2021 EAGE ANNUAL
82ND CONFERENCE & EXHIBITION
AMSTERDAM | THE NETHERLANDS

DATA-DRIVEN METHOD FOR TRAINING DATA SELECTION FOR DEEP LEARNING

C. Lacombe[1], I. Hammoud[1], J. Messud[1], H. Peng[1], T. Lesieur[1], P. Jeunesse[1]

[1] CGG

Summary

Convolutional based deep neural networks can be used in addition to existing workflows, to improve turnaround or as a 'guide' for further processing. Whilst a lot of effort has been made to try to improve the DNN architecture for processing tasks or to understand their physical interpretation, the choice of the training-set is rarely discussed. For a good quality DNN result, the training-set must be representative of the variability (or statistical diversity) of the full dataset, and the question of the choice of this dataset for seismic data is discussed in this paper. We present two methods for the selection of the training set. The first one is based on proxy attributes and their clustering. Our clustering approach is not only using the clusters themselves but also the information on the distance to the centroid for the cluster definition. The other method is based on the data themselves. It starts from a predefined training set and then scans through the full dataset to identify additional training points that will be used to augment the initial training set.

Data-driven method for training data selection for deep learning

Introduction

Deep Learning (DL) for seismic processing has gained interest in the last few years and is an active field of research. Convolutional-based deep neural networks (DNNs) can be used to learn to mimic physics-based processing algorithms results, to improve turnaround or as a 'guide' for further processing (see Mandelli et al. 2019, Richarson and Feller 2019). However, three key challenges must be addressed to build confidence in adopting DNNs routinely in seismic processing workflows: (1) an intelligent selection of the training set with minimal human intervention; (2) the robustness of the DNN results to noise and uncertainties; (3) the design of optimum DNN architectures and its components (Sun and Demanet 2018, Chambefort and Messud 2020, Messud and Chambefort 2020). In this paper, we focus on point (1). Although the choice of the training set can be a main determinant of the quality of DNN predictions, a discussion of this aspect seems so far absent in the seismic field.
We consider the case where we dispose of a full (raw or input) dataset and use a small subset of it as a training data. The subset is then processed by a physics-based algorithm to provide the labels ("ground truth" - see Figure 1 in the case of deghosting process). Finally, we apply the trained DNN model on the full data with the objective that the results have similar quality as the physics-based method.
Obviously if the training set is too small or insufficiently sampled, it will not capture the variability of the full data, leading to poor result. Conversely, if we (brute force) sample the whole dataset and create a large training set, the quality of the DNN results should be very good but will come at a prohibitive cost on label generation and training process (Hou and Hoeber 2020).
In the following sections, we start by showing that using regularly spaced sail-lines as training data can fail to capture the data variability needed for seismic deghosting. We then introduce two new approaches for potential improvement. The first approach uses a prior attribute clustering method based on geological horizons. We use the distance to centroids in the clustering method and do not put any constraints on the location of the selected data. The second method starts from a predefined training set, and then scans through the full dataset to identify additional training samples to augment the initial training set. Note that both methods allow irregularly sampled selection, in order to better capture the data variability.

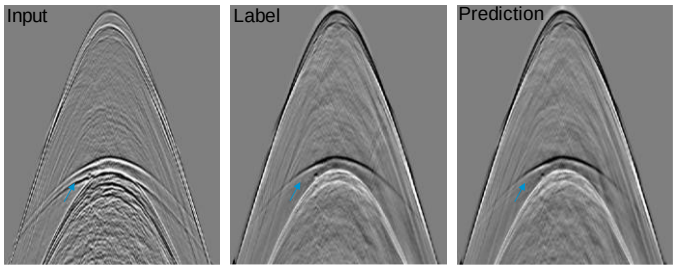

Figure 1 *– Shot point (Left) before deghosting, (Middle) after physics-based source and receiver deghosting (label) and (Right) after DNN prediction.*

Impact of the training set on DNN results

We use a marine dataset acquired with the source-over-spread acquisition technology (Vinje et al. 2017) to demonstrate the impact of the training set on the source and receiver deghosting DNN result. As the whole 5200 km^2 of the dataset (1.4 million shots) has been processed with physics-based deghosting workflow (Wang et al. 2013), this gives us a complete labelled dataset to test various training set configurations and to compare the DNN results with the labelled data on the entire volume. In this study, we adopted the DUnet DNN architecture described in Peng et al. (2021).
Figure 1 shows a raw shot point (Left), the corresponding label (Middle) and the DNN prediction (Right). To assess the quality of the results, we calculate for each shot record the RMS

