

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

| Project Title | 딥러닝을 활용한 노래 가사 분석 및 추천 시스템 |
| --- | --- |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-05-13 ~ 2022-05-19 |
| Project Manager of Reporting Week | 유창현 |
| Faculty Supervisor | 유창현 |

# **Tasks Outlined in Previous Weekly Progress Report** (Provide detailed information on the tasks to be completed in this week)

* 행복 이외의 키워드에 대한 가사를 추가로 수집하여 감정사전을 재구축
* 행복으로 분류되는 단어중 ‘나’를 제거한 후 데이터 불균형 문제 해결
* 감정 분류에 활용할 4가지 딥모델 선정 후 각자 맡은 모델 조사 및 구현시작

# **Progress Made in Reporting Week** (Provide detailed information on the progress that you made in the reporting week. Limit your write-up to no more than two page)

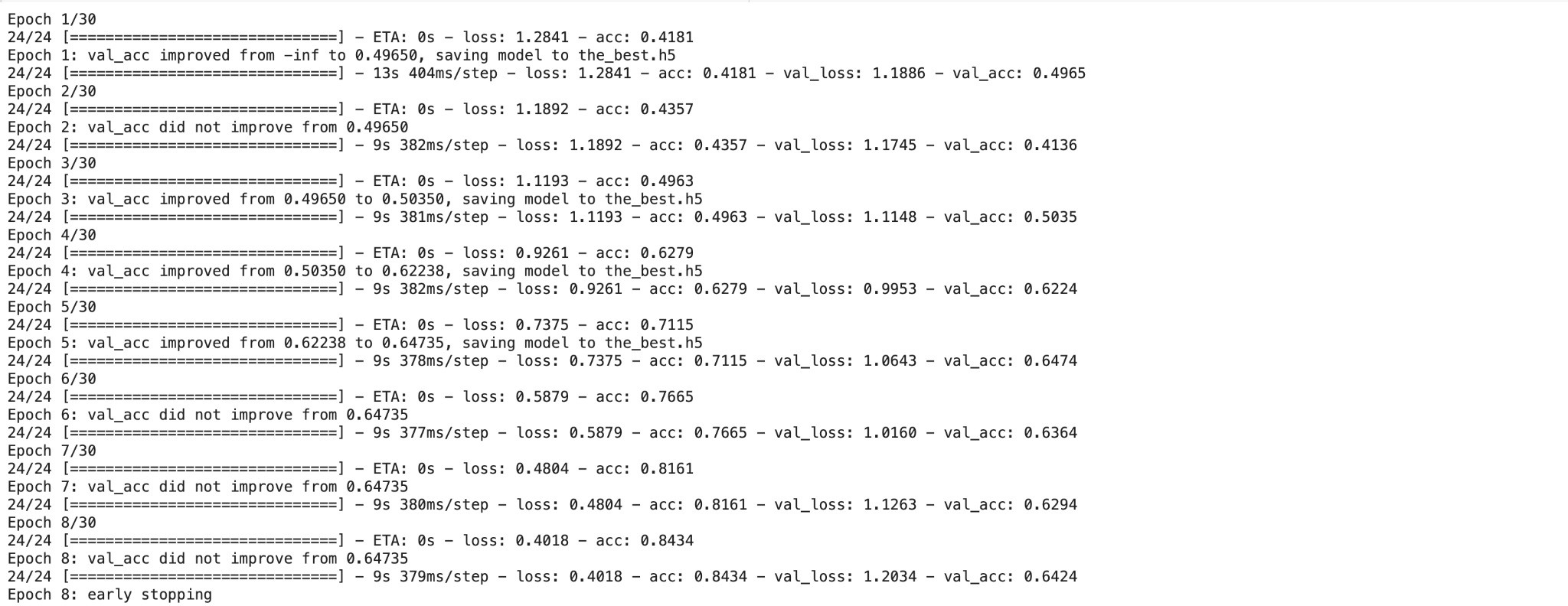
1. 행복을 제외한 다른 감정들에 관련된 키워드로 노래 가사를 추가로 수집하여 감정사전 재구축

기존에 ‘사랑’과 관련된 노래의 가사 데이터와 TOP100 데이터의 경우 행복과 관련된 가사와 단어가 주를 이루었기 때문에 그 외에 두려움, 분노, 슬픔과 관련된 노래 가사를 추가로 수집하여 감성 사전의 단어를 추가하고 다시 한번 라벨링 작업을 수행하였다.

