

Department of Software Convergence

Data Analysis Capstone Design Project   
Weekly Progress Report

| Project Title | 딥러닝을 활용한 노래 가사 분석 |
| --- | --- |
| Student / ID | 김은비 / 2019100858  유창현 / 2018102121  이은경 / 2019100896  이인석 / 2018110659 |
| Reporting Week | 2022-05-27 ~ 2022-06-02 |
| Project Manager of Reporting Week |  |
| Faculty Supervisor | 유창현 |

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

* 모델 학습에 활용되는 데이터셋 추가적으로 개선
* 4가지 모델에 대한 성능 개선

# **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)

* train, validation, test set 비율 조정 및 감정별 데이터 비율 개선

이전 주차에는 train : validaion : test set = 5: 2: 3으로 설정하여 나누었는데, train set의 비율이 너무 작다는 교수님의 조언에 따라 7: 1: 2로 변경하여 재분할하고 각 모델 학습에 활용하였다. 하지만, 비율을 변경한 이후에 정확도가 조금씩 높아지긴 했으나 성능 크게 향상되지 않았다. 그래서 다른 방법으로 4가지 감정에 해당하는 데이터의 비율을 조정하여 4가지 감정 데이터가 균일한 비율로 학습에 포함되도록 조정하였다. 기존 데이터셋의 경우 슬픔 3234곡, 행복 2356곡, 분노 956곡, 두려움 792곡으로 슬픔과 행복에 해당하는 곡이 지나치게 많고 분노와 두려움에 해당하는 곡은 상대적으로 매우 적어서 이러한 불균형이 모델의 정확도에 영향을 주었을 것으로 판단했기 때문이다. 해당 문제를 해결하기 위해 1000곡을 기준으로 하여 이보다 많은 곡이 포함되어 있는 경우 제거, 적은 곡이 포함된 경우 크롤링을 통한 데이터 추가를 진행하였다.

