ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 49, No.4: 569-579, December 2021
https://dx.doi.org/10.7469/JKSQM.2021.49.4.569

Attention 기반 Encoder-Decoder 모델을 활용한 작물의 생산량 예측

강수람* · 조경철** · 나명환***†

* 전남대학교 수학통계학과
** 전라남도농업기술원
*** 전남대학교 수학/통계학과

Forecasting Crop Yield Using Encoder-Decoder Model with Attention

Kang, Sooram* · Cho, Kyungchul** · Na, MyungHwan***†

* Dept. Mathematics & Statistics, Chonnam National University
** Jeonnam Agricultural Research & Extension Services
*** Dept. Statistics, Chonnam National University

ABSTRACT

Purpose: The purpose of this study is the time series analysis for predicting the yield of crops applicable to each farm using environmental variables measured by smart farms cultivating tomato. In addition, it is intended to confirm the influence of environmental variables using a deep learning model that can be explained to some extent.

Methods: A time series analysis was performed to predict production using environmental variables measured at 75 smart farms cultivating tomato in two periods. An LSTM-based encoder-decoder model was used for cases of several farms with similar length. In particular, Dual Attention Mechanism was applied to use environmental variables as exogenous variables and to confirm their influence.

Results: As a result of the analysis, Dual Attention LSTM with a window size of 12 weeks showed the best predictive power. It was verified that the environmental variables has a similar effect on prediction through wieghtss extracted from the prediction model, and it was also verified that the previous time point has a greater effect than the time point close to the prediction point.

Conclusion: It is expected that it will be possible to attempt various crops as a model that can be explained by supplementing the shortcomings of general deep learning model.

Key Words: Crop Yield, Tomato, Attention Mechanism, Dual Attention, LSTM

● Received 15 November 2021, 1st revised 1 December 2021, accepted 7 December 2021
† Corresponding Author(nmh@chonnam.ac.kr)

* 본 논문은 강수람의 2021년도 석사 학위논문에서 발췌 정리하였음.
* 본 논문은 농촌진흥청 공동연구사업의 지원을 받아 연구되었음(과제번호 : PJ01455903).

1. 서 론

1.1 연구배경

영세농 위주의 우리나라 농업은 특히 농촌인구 감소와 노동인구의 고령화로 인해 농업 경쟁력이 지속적으로 저하되고 있다. 이를 극복하기 위하여 최근 수년간 정부의 적극적 추진 정책 중 하나인 스마트팜 사업이 우리나라 농업의 경쟁력 제고의 방법으로 주목받고 있다. 스마트팜은 IoT 기술과 클라우딩 시스템 등의 기술이 접목되어 정밀한 데이터가 생성되고 관리되며 작물의 생육 환경 관리를 자동화·원격화·지능화할 수 있는 시설을 의미한다. 스마트팜 농가에서 수집되는 데이터는 개별 농가별로 그리고 실시간 또는 정기적으로 수집되기 때문에 시간에 따른 작물의 생장과 생산량의 변화를 분석하기 매우 유리하다. 특히 작물의 생산량 예측은 관련 기관뿐만 아니라 농가에서의 필요성 또한 매우 높다. 농업의 산업적·경제적 측면, 식량 자급 문제 등과 더불어 개별 농가의 경영적 판단의 중요한 수단이기 때문이다. 본 연구는 스마트팜 농가에서 수집되는 데이터를 이용하여 특정 환경에서의 작물의 생산량을 예측하는 시계열 분석을 수행하였다. 뿐만 아니라 생산량 예측에 있어 중요한 환경요인과 시점에 대한 설명을 통해 설명 가능한 모델로서 개별 농가에서 이용할 수 있고 작기 중 특정 시점 이후의 생산량을 예측하여 농가의 경영판단에 도움이 될 수 있을 것으로 기대된다.

1.2 선행연구

시계열 분석 기법을 적용한 작물의 생산량 예측은 많은 연구가 진행되었다. 특히 생산량만을 이용한 예측이 아닌 생산량에 영향을 미치는 환경정보를 활용하는 다변량 시계열 기법을 통해 인과적 관계를 추정하면서 예측의 정확성을 높이려는 방향으로 발전하였다. 정대희(2018)는 환경정보를 활용하여 국내 쌀 토지 생산성을 예측하였다. 1946년에서 2017년까지의 데이터를 이용하였고 예측대상은 전국의 연간 면적당 쌀 생산량이었다. 온도, 강수량, CO2 농도의 환경변수를 이용하였는데 온도와 강수량은 벼 생장 시기인 4월에서 10월까지의 전국 평균을 사용하였다. ARIMA, ARIMA-X, ARDL, GARCH-M 등의 모형을 비교하였으며 환경변수를 외생변수로 사용한 ARIMA-X가 가장 좋은 예측력을 보였다. 오승원(2017)은 공간정보와 환경정보를 활용하여 국내 양파 생산량을 예측하였다. 예측대상은 전국 연간 면적당 양파 생산량이었고 주산지의 공간정보를 이용한 STARMA, 지역별 생산량의 상관성을 이용한 VAR, 생산량의 상관성과 더불어 환경정보를 외생변수로 추가한 VAR-X 등의 모형을 비교하였다. VAR과 VAR-X는 지역별로 2개~3개의 시군을 하나의 모형으로 적합하였으며 생산량의 상관성과 1월 일조량을 사용한 VAR-X의 예측력이 가장 우수하였다.