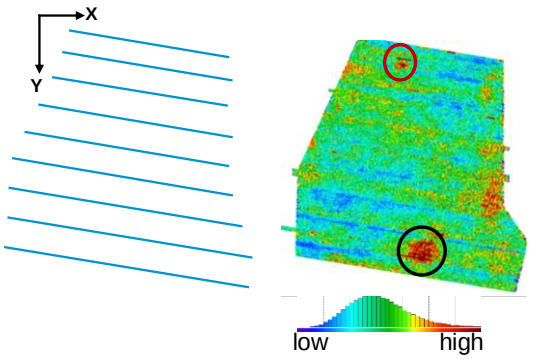

Figure 2 *– (Left) Training dataset (regular lines). (Right) RMS error for the initial training set. Blue and red indicate low and high error respectively.*

of the difference between the DNN prediction and the label on a given time window and define this as the "RMS error". In Figure 2 (Right), we show the distribution of RMS error (shot-record x-y) for all the shots across the whole survey, using a training set from regularly (every 15km) spaced sail-lines as outlined on Figure 2 (Left). Blue and red indicate respectively areas of low and high error. A salt feature (red circle) and some gas inclusions (black circle) are particularly badly predicted as these localized features are not represented in the given training set. This indicates that geology is one of the main factors influencing data variability. Clearly, we could improve the result by choosing a much more densely sampled training dataset, but this would not be cost effective and would need repetitive human quality control. Instead, we discuss two alternative approaches to improve the training data variability.

Training data selection from clusters defined using proxy attributes

We propose a method whereby the shot points to be included in the training set are selected based on (geologic) proxies, with no requirement for them to be located on a regular grid. Proxies are priors that must be defined according to the processing task, e.g., dispersion panels for surface wave attenuation (Masclet et al. (2020)). Assuming geology as a dominant contributor to data variability, we choose four geological time horizons as proxy for data variability. A k-mean algorithm is used to geologically cluster the four time horizons simultaneously. Conventional analysis to assess the optimum number of classes

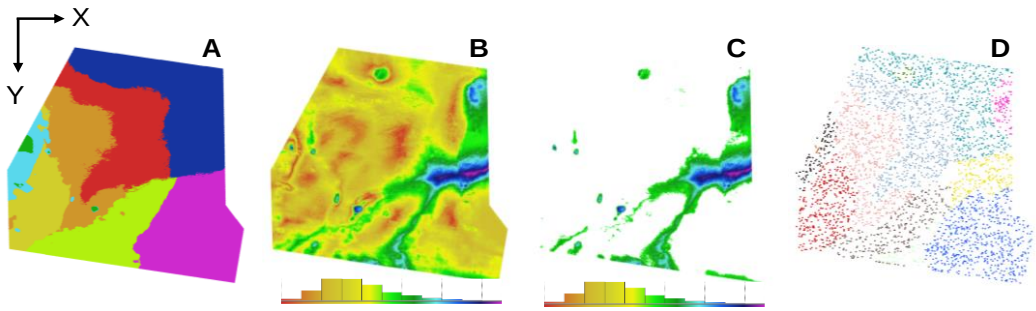

Figure 3 *– A- Clusters obtained using the four horizons (8 clusters, each cluster is represented by one color). B- Distance to the centroid of the cluster. Blue indicates high distance to the centroid. C- Outliers from map B used to form new clusters. D- Positions of the training points (when selecting a percentage of data points per cluster), the colors corresponding to the 12 updated clusters.*

gives around 20 clusters. However, we observed large variations in the cluster distribution for small changes in the number of clusters, so we use only 8 clusters as shown in Figure 3A (each color represents a cluster). The distance from each shot point to its own cluster's centroid is then analyzed (Figure 3B). Areas exhibiting distances to centroid larger than one standard deviation from the median value were isolated (Figure 3C), and four new clusters were added, bringing the total number of clusters to 12 (Figure 3D, showing individual data points, colors corresponding to the clusters).

