

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 | 2022-05-20 ~ 2022-05-26 |
| Project Manager of Reporting Week | 이인석 |
| Faculty Supervisor | 유창현 |

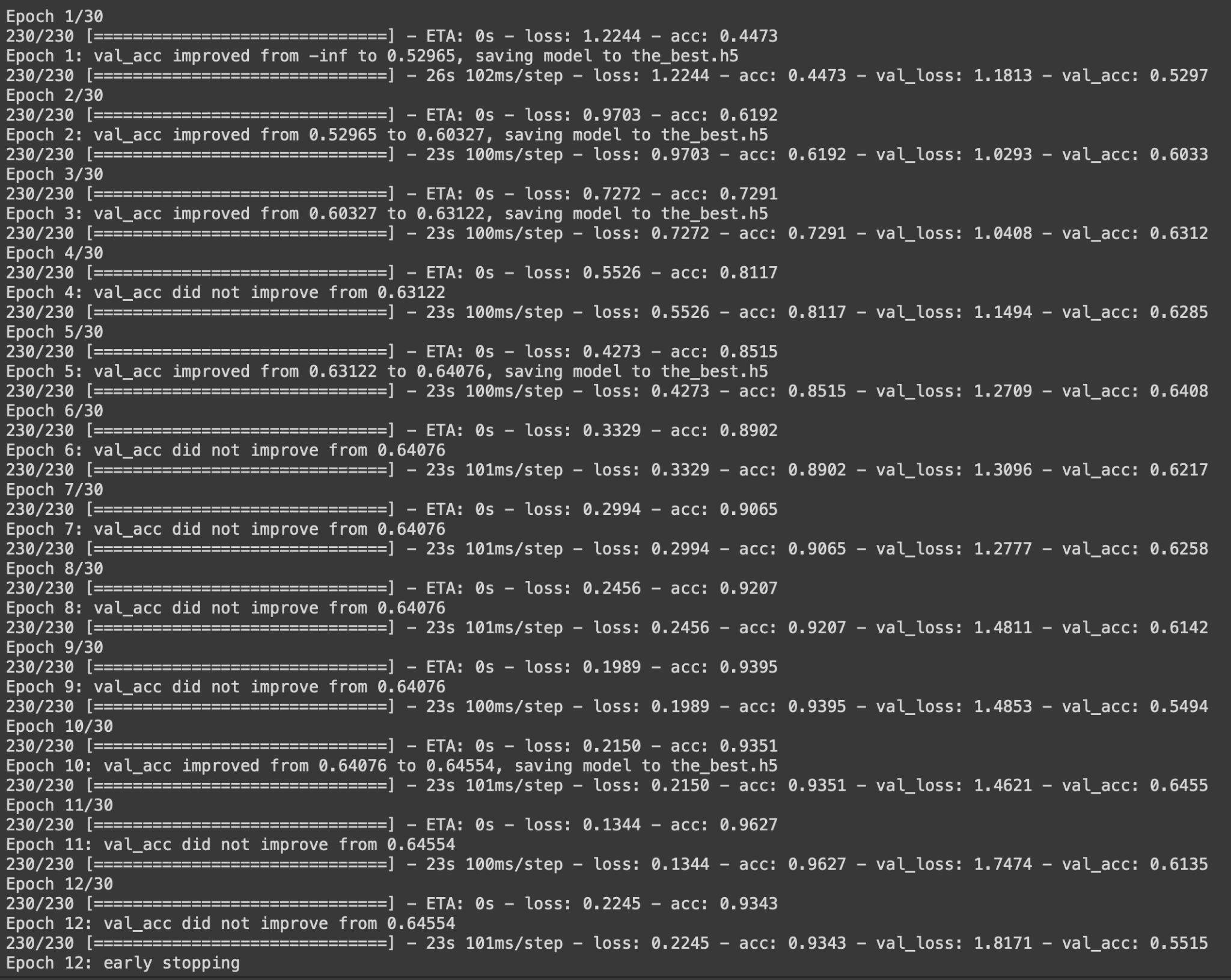
# **Tasks Outlined in Previous Weekly Progress Report** (Provide detailed information on the tasks to be completed in this week)

* train, validation, test dataset 비율 결정 및 분할 (train:val:test = 5:2:3)
* 4가지 딥러닝 모델(LSTM, bi-LSTM, GRU, CNN)성능 개선 및 평가

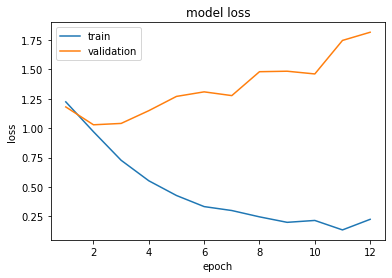
# **Progress Made in Reporting Week** (Provide detailed information on the progress that you made in the reporting week. Limit your write-up to no more than two page)

* LSTM

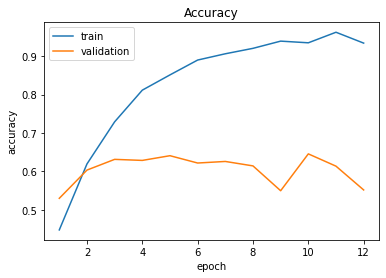
대부분의 모델에서 비슷한 test 정확도가 나왔으며, 가장 높은 정확도를 보인 모델은 hidden units = 128, epochs = 30, batch\_size = 32, Dropout = 0.3 인 모델



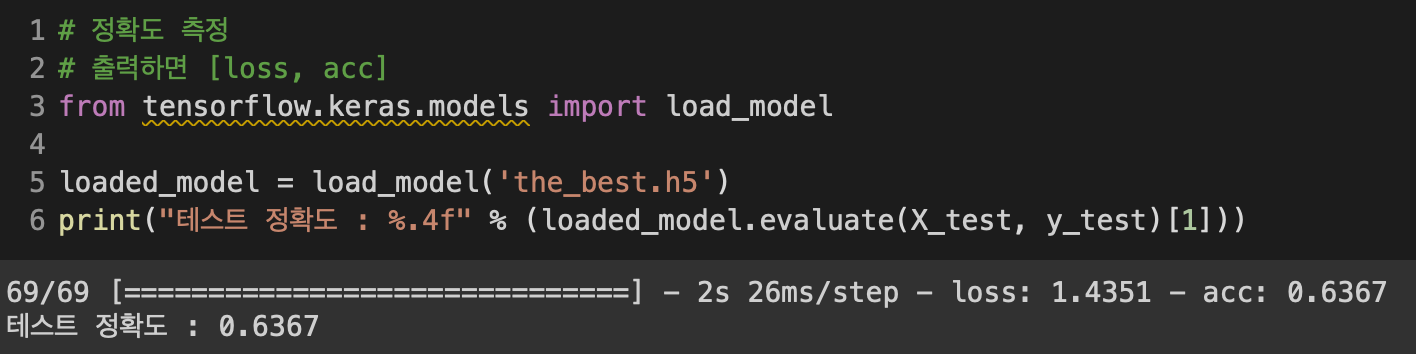
epoch 12 에서 early stopping



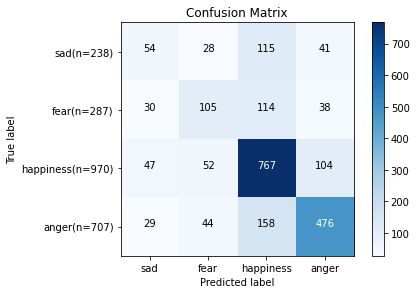
loss



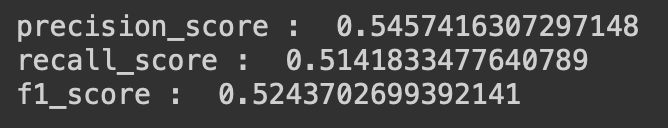
accuracy



test 정확도 : 0.6367

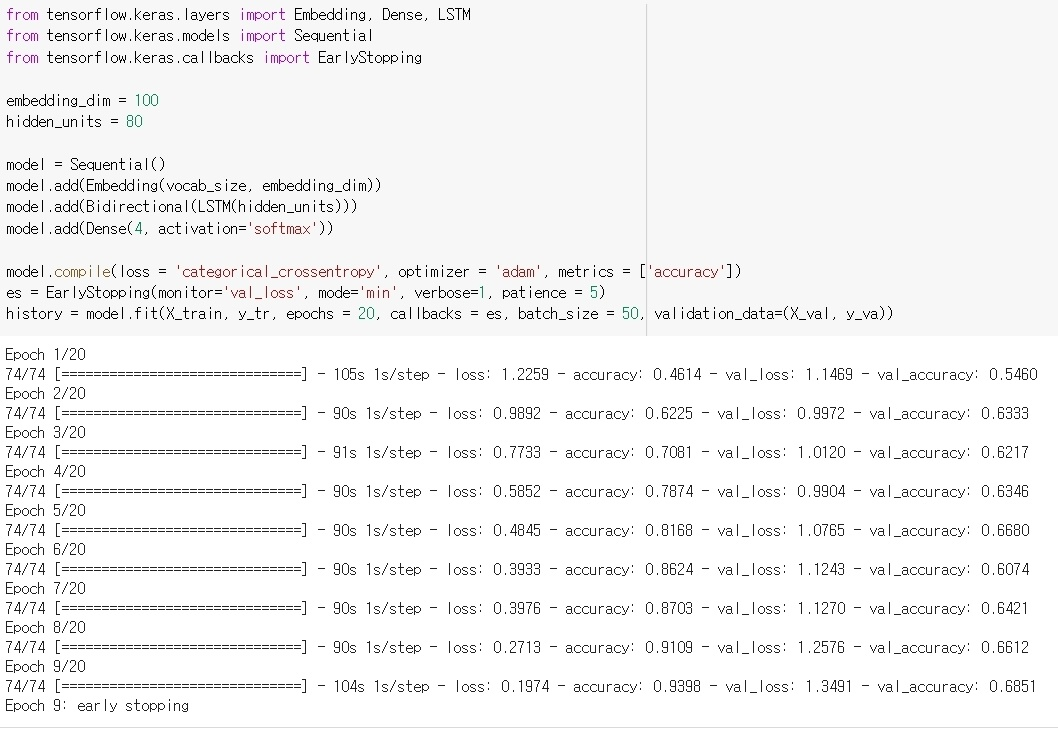


confusion matrix

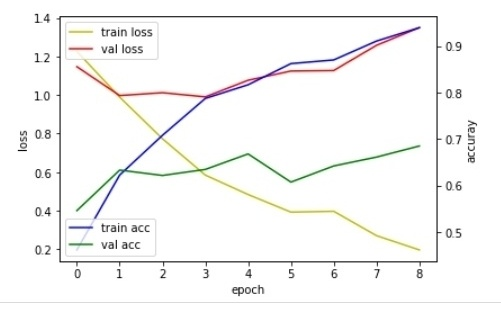


* bi-LSTM

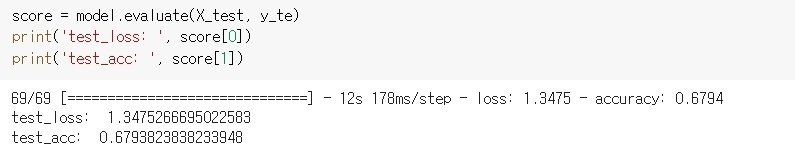
모델의 hidden units(70~200)의 개수, early stopping 옵션, epochs, batch\_size 등 다양한 파라미터를 바꾸어가면 모델을 실행해본 결과 가장 좋은 정확도를 보인 모델은hidden units = 80, epochs = 20, batch\_size = 50, early stopping 옵션을 적용한 경우였다.

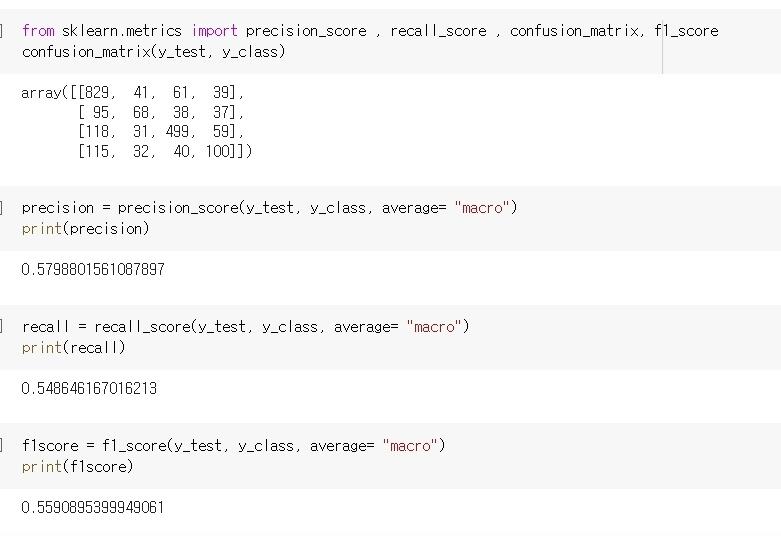


epoch 9에서 early stopping 되었으며, 정확도와 loss값의 그래프는 다음과 같다.



