# LSTM과 GRU를 활용한 업종별 전력사용량 데이터 보간 방법

장민영, 고영준
충남대학교

# The Interpolation Method of Electricity Consumption Data by Business Using LSTM and GRU

Min-Young Jang, Yeong-Jun Koh
ChungNam National University

**Abstract** - 최근 전력공급 사업자들은 원격검침 인프라를 기반으로 전기사용자의 전력사용량 정보를 시간 단위로 수집하여 전기요금 과금, 실시간 조회 서비스 등을 제공하고 있다. 특히 전기사용자의 전력사용량 데이터는 국가적인 에너지 절감과 소비 효율성 향상을 위해 전기에너지 이용 패턴 분석에 많이 활용되고 있다. 전력사용량 데이터는 원격검침계기(스마트미터)를 통해 15분 또는 1시간 단위로 취득되고 있으며 인프라 특성상 일부 데이터의 유실이 발생하여 데이터 품질에 영향을 준다. 본 연구에서는 전력사용량 데이터의 품질 향상을 통해 에너지 이용 패턴 분석 활용에 도움이 되고자 전력사용량 데이터 보간 방법을 개선하고자 한다. 전력사용량 데이터는 업종별로 다양한 패턴을 가지고 있다. 이번 연구에서는 10개의 대표 업종을 선별하고 딥러닝을 활용한 업종별 특징 추출 및 학습을 통해 전력사용량 데이터 보간 방법의 가능성을 제시한다.

## 1. 서 론

전력사용량 데이터(Load Profile, LP)는 시간 단위로 측정된 대표적인 시계열 데이터이다. 전기사용자의 원격검침계기에서 생성되는 데이터는 유·무선 통신망을 이용하여 수집되며, 원격검침계기 상태와 통신망 환경에 따라 결측치 비율이 달리 나타난다. 전력사용량 데이터는 자기상관성이 존재한다. 그간 연구에서는 자기상관성을 기반으로 결측치 보간 모델을 제시하기도 하였다. 하지만 데이터에 의존적인 인공지능 기법에서는 자기상관성이 오히려 일반화 성능을 떨어뜨려 문제가 되기도 한다. 그렇기 때문에 유사 패턴의 데이터끼리 그룹화하여 학습한다면 인공지능의 일반화 성능을 향상시킬 수 있다. 본 연구는 이런 아이디어의 실현과 검증 차원에서 시작하였다. 시계열 데이터 분석기법에서 주로 활용되는 RNN 계열의 LSTM, GRU를 활용하여 다양한 경우의 전력사용량 데이터 결측치 보간을 실험하였다.

## 2. 본 론

### 2.1 LSTM, RNN 모델 적용 개요

LSTM(Long Short-Term Memory)는 RNN(Recurrent Neural Network)의 한 종류로 기존의 RNN에 있던 vanishing gradient 문제를 효율적으로 해결할 수 있는 모델이다. LSTM 아키텍처의 특징은 다음과 같다. 메모리 역할을 하는 각각의 뉴런에 학습이 가능한 '입력', '출력', '망각' 게이트가 추가되며, 현재 시간 스텝(t)의 입력 값이 메모리에 적용되는 여부는 '입력' 게이트가, 다음 은닉층(혹은 출력층)으로의 전파 여부는 '출력' 게이트가, 현재까지 순차적 데이터가 적용된 메모리 값의 유지 여부는 '망각' 게이트가 담당한다.

GRU(Gated Recurrent Unit)은 2014년 발표된 모델로 LSTM의 구조를 보다 단순하게 처리한 LSTM 변형모델의 하나이다. GRU의 구조는 LSTM과 마찬가지로 게이트를 이용하여 정보의 양을 조절하는 것은 같지만, '망각' 게이트와 '입력' 게이트를 '갱신' 게이트로 통합하고, 이를 기준으로 상호 작용하는 4개의 층이 존재한다. 입력 데이터를 선택적으로 학습시키기 위해 세 개의 셀 상태의 게이트에서 정보를 더하거나 지우는 구조를 가지고 있다. LSTM보다 단순한 구조로 가중치 수가 작으므로 학습이 더 빠르지만 LSTM과 거의 같은 성능을 보인다.

본 연구에서는 업종별 전력사용량 데이터 보간을 위해 LSTM과 GRU를 활용하였으며, 두 모델의 정확도와 속도를 같이 비교분석하였다.

