Identifying Geologic Facies Through Seismic Dataset-to-Dataset Transfer Learning using Convolutional Neural Networks

Rachael Wang
06214609
rachaelw@stanford.edu

Joseph Stitt
06175027
jdstitt@stanford.edu

Adam Shugar
06164615
ashugar@stanford.edu

Problem Statement and Motivation

One of the most time consuming tasks in the reservoir characterization and interpretation phase of exploration projects within energy companies is to correctly interpret seismic facies in order to determine source, reservoir, and seal potential. This process is costly since interpretation is completed manually. By using a supervised method, U-Net CNN segmentation, we allow for the automatic classification of seismic facies, which could provide a good starting point for geologists and geophysicists in the upstream sector. We also implement a transfer learning algorithm that can carry over learned model parameters directly from one 3-D seismic dataset to another. These weights were then used as a starting point to help train on the second seismic dataset in order to save resources and improve efficiency by using a combination of the already learnt trivial features of subsurface geology to recognize new facies types. To our groups knowledge, this is the first known attempt at applying transfer learning between the Parihaka data set and the Netherlands F3 data set and one of few documented attempts at implementing transfer learning directly from one 3-D seismic dataset to another [1]. Others have tried to implement transfer learning with a seismic dataset, but used well known pretrained models such as Imagenet instead of direct seismic datasets [2]. We took inspiration from this, and decided to try to do it with direct dataset-to-dataset transfer learning. Our main tuning goal with the implementation of transfer learning is to find the best test prediction result, with the lowest amount of training and validation examples used. With a successful result using a low amount of training and validation examples from the applied seismic dataset, this will allow geoscientists with a limited amount of labeled data to be able to still train their small dataset, and with the power of transfer learning, be able to get high accuracy facies prediction results. This would then allow geoscientists to start their projects' with a solid starting interpretation of the rock layers with a minimal amount of hand labeling done for the the training process of their own dataset.

Dataset Information

Stage I: Parihaka

For our training data, we used a public-domain survey called, "Parihaka," which contains offshore seismic data from the New Zealand government [4]. The dataset consists of .segy files that were converted into .npz files for easier use with Python. The entire dataset is a 3-D seismic cube which consists of contiguous 2-D slices. A single 2-D slice of dimensions 590 x 1006 is used as the input image, meaning we have 593,540 input features. There are 782 2-D slices which combine to create the 3-D cube. The labeling scheme is one where each pixel in an input slice is classified into a distinct facies category out of six possible: "basement", "slope mudstone A", "slope mudstone B", "mass transport deposit", "slope valley", and "submarine canyon system". The label data has the same dimensions as the features since each pixel corresponds to a category label. The test and training datasets were provided in separate files and they come from different sections of the 3-D dataset. Unfortunately, labels for the test dataset were not provided to compare our predictions. Thus, manual inspection of the true 3-D seismic cube was required after we received our predictions from our test set to compare the facies. We used GAIN and RMS to improve the amplitudes within the image, both normalizing the amplitudes and improving the contrast between differing amplitudes.

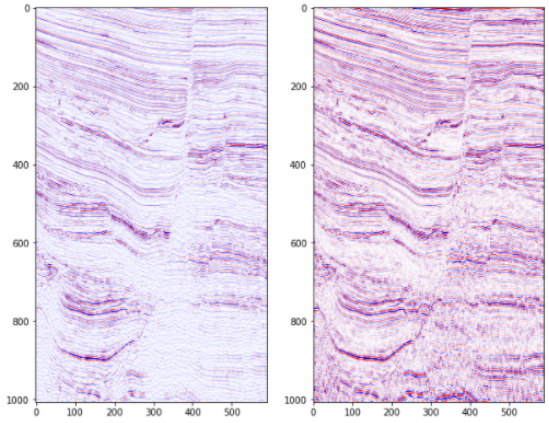


Figure 1: Training dataset examples from Parihaka (left) is before pre-processing (right) is the image after gain/RMