Machine Learning Engineer Nanodegree

Capstone Proposal: Icebergs vs Ships

Antoine Diffloth
November 6, 2019

Domain Background

My project proposal is based on the Statoil/C-CORE Iceberg Classifier Challenge on Kaggle[1].

The shipping industry is a key enabler of the modern global economy. It moved 10.7 billion tons of goods in 2017 and is expected to grow 3.8% annually between 2018 and 2023[2].

In some parts of the world's oceans, icebergs can disrupt commercial vessels by blocking shipping lanes and threatening physical damage to ships. Beyond the obvious collision risk, icebergs can cause delays and reroutes which are very costly to the shipping operators. As the climate warms and arctic ice shrinks, new trade routes will open up[3], but the risk of drifting icebergs will increase. Shipping companies will need better ways of tracking these building-sized chunks of ice.

The shipping industry uses various methods to detect and track icebergs, including airplane-borne and shore-based monitoring. Some parts of the ocean, however, are too distant from land to allow this, and satellite-based monitoring is the only option. The data generated from these satellites is not as easy to interpret as the optical images we're used to seeing but could prove very valuable in detecting problematic icebergs.

As a parent of young children, I feel a personal motivation to leave the planet to future generations in as good a condition as we found it, if not better. I believe that all research, no matter how small, that improves our understanding and raises awareness of climate related topics will improve the odds that we don't ruin this planet for our children and our children's children. Although this Kaggle challenge is

[1] www.kaggle.com/c/statoil-iceberg-classifier-challenge/overview
[2] United Nations Conference on Trade and Development, *Review of Maritime Transport 2018,* unctad.org/en/PublicationsLibrary/rmt2018_en.pdf
[3] Patel, J. (May 3, 2017) *As Arctic Ice Vanishes, New Shipping Routes Open,* www.nytimes.com/interactive/2017/05/03/science/earth/arctic-shipping.html

not directly related to climate change, it is related to our ability to measure changes to our planet from space, particularly related to icebergs and the warming planet.

Problem Statement

This challenge is a binary classification problem which consists of identifying whether a given satellite image contains an iceberg or a ship. We will be presented with a collection of images that are labeled as containing an iceberg or not containing an iceberg. We will train a machine learning model on this data, then use the trained model to classify a different, unlabeled collection of images based on whether we think it contains an iceberg.

If this approach proves successful, shipping companies could more easily protect their fleet from the threat of drifting icebergs.

Datasets and Inputs

The data sets are provided by Kaggle and can be downloaded here: https://www.kaggle.com/c/7380/download-all.

The data consists of 75x75 pixel images of the radar backscatter patterns from the Sentinel-1 satellite. There are two bands of data provided, one for the horizontal polarization and one for the vertical polarization. In addition to the backscatter data, the angle of incidence of the radar beam to the object is provided.

There are 1604 images in the training data set and 8424 images in the test data set. The data is provided in json format.

The training data has an additional feature “is_iceberg” which is 0 if the image is a ship and 1 if it is an iceberg. This is the dependent (target) variable of the study.

Solution Statement

The solution I’m proposing is to train a convolutional neural network to learn the difference between ships and icebergs. CNNs are a commonly used in computer vision and image classification tasks. The NN basis of CNNs allow them to model highly complex and non-linear relationships. The convolutions allow the model to look at the data surrounding each pixel, which enables the identification of more and more complex structures within the image.

Although we’re dealing with radar backscatter images rather than visible light images, CNNs seem like they are still an appropriate solution. There may be some additional processing required to adapt radar data to optical image manipulation tools.

Benchmark Model

I will use the median score of the Kaggle public leaderboard as my benchmark. The Kaggle community has many highly skilled data scientists, but if I aspire to join their ranks, I think beating half of them is a reasonable goal. At the time of this writing, there were 3339 teams on the leaderboard and the 1670^{th} team has a log loss score of 0.19387. This is the score I will aim to beat.

Evaluation Metrics

In order to compare to the Kaggle leaderboard, I will use log loss as the primary evaluation metric.

Since this is a classification problem, I will also calculate precision, recall and F1 score in order to understand the kinds of errors that my model is making. I think looking at recall is relevant here since correctly identifying all the icebergs in the test set is an important goal.

Project Design

The first task will be to explore the data. I will look for missing data, imbalanced classes, examine the summary stats of the data and generate some visualizations to understand the structure of the images and identify patterns.

Next I will perform pre-processing of the image data. From looking at a few of the images in the data set, there is a lot of noise, so I will have to apply a de-noising filter. Since there are only 1600 records in the training set, I will apply some data augmentation techniques to the images in order to get more data to train on.

Finally, I will train a CNN on the training images and make classification predictions on the test images. I plan on using Keras with Tensorflow. I will take a transfer learning approach in order to shorten training time and take advantage of the work that has been put into networks like VGG16 and Inception.