손실 데이터를 포함한 교통량 데이터의 분석을 위한 어텐션 메커니즘 및 CNN 기반의 패턴 분류 모델*

이은진°, 오소연, 이민수
이화여자대학교 컴퓨터공학과
leee5495@ewhain.net, soyeon.oh@ewhain.net, mlee@ewha.ac.kr

Pattern Classification Model based on Attention Mechanism and CNN for Analysis of Traffic Data including Missing Values

Eunjin Lee°, Soyeon Oh, Minsoo Lee
Dept. of Computer Science and Engineering, Ewha Womans University
leee5495@ewhain.net, soyeon.oh@ewhain.net, mlee@ewha.ac.kr

요 약

교통량 데이터의 분석은 도시 계획, 교통 공학, 다양한 교통 및 위치 기반 서비스의 구현 등에 다양하게 활용될 수 있다. 그러나 교통량 데이터는 그 측정 과정에서, 센서와 서버 간의 통신 오류 또는 부정확한 측정 등으로 인한 손실 데이터를 포함하는 경우가 대부분이다. 이러한 손실 데이터는 교통량 예측이나 교통량 패턴 분류 등 다양한 교통량 데이터 분석 기법들의 성능 저하를 일으키는 원인이 된다. 본 논문에서는 이러한 손실 데이터를 포함한 교통량 데이터 분석에 활용될 수 있는 패턴 분류 모델을 설계 및 제안한다. 제안하는 모델은 CNN(Convolutional Neural Networks) 기반의 패턴 분류 모델로, 어텐션 메커니즘 기반의 신경망(Attention Neural Networks)을 적용하여 손실 데이터에 대한 가중치를 부여함으로써 모델의 패턴 분류 성능을 향상 시킬 수 있도록 하였다.

1. 서 론

교통량 데이터는 도로를 비롯한 다양한 도시 환경 및 공간을 효과적으로 배치하는 도시 계획, 교통 상황을 분석하고 문제점을 도출 및 이에 대한 해결 방안을 찾는 교통 공학, 교통 체증 정보 또는 여행 경로 등에 기반한 다양한 형태의 교통 및 위치 기반 서비스 구현에 활용될 수 있다. 이러한 이유로 교통량을 측정할 수 있는 루프 센서들이 많은 도로에 설치되었고, 공공 데이터의 형태로 교통량 데이터를 제공하는 사례도 많아졌다.

그러나 교통량 데이터의 수집 과정에는 센서와 서버 간의 통신 오류 또는 부정확한 측정으로 인한 부분적인 데이터의 손실, 유의미한 분석을 위한 충분한 수의 데이터를 수집하기까지의 시간적 제약 등의 한계점들이 존재한다[1]. 따라서 손실된 값을 포함한, 제한된 수의 데이터 셋을 기반으로 한 패턴 분류 등의 교통량 데이터 분석 기법에 대한 연구는 매우 필수적이고 중요하다.

본 논문에서는 앞서 언급한 교통량 데이터 분석 문제가 지닌 한계점들을 개선하여, 교통량 데이터 분석에 활용될 수 있는 교통량 패턴 분류 모델을 설계 및 구현하였다. 구체적으로, 캘리포니아 교통량 정보를 제공하는 Caltrans PeMS[2]의 데이터 중 다저스 경기장과 가까운 거리에 설치된 센서의 교통량 측정 데이터 셋에 대하여, 다저스 경기장에서의 경기 개최일 여부와 주말/평일을 높은 정확도로 분류할 수 있는 딥러닝 기반의 모델을 설계하고 실제 수집된 교통량 데이터에 대해 실험한 뒤, 그 결과를 분석한다.

2. 관련 연구

본 논문에서 분석하고자 하는 교통량 데이터는 시계열 데이터의 한 종류이다. 기존에는 통계적 관점에서 가우시안 기반의 선형 모델을 활용하는 기법들이 시계열 데이터의 분석을 위해 활발히 연구되었다. 하지만 이러한 선형 모델 기반의 분석 기법들은 비선형적인 데이터 패턴 분류 문제에서는 제한적인 성능을 보인다는 한계점이 존재한다[3].

최근에는 딥러닝 기반의 분석 기법들이 시계열 데이터의 예측 및 분류를 위하여 활발히 연구 및 활용되고 있다[4]. RNN(Recurrent Neural Networks)은 순차적인 데이터에 대한 회귀적인 특징을 학습하여, 시계열 데이터의 미래 원소를 예측하는 문제에 주로 활용되고[5], 앞먹임 신경망과 CNN(Convolutional Neural Networks)은 주어진 시계열 데이터 패턴을 특정 클래스로 분류하는 문제에 주로 활용된다[6]. 이러한 딥러닝 기반의 분석 기법들을 통하여 기존의 선형 모델 기반의 분석 기법과는 달리, 비선형적인 분석 문제까지 해결할 수 있게 되어, 보다 정확한 시계열 데이터의 분석이 가능해졌다.

