Natural Resources Canada | Ressources naturelles Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 8877

Predictive models for first arrivals on seismic reflection data, Manitoba, New Brunswick, and Ontario

B. Patwa, P.L. St-Charles, G. Bellefleur, and B. Rousseau

2022

Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 8877

Predictive models for first arrivals on seismic reflection data, Manitoba, New Brunswick, and Ontario

B. Patwa[1,2,3], P.L. St-Charles[2], G. Bellefleur[3], and B. Rousseau[2]

[1]Université de Montréal, 2900 Edouard Montpetit Boulevard, Montréal, Québec
[2]Mila - Québec AI Institute, 6666 St-Urbain Street, Montréal, Québec
[3]Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario

2022

Permanent link: https://doi.org/10.4095/329758

This publication is available for free download through GEOSCAN (https://geoscan.nrcan.gc.ca/).

Recommended citation

Patwa, B., St-Charles, P.L., Bellefleur, G., and Rousseau, B., 2022. Predictive models for first arrivals on seismic reflection data, Manitoba, New Brunswick, and Ontario; Geological Survey of Canada, Open File 8877, 41 p. https://doi.org/10.4095/329758

Publications in this series have not been edited; they are released as submitted by the author.

Abstract

First arrivals are the primary waves picked and analyzed by seismologists to infer properties of the subsurface. Here we try to solve a problem in a small subsection of the seismic processing workflow: first break picking of seismic reflection data. We formulate this problem as an image segmentation task. Data is preprocessed, cleaned from outliers and extrapolated to make the training of deep learning models feasible. We use Fully Convolutional Networks (specifically UNets) to train initial models and explore their performance with losses, layer depths, and the number of classes. We propose to use residual connections to improve each UNet block and residual paths to solve the semantic gap between UNet encoder and decoder which improves the performance of the model. Adding spatial information as an extra channel helped increase the RMSE performance of the first break predictions. Other techniques like data augmentation, multitask loss, and normalization methods, were further explored to evaluate model improvement.

Contents

List of tables

List of figures

Introduction

Seismic reflection data are used to generate images of the subsurface. Field seismic data typically comprise millions of records called seismic traces that need to be processed and combined to provide an image of the subsurface. One of the processing steps involves the selection of first arrivals or first breaks on a large proportion of the seismic traces using a combination of semi-automated and semi-manual operations. First break selection is a tedious and time-consuming task that can be improved and automated with deep learning.

Results presented in this Open File were obtained during a 6-month internship at the Central Division of the Geological Survey of Canada. The main objective of the internship was to develop machine learning models to predict and select first arrivals automatically. The data sets consisted in raw field seismic data with approximately six million pre-selected first arrivals that were used as labels for training and validation of the models. The internship is an essential part of the Master's in Computer Science curriculum at the University of Montreal (IFT 6917 - Machine Learning Internship) and was supervised by Mila's Applied Research Team and the Geological Survey of Canada.

All the experiments throughout the report were performed on AWS Cloud EC2 Instances with1-4 Tesla T4 GPUs (8 GB). The Geological Survey of Canada/ Natural Resources Canada provided access to the AWS cloud computing infrastructure.

Chapter 1

Seismic problem formulation

Geophysicists study the structure of the Earth's crust by collecting and analyzing a variety of data providing information about the physical properties of the subsurface. For example, vibrations or ground motion produced by seismic waves and recorded with seismographs can be used to infer areas of the crust with slow or fast seismic velocities. Those velocities are commonly associated with different rock types or physical conditions and improve the knowledge of the subsurface. In hydrocarbon or mineral exploration, physical properties can sometimes be used to detect and locate economical accumulation of oil, gas, or minerals. In the context of this project, we used 3D seismic reflection data acquired for mineral exploration at four different mining sites. The seismic waves were generated from dynamite explosions placed at different locations within a survey area. The produced shockwaves are reflected and refracted as they traverse the subsurface and then detected at the surface by geophones along receiver lines. These receivers produce collections of time series of ground motion called traces. The data is then used for many purposes, including classifying different geological layers and predicting the underlying structure and their constituents. Several data processing steps are needed to generate images of the subsurface with seismic reflection data. One of the necessary preprocessing steps for these data is to pick the first arrivals of direct and refracted seismic signals from all the signals detected by the receiver arrays.

1.1. Nomenclature

Seismology and applied seismology have specific terminology to describe the dataset, its properties and their physical and geological relevance. Here, a trace is a sequence of recorded ground vibrations through time, and a gather is a collection of traces. These terminologies will be used throughout the Open File report to explain phenomena and data properties. The most frequent terms used in this report are defined below.

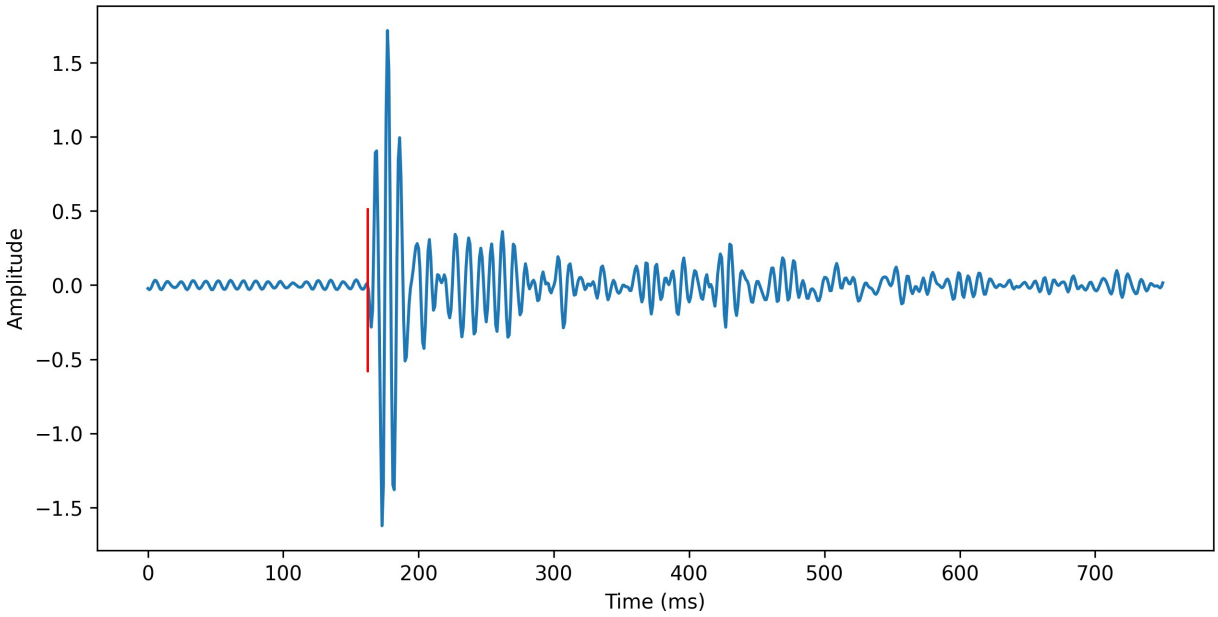


Fig. 1.1. A seismic trace recorded for 751 ms with a sampling rate of 1ms. We can observe a sharp response at roughly 165 ms that corresponds to the first arrival of the seismic shockwave.

1.1.1. Seismic trace

Seismic traces are one-dimensional arrays of recorded measurements from a receiver; an example is shown in Figure 1.1. It contains information that can be associated with changes in rock properties in the subsurface. A single seismic trace is made up of amplitudes of the ground motion recorded at a fixed sampling rate (typically around 1 ms) as a function of time. Units of standard receivers (i.e. geophones) are recorded in mV, but traces are often displayed as relative amplitudes (unitless).

1.1.2. Receiver

Geophones are instruments that are coupled to the ground and record vibrations of passing seismic waves. The device in turn, induces an electric current which is recorded as a seismic trace. Receivers strung together on a single line form a receiver line.

1.1.3. Shot

The initiation of seismic waves through the detonation of an explosive from a known point is called a shot. This produces the shockwaves that travel through the Earth's crust and reflect back for recording on geophones.

Shot Offset: The relative distance between a receiver and the shot origin is the offset distance. Figure 1.2 explains this property.

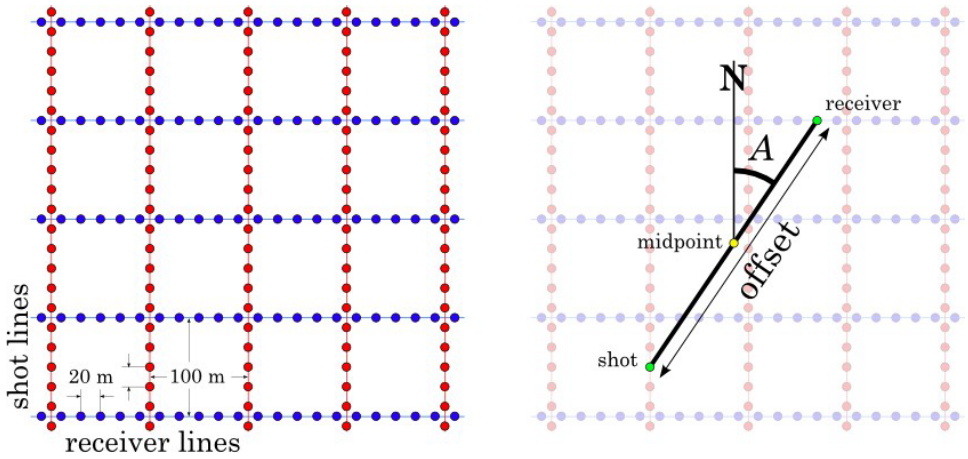


Fig. 1.2. An example of offset between a shot and a receiver (Hall, 2013). The scales given in the figure (20 m, 100 m) are examples of a typical setup used in the studied survey areas.

Shot line: Multiples shots detonated sequentially in a single line perpendicular to the receiver lines form a shot line. A typical 3D seismic survey comprises several shot lines parallel to each other on a predefined area.

Shot gather: A gather is a collection of seismic traces based on some common geometric attribute, like a common shot for multiple receivers, a common receiver for multiple shots, common offset and various other "common" possibilities of the survey geometry (Bianco, 2011).

