LSTM을 이용한 강남역 시간별 탑승인원 예측

영문제목

요 약

한국정보과학회는 정보과학에 관한 기술을 발전, 보급시키고 회원상호간의 친목을 도모하기 위하여 1973년 3월 3일에 설립되었으며, 정보통신부에 '사단법인 한국정보과학회'로 등록되었다. 학회의 주요 활동은 1) 컴퓨터 기술 및 이론에 관한 새로운 연구결과를 발표하는 기회를 제공하고, 2) 국내의 컴퓨터 관련 기술 개발에 참여 하며, 3) 국제적 학술 교류 및 협력 증진을 도모 하고, 4) 회원 상호간의 친목을 증진시키는 것이다.

1. 서 론

출퇴근 시간대 혹은 특정 호선들의 차량은 수용률이 150% 이상을 넘어가며, 현대인들이 대중교통 이용에 불편함을 겪고 있다. 광역 버스에는 남은 좌석수가 표시 되지만 지하철에는 도착 시간 정보 밖에 존재 하지 않는다. 본 연구는 지하철의 탑승인원을 예측 할 수 있는 방안을 찾고, 이용객들로 하여금 이동수단에 대한 최소한의 편의성을 보장하기 위하여 연구한다. 본 논문에서는 서울 특별시에서 제공받은 지하철 호선별 역별 승하차 인원 정보 데이터를 전처리 한뒤 시계열 데이터에 예측에 적합한 Long Short-Term Memory(LSTM) 모델을 사용해 강남역의 탑승 인원을 최대한 정확하게 예측하기위해 실험을 하였고, 만족할만한 정확도와 추후 연구 계획을 세웠다.

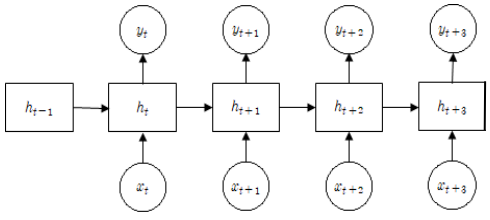
2. 데이터셋

서울시에서 운영하는 홈페이지인 정보소통광장 에서 제공하는 서울시 지하철호선별 역별 승하차 인원 정보를 데이터를 이용하였다. 사용한 데이터의 기간은 2008년 1월 1일부터 2017년 09월 30일 까지이며, 지하철 운행 시간대 별로 구간을 나누어 승차 및 하차 인원을 역별로 나누어 엑셀형태로 저장하였다. 본 연구에서 원하는 예측값은 각 시간대별로 탑승인원을 구하는 것이므로, 승차인원에서 하차인원을 뺀 값에서 시간대 별로 나누어진 데이터를 하나의 일일 데이터로 만들기 위하여 pandas의 Dataframe을 다루어 처리하였다.

3. 실험환경

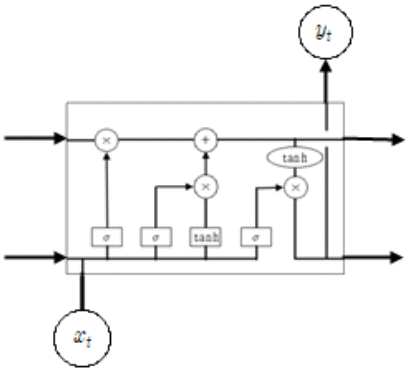
본 실험에서는 python3.6과 Keras를 이용한다. 데이터셋은 승차인원-하차인원(탑승인원)의 값을 입력으로 받는다. 데이터는 출퇴근 시간대에 가장 큰 값을 가지고, 05시, 11시 이후에는 급격히 감소하는 것을 알수 있었다. 입력의 범위의 격차가 크다면 Gradient Descent를 적용하기 까다로워지지만, 데이터를 정규화 하면 쉽고 빠르게 최적화 지점을 찾을수 있다. 따라서 본 논문에서는 원본(raw)데이터를 0과 1사이의 값으로 정규화를 실시하였다. 이중에서 2018년기준 가장 많은 승하차 인원을 가진 강남역을 기준으로 삼아 실험을 진행한다. 강남역의 데이터의 총 개수는 71,220개이며, training에 80%, test에 20%의 데이터를 할당하였다. 1일을 20개의 시간으로 나누었기 때문에 look_back을 20으로 설정하였다.

RNN은 은닉층의 결과가 다시 같은 은닉층의 입력으로 들어가도록 연결되어 있다. 따라서 글, 유전자, 음성 신호, 주가 등 시계열 데이터의 형태를 갖는 데이터에서 패턴을 인식하는데 우수한 성능을 지닌 인공 신경망이다. 패턴을 기억할 수 있는 능력이 있어 마치 사람의 기억력에 비유할 수 있다.