### 2.2 전력사용량 데이터 결측치 형태

결측 데이터 종류는 크게 3가지 형태로 나눌 수 있다. 첫 번째, 완전 무작위 결측(MCAR : Missing Completely At Random)은 전체에 걸쳐 무작위하게 누락된 경우로 변수의 종류, 변수의 값과 상관없이 비슷한 분포로 누락된 데이터를 의미한다. 두 번째, 무작위 결측(MAR : Missing At Random)은 어떤 특정 변수에 대하여 데이터가 누락되는 경우를 의미하며, 결측값의 경우가 자료 내의 다른 변수와 관련이 있다. 세 번째, 비무작위 결측(MNAR : Missing Not At Random)은 누락되는 부분들이 무작위로 누락되는 것이 아닌 누락된 변수의 값이 누락된 이유와 관련이 있는 경우이다.

전력사용량 데이터는 원격검침계기(스마트미터)나 통신망의 품질 문제로 데이터가 누락되며, 이는 비무작위 결측(MNAR)과 같다. 그렇기 때문에 결측값이 있는 데이터를 제거하고 분석을 진행할 경우, 모델이 편향적으로 학습될 수 있기 때문에 상황에 맞는 결측치 보간 및 처리 방법이 중요하다. 이러한 이유로 본 연구에서 전력사용량 데이터의 보간은 결측치가 거의 없는 고품질 데이터를 활용하여 시계열 모델링을 통해 미래값을 예측하는 방법을 선택하였다.

### 2.3 보간 모델의 입력 데이터 설계

전력사용량 데이터 보간을 2가지 방식으로 검토하였다. 첫 번째는 전력사용량 데이터의 자기상관성을 이용하여 전기사용자 개별의 전력사용량 데이터를 학습 데이터로 보간하는 방법이다. 두 번째는 업종별 전력사용 패턴의 상관성을 이용하여 업종별 전력사용량 데이터를 그룹화하여 학습하는 방법이다. 2가지 방식의 타당성 검증을 위하여 전력사용량 데이터의 상관계수를 분석하였다. 상관계수는 pearson 상관계수를 이용하여 일(24시간) 단위로 계산하였다.

**<표 1> 전력사용량 데이터 자기상관계수와 업종별 상관계수 비교**

| 구 분 | 최소값 | 최대값 | 평균값 |
|---|---|---|---|
| 전력사용량 데이터 자기상관계수 | 0.035 | 0.969 | 0.616 |
| 업종별 전력사용량 데이터 상관계수 | 0.347 | 0.537 | 0.459 |

전력사용량 데이터의 자기상관계수의 평균값은 0.616으로 업종별 전력사용량 데이터 상관계수보다 약 0.157이 높게 나왔다. 하지만 최소값은 0.035로 상관성이 매우 낮은 전기사용자 데이터도 존재하여 학습 데이터로서 다소 부적합하다. 그렇기 때문에 업종별 전력사용량 데이터를 보간 모델의 학습 데이터로 선정하였다.

전력사용량 데이터는 기상 데이터와 상관관계는 이전의 많은 연구에서 다루어진 내용이다. 본 연구에서는 입력변수 설계와

검증 차원에서 업종별 전력사용량과 기상 데이터간의 상관관계를 분석하였다. 기상 데이터는 기상청에서 개방한 데이터를 수집하여 온도, 강수량, 습도 데이터를 요인으로 선정하였다.

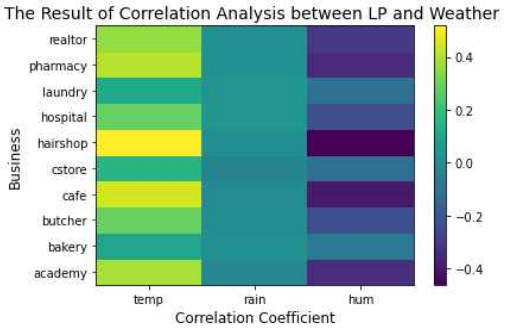


**그림 1. 전력사용량(LP)과 기상 데이터간의 상관분석 결과**

상관관계 분석결과 전력사용량은 온도와는 양(+)의 상관관계를 보이고, 습도와는 음(-)의 상관관계를 보이고 있다. 세탁소, 편의점, 빵집 업종은 다른 업종과 비교하여 약한 상관관계를 가지고 있다. 업종에 따라 일부 다른 특성을 보이지만 다수 업종에서 전력사용량과 온도, 습도 데이터의 상관관계는 유의미하다고 할 수 있다. 특히 미용실의 경우 온도, 습도와 모두 가장 강한 상관계를 보이고 있다.

상기 분석결과에 따라 보간 모델의 입력 변수는 분석의 효율성을 고려하여 전력사용량과 온도 데이터로 선정하였다.