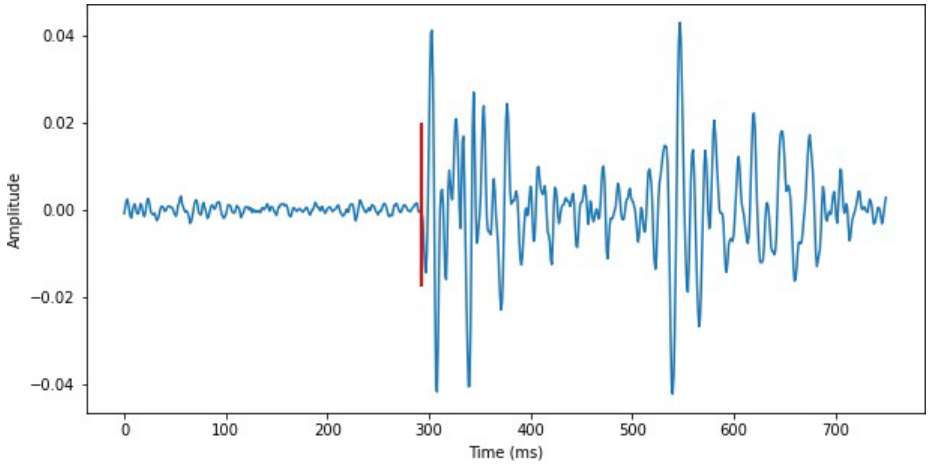
1.1.4. First arrivals

The first arrivals are the earliest information recorded at geophones and correspond to direct or refracted seismic waves that propagated in the near surface. The task of picking the onset of these first arrivals is called first-break picking. They are usually selected automatically and manually edited prior to being used for further processing by geophysicists. The quality of selected picks often depends on the expertise level of the annotator but primarily on the quality (i.e. signal-to-noise ratio) of the data.

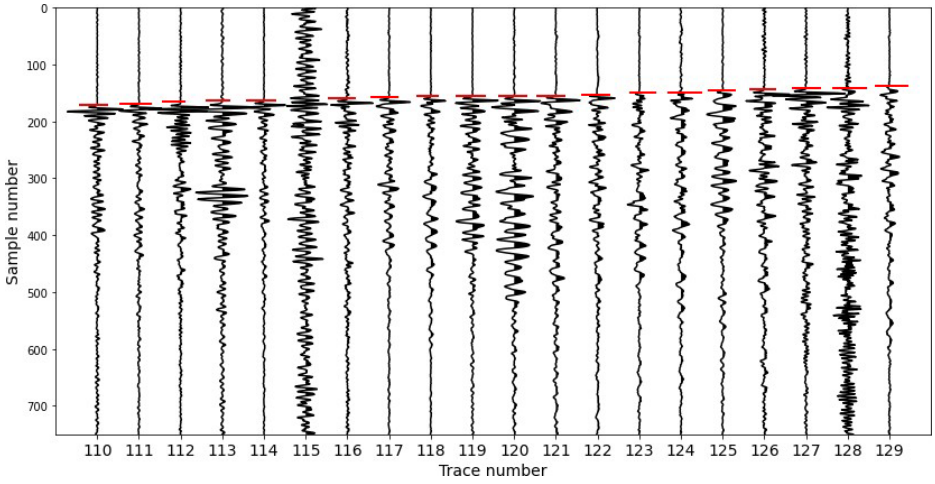
1.1.5. First-break (FB) picking

The onset of the first arrivals of a wave is called the first break (FB). First breaks are produced by the first kind of body waves, called the P-waves (primary or pressure waves). This is the fastest wave and, thus, the first to arrive at receivers. The waves which arrive later contain information used by geophysicists to generate images of the subsurface. For this, accurate picking of first arrivals is important as travel times of first breaks are used to correct effects of the soft, unconsolidated near-surface layer which distort imaging of deeper geological bodies and structures.

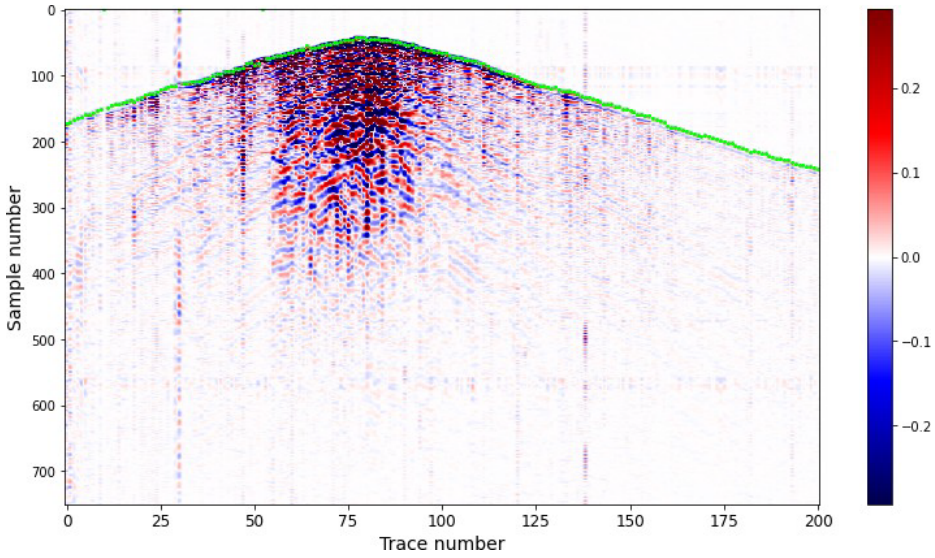
Figure 1.3(a) shows the first break in a single seismic trace. In Figure 1.3(b), we can observe that the 6^{th} trace (trace #115 in Figure 1.3(b)) is noisy and unusable without spatial context, and this is why first break picking is not a trivial task. The dome-like shape of the recorded shot in Figure 1.3(c) is the result of the offset distance between the shot point and the receiver line array. Although it is not universal and not expected in the case of highly uneven subsurface geometry, the receivers nearest to the shot origin will usually have the least time for first-break arrival. Whereas for far away receivers, the first-break arrival times would gradually increase due to offset-distance delaying the recording of seismic waves.



(a) First break event in a trace



(b) First break events in traces collected from a subset of a receiver line



(c) First break events in a shot gather of 1 receiver line array

Fig. 1.3. The markers in (a) red, (b) red, and (c) green are first breaks for each signal.

Chapter 2

Literature review

First break picking is a relatively straightforward task for data with high signal-to-noise ratio (S/N). However, a variety of noise typically affects seismic records and complicates the selection of first arrivals. It is also a simultaneously tedious and time-consuming task for the annotator, especially when the amount of data is huge. This can lead to biased and inconsistent picks since it relies on the subjectivity and attention of the annotator. With the advent of dedicated software for interactive FB picking, the procedure has been somewhat facilitated, but it remains time-consuming and subjective, especially for the manual editing of automatically selected picks. Thus, there is a need for automated computed-based algorithms to rapidly pick FBs in a consistent and objective way.

2.1. Algorithmic solutions

Several algorithmic attempts have been made in the past to tackle the first-break picking problem by transforming the signal into different domains, thus making the FB more easily interpretable. Semi-automatic methods like short and long-term average ratio (STA/LTA) (Coppens, 1985), trace entropy method (Sabbione et al., 2010), and fractal dimension method (Korvin, 1992) have been employed. They are applied on a single trace and, more often than not, are still prone to noise. Thus, the robustness of these semi-automatics methods relies highly on the signal-to-noise ratio. The large number of tuning parameters of sophisticated picking tools also require time to be appropriately configured. The visualizations of attributes generated by some first break picking algorithms are shown in Figure 2.1. The principles of those algorithms are explained below.

2.1.1. Entropy Method (EM)

The trace entropy method (Sabbione et al., 2010) computes the entropy of a time series as a function of time. It is approximated by the sum of absolute differences of seismic trace amplitudes. For a seismic trace $s(t)$, entropy at time t is estimated through a moving window of a fixed length n_h.For the moving average window, we get the logarithm of the sum of the absolute difference between amplitudes, which is then assigned to the last point of the window. This method helps

detect first breaks by taking advantage of the variability and correlation structure of the trace signal. A sample output is visualized in Figure 2.1(a). The entropy is defined by:

$$H(t) = \log\left(\frac{1}{n_h}\sum_{t=t-n_h+1}^{t-1}|s_{i+1} - s_i|\right)$$

2.1.2. Fractal-dimension method (FDM)

A procedure similar to the EM method is applied to transform the seismic trace into its fractal dimension as proposed in Korvin (1992) and Sabbione et al. (2010). In the case of one-dimensional time-series, a fractal curve has a fractal dimension D that is in the range of $1 \leq D \leq 2$. It measures the degree of complexity of a fractal curve. Thus, the onset of a first arrival is determined by detecting the change in fractal dimension between the noise and noise plus signal. We sum amplitudes within a sample-by-sample moving window (n_f) of the seismic trace $s(t)$ for various lags h to estimate its Variogram $V(h, t)$. The fractal dimension D is then estimated after fitting a line's slope b to the log vs log plot of $V(h)$ and h. Figure 2.1(b) shows a sample output of the FDM algorithm. The equations are given by:

$$V(h,t) = \frac{1}{\mathrm{n}_f - \mathrm{h}}\sum_{i=t-n_f+1}^{t-h}(s_{i+h} - s_i)^2$$

$$D = 2 - \frac{b}{2}$$

2.1.3. STA/LTA

The short-term average and long-term average (STA/LTA) ratio which is an energy ratio method (Coppens, 1985), calculates the ratio of the energies of two windows: short term for ns points and long term for nl points along a seismic trace X. A numerical derivative d_i of the STA/LTA ratio r_i is computed for each index i, and the maximum value is given the first-break arrival time for the signal. Figure 2.1(e) presents an output from the equations which are given by:

$$\mathrm{STA}_i = \frac{1}{ns}\sum_{j=i-ns}^{i} X_j^2 \text{ (Short term average)}$$

$$\mathrm{LTA}_i = \frac{1}{nl}\sum_{j=i-nl}^{i} X_j^2 \text{ (Long term average)}$$

$$r_i = \frac{\mathrm{STA}_i}{\mathrm{LTA}_i} \text{(Energy Ratio)}$$

$$d_i = r_{i+1} - r_i;\ i = 1,2,\ldots(N-1);\ d_N = 0$$

A modification was proposed in 2009 (Han et al., 2009) over STA/LTA called the modified energy ratio (MER) uses a fixed size window n_e instead of n_s and n_l, and computes the MER $er3_i$ as:

$$er3_i = (abs(\mathrm{X}_i) * r_i)^3 \qquad \text{(Modified Energy Ratio)}$$

2.1.4. Z-detector

A Z-detector with Z as the standardized variable of the *STA* equation in 2.1.3 was used for comparison in Withers et al. (1998), and the authors found that STA/LTA met the requirements better in an event detection setting. In the equation below, μ is the mean and σ is the standard deviation. Figure 2.1(f) shows the output of the Z-detector on a sample shot gather.

$$Z(\mathrm{STA}_i) = \frac{\mathrm{STA}_i - \mu(\mathrm{STA})}{\sigma(\mathrm{STA})}$$

