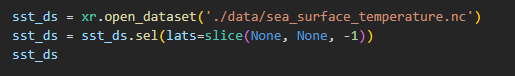
제**2**회 **Planet A** 지구환경과학 온라인 해커톤

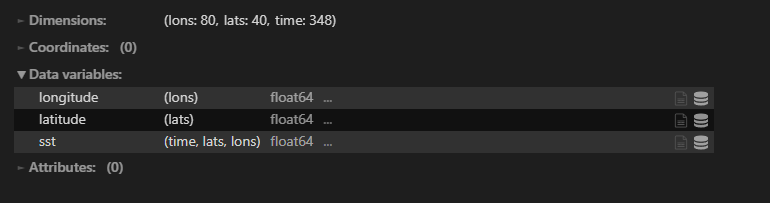
코드 설명자료

대회 참가자 **김준하, 박주현, 배병욱 2022**년 **09**월 **13**일

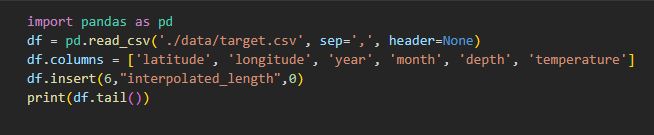
1. 라이브러리 및 데이터 **(Library & Data)**

● train data를 불러올 때에 nc file을 process하기 위해 Xarray 라이브러리를 사용





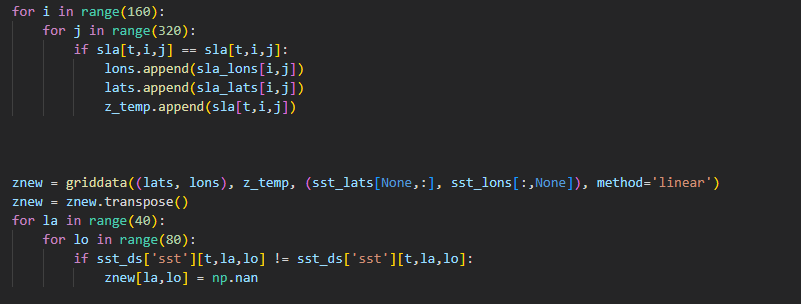
● target data를 불러올 때에는 csv file을 process하기 위해 pandas 라이브러리를 사용



1. 데이터 전처리 **(Data Cleansing & Pre-Processing)**

● sea surface temperature dataset(이하 sst dataset)과 sea level anomaly dataset(이하 sla dataset)에서 latitude, longitude의 좌표가 서로 다른 것을 확인함, 이를 더 균일하게 좌표가 설정된 sst datset의 좌표에 맞춰 sla dataset을 coordinate하기로 결정

● 기존에 존재하지 않는 좌표에 대해 sla dataset을 만들어야 하기에 이를 위해 scipy의 interpolation을 사용함. 이 때에 육지를 나타내는 Nan 값은 제외하고 interpolation을 진행함.

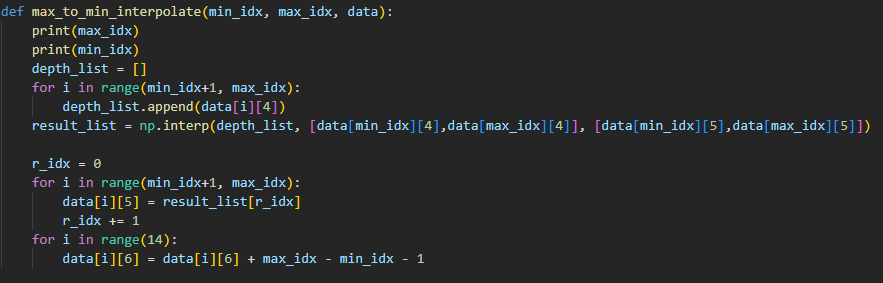
● 그 후 interpolation된 sla dataset에 대해 sst dataset에서 육지인 곳을 참고하여 해당 data를 다시 Nan 값으로 처리하는 과정을 거침. 

● target이 되는 수온 수직 분포 data의 경우 누락된 값이 많은 것을 확인함.