최근에는 딥러닝 모델을 활용하여 작물의 생산량을 예측하는 연구가 진행되고 있다. 특히 딥러닝 시계열 분석 모델 RNN의 일종인 LSTM을 활용하였으며 LSTM의 경우 긴 시계열 데이터에도 적용하기 적합하고 통계적 시계열 기법을 이용한 연구에 비해 많은 외생변수를 사용하여 예측력을 높였다. Jiang(2018)은 LSTM을 적용하여 미국 Corn Belt 지역의 연간 옥수수 생산량을 예측하였다. 강수량, 풍속, 온도, GDD(Growing Degree Days, 유효적산온도), 토양습도, 토양품질 정보 등의 변수를 일 단위의 3D 텐서로 만들어 이를 각 해의 생산량과 연결하여 예측에 이용하였다. Alhnaity(2020)는 연간 생산량 예측 위주의 연구와 달리 시설 재배 농가의 데이터를 이용하여 일별 생산량 및 생장을 예측하였다. 예측대상은 두 농가의 일별 토마토 생산량과 일별 벤자민 고무나무 줄기 직경으로 LSTM, SVM, 랜덤포레스트를 적용하여 예측력을 비교하였다. CO2 농도, 내부온도, 외부온도, 일사량의 환경변수를 외생변

수로 사용하였으며 두 작물 모두 이러한 환경변수를 사용한 LSTM의 예측력이 가장 우수하였다.

1.3 연구목적

작물의 생산량을 예측하는 기존의 연구들은 주로 전국단위의 연간 생산량을 대상으로 하였다. 따라서 해당 작물의 전국적인 수급 경향을 파악하여 관련 기관에서 정책을 결정하거나 농가에서 작물을 선정하고 경영판단을 내리기에 유용하였다. 하지만 농가에서 이미 작물을 선정하고 재배하고 있는 상황에서 해당 농가의 생산량을 예측하는 것은 불가능하였다. 또한, 기상청 등에서 제공하는 지역별 혹은 전국단위의 환경정보를 외생변수로 사용하여 예측대상과의 밀접성이 떨어지며 특히 외부환경의 영향을 적게 받으며 특정 환경을 관리하는 시설 재배 작물과 농가에서의 유용성은 떨어진다. 기존 연구에서 농가별 실측 데이터를 분석한 Alhnaity(2020)의 연구에서는 두 작물에 대해 각각 하나의 농가에서 수집한 데이터만을 사용하여 결과의 일반화가 어렵다. 이와 더불어 예측에 사용한 기법의 한계 또한 존재하였다. 통계적 기법을 적용한 연구에서는 예측에 필요한 외생변수를 충분히 사용한 경우가 적었고 딥러닝 모델은 기존의 기법에 비해 좋은 예측결과를 보였지만 예측결과와 외생변수에 대한 설명이 불가능하였다.

본 연구에 이용된 연구자료는 각 스마트팜 농가에서 수집된 생산량과 환경 데이터로써 예측대상과 변수 간의 관련성이 높으며 따라서 기존연구에 비하여 설명력을 높이고 작기 동안 개별 농가의 생산량을 예측할 수 있도록 하는 것을 목적으로 하였다. 또한, 일반적인 딥러닝 모델에서는 설명이 어려운 것에 반하여 생산량을 예측하면서 예측에 있어 중요한 환경변수와 시점에 대한 설명이 비교적 가능하도록 하고자 한다. 이러한 목적으로 Dual Attention 구조를 적용한 LSTM 모델을 이용하여 작물의 생산량을 예측한다.

2. 이론적 배경

2.1 LSTM(Long-Short Term Memory)

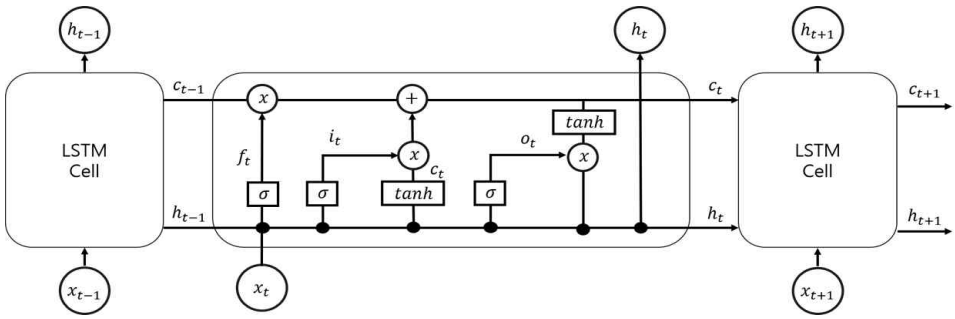

Figure 1. Structure of LSTM

LSTM은 RNN(Recurrent Neural Network)에서 발전한 모델로 시계열 예측과 자연어 처리에서 주로 이용된다. 전통적인 뉴럴 네트워크는 이전 시점의 출력과 현재 시점의 입력이 독립적으로 작용하지만 RNN은 이전 시점의 출력이 현재 시점에 입력되는 순환적 구조가 특징이다. 이러한 구조로 시간 혹은 순서의 영향을 분석하기 용이하기 때문에 시계열 데이터와 자연어 데이터 분석에 많이 이용되고 있다. 또한 RNN 계열의 시계열 모델은 그 자체만으로도 분석에 이용되지만 인코더 디코더와 같은 여러 가지 구조의 내부를 구성하기도 한다.

