IPTC-21940-EA

Improving Machine Learning Approaches to Seismic Fault Imaging Through Training Augmentation

Philip Norlund and Fan Jiang, Halliburton Landmark

Copyright 2022, International Petroleum Technology Conference DOI 10.2523/IPTC-21940-EA

This paper was prepared for presentation at the International Petroleum Technology Conference held in Riyadh, Saudi Arabia, 21 - 23 February 2022.

Abstract

Fault interpretation is critical for understanding subsurface challenges such as fluid migration and avoiding drilling hazards. Recently, Convolutional Neural Networks (CNN) have been shown to be effective tools for identifying faults in seismic data by utilizing image segmentation. However, selecting the optimal training data for building a model can be challenging. Fault shapes and relationships are highly variable due to the complexity of the regional geodynamic processes happening within the earth over time. How to properly select a training subset from a real (i.e., non-synthetic) data-set is crucial for the success of machine learning for seismic interpretation. In this paper, we attempt to quantify how models can be improved by augmenting the training data in various ways. These augmentations include varying the amount of real and synthetic data and including combinations of seismic attributes.

Summary

Fault interpretation is critical for understanding subsurface challenges such as fluid migration and avoiding drilling hazards. Recently, Convolutional Neural Networks (CNN) have been shown to be effective tools for identifying faults in seismic data by utilizing image segmentation. However, selecting the optimal training data for building a model can be challenging. Fault shapes and relationships are highly variable due to the complexity of the regional geodynamic processes happening within the earth over time. How to properly select a training subset from a real (i.e., non-synthetic) data-set is crucial for the success of machine learning for seismic interpretation. In this paper, we attempt to quantify how models can be improved by augmenting the training data in various ways. These augmentations include varying the amount of real and synthetic data and including combinations of seismic attributes.

Methods

The prediction accuracy from supervised deep learning models largely depends on the amount, condition, and diversity of data available during training. The quality of training data is key to achieving high performance for complex tasks, such as seismic fault interpretation. Additionally, a regular CNN has a large number of hidden neurons, which also increases trainable parameters and makes it difficult to reduce overfitting. Data augmentation in a machine learning project is used to increase the amount of data by adding modified copies or newly created synthetic data, as regularization, to help reduce overfitting. In image processing, this problem could be addressed by data augmentation schemes, such as adversarial training-based augmentation (Baek et al., 2019) or neural style transfer-based augmentation (Zheng et al., 2019). In a deep learning architecture, those schemes include geometric transformation, color modification, rotation, and cropping, which are widely used to augment training data (Figure 1).

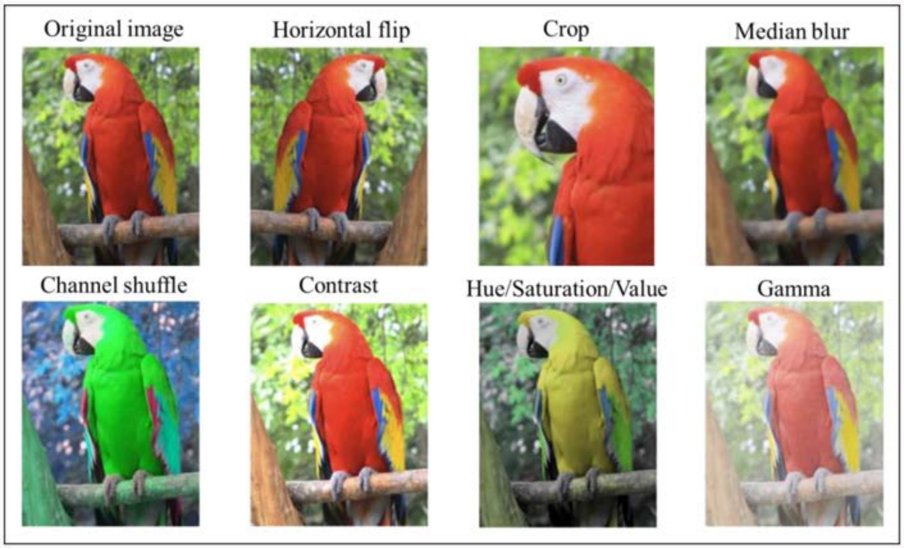

Figure 1—Image augmentation used for training a machine learning image segmentation model (Modified from Buslaev et al., 2018).