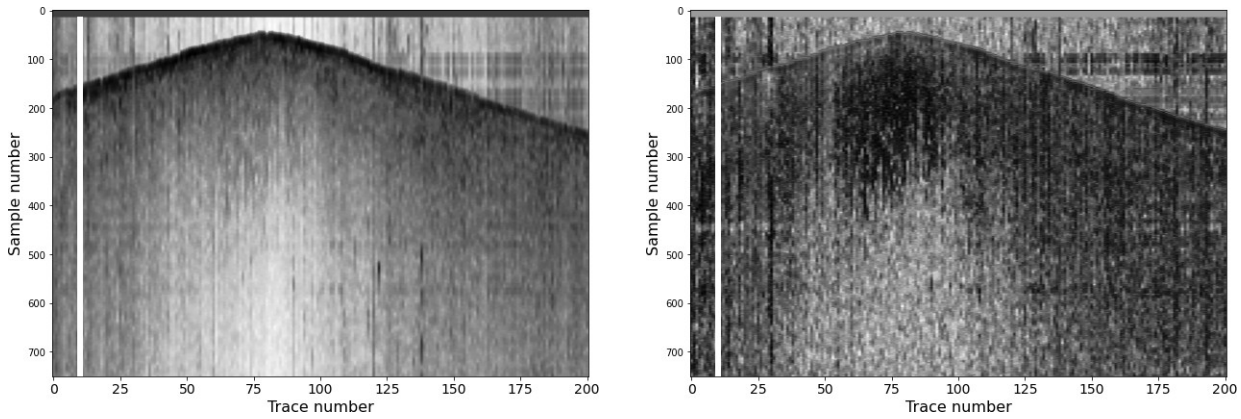
2.1.5. Summary

Apart from the above techniques, other statistical transformations like kurtosis and skewness have been used in seismology as an input to machine learning models (Mężyk et al., 2019) to predict the onset of first-arrivals (see Figure 2.1(c) and 2.1(d)). We will discuss that in the next section.

Sabbione et al. (2010) also improve on the above methods by proposing an additional stage of edge-preserving smoothing (EPS) filter to enhance the transition from noise to noise plus signal. They conclude that the MER method is quite effective for first breaks with strong energy arrival, whereas FDM provides excellent results when other methods are less effective. They still recommend running the three methods (EM, FDM and MER) and selecting the ones that appear to yield the best results, which still, while significantly improved, is a manual task for geophysicists. It is also important to note that algorithms described above are applied to individual traces and do not take advantage of the spatial information (i.e. parabolic alignment of FB's on receiver lines as a result of offset geometry).

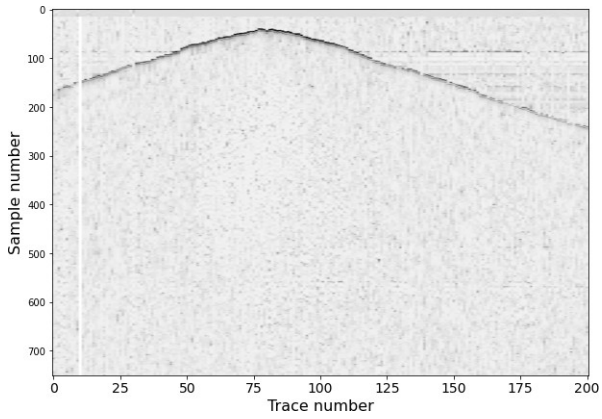
2.2. Machine Learning

Automated first-break picking using Artificial Neural Networks (ANNs) (McCormack et al., 1993) have been proposed as far back as 1993. Recently, the use of Convolutional Neural Networks (CNNs) in an encoder-decoder setup (for example, based on UNet (Yuan et al., 2018)) has led to promising results in solving the problem of seismic segmentation, which also benefits first-break picking.

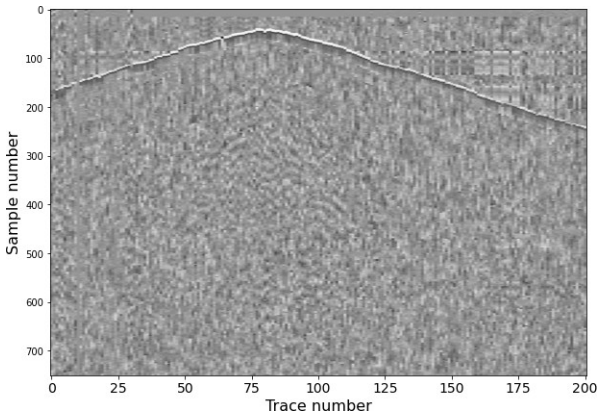


(a) Entropy

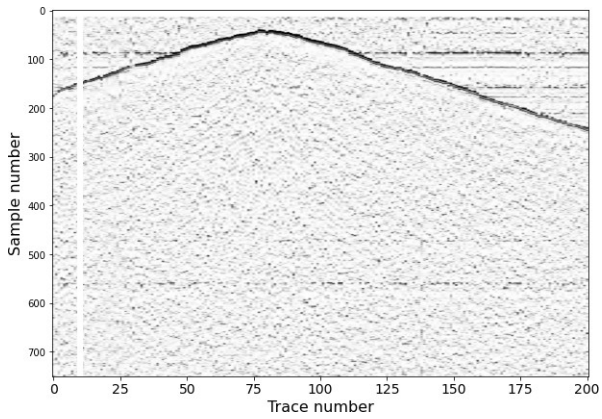
(b) Fractal Dimension



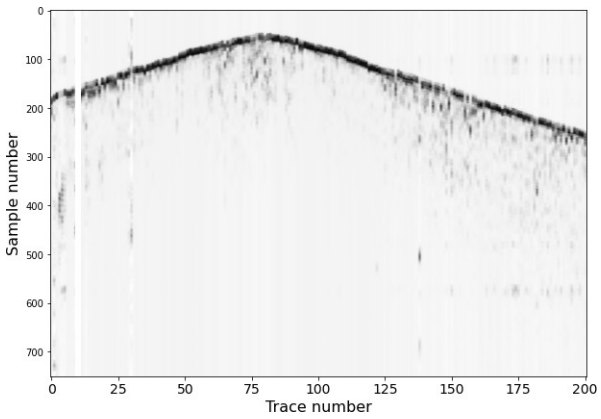
(c) Kurtosis



(d) Skewness



(e) STA/LTA



(f) Z-detect

Fig. 2.1. Different attributes generated by various methods for the onset picking of first arrivals. Notethat in (a), (b), (c), (e) and (f) the start of the darker regions correspond to the arrival of the FBs while in (d), it is the opposite. Regardless, the FB is assigned to the time-stamp corresponding to maximum value in change between the noise and noise plus signal of the transformed data.

2.2.1. Artificial Neural Networks

First break picking has been formulated as a binary classification problem (Mężyk et al., 2019), where the manually picked time sample has been assigned to the positive class and a randomly selected trace event as the negative class. The authors generated the data into a feature matrix with 11 types of signal transformations (Mężyk et al., 2019, Figure 2). The seismic traces for before (pre-FB model) and after (post-FB model) the FB and non-FB events were sampled (Mężyk et al., 2019, Figure 4) to train an ensemble Deep Neural Network model. The limitations of this form of model which does not incorporate spatial information, have been acknowledged by the authors.

2.2.2. Convolutional Neural Network

Models based on Convolutional Neural Networks (CNNs) have been introduced due to limitations of the earlier solutions, which do not take advantage of the spatial information of the data. Yuan et al. (2018) introduced CNNs to design FB picking solutions that output first break confidence scores (in [0, 1]) for every pixel. By using score thresholding, first local minimum rule, and a median filter to smoothen outputs, the final FBs are predicted for the seismic input image.

Following this, Alaudah et al. (2019) present work based on deconvolutional neural networks for facies classification, a task similar to FB picking but of higher complexity as the link between facies classes and input features is more subtle. Chevitarese et al. (2018) introduce residual units in a CNN, which reduces the number of trainable parameters by 5 times and operations by 2 times compared to topologies with fully-connected layers. The same authors continue on the same line of ideas by introducing transposed convolutions at the upscaling expansive path in the encoder-decoder style architecture (Civitarese et al., 2019), similar to that of the UNet architecture (Ronneberger et al., 2015).

2.2.3. UNet

UNet is an encoder-decoder style architecture that has a contracting path on the encoder-side and an expansive path on the decoder side. The encoder reduces the spatial dimensions while increasing channels whereas the decoder increases the spatial dimensions while decreasing the channels, leading to a 1x1 convolution layer that converts feature maps into class score maps. To bridge the information gap on the decoder side, outputs of each encoder is given to the respective decoder as a skip connection as shown in Figure 4.1.

MultiResUNet, as defined in Ibtehaz (2020), is a novel architecture that proposes Inception-like blocks called *MultiRes* block (Szegedy et al., 2015) at each convolutional layer of UNet (Ronneberger et al., 2015) in order to learn features from the image at different scales. The 3×3, 5×5 and 7×7 convolutional filters are converted to a succession of 3×3 filters with intermediate outputs that are concatenated to extract spatial features. Residual path (termed as *Res* path in

Ibtehaz, 2020) with extra convolutional blocks at the shortcut connection path is also proposed to address the semantic gap between the downsampling layer (encoder) and upsampling layer (decoder) of the same level. Moreover, both *MultiRes* block and *Res* path contain residual connections.

As a result of the architecture and improvements available from these latest works, we try to continue working on the same style of architectures on the datasets provided by NRCan.