Each cluster now needs to be 'sufficiently represented' inside the training set. Since within each cluster the geology and hence the data are similar, there are two methods to sample these clusters: (A) by taking a fixed constant percentage of data points (Figure 3D) or (B) a fixed constant number of data points per cluster. Both selection methods were tested to create two training sets of similar size. The prediction RMS errors for both cases are shown in Figure 4A and 4B for sampling methods (A) and (B). The level of average error is similar in both cases and a small improvement is seen (black circles) at some locations when a constant number of points is used

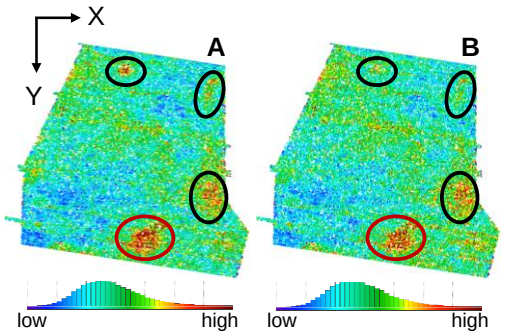

Figure 4 *– RMS error between physics-based deghosting and DNN prediction when the training set contains (A) 1% of data in each cluster (B) a constant number per cluster. Blue and red indicate low and high error respectively.*

per cluster (sampling method B). The black circled areas correspond to the extra clusters introduced by the distance to the centroid, indicating that these areas were indeed not well represented by the initial clustering and that more importance is given to them when a constant number of points per cluster (method B) is used. The area circled in red in Figure 4 is poor with both of these methods. It contains some gas inclusions and is not seen as a separate cluster, as these gas inclusions are not included in the horizon information.

The main shortcoming of the method is that it relies on the chosen prior proxies. Here we have chosen time horizons as proxies for geology; this has succeeded for the main features of the data, but not for gas inclusions; AVO or other noise sources would also be problematic. To overcome this limitation, the method could be extended to other proxies or multi-proxies, but it may prove difficult to find proxies representative of all features of the data. Also, in this deghosting process, the data are grouped on a sail-line order, which could pose operational challenges for geological proxy-based data selection. Therefore, a second method which is augmenting a predefined initial training set (here sail-lines) is proposed.

Semi-automatic training data augmentation from a predefined initial training set

This method relies directly on the data similarity rather than on proxies and can be easily integrated in a processing sequence. We start by defining as initial dataset, hereafter called 'anchor', some specified subset of the data. The anchor dataset is flexibly defined according to the acquisition set-up (e.g., sail-lines, cross-spread) or even randomly if the processing flow allows it. In our case, as deghosting is applied in a sail-line manner, we use regularly spaced sail-lines.

A pairwise similarity measurement is then calculated between all input shot points and each "anchor" shot point. This similarity measurement could be an Lp distance, a Wasserstein distance or any other similarity metric. We found that the L2 distance worked well in our study.

Figure 5 shows the smallest similarity measurement value for each shot of the entire volume, with large outliers in red. The location of the anchor dataset is represented by the white lines. We observe that the method identifies the gas inclusion area (polygon at the bottom) as an outlier requiring additional training data. This was not the case with the proxy method presented in the previous section. From this map, black polygons outline the areas where there is a large difference compared to the anchor data (in black on Figure 5) and a subset of shot points in the polygons is added to augment the sail-line based anchor training set as shown as red dots in Figure 6 (Left).

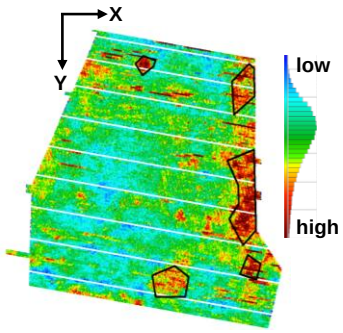

Figure 5: *Map of the smallest similarity measurement (L2) value for each shot. Polygons around areas of poor similarity (large values) are used to select extra shots for the training.*

Figure 6 (Right) show the RMS error of the DNN prediction after training with the augmented training sets. This error map can be compared to the one from Figure 2 (Right), where the error for the initial anchor training set is shown. The mean RMS error value is at a similar level for both methods and a reduction of the outliers is observed when the updated training set is used. Around the gas inclusion the RMS error has not been completely reduced.