하지만 RNN은 과거 입력 시점과 현재 입력 시점이 멀어질수록 과거 학습의 영향이 줄어드는 기울기 소실(vanishing gradient) 문제로 학습 능력이 저하되었다. LSTM은 셀(cell) 내부의 게이트(gate)를 통해 장기의존성(long-term dependency)을 학습하여 이러한 문제를 극복하였다. 각각의 게이트는 시점 t의 입력벡터인 X_t와 이전 시점의 은닉상태(hidden state)인 h_{t-1}을 입력받는다. 입력받은 내용을 학습 가능한 가중치와의 연산과정과 활성함수를 거쳐 각 게이트의 값을 갖게 된다. 입력 게이트는 현재 정보를 업데이트하는 게이트로 시점 t의 입력벡터인 X_t와 이전 시점의 은닉상태인 h_{t-1}을 입력받아 가중치 연산과 시그모이드(sigmoid) 연산을 통해 0과 1 사이의 값을 갖는 i_t가 된다. 망각 게이트는 이전 셀에서 넘어온 정보를 얼마나 기억할지 결정한다. f_t가 0이면 모든 정보를 기억하지 않고 1이면 모든 정보를 기억한다. 출력 게이트는 어떤 정보를 출력으로 내보낼지 정하는 게이트이며 $\tilde{c}_t$는 하이퍼볼릭 탄젠트 활성 레이어를 거쳐 연산된 셀 상태(cell state)의 후보로 입력 게이트, 망각 게이트와의 연산을 거쳐 시점 t의 셀 상태 c_t가 생성된다. 마지막으로 셀 상태 c_t에 하이퍼볼릭 탄젠트 활성 레이어를 거치고 출력 게이트에서 계산된 o_t와의 연산을 거쳐 시점 t의 은닉상태인 h_t가 생성되어 다음 시점에 전달된다.

2.2 Attention Mechanism

일반적으로 딥러닝을 이용한 시계열 분석의 경우 일차원 구조의 데이터를 이용하며 모델이 데이터의 패턴을 파악할 수 있을 만큼 충분히 긴 시퀀스를 필요로 한다. 데이터의 특성과 연구자의 문제 정의에 따라 그 길이는 달라진다. 하지만 자연어 처리의 경우 긴 문장보다는 많은 개별 문장을 학습하는 것이 좋은 결과를 보인다. 본 연구에서는 시퀀스 길이가 평균 30주 정도로 길지 않지만 많은 농가의 케이스로 이루어져 있기 때문에 자연어 처리의 경우와 유사하게 인코더 디코더 구조를 갖는 시계열 모델이 좋은 성능을 보였다. 기계번역, 챗봇, 내용 요약 등의 자연어 처리에 주로 이용되는 모델인 시퀀스 투 시퀀스 모델은 인코더와 디코더의 구조로 이루어져 있으며 인코더는 문장을 입력받아 내부의 LSTM 셀에서 단어 배열 순서대로 순차적 연산을 수행하여 정보를 압축한다. 반대로 디코더는 인코더에서 압축된 정보를 입력받아 문장을 출력한다. 이때 인코더에서 압축된 정보는 인코더의 마지막 셀의 은닉상태이다. 이를 context 벡터라고 부르는데 기울기 소실 문제와 고정된 크기의 벡터에 압축되어 생기는 정보손실의 문제가 존재하였고 attention 구조로 발전하였다.

attention 구조는 예측에 있어 중요한 내용에 집중하여 학습하자는 아이디어로 디코더에서 예측값을 출력하는 매 시점마다 인코더의 전체 값을 일정 비율로 매번 참고하는 구조다. attention 스코어는 인코더의 모든 은닉상태와 디코더의 예측 시점의 은닉상태 간의 유사도로 스코어를 구하는 방법에 따라 attention 구조의 종류를 구별하기도 한다. 본 연구에서는 attention 구조에 많이 사용되는 바나다우 attention을 적용하였다. 바나다우 attention은 인코더와 디코더 각각의 은닉상태에 학습 가능한 가중치를 곱하고 하이퍼볼릭 탄젠트 함수를 적용한 것이 특징이다. attention 가중치는 이렇게 구한 스코어를 소프트맥스 함수에 적용하여 모든 값을 합하면 1이 되도록 만들어진 일종의 가중치로 인코더의 각 은닉상태와 각각 곱하고 더하여 가중합인 attention value가 생성된다. 그리고 디코더의 예측 시점의 은닉상태와 결합하여 학습에 이용되는데 이것이 시퀀스 투 시퀀스 모델에서의 context 벡터와 같다. 다만 attention 구조의 context 벡터는 디코더의 매 시점마다 생성된다는 점에서 차이가 있다.