* 이 논문은 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1D1A1B03034691)

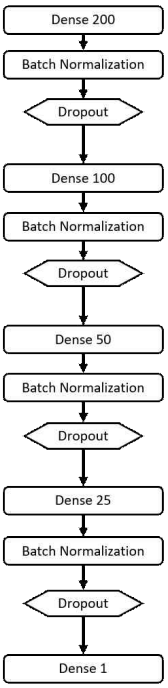
본 논문에서는 CNN과 어텐션 메커니즘 기반의 신경망(Attention Neural Networks)을 활용하여 교통량 데이터 분류 모델을 설계하였다. CNN은 커널을 통해 주변 데이터들과의 연관성을 학습할 수 있기 때문에, 이미지 데이터뿐만 아니라, 여러 채널로부터 수집되는 다변량 시계열 데이터 처리에 효과적임이 여러 연구들에 의해 증명되었다[7]. 어텐션 메커니즘 기반의 신경망은 주로 순차적인 입력 데이터에 대한 회귀 분석 문제에서, 입력 데이터에 가중치를 부여하기 위하여 활용된다. 이를 통하여 데이터에 대한 일반화를 위한 특징 학습을 용이하게 하여, 회귀 분석을 위한 RNN 기반 모델의 성능을 향상시키는 데에 큰 도움을 줄 수 있음이 알려져 있다[8, 9]. 또한, 조건부 드롭아웃(conditional dropout)을 활용하여 불필요한 데이터에 대해서는 연산을 수행하지 않음으로서, 모델의 성능을 낮추지 않으면서 학습 속도를 높일 수 있다[10]. 본 논문에서는 어텐션 메커니즘 기반의 신경망과 조건부 드롭아웃을 활용한 분류 모델을 설계 및 구현하였다.

3. 제안하는 패턴 분류 모델

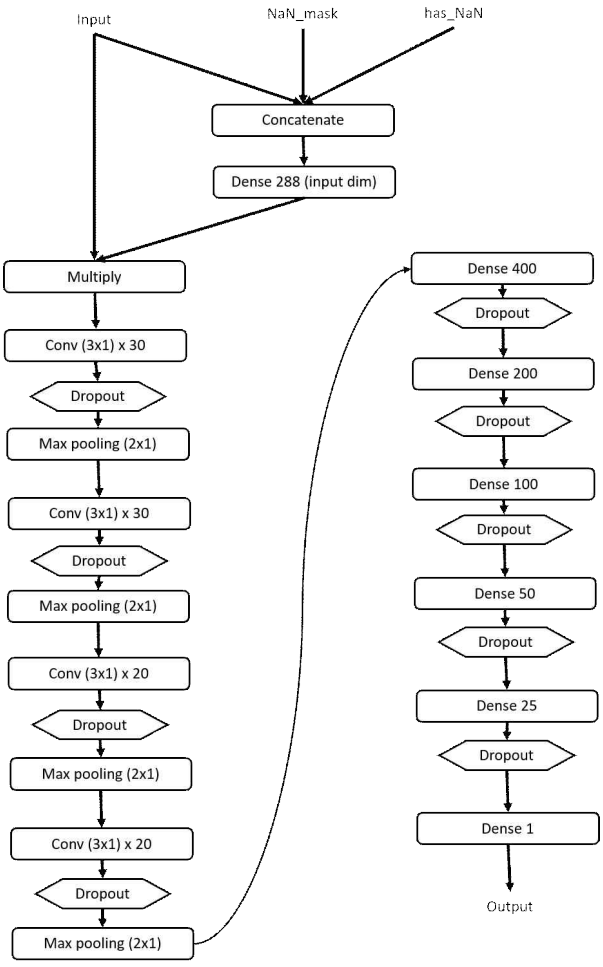
본 논문에서는 캘리포니아 주의 최대 도시인 LA의 교통량 데이터를 대상으로, 교통량 패턴에 기반 하여 ① 주말/평일을 구분하는 모델과 ② 다저스 경기장에서의 경기 개최일 여부를 구분하는 모델을 각각 설계 및 구현하였다. [그림 1]은 주말/평일 분류 모델의 구조를, [그림 2]는 경기 개최일 여부 분류 모델의 구조를 도식화 한 것이다.