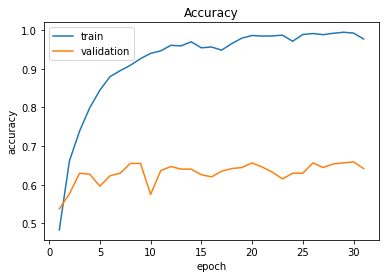
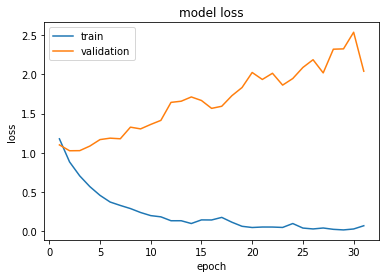
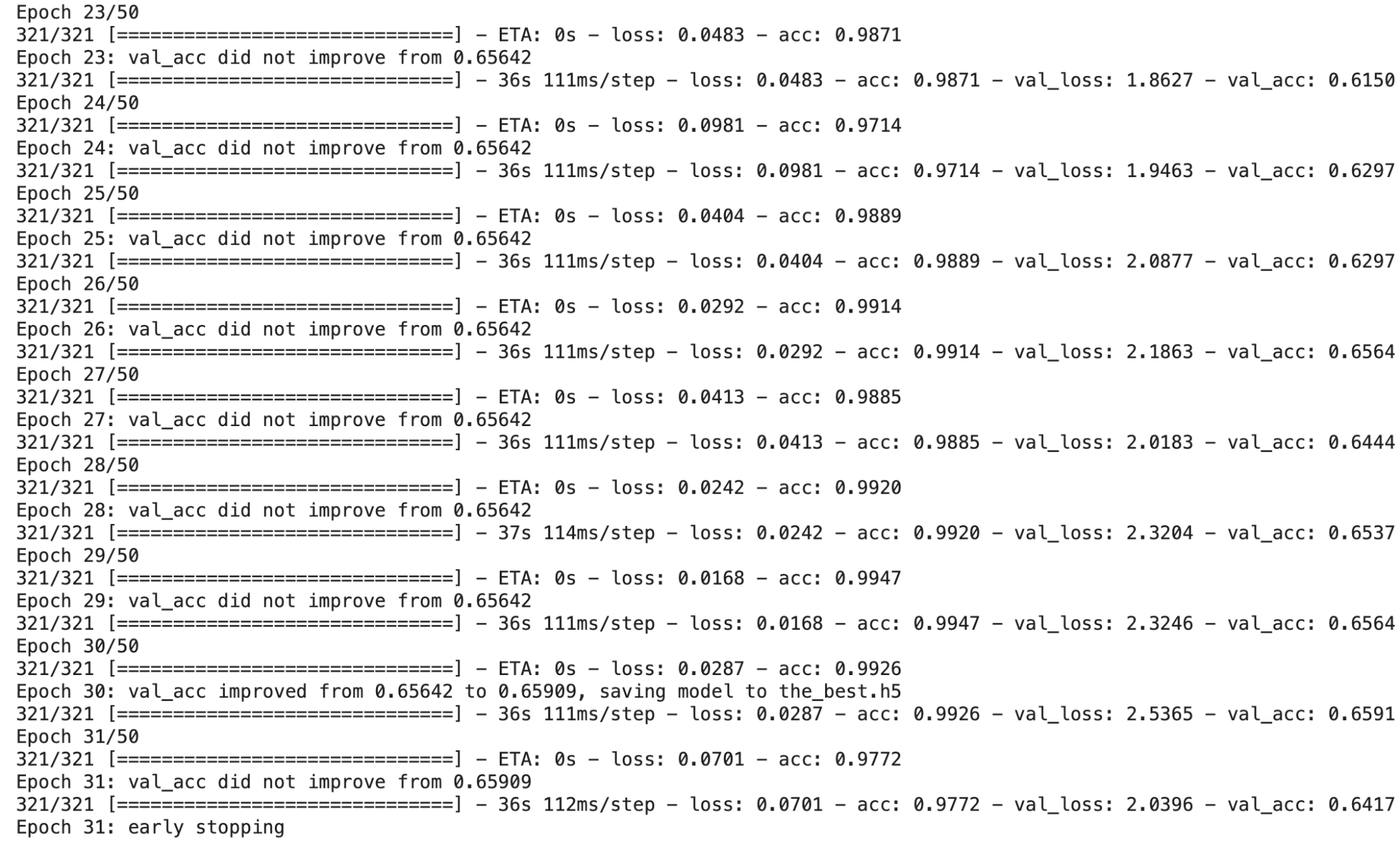
각각 2가지 데이터셋을 활용하여 모델을 학습시킨 결과, 1000곡이라는 동일한 곡수로 조정했을 때보다 기존에 불균형하지만 데이터의 개수가 더 많은 데이터셋을 활용한 경우의 정확도가 더 높게 나타났다.

* 데이터셋 구성

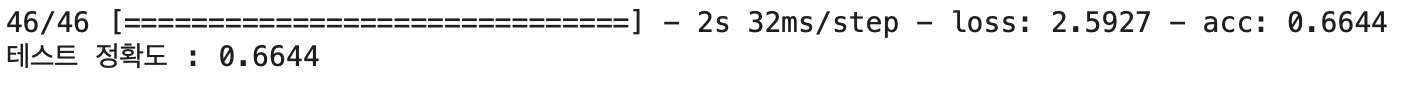
| 감정 | 데이터 수8,394(비율) |
| --- | --- |
| 슬픔 | 4,0945곡(44%) |
| 두려움 | 865곡(11%) |
| 행복 | 2,406곡(32%) |
| 분노 | 1,028곡(13%) |

-LSTM

이번에는 embedding\_dim = 100, hidden\_units = 128, epochs = 50, batch\_size = 16으로 설정하고, 모델을 최대한 훈련시킨 후의 결과를 보기 위해 earlystopping 옵션을 patience = 30 으로 설정하였다.

