1. 2018-1 머신러닝 최신기술 특론 보고서

- Github URL: https://github.com/bonopi07/2018-1_advML_project

구성원

- Department: Computer Science in Kookmin University
- Student name: Seongmin Jeong (English name: Allen)
- Student ID: M2018075
- Laboratory: Information Security Lab

프로젝트 환경

- Python v3.6.2
- Tensorflow v1.8

프로젝트 목표

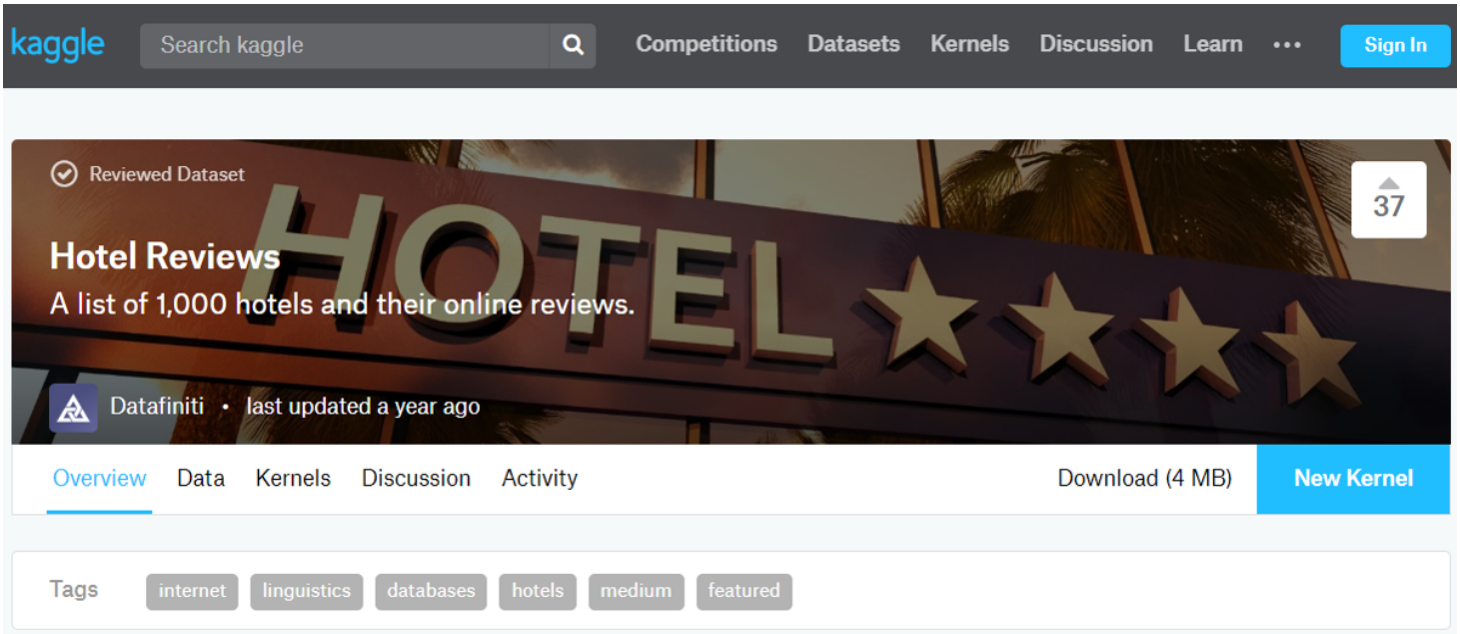
- 호텔 리뷰가 주어지면 평점을 예측할 수 있는 텐서플로우 기반 RNN(or LSTM) model을 학습한다.
- 자연어 처리 기반의 many-to-one 방식 RNN을 학습한다.
  - RNN 실습을 통해 동적 입력 크기에 대한 처리 방법을 익힌다.
  - char-level, word-level 의 다양한 자연어 처리 방법을 공부한다.
  - deep, wide한 방법(e.g.stacked RNN)과 다양한 딥러닝 테크닉(e.g.dropout, word embedding)을 적용해본다.
  - RNN model의 뒤에 softmax나 FC layer를 붙혀본다.