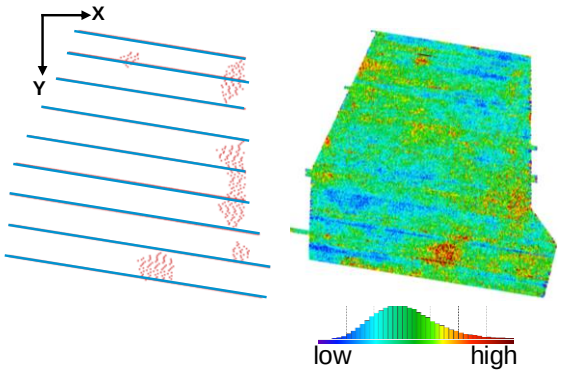

Figure 6: *(Left) Initial training dataset (blue) and data added after further data selection (red). (Right) RMS error for the updated training dataset. Blue and red indicate low and high error respectively.*

This could be related to the sampling density of the polygon areas.
In this comparison, the number of training points is larger for the updated dataset than for the initial one by 9%. However, we observed that increasing the number of points in the initial training set by simply densifying the sampling along the lines does not improve the outlier areas, meaning that the impact seen is indeed coming from added variability in the training set.
Not surprisingly, we can also observe stripes on these RMS error maps which are linked to better prediction around the densely sampled lines of the training set. Here we have chosen sail-lines as anchor, but other anchor geometries could have been chosen, for example random shot points. Alternatively, to avoid possible bias linked to the choice of the anchor dataset, the method can be iterated.
Finally, as a similarity is measured with this method, both the signal and noise variability can theoretically be assessed. The impact of the noise content on the method will be dependent on the chosen similarity measurement. This has not been studied here and would require further investigation.

Conclusion and way forward

In this study, we have highlighted the importance of data variability in DNN training. We have tested two approaches to select training sets representative of data variability. The first one is based on proxy attributes. This is a very simple method in principle, but it needs a prior understanding on the cause of data variability. We assumed a single cause, geology, and a single proxy, time horizons; this did not give fully satisfactory results in the deghosting case. The second method is based on the data themselves and needs no a-priori knowledge. It can be used to augment the initial training set, as it identifies data that are less represented by the anchor (initial) data. Further analysis is required to understand how to identify the sources of data variability, and deciding the optimum selection scheme and training size. As more complex training scheme and larger training data comes at a cost, this remains an important open question to be addressed for DNN to become a cost-effective alternative in seismic processing.

Acknowledgements

We are grateful to CGG, CGG Multi-Client and TGS for the permission to publish this work, and to M. Chambefort, D. Le Meur, G. Poulain, N. Salaun, G. Lambaré and H. Hoeber for their involvement.

References

Chambefort, M. and J. Messud [2020] Building and understanding deep neural network components for seismic processing: lessons learned, *82nd EAGE Conference & Exhibition Workshop Programme*.
Mandelli, S., Lipari, V., Bestagini, P. [2019] Interpolation and Denoising of Seismic Data using Convolutional Neural Networks. *arXiv:1901.07927v4*.
Masclet, S., T. Bardainne, V. Massart and H. Prigent [2019] Near surface characterization in Southern Oman: Multi-Wave Inversion by Machine Learning, *81st EAGE Conference & Exhibition, Expanded Abstracts*.
Messud, J. and M. Chambefort [2020] Understanding how a deep neural network architecture choice can be related to a seismic processing task, *First EAGE Digitalization Conference and Exhibition.*
Peng,H. Messud, J., Salaun, N., Hammoud., I., Jeunesse, P., Lesieur, T. and C. Lacombe [2021] DUnet architecture for seismic processing tasks - Proposal and theoretical analysis, *83rd EAGE Conference & Exhibition, Submitted.*
Richardson, A. and C. Feller [2019] Seismic data denoising and deblending using DL, *arXiv:1907.01497*
Sun, H. and L. Demanet, [2018] Low frequency extrapolation with deep learning. *SEG Technical Program Expanded Abstracts*, 2011-2015.
Hou, S. and H. Hoeber [2020] Seismic processing with deep convolutional neural network; opportunities and challenges, *82nd EAGE Conference & Exhibition, Expanded Abstracts.*
Vinje, V., Lie, J.E., Danielsen, V., Dhelie, P.E., Siliqi, R., Nilsen, C.I., Hicks, E. and A. Camerer [2017] Shooting over the seismic spread, *First Break*, vol 35, pp97-104
Wang,P., Ray, S., Peng, C., Li, Y., and G. Poole [2013] Premigration deghosting for marine streamer data using a bootstrap approach in tau-p domain, *SEG Technical Program Expanded Abstracts*, 4221-4225.