### 2.4 LSTM, GRU 활용 전력사용량 데이터 보간 모델링

LSTM, GRU를 이용한 데이터 보간 방법은 이전 연구의 사례를 보면 과거 시퀀스 데이터를 이용하여 결측치를 예측하는 방식을 제안하였다. 본 연구에서도 이와 같은 방식을 채택하여 과거 **24시간**의 데이터 시퀀스를 기반으로 결측치를 예측하는 형태로 데이터 보간 모델링을 수행하였다. 24시간 데이터를 사용하는 이유는 전기사용자의 전력사용 패턴이 하루를 주기로 반복적인 패턴을 보이기 때문이다.

이번 단계에서는 LSTM, GRU를 활용하여 전력사용량 보간 모델링을 실험하였다. 두 알고리즘에 동일한 학습 데이터와 하이퍼 파라미터를 적용하여 정확도(Accuracy)와 학습 실행시간의 차이를 비교 분석하는 것을 목표로 하였다.

모델 정확도의 평가지표는 MAPE(Mean Absolute Percentage Error), R2 Score를 검토하였다. MAPE는 직관적으로 에러율을 판단하기에 좋으나, 학습 데이터가 0값에 가까운 값이 존재하여 스케일 측면에서 부적합하여 R2 Score를 평가지표로 선정하였다.

모델 정확도와 학습 실행시간 측정 결과는 다음과 같다.

**<표 2> LSTM, GRU 모델 정확도, 학습 실행시간 비교표**

| 구 분 | LSTM | GRU | 차 이 |
|---|---|---|---|
| 정확도(R2 score) | 0.9628 | 0.9589 | 0.0039 |
| 학습 실행시간(초) | 1271.0944 | 1085.8544 | 185.24 |

측정 결과 학습 정확도는 LSTM이 GRU와 비교하여 정확도는 0.0039 높았고, 학습 실행시간은 185.24초 늦었다. 보간 모델의 경우 결측치의 정확한 실제값을 예측하는 문제로 에러율이 무엇보다 중요하다. GRU가 속도 측면에서 이득이긴 하지만 에러율이 LSTM보다 높기 때문에 보간 모델에서는 LSTM을 사용하는 이득이 더 크다. 다음 단계에서는 LSTM을 활용하여 보간 모델 개선을 수행하였다.

### 2.5 양방향 LSTM 활용 전력사용량 데이터 보간 모델링

보간 모델은 일반적인 회귀 모델과는 달리 결측 데이터의 시점을 중심으로 과거와 미래 데이터가 존재한다. 그렇기 때문에 LSTM을 양방향으로 적용하여 결측치 예측이 가능하다. 다른 연구 사례에서도 시계열 데이터에서 양방향 LSTM 모델이 조금 더 나은 성능을 보인 결과가 있다.

본 연구에서도 상기 연구 결과를 적용하여 Forward, Backward 방향으로 LSTM 모델을 적용해 보기로 하였다. forward 모델($x_i$), backward 모델($y_i$), foraward와 backward 모델의 결과를 합친 bidirectional 모델($z_i$)로 정확도를 측정하였다. bidirectional 모델은 forward 모델과 backward 모델에 동일한 weight($z_i = 0.5x_i + 0.5y_i$)를 적용하여 수행하였다.

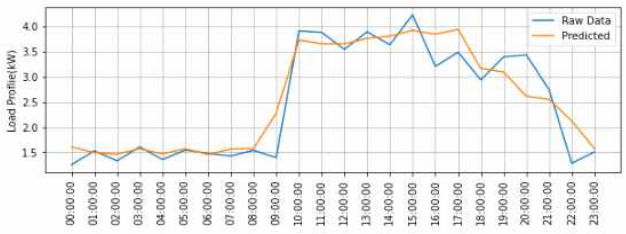


**그림 2. Forward LSTM 모델($x_i$) 예측 결과**

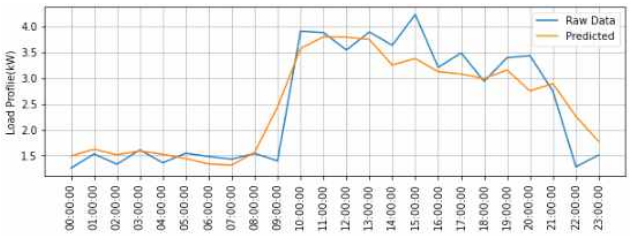


**그림 3. Backward LSTM 모델($y_i$) 예측 결과**

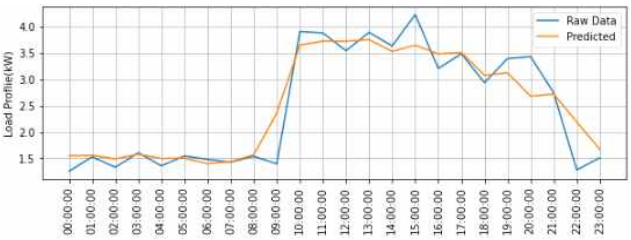


**그림 4. Bidirectional LSTM 모델($z_i$) 예측 결과**

1시간 데이터(결측치) 예측 모델의 정확도 측정 결과는 다음과 같다.