RNN구조

LSTM은 여러 개의 게이트(gate)가 붙어있는 셀(cell)로 이루어져 있으며 이 셀들이 하는 역할은 정보를 버리고, 저장하고, 업데이트하고, 내보내는 기능들로 이루어져 있다. 각 셀은 셀에 연결된 게이트의 값을 보고 무엇을 저장할지, 언제 정보를 내보낼지, 언제 쓰고 언제 지울지를 결정한다. LSTM은 이러한 셀 구조를 가짐으로써 RNN의 대표적인 문제인 기울기 소실 및 발산 문제와 장기 의존성 문제를 해결할 수 있다.



LSTM구조

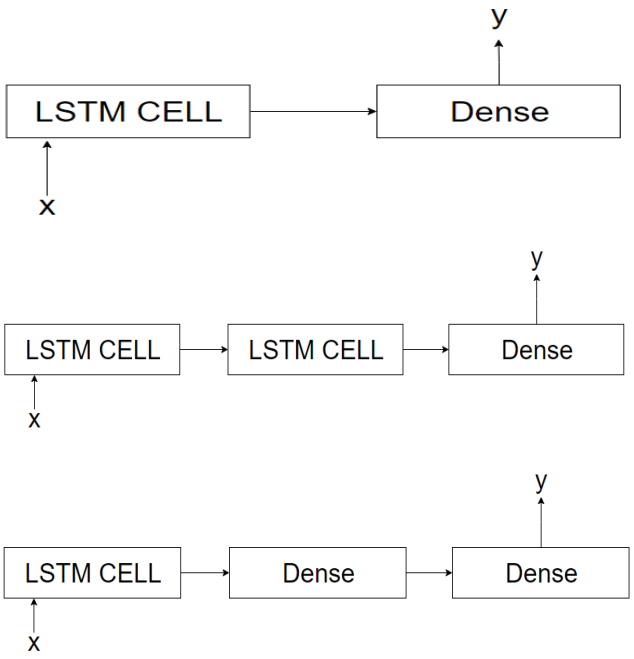
따라서 본 실험에서는 데이터의 학습 모델을 LSTM으로 구성한다. 모델의 성능 평가 척도는 평균 제곱오차(MeanSquareError, MSE)를 사용한다. MSE는 실제 데이터 값과 예측 값의 차이의 제곱을 의미한다. 차이가 크면 클수록 MSE는 더욱 커지게 된다. 따라서 이 값의 차이를 최대한 줄임으로써 실제 값과 예측 값의 차이를 줄이는 것이 이번 실험의 목표이다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

위 수식과 같이 입력값 y_i에서 예측값 $\tilde{y}_i$를 뺀값을 제곱한뒤 모두합하여 평균을 낸다.

학습의 loss는 mean_squer_error, optimizer는 rmsprop, epochs는 10, batch_size는 16으로 설정하였다. 모델의 마지막에는 하나의 출력값을 내보내기 위해 Dense(1)을 추가하였다.

따라서 본 실험의 예측 모델은 다음과 같이 구성할 것이다.



x는 입력 데이터이고, y는 학습된 모델에서의 예측값이다.

4. 실험 및 분석

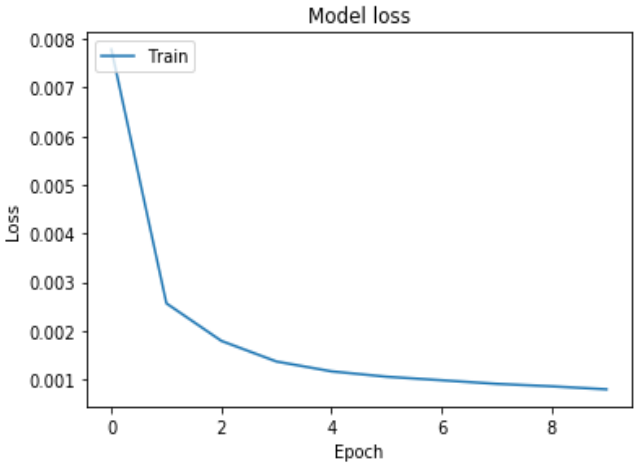
본 실험에서는 LSTM Layer를 최소값의 MSE를 얻을 수 있는 구체적인 구현 모델을 실험하였다.

