Machine Learning Engineer Nanodegree

Capstone Proposal

Bhavani Chillapalli
August 1st, 2018

Kaggle's TGS Salt Identification Challenge

Domain Background

Several areas of Earth with large accumulations of oil and gas also have huge deposits of salt below the surface.But unfortunately, knowing where large salt deposits are precisely is very difficult. Professional seismic imaging still requires expert human interpretation of salt bodies. This leads to very subjective, highly variable renderings. More alarmingly, it leads to potentially dangerous situations for oil and gas company drillers.

Seismic data is collected using reflection seismology, or seismic reflection. The method requires a controlled seismic source of energy, such as compressed air or a seismic vibrator, and sensors record the reflection from rock interfaces within the subsurface. The recorded data is then processed to create a 3D view of earth's interior. Reflection seismology is similar to X-ray, sonar and echolocation.

A seismic image is produced from imaging the reflection coming from rock boundaries. The seismic image shows the boundaries between different rock types. In theory, the strength of reflection is directly proportional to the difference in the physical properties on either sides of the interface. While seismic images show rock boundaries, they don't say much about the rock themselves; some rocks are easy to identify while some are difficult.

The following paper provides a deeper dive into the domain:
https://www.iongeo.com/content/documents/Resource%20Center/Articles/INT_Imaging_Salt_tutorial_141101.pdf

Problem Statement

The main problem to be solved is to identify if a subsurface target is a salt.There are several areas of the world where there are vast quantities of salt in the subsurface. One of the challenges of seismic imaging is to identify the part of subsurface which is salt. Salt has characteristics that makes it both simple and hard to identify. Salt density is usually 2.14 g/cc which is lower than most surrounding rocks. The seismic velocity of salt is 4.5 km/sec, which is usually faster than its surrounding rocks.

This difference creates a sharp reflection at the salt-sediment interface. Usually salt is an amorphous rock without much internal structure. This means that there is typically not much reflectivity inside the salt, unless there are sediments trapped inside it. The unusually high seismic velocity of salt can create problems with seismic imaging.

Datasets and Inputs

The data is a set of images chosen at various locations chosen at random in the subsurface. The images are 101 x 101 pixels and each pixel is classified as either salt or sediment. In addition to the seismic images, the depth of the imaged location is provided for each image. The goal of the competition is to segment regions that contain salt.

Solution Statement

A CNN such as a U-Net can be used to perform Image Segmentation on Keras.After resizing the image to fit the U-Net model, the next step will be to build the neural network and perform hyper parameter tuning. As i see it, the most challenging aspect of this project will be to arrive at the right combination of neural network layers and activation functions.
I will also be experimenting with various Transfer-learning models to achieve the best performance.

Benchmark Model

This basic CNN achieves a Kaggle score of 0.23.
It is a baseline model forked from this kernel :
https://www.kaggle.com/christofhenkel/keras-baseline/data

My goal will be to improve the accuracy of salt detection by enhancing and optimizing this benchmark model.

Evaluation Metrics

This competition is evaluated on the mean average precision at different intersection over union (IoU) thresholds. The IoU of a proposed set of object pixels and a set of true object pixels is calculated as:

$$IoU(A,B) = \frac{A \cap B}{A \cup B}.$$

The metric sweeps over a range of IoU thresholds, at each point calculating an average precision value. The threshold values range from 0.5 to 0.95 with a step size of 0.05: (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). In other words, at a threshold of 0.5, a predicted object is considered a "hit" if its intersection over union with a ground truth object is greater than 0.5. At each threshold value t, a precision value is calculated based on the number of true positives (TP), false negatives (FN), and false positives (FP) resulting from comparing the predicted object to all ground truth objects:

$$\frac{TP(t)}{TP(t)+FP(t)+FN(t)}.$$

A true positive is counted when a single predicted object matches a ground truth object with an IoU above the threshold. A false positive indicates a predicted object had no associated ground truth object. A false negative indicates a ground truth object had no associated predicted object. The average precision of a single image is then calculated as the mean of the above precision values at each IoU threshold:

$$\frac{1}{|thresholds|}\sum_t \frac{TP(t)}{TP(t)+FP(t)+FN(t)}.$$

Lastly, the score returned by the competition metric is the mean taken over the individual average precisions of each image in the test dataset.

Project Design

1. Data Exploration
2. Data Visualization
3. Model Building
4. Model Evaluation : Comparing the results of the Benchmark Model and new model.
5. Hyperparameter Tuning

****References****

Kaggle Competition : https://www.kaggle.com/c/tgs-salt-identification-challenge