Chapter 3

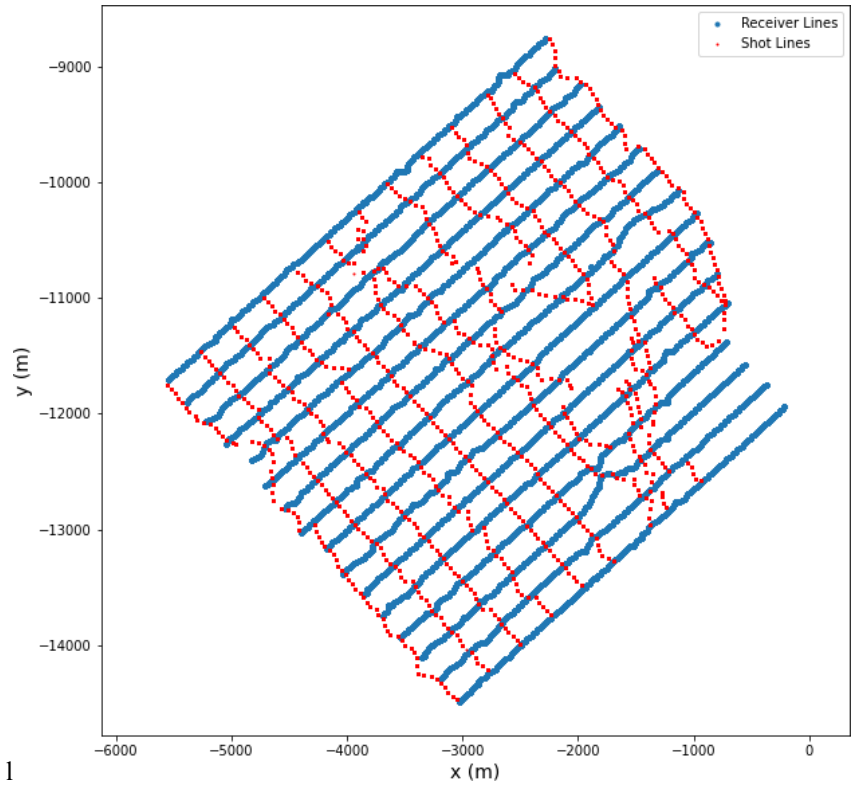
Dataset

3.1. Overview

The seismic datasets provided by NRCan were both publicly available and proprietary. All datasets are now available on GitHub (https://github.com/mila-iqia/hardpicks) through open access licenses. The seismic datasets used for FB predictions included surveys from 4 different locations and were collected as stated in Section 1.1. The amount and quality of available FB annotations for each data set (Table 3.1) give a general overview of the data. A more holistic view of annotations per shot-receiver line pair will be discussed later.

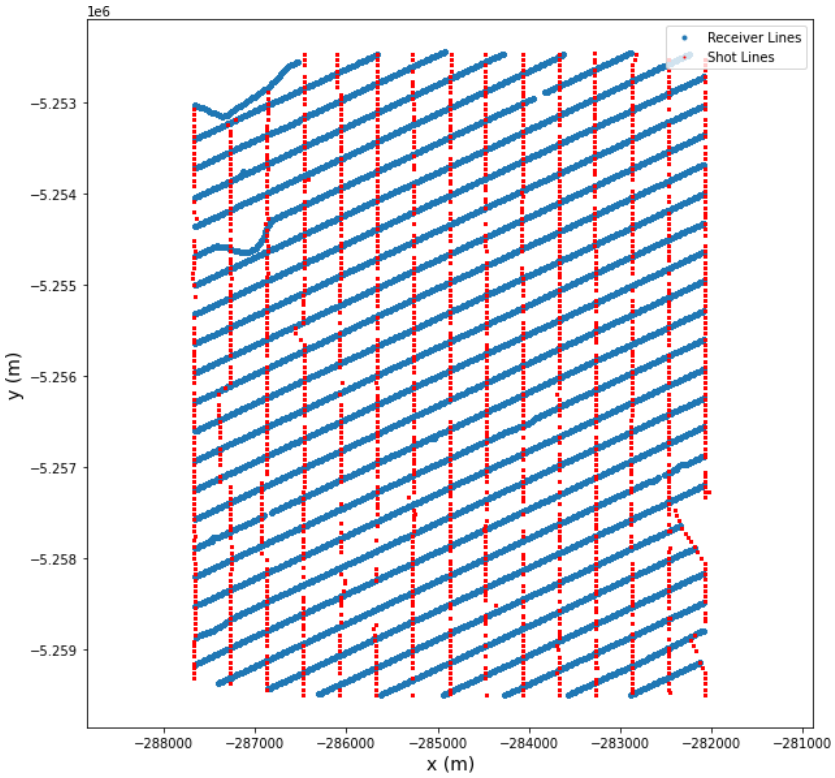
Name	Traces	Traces with picks	% of traces with picks	Quality
Brunswick 3D	4496540	3737844	83%	Picks are PhD student grade.
Halfmile 3D	1099559	925259	90%	Picks are post-doc grade.
Lalor 3D	2424923	1214572	50%	Picks are commercial grade followed by in-house editing.
Sudbury 3D	1810220	235124	13%	Picks are of unknown source.

Table 3.1. Dataset description.



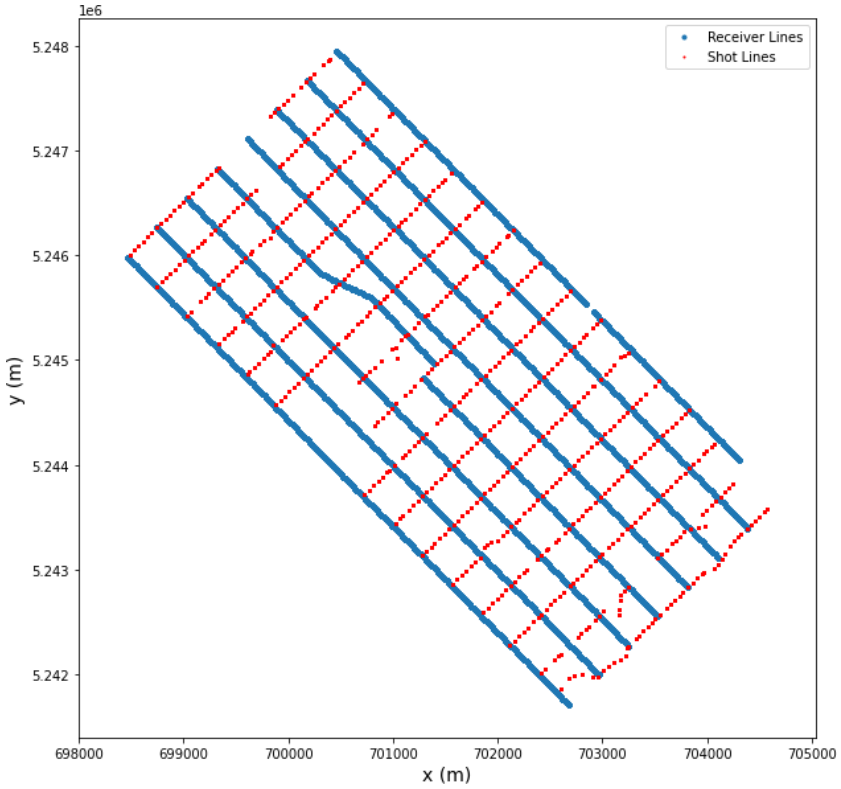
l

(a) Lalor 3D

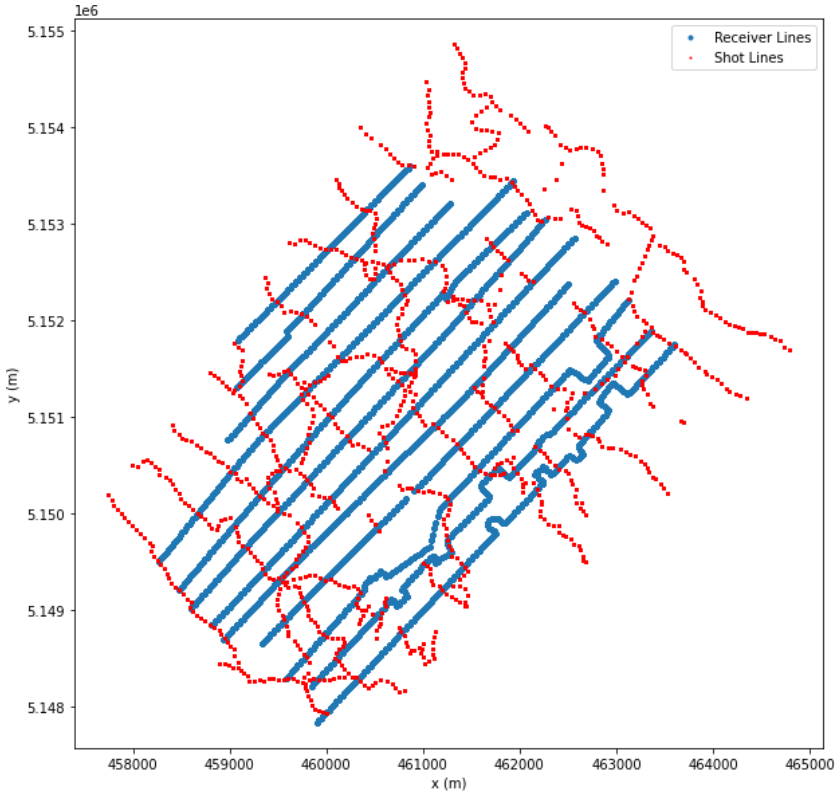


(b) Brunswick 3D

Fig. 3.1. Receiver and shot line layout over the Lalor and Brunswick 3D seismic survey areas.



(a) Halfmile 3D



(b) Sudbury 3D

Fig. 3.2. Receiver and shot line layout over the Halfmile and Sudbury 3D seismic survey areas.

3.2. Dataset Structure

As seen in Figures 3.1 and 3.2, we can conclude that the receiver lines are largely parallel in 3.1(a), 3.1(b), 3.2(a), with some acceptable discontinuities, whereas 3.2(b) has notable irregularities in the setup of the receiver lines. Note that shot lines having irregularities would not matter in the context of first-break picking since a shot gather would only pertain to an individual shot (single red dot) to a receiver line (array of blue dots).

When collecting the data, not all receivers or receiver lines are necessarily active for every shot point. The subset of active receivers for each shot point is predetermined based on survey objectives prior to data acquisition.

3.3. Preprocessing

Both seismic traces and labels (FB) were preprocessed before they were used for training. Preprocessing of seismic data aimed primarily at normalizing amplitudes, whereas preprocessing of FB consisted in finding and removing outliers. Details of various preprocessing steps are provided below.

3.3.1. Seismic Trace Normalization

All traces were processed with maximum amplitude normalization due to some significant amplitude variations between seismic traces. These large amplitude variations in both directions differed from trace to trace as they depend on multiple physical factors recorded by individual receivers (i.e. coupling, near-surface ground conditions, etc.).

3.3.2. First break picks

The annotations were cleaned to remove empty and erroneous picks and were replaced with different techniques. Sabbione et al. (2010) propose a miss-pick correcting algorithm with 5 steps but we find this quite computationally expensive and devise a more lightweight technique using the difference in velocity between seismic arrivals, and determined from arrival times and offsets. This produces more false positives, but the regression models used during preprocessing correct them quite accurately.

Removal of invalid FBs and filtering shot gathers on FB ratio: This was fairly straightforward, any annotations having a time-stamp of 0 or lower were discarded. Any shot gathers that had less than 60% of first-break annotations were discarded. This is discussed in the next section.