대부분의 교통량 데이터에는 센서와 서버 간의 통신 오류 및 센서의 하드웨어적 결함으로 인하여 누락된 값들이 존재한다. 따라서 두 모델을 설계하기에 앞서, 손실 데이터 즉, NaN에 대한 전처리를 먼저 수행하였다. 본 논문에서는 ① 손실 데이터를 0으로 치환, ② 손실 데이터에 대하여, 각 데이터 패턴의 평균 교통량 값으로 치환, ③ 해당 손실 데이터의 왼쪽 값과 오른쪽 값을 이용하여 직선 보간법을 수행하여 값을 치환하는 세 가지 전처리 기법을 적용하여 보았다. 그 결과, ③ 해당 손실 데이터의 왼쪽 값과 오른쪽 값을 이용하여 직선 보간법을 수행하여 값을 치환하는 전처리 기법을 수행했을 때, 가장 높은 분류 정확도를 얻을 수 있음을 확인하였다.

주말/평일 분류 모델의 경우, 손실 데이터에 대한 전처리만 적절하게 수행하면 [그림 1] 에서와 같은 간단한 구조의 앞먹임 신경망 모델을 적용했을 때에도 높은 분류 정확도를 얻을 수 있었다. 이 앞먹임 신경망 모델의 경우, 과적합을 방지하기 위하여 각 레이어에 대해 0.2 비율의 드롭아웃(Dropout)을 적용하였고, 배치 정규화(Batch Normalization)를 수행했다[11, 12]. 반면, 경기 개최일 여부를 분류하는 모델의 경우, [그림 2]와 같이 CNN을 기반으로 하여 모델을 설계하였다. 교통량 데이터 수집을 위한 다저스 경기장 주변의 루프 센서가 경기 개최 여부에 따른 교통량 변화를 반영할 수 있을 만큼의 거리 반경 안에 설치되었다고 보기 어렵기 때문에, 경미한 교통량 차이에도 정확한 예측을 할 수 있는 모델을 설계하기 위해서이다[3].



[그림 1] 주말/평일 분류 모델



[그림 2] 다저스 경기장에서의 경기 개최일 여부 분류 모델

또한 경기 개최일 여부 분류 모델의 경우, CNN 모델을 활용한 분류를 수행하기에 앞서, 입력 데이터 패턴에 어텐션 메커니즘 기반의 신경망을 적용하였고, 그 결과로 얻어진 데이터를 CNN 모델의 입력 데이터 구성에 활용하였다. 다저스 경기장에서의 경기는 정해진 시간대에 개최되므로, 경기가 개최되는 시간대의 데이터 패턴에 더 높은 가중치를 부여하고, 전처리 된 손실 데이터에 대해서는 낮은 가중치를 부여함으로서 경기 개최일 분류 모델의 정확도를 높일 수 있다고 판단했기 때문이다.

먼저, [그림 2]의 상단에 도식화 된 것과 같이, ① 기존 입력 데이터, ② 기존 입력 데이터를 구성하는 값들이 각각 손실 데이터인지의 여부를 0과 1로 표현하는 벡터, ③ 기존 입력 데이터에 손실 데이터가 포함되어 있는지의 여부를 one-hot encoding으로 표현하는 벡터를 순차적으로 연결하여 하나의 벡터를 구성한다. 그런 뒤, 이렇게 구성된 벡터를 기존 입력 데이터의 차원과 같은 수의 뉴런으로 구성된 1개 계층 앞먹임 신경망에 입력 데이터로 전달한다. 그리고 앞먹임 신경망의 출력으로 얻어진 가중치 실수 벡터와 ① 기존 입력 데이터에 대해 곱셈을 수행한 결과 벡터를 경기 개최일 여부 분류를 위해 설계된 합성곱 신경망에 입력 데이터로 전달한다.

어텐션 메커니즘 기반 신경망의 학습 성능을 높이기 위하여, 추가적인 학습 데이터 셋을 구성하였고 이를 학습 과정에 활용하였다. 추가적인 학습 데이터 셋은 다음과 같이 구성하였다. 먼저, 학습 데이터 셋에서 무작위로 선정한 부분 데이터 셋을 구한다. 이 부분 데이터 셋의 각 데이터에 대하여, 무작위로 선택한 원소들을 손실 데이터 즉, NaN으로 치환하고 직선 보간법 기반의 전처리를 수행한다. 이러한 과정을 거쳐 구성된 추가적인 학습 데이터 셋을 기존 학습 데이터 셋에 더하여 어텐션 메커니즘 기반의 신경망 학습에 활용한다. 이는 교통량 분석을 실제로 수행하는 경우, 분석을 위한 교통량 데이터 수집 기간이 짧아서 제한된 크기의 학습 데이터 셋을 구성하게 되더라도, 전체 성능 향상에 많은 영향을 미치는 어텐션 메커니즘 기반 신경망을 효과적으로 훈련할 수 있도록 하기 위함이다. 실험을 수행한 결과, 이렇게 추가적인 학습 데이터 셋을 구성하여 어텐션 메커니즘 기반 신경망을 훈련시킨 경우, 경기 개최일 여부 분류 모델의 정확도가 향상되는 것을 확인할 수 있었다.