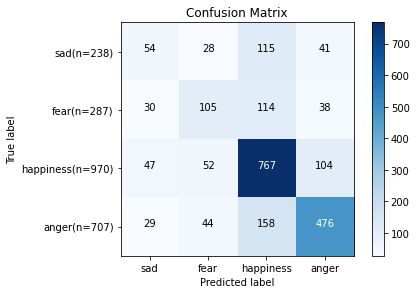


하지만 validation\_loss는 계속해서 발산하는 모습을 보여주었다.

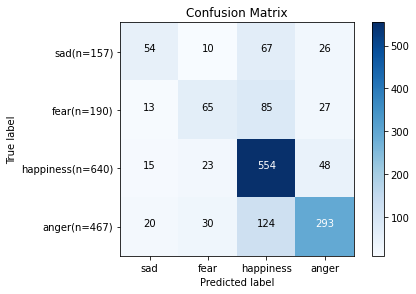


테스트 정확도는 지난 번 0.6367 보다 3% 정도 증가하였다.

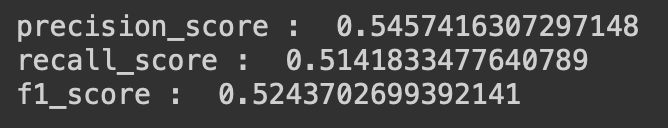
지난 번 confusion matrix



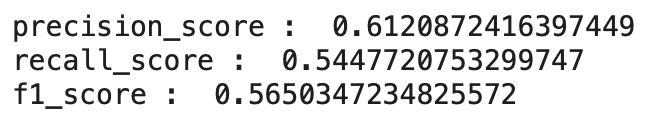
이번 confusion matrix



지난 번 score



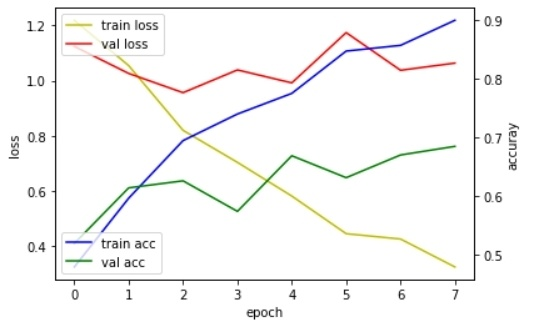
이번 score

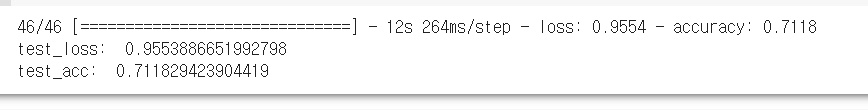


대체로 저번보다 모델의 성능이 증가한 모습을 볼 수 있었지만, validation loss 가 epoch 마다 증가하는 모습을 볼 수 있었고, 아직 기대할만한 성능에 못 미치는 성능과 정확도를 보임

