

I have included a screenshot of the data tab from the Kaggle competition page. You can see from this image that there are over 193,000 image files included in the training set. You can also see that the entire dataset is nearly 30GB in size.


Data Explorer
29.25 GB

- test_v2
- train_v2
- sample_submission_v2.c...
- train_ship_segmentation...


Summary

- 208k files
 - .jpg 208k
 - .csv 2
- 4 columns


< train_v2 (193k files)




00003e153.jpg
128.94 KB




0001124c7.jpg
76.06 KB




000155de5.jpg
147.63 KB




000194a2d.jpg
75.22 KB




0001b1832.jpg
95.63 KB




00021ddc3.jpg
242.91 KB




0002756f7.jpg
287.62 KB




0002d0f32.jpg
125.6 KB




000303d4d.jpg
205.59 KB



00031f145.jpg
232.9 KB



00052ed46.jpg
303.06 KB



000532683.jpg
166.85 KB

The approach I have taken to handle this very large amount of data is to store it in the Google Drive cloud storage service. This service interfaces directly with Google's CoLab service.