Outlier Removal: For outlier removal, different strategies were tried and tested such as 1) finding the mean and standard deviation of first break time-stamps from n bins in each shot and discarding outliers outside three standard deviations from the mean, 2) finding the global mean

and standard deviation and then using the same strategy, or 3) using velocity $v = \frac{offset}{fbtimestamp}$ as the magnitude at each receiver offset and assuming a linear relationship between velocity vs offset of the travelling P-wave to find outliers.

The issue with the first approach was that it did not take into account the continuous nature of the FB and its curvature properties (i.e. time moveout with offset on consecutive traces), so any abrupt change in FB or the endpoints of the bin were falsely classified as outliers. With the second approach, the nature of the wave was different for every dataset and the lowest shot offset, and the same issues as the previous one occurred again. In the third approach, the similar offset distances from opposite sides of the shot origin within the same receiver line made it tricky to resolve.

So, we used a unique and computationally feasible method using an n-length window of the velocity v and finds a median instead of the mean. A median filter was applied to calculate the sequence of median values along a receiver line. The absolute differences of the median curve and the FB velocity were taken and anything greater than a predetermined threshold was counted as an outlier. Based on empirical analysis, a window size of 7 for the median filter and a threshold of 0.07 worked best for the available datasets.

Interpolation and Extrapolation: To fill gaps, the missing first breaks after outlier removal were filled with polynomial interpolation. Empirically, cubic polynomial for interpolation and squared polynomial for extrapolation worked best for filling missing first break values. The receiver distance relative to the first receiver in a receiver array was used as the X value and the first breaks as the Y values for the estimation of the polynomial parameters.

3.4. Dataset Preparation

All the datasets had distinct sampling rates and trace lengths, so to ensure consistency in image size, all traces were resampled to have exactly 751 time samples (without changing the total time length), as shown in Table 3.2. The width of receiver panels was padded to a maximum size of 280, which is also the maximum size among all the receiver lines in the 4 datasets. The first break values were also scaled according to the new time delta. The formula was: $\Delta t = \frac{\text{Trace length} * \text{Sampling rate}}{751}$. Table 3.1 describes the modified sampling rate of the datasets.

As per Figure 3.3, the percentage of annotations in every shot gather remained consistent until the 60% mark on the x axis and then started reducing gradually. So, a threshold of 60% was set in order to remove shot gathers with low annotations that could seriously affect the interpolation/extrapolation of missing first breaks. The possibility of using them without interpolation/extrapolation was also considered by applying masks, although this was not explored.

The output image was a segmentation map with the area above the first-break patch having an index of 0, below the first-break patch having an index of 1 and the first-break patch having an index of 2. The First-break patch was initially set to a size of 10 samples so that the model could

Name	Sampling Rate (ms)	Trace Length	New Sampling Rate (ms)
Brunswick 3D	2	751	2
Halfmile 3D	2	751	2
Lalor 3D	1	1501	2
Sudbury 3D	2	1001	2.666

Table 3.2. Modifications in sampling rate and time delta to ensure consistency among datasets. Although this changes the frequency content of the data, its impact was not assessed.

sufficiently capture the change in transition of the seismic traces during drastic discontinuities in either subsurface structure or receiver line geometry. However this can be changed as per empirical results.

3.5. Train-Val-Test Split

As shown in Figure 3.3, the Halfmile 3D dataset has the highest percentage of annotations so it was kept as the test set. The amount of annotations for Sudbury 3D is quite low for the 60% minimum annotation requirement, so we removed this dataset. For train and validation, we use a split of 75-25% from the Brunswick and Lalor dataset. The data is split randomly among all shot gathers.

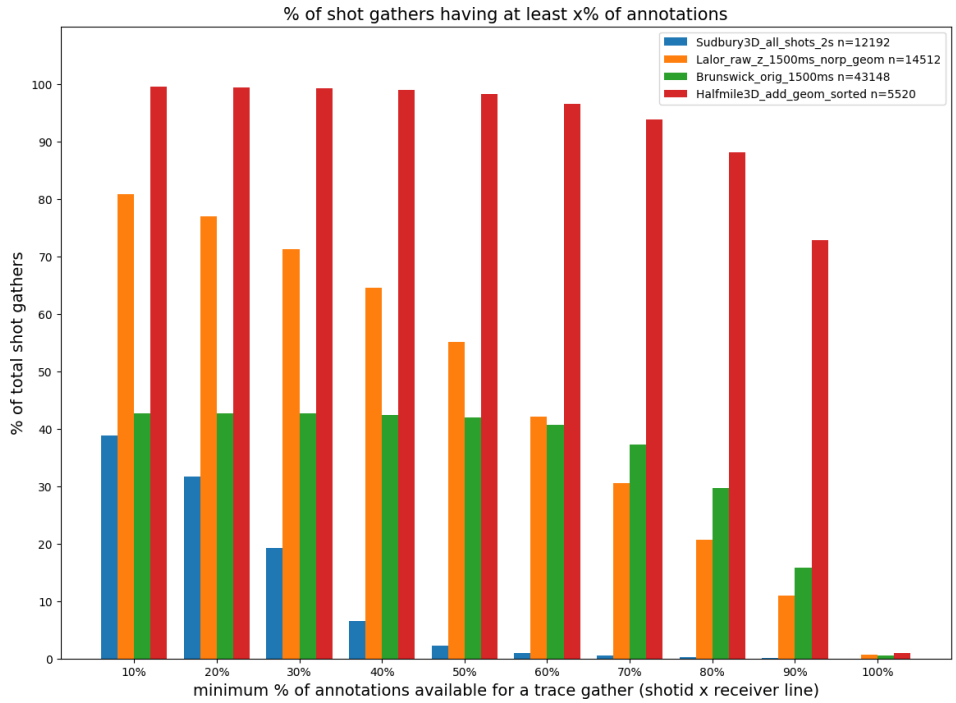


Fig. 3.3. 10-bin distribution of the percentage of total shot gathers w.r.t annotation percentage of a trace gather for each dataset.

Chapter 4

Models

4.1. UNet Architecture

The UNet is a type of CNN architecture designed to perform the task of semantic segmentation. The architecture is shown in Figure 4.1. It is a symmetric architecture, having an encoder network that extracts high-level features from the input images and a decoder that converts these high-level features into the segmentation map. The encoder follows a sequence of two 3×3 convolution operations, each followed by Batch Normalization and ReLU (Rectified Linear Unit) activation. Batch Normalization standardized the layer outputs before passing it into the next layer and helps in reducing overfitting during training. ReLU activation helps in avoiding the vanishing gradient problem and introduces sparsity in the network, which has been shown to be advantageous by Glorot et al. (2011). It is followed by max pooling with a pooling size of 2×2 and stride of 1. This sequence of operations is repeated 4 times, and at each downsampling event, the number of convolutional filters are doubled.

For the decoder, the upsampling event is made using either bilinear upsampling or using transposed convolution with 2×2 kernel size and stride of 2. Then, the same sequence of two convolutions is applied like the encoder but with halving the number of filters at each level. A 1×1 convolutional layer at the final output generates the logits of the final segmentation map using a *Sigmoid* activation function.

At each depth level, the output of the encoder before the pooling operation is provided to the decoder through a skip connection. Skip connections link feature maps of similar sizes between encoder-decoder layers and help with gradient flow to the topmost nodes of the network. It also allows features to be learned across different ranges of receptive fields and ensures feature reusability. These feature maps are concatenated to the output of the decoder (upsampling operation) and propagated to the subsequent layers.

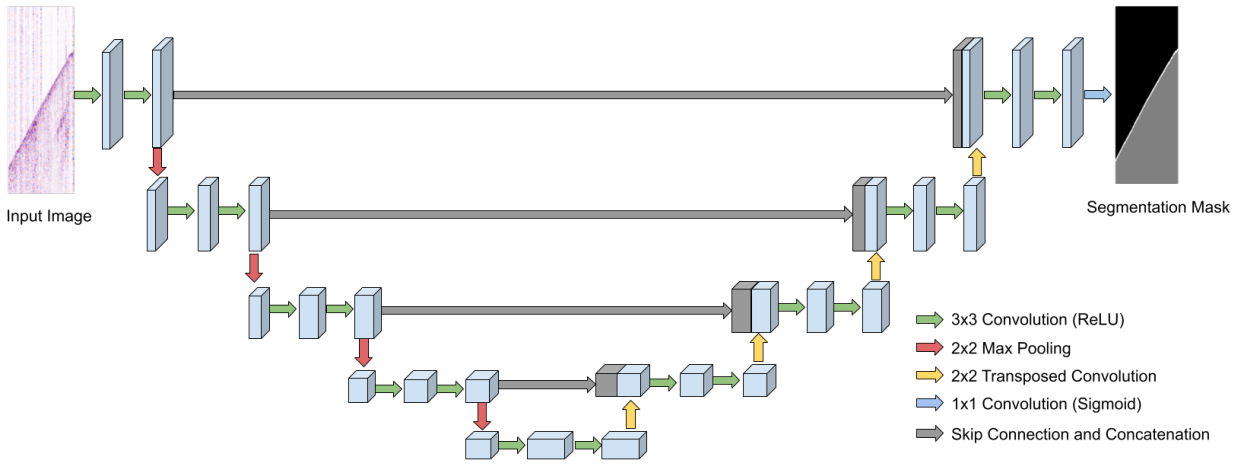


Fig. 4.1. The U-Net architecture. Model consists of downsampling and upsampling layers with skip connections between the two corresponding layers. Image regenerated from (Ibtehaz, 2020)

4.2. MultiResUNet

4.2.1. Res Block

Inception-like blocks have the structure to facilitate learning features of images at different scales as described in Szegedy et al. (2015). But they have a high-memory requirement due to convolutional layers with 5×5 and 7×7 kernels. To alleviate this, MultiResUNet (Ibtehaz, 2020) factorize the larger convolutional layers into a sequence of 3×3 kernels.

In order to incorporate Inception-like blocks with 3×3 and 7×7 convolutions in parallel to the 5×5 one, (Ibtehaz, 2020) proposes 3 consecutive 3×3 convolutional blocks. Their 2nd and 3rd intermediary outputs achieve the 5×5 and 7×7 convolutional operations respectively. They also add a residual path from input to output with a 1×1 convolutional layers. Also, the number of filters for all 3 consecutive convolutional layers increase gradually instead of keeping them the same to prevent memory requirement from dramatically growing. The whole configuration is called a "*MultiRes block*" as shown in Figure 4.2.

4.2.2. Res Path

Furthermore, there is disparity in the features at the same level of encoder and decoder (Ibtehaz, 2020). The processing at the first level of encoder is at the initial stages, but for the first level decoder it is at the final stage. So to bridge this possible semantic gap in the shortcut connections, MultiResUNet incorporates more convolutional layers at the shortcut path to account for more processing done at the decoder stage therein. The layers also come with residual connections. The whole arrangement is termed as "*Res path*". This module is shown in Figure 4.3.