Model	MSE
LSTM(32)	1.5367935e-3
LSTM(64)	2.1808513e-3
LSTM(128)	1.0510251e-3
LSTM(256)	9.3585515e-4
LSTM(512)	7.616265e-4
LSTM(1024)	5.893569e-4
LSTM(512)+LSTM(32)	1.693891e-4
LSTM(512)+LSTM(64)	2.1497853e-4
LSTM(512)+LSTM(128)	1.7026807e-4
LSTM(512)+LSTM(256)	1.861279e-4
LSTM(512)+LSTM(512)	1.3251218e-4
LSTM(512)+LSTM(1024)	1.1900409e-4
LSTM(512)+Dense(32)	7.3264446e-4
LSTM(512)+ Dense(64)	6.963257e-4
LSTM(512)+ Dense(128)	6.466417e-4
LSTM(512)+ Dense(256)	5.3646386e-4
LSTM(512)+ Dense(512)	8.9407247e-4
LSTM(512)+ Dense(1024)	8.351417e-4

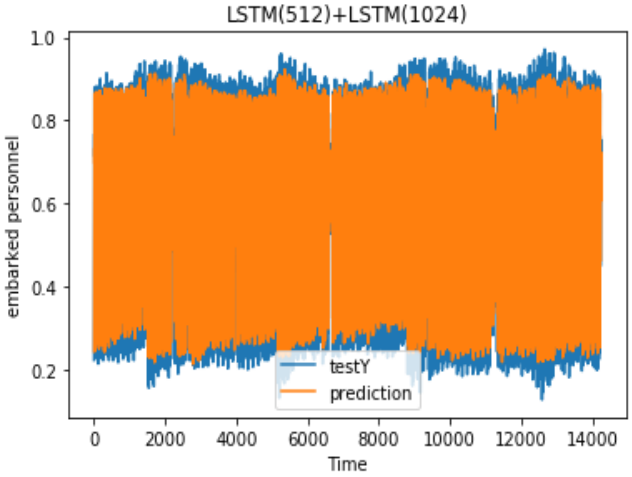
실험 결과를 분석해보면 단순 LSTM layer만으로도 괄목할만한 성과를 보이고있다. LSTM layer와 Dense layer의 결합 모델의 MSE가 높은 이유는

Fully-Connected 만으로는 시계열 데이터의 예측 정확성이 떨어진다는 것을 의미한다. 과거의 데이터를 고려하지 않았기 때문이다. 또한 LSTM을 중첩하였음에도 MSE가 크게 떨어지지 않는 것은 입력 데이터의 크기에 비해 모델이 너무 복잡하여 과도한 파라미터를 가지므로 학습 능력이 줄어든다는것을 알 수 있다.

가장 오차가 적은 LSTM(512)+LSTM(1024)모델의 loss 그래프는 다음과 같다



이 모델의 정확도 그래프는 다음과 같다.



testY는 test데이터의 정답 값이고, prediction은 학습된 모델에 test데이터의 입력 값을 넣어서 예측한 값이다. 이 그래프에서의 MSE는 1.1900409e-4가 나왔다.

5. 결론 및 향후 연구

본 논문에서는 강남역의 시간별 탑승인원에 대해 Keras의 LSTM layer와 Dense layer를 이용하여 MSE를 줄이는 방향으로 연구를 진행하였다.
이번 실험에서는 강남역 하나에 대한 탑승 인원 예측을 진행 했다. 이를 확장시켜 모든 역에 대한 탑승 인원을 예측하면 지하철 수용인원 대비 탑승인원을 계산하여 지하철 포화도를 계산할 수 있을 것이다. 이를 통해 이용객들에게 출퇴근 시간대의 특정 시점에서의 지하철 포화도를 제공할 수 있으며, 비교적 적은 포화도에서 편안한 지하철 이용을 가능하게 하는 서비스를 만들 수 있다. 많은 사람들이 이러한 포화도를 보고 지하철을 이용한다면 분산효과를 기대할수 있으며 이는 지하철 이용의 불편 감소로 이어지는 긍정적인 영향을 가져올수 있다. 또한 정부기관에서는 미래의 탑승 인원을 예상하여 지하철 시간표를 조정할 수 있고, 예산할당의 지표로 삼을 수 있다. 현재 데이터셋은 각 시간별로 구간을 나누어서 예측한 모델이지만, 이를 선형 회귀 모델로 바꾸어 분 단위의 탑승인원 예측을 가능하게 할 수 있다. 또한 LSTM을 사용했던 모델의 개선을 통해 모델의 정확도를 더욱 높일 수 있다.

참 고 문 헌