1. 행복으로 분류되는 단어중 ‘나’를 제거한 후 데이터 불균형 문제 해결

데이터 불균형 문제 해결을 위해 구축한 감성사전을 확인하던 중 행복에 해당하는 단어에 ‘나’라는 단어가 포함되어 있음을 확인했다. 이 단어는 행복과 직접적인 연관성이 있는 단어가 아닐 뿐더러 다양한 노래 가사에 여러번 등장할 수 있는 단어이기에 삭제하였다. 그 결과 지나치게 행복에 해당하는 노래가 많았던 데이터 불균형 문제가 해결되었고, 1)에 언급했듯 감정별로 추가로 수집한 노래들도 데이터셋에 포함하여 4가지 감정에 해당하는 노래들이 균등하게 포함된 최종 데이터셋을 구축할 예정이다. 기존에 5가지로 분류하기로 했던 기준으로 4가지로 수정한 이유는 ‘무관심’에 해당하는 노래가 극히 드물어 데이터셋을 구축하기에 어려움이 크다고 판단했기 때문이다.

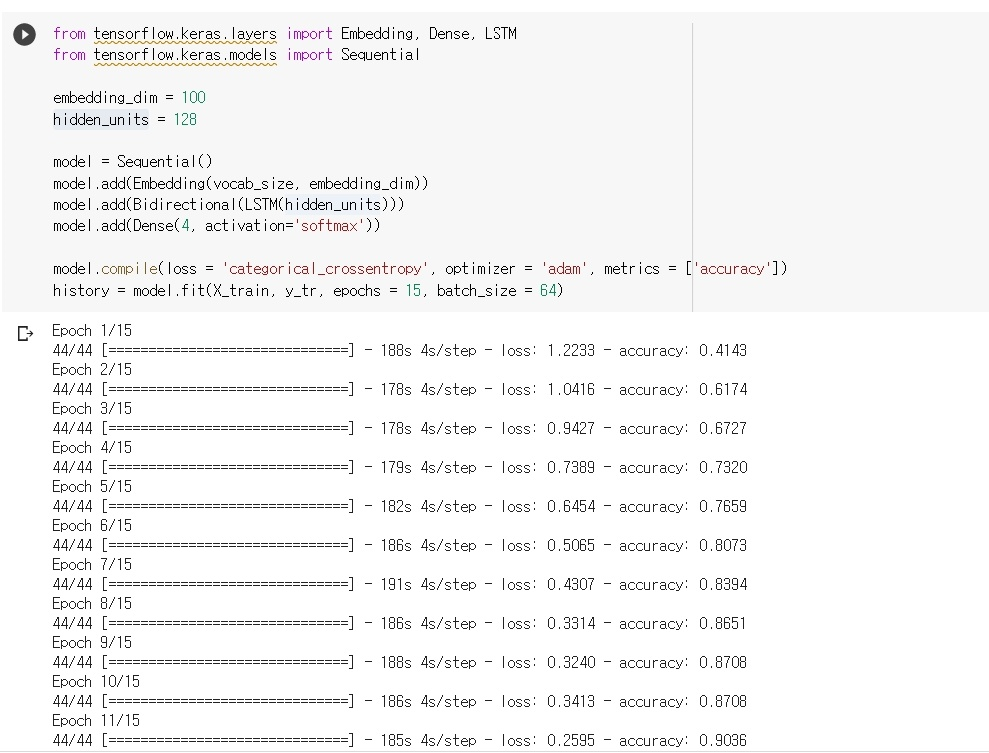
1. 감정 분류에 활용할 4가지 딥모델 선정 후 각자 맡은 모델에 대해 자료 조사

* LSTM: LSTM은 RNN(Recurrent Neural Network)의 한 종류로, RNN은 전통적인 뉴럴 네트워크와 달리 이전에 일어난 사건을 바탕으로 나중에 일어나는 사건을 생각할 수 있다. LSTM은 다른 RNN에 비해 긴 문맥에 대해 학습하는 데에 더 좋다.

