

GRU 기반의 도시부 도로 통행속도 예측 모형 개발

Development of a Speed Prediction Model for Urban Network
Based on Gated Recurrent Unit

김 호 연* · 이 상 수** · 황 재 성***

* 주저자 : 아주대학교 교통공학과 석사과정

** 교신저자 : 아주대학교 교통공학과 교수

*** 공저자 : 아주대학교 교통공학과 연구원

Hoyeon Kim* · Sangsoo Lee* · Jaeseong Hwang*

* Dept. of Transportation Eng, Univ. of Ajou

† Corresponding author : Lee Sangsoo, sslee@ajou.ac.kr

Vol. 22 No.1(2023)
February, 2023
pp.103~114pISSN 1738-0774
eISSN 2384-1729
<https://doi.org/10.12815/kits.2023.22.1.103>Received 30 November 2022
Revised 2 December 2022
Accepted 25 December 2022© 2023. The Korea Institute of
Intelligent Transport Systems. All
rights reserved.

요 약

본 연구에서는 도시부 도로의 다양한 자료를 수집하여 통행속도 변화에 대한 영향을 분석하였고, 이와 같은 빅데이터를 활용하여 GRU 기반의 단기 통행속도 예측 모형을 개발하였다. 그리고 Baseline 모형과 이중지수평활 모형을 비교 모형으로 선정하여 RMSE 지표로 예측 오차를 평가하였다. 모형 평가 결과, Baseline 모형과 이중지수평활 모형의 RMSE는 평균 7.46, 5.94 값으로 각각 산출되었다. 그리고 GRU 모형으로 예측한 평균 RMSE는 5.08 값이 산출되었다. 15개 링크별로 편차가 있지만, 대부분의 경우 GRU 모형의 오차가 최소의 값을 나타내었고, 추가적인 산점도 분석 결과도 동일한 결과를 제시하였다. 이러한 결과로부터 도시부 도로의 통행속도 정보 생성 과정에서 GRU 기반의 예측 모형 적용 시 예측 오차를 감소시키고 모형 적용 속도의 개선을 기대할 수 있을 것으로 판단된다.

핵심어 : 딥러닝, 교통 혼잡, 통행속도, GRU, 빅데이터

ABSTRACT

This study collected various data of urban roadways to analyze the effect of travel speed change, and a GRU-based short-term travel speed prediction model was developed using such big data. The baseline model and the double exponential smoothing model were selected as comparison models, and prediction errors were evaluated using the RMSE index. The model evaluation results revealed that the average RMSE of the baseline model and the double exponential smoothing model were 7.46 and 5.94, respectively. The average RMSE predicted by the GRU model was 5.08. Although there are deviations for each of the 15 links, most cases showed minimal errors in the GRU model, and the additional scatter plot analysis presented the same result. These results indicate that the prediction error can be reduced, and the model application speed can be improved when applying the GRU-based model in the process of generating travel speed information on urban roadways.

Key words : Deep learning, Traffic congestion, Traffic speed, Gated recurrent unit, Big data

I. 서론

도시부 도로의 교통 혼잡은 해결하기 어려운 사회적 과제 중 하나로, 삶의 질을 저하시키는 각종 교통사고나 대기오염 문제를 유발하는 요인으로 평가되고 있다. 한국교통연구원이 추정한 국내 교통혼잡 비용 분석 결과에 따르면, 최근 5개년간 국내 교통혼잡 비용은 꾸준히 증가하고 있고, 2019년도에는 약 70.6조 원으로 나타났다. 이는 도로를 주행하는 차량이 교통혼잡으로 인해 정상 속도 이하로 운행하게 됨으로써 발생하는 시간가치의 손실 및 차량 운행비 증가의 비용을 산정한 것으로, 이러한 혼잡비용은 앞으로도 계속 증가할 것으로 전망되고 있다. 전통적으로는 도로 시설에 관한 인프라를 확장하는 방식으로 교통 혼잡을 해결하고자 하였으나, 이는 곧 신규도로 건설 비용보다 차량증가율이 앞서게 되면서 지능형 교통 시스템의(ITS, Intelligent Transport System) 도입이 추진되었다(Lee et al., 2014). 지능형 교통 시스템의 주요 역할 중 하나로 교통 정보 예측이 있으며, 효율적인 교통운영 및 통행 수요 관리를 위해서는 교통 정보 예측의 정확도가 중요하다.

그러나 현재 국내 교통정보센터에서 활용되는 교통 정보 예측 모형을 검토한 결과, 교통 정보 수집 장치에서 수집되는 단일 데이터만을 활용하여 정보를 생성하고 있다. 연속류 도로의 경우, 이와 같은 접근 방법을 통하여 의미 있는 정확도를 얻을 수 있지만, 단속류인 도시부 도로의 경우 신호교차로를 포함하여 다양한 요소들이 예측 정보에 큰 영향을 주고 있다. 일반적으로 도시부 도로의 교통 정보 예측 정확도는 신호교차로를 포함한 외부요인의 영향으로 연속류 도로와 비교하여 오차율이 매우 높은 것으로 알려져 있지만 이에 관한 연구가 상대적으로 부족한 실정이다. 기존에 사용되는 도시부 통행속도 예측 모형은 다양한 요소들의 연계성이 고려되지 않는 것으로 확인되었다. 최근에 국내에서는 도시부 도로에 스마트 교차로를 구축하여 다양한 자료를 수집할 수 있도록 노력하고 있다. 이러한 정보에 기상적 자료, 운전자 행태를 반영하는 자료, 돌발상황 이벤트 자료등을 포함하여 교통 정보를 생성하는 모형을 개발한다면 정보 예측의 정확도가 개선될 것으로 기대할 수 있다. 또한 향후 자율협력주행을 지원하기 위한 대규모의 데이터가 축적될 것으로 예상되므로 다양한 자료를 융합하는 모형의 개발에 대한 필요성이 요구되며, 이러한 새로운 방향에 대한 탐색이 필요하다.

