Skoltech

Skolkovo Institute of Science and Technology

Introduction

Hosam Asaad Taha

Detect acoustic waves in real time from microseismic

Intro

Micro-seismic monitoring hydraulic-fracturing. While fracturing the rock micro-seismic events happen to emit comprised waves in form of “P-, S-“and other wave modes. Those waves can be detected by a group of “geophones” placed a few-hundreds of (m) away from the event locations. And by talking about making a real-time algorithm or code to detect P- and S- waves it will be like an instrument to help us to take (real-time drilling and stimulation decisions). As Micro-seismic event will be easily located, determine the fracture growth, and possible drilling hazard. Now a days, the operation of picking P- and S-wave arrivals can be done by experienced analysts manually, or automatically picked by phase picking algorithms such as the (STA/LTA or Phase-Net). The arrival through those algorithms is not showing high accuracy as it takes 10’s of seconds for the earthquake signal to be transmitted from source to geophones while the hydraulic fracturing induced micro-seismic event in means of milliseconds to be transmitted to geophones. In simple words this time delay will cause miss location of the event as the delay of 10’s of seconds will cause miss location of 10’s of meters.

In this project, by building a machine learning code that relays on Random forest algorithm and two CNN (2D and 3D) we will try to pick P- and S- waves in real-time and after that we will measure the satisfaction of the readings and prediction to show the efficiency of our model then we will perform some test to make sure the model is robust enough.

Data-Acquisition and Representation

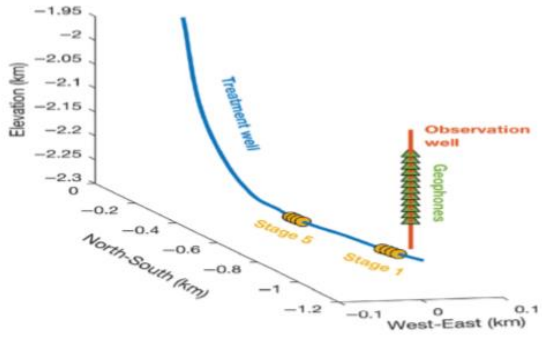

Figure 1 shows the way micro-seismic operation and survey setup.

The data is collected through 12 geophones place is all the four main directions. Each wave form is shown 2000 time-lap which is equal to 500 Ms. In 2D-CNN 3 dimensional collection reading is taken from the 12 geophones (2000, 12, 3) to make time series full vision. On the other hand, in the random forest models we treated the data as a single sample from each geophone (2000, 3) and we flatted it in the vector form of (6000, 1).

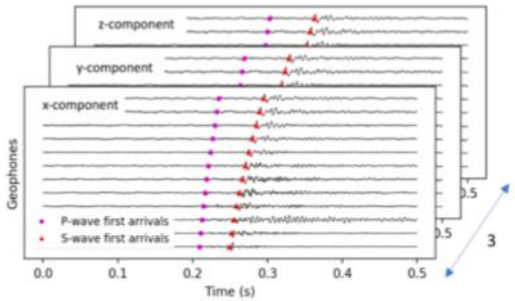

Figure 2 representation of a single sample in the time domain

Data Pre-processing

Waveform scaling of the amplitude in the means of -1 and 1 by dividing max/min amplitudes then we converted to time means to work over the time domain approach. Then we used Hamming filter with a window size “200”and a number of overlap of “195” as constant constrains proven to maintain the resolution of 0-500 Hz to maintain the wave bandwidth as a result the signal form mentioned before (2000,12,3) will be represented as time domain of (12,361,25,3).

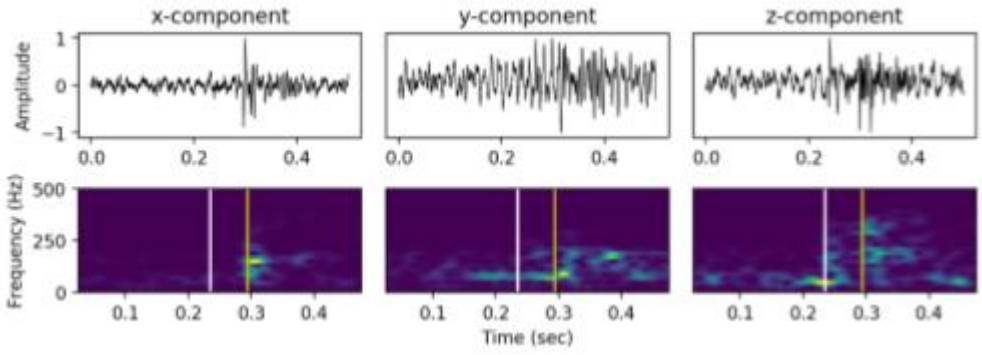

Figure 3 Time series and spectrograms

Methods

Random Forest Regression

Random forest a tree model based on multiple decision trees each one is trained over bootstrapped samples “raw data” while each training point is restricted to a random sub-feature if splitting a branch is done. In the prediction output we take the mean average as the final result of each tree. The choice of random forest was based on its high flexibility it does not require any assumptions and high running time. In addition, for over fitting problem is maintained by random sub-feature space splitting and data bootstrapping. And those are the assumptions we considered.

n_estimators = 100 | max_features = 80| min_samples_split = 30 |min_samples_leaf = 10

And it was considered as the baseline to compare with.