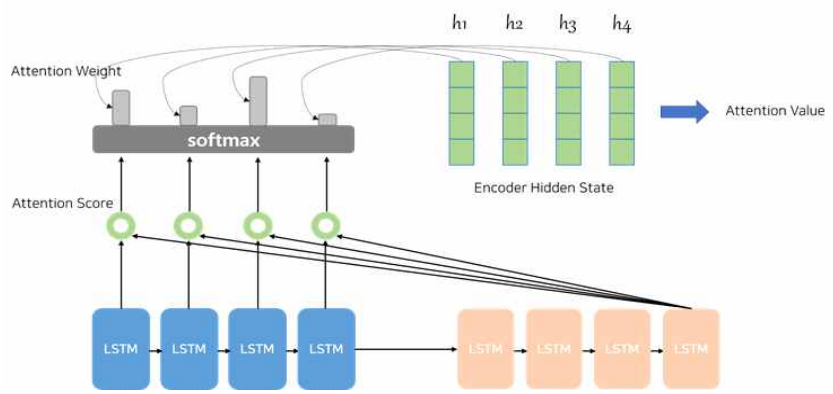

Figure 2. Structure of Attention Mechanism

2.3 Dual Attention LSTM

Dual Attention LSTM은 attention 구조를 두 번 적용한 것으로 다변량 시계열 분석에서 적용가능하다. 자연어 처리의 경우 모델의 입력이 문장이라는 단일한 형태를 갖지만 외생변수를 이용하는 시계열 분석의 경우, 다차원의 구조를 갖는다. Dual Attention LSTM은 input attention을 통해 다차원의 외생변수를 하나의 차원으로 결합하여 새로운 입력을 생성하여 인코더 디코더 구조에 적용할 수 있도록 한다. 다시 말해 input attention은 외생변수 중 의미있는 변수에 집중하기 위하여 적용하는 것으로 각 T의 길이를 갖는 n개의 외생변수들이 가중합을 통해 $\tilde{x}_t$라는 새로운 변수가 생성된다. 그림 3은 input attention의 구조를 보여준다. 이후 새로 생성된 $\tilde{x}_t$를 입력으로 일반적인 attention 구조를 적용한다.

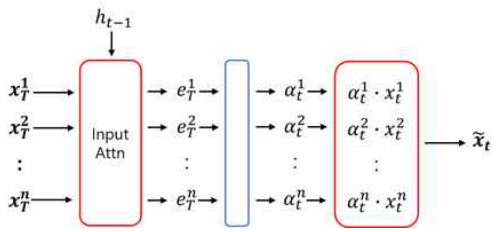

Figure 3. Structure of Input Attention

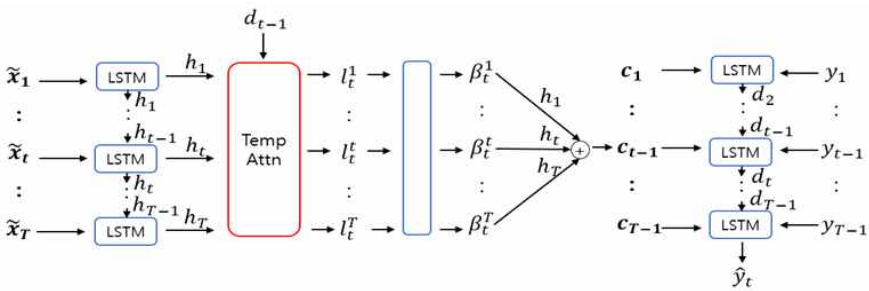

Figure 4. Structure of Temporal Attention and Encoder-Decoder

3. 연구자료

3.1 수집데이터

본 연구의 분석 작물인 토마토는 7월 혹은 8월에 정식하여 10월부터 평균 30주, 길게는 40주가량 긴 기간 동안 수일 간격으로 지속적으로 생산되는 작물이다. 따라서 작기 내 생산량 예측의 필요성이 크다고 할 수 있다. 수집된 데이터는 2017년에서 2018년, 2018년에서 2019년 두 작기 동안 경남, 전남, 전북 95개 스마트팜 토마토 재배 농가에서 수집되었다. 각 농가에서 수집된 데이터는 생육정보, 생산량, 환경정보로 이루어져 있으며 생육정보는 생육조사원에 의해 주 1회 측정 및 기록되었고 생산량은 각 농가의 수확면적과 실제 수확하는 일자 및 수확량이 기록이 되었는데 1일에서 3일 길게는 5일 간격으로 다소 불규칙하게 수확이 이루어졌다. 환경정보는 1시간 단위로 수집되었으며 온도, 습도, 절대습도, 이슬점, CO2 농도, 습도, 감우, 풍향, 풍속, 지온, 일사량, 광량 등의 측정정보와 환기온도, 난방온도, 공급온도, 제어온도, 제어습도 등 스마트팜 시설의 설정정보로 이루어져있다. 이 중 측정정보는 스마트팜 시설 내 설치된 센서장비에 의해 측정되었고 측정된 정보는 클라우드에 자동 기록 및 저장되었다. 실제 분석은 온전한 데이터를 갖는 농가이면서 생산 주의 길이가 20주 이상인 75개 농가의 생산량과 환경정보의 측정정보를 대상으로 하였다.