해당 모델에 test dataset을 입력하여 성능을 측정한 결과 정확도는 67.9%, loss값은 1.34, precision은 0.57, recall = 0.54, f1-score는 0.55였다.



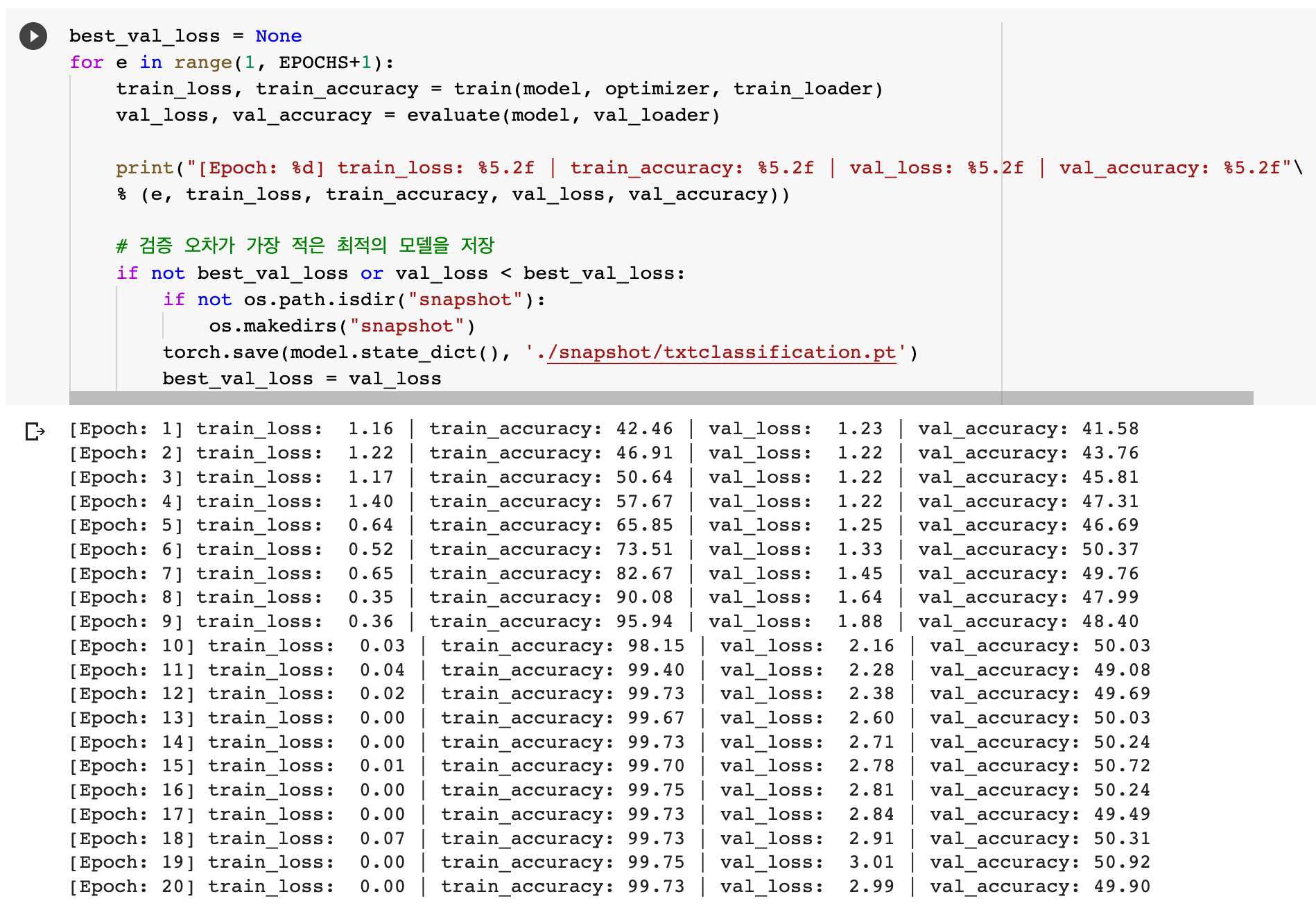


