Machine Learning Engineer Nanodegree
Capstone Proposal

Irina Gruzinov

January, 5, 2020

1 Domain Background

Seismic data acquisition and processing is an important and very sophisticated tool in oil and gas exploration. Seismic data acquisition involves generating a sharp sound (a shot) at a location (seismic source), and recording the reflected waves with geophones (receivers) placed in multiple locations around the source. Then the source and the receivers are relocated and the process is repeated. A survey can contain thousands of such shot records. It's desirable that the sources and receivers are placed on a regular grid, but it's never possible in real world. Typical distance between shots is 37.5 meters and that between receivers is 25-200 meters. The duration of the recorded seismograms can be 12 seconds. Such a duration permits recording seismic waves reflected from a subsurface layer at 10 km depth. https://en.wikipedia.org/wiki/Reflection_seismology

2 Problem Statement

During seismic data processing, recorded seismic signals are used to reconstruct the subsurface Earth structure that would explain observed records. https://wiki.aapg.org/Seismic_migration Many algorithms and methods utilized in seismic data processing require data on a regular grid. The field data are never on a regular grid and often incomplete, with some or many traces missing, because, for example, some locations were not accessible during acquisition (no permit to a location, or buildings and obstacles, difficult terrain and other reasons). The problem of regularization, that is, interpolation of traces into a regular grid, and a related problem of interpolating missing traces is very challenging. The signal varies rapidly from one trace to another and simple interpolation never works. The problem of regularization is very important, and there are very complicated methods that attempt to reconstruct the waveforms and interpolate along these forms.

3 Solution Statement

In this project, I will approch the problem of interpolation of missing traces on seismic data with Machine Learning tools. I will apply a special architecture of a Deep Convolutional Neural Network that was successfully used to denoise and to upsample images. https://ieeexplore.ieee.org/document/7839189 The special features of this architerure is that there is no pooling layers, and it utilises batch normalization (BN) at every hidden layer. This BN prevents the problem of vanishing gradients and also accelerates the learning. This allows to have a deep CNN with wide perception area capable to learn large-scale structures on the input images that span the whole image. This is important for seismic images. Another interesting property of this architecture is that since there are no pulling layers, and feature maps are of the same size as the input. It might be interesting to pull this feature maps at different depth and try to interpret them.

4 Datasets and Inputs

Field seismic data are very expensive to acquire and companies don't share them. The Netherlands government made public a large 3D survey of North Sea in Europe. These data are extensively used in the geophysical community and in universities.
https://terranubis.com/datainfo/Netherlands-Offshore-F3-Block-Complete

From the North See data, the OLIVES lab at Georgia Institute of Technology prepared two datasets suitable for ML studies. These datasets are 2D seismic slices taken from a processed 3D seismic volume. There are 18000 images of size 99×99, and 4000 images of size 300×150, see Figure 1. These datasets are publicly available upon request through their web site.
https://github.com/olivesgatech/LANDMASS/

I decided to use these datasets because (i) they are from real field data, and are more complex than synthetic data; (ii) they are processed, i.e. clean and balanced, so I can concentrate on ML task in my feasibility study.

Real seismic data are intrinsically 3D, and using only 2D slices is a major simplification of this study which allows me to apply "of the shelf" DCNN architecture proven working well on images.

5 Benchmark Model

I am not aware of any application of Machine Learning to seismic data interpolation that is publicly accessible. I will use the interpolation by SciPy driddata module as benchmark for this study.
https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html

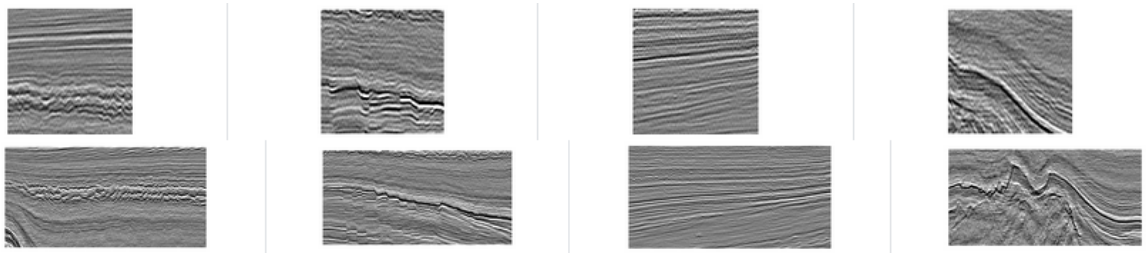

Figure 1: Examples of 99×99 images and of 300×150 images from two datasets

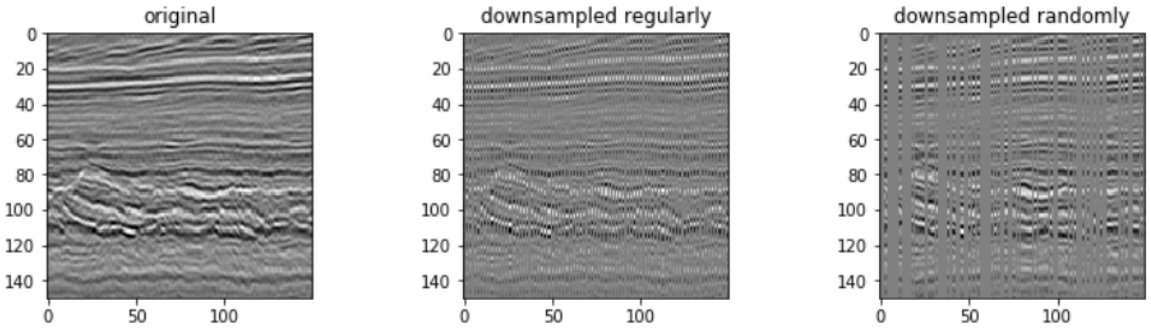

Figure 2: Input generation, 50% downsampling example

6 Evaluation Metrics

The ultimate metrics is to compare the restored seismic sections to the original ones.

7 Project Design

I will be using the first dataset with 18000 99×99 seismic images. To generate an input dataset, I will remove part of the traces regularly or randomly, see Figure 3. I will test removing 30%, 40%, 50% of the traces. Downsampled images will be the input to the DNN, and the output produced will be compared to the original "complete" images to calculate the loss function. I will use a repo for the aforementioned paper and utilize their implementation of the DCNN, but keeping number of layer as a hyperparameter. I will train for a few epochs testing different numbers of layers to see if I am getting anywhere. If I cannot get a good result with interpolation, I will try to reproduce the paper's denoising result.