4. 실험 결과 및 분석

본 논문에서는 하루 288개 값으로 측정된 LA의 날짜별 교통량 데이터와 다저스 경기장에서의 경기 개최 정보로 구성된 부가적인 데이터를 결합하여 (1×288) 차원의 레이블링 된 하루 동안의 교통량 패턴 데이터를 구성하여 제안하는 두 모델의 학습 및 평가에 활용하였다. 총 175일의 데이터가 사용되었고 모든 데이터는 다저스 경기장과 인접해 있는 고속도로의 센서에서 측정되었다. 그중 주말의 비율은 28.57%, 주중의 비율은 71.43%이고, 경기가 있는 날은 44.57%, 경기가 없는 날은 55.43%의 비율로 나누어진다.

데이터 셋 구분	학습	검증	테스트	전체
전체 데이터 셋에서의 비율	0.6	0.2	0.2	1
정확도 (%)	100	100	100	100

[표 1] 주말/평일 분류 모델의 정확도

데이터 셋 구분	학습	검증	테스트	전체
전체 데이터 셋에서의 비율	0.65	0.15	0.2	1
정확도 (%)	100	100	97.06	99.43

[표 2] 경기 개최일 여부 분류 모델의 정확도

	어텐션 메커니즘 적용한 모델			어텐션 메커니즘 적용하지 않은 모델		
	경기 개최일	개최일 아님	전체	경기 개최일	개최일 아님	전체
바르게 분류된 개수	77	97	174	74	94	168
잘못 분류된 개수	1	0	1	4	3	7
합계	78	97	175	78	97	175
정확도 (%)	98.72	100	99.43	94.87	96.91	96.00

[표 3] 어텐션 메커니즘의 적용 여부에 따른 경기 개최일 여부 분류 모델의 정확도 비교

	추가적인 학습 데이터 셋을 학습에 사용			기존 학습 데이터 셋만을 학습에 사용		
	경기 개최일	개최일 아님	전체	경기 개최일	개최일 아님	전체
바르게 분류된 개수	77	97	174	76	93	169
잘못 분류된 개수	1	0	1	2	4	6
합계	78	97	175	78	97	175
정확도 (%)	98.72	100	99.43	97.44	95.88	96.57

[표 4] 어텐션 메커니즘 기반의 신경망 학습 과정에 추가적인 학습 데이터 셋 사용 여부에 따른 경기 개최일 여부 분류 모델의 정확도 비교

먼저, 주말/평일 분류 모델과 경기 개최일 여부 분류 모델 각각에 대하여 정확도를 측정해 보았다. [표 1]은 주말/평일 분류 모델의 정확도 측정 결과를 요약한 것이다. 이 경우, 전체 데이터 셋의 60%를 학습 데이터 셋으로, 20%를 검증 데이터 셋, 나머지 20%를 테스트 데이터 셋으로 활용하였다. 모델 학습에 있어서는, [그림 1]과 같

이 구성된 앞먹임 신경망에 대하여 에폭(epoch) 100, 배치(batch) 크기 20, 드롭아웃 비율(dropout rate) 0.2, 학습률(learning rate)은 0.001로 설정하여 Adam 최적화 알고리즘으로 학습을 수행하였다. 그 결과, 학습/검증/테스트 데이터 셋 모두에서 100%의 분류 정확도를 보였다[13].

[표 2]는 경기 개최일 여부 분류 모델의 정확도 측정 결과를 요약한 것이다. 이 경우, 전체 데이터 셋의 65%를 학습 데이터 셋으로, 15%를 검증 데이터 셋, 나머지 20%를 테스트 데이터 셋으로 활용하였다. 모델 학습에 있어서는, [그림 2]와 같이 구성된 모델에 대하여 에폭 100, 배치 크기 20, 드롭아웃 비율 0.2, 학습률 0.001로 Adam 최적화 알고리즘[13]을 사용하여 학습을 수행하였고 그 결과, 학습 데이터 셋에서 100%, 검증 데이터 셋에서 100%, 테스트 데이터 셋에서 97.06%의 분류 정확도를 보였다.

