

capstone_proposal

August 15, 2018

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Shane Moloney
August 13th, 2018

1.2 Proposal

(approx. 2-3 pages)

1.2.1 Domain Background

(approx. 1-2 paragraphs)

I found this project on Kaggle, the domain is underground probing by seismic imaging and the processing of the resulting images to find salt deposits. Areas of Earth with large quantities of oil and gas also contain large deposits of salt, detecting these in the seismic images currently requires experts to manually analyse all the images. This can be very subjective, varying conclusions, making it very difficult to find the exact position. Drilling an incorrect position can cost huge amounts in wasted funds and can result in dangerous work conditions for the drilling crew.

TGS is hoping to use machine learning to find the deposits accurately and have provided a large dataset of thousands of image samples and masks containing the salt deposit locations. This can be solved with the use of a CNN (Convolutional Neural Network), and this is the approach I plan to use. All datasets and requirements can be found at the [TGS Salt Identification Challenge](#) kaggle page.

1.2.2 Problem Statement

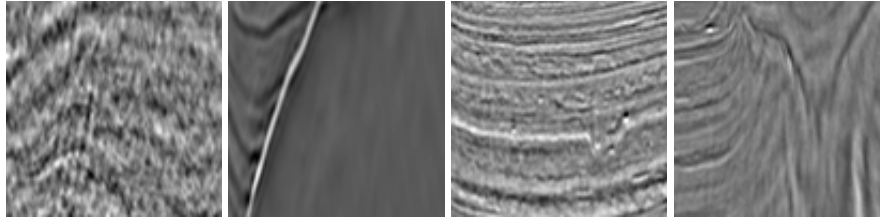
(approx. 1 paragraph)

The exact problem is to take in a seismic image of a cross-section of an area underground and to produce a mask of said image where white pixels represent salt deposits and black pixels represent other rocks and minerals. The difficulty with this problem is that of all image recognition of natural phenomena challenges, there are no straight lines in nature, the model must learn to recognise patterns based on a wide variety of images where no two deposits are the same. This makes the problem much more difficult than man-made object detection.

1.2.3 Datasets and Inputs

(approx. 2-3 paragraphs)

The images themselves are greyscale and the pattern that you see is most earth is layered and so appears stripped, whereas the salt deposits tend to appear either smooth or chaotic. There are 18,000 images in the training set and 8000 images in the testing set. Below are some examples from the training set:



As you can see this is the perfect problem for a CNN due to the high reliance on finding edges both vertical and horizontal. The submission for the competition is to use run-length encoding, meaning instead of submitting a folder of mask images the submission is a CSV file where the first column is the image name and the second is the range of the white pixels in the mask, ie. example 1 3 10 12 would mean the image named example contains a salt deposit at pixels 1, 2, 3, 10, 11 and 12. The provided masks are also in this for so this requires some pre and post processing.

1.2.4 Solution Statement

(approx. 1 paragraph)

The solution is clearly to use a CNN to recognise salt deposits. The model may not even need to be incredibly accurate, ie finding the exact dimensions and borders of every deposit, instead it could simply need to flag an area that may contain a deposit and then a human expert can be brought in to define the exact position for drilling to occur.

1.2.5 Benchmark Model

(approximately 1-2 paragraphs)

Currently there are not many models in use for salt deposit identification, there are some models in use for general seismic imaging but the difficulty is getting data sets large enough to accurately train a model. While there are plenty of images out there accurately labelled features require an expert in the field and as stated previously this is a time consuming process and so makes it difficult to build up data sets for the sector.

Most recently one Anders Waldeland published his thesis on salt classification using deep learning, where he used a CNN to accomplish a similar task to this one. His model was designed to take a single, feature labelled slice of a 3D seismic scan and use it to detect all salt deposits within that specific scan. This would allow the experts to only have to find a salt deposit in a single slice of a 3D scan and then use the model to classify the rest. However the problem we are presented with here is to detect salt in any seismic image in a much more general dataset meaning the model needs to be more thorough and complex. Unfortunately this model does not use the same data set as the competition provides, meaning we can't use this as a direct benchmark.

While there is no benchmark model provided, a group has created an open source project for this competition to be used as a benchmark or as baseline code to be used as a space for experimentation or just to be used as a starting off point. The project is being run by a group called neptune.ml and can be found on their site [here](#). I think this will make a good benchmark

as it implements a Unet as well and is not like other baseline code examples out there as the goal for the project is to use an open source platform to work on the challenge together so it does not stop at the most basic implementation and uses augmentation and other methods to improve the score.

1.2.6 Evaluation Metrics

(approx. 1-2 paragraphs)

The evaluation for this challenge is done using the mean average precision at different intersection over union thresholds. This is done by first getting the prediction image A and the ground truth image B and calculating their intersection over their union:

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

The IoU is calculated based on a threshold, ie. if the threshold is 0.5 the IoU has to be over 0.5 to be considered a hit. The thresholds in this case range from 0.5 to 0.95 with a step size of 0.05. For each threshold, t , we calculate a precision value using the number of true positives (TP), false negatives (FN) and false positives (FP). These are gathered when comparing the values of the prediction and the ground truth:

$$\frac{TP_t}{TP_t + FP_t + FN_t}$$

The average precision value for a single image is calculated by the mean of the precision values for each IoU threshold calculated above, so the equation looks like this:

$$\frac{1}{thresholds} \sum_t \frac{TP_t}{TP_t + FP_t + FN_t}$$

Finally the models score is the mean of all the average precision values of all the images in the test set.

1.2.7 Project Design

(approx. 1 page)

I plan to use a CNN as this is the perfect problem to apply such a model to. To ensure exploration The data will likely require augmentation on the data, but only by flipping on the horizontal axis as the orientation of the layers in the earth is a key feature of the images. I will also apply dropout to allow for full exploration of the features.

For specifications on the CNN I am going to attempt a Unet implementation, as it is a relatively new CNN framework that has recently been performing very well in competitions and applications. The basis is to reduce the layer size exponentially and similarly increase the channel count until the layer size reaches some minimum, ie. 32x32, and then to use Conv2DTranspose layers to exponentially increase the layer sizes and reduce the channel count to the original size. This allows for fast, accurate image segmentation and has been used for many industries and areas of study. A diagram of a Unet architecture that was used for [Biomedical Image Segmentation](#) can be seen below:

[unet] [unet]: u-net.png

The data itself will need to be preprocessed as a Unet implementation requires the image dimensions be divisible by two. I will need to apply the same to the masks as well as loading them

in and creating image arrays from them as they are provided in the run-length encoded form described above. I will also need to do some exploration of the data to expand on my process further.

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?