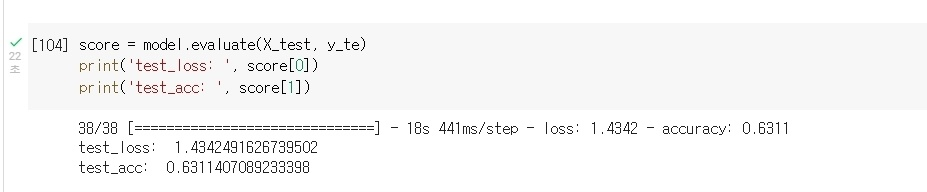
train data의 비율은 75%, 임베딩 차원은 128, hidden\_unit은 128로 설정하였고, train data에 대한 정확도는 약 84%, test(validation) 정확도는 약 64% 였다.

* BiLSTM: 양방향 LSTM은 두 개의 독립적인 LSTM 구조를 함께 사용하는 구조이다. 양방향 LSTM은 앞뒤의 문맥을 모두 고려하기 위해서 문장을 왼쪽에서 오른쪽으로 읽는 정방향 셀과 오른쪽에서 반대로 읽는 역방향의 셀을 함께 사용한다. 이 두 가지 정보를 고려하여 출력층에서 두 가지 정보를 모두 활용하여 예측한다.

모델을 설계 하기에 앞서 가사 데이터 토큰화와 정수 인코딩, 패딩 과정을 진행했고 그 이후 모델 구현과 학습과정을 진행하였다.

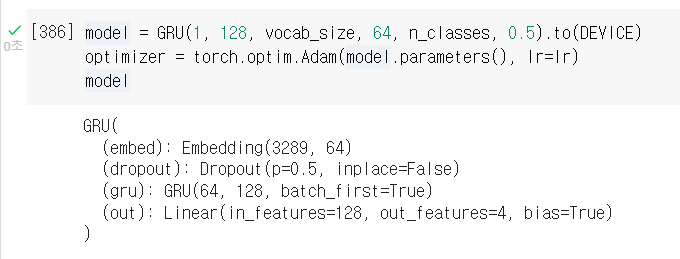


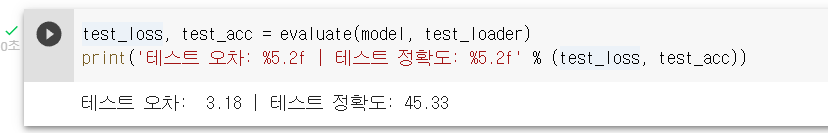
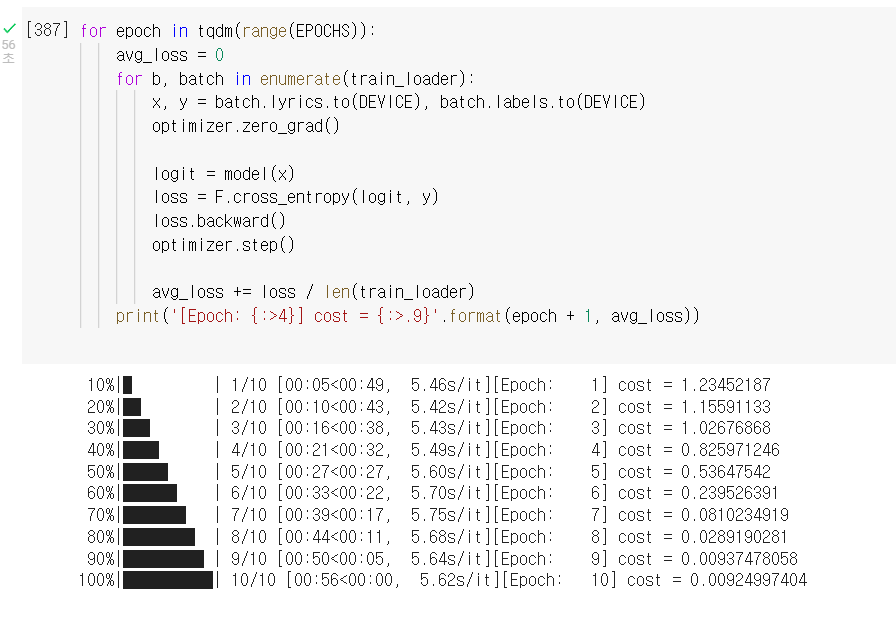
구현 시작 단계이기 때문에 다양한 파라미터와 모델 설계를 시도하는 과정은 아직 수행 전이며, train data의 비율은 70%, 임베딩 차원은 100, hidden\_unit은 128로 설정하였다. 15 epoch 학습 결과 train data의 정확도는 약 92% 수준이었다. test data로 모델을 성능을 평가한 결과 정확도는 약 63%로 나왔다.