In seismic interpretation, there are two augmentation methods developed to improve the training accuracy and reduce overfitting. The first augmentation method is to increase the amount of training data by adding synthetic data and their attributes to expand the variety of seismic characteristics. The advantage of using synthetic data to augment the training model is that we have a ground-truth label to be trained and thus should be able to minimize uncertainty. However, due to the complexity of the subsurface structures (AlRegib et al., 2018), augmentation by synthetic data could also miss providing specific seismic features in the survey area, which may result in an incomplete training model. In this case, a small set of manual interpretation from real, field data is useful to help augment the training model and provide the necessary data types in the testing area.

For a complete assisted fault interpretation process, training a model with all kinds of fault types, such as normal faults, reverse faults, or listric faults, is necessary. Figure 2 shows several synthetic seismic data examples and their fault labels used to train a model to predict fault probability maps. Different fault types could help a model learn and better teach it how to distinguish a "fault" or "no-fault" by optimizing the weighting coefficient and minimizing the loss function. Wu et al (2019) used a class-balanced binary cross-entropy loss function to adjust the imbalance so that the network is not trained or converged to predict only zero. This helps to reduce overfitting and train a more accurate model.

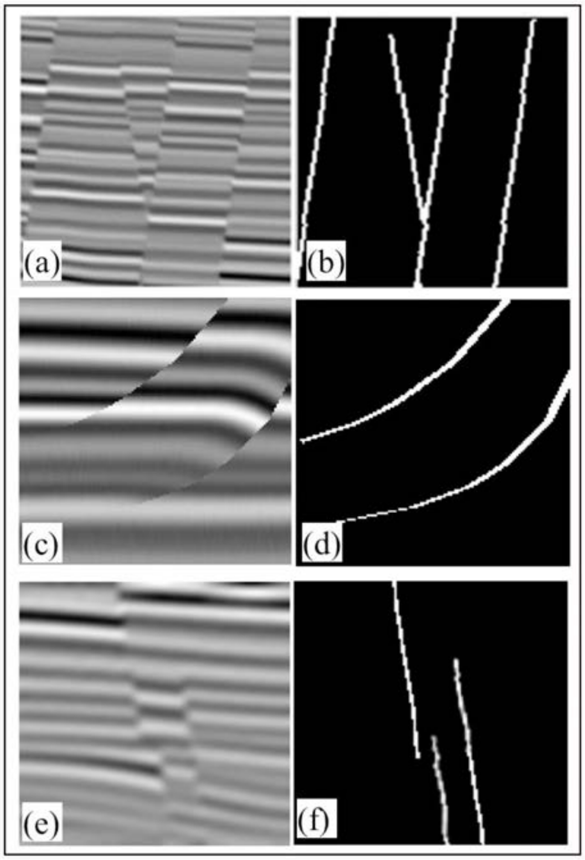

Figure 2—Synthetic seismic data and fault labels used to train a machine learning model. Those data show some known fault types such as normal faults (a) and their label (b), listric faults (c) and their label (d), reverse faults (e), and their label (f).

Once we have trained a machine learning model using synthetic data, it is possible to predict fault probability maps from different survey areas. However, due to the complexity of fault formations in the real world, synthetic data may not correctly represent all potenmtial fault images, which can subsequently lead to flawed or erroneous interpretation. Augmenting the pre-trained model by including local, manually interpreted fault data could help reduce uncertainty and remove problems such as inconsistent fault throws and other non-geological representations. Therefore, in the second step, we manually interpret several fault planes from a real dataset, then convert those fault planes to a binary fault mask which is used to distinguish a "fault", as "1", or "no-fault" as "0". This processed real data is combined with synthetic data together to train an augmented machine learning model. There is an argument that if this data is needed to train a model from scratch, or implement a transfer learning scheme, then we should only update the last few layers from a neural network. However, it depends on how much new data is provided. If there is only a small amount of new data interpreted, the transfer learning scheme could perform more effectively. If a large amount of new interpreted data exists, training from scratch may be the better choice with improved performance as well as bringing low variance and bias.