본 연구의 목적은 도시부 도로망을 대상으로 다양한 자료를 수집하여 통행속도 변화에 대한 영향 분석을 수행하고, 이러한 빅데이터(Big Data) 자료를 이용하여 통행속도 예측 모형을 개발하는 것이다. 이를 위해, 스마트 교차로가 설치된 도심부 도로에서 신호 운영 자료, 교통류 자료, 상·하류 인접 교통류 자료와 기상 자료 등을 결합하여 빅데이터 자료를 구축하였으며, PFI(Permutation Feature Importance)를 통한 데이터 선별 과정을 수행하였다. 이러한 빅데이터를 처리하기 위하여 Deep Learning 기법 중 GRU(Gated Recurrent Unit) 기반의 단기 예측 모형을 구축하였다. 또한 모형 간 비교 분석을 통해 최적의 예측률을 보이는 Model Structure 및 Hyper Parameter를 제시하였다. 본 연구를 통하여 도시부 도로의 통행속도를 예측하는 빅데이터 기반의 연구가 활발하게 진행될 것으로 기대한다.

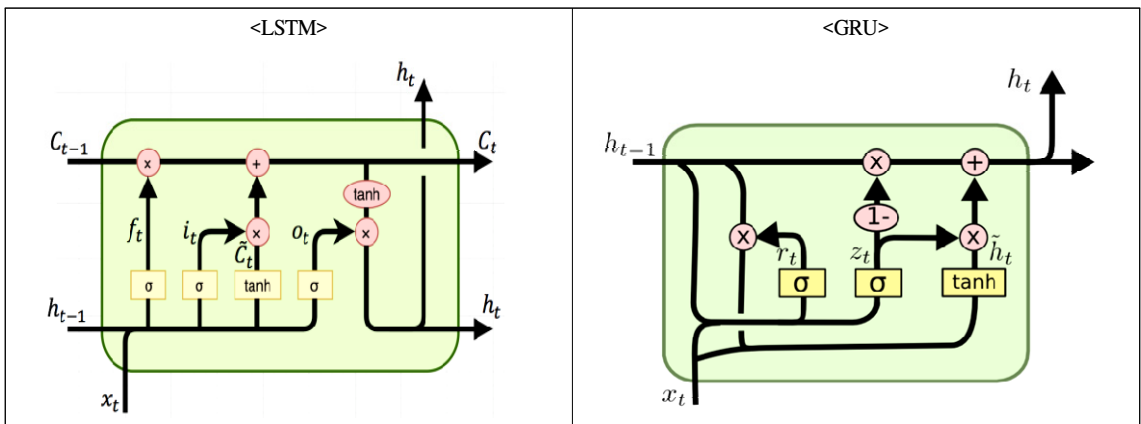
II. 이론적 배경

1. GRU(Gated Recurrent Units)

딥러닝 기법 중 RNN(Recurrent Neural Network)은 시계열 데이터 및 언어 학습에서 활용되었으나, 이내 가

중치 조절을 위한 역전파(Backpropagation) 과정에서 심층적인 신경망으로 전개됨에 따라 장기 의존성 학습이 어려운 문제에 직면하게 되었다. 이후 이러한 Vanishing Gradient 문제를 극복하고자 LSTM(Long Short Term Memory) 구조가 제시되었다. LSTM은 Input Gate, Forget Gate 그리고 Output Gate까지 총 3개의 gate로 구성되어 있다. 학습 과정에서 Input Gate 및 Forget Gate를 통해 장기 기억 값이 출력되며, Cell State 내에서 (+) 연산을 거쳐 Gradient 값을 효과적으로 지켜낸다. 또한 이 값을 Output Gate를 거쳐 다음 Cell State를 위한 Input 값으로 활용되며 학습이 진행된다.

이후 LSTM의 학습 과정을 개선하고자 더욱 간소화한 구조로 변형하여 GRU가 제안되었다. 일반적으로 GRU는 LSTM과 동등한 기능을 낼 수 있으며, 계산 속도 면에서는 더욱 효율적인 것으로 알려져 있다(Jeong et al., 2021). GRU는 Reset Gate와 Update Gate 2개의 Gate로 구성되어 있으며, LSTM과 비교한 GRU의 구조는 <Fig. 1>에 제시하였다.



<Fig. 1> LSTM and GRU cell architectures

GRU에서는 LSTM에서의 Cell State와 Hidden State가 Hidden State로 통합되고, LSTM의 Output Gate는 GRU에 존재하지 않는다. Reset Gate에서는 전 시간대에서의 Hidden State Output 값과 현 시간대의 Input 시계열 값을 현 Hidden State에 얼마나 반영할지 결정하며, Update Gate는 LSTM의 Forget Gate와 Input Gate와 유사한 기능을 수행하여 정보의 최신화 비율을 결정한다. GRU의 상세한 계산 과정은 다음 식 (1)과 같다.

$$\begin{aligned}
 z_t &= \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \dots\dots\dots (1) \\
 r_t &= \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \\
 \hat{h}_t &= \phi_h(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\
 h_t &= z_t \odot \hat{h}_t + (1 - z_t) \odot h_{t-1}
 \end{aligned}$$