2D-and 3D-CNN

The use of the 2D was based on the relation of our input structure which is time and spatial so it will offer 2D scan over the input. In addition the 2D CNN prove the perfection over RGB. Then we added the 3D CNN to work back to back with 2D CNN to pick additional features along the spectrogram to offer better pattern recognition dataset. Both of them are built over the same architecture in which applying gradually increase in filters while filter size decease in the forward-direction. Secondly phase is (ReLu) activation function after each layer then (batch normalization) and sometimes “max-pooling and dropout layers” which was add also in 3D CNN to reduce high variance in the architecture while less layers was used. Adam optimizer also was used to fit the models to initial learning rate of 0.001 followed by linear decays after each epoch.

Sensitivity Analysis

"hyper-bolic pattern" was recognized by each arrival of P- and S- wave in the time domain along the geophones. So we got benefit of it by apply it to more application scenarios. And due to the hard environment of work we assumed the probability of a broken geophone. As mentioned our input is from all the geophones so what if one or more is missing so we put a case that our network could realize and detect the patter even if a geophone is broken. So by designing a case where there are one or more geophone to our training sample while testing were considered. By making something like a blind eye by blocking one geophone and set its signal to zero then also over three geophones to check the robustness of the model.

Results

The training was separately taken for P- and S- individually. Random forest was trained by 80% and tested to 20% of data. And for each CNN the train to test was (train-70 / test 15 / dev 15) to 120 epochs and a mini-batch size of 128.

We used mean square error as the “optimizing metric” as that’s because it a regression problem. Also we defined satisfaction levels (5 Ms, 10 Ms and 20 Ms) to show the accuracy of our prediction to show the deference rate of less than (5 Ms, 10 Ms and 20 Ms).

model	A5	A10	A20	TR-MSE	TT-MSE
RF_P	26%	46%	65%	6223	16004
2D_CNN_P	45	70	89	533	2082
2D_CNN_P_B1	42	67	87	437	2153
2D_CNN_P_B3	24	51	81	563	8282
3D_CNN_P	60	82	91	27	141
3D_CNN_P_B1	52	79	89	22	139
3D_CNN_P_B3	25	52	88	41	146

model	A5	A10	A20	TR-MSE	TT-MSE
RF_S	27%	48%	70%	5027	15360
2D_CNN_S	59	86	97	398	1179
2D_CNN_S_B1	56	87	96	826	1854
2D_CNN_S_B3	40	68	90	618	4699
3D_CNN_S	55	86	95	52	410
3D_CNN_S_B1	52	82	95	71	365
3D_CNN_S_B3	34	65	89	47	215

- “P" P-wave arrivals
- “S" S-wave arrivals
- “B" no. blocked geophones
- “TR” train
- “TT” test

By watching the results we can notice an over-fitting problem so I went for tuning of (max_features, min_samples_split and min_samples_leaf) for random forest model, and by adding more (maxpooling and dropout layers) for the CNN models but it also shows a difference which may be due to the size of samples “1660 micro-seismic event”. And that was the prediction of the CNN

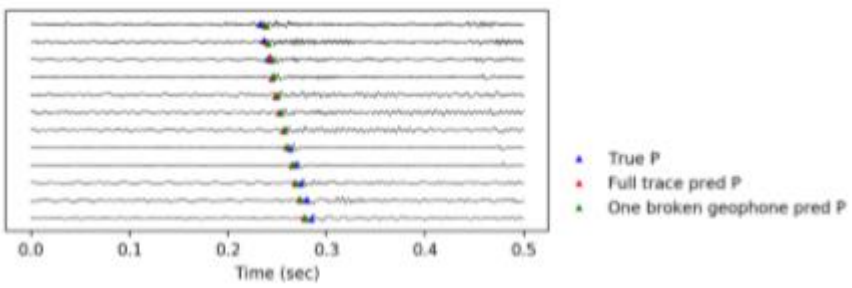

Figure4 S-arrival prediction

Conclusion

• We investigated three models for micro-seismic P- and S-wave entry picking. Both 2D- and 3D-CNN models outperformed the random forest model and accomplished exceptionally high prediction accuracy with A20 higher than 90%.

• The 3D-CNN beats the 2D-CNN in P-wave entry picking since the P-wave irritation is much less than that of the S-wave within the time-space.

• Both CNN models are vigorous indeed with one geophone completely lost. This is often because the models take advantage of the spatial course of action geophone.