● 우선 해당 data를 년도별, 월별로 정렬 후, 누락된 값의 경우 excel을 통해 바다의 온도가 영하가 될 수 없다는 점을 참고하여 -1을 할당함.

● target의 수직 분포에서 중간에 누락된 값이 많아 loss 계산이 잘 되지 않을 것을 고려하여 해당 값을 interpolation 해주기로 결정함.

● 시작 지점인 0m에서 -1인 값을 제외하고 양 끝 지점에서 관측치가 있는 수온 분포에 대해 그 사이의 누락된 값을 numpy의 interpolation을 진행함.



1. 탐색적 자료 분석 **(Exploratory Data Analysis)**

● 해양 표층 수온의 data의 경향성을 확인하기 위해 분기별 sst data를 plot함

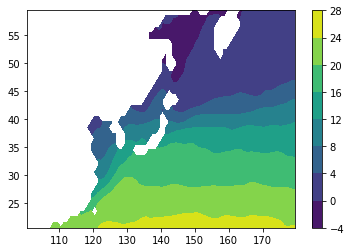
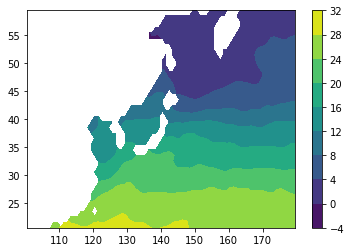
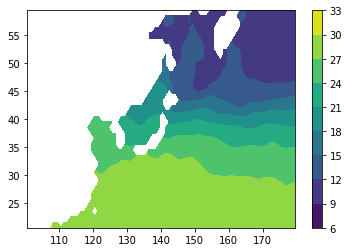
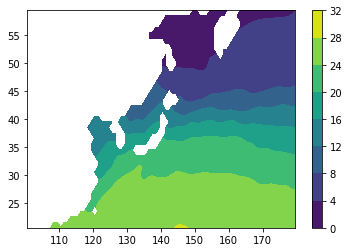
   

그림 1~4 1993년 2월, 5월, 8월, 11월의 표층 수온 데이터

● 해수면 높이 자료의 data의 경향성을 확인하기 위해 분기별 sla data를 plot함

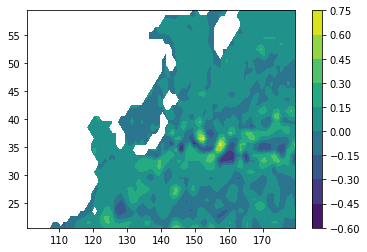
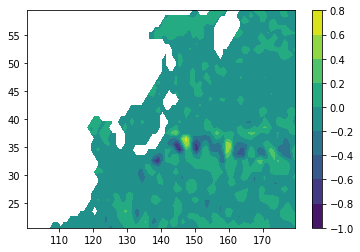
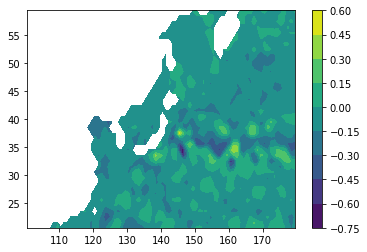
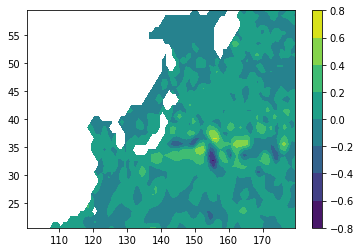
 

그림 5~8 1993년 2월, 5월, 8월, 11월의 해수면 높이 데이터

● 위의 데이터를 통해 해양 표층 수온, 해수면 높이 모두 위도, 경도, 날짜에 따라 크게 변한다는 경향성을 확인할 수 있었음.

● target data와 train data사이의 관계를 파악하기 위해 비슷한 sst data에 대해 다른 sla data를 갖는 지점의 수온 수직 분포와 비슷한 sla data에 대해 다른 sst data를 갖는 수온 수직 분포를 엑셀을 통해 확인한 결과는 아래와 같음.

