Wang, Z. Lei, X. Zhang, B. Zhou, J. Peng, A review of deep learning for renewable energy forecasting, *Energy Conv. Manag.*, 198, 111799(2019).
2. Ministry of Trade, Industry and Energy, Renew Energy 3020, Ministry of Trade, Industry and Energy, Sejong, Korea(2017).
3. M. A. F. B. Lima, P. C. M. Carvalho, L. M. Fernandez-Ramírez, A. P. S. Braga, Improving solar forecasting using deep learning and portfolio theory integration, *Energy*, 195, 117016(2020).
4. Korea Energy Agency, Renewable Energy Center, New renewable energy supply statistics, Korea Energy Agency, Renewable Energy Center, Ulsan, Korea(2019).
5. P. Li, K. Zhou, X. Lu, S. Yang, A hybrid deep learning model for short-term pv power forecasting, *Appl. Energy*, 259, 114216(2020).
6. C. Voyant, G. Notton, S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: a review, *Renew. Energy*, 105, 569-582(2017).
7. M. G. D. Georgi, P. M. Congedo, M. Malvoni, Photovoltaic power forecasting using statistical methods: impact of weather data, *IET Sci. Meas. Technol.*, 8(3), 90-97(2014).
8. S. Sobri, S. Koohi-Kamali, N. A. Rahim, Solar photovoltaic generation forecasting methods: a review, *Energy Conv. Manag.*, 156, 459-497(2018).
9. G. M. Yagil, D. Yang, D. Srinivasan, Automatic hourly solar forecasting using machine learning models, *Renewable Sustainable Energy Rev.*, 105, 487-498(2019).
10. S. K. Aggarwal, L. M. Saini, Solar energy prediction using linear and non-linear regularization models: a study on AMS (American Meteorological Society) 2013-14 solar energy prediction contest, *Energy*, 78, 247-256(2014).
11. J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Prison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, *Solar Energy*, 136, 78-111(2016).
12. M. W. Ahmad, M. Mourshed, Y. Rezgui, Tree-based ensemble methods for predicting pv power generation and their comparison with support vector regression, *Energy*, 164, 465-474(2018).
13. L. Oneto, F. Laureri, M. Robba, F. Delfino, D. Anuita, Data-driven photovoltaic power production nowcasting and forecasting for polygeneration microgrids, *IEEE Syst. J.*, 12(3), 2842-2854(2018).
14. A. Yona, T. Senjyu, T. Funabashi, C. H. Kim, Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead pv power output correction, *IEEE Trans. Sustain. Energy*, 4(2), 527-533(2013).
15. M. A. Nasser, K. Mohamoud, Accurate photovoltaic power forecasting models using deep LSTM-RNN, *Neural Comput. Appl.*, 31, 2727-2740(2019).
16. K. Wang, X. Qi, H. Liu, Photovoltaic power forecasting based LSTM-Convolution Network, *Energy*, 189, 116225(2019).
17. H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan, Y. Du, Short-term photovoltaic power forecasting based on long short term memory neural network and attention mechanism, *IEEE Access*, 7, 78063-78074(2019).
18. Y. Wang, W. Liao, Y. Chang, Gated recurrent unit network-based short-term photovoltaic forecasting, *Energies*, 11(8), 2163(2018).
19. Y. Liu, H. Qin, Z. Zhang, S. Pei, C. Wang, X. Yu, Z. Jiang, J. Zhou, Ensemble spatiotemporal forecasting of solar irradiation using variational Bayesian convolutional gate recurrent unit network, *Appl. Energy*, 253, 113596(2019).
20. M. Aslam, J. M. Lee, H. S. Kim, S. J. Lee, S. W. Hong, Deep learning models for long-term solar radiation forecasting considering microgrid installation: a comparative study, *Energies*, 13(1), 147(2020).
21. L. Breiman, Random forests, *Mach. Learn.*, 45(1), 5-32(2001).
22. C. L. Giles, S. Lawrence, A. C. Tsoi, Rule inference for financial prediction using recurrent neural networks, in *Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering (CIFER)*, IEEE, New York, USA, pp. 253-259(1997).
23. Z. Pang, F. Niu, Z. O'Neill, Solar radiation prediction using recurrent neural network and artificial neural network: a case study with comparisons, *Renew. Energy*, 156, 279-289(2020).
24. R. E. H. Mohamad, M. Skafi, A. M. Haidar, Predicting global solar radiation using recurrent neural networks and climatological parameters, *Int. J. Comput. Inf. Eng.*, 8(2), 331-334(2014).
25. S. Hochreiter, J. Schmidhber, Long short-term memory, *Neural Comput.*, 9(8), 1735-1780(1997).
26. A. Graves, Supervised sequence labelling with recurrent neural networks, Springer-Verlag Berlin Heidelberg, Berlin, Germany, pp. 1-146(2012).
27. M. Husein, I. Y. Chung, Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: a deep learning approach, *Energies*, 12, 1856(2019).
28. K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1724-1734(2014).
29. J. J. Roberts, A. A. M. Zavallos, A. M. Cassula, Assessment of photovoltaic performance models for system simulation, *Renewable Sustainable Energy Rev.*, 72, 1104-1123(2017).
30. M. Gao, J. Li, F. Hong, D. Long, Short-term forecasting of