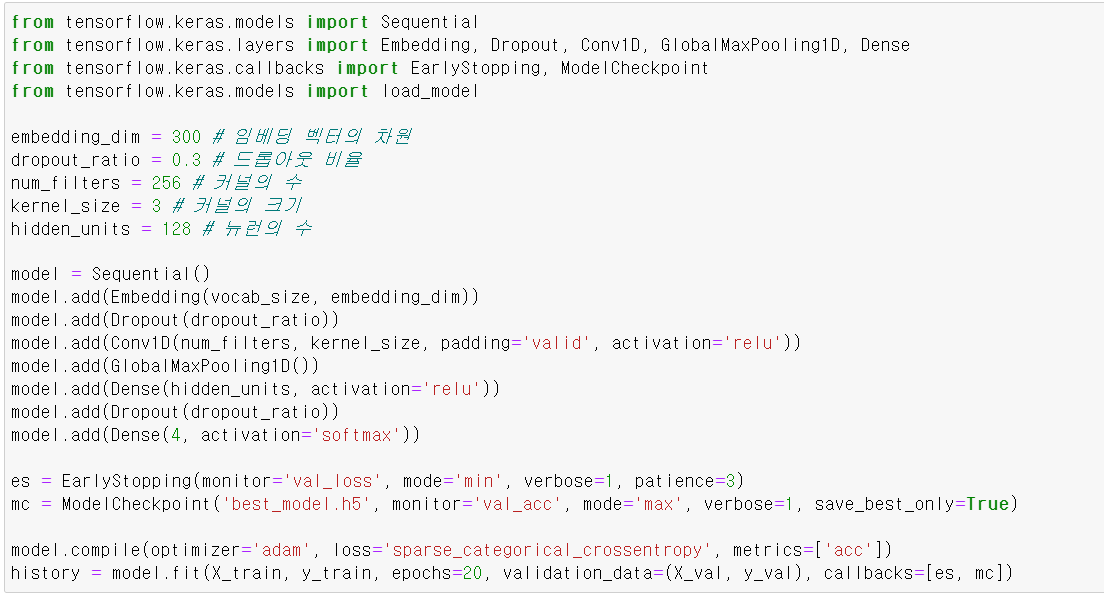
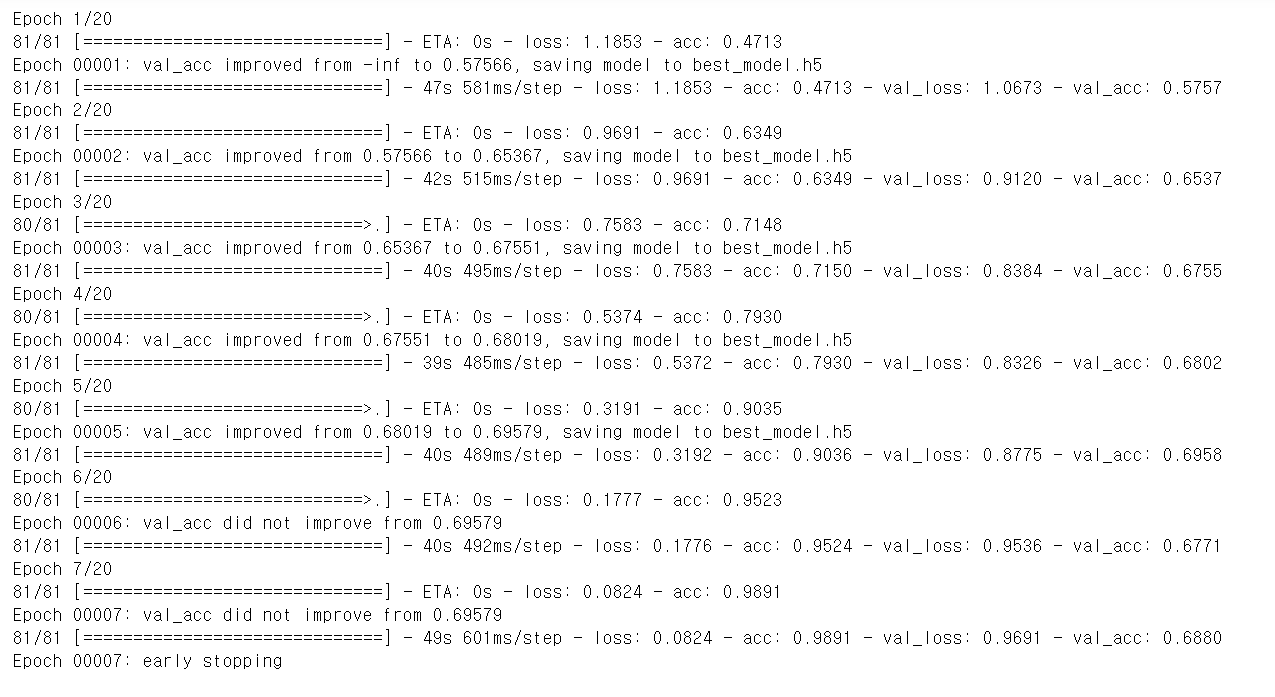