그래프 1,2 유사한 sla, sst에 대한 수온의 수직 분포 그래프

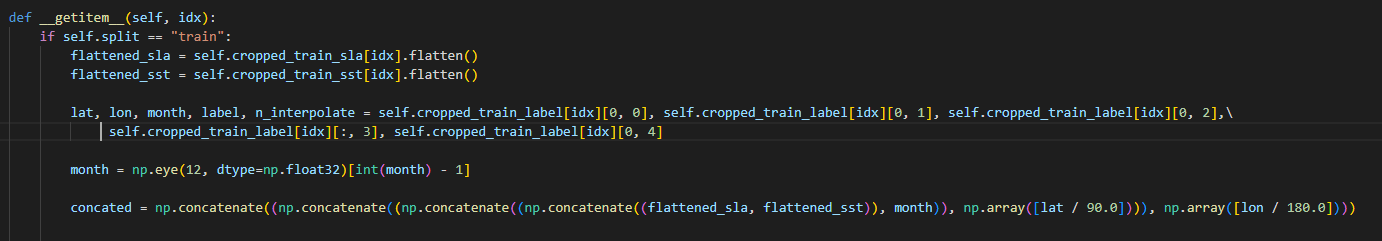
● sst의 경우 표층 수온이기 때문에 수온의 수직분포에서 시작점을 결정하며, sla의 경우 그래프의 기울기에 영향을 미치는 것을 알 수 있음.

● 하지만 target data에 결측치가 다수 존재하기에 각각의 sla, sst에 가중치를 다르게 설정하기에는 어려움이 있음

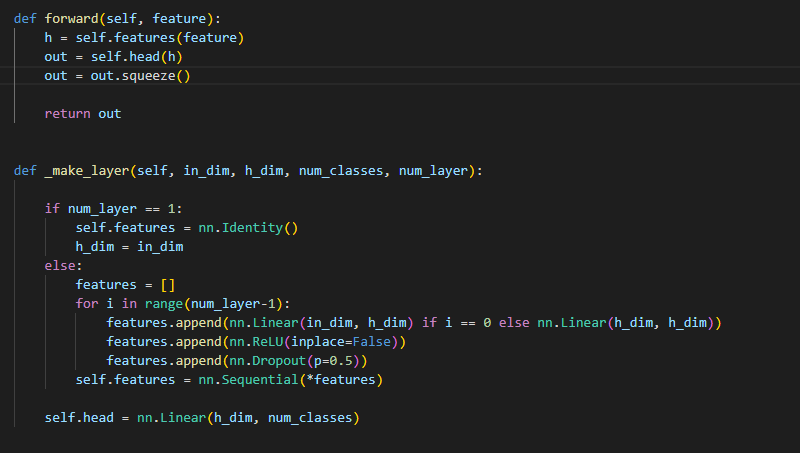
1. 변수 선택 및 모델 구축 **(Feature Engineering & Initial Modeling)**

● MLP와 CNN의 총 두가지 method를 시도해 보기로 결정

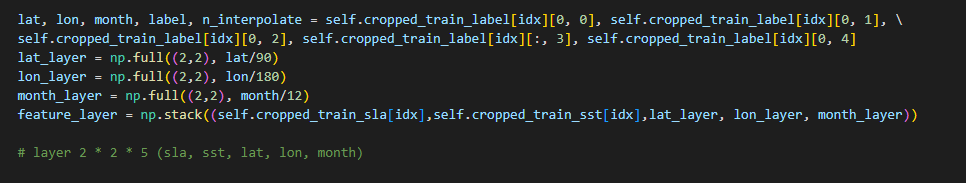
● MLP의 경우 feature는 표층 수온, 해수면 높이, 위도, 경도, 월을 feature로 사용하기로 결정, 각각을 flatten하여 vector로 만들어 주었으며 위도, 경도, 월의 경우 normalization을 거침. 각 feature vector를 concat하여 model에 feed함.