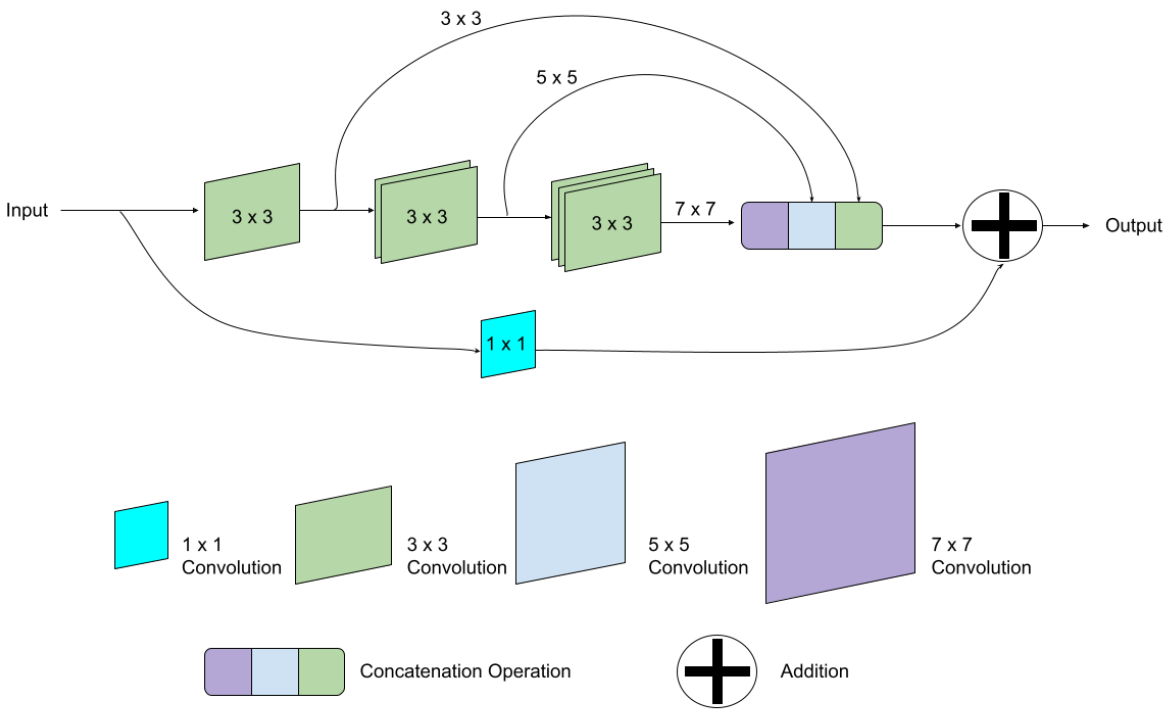


Fig. 4.2. The *MultiRes block* as proposed in (Ibtehaz, 2020) outputs at each subsequent 3 x 3 convolutional layer which are concatenated at the output.

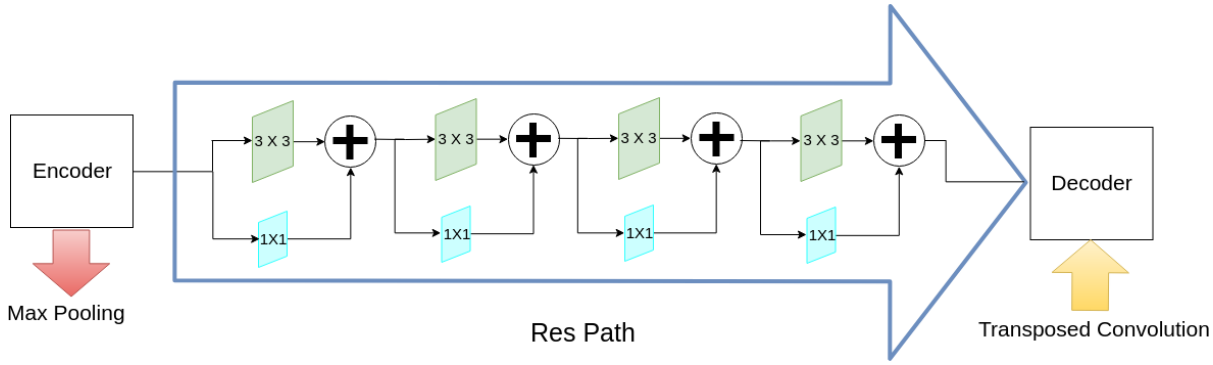


Fig. 4.3. The *Res path* as proposed in (Ibtehaz, 2020). The additional non-linear operations as convolutional layers are supposed to reduce the semantic gap between encoder/decoder.

Chapter 5

Experiments and Analysis

5.1. Training setup

Training was done in Python using PyTorch as the framework. The UNet models were based on Milesi (2017) and the MultiResUNet was built upon that.

The UNet and MultiResUNet models were trained on different objectives, including: learning rate scheduling, with additional input channels, total number of classes, and with multiple post processing options. The Focal Loss was used as the training loss as it addresses the issue of class imbalance (Lin et al., 2017). The metrics used to report the accuracy are the mean Intersection Over Union (Jaccard Similarity/mIOU) and RMSE scores.

We used an improved version of the Adam optimizer called AdamW, for which the weight decay does not end up in the moving averages and is added only after controlling the parameter-wise step size (Loshchilov et al., 2019). Thus, the term remains proportional to the weight itself. An optimal learning rate was computed as per the algorithm proposed by Smith (2017) for the cyclical learning rate.

All training was done on the NRCan-AWS cloud on the EC2 instances with 1-4 GPU combinations. Nvidia's Automatic Mixed Precision (AMP) available via the CUDA library through PyTorch was used to speed up training while maintaining accuracy.

5.1.1. Prediction

The time-stamp of the first index at which we get prediction probability $p > 0.5$ for class=2 is chosen as the first break. The definitions of these classes are elaborated in the next subsection. In case we do not find a positive prediction for class=2, we take the time-stamp of the index after the last positive index for class=0 to ensure consistency. The time-stamp is then converted into the original sample space.

5.1.2. Focal Loss

Segmentation masks are known to be prone to imbalanced classes and thus skew the data balance towards the majority class. This also means the loss computed on the final segmentation mask heavily relies on the majority class. To combat this, focal loss reduces the weight on the loss of correctly classified examples during training, in addition to balancing the loss based on the inverse class frequency. This allows the model to focus learning on our target class as it is not a majority class. Lin et al. (2017) add a modulating term $(1-p_t)^\gamma$ to the cross-entropy loss $\mathrm{CE}(p_t) = -\log(p_t)$, with a tunable focusing parameter $\gamma \geq 0$. In a binary classification scenario, the model's predictions for the class with label y=1, where $y \in \{\pm 1\}$ is $p \in [0,1]$. Also, α is set by inverse class frequency or as hyperparameter. The focal loss for binary classification is defined as:

$$\begin{array}{ll} \mathrm{FL}(p_t) & -\alpha_t(1-p_t)^\gamma \log(p_t) \\ p_t & \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise,} \end{cases} \end{array}$$

5.1.3. Intersection over Union (IOU)

This metric allows us to compute the accuracy of the prediction class over the target class. It is calculated by taking the ratio of area intersection between the predicted (A) and target segments (B), and the area of overlap of the two segments.

$$IOU = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

While IOU is calculated for each class, mIOU is the mean IOU among all classes.

5.1.4. Root Mean Squared Error (RMSE)

RMSE is a crucial metric, as it will quantify the accuracy of the first-break predictions ($\hat{y}$) against the actual target predictions (y) over T receivers in a gather. We calculate RMSE for each gather and report the average RMSE score during evaluation.

$$RMSE = \sqrt{\frac{\sum_{i=i}^{T}(\hat{y}_i - y_i)^2}{T}}$$

5.2. Configurations

The preliminary experiments were done on UNets, which were small-scale with a 2-layer setup in both encoder and decoder, then gradually increasing depths of UNet as better configurations emerged.

5.2.1. Two Class, Three Class and Multitask Loss

Initial experiments were done to find out which segmentation mask facilitate the training of the model. We later show using empirical results that a 3-class segmentation setup gave better results than a 2-class setup. For 3-class, the third class was chosen as the first-break patch with a thickness of 10-pixels. A combined multitask loss of 2-class and 3-class was also taken into consideration.

5.2.2. Normalization vs No Normalization

Experiments were done to analyze the effects of min-max normalizing each trace versus keeping the data as it is with just trace amplitude normalization.

5.2.3. Bilinear upsampling vs Transposed Convolutions

For the decoder section, a comparison was done to determine whether deterministic upsampling like bilinear upsampling was a better choice than transposed convolutions which are learned during training.