* GRU: GRU(Gated Recurrent Unit)는 LSTM의 장기 의존성 문제에 대한 해결책을 유지하면서, 은닉 상태를 업데이트하는 계산을 줄였다. 다시 말해서, GRU는 성능은 LSTM과 유사하면서 복잡했던 LSTM의 구조를 단순화 시켰다. LSTM에서는 출력, 입력, 삭제 게이트라는 3개의 게이트가 존재하는 반면, GRU에서는 업데이트 게이트와 리셋 게이트 두 가지 게이트만이 존재한다.

감정 컬럼의 클래스를 숫자로 레이블링 하고 torchtext를 통해 토큰화, 전처리를 진행했다. train:set = 8:2 의 비율로, 임베딩 차원은 64, hidden unit은 128로 설정했다. 10 epoch로 학습을 진행하였고 testset으로 모델 성능을 평가한 결과 정확도는 45%로, 오차는 약 3%로 나왔다.





* CNN : 이미지 처리에 주로 사용하는 모델이라 정확도가 크게 나오지 않았다.train set 98% 정확도. / test set 48% 정확도
* 
* 

# **Difficulties Encountered in Reporting Week** (Provide detailed information on the difficulties and issues that you encountered in the reporting week. Limit your write-up to no more than one page)



# **Tasks to Be Completed in Next Week** (Outline the tasks to be completed in the following week)

* 모든 모델에 동일하게 적용할 train/test set 비율 설정
* 성능 평가 지표 선정
* 딥러닝 모델 별 성능 비교