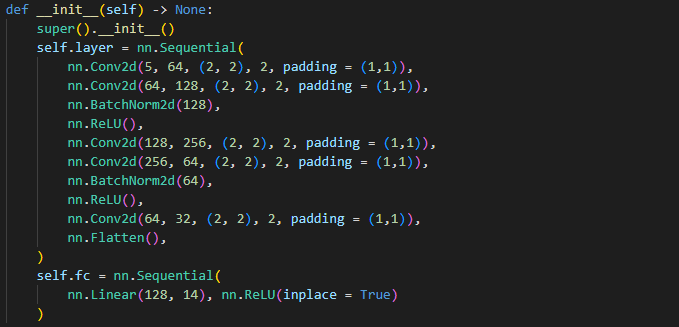
● Hidden dimension이 256인 2 layer MLP를 사용하였음. 마지막에 깊이 data인 14개의 vector가 출력되어야 하기에 Flatten한 후 linear layer를 거치도록 함.



● CNN의 경우 MLP와 마찬가지로 같은 feature를 사용, 각각을 2\*2 layer로 만들어 주었으며 위도, 경도, 월 layer의 경우 normalization을 거침. 해당 layer를 stack하여 2\*2\*5의 5개의 채널을 갖는 feature layer를 구성함.



● 2D convolution을 사용했으며 Batch normalization과 Relu activation function을 사용, 이 때에 layer의 크기가 2\*2로 작기 때문에 더 작아지지 않도록 padding을 통해 kernel size를 2\*2로 고정시키고, dropout 등의 방법은 사용하지 않음. 마지막에 MLP와 마찬가지로 Flatten한 후 linear layer를 거치도록 함.

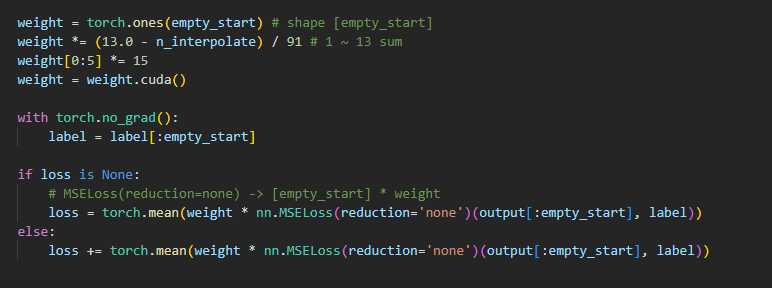


1. 모델 학습 및 검증 **(Model Tuning & Evaluation)**

● Optimizer는 Adam을 이용하였고, scheduler는 CosineAnnealingScheduler를 이용



● loss의 경우 Mean Squared Error loss function을 사용하였으며 label의 데이터 전처리 과정에서 interpolate된 길이가 작을수록 weight를 더 크게 주었음. 또한 처음 표층에 가까운 수온 분포에 대해 틀리게 될 경우 어긋나는 양상을 보였기에 처음의 포인트들에 더욱 가중치를 주는 방식을 택함.



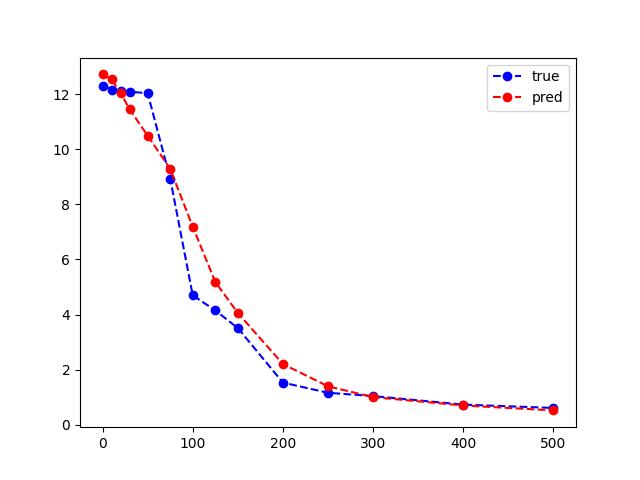
● 총 1000 epoch의 학습을 진행하며 best\_valid\_loss가 갱신될 경우 모델을 저장하는 방식을 택함.

1. 결과 및 결언 **(Conclusion & Discussion)**

● 학습 결과 MLP를 사용한 model의 경우 6.6123의 validation loss를 가지며 CNN 을 사용한 model의 경우 3.8907를 갖는 것을 확인

● 학습한 model은 result 폴더 안에 pt 파일로 저장되어 있음.