2. Processing Data

데이터셋 정보

- 프로젝트를 위한 dataset은 kaggle dataset에서 제공하는 정보를 사용하였음.
  - URL: https://www.kaggle.com/datafiniti/hotel-reviews
- Datafiniti의 비즈니스 DB에서 제공하는 1000개의 hotel review list이다.
- 데이터셋은 하나의 csv file로 주어지며, 약 40000개의 정보를 포함한다.
- 데이터셋은 호텔 이름, 호텔 위치, 작성자 이름, 평점, 리뷰 등의 정보를 포함한다.
  - 본 프로젝트는 평점(rating)과 리뷰(review text)만을 사용한다.

- 데이터셋의 리뷰 정보는 영어와 스페인어로 표기되어 있다. (인코딩: latin-1)



파일 설명

- data.csv: 제공하는 원본 dataset file
- data_rating_text.csv: data.csv에서 rating과 review text만 추출한 file
- data_final.csv: 데이터 정제 방법을 거친 후 최종적으로 사용하는 dataset file
- char2idx.pickle: 리뷰 기준에 대한 character 별로 숫자를 mapping한 python dictionary 객체 (pickle format)
- word2idx.pickle: 리뷰 기준에 대한 word 별로 숫자를 mapping한 python dictionary 객체 (pickle format)
- data_final.csv 데이터 화면(좌측 열: 평점, 우측 열: 리뷰 텍스트)

	A	B	C	D	E	F	G
43	2	you need to up grade the room the carpet in my room was p					
44	1	you need to post pictures of both location. the other locatior					
45	3	you need to pay more attention to the cleanliness of the bed					
46	4	you might struggle pronouncing the name of this city but yo					
47	0	you might as well put a sleeping bag in the middle of the hi					
48	4	you may want to re-think this decision.					
49	3	you may think youre lost as you drive the long and winding r					
50	2	you know... i read the reviews and was worried about staying					
51	2	you know... i read the reviews and was worried about staying					
52	4	you know youre welcome after a fun and friendly check-in w					

데이터 정제 방법

- 평점 기준: 1 ~ 5 점의 정수형 데이터로 한정한다. [data에는 0 ~ 4점으로 표시되어 있다.]
  - 그 외의 평점(e.g. 0점, 실수 평점)은 폐기 및 반올림한다.
- 보장된 데이터 사용: 평점, 리뷰가 하나라도 존재하지 않은 정보는 사용하지 않는다.
- 리뷰 기준
  - 리뷰는 미리 정한 character만 취급한다. (그 외의 문자는 사전 삭제한다.)
  - 알파벳은 모두 소문자로 취급한다.
  - 미리 정한 character 기준 (44개): 숫자(0-9), 알파벳(a-z), 공백, 특정 특수문자(온점(.), 반점(,), 하이픈(-), 언더스코어(_), 느낌표(!), 물음표(?), 슬래시(/))