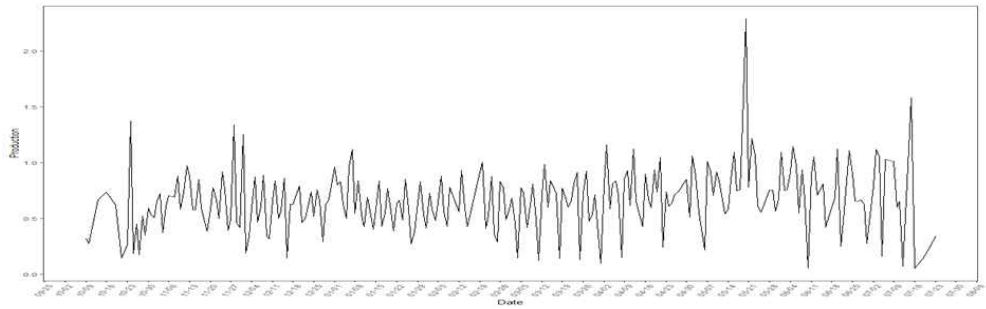

Figure 5. Tomato Yield in Farm A

3.2 데이터 전처리

생산량과 환경정보의 시간단위가 수일 간격과 1시간 간격으로 다르기 때문에 실제 분석에서는 일주일 단위로 가공하여 사용하였다. 기준은 해당연도의 주 번호 즉, 일주일 단위로 하였으며 생산량은 일주일 합계를 내었고 환경은 일주일 평균을 내었다. 예측대상은 총 생산량이 아닌 각 농가의 재배면적으로 생산량을 나눈 주별 면적당 생산량으로 하였다. 외생변수는 모든 농가에서 공통적으로 측정되었으면서 결측값이 적은 항목인 내부온도, 외부온도, 내부습도, CO2 농도를 선정하였고 이들의 최소, 최대, 평균으로 총 12개 변수를 사용하였다. 일부 농가에서 환경정보의 결측값이 존재하였고 이를 해당 농가와 같은 도(道)에 속한 농가들의 평균으로 대체하였다. 또한, 변수들 간의 단위와 값의 크기가 다르기 때문에 원 데이터를 분석할 시 나타나는 왜곡을 없애기 위해 각 변수들의 평균과 분산을 각각 0과 1로 표준화하였다. 전체 75개 농가 중 62개 농가로 모델을 훈련하였고 13개 농가로 모델을 예측성능을 테스트하였다.

Table 1. Environment Variables

Environment Variables	Unit	Names of Variables in Experiment
Minimum inside temperature	°C	in.temp_min
Mean inside temperature	°C	in.temp_mean
Maximum inside temperature	°C	in.temp_max
Minimum outside temperature	°C	out.temp_min
Mean outside temperature	°C	out.temp_mean
Maximum outside temperature	°C	out.temp_max
Minimum humidity	%	hum_min
Mean humidity	%	hum_mean
Maximum humidity	%	hum_max
Minimum CO2 level	ppm	CO2_min
Mean CO2 level	ppm	CO2_mean
Maximum CO2 level	ppm	CO2_max

4. 생산량 예측

4.1 예측모델

본 연구의 예측모델은 2주에서 14주의 데이터를 학습하여 다음 주의 생산량을 예측하는 단기 예측모델이다. 예측모델에 이용되는 LSTM은 전체 길이의 데이터를 한 번에 학습하는 것이 아니라 일정 길이의 데이터를 학습하여 다음 시점의 생산량을 예측하는 구조로 학습할 데이터의 길이인 윈도우 사이즈를 설정하여야 한다. 본 연구에서는 여러 가지 윈도우 사이즈로 각 농가의 데이터를 분할하여 모델의 입력으로 사용하였다. 분석에서는 3주, 5주, 8주, 12주, 15주로 달리 학습하여 예측력이 가장 좋은 윈도우 사이즈를 선택하였다. 각 농가의 데이터를 구분하여 모델의 입력으로 사용하기 때문에 예측결과 또한 각 농가별로 구분할 수 있다. 모델의 예측력은 회귀문제에서 많이 이용되는 MSE(Mean Squared Error)를 기준으로 하였다. Dual Attention LSTM은 생산량 예측과 함께 매 예측 시점에서의 외생변수와 시점의 중요도를 확인할 수 있다. 모델 내의 두 attention 층인 input attention과 temporal attention에서 추출되는 가중치가 각각의 중요도를 나타낸다. input attention 가중치는 예측 시점에서 12개 변수들의 예측 시점에서의 중요도를 나타낸다. temporal attention은 $\tilde{x}_t$에서 시점 즉, 전체 길이 T에서 각 시점(time step)의 예측 시점에서의 중요도를 계산하여 temporal attention 가중치로 나타낸다. 마지막으로 모델의 하이퍼 파라미터인 은닉층, 은닉노드, 에포크, 배치크기, 옵티마이저, 학습률 등은 그리드 탐색을 통하여 최적화하였다.