5.2.4. Offset distance as input channel

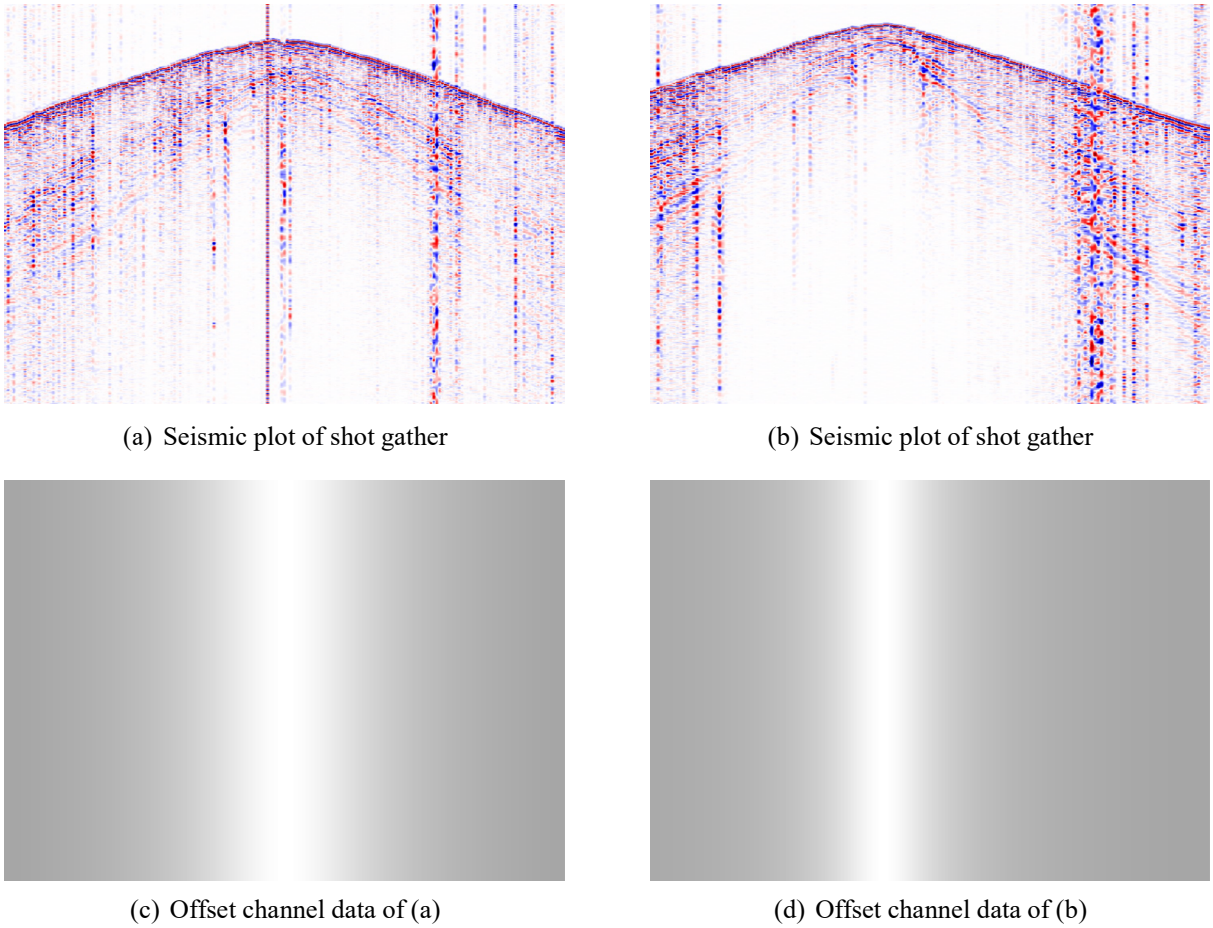
Liu et al. (2018) propose a coordinate convolutional layer by providing coordinates in (x,y) Cartesian space as hard-coded input channels. This allows CNNs to learn a mapping between the inputs and the spatial coordinates. The benefit here is that the closest offset distances are at the peak of the first break curve as they are the closest to the shot point. In contrast, receivers with large offset distances OD records the first break much later as they are farther away from the shot origin. So clearly, there is a correlation between offset and time of first-break arrivals.

Since the datasets were of different origins and model generalization is one of our primary goals, we opted for a solution that works independently of the dataset's physical layout. An extra input channel for offset distances that are inverse-scaled and normalized was considered as the spatial dimension. Inverse-scaled normalized offset distances were repeated for every row OR up to the height of the input image to ensure cohesive dimensions. The closest offset distances are the brightest at 1. Figure 5.1 (c) and (d) show the usage of the offset distance as an additional channel and its relationship with the input image shown in (a) and (b), respectively. The formula for generating the offset row OR is:

$$\mathrm{OR}_i = \frac{min(\mathrm{OD})}{\mathrm{OD}_i}$$

5.2.5. Data Augmentation

The input images were horizontally mirrored with a random probability of 50%. This is to ensure padding information at the end is not learned by the model, and the augmented data



(a) Seismic plot of shot gather

(b) Seismic plot of shot gather

(c) Offset channel data of (a)

(d) Offset channel data of (b)

Fig. 5.1. The correlation between the top dome of first break arrivals in the shot gather and offset channel is visualized here. The offset image can be visualized as a band repeating itself for all rows of the image dimension.

helps the model learn from a different view. Other augmentation operations were not explored due to a lack of time.

5.2.6. Learning Rate and Scheduling

The OneCycle Learning rate scheduling proposed by Smith (2018) was chosen over Plateau scheduling due to its cyclical learning rate which increases until halfway through the total number of epochs and reduces from then on. This configuration has been shown to reach super convergence as proposed first in Smith et al. (2018). The optimal initial learning rate value was found by the Learning rate range test in Smith (2017). The batch size used were 56, 32, or 16 depending on the GPU configuration used. The training was done for 25 epochs first and 50 epochs for later experiments.

5.3. Results

	UNet (1-5)					MultiResUNet (6-13)							
Learning Rate	6.00E-04				2.85 E-03	6.33E-03							
Class size	3	2			3								
Loss over target	Single task loss (S)					Multitask loss (M)		S					M
Channels	1					2	1	2	1	2	1		
Normalize	Y					N	Y			N			
UpSample	Bilinear			Transposed Convolutions									
Data Aug.	N					Y							N
UNet Depth	2			3	4								
Scheduling	Plateau				Tri ang ular	OneCycle							
Kernel Size	(3,3)		(5,3)	(3,3)									
mIOU	0.808	0.814	0.791	0.813	0.884	0.915	0.923	0.918	0.921	0.925	0.913	0.911	0.926
avg. RMSE	15.06	28.24	45.07	31.10	11.79	**5.52**	7.66	10.88	7.79	10.47	10.41	8.81	8.51
Exp. No	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 5.1. Performance of different models over different configurations on the validation set. The last row defines the index for each experiment for convenience during analysis. It is important to note that for x_i number of filters every layers in UNet, they are halved to $x_i/2$ for MultiResUNet. Further analysis show that halving did not degrade model performance.

5.3.1. Analysis

During initial experiments on UNet, empirical results showed that using Transposed Convolutions during upsampling input feature maps improved results. This is because it is able

to learn the upsampling features rather than just rely on linear interpolation. As UNet depth size was increased, the mIOU metric gradually improved. The original UNet had 31.04 million parameters whereas the MultiResUNet had 3.59 million trainable parameters with half the number of filters at every layer compared to UNet due to GPU-memory constraints, both with the same depth of 4. An equivalent MultiResUNet with no halving of filters will have 14.35 million parameters, a reduction of 54%.

Figure 5.2(a) and 5.2(c) show what an ideal prediction would look like as it nearly matches the target predictions. In contrast, we can see the prediction markers not following the first breaks over the last 20-30 receiver indexes in Figure 5.2(c) and the first 20-30 indexes in Figure 5.2. A clear distortion in pick-time of the first arrivals can be seen. Even though we do not have correct target predictions for the majority of receivers in Figure 5.2(d), the errors for predicted first break times can still be inferred from visual observation. There is a presence of a vertical patch of unknown noise/disturbance observed in the focused receiver indexes of Figure 5.2(c) near the end of the receiver line. A milder version of the same noise is also observed in Figure 5.2(d) for low receiver indexes. Due to the presence of this noise, the model struggles to segment the region and thus predicts first break times with lower accuracy in both cases. A more vivid visualization of such disturbance can be observed in Figure 5.3.

During early experiments, the mIOU scores prove to be a reliable metric. However, mIOU lacked a strong correlation with first break RMSE scores beyond a certain improvement threshold. It became apparent that mIOU only gave a general performance of the segmentation, whereas the performance of class boundaries for the 3-class setup was better judged by relying on RMSE scores.

Using the equivalent MultiResUNet model with the halved filters configuration as the best performing UNet (5) reduced the RMS error by a little bit more than 1-pixel down to 10.47. Using the Multitask loss as a combined loss of 2-class and 3-class segmentation also did not improve the scores as much when comparing experiments with identical configurations (apart from the loss; see results for experiments 8 vs 10, 6 vs 11, and 7 vs 9 in Table 5.1). One reason could be that making the model learn conflicting objectives (i.e. the class-2 10-pixel buffer of the 3-class setup overlapping with the class-1 objective of the 2-class setup) would confuse it during training. As the image sizes in this project were rectangular, we also tested a rectangular kernel of size (5,3) across the entire network instead of the (3,3) kernel. The rectangular kernel was not preserved as it degraded the network performance.

The use of offset distance in the form of an additional channel significantly improved the metric by 3 pixels, at the 7-pixel range when comparing experiments 7 vs 10 and 8 vs 9. This hypothesis turned out to be true as offset distance strongly correlates with first break arrival

times, as can be seen in Figure 5.1. Furthermore, experiments that tested normalization (experiments 6 vs 7 and 9 vs 11) could not provide concluding remarks on the models' performance.

We get a Test RMSE score of 8.8 and a mIOU score of 0.9196 on the Halfmile Lake dataset on the best model.