Figure 3 shows a modified workflow from Norlund and Jiang (2021) to train a synthetic model which includes field data to augment the pre-trained synthetic model. In this experiment, we first applied multiple attributes to augment the original training data, then provided manually interpreted faults to augment the machine learning model trained by synthetic data. The fault probability volume predicted by a synthetic-data-trained model could be used by geoscientists during the initial interpretation phase to guide their work, then the second prediction by the augmented model from real data could help to identify any missed or alternate interpretations. This further analysis can both improve the quality of the interpretation and bring insight into any uncertainty around the interpretations (Jiang and Norlund, 2021).

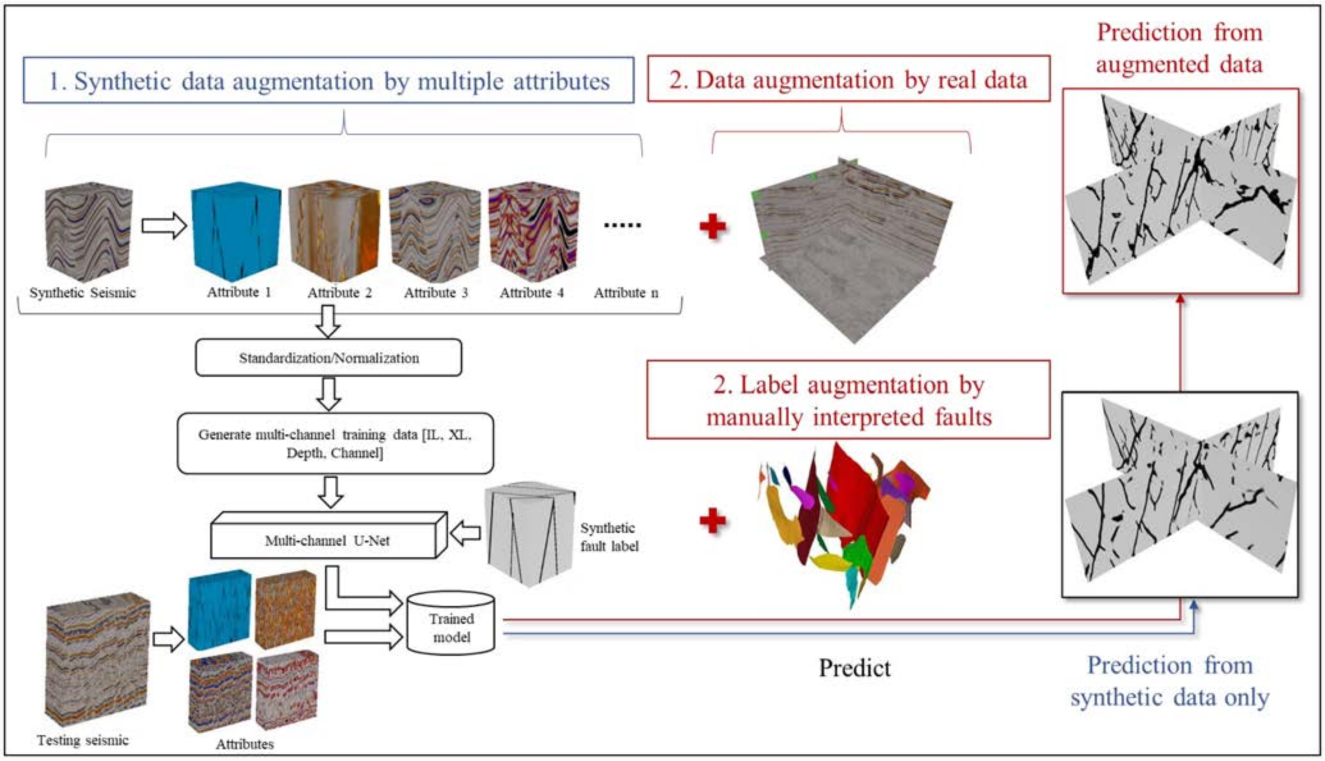
1. Synthetic data augmentation by multiple attributes
2. Data augmentation by real data
Prediction from augmented data
Synthetic Seismic
Attribute 1
Attribute 2
Attribute 3
Attribute 4
Attribute n
Standardization/Normalization
Generate multi-channel training data [IL, XL, Depth, Channel]
2. Label augmentation by manually interpreted faults
Multi-channel U-Net
Synthetic fault label
Trained model
Testing seismic
Attributes
Predict
Prediction from synthetic data only

Figure 3—A workflow that describes two data augmentation schemes. The first is data augmentation by multiple attributes, and the second is data augmentation by real data interpretation. (Modified from Norlund and Jiang, 2021)

