자율주행을 위한 딥러닝 기반 다변량 회귀 기반의 저복잡도 다중 시계열 데이터 예측 모델

강승우, *조오현

충북대학교, 컴퓨터과학 전공

SWKang@chungbuk.ac.kr, *ohyunjo@chungbuk.ac.kr

Low Complex Multi-Spatial Time Series Data Prediction Model Using Deep Learning-based Multivariate Regression for Autonomous Vehicle

Seungwoo Kang, *Ohyun Jo

Chungbuk National University, Department of Computer Science

요 약

향후 미래 교통량 예측은 스마트 시티의 자율주행 차량 운행과 지능형 교통 서비스에 유용하게 활용될 수 있다. 이를 위해 여러 위치에서의 실시간 시계열 데이터 동시 처리가 필요하며 정확도를 보장하면서 빠르고 효율적인 예측을 해야 한다. 본 논문에서 제안하는 딥러닝 기반 다중 시계열 데이터 예측 모델은 여러 위치의 실시간 데이터를 이미지화하여 저복잡도를 갖는 CNN(Convolution Neural Network)에 적용함으로써 다수의 시계열 종속변수를 동시 적용함으로 복수 개의 미래 구간의 교통량 정보를 회귀 분석한다. 제안하는 모델은 예측하고자 하는 시계열 교통량 데이터를 시간 정보와 함께 이미지화시켜 학습하는 것이 특징이다. 본 연구에서는 1시간 단위로 6시간 이후까지 총 6구간의 평균 교통량을 예측하며 각 구간을 단일의 종속변수로 가진 단변량 회귀 모델들과 비교한다. 제안하는 모델은 단변량 회귀 모델과 비교했을 때 정확도를 유지하면서 학습 속도 면에서는 4.27배 향상된 성능을 나타낸다.

Ⅰ. 서 론

미래의 교통량을 정확하게 예측할 수 있다면 자율 주행 차량과 스마트 시티의 자동화 도시에서 적용되는 교통 서비스의 설계와 관리를 최적화할 수 있다. 또한 운전자의 이동시간 단축, 오염 물질 배출 감소, 운전자 스트레스 감소 등 여러 이점을 불러올 수 있다[1][2]. 하지만 이런 실시간 서비스의 목적에 부합하기 위해서는 교통량 예측에 있어서 정확함을 유지해야 할 뿐 아니라 실시간 서비스 제공을 위해 빠르고 효율적이어야 한다.

CNN(Convolution Neural Network)은 이미지 데이터의 특징을 추출하여 학습하기 때문에 적은 매개변수로 학습할 수 있으므로 학습에 있어서 저복잡성을 띄고 그 성능의 우수성은 입증되었다[3]. 그리고 다수의 종속변수가 있을 때는 각각에 대해 따로 회귀 분석하는 것이 일반적이지만, 종속 변수들 간 상관관계가 있다면 다수의 종속변수를 동시에 분석하는 다변량 회귀 분석을 적용하는 것이 더 효율적이다[4]. 본 논문에서는 교통량 예측에 있어서 효율성을 높이고 복잡도를 개선하기 위해 여러 시계열 데이터들을 CNN을 활용하여 동시에 다변량 다중 회귀 분석할 수 있는 예측 모델을 고안하고 그 성능을 평가한다.

Ⅱ. 본론

1. 데이터 전처리

본 논문에서 제안하는 모델은 실험 데이터로 공공 데이터인 서울특별시 고속도로 일별, 시간대별 교통소통 정보가 사용되었다. 이 데이터셋은 서울특별시 내에 존재하는 모든 고속도로 구간들의 1시간 단위 교통 속도 누적 데이터를 가진다. 이 중 실험에서 사용된 데이터는 2017년 1월 1일부터 2018년 7월 9일까지 총 555일간, 160개의 도로구간의 데이터다. 또한 데이터 특성을 보조하기 위해 해당 데이터들의 0부터 23까지의 시간 정보(Hour)를 사용한다. 도로 구간의 데이터를 분석하였을 때 교통 체증이 심하고 교통량 변화가 잦다고 판단된 강변 북로의 구간 중 하나를 예측 대상으로 선택하였다.