● 더 낮은 validation loss를 갖는 CNN model에 대해 true label과 prediction label에 대한 plotting을 진행한 결과는 아래와 같음.



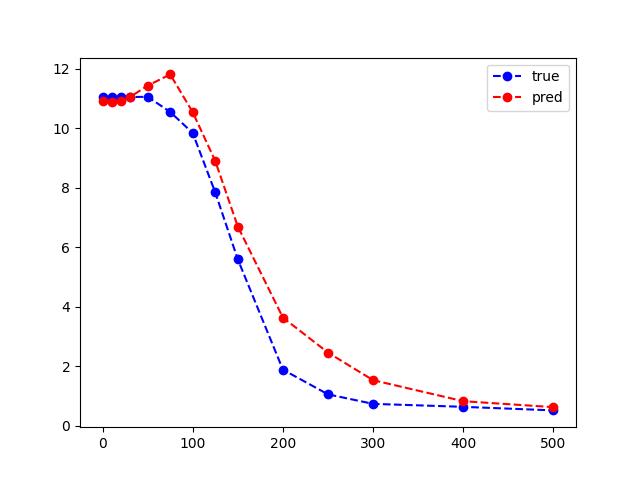


그림 9,10 interpolate되지 않은 true label에 대한 prediction plot

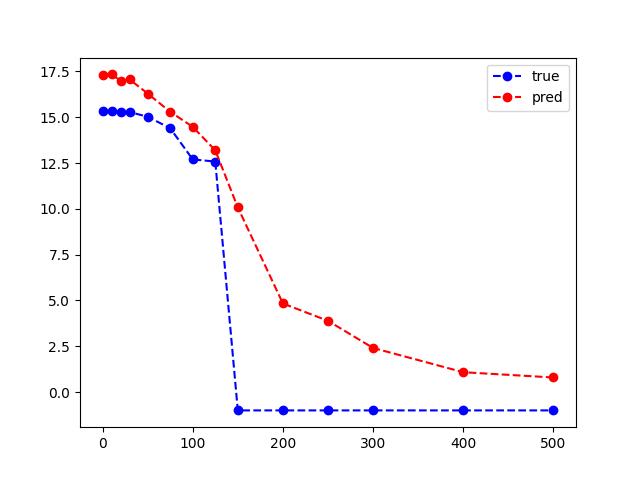
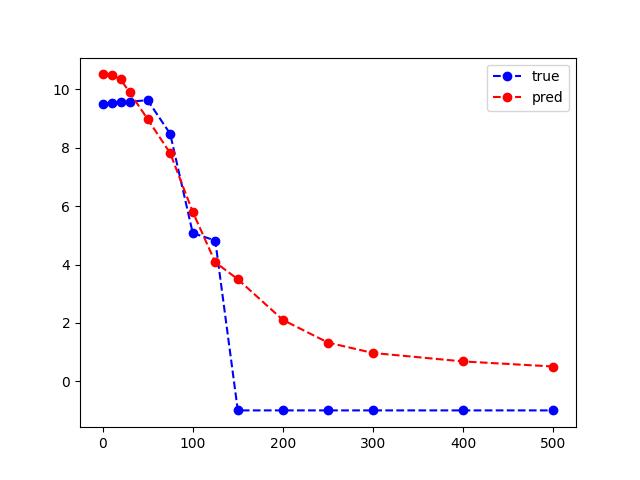


그림 9,10 결측치가 6개인 true label에 대한 prediction plot

● intepolate되지 않은 true label에 대해서는 어느 정도의 경향성을 잘 따라가고 있음을 알 수 있음.

● 결측치가 존재하는 true label에 대해서는 결측치가 존재하기 전까지의 label에 대해 robust하게 model을 예측하고 있는 모습을 확인할 수 있음.

● 해당 model을 fitting하는 데에 있어 결측치가 가장 큰 문제점을 야기함. 더 accurate한 target data가 존재한다면 CNN이 더 잘 fitting될 수 있을 것이라 예상 가능함.

● 아울러 train data와 target data의 학습을 위해 결측치를 linear interpolation을 통해 보간하였지만 더 나은 방법이 존재할 수 있음.