- power production in a large-scale photovoltaic plant based on lstm, *Appl. Sci.*, 9 (15), 3192(2019).
31. V. J. Fesharaki, M. Dehghani, J. J. Fesharaki, H. Tavasoli, The effect of temperature on photovoltaic cell efficiency, in *Proceedings of the 1st International Conference on Emerging Trends in Energy Conservation-ETEC*, Renew. Energy, Tehran, Iran, pp.20-21(2011).
 32. T. Alskaf, S. Dev, L. Visser, M. Hossari, W. V. Sark, A systematic analysis of meteorological variables for pv output power estimation, *Renew. Energy*, 153, 12-22(2020).
 33. S. Mekhilef, R. Saidur, M. Kamalisarvestani, Effect of dust, humidity and air velocity on efficiency of photovoltaic cells, 16(5), 2920-2925(2012).
 34. C. Chen, S. Duan, T. Cai, B. Liu, Online 24-h solar power forecasting based on weather type classification using artificial neural network, *Sol. Energy*, 85(11), 2856-2870(2011).
 35. S. Bhardwaj, V. Sharma, S. Srivastava, O. S. Sastry, B. Bandyopadhyay, S. S. Chandel, J. R. P. Gupta, Estimation of solar radiation using a combination of Hidden Markov Model and generalized Fuzzy model, *Sol. Energy*, 93, 43-54(2013).
 36. Y. S. Lim, J. H. Tang, Experimental study on flicker emissions by photovoltaic systems on highly cloudy region: a case study in Malaysia, *Renew. Energy*, 64, 61-70(2014).
 37. S. Shivashankar, S. Mekhilef, H. Mokhlis, M. Karimi, Mitigating methods of power fluctuation of photovoltaic (PV) sources- a review, *Renewable Sustainable Energy Rev.*, 59, 1170-1184(2016).
 38. U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. V. Deventer, B. Horan, A. Stojcevski, Forecasting of photovoltaic power generation and model optimization: a review, *Renewable Sustainable Energy Rev.*, 81, 912-928(2018).
 39. I. Orlandi, L. Oneto, D. Auguita, Random forests model selection, in *Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN*, Bruges, Belgium, pp. 441-446 (2016).
 40. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, *Mach. Learn.*, 36(1), 3-42(2006).
 41. G. Panchal, A. Ganatra, Y. P. Kosta, D. Panchal, Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers, *Int. J. Comput. Theory Eng.*, 3(2), 332-337(2011).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors and Contribution Statement

Ramek Kim

Department of Environmental Engineering, Kangwon National University, Bachelor, ORCID® 0000-0001-7525-6527: Conceptualization, Data curation, Data analysis, Visualization, Writing - original draft.

Kyungmin Kim

Department of Integrated Energy and Infra System, Kangwon National University, Master, ORCID® 0000-0002-7348-9367: Methodology, Data analysis, Resources, Validation, Writing - review and editing.

Johng-Hwa Ahn

Department of Environmental Engineering, Kangwon National University, Department of Integrated Energy and Infra System, Kangwon National University, Professor, ORCID® 0000-0001-5745-5338: Supervision, Project administration, Funding acquisition, Resources, Writing - review and editing.