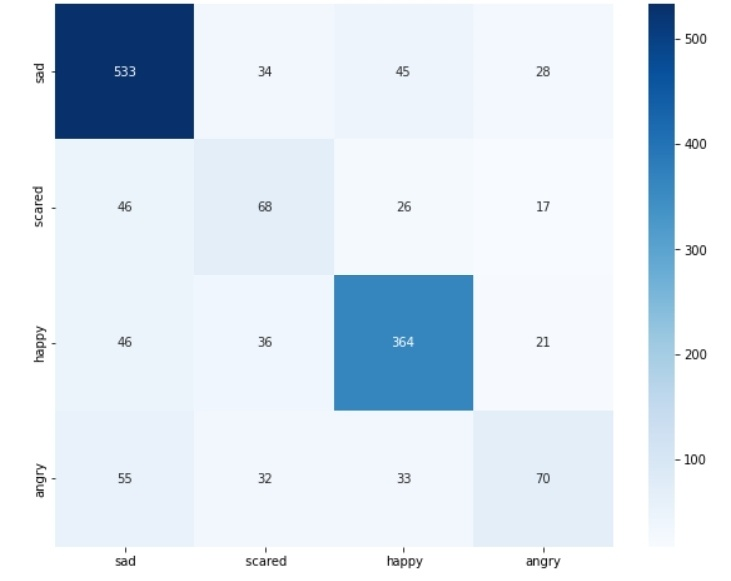
-Bi-LSTM

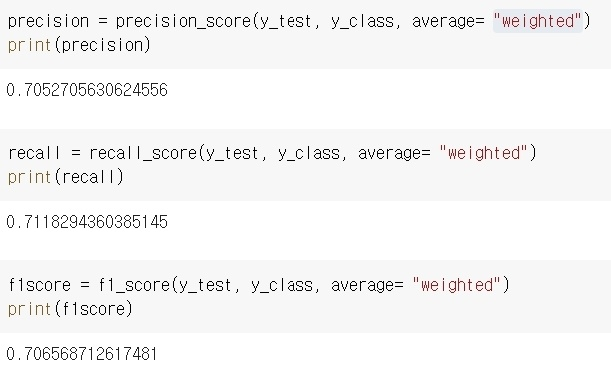
가장 정확도가 높은 모델은 embedding\_dim = 120, hidden\_units = 100, epochs = 20, batch\_size = 80으로 설정하고 earlyStopping옵션은 patience = 5로 설정한 모델이다. 모델 실행결과 Epoch 8에서 early stopping되었다.



train loss는 지속적으로 감소했지만 validation loss는 Epoch 2이후에 계속 증가하는 양상을 보였다. test set을 통해 모델 평가를 진행한 결과 test accuracy는 71.1%로 나타났다.

confusion matrix와 precision, recall, f1-score는 다음과 같다.