데이터 입력 단위

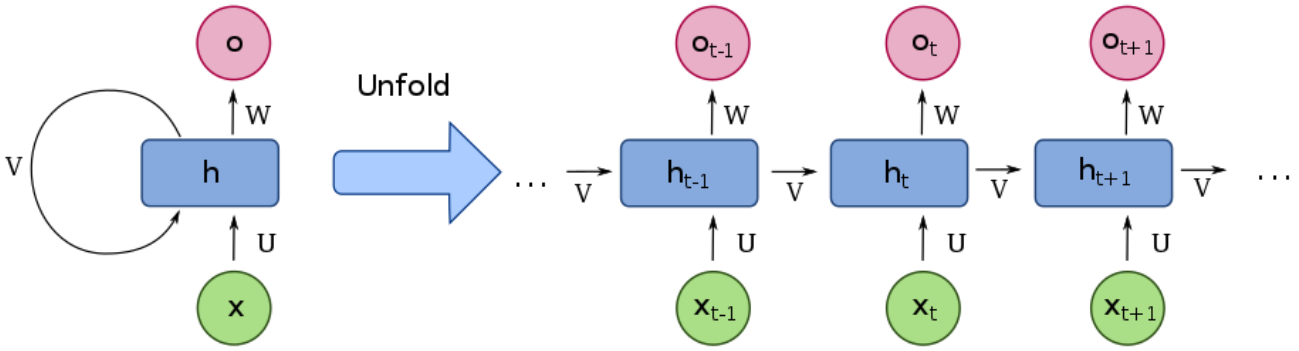
- Character-level: 각 문자를 one-hot encoding한 후 (44-D) RNN의 data-dim(=hidden size)로 사용한다.
- Word-level: 가지고 있는 데이터셋에서 나오는 모든 단어에 대한 word2idx 사전 객체를 생성한다 (개수: 58,694). 이후에는 각 단어에 대한 숫자를 tf에서 지원하는 embedding lookup 메서드를 통해 벡터를 생성하고(300-D), 학습을 통해 벡터를 최적화한다.

3. Algorithms

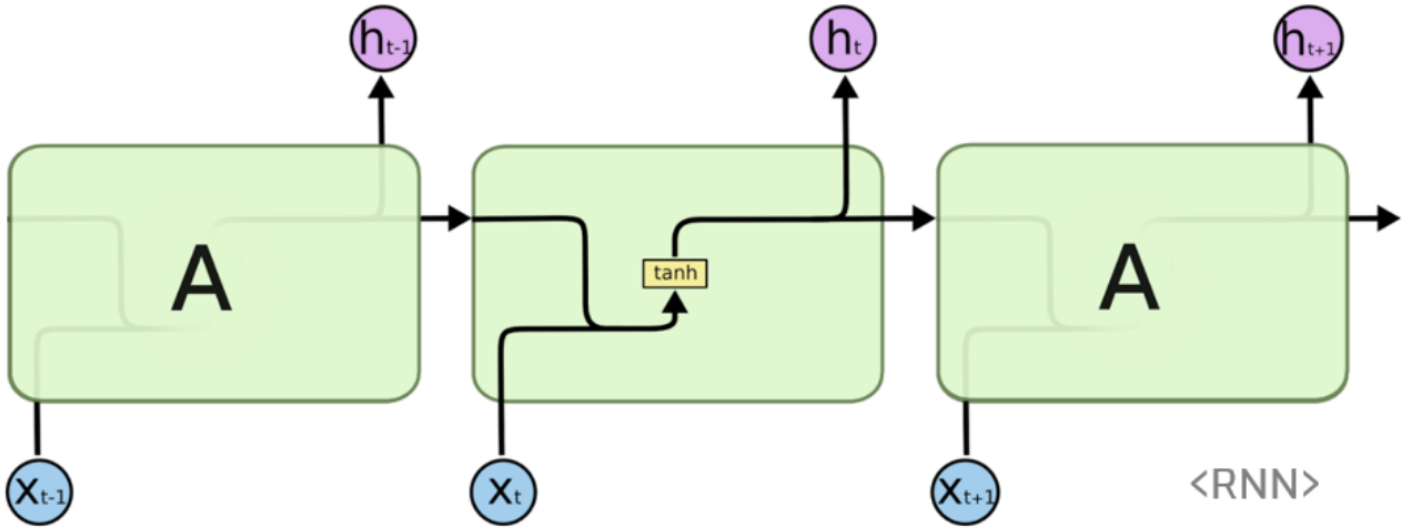
- 본 프로젝트는 리뷰 데이터(data)와 평점 데이터(label)를 이용하여 LSTM 기반의 예측모델을 만들 것이다.