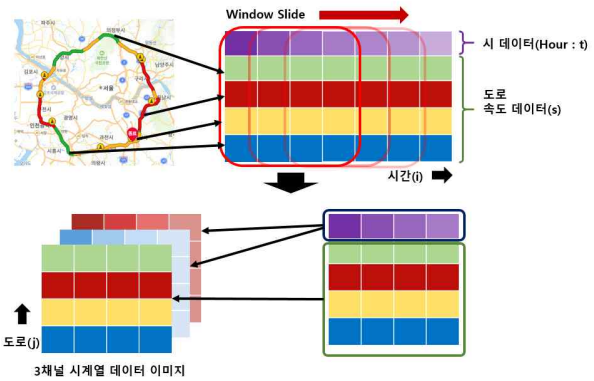

그림 1. 학습을 위한 여러 시계열 데이터의 이미지화 전처리 과정

학습을 위한 이미지는 윈도우 슬라이드(Window Silde)기법으로 그림 1처럼 시계열 구간을 하나씩 옮겨가며 생성하였다. 그림 1의 가로축과 세로축은 각각 시간(i)과 도로(j)의 인덱스를 의미한다. 이미지화 된 데이터는 사용될 도로에 대해 예측시점(C) 기준으로 윈도우 크기(W)만큼 이전의 데이터를 사용한다. 그리고 이 이미지는 3개의 채널로 구성된다. 하나는 속도 값(s_{ij})을 나타내고 나머지 두 개는 해당 시간 정보(t_{ij} : Hour)를

나타낸다. 속도의 경우 고속도로의 최대속도와 가까운 120을 나눈 값을 이용한다. 시간 정보(Hour)의 경우는 0부터 23까지의 값을 연속적으로 나타내기 위해 각으로 표현하고, sin과 cos 값을 각각의 채널에 사용했다. 이 데이터로 대상 도로구간($Target$)에 대해 예측 구간 크기(P)만큼 미래 교통량을 예상한다. 사용될 도로의 집합을 R이라고 했을 때, 학습 데이터 1개의 입력 데이터(I_C)와 다중 출력 데이터(O_C)는 각각 식 (1), 식 (2)과 같이 나타낸다.

$$I_C = \left\{ \left(\frac{s_{ij}}{120}, \sin\frac{\pi}{12}t_{ij}, \cos\frac{\pi}{12}t_{ij} \right) | C - W < i \leq C, j \in R \right\} \quad (1)$$

$$O_C = \{ s_{ij} | j = Target, C < i \leq C + P \} \quad (2)$$

위의 다변량 회귀와 단변량 회귀의 성능을 비교하기 위한 단일 출력 데이터(O'_{kC})가 필요하다. 그래서 같은 입력 데이터에 대해 예측 구간을 1시간 단위로 나누고, 각 구간(k : $C+1, C+2, \ldots, C+P$)의 단일 출력 데이터를 만들었다. 모든 전처리 데이터를 테스트 데이터 셋과 훈련 데이터 셋을 시간 별로 고르게 2 : 8로 분할하여 학습을 수행하였다.

2. 학습 모델

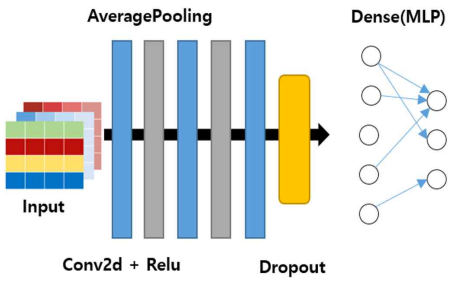

그림 2. CNN 학습 모델 구성도

그림 2는 제안하는 모델의 구성을 나타낸 것이다. CNN 학습 구성에는 3개의 컨볼루전 계층이 있고, 그 사이에 이미지 특징을 잡는 평균 Pooling 계층이 있다. 과적합을 피하기 위한 Drop-out 계층을 지나고 다중 레이블 퍼셉트론(Multi-Layer Perceptron)에 전달하여 학습을 수행한다. 마지막 Dense 계층은 노드 개수를 출력 데이터 개수에 맞춰 사용하고, 단일 출력 데이터의 경우 1개의 노드를 사용한다.