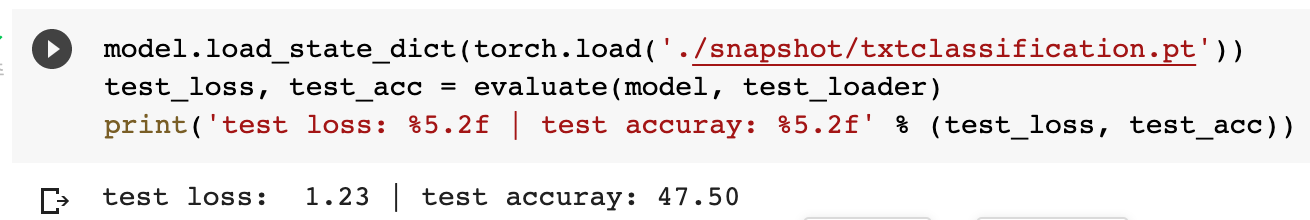
* GRU

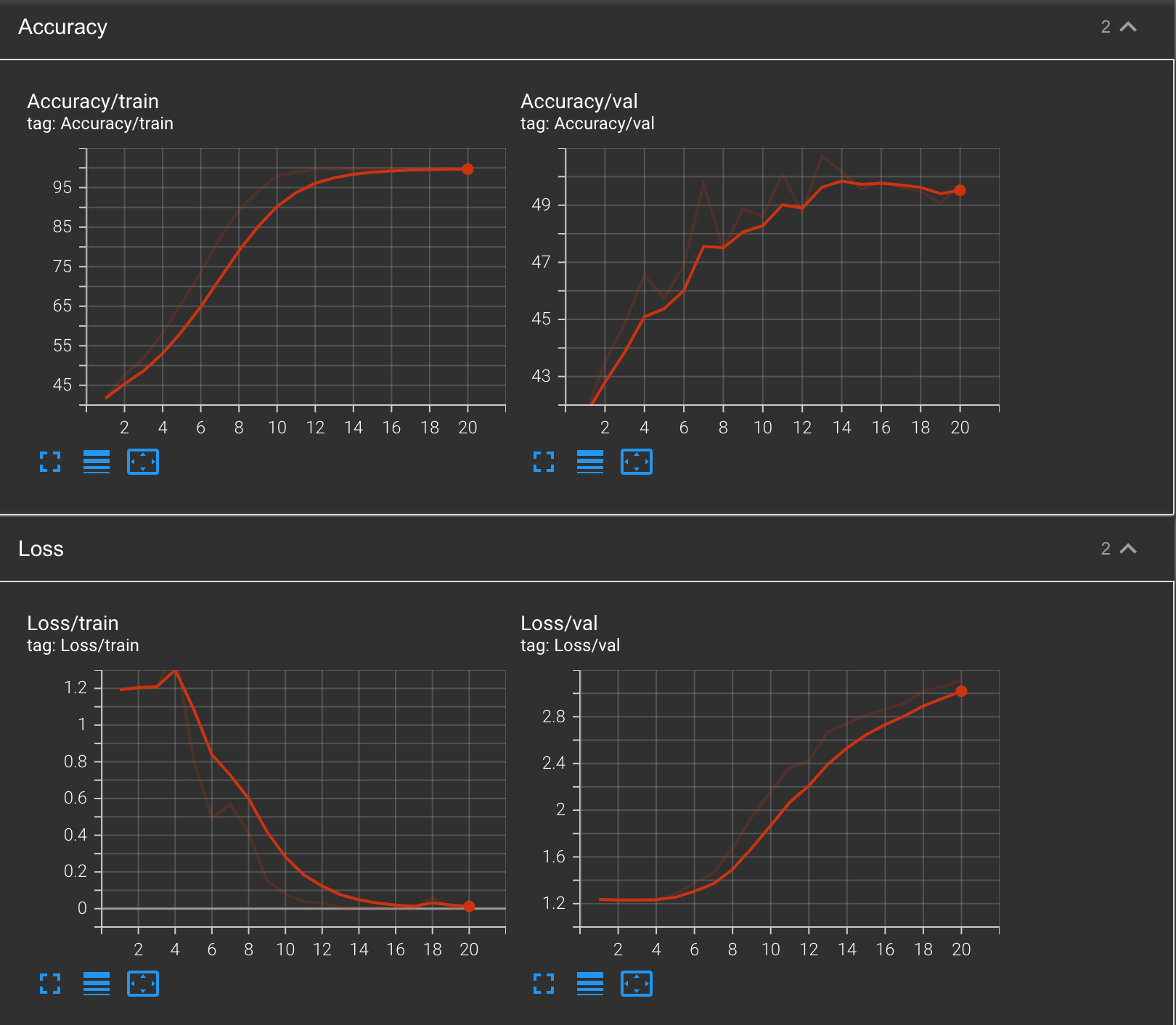
hidden\_units=128, embedding\_dim=128, drop\_out=0.5, batch\_size=64, epochs=20 일때 가장 좋은 성능을 보여줬다.