**<표 3> LSTM 1시간 결측치 예측 모델 정확도**

| 구 분 | Forward | Backward | Bidirectional |
|---|---|---|---|
| 정확도(R2 score) | 0.9628 | 0.9706 | 0.9753 |

1시간 데이터(결측치) 예측 모델 정확도는 Bidirectional > Backward > Forward 순으로 나타났다. Bidirectional 모델이 Unidirectional 모델보다 더 좋은 성능을 보였다.

실제 전력사용량 데이터의 결측은 연속적으로 나타날 수 있다. LSTM 활용 전력사용량 보간 모델의 실효성을 검증하기 위하여 2, 3시간 데이터(결측치) 예측 모델링을 추가 수행하였다. 최종 결과는 다음과 같다.

**<표 4> LSTM 1~3시간 결측치 예측 모델 정확도(R2 Score)**

| 구 분 | Forward | Backward | Bidirectional |
|---|---|---|---|
| 1시간 결측치 예측 모델 | 0.9628 | 0.9706 | 0.9753 |
| 2시간 결측치 예측 모델 | 0.9619 | 0.9605 | 0.9526 |
| 3시간 결측치 예측 모델 | 0.9633 | 0.9643 | 0.9747 |

1시간과 3시간 예측 모델은 유사한 정확도가 측정되었다. 2시간 예측 모델은 정확도가 1시간, 3시간 보다 낮게 측정되었고, Bidirectional 모델의 결과가 더 낮게 나왔다. 약간의 변동폭이 있긴 하나, 3가지 예측 모델 모두 R2 Score가 0.95 이상으로 안

정된 모델 성능을 보였다.

## 3. 결 론

전력사용량 데이터는 다양한 분야의 데이터와 결합하여 활용될 수 있는 가치가 높은 데이터이다. 특히 최근에는 전력과 통신 데이터를 결합하여 사회취약계층을 모니터링하는 서비스를 개발하는데 이용 되기도 하였다. 전력사용량 데이터의 결측치 보간 방법의 정확도를 향상시키면 고품질 데이터를 이용하여 다양한 분야의 서비스를 개발하는데 활용할 수 있다. 이번 연구에서는 LSTM, GRU를 활용하여 전력사용량 데이터 보간 성능을 향상시킬 수 있었다. 특히 LSTM, GRU의 일반화 성능을 향상을 위하여, 전력사용량 데이터를 업종별로 그룹화하여 사용하였고, 그 결과를 높은 성능의 딥러닝 모델을 개발할 수 있었다. 또한 양방향(Bidirectional) LSTM을 적용하여 일반적인 LSTM보다 높은 성능에 도달할 수 있었다. 이번 모델에서는 입력 데이터로 전력사용량 데이터와 온도 데이터만 활용하였으며, 딥러닝 모델의 특성상 입력 변수를 보강한다면 좀 더 나은 성능을 기대할 수 있을 것으로 보인다.

**감사의 글**

본 연구는 2022년 한국전력공사에서 운영중인 전력데이터 공유센터 내부에서 학술연구 목적으로 공유된 데이터를 활용하여 연구하였음을 알려드립니다.

## [참 고 문 헌]

[1] Wei Song, Chao Gao, Yue Zhao and Yandong Zhao, “A Time Series Data Filling Method Based on LSTM－Taking the Stem Moisture as an Example, MDPI Sensors, 20(18), pp.5045, 2020

[2] 이원규, 안소영, 임민섭 외 1명, “LSTM을 이용한 전력 데이터 예측”, 한국정보과학회 학술발표논문집, 2016.12, pp.693-695, 2016

[3] 류경근, 최용철, 이덕규, “GRU를 활용한 악성코드 탐지의 관한 연구”, 2020 온라인 춘계학술발표대회 논문집, 제27권 제1호, pp.254-257, 2020

[4] 김에덴, 고석갑, 손승철, 이병탁, “시계열 데이터 결측치 처리 기술 동향”, ETRI DOI, S0317-21-1001, 2021