4.2 예측결과

기본 모델을 통해 은닉층의 개수와 에포크, 배치크기, 옵티마이저, 학습률에 따라 예측력의 큰 차이가 없었기 때문에 이들을 고정하고 은닉노드와 윈도우 사이즈를 탐색하였다. 테스트 데이터에 대한 예측력이 가장 좋은 모델은 MSE가 0.678로 나타났고 이때 하이퍼 파라미터는 은닉층의 개수가 인코더와 디코더 각각 1개, 에포크는 100번, 배치크기는 128, 옵티마이저는 Adam, 학습률은 0.01이었으며 은닉노드는 256개이고 윈도우 사이즈가 12인 모델이었다. 대부분의 경우 은닉노드가 32개 혹은 64개인 경우보다 128개 혹은 256개인 경우가 예측력이 더 좋은 것으로 나타났다. 윈도우 사이즈는 15주를 제외하고는 커질수록 결과가 좋았다. 앞에서 설명하였던 것처럼 attention 가중치는 디코더의 예측 시점의 각 셀마다 생성되므로 매 예측 시점 마다 다른 가중치가 생성된다. 따라서 테스트 농가의 수인 13과 각 농가의 예측 시점의 개수를 곱한 수만큼의 다른 가중치가 생성된다. 그러므로 전체 농가, 전체 예측 시점에서의 가중치의 평균과 분산을 계산하였다. input attention에서는 각 변수들 각각의 가중치의 평균과 분산을 계산하였고 temporal attention에서는 각 시점(time step)의 가중치의 평균과 분산을 계산하였다. 계산 결과 분산의 가장 큰 값이 0.003 정도로 매우 작게 나타남에 따라 농가와 예측 시점에 따라서 가중치의 차이가 거의 없는 것을 확인하였다. 이에 따라 농가나 예측 시점에 상관없이 외생변수와 시점의 영향력은 같다고 할 수 있으며 모델의 일반화가 가능함을 확인하였다.

먼저 input attention 가중치는 길이 T의 $\tilde{x}_t$를 구성하는 가중치이며 12개 외생변수가 예측에 주는 영향의 정도를 의미한다. 실제 결과에서 변수들의 중요도의 큰 차이는 나타나지 않았다. 따라서 생산량을 예측하는데 있어서 전체 외생변수가 비슷한 정도로 중요하다고 할 수 있다. temporal attention 가중치는 각 시점이 예측에 주는 영향의 정도를 의미한다. 실제 작물 재배에 있어서 급격한 환경의 변화가 일어나지 않는다면 근시일 전의 환경보다 이전의 환경이 더 큰 영향을 미치는 것으로 추정한다. 본 연구에서도 이와 같은 결과가 나타났다. 특히 9주전의 환경이 가장 0.128로 가장 영향을 많이 주는 것으로 나타났으며 그 뒤로 8주전(0.110), 4주전(0.93), 3주전(0.92)의 순서로 영향력이 높게 나타났다.

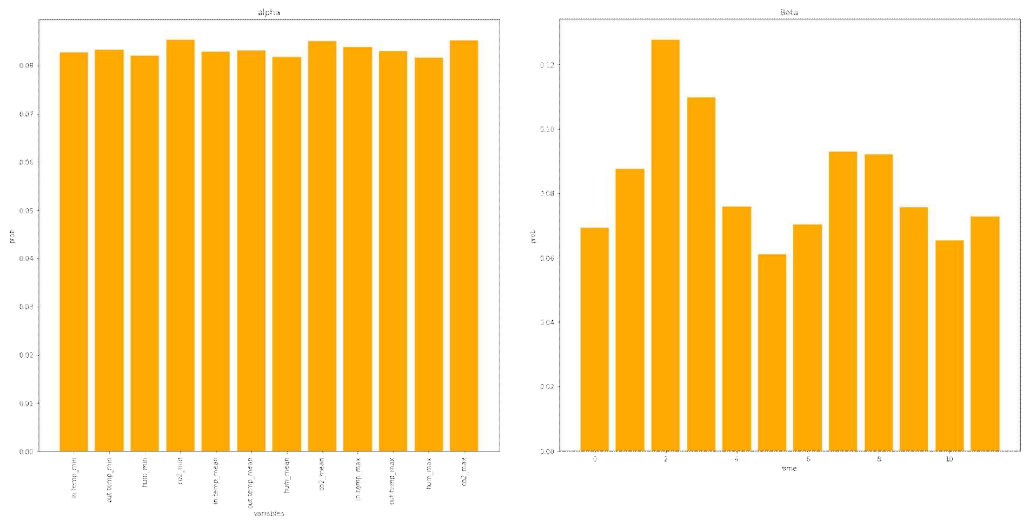

Figure 6. Input Attention Weights(Left) and Temporal Attention Weights(Right)