epoch별 train 데이터에 대한 손실값, 정확도는 그리고 validation 데이터에 대한 손실값, 정확도는 다음과 같다. 학습된 모델에 대한 test 데이터 셋의 정확도는 47%가 나왔다. 하이퍼 파라미터 조정을 통해 여러가지 경우의 수를 확인해 보았지만 50%를 넘는 사건은 없었다.





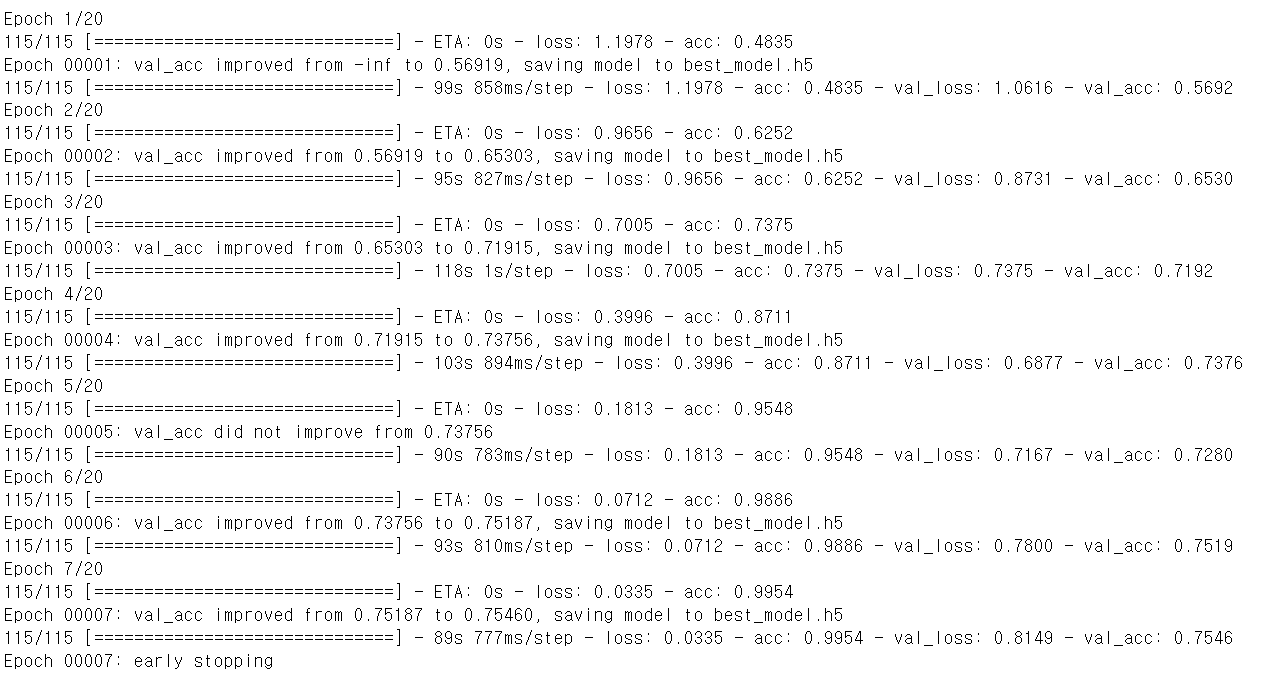


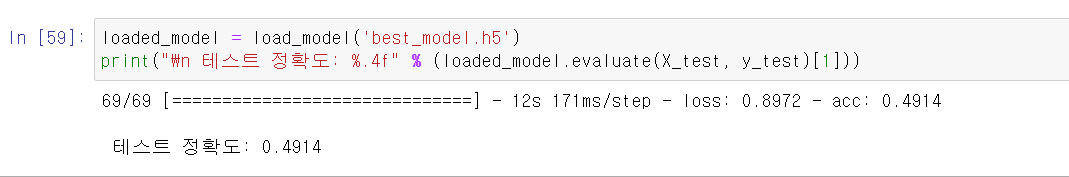
* CNN

아래와 같이 임베딩 벡터의 차원은 300, 드롭아웃 비율은 0.3, 커널의 수는 256, 커널의 크기는 3\*3, 뉴런의 수는 128개로 설정하였다.