[표 3]은 경기 개최일 여부 분류 모델에 있어서, 손실 데이터에 대한 가중치를 부여하는 어텐션 메커니즘 기반의 신경망을 적용한 경우와 그렇지 않은 경우 각각에 대한 분류 정확도를 비교한 실험 결과이다. 그 결과, 어텐션 메커니즘 기반의 신경망을 적용한 경우에는 전체 데이터 셋의 분류 예측에 있어서 99.43%의 정확도를 보였다. 반면, 어텐션 메커니즘 기반의 신경망을 적용하지 않은 경우에는 전자의 학습 방식을 따라 추가된 학습 데이터 셋을 이용해 훈련을 시켰음에도 불구하고 전체 실데이터 셋의 분류 예측에 있어서 96.00%의 정확도를 보였다. 따라서 어텐션 메커니즘 기반의 신경망을 통한 손실 데이터 및 특정 패턴에 대한 가중치 부여가 경기 개최일 여부 분류 모델의 정확도 향상에 기여함을 알 수 있다.

[표 4]는 제한된 크기의 학습 데이터 셋을 활용한 어텐션 메커니즘 기반의 신경망 학습에 있어서, 3절에서 언급한 방식에 의하여 추가적인 학습 데이터 셋을 구성하고 이를 학습에 활용한 경우와 기존 학습 데이터 셋만을 학습에 활용한 경우 각각에 대해서 경기 개최일 여부 분류 모델의 분류 정확도를 비교한 실험 결과이다. 그 결과, 추가적인 학습 데이터 셋을 학습에 활용한 경우, 전체 데이터 셋에 대하여 99.43%의 분류 예측 정확도를 보였다. 반면, 추가적인 학습 데이터 셋을 활용하지 않고 기존 학습 데이터 셋만을 학습에 활용한 경우, 전체 데이터 셋에 대하여 96.57%의 분류 예측 정확도를 보이는 것에 그쳤다. 따라서, 제한된 크기의 학습 데이터 셋에 대해서, 어텐션 메커니즘 기반의 신경망을 분류에 활용하는 경우에는 3절에서 언급한 방식에 따라 추가적인 학습 데이터를 구성하여 학습에 활용하는 것이 모델의 정확도 향상에 도움을 준다고 볼 수 있다.

5. 결 론

본 논문은 데이터 수집 과정에서 발생하게 되는 손실 데이터를 포함한 교통량 데이터의 패턴 분류 모델을 설계하고 구현하였다. 이 과정에서 패턴 분류 성능에 영향을 미치는 손실 데이터를 처리하기 위한 세 가지 전처리 기법을 적용하여 보았다. 또한, 어텐션 메커니즘 기반의 신경망을 적용하여, 패턴 데이터의 특징에 영향을 미치는 부분 패턴 데이터와 손실 데이터에 대하여 가중치를 조정함으로서 분류 성능을 높일 수 있음을 확인하였다. 또한 제한된 크기의 데이터 셋으로 어텐션 메커니즘 기반의 신경망을 적용한 분류 모델의 정확도를 높일 수 있는 학습 기법을 제안하였다. 본 논문의 연구 결과는, 실제 센서 데이터 수집 과정에서 흔히 발생하는 손실 데이터를 고려하는 동시에, 주로 비선형적인 특성을 갖는 실세계 문제 해결을 위한 향후 연구들에 활용될 수 있을 것이라 생각한다.

참 고 문 헌

[1] Chen, C. (2003). Freeway performance measurement system (PeMS).

[2] Caltrans PeMS. (2019). Pems.dot.ca.gov. Retrieved 13 August 2019, from http://pems.dot.ca.gov/

[3] Fan, J., & Yao, Q. (2008). Nonlinear time series: nonparametric and parametric methods. Springer Science & Business Media. pp.10-15

[4] Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42, 11-24.

[5] Zhang, J. S., & Xiao, X. C. (2000). Predicting chaotic time series using recurrent neural network. Chinese Physics Letters, 17(2), 88.

[6] Gamboa, J. C. B. (2017). Deep learning for time-series analysis. arXiv preprint arXiv:1701.01887.

[7] Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015, June). Deep convolutional neural networks on multichannel time series for human activity recognition. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

[8] Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

[9] Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In Advances in neural information processing systems (pp. 2204-2212).

[10] Bengio, E., Bacon, P. L., Pineau, J., & Precup, D. (2015). Conditional computation in neural networks for faster models. arXiv preprint arXiv:1511.06297.

[11] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[12] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[13] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.