Results

We applied the proposed augmentation schemes to field data from Northwest Australia. In this experiment, multiple machine-learning models for fault prediction wer created using variations in the input training data. A multi-channel CNN architecture was used for all the models (Ronneberger et al., 2015; Jiang and Norlund, 2020). Our first models were trained on synthetic data and only included the amplitude seismic data. Our next models were trained with synthetic data including varying sets of seismic attributes. Our final models included both synthetic data and varying combinations of real data. All models were then run on a seismic dataset from the study area to generate fault-probability volumes. These volumes were then compared to professionally interpreted fault-planes (Figure 4).

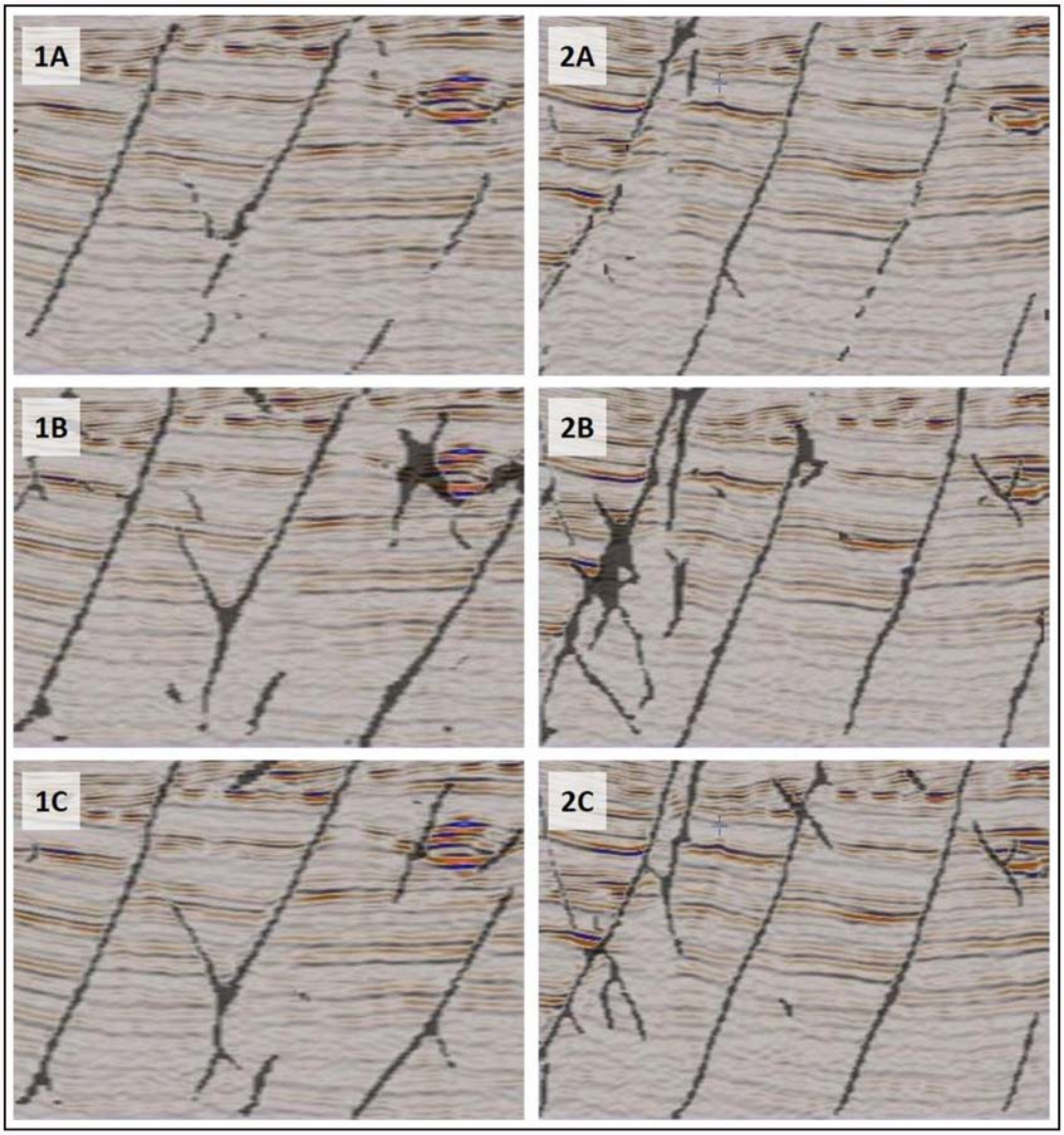

Figure 4—Examples of machine learning fault prediction volumes created using; synthetic data from a single attribute only (1A, 2A), synthetic data from multiple seismic attributes (1B, 2B), a model augmented with synthetic and real data and with multiple seismic attributes (1C, 2C). Data courtesy of Geoscience Australia.