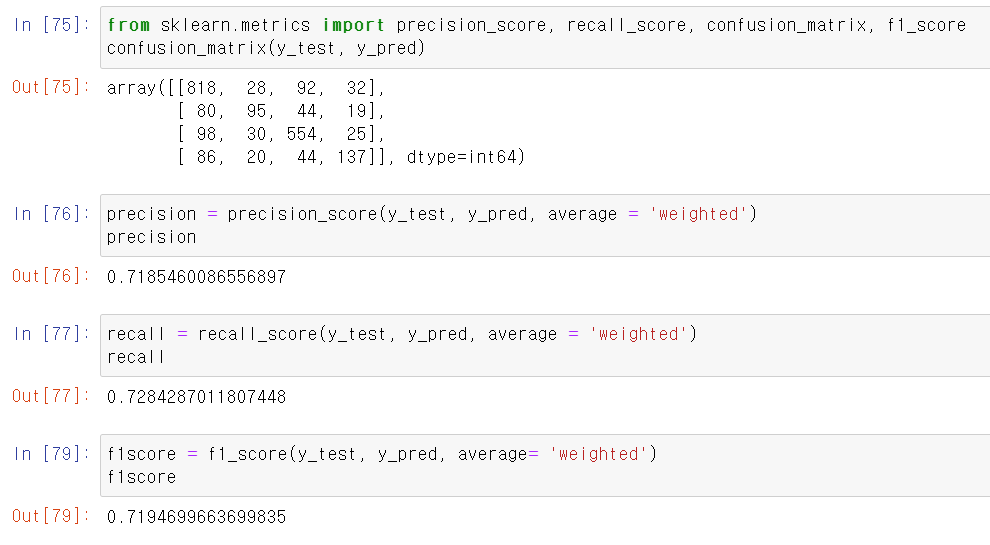
* 

training set 에 대한 accuracy는 87%, validation set에 대한 accuracy는 75% 이다.

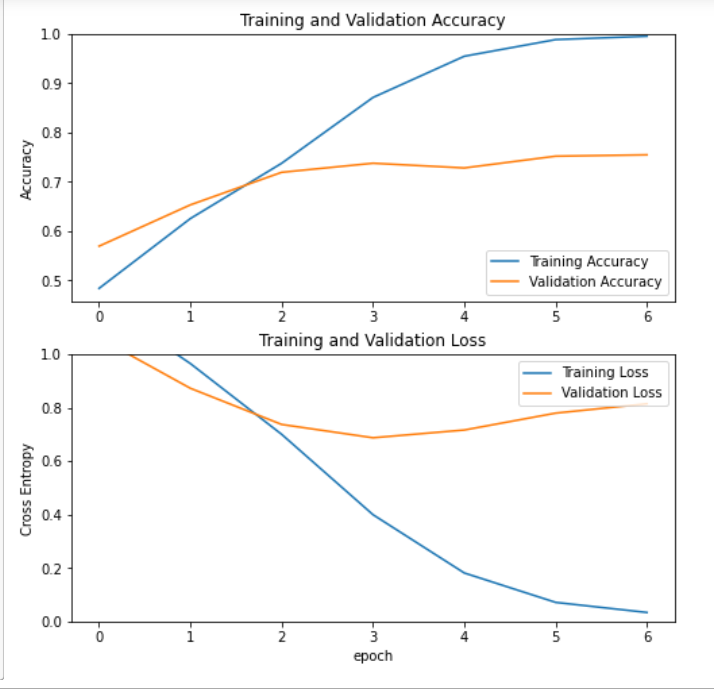




test set 에 대해서는 49% 의 정확도를 보인다.



precision, recall, f1 score 모두 0.71 정도의 값을 가진다.



# **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)

* 성능지표 학습 및 프로젝트에 가장 적합한 성능지표 선정
* 선정한 성능지표에 따른 가장 좋은 모델 선정
* 데이터셋 및 모델 추가적으로 개선