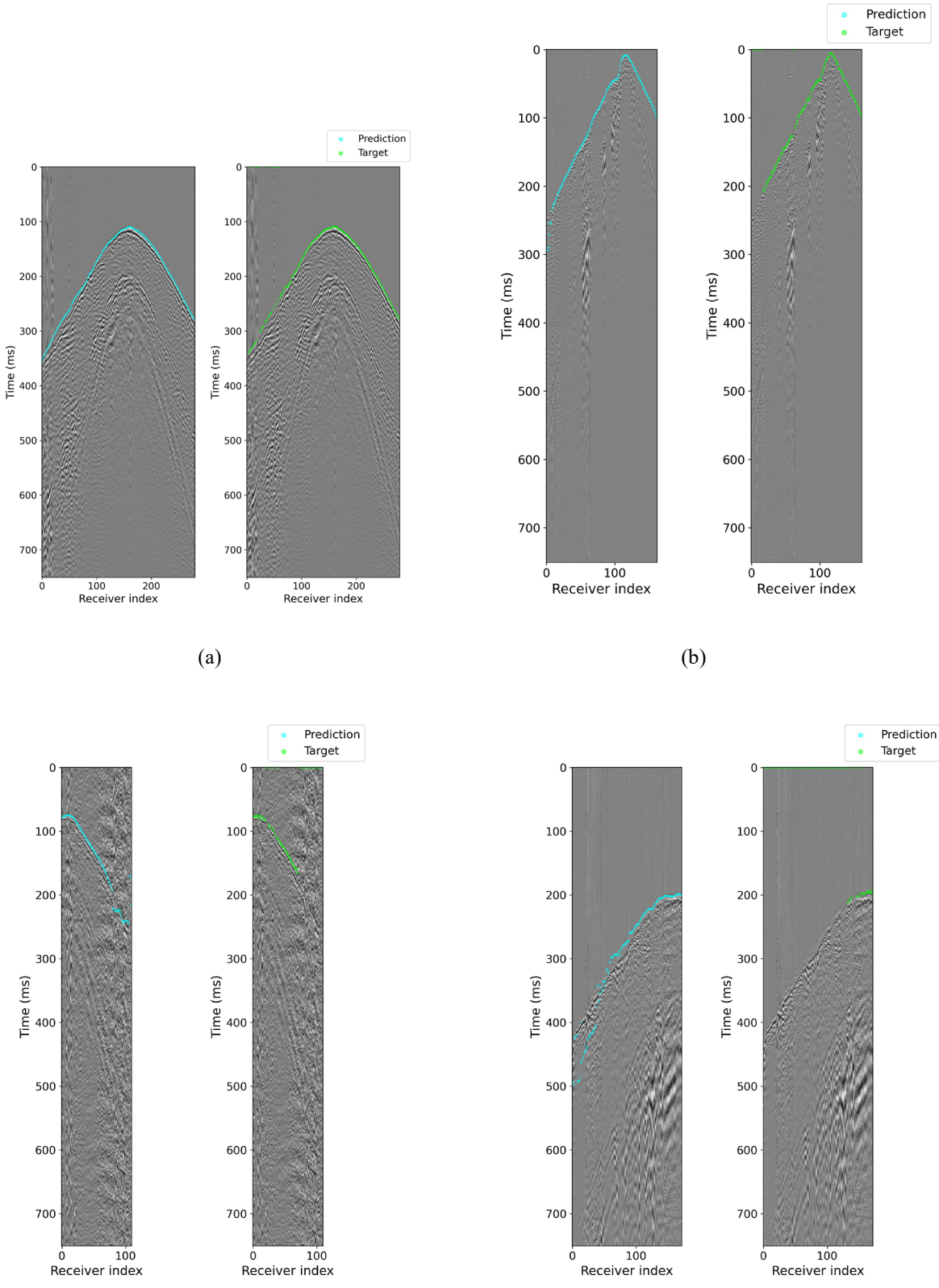
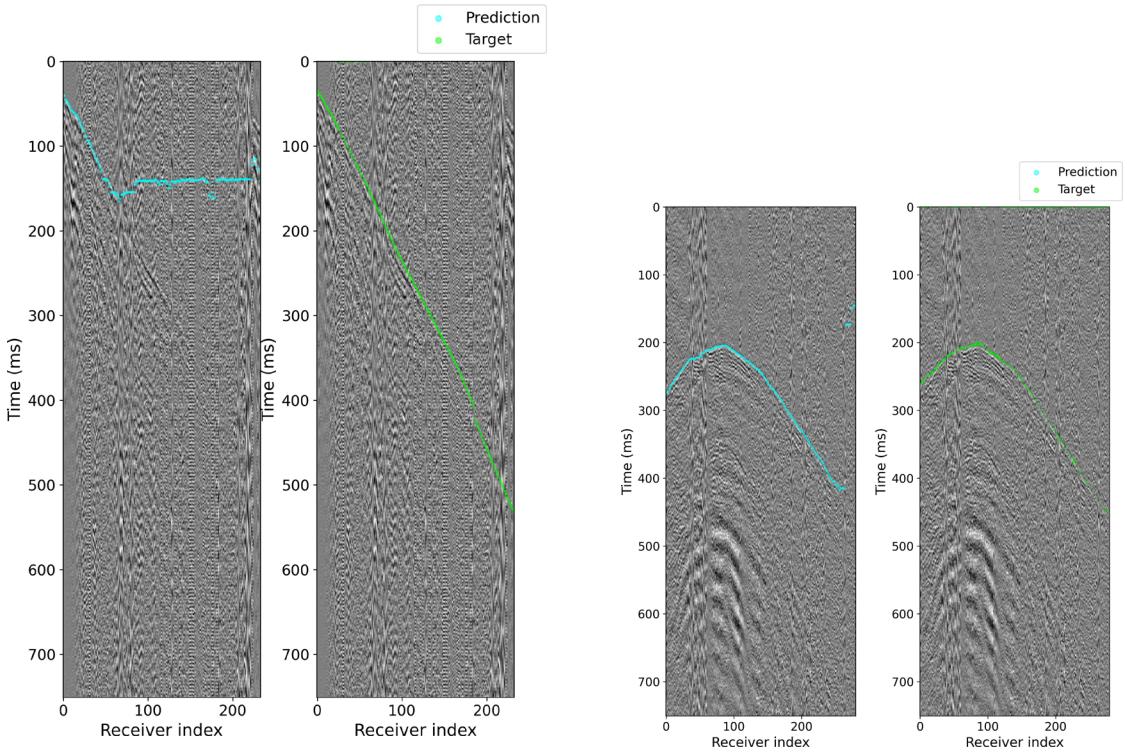


Fig. 5.2. Figure (a) and (b) show the typical prediction results, while (c) and (d) showcases some of the failure cases of the model.



(a) (b)

Fig. 5.3. Strong disturbance of unknown origin affects the model performance in (a), whereas not much in (b). Model can also falsely predict presence of unknown signal as class-3 i.e. belonging to the first-break event due to the absence of strong FB arrival signals that are indicator of first-breaks.

Chapter 6

Conclusions

The MultiResUNet approach reduces the RMSE loss below the 11-pixel threshold which was easily crossed by standard UNets. Our best model with a training RMSE of 5.52 gives a test RMSE of 8.8. We looked into various techniques to tackle the first break picking problem such as using increased UNet layer depths, better scheduling, data augmentation, additional input channels, multitask loss, different ways of upsampling, class size and kernel size. We also analyzed some of the results, which provided insights into the weaknesses and failures of the model.

Overall, using the offset-channel as an extra spatial dimension helped reduce RMSE by a few pixels. No specific analysis on normalization could be reached due to the lack of strong evidence to comment on it. The mIOU has been a meaningful metric until reaching an improvement threshold of 0.90-0.92. After that, RMSE provides better insights on first break picking accuracy.

Multitask loss did not worsen but also did not improve the scores. Another type of joint loss that does not conflict with the primary loss objective could be a better alternative. Our results have shown that using additional input channels that augment information can help the model performance. This was the case with offset distance added as an additional channel. A similar approach could also be devised to employ receiver-line geometry. This would help reduce the disparity in seismic gathers due to uneven distances between adjacent receivers. This can be achieved through scaling and processing, similar to the offset channel.

Further tuning the model by reducing the width of the class-3 patch (i.e. first break class) of the input image can help delineate class boundaries better. Using an ensemble-model of the available outputs can also be advantageous. Having better techniques to predict and resolve outliers for noisy first break picks will improve the quality of the dataset and, ultimately the performance of the model.

Limitations

The experiments in this report were performed on datasets converted to a certain sampling rate and length. This configuration could prove to be an obstacle for predicting on unknown datasets of widely different physical characteristics and geometry. Moreover, we could only reach conclusions on certain hyperparameters and not all. Many experiments ran on MultiResUNet with different configurations could not be directly compared with UNet due to the lack of experiments with the same hyperparameters. Thus, our work, while showing promising results, also requires more data on varied experiments before any strong recommendations can be made. Applicability to seismic data acquired in different geological environments or with other types of sources (i.e. other than explosives) also remains to be tested.

Acknowledgements

I (BP) would like to gratefully thank my NRCan supervisor Gilles Bellefleur, and Mila supervisors Pierre-Luc St-Charles and Bruno Rousseau for their mentorship, time and reviewing this report. I offer my sincere appreciation for the learning opportunities provided during my time at NRCan. We thank E. Schetselaar for the review of the manuscript.

References

Alaudah, Y., Michalowiczy, P., Alfarraj, M., AlRegib, G., 2019. A Machine Learning Benchmark for Facies Classification. Interpretation, 7, 3, pp. SE175–SE187. https://doi.org/10.1190/INT-2018-0249.1

Bianco, E., 2011. G is for Gather. URL: https://agilescientific.com/blog/2011/9/14/g-is-for-gather.html.

Chevitarese, D.S., Szwarcman, D., Mozart, R., Brazil, E. V., 2018. Deep Learning Applied to Seismic Facies Classification: a Methodology for Training. In: Saint Petersburg 2018. EAGE Publications BV. DOI: 10. 3997/2214-4609.201800237.

Civitarese, D., Szwarcman, D., Brazil, E. V., Zadrozny, B., 2019. Semantic Segmentation of Seismic Images. arXiv:1905.04307 [eess.IV].

Coppens, F., 1985. First arrivals picking on common-offset trace collections for automatic estimation of static corrections. Geophysical Prospecting, 33, 1212–1231.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep Sparse Rectifier Neural Networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Ed. by Geoffrey Gordon et al. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, pp. 315–323. URL: http://proceedings.mlr.press/v15/glorot11a.html.

Hall, M., 2013. Offset Geometry: URL: https://subsurfwiki.org/wiki/File:Offset_ geometry.png.

Han, L., Wong J., Bancroft J. C., Stewart, R., 2009. Automatic time-picking of first arrivals on noisy microseismic data. CSEG CSPG CWLS Convention, Conference Abstracts.

Ibtehaz, N., Rahman, M.S., 2020. MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural networks : the official journal of the International Neural Network Societya.

Korvin, G., 1992. Fractal models in the earth sciences. Elsevier Amsterdam, 396 p.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal Loss for Dense Object Detection, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Tosinski, J., 2018. An Intriguing Failing of Convolutional Neural Networks and the Co-ordConv Solution. NIPS'18. Montréal, Canada: Curran Associates Inc., 9628–9639.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. ICLR 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

McCormack, M. D., Zaucha, D.E., Dushek, D.W., 1993. First-break refraction event picking and seismic data trace editing using neural networks. Geophysics, 58, 1, pp. 67–78. DOI: 10.1190/1. 1443352.

Mężyk, M., Malinowski, M., 2019. Multi-pattern algorithm for first-break picking employing open- source machine learning libraries. Journal of Applied Geophysics 170, p. 103848. DOI: https://doi.org/10.1016/j.jappgeo.2019.103848.

Milesi, A., 2017. Pytorch-UNet. URL: https://github.com/milesial/Pytorch-UNet.

Ronneberger O. Fischer P., Brox T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. https://arxiv.org/pdf/1505.04597.pdf

Sabbione, J. I., Velis, D., 2010. Automatic First-Breaks Picking: New Strategies. Geophysics, 75, 4, pp. V67-V76..

Smith, L. N., 2017. Cyclical Learning Rates for Training Neural Networks. arXiv:1506.01186 [cs.CV].

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv:1803.09820 [cs.LG].

Smith, L. N., Topin. N., 2018. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv:1708.07120 [cs.LG].

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, pp. 1–9. DOI:10.1109/CVPR.2015.7298594.

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S.,Trujillo, J., 1998. "A comparison of select trigger algorithms for automated global seismic phase and event detection". Bulletin of the Seismological Society of America 88, 1, pp. 95–106. eprint:https://pubs.geoscienceworld.org/ssa/bssa/articlepdf/88/1/95/2709205/BSSA0880010095.pdf

Yuan, S., Liu, S., Wang, S., Wang, T., Shi, P., 2018. Seismic Waveform Classification and First-Break Picking Using Convolution Neural Networks. IEEE Geoscience and Remote Sensing Letters, 15, 2, pp. 272– 276. DOI: 10.1109/LGRS.2017.2785834.