As we augmented the training data with seismic attributes and real data (manual fault interpretations from the survey), we generally saw a progressive improvement in the prediction quality. More faults were detected, and the fault imaging became clearer and more continuous. However, we also saw distinct variations in the prediction quality dependent on how much real data was used for augmentation, and what areas of the survey the manual interpretation came from. The study showed that it is important to understand the limitations of combining real and synthetic data. Any bias or errors in the manual interpretations can be incorporated into the model and thus be propagated throughout the output fault probability volume. Also, it is unlikely all faults in a seismic subset will be interpreted as there is often neither the time nor need, to interpret every fault present. In these situations, the model is trained not to recognize valid faults which further compromises the results. While these challenges do not invalidate the use of real data in training models for seismic interpretation, it is crucial to understanding the limitations of this approach to extract the full value from the machine learning predictions.

The reason for using synthetic data is that examples can be created where there is only one possible correct interpretation, the model is always being trained with the correct answer. Real seismic data, however, is not so simple. Interpretation in real seismic data is often a non-deterministic challenge with multiple answers available for every seismic section analyzed (Alcalde et al., 2017, Jiang and Norlund, 2021). This uncertainty also highlights some of the limitations of including non-synthetic data. First, even in the small subset of the volume, we did not select every fault in the seismic that was interpreted. Those faults could reduce the model's ability to identify all other faults in the volume. Second, interpretation can be a highly subjective process. Even experienced interpreters could interpret the same seismic section in many ways. By including just one interpreter's interpretation into the model, you are also adding that person's biases into the model and thus limiting the ability of the model to generate genuine and valid predictions.

Conclusions

Using Convolutional Neural Networks to image faults in seismic has become increasingly common over the past few years. However, little research effort has been put into understanding the effects of combining real and synthetic data in the training process. In this paper, we proposed several augmentation methods aiming to quantify the ways varying training inputs affect prediction results and recommend best practices for optimizing results.

References

Alcalde, J., C. E. Bond, G. Johnson, R. W. H. Butler, M. Cooper, and J. Ellis, 2017, The importance of structural model availability on seismic interpretation: *Journal of Structural Geology*, **97**, 161–171.

AlRegib, G., Deriche, M., Long, Z., Di, H., Wang, Z., Alaudah, Y., Shafiq, M., and Alfarraj, M., Subsurface structure analysis using computational interpretation and learning: A visual signal processing perspective, *IEEE Signal Processing Mahazine*, v**35**, p82–98.

Baek, F., Park, S., and Kim, H., 2019, Data augmentation using adversarial training for construction-equipment classification, *Image and Video Processing*, arXiv:1911.11916.

Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V. and Kalinin, A., 2018, Albumentations: fast and flexible image augmentations, arXiv:1809.06839v1.

Jiang, F. and Norlund, P., 2020, Seismic attribute guided automatic fault prediction by deep learning, EAGE Extended Abstract.

Jiang, F. and Norlund, P., 2021, Assisted fault identification and surface extraction by machine learning, a case study from Oman, *First International Meeting for Applied Geosciences and Energy*, Expanded Abstract.

Norlund, P. and Jiang, F., 2021, The evolution of assisted fault interpretation – Part 2, *Subsurface Insights*, p13–19.

Ronneberger, O., Fischer P., and T. Brox, 2015, U-net: Convolutional networks for biomedical image segmentation: *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Wu, X., L. Liang, Y. Shi, and S. Fomel, 2019, FaultSeg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation: *Geophysics*, **84**, no. 3, IM35–IM45,

Zheng, X., Chalasani, T., Ghosal, K., Lutz, S. and Smolic, A., 2019, STaDA: Style Transfer as Data Augmentation, Computer Vision and Pattern Recognition, arXiv:1909.01056.