4.3 성능비교

같은 데이터를 이용하여 Dual Attention LSTM과 다른 모델의 예측력을 비교하였다. 비교모델은 일반적인 LSTM과 input attention만을 적용한 모델로 Input Attention LSTM은 12개의 변수를 길이 T의 라는 새로운 변수를 생성하여 LSTM에 적용한 모델이다. 두 모델에 대하여 Dual Attention LSTM과 같은 방식으로 최적화하였다. 세 모델의 예측력은 MSE를 기준으로 비교하였으며 결과는 표 3과 같다. LSTM의 MSE는 1.22로 가장 크게 나타났고 Input Attention LSTM과 Dual Attention LSTM이 각각 0.789와 0.678로 나타났다. 일반적인 LSTM에 비하여 attention 구조를 적용하였을 때 더 좋은 결과를 갖으며 input attention만을 적용할 때보다 temporal attention까지 적용한 모델이 좋은 성능을 보임을 확인하였다.

Table 2. Comparison Between Models

Measurement	LSTM	Input Attention	Dual Attention
MSE	1.220	0.798	0.678

Table 3. Input Weights(Left) and Temporal Weights(Right)

Environmental Variables	Weights	Time Steps	Weights
Minimum inside temperature	0.083	11 weeks ago	0.069
Mean inside temperature	0.083	10 weeks ago	0.088
Maximum inside temperature	0.084	9 weeks ago	0.128
Minimum outside temperature	0.083	8 weeks ago	0.110
Mean outside temperature	0.083	7 weeks ago	0.076
Maximum outside temperature	0.083	6 weeks ago	0.061
Minimum humidity	0.082	5 weeks ago	0.070
Mean humidity	0.082	4 weeks ago	0.093
Maximum humidity	0.082	3 weeks ago	0.092
Minimum CO2 level	0.085	2 weeks ago	0.076
mean CO2 level	0.085	1 week ago	0.065
Maximum CO2 level	0.085	0 week ago	0.073

5. 결 론

본 연구는 작기 동안 개별 농가에서 작물의 생산량을 예측하면서 예측에 중요한 환경변수와 시점에 대한 정보를 함께 추정하여 일반적인 딥러닝 방식과 달리 예측에 대한 설명이 가능한 작물의 생산량 예측 모델을 제안하였다. 전국단위의 연간 생산량을 대상으로 한 기존의 연구와 달리 각 농가별로 실측된 생산량과 환경정보를 이용하여 예측대

상과의 밀접성을 높이고 시설 재배 작물의 예측에 유용하다는 장점이 있다. 또한 많은 외생변수를 사용하기 힘든 통계적 모델에 비해 외생변수를 활용하기 쉬울뿐더러 예측결과에 대한 해석이 불가능한 딥러닝 모델에 비해 예측에 대해 비교적 설명이 가능한 장점이 있다. 예측모델로 자연어 처리에서 주로 이용하는 attention 구조를 적용한 Dual Attention LSTM을 이용하였으며 12주의 윈도우 사이즈를 갖는 모델의 MSE가 0.678로 LSTM, Input Attention LSTM과의 비교를 통해 가장 좋은 예측력을 보임을 확인하였다. 추출된 attention 가중치는 농가와 예측 시점에 따른 차이가 없이 거의 비슷한 결과를 보여주어 모델의 일반화가 가능한 것을 확인하였다.

모델에서 추출한 input attention 가중치은 12개 외생변수의 값이 크게 차이나지 않는 것으로 나타났다. 다시 말해 외부온도, 내부온도, 습도, CO2 농도가 토마토의 생산량 예측에 비슷한 정도로 중요하다는 것을 의미하며 결과적으로 토마토 생산량이 특정 환경변수에 의해 좌우되기 보다는 여러 요인들이 적절하게 유지되어야 함을 알 수 있다. 반대로 temporal attention 가중치은 예측 시점 이전의 여러 시점들 중 특정 시점의 중요도가 매우 높게 나타났다. 특히 9주전, 8주전의 환경이 가장 중요하게 나타났으며 이어서 4주전, 3주전 등의 순서였다. 이는 토마토 생산량이 특정 환경변수에 영향을 크게 받기보다는 여러 요인들에 의해 비슷한 영향을 받는 것과는 반대로 특정 시점의 환경에는 크게 영향을 받는다는 것으로 시점의 영향력이 다름을 의미한다. 따라서 기상이변이나 천재지변 같은 매우 큰 환경변화가 일어나지 않는다고 한다면, 오늘의 환경요인이 당장 내일의 토마토 생산량에 바로 영향을 끼치지는 않지만 몇 주 뒤의 생산량에는 영향을 크게 끼칠 수 있음을 알 수 있다.

본 연구에서 제안한 예측모델을 통해 개별 농가에서 수확이 시작된 후 12주 이후의 토마토 생산량을 예측할 수 있음은 물론 데이터가 확보된다면 다른 작물에도 쉽게 적용할 수 있을 것으로 기대된다. 또한 토마토와 같이 일정기간 지속적으로 생산되는 작물 뿐 아니라 예측대상의 설정에 따라 작기 동안 한 번의 수확이 이루어지는 작물의 생산량이나 생육에도 적용할 수 있을 것으로 기대된다.