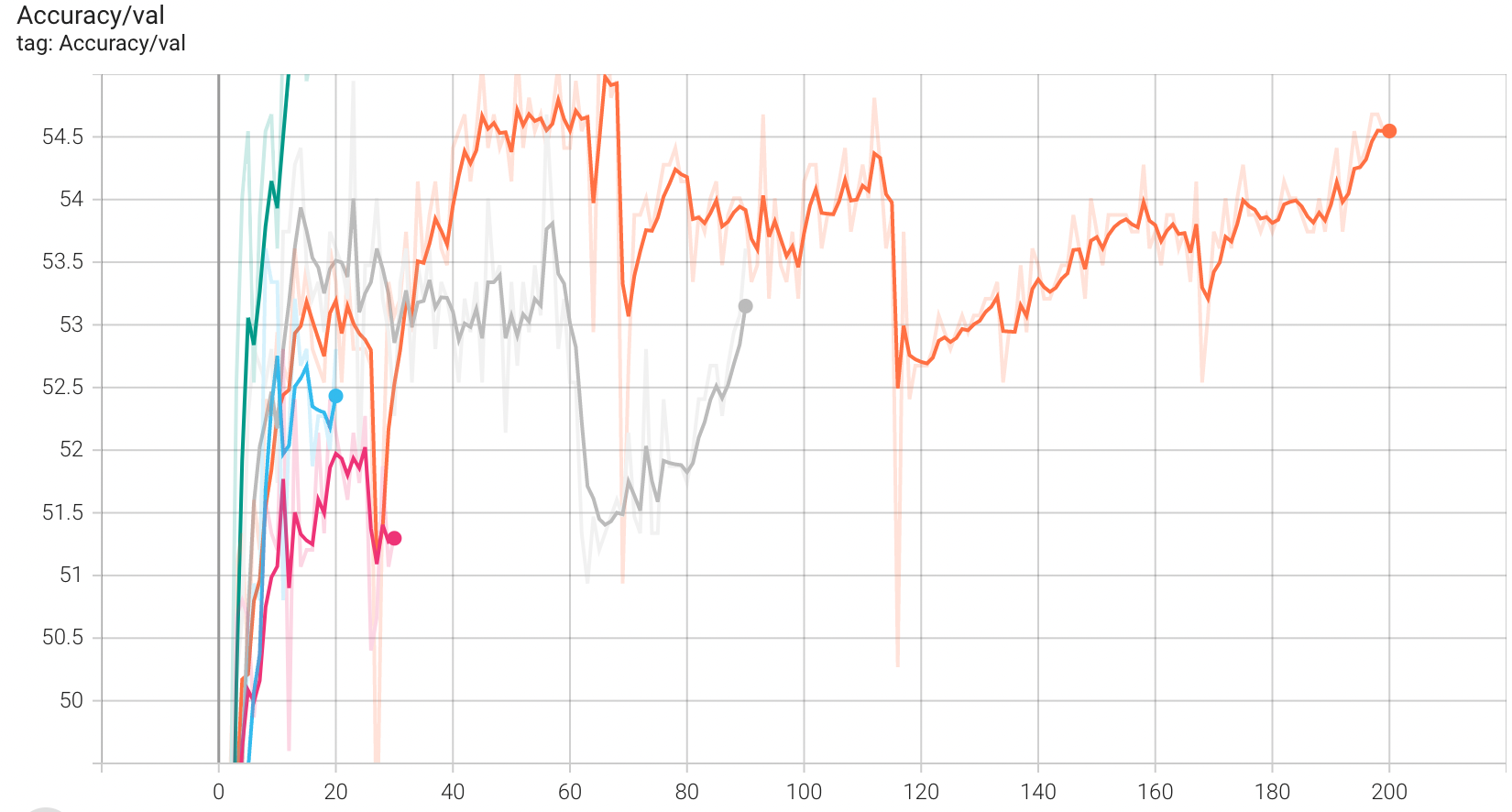


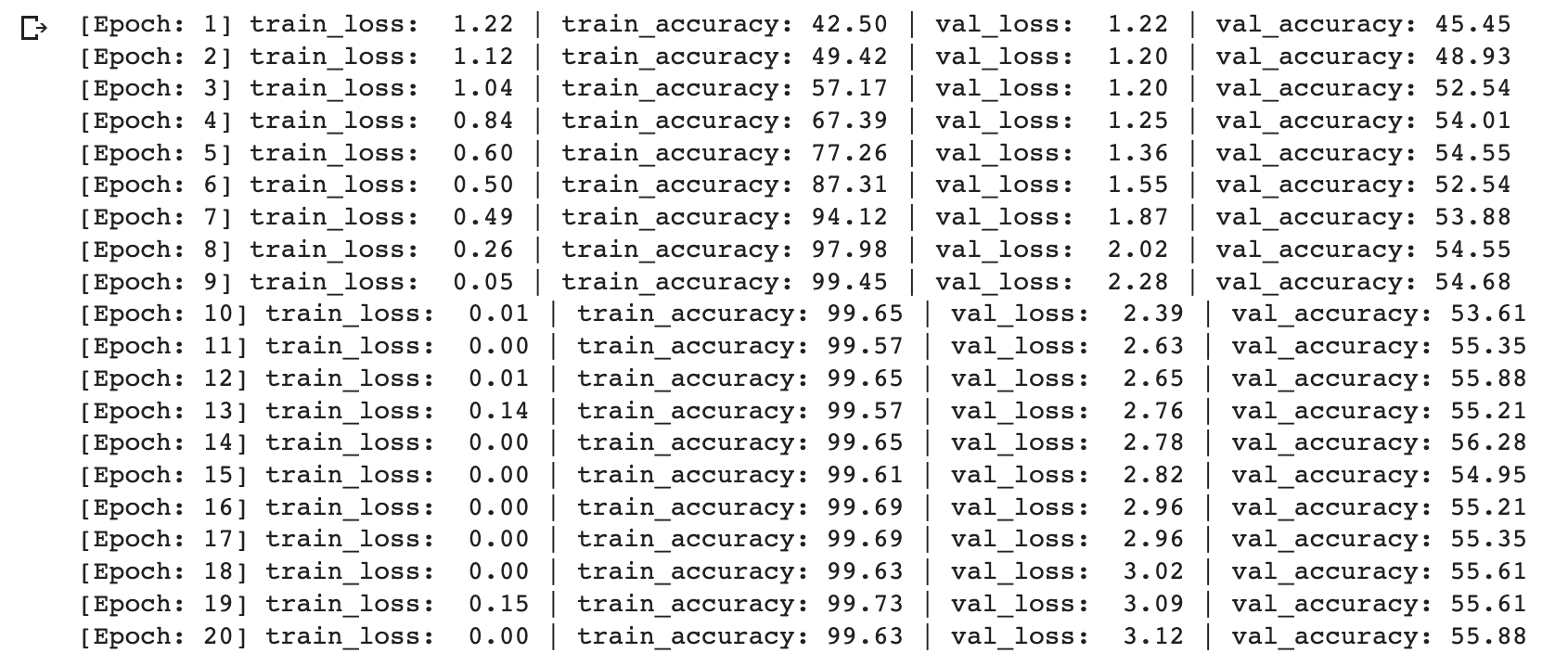
지난주 성능과 비교했을때 정확도는 67.9%에서 71.1%로 3.2% 올랐고, loss값은 1.34에서 0.95로 감소하였다. precision은 0.57에서 0.7, recall = 0.54에서 0.71, f1-score는 0.55에서 0.7로 모두 개선되었다.