- 여기서, x_t = 시간 t의 입력값
- h_t = 시간 t의 출력값
- \hat{h}_t = 후보군 결과값
- z_t = 업데이트 게이트 결과값
- r_t = 리셋 게이트 결과값

W, U, b = 매개변수 가중치
 σ_g = 시그모이드 함수
 ϕ_h = 하이퍼볼릭 탄젠트 함수
 \odot = 점별 연산

2. Brown's Double Exponential Smoothing

단순 지수평활법에서는 데이터에 대해 α 가중치를 곱해줌으로써 최근 값과 과거 값의 반영 비율을 조정하여 예측에 반영한다. 그러나 이는 데이터에 추세가 존재하면 예측력이 떨어진다는 단점이 존재하였다. 이에 데이터의 추세를 예측에 반영하고자 다양한 이중지수평활법이 제안되었다. 이 중 Brown(1959)의 이중지수평활법은 단순한 모형 식을 통해 예측치를 추정할 수 있어 최근까지도 시계열 데이터 분석 시에 활용되고 있다. 이중지수평활법에서는 데이터의 추세 흐름을 반영하기 위해 가중치 계산을 반복하여 수행하는 것이 특징이며, Brown 기법의 구체적인 계산 과정은 아래 식 (2)에 제시하였다.

$$\begin{aligned}
 S'_t &= \alpha_p X_t + (1 - \alpha_p) S'_{t-1} \dots\dots\dots (2) \\
 S''_t &= \alpha_p S'_t + (1 - \alpha_p) S''_t \\
 a_t &= S'_t + (S'_t - S''_t) = 2S'_t - S''_t \\
 b_t &= \frac{\alpha_p}{1 - \alpha_p} (S'_t - S''_t) \\
 F_{t+m} &= a_t + b_t m
 \end{aligned}$$

여기서, S'_t = 단순지수평활값
 S''_t = 반복지수평활값
 m = 예측 기간
 α_p = 지수평활 가중치, $0 < \alpha_p < 1$
 a_t, b_t = 평활 상수

3. Baseline Model

딥러닝 모형은 대개 복잡하고 모형 구축 과정에서 데이터 학습 등에서 다소 시간이 소요되므로, 구축된 예측 모형이 최소 성능 기준 이상의 효과를 내는지 확인하는 것이 중요하다. 이를 위해 간단하고 직관적인 비교 모형을 구축하여 예측값의 비교 평가를 수행하는데, 이러한 모형을 Baseline Model(모형)이라고 한다. 특히 시계열 데이터 학습에서 Baseline 모형은 주로 전 주기의 값이 예측값으로 활용되고 있다. 본 연구에서도 전 주기의 값으로 예측값을 추정하는 Baseline 모형을 구축하였고, 다음 식 (3)과 같다.

$$\text{시간 } t \text{에서, } \hat{y}_{t+1} = y_t \dots\dots\dots (3)$$

여기서, \hat{y}_{t+1} = 한 주기(5분) 이후 예측 통행속도
 y_t = 현재 통행속도

4. 교통 혼잡 예측 관련 연구

Abduljabbar et al.(2021)은 고속도로 실시간 속도 데이터의 공간적, 시간적 특성을 모두 고려하는 LSTM 모형을 구축하였다. 고속도로 각 구간의 4개의 검지기를 대상으로 데이터가 10분 이상 누락될 경우 Data Set을 제거하도록 처리하였으며, 5분에서 60분 사이의 예측을 수행하였다. 예측 결과를 RNN, GRNN, MNN 등 타 신경망 모형과 비교하여 분석한 결과 특정 검지기(14031)를 제외하고 타 모형에 비해 우수한 예측 정확도를 나타냈다. 또한 향후 예측 정확도 향상을 위해서는 날씨와 같은 외부요인뿐만 아니라 교통량, 점유율 등의 다양한 입력 데이터를 구축할 필요가 있다고 하였다.

Gao et al.(2022)은 교통 속도 데이터 없이 교통량 데이터를 기반으로 단기 통행속도를 예측할 수 있도록 예측 모형을 구축하여 비교 평가하였다. 2분 단위의 교통량 및 속도 데이터를 시간 흐름에 따른 패턴 데이터로 구성하여 교통량 변동 패턴에 따라 속도를 예측하도록 학습하였다. 최적의 패턴 데이터 구축 주기를 분석한 결과 3주기(6분) 단위의 데이터 패턴 구축이 속도 예측 시 가장 효과적인 것으로 나타났다. 예측 모형으로 LSTM 모형과 CART, KNN, SVR 등의 머신러닝 모형을 비교한 결과 LSTM의 MAPE(Mean Absolute Percentage Error)가 변동폭이 안정적이며 평균 MAPE가 가장 낮은 3.29%로 도출됨을 확인하였다.

Jeon(2018)은 통행속도 예측 시 시공간적 상관을 고려하고자 이미지 인식에 뛰어난 CNN(Convolution Neural Network)모형을 구축하였으며, 학습을 위해 속도 데이터를 도로 이미지 픽셀로 변환하였다. 15분 통행 시간을 예측하여 KNN, SVM, ARIMA 모형과 비교한 결과, 테스트 데이터의 산점도에서 확인한 차이를 알 수 있었으며, RMSE가 2.82로 나타나 가장 낮은 오차율로 나타났다.