REFERENCES

Alhnaity, B., Pearson, S., Leontidis, G., and Kollias, S. 2020. Using Deep Learning to Predict Plant Growth and Yield in Greenhouse Environments. Acta Hortic 1296:425-432.

Chen, T., Yin, H., Chen, H., Wu., L., Wang, H., Zhou, X., and Li, X. 2018. TADA : Trend Alignment with Dual-Attention Multi-Task Recurrent Neural Networks for Sales Prediction. 2018 IEEE International Conference on Data Mining 49-58.

Chen, Y., Lin and W., Wang, J. 2019. A dual-Attention-Based Stock Price Trend Prediction Model With Dual Features. IEEE Access 7:148047-148058.

Choudhury, A. and Jones, J. 2014. Crop Yield Prediction Using Time Series Models, Journal of Economic and Economic Education Research 15(3):53-68.

Chung, D. and Han, D. 2018. Evaluation of Forecasting Performance of Rice Yield Models under Climate Change. Korea Environmental Policy And Administration Society 26(4):197-222.

Dharmaraja, S., Jain, V., Anjoy, P., and Chandra, H. 2020. Empirical Analysis for Crop Yield Forecasting in India, Agric Res. 9(1):132-138.

Farook, A., J., Kannan, K., S. 2014. Climate Change Impact on Rice Yield in India-Vector Autoregression Approach, Sri Lankan Journal of Applied Statistics 16(3):161-178.

Feng, L., Zhao, C., Sun, Y. 2020. Dual attention-based encoder-decoder: A customized sequence-to-sequence

learning for soft sensor development. IEEE Transactions on Neural Networks and Learning Systems 1-12.

Han, M. and Yu, S. 2019. Prediction of Baltic Dry Index by Application of Long Short-Term Memory. Journal of Korean Society for Quality Management 47(3):497-508.

Hossain, M., M., Abdulla, F. 2015. On the Production Behaviors and Forecasting the Tomatoes Production in Bangladesh Journal of Agricultural Economics and Development 4(5):66-74.

Jiang, Z., Liu, C., Ganapathysubramanian, B. Hayes, and D., Sarkar, S. 2020. Predicting county-scale maize yields with publicly available data, Scientific Reports 10:14957.

Kim, N. and Lee, Y. 2016. Experimental Predictions of Crop Yields Using Time-Series Modeling of Climate Reanalysis Data: A Case of Iowa, USA, 1960-2009. The Korean Cartographic Association 16(2):115-126.

Lee, S., Yoon, Y, Jung, J., Sim, H., Chang, T., and Kim, Y. 2020. A Machine Learning Model for Predicting Silica Concentrations through Time Series Analysis of Mining Data. Journal of Korean Society for Quality Management 48(3):511-520.

Li, L., Wu, Y., Zhang, Y., and Zhao, T. 2018. Time+User Dual Attention Based Sentiment Prediction for Multiple Social Network Texts With Time Series IEEE Access 7:17644-17653.

Na, M., Cho, W., and Kim, S. 2020. A Construction of Web Application Platform for Detection and Identification of Various Diseases in Tomato Plants Using a Deep Learning Algorithm. Journal of Korean Society for Quality Management 48(4):58-596.

Oh, S. and Kim, M. 2017. Predicting Onion Production by Weather and Spatial Time Series Model. Journal of The Korean Data Analysis Society 19(5):2447-2456.

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, and G., Cottrell, G. 2017. A dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. arXiv preprint:1704.02971.

Shook, J., Gangopadhyay, T., and Wu, L., Ganapathysubramanian, B., Sarkar, S., Singh, A., K. 2020. Crop Yield Prediction Integrating Genotype and Weather Variables Using Deep Learning arXiv preprint:2006.13847.

저자소개

강수람 전남대학교 사회학과 학사, 전남대학교 수학통계학과 이학석사, 현재 전남대학교 통계학과 농업빅데이터 연구실에서 연구활동을 하고 있다. 주요 관심분야는 통계학, 빅데이터, 딥러닝 등이다.

조경철 전남대학교 원예학과에서 학사(1996년), 석사(1999년), 박사(2005년) 학위를 취득하고 현재 전남농업기술원에서 농업연구사로 재직하고 있다. 주요 연구분야는 농업빅데이터 수집 및 활용방안 등이다.

나명환 서울대학교 수학교육학과를 졸업하고 통계학과에서 석사와 박사학위를 취득하였다. 현재 전남대학교 통계학과에 재직하고 있으며, 전남대학교 통계연구소 소장, 농업빅데이터 연구회 회장, 스마트팜빅데이터연구실 지도교수, 인공지능 의학연구회 학술이사, 광주전남과총 기초과학분과 위원장, 한국품질경영학회 부회장・광주전남제주지회장, 한국신뢰성학회 운영이사, 한국통계학회 호남제주지회장, 한국표준협회 창의융합개발센터 전문위원으로 활동하고 있으며, 주요 관심 분야는 스마트팜, 스마트팩토리, 그린뉴딜 관련 빅데이터분석이다.