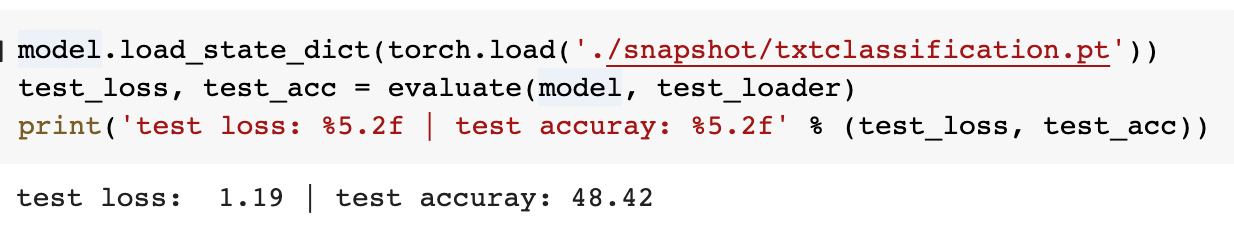
각 감정의 곡들이 10000곡씩 포함된 데이터셋을 활용한 경우의 test 정확도는 61.3%로 오히려 더 낮은 정확도를 보였다.

* GRU

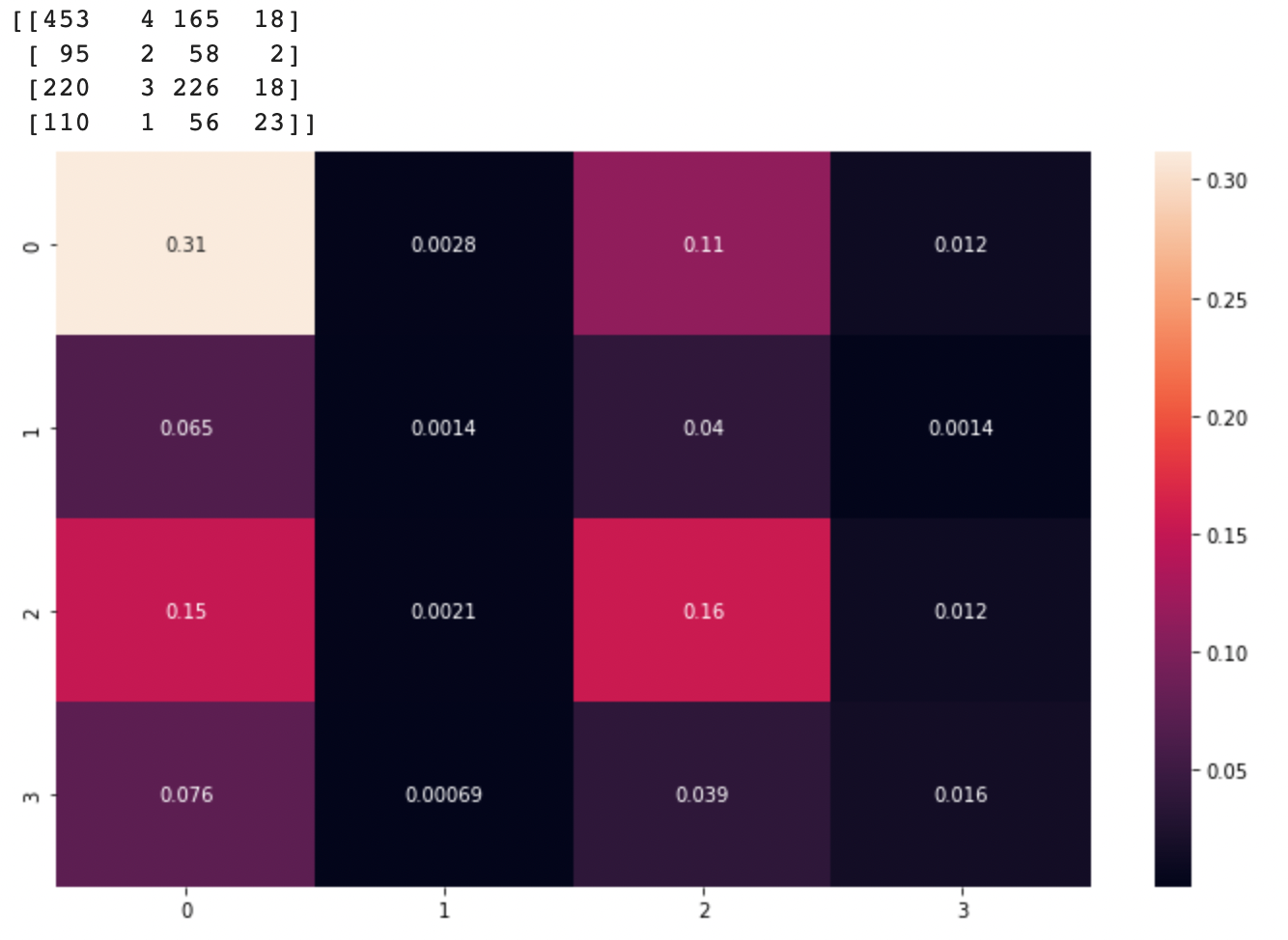
hidden\_units=128, embedding\_dim=100, drop\_out=0.5, batch\_size=32, epoch=20일때 가장 좋은 성능을 보여주었다. 다음은 validation의 accuracy의 그래프이다. 하이퍼 파라미터 조정을 통해 여러가지 경우의 수를 확인해 보았지만 큰 변화가 없었다. 또한 test의 accuracy도 47-49%를 유지하였다.