Jeong et al.(2021)은 고속도로 DTG 속도 데이터를 활용하여 구간속도를 예측하는 모형으로 GRU, LSTM, ARIMA 모형을 구축하여 각 성능을 비교하였다. 시계열 데이터의 구성은 15분의 시간 간격을 갖는 데이터를 Training, Validation, Test Set으로 각각 10일씩 구분하였다. 버스전용차로와 일반차로를 구분하여 예측 모형을 구축하여 MAE 값을 비교한 결과 GRU, LSTM, ARIMA 순으로 높은 성능을 보였다. 각 모형의 예측치에 대한 ANOVA Test를 수행한 결과, 버스전용차로 Data Set에서는 그룹 간 유의미한 차이가 없지만, 일반차로 Data Set에서는 그룹 간의 유의미한 차이가 있음을 나타냈다. 학습 시간 분석 결과, GRU 모형의 학습 시간 단축이 통계적으로 유의미한 것으로 나타났다.

Kim et al.(2020)은 5분 단위 구간통행속도 정보를 활용하여 LSTM 기반의 통행속도 예측 모형을 개발하였다. 입력변수로 전날 동 시간대의 과거 통행속도 등을 활용하여 시간대에 따라 일정한 교통류 패턴이 발생하는 점을 고려하였으며, 상·하류의 직전 시간대 통행시간을 통해 공간적 요소의 영향을 반영하고자 하였다. 비교 모형으로 직전 시간대의 통행속도 한 가지 변수로만 학습한 LSTM 모형을 구축하고 서울시 교통정보센터에서 제공되는 예측치와 비교한 결과, 해당 모형이 두 예측치에 비해 MAPE가 감소하여 예측력이 개선된 것으로 나타났다.

Mahmoud et al.(2021)은 단일로에서의 이동류별 교통량을 예측하기 위해 LSTM, GRU, XGBoost 모형을 구축하여 비교·분석하였다. 데이터는 플로리다 도심부의 16개 교차로를 대상으로 1개월간 교차로별 신호 주기, 녹색 시간, 방향별 교통량의 데이터를 수집하여 학습하였다. 연속된 3개의 교차로를 한 그룹으로 하여 중간 교차로를 예측 대상으로 선정하고, 직전 교통량과 좌회전 교통량 예측 모형을 각각 구축하였다. 모형의 평가 지표로 MAE 및 RMSE를 선정하여 비교한 결과, 모형 간 성능 지표는 크게 다르지 않았으나 6개 주기를 학습한 GRU 모형이 최적 모형인 것으로 나타났다.

Mussumeci and Coelho(2020)은 계절성 질병의 정확한 발병 예측을 위한 머신러닝 모형으로 LASSO 회귀모형, Random Forest 회귀모형, LSTM 기반 모형을 구축하여 각 모형의 예측 성능을 비교하였다. 분석 결과, 예

측 정확도 및 편향도 면에서는 LSTM 모형의 예측력이 가장 우수하며, RF, LASSO 순으로 우수하게 도출되었다. RF 및 LASSO 회귀모형은 자료 훈련에 몇 초의 시간이 소요되었으나 LSTM 모형은 동일 자료에서 10분 이상 소요됨을 확인하였다.

Peng et al.(2018)은 교통량을 예측함에 있어 강수 데이터를 예측 모형에 통합할 필요가 있다고 판단하였다. 5개월간의 교통 데이터 중 주말 데이터 및 야간 데이터를 제외하여 혼잡한 시간대의 예측을 수행하였다. 12주간 평균치를 성능 기준 값으로 한 Baseline model과 SARIMA, Exponential Smoothing, Neural Network 모형을 구축하여 다음 날 1시간 단위의 24시간의 MAPE를 평가하였다. 그 결과 모든 모형에서 강수량을 포함한 데이터로 학습한 시나리오가 개선된 MAPE를 나타냈고, 그 중 Neural Network의 예측 모형이 가장 낮은 오차율을 보였다.

Xu et al.(2022)은 기존의 머신러닝 기반의 예측 방법이 교통 데이터의 시공간적 상관관계 및 비선형관계로 인해 예측 정확도가 낮으며, 장기적인 시간 변화를 학습할 필요가 있다고 하였다. 해당 연구에서는 5분 간격의 고속도로 속도 데이터를 활용하였다. GCN 및 GRU를 결합하여 모형을 구축하였고 15분, 30분, 45분 후를 예측하였다. 비교 모형으로 예측 기간의 평균치인 History average(HA), ARIMA, 딥러닝 기반의 STGCN, GMAN, W-GRU-ARMA를 비교한 결과, 훈련 시간 및 예측력을 종합적으로 평가할 때 해당 모형의 효율이 가장 높다고 평가하였다. 또한 GRU의 경우 대개 LSTM과 예측력은 동등하나 훈련 시간 면에서 크게 향상될 수 있음을 제시하였다.

Zhang et al.(2022)은 기존에 개발된 속도 예측 모듈에서 발생하는 중장기 예측 오차를 극복하기 위해 시공간적 특성을 추출하여 결합하는 Temporal Attention Convolutional Network 모형을 구축하였다. 모형은 과거 24개의 속도 시퀀스 및 노드 간의 거리에 따라 가중치 행렬을 학습하여 미래 24개의 속도 시퀀스를 예측하도록 하였다. 고속도로와 도시부 자료를 대상으로 각각 실험한 결과 도시부 자료의 예측 오차가 높은 것으로 나타났다.

5. 연구의 차별성 및 시사점

기존 연구의 경우 고속도로와 같은 연속류 도로 구간을 대상으로 하는 연구가 대다수이고, 교통량, 혹은 통행속도 등 교통 혼잡 예측을 위한 목표 변수는 다양하나 예측 모형의 학습데이터 구축 과정에서 주로 단편적인 데이터만을 고려하여 모형을 구축하였다. 그러나 각 변수는 인접 도로의 교통류 흐름 변화나 기상, 돌발상황 등 다양한 외부요인에 의해 영향을 받으며, 이런 특성은 도시부 단속류 도로에서 더욱 두드러져 연속류 도로와 비교하여 예측 오차율이 큰 것으로 확인된다. 본 연구에서는 신호 운영 변수, 교통류 변수, 상하류 인접 교통류 변수와 기상 변수 등을 결합한 빅데이터 자료를 구축하여 다양한 변수 간의 상관관계를 분석하였고 이를 반영하여 모형을 구축하였다. 예측 모형은 딥러닝 기법 중 기존 LSTM의 구조를 경량화하여 모형의 학습 속도를 개선한 GRU를 적용하였다. 이는 딥러닝 모형의 경우 모형 layer가 깊어지고 심층적인 학습이 진행됨에 따라 계산 속도가 중요하고, 또한 공간적인 범위를 네트워크로 확장 시 예측 모형의 계산 속도가 시스템 효율을 결정하는 주요 요인으로 작용할 것으로 판단하였기 때문이다.

Ⅲ. 모형 구축

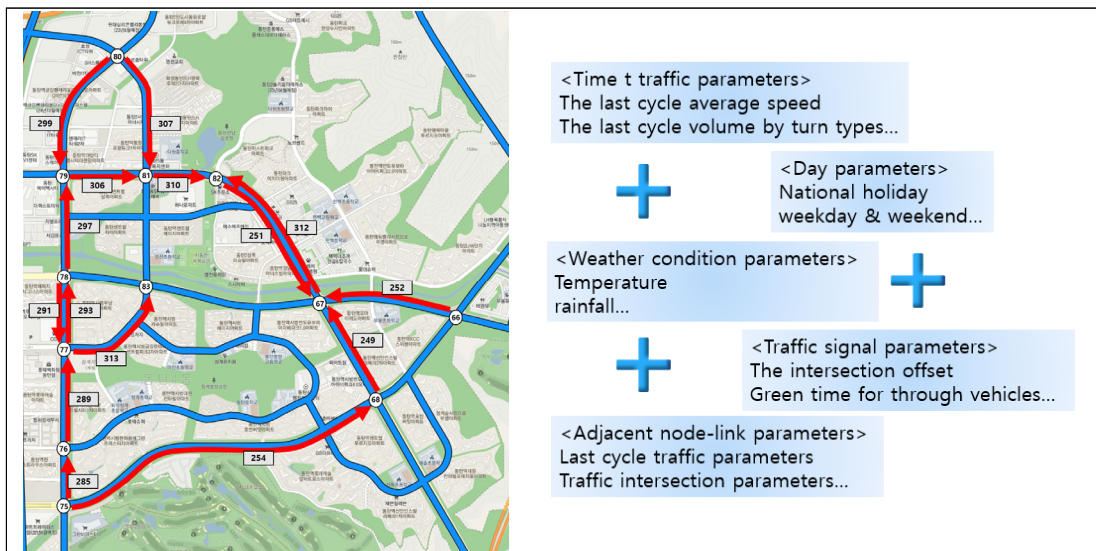
1. 데이터 수집 및 구축

분석에 활용된 수집 자료는 경기도 화성시로부터 제공된 스마트 교차로 데이터 및 ITS(VDS, DSRC) 기반

의 데이터로, 구체적인 공간적 범위는 경기도 화성시 동탄 2지구의 15개 링크이다. 시간적 범위는 2022/05/01 부터 2022/07/31까지 3개월이다. 스마트 교차로 데이터의 경우 5분 단위의 방향별 교통량, 통행속도, 점유율, 대기길이, 돌발상황 데이터로 구성되어 있다. 그러나 이는 최근 신규 설치되어 관심 영역(ROI, Region of Interest) 세팅이 불완전하기에 점유율 및 대기길이 데이터에서 결측이 다수 발생하여 본 연구에서는 포함하지 못하였다.

링크 통행속도 데이터 또한 ROI 세팅 문제로 전 시간대에서 30km/h로 표기되어, 화성시에서 수집되는 ITS(VDS, DSRC) 기반 링크 속도 데이터를 활용하였다. 또한 검지 시스템의 특성상 차량이 지나가지 않는 시간대는 생략되어 기록되지 않는 현상이 발생하는 것을 확인하였다. 시계열 데이터 분석에서는 완전한 시간 시퀀스를 요구하고 있어 이에 대한 결측치 보정을 진행하였다. 교통량의 경우 결측치를 0으로 대체하였으며, 이때의 통행속도는 자유속도로 가정하여 해당도로의 제한 속도 값인 50km/h를 적용하였다.

돌발상황 데이터의 경우 본 연구에서 수집된 자료 기간에는 미발생하였다. 그리고 전·후방 교통류 변동성을 고려하기 위해 상·하류 링크의 교통류 데이터를 결합하였으며, 신호제어에 의한 영향은 각 교차로의 직진 현시 기준의 녹색 시간 및 offset을 결합하여 고려하였다. 이외에도 기타 외부요인에 의한 운전자 통행 행태 패턴을 고려하고자 요일 변수와 공휴일 변수는 더미변수로 변환하여 결합하였으며, 해당 지역 인근의 기온, 습도, 풍속 등의 기상 변수를 결합하였다. 본 연구는 목적변수 데이터로 5분 이후의 링크 통행속도 데이터를 선정하였으며, 다양한 설명변수 데이터와 결합하여 최종 Data Set을 구축하였다. 아래 <Fig. 2>에서는 본 연구의 분석 대상 네트워크와 데이터 병합 과정을 제시하였다.

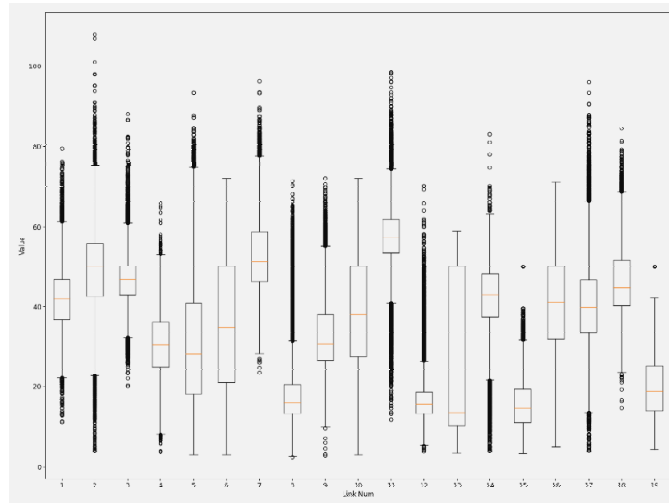


<Fig. 2> Road network and input data features

2. 모형 구축

모형 구축에 앞서 각 링크 별 통행속도 분포를 시각화해본 결과, <Fig. 3>과 같이 상이하게 나타나고 있어 링크 별로 모형을 학습할 필요가 있음을 확인하였다. 각 링크의 데이터는 모형 학습을 위해 Training Data(80%), Test Data(20%)로 구성하고, Training Data 내에서 다시 Training Data(80%), Validation Data(20%)로

구분하였다.



<Fig. 3> Different range of the traffic speed data per road section

또한, 모든 특성 변수를 모형에 적용하면 데이터 과적합 문제와 모형 계산 성능 문제가 발생하게 될 우려가 있을 것으로 판단하였다. 따라서 Keras의 Scikit-learn 모듈과 Random Forest 모형 구축을 통해 종속변수인 예측 통행속도 데이터(\hat{y}_{t+1})와 Input Data 간의 PFI를 추출하였다. 이는 특정 변수의 데이터를 무작위로 섞은 Data Set을 모형에 학습시켜 Loss 값의 변화를 통해 변수의 중요도를 파악하는 것으로, 변수 간 상관관계가 높을 경우 중요도 추출의 정확도가 하락한다. Training Data를 대상으로 변수 간 상관관계를 분석한 결과, 직진 교통량과 상·하류 직진 교통량 간 상관관계가 0.5 이상으로 나타났으며, Offset과 상·하류 Offset 간 상관관계 역시 0.5 이상으로 나타났다. 직진 교통량 변수와 Offset 변수를 제거하고 상관관계를 다시 분석한 결과 변수 간의 상관관계가 모두 0.5 이하로 도출되었다.

이후, Python의 GridSearchCV 모듈을 통해 Random Forest 함수의 최적의 Hyper Parameter를 도출하였다. 도출된 Hyper Parameter는 Max depth=16, Minimum Samples Leaf= 8, Minimum Samples Split=48, Random State=0, N_Estimators=300이다. 해당 파라미터로 Random Forest 모델을 구축하고, 해당 모델을 통해 PFI를 도출하였다. 26개의 설명변수 중 PFI값의 비율이 높은 순서대로 누적하였을 때 99% 이내에 해당하는 변수만을 추출하였고, 총 15개의 변수가 선별되었다. 15개 변수는 순서대로 t 시간의 통행속도, 상류 링크 통행속도, 기온, 습도, 풍속, 상류 링크 직진 교통량, 하류 링크 통행속도, 하류 링크 직진 교통량, 좌회전 교통량, 우턴 교통량, 우회전 교통량, 상류 링크 Offset, 월요일, 직진 방향 녹색시간, 하류 링크 Offset이다, 이 중 t 시간 통행속도 1개 변수가 변수 중요도의 약 82%를 차지하고 있는 것으로 나타났다.

이후, 위의 15개 변수로 이루어진 Data Set을 활용하여 각 링크의 GRU 모형을 구축하고, 학습 능력 향상을 위해 0과 1 사이로 Data Scaling을 진행하였다. 이후 Inverse Scaling 과정을 거쳐 실제 통행속도 데이터와 의 평균 제곱근 오차(RMSE, Root Mean Square Error) 값을 도출하였다. RMSE는 극단적인 이상값에 민감하지 않아 예측 모형의 주된 평가지표로 활용되고 있으며, RMSE 값의 산출 과정은 식 (4)에 제시하였다.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - y_p)^2} \dots\dots\dots (4)$$

y_i = 실제 측정값
 y_p = 예측값
 N = 데이터의 수

변수 구성에 대한 검증을 수행하기 위해 변수를 모두 포함하는 Data Set, 추출된 15개 변수만을 포함하는 Data Set으로 구분하였다. 249번 링크 Training Data를 대상으로 GRU 1 Layer에 대한 RMSE를 도출하였다. 비교 결과 각각 0.042로 동일한 값을 보였으며, 계산 속도는 전 변수 데이터의 모형이 50.60초, 추출된 변수 데이터의 모형이 30.98초가 소요되어 추출된 변수 데이터를 적용한 모형이 더욱 효율적인 것으로 확인하였다.

GRU 모형의 Layer 구조 및 Hyper Parameter 선정은 GridSearchCV 모듈을 통해 구조별로 Validation Data의 RMSE를 분석하여 선정하였다. 최종 선정된 모형 구조와 Hyper Parameter는 <Table 1>에서 제시하였다. 본 연구 모형의 예측 성능을 비교하기 위한 모형은 Baseline Model과 전통적인 시계열 모형인 이중지수평활 모형을 앞에 제시된 식에 따라 구축하였다.

<Table 1> GRU Model Structure

[Model Structures]			[Hyper Parameters]
Model: "sequential_1633"			
Layer (type)	Output Shape	Param #	
gru_5467 (GRU)	(None, 1, 30)	4230	<input type="checkbox"/> Batch Size : 32
dropout_3845 (Dropout)	(None, 1, 30)	0	<input type="checkbox"/> Epoch : 100
gru_5468 (GRU)	(None, 1, 45)	10395	<input type="checkbox"/> Optimizer : Adam
dropout_3846 (Dropout)	(None, 1, 45)	0	<input type="checkbox"/> Learning Rate : 0.01
gru_5469 (GRU)	(None, 1, 15)	2790	<input type="checkbox"/> Activation Function : Leaky ReLU(Alpha=0.01)
flatten_1629 (Flatten)	(None, 15)	0	<input type="checkbox"/> Dropout : 0.1
dense_1629 (Dense)	(None, 1)	16	<input type="checkbox"/> Early Stopping : Patience 20
-----			<input type="checkbox"/> Loss Function : MSE
Total params: 17,431			
Trainable params: 17,431			
Non-trainable params: 0			

IV. 결과 분석

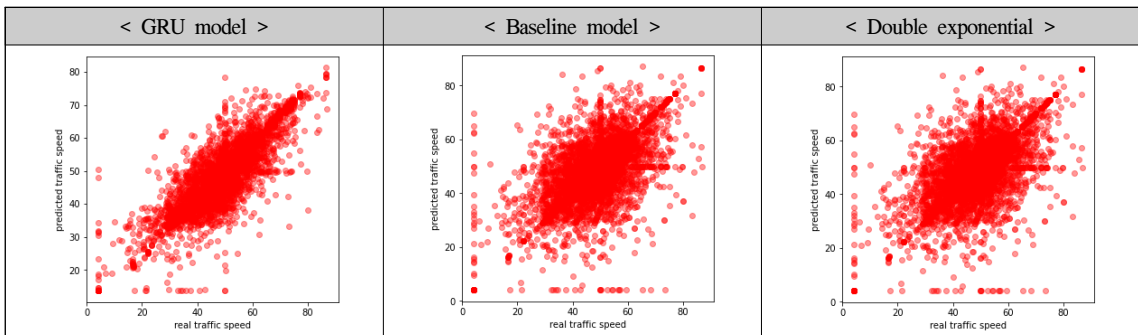
Test Data에 대한 RMSE를 통해 GRU 모형과 비교 모형들을 평가한 결과는 <Table 2>와 같다. 제시된 바와 같이 GRU 모형의 RMSE 값이 14개 링크에서 비교 모형보다 낮게 도출되어 예측 성능이 우수한 것으로 나타났다. 다만, 293번 링크에서는 이중지수평활 모형이 가장 우수한 것으로 나타났으며 Baseline 모형도 GRU 모형보다 낮은 오차값으로 나타났다. 모든 링크 데이터 예측치에 대한 RMSE 값을 평균한 결과, GRU, Baseline

모형, 이중지수평활 모형의 RMSE 평균이 각각 5.08, 7.46, 5.94로 도출되어 전반적으로 GRU 모형이 비교 모형보다 예측력이 개선된 것으로 나타났다. Baseline 모형의 RMSE 대비 GRU 모형의 RMSE 감소율을 산정해 본 결과, Baseline 모형보다 약 32% 감소한 것으로 나타났으며 이중지수평활 모형의 RMSE 대비 GRU 모형의 RMSE 감소율은 약 15%로 나타났다.

<Table 2> Performance Comparison of Different Models

Link ID	GRU	Baseline Model	Double Exponential Smoothing Model
249	4.00	6.23	4.93
251	6.31	10.67	8.09
252	4.11	6.34	5.19
254	4.56	7.63	5.91
285	7.03	9.85	7.59
289	5.28	7.42	6.43
291	4.91	7.76	6.09
293	8.34	7.29	5.79
297	4.95	6.62	5.80
299	3.70	5.57	4.41
306	5.79	9.72	7.51
307	3.57	5.62	4.43
310	5.60	9.33	7.20
312	4.38	6.09	5.33
313	3.60	5.77	4.47
Average RMSE	5.08	7.46	5.94

또한, 각 링크별로 실제 통행속도 값과 예측값 간의 일치하는 정도를 산점도를 통해 시각화하였다. 산점도에서는 x축값이 실제값, y축값이 예측값으로 이루어져 있다. 따라서, 실제 값과 예측치가 정확하게 일치할수록 산점도가 $y=x$ 함수 그래프에 집중되어 나타나게 된다. 이러한 결과로써 다음 <Fig. 4>는 251번 링크에 대한 산점도를 나타내었고, GRU 모형의 예측치가 $y=x$ 함수에 더욱 밀집되어 나타남을 확인하였다. 한편, Baseline model은 이중지수평활 모형의 결과보다도 더욱 분산된 산점도를 보이는 것으로 판단된다. 나머지



<Fig. 4> Scatter plots of different models

대상 링크에서도 대부분의 경우 GRU 모형의 산점도가 밀집된 결과로 나타남을 확인하였다.

V. 결 론

본 연구는 도시부 도로의 다양한 자료를 수집하여 통행속도 변화에 대한 영향을 분석하였고, 수집된 빅데이터 자료를 이용하여 GRU 기반의 통행속도 예측 모형을 개발하였다. 자료는 스마트 교차로가 설치된 도심부 도로에서 신호 운영 변수, 교통류 변수, 상·하류 인접 교통류 변수와 기상 변수 등을 결합하여 빅데이터 자료를 구축하였으며, PFI를 통한 데이터 선별 과정을 수행하였다. 그리고 Baseline 모형과 이중지수평활 모형을 비교 모형으로 선정하여 RMSE 지표로 예측 오차를 평가하였다.

PFI 분석 결과, 총 26개의 설명변수 중 t 시간의 통행속도, 상류 링크 통행속도, 기온, 습도, 풍속, 상류 링크 직진 교통량, 하류 링크 통행속도, 하류 링크 직진 교통량, 좌회전 교통량, 우턴 교통량, 우회전 교통량, 상류 링크 Offset, 월요일, 직진 방향 녹색시간, 하류 링크 Offset의 총 15개 변수가 선정되었다. 이 중 t 시간 통행속도 변수가 변수 중요도의 약 82%를 차지하였다.

모형 평가 결과, Baseline 모형의 RMSE는 평균 7.46, 이중지수평활 모형의 RMSE는 평균 5.94 값이 산출되었다. 그리고 GRU 모형으로 예측한 평균 RMSE는 5.08 값이 산출되었다. 15개 링크별로 편차가 존재하나, 14개 링크에서 GRU 모형의 오차가 최소의 값을 나타내었고, 추가적인 산점도 분석 결과도 동일한 결과를 제시하였다. 이러한 결과로부터 도시부 도로의 통행속도 정보 생성 과정에서 GRU 기반의 예측 모형 적용 시 예측 오차를 감소시키고 모형 적용 속도의 개선을 기대할 수 있을 것으로 판단된다.

본 연구에서는 국내에 구축된 스마트 교차로에서 자료를 수집하여 활용하고자 하였으나 자료의 일부 항목의 경우 많은 결측치가 발생하여 연구에 적용하지 못한 변수들이 있다. 특히 도시부의 경우 대기질이나 사고 자료는 통행속도에 큰 영향을 미치는 변수이지만, 자료의 한계로 인하여 본 연구에는 적용하지 못하였다. 향후 보다 다양한 자료가 수집되는 환경에서 본 연구에서 개발된 모형을 적용한다면 예측오차를 더욱 감소시킬 수 있을 것으로 판단된다. 또한 본 연구에서는 GRU 기반 모형으로 5분 이후의 단기예측만 수행하였으나, 향후 장기적인 예측에 관한 연구가 진행되면 모형의 범용성이 크게 높아질 수 있을 것으로 생각한다.

ACKNOWLEDGEMENTS

이 논문은 2022년도 정부(경찰청)의 재원으로 과학치안진흥센터의 지원을 받아 수행된 연구임.
(No.092021C28S01000, 자율주행 혼재 시 도로교통 통합관제시스템 및 운영기술 개발)

REFERENCES

- Abduljabbar, R. L., Dia, H., Tsai, P. W. and Liyanage, S.(2021), “Short-term Traffic Forecasting: An LSTM Network for Spatial-temporal Speed Prediction”, *Future Transportation*, vol. 1, no. 1, pp.21-37.
- Brown, R. G.(1959), “Statistical Forecasting for Inventory Control”, *Journal of the Royal Statistical*

Society. Series A, vol. 123, no. 3, p.232.

- Gao, Y., Zhou, C., Rong, J., Wang, Y. and Liu, S.(2022), “Short-Term Traffic Speed Forecasting Using a Deep Learning Method Based on Multitemporal Traffic Flow Volume”, *IEEE Access*, vol. 10, pp.82384-82395.
- Jeon, H.(2018), *A Deep-learning Approach to Predict Short-term Traffic Speeds Considering City-wide Spatio-temporal Correlations*, Master's Thesis, Jung-ang University.
- Jeong, M., Lee, T., Jeon, S. and Youm, M.(2021), “Highway Speed Prediction using Gated Recurrent Unit Neural Networks”, *Applied Sciences*, vol. 11, no. 7, p.3059.
- Kim, Y., Kim, J., Han, Y., Kim, J. and Hwang, J.(2020), “Development of Traffic Speed Prediction Model Reflecting Spatio-temporal Impact based on Deep Neural Network”, *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 19, no. 2, pp.1-16.
- Lee, S., Oh, D., Seo, J., Yoon, J., Choi, C. and Yoon, S.(2014), *Establishment of Intelligent Transport Systems(ITS)*, Ministry of Strategy and Finance, 11-1051000-000460-01.
- Mahmoud, N., Abdel-Aty, M., Cai, Q. and Yuan, J.(2021), “Predicting Cycle-level Traffic Movements at Signalized Intersections using Machine Learning Models”, *Transportation Research Part C: Emerging Technologies*, vol. 124, 102930.
- Mussumeci, E. and Coelho, F. C.(2020), “Machine-learning Forecasting for Dengue Epidemics-Comparing LSTM, Random Forest and Lasso Regression”, *MedRxiv*, pp.1-19.
- Peng, H., Bobade, S. U., Cotterell, M. E. and Miller, J. A.(2018), “Forecasting Traffic Flow: Short Term, Long Term, and When It Rains”, *International Conference on Big Data*, pp.57-71.
- Xu, C., Zhang, A., Xu, C. and Chen, Y.(2022), “Traffic Speed Prediction: Spatiotemporal Convolution Network based on Long-term, Short-term and Spatial features”, *Applied Intelligence*, vol. 52, no. 2, pp.2224-2242.
- Zhang, A., Liu, Q. and Zhang, T.(2022), “Spatial - temporal Attention Fusion for Traffic Speed Prediction”, *Soft Computing*, vol. 26, no. 2, pp.695-707.