3. 성능 평가

다변량 회귀 성능 평가를 위한 실험에서 예측 구간 크기(P)를 6으로 고정하였다. 다변량 회귀 1번과 각각의 구간에 대한 6번의 단변량 회귀 포함해 총 7번의 실험을 진행했다. 회귀 정확도 지표는 예측에 사용된 N개의 테스트 데이터 중, 실제 데이터(T_l : $1 \leq l \leq N$)의 오차 범위 ±10km/h 이내에 있는 예측 값(F_l) 개수의 비율을 사용하였다. 정확도는 식 (3)으로 나타내며, 다변량 회귀 정확도는 전체 예측 구간에 대해 73.2%였다.

$$\frac{n(\{F_l | 1 \leq l \leq N, |F_l - T_l| < 10\})}{N} \quad (3)$$

그림 3은 각각의 예측 구간에 대해 다변량 회귀의 정확도를 구한 후 단변량 회귀 정확도와 비교한 그래프다. 서로 정확도가 비슷하며 둘 중 더 나은 것을 판단하기 힘든 지표이다. 반면 표 1은 학습 속도를 비교하며 두 회귀 모델의 성능 차이를 나타낸다. 단변량 회귀의 경우 6구간 전부 예측을 하면 약 2618초가 걸리지만, 다변량 회귀의 경우 612초로 측정되어 학습 속도가 4.27배 개선됨을 확인하였다. 결론적으로, 다변량 회귀 모델과 단변량 회귀 모델의 경우 정확도 성능에서는 차이가 없지만 학습시간 효율은 단변량 회귀 모델보다 제안하는 다변량 회귀 모델 쪽에서 더 높게 나타난다.

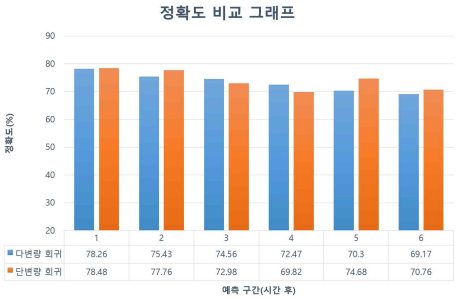

그림 3. 다변량 회귀와 단변량 회귀의 정확도 비교 그래프

학습모델 종류	1시간 후 예측모델의 경우	2시간 후 예측모델의 경우	3시간 후 예측모델의 경우	4시간 후 예측모델의 경우	5시간 후 예측모델의 경우	6시간 후 예측모델의 경우	제안하는 다변량 회귀 모델의 경우
학습시간(초)	440.03	434.29	447.69	440.06	413.04	443.30	612.47
	2618.41						

표 1. 다변량 회귀와 단변량 회귀의 학습 속도 비교표

Ⅲ. 결론

본 논문에서는 이미지화 시킨 시간 정보 데이터와 시계열 교통량 데이터를 CNN모델을 통해 한 번에 넓은 구간의 미래 교통량 예측이 가능한 다변량 다중 회귀 모델을 제안한다. 같은 모델 사용 시, 단변량 회귀를 사용했을 때의 정확도 차이가 거의 없고 학습시간 단축 면에서 상대적으로 높은 성능을 나타냈다. 제안하는 모델은 저복잡도를 갖는 CNN모델과 더불어 정확도는 유지되고 학습 속도는 4.27배로 대폭 향상되었다.

Acknowledgement

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2020R1A6A1A12047945). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 Grand ICT연구센터지원사업의 연구결과로 수행되었음(IITP-2020-0-01462). 또한, 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00165, 5G+ 지능형 기지국 소프트웨어 모뎀 개발)

참 고 문 헌

[1] Stephen Clark," Traffic Prediction Using Multivariate Nonparametric Regression," Journal of Transportation Engineering, volume 129 issue 2, Mar. 2003.

[2] Attila M.Nagy, VilmosSimon "Survey on traffic prediction in smart cities," Pervasive and Mobile Computing, volume 50, pp. 148-163, Oct. 2018.

[3] Jung Hun Byun, Ohyun Jo "Low Complexity CNN Model by Using 2D Imaging od Time Series Traffic Data," Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 98-99, Aug. 2020.

[4] Joachim Hartung, Guido Knapp "Multivariate Multiple Regression", https://doi.org/10.1002/9781118445112.stat06583, Sep. 2014.