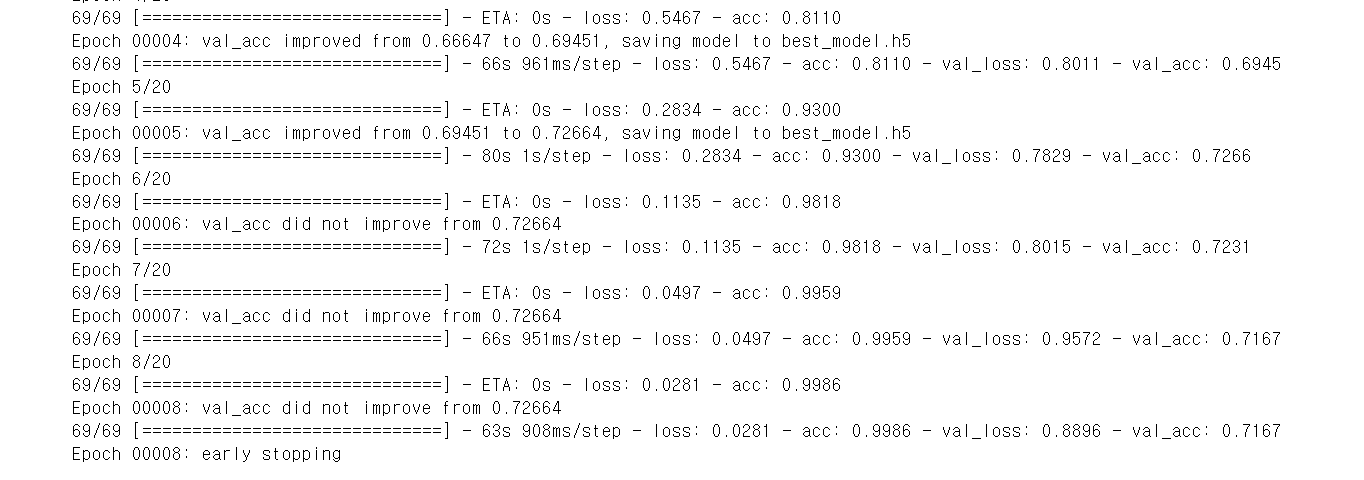


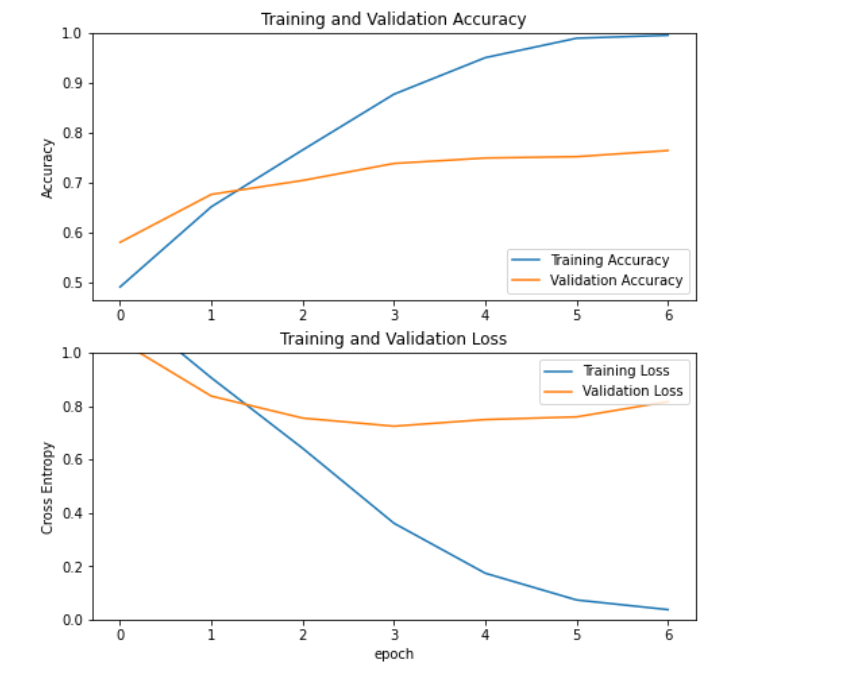


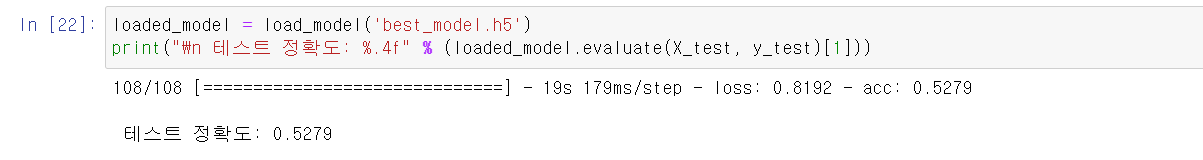


confusion matrix와 precision, recall, f1-score는 다음과 같다.



* CNN
* 
* 







# **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)

* train, validation, test set 비율 조정과 불균형 해소 두가지 방법 모두 정확도 향상에 도움을 주지 않았다. 특히 각 감정 당 1000곡 씩 구축한 데이터셋에 대해서는 4가지 모델 모두 오히려 성능이 떨어졌다.

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

* 성능 개선을 위한 방법 추가적으로 논의
* 데이터셋과 모델 추가적으로 개선
* 성능지표 학습 및 가장 best 모델 선정