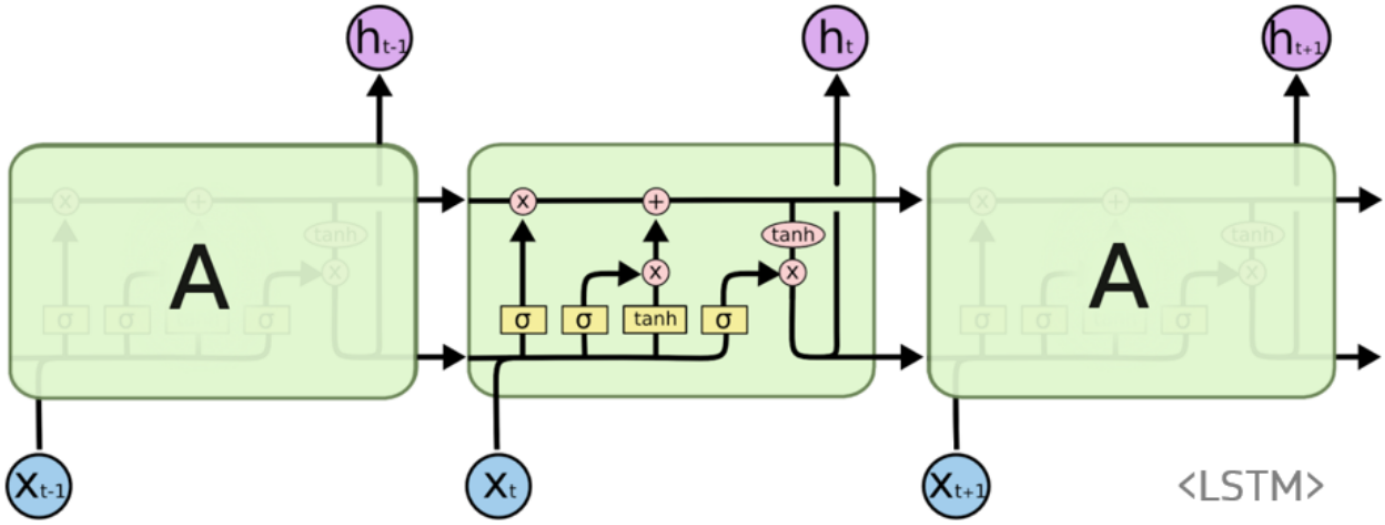
RNN

- RNN은 인공 신경망의 한 종류로, 유닛 간의 연결이 순환적 구조를 갖는 특징을 갖고 있다. 이러한 구조는 순차적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해준다. Feedforward 신경망과 달리, recurrent 인공 신경망은 내부의 메모리를 이용해 시퀀스 형태의 입력을 처리할 수 있다. 따라서 recurrent 인공 신경망은 필기체 인식이나 음성 인식과 같이 순차적 특징을 가지는 데이터를 처리할 수 있다.



- 하지만 RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 back propagation시 gradient가 점점 줄어 학습 능력이 크게 저하되는 현상을 가진다. 이를 vanishing gradient problem 이라고 한다. 이 문제를 보완하기 위해 RNN의 hidden state와 cell state를 추가로 가지고 있는 LSTM 모델을 사용할 것이다.
- LSTM은 cell state와 hidden state를 통해 추가적인 연산을 하게 되는데, 이를 통한 3가지의 gate가 생성된다. Forget gate, input gate, output gate인데, 이 gate들은 이전 값을 얼마나 기억할지, 그리고 현재 값을 얼마나 기억할지 등의 값을 제어하기 위한 목적으로 사용된다. 본 프로젝트는 간단한 LSTM model을 설계함으로써 호텔 리뷰와 같은 다대일(many-to-one) 문제를 해결한다.





Model Architecture

- LSTM Layer #
  - Single (1 Layer) --> Stacked (3 layer)
- Data Dimension
  - char-level : 44차원
  - word-level : 300차원
- Loss Function
  - Softmax cross entropy (with Adam Optimizer)
- 데이터셋 검증 및 모델 파라미터 검증 방법
  - K-fold Cross Validation (K = 5)

Result

실험 step	정확도 (Accuracy)

실험 step	정확도 (Accuracy)
1	89.17
2	88.42
3	91.03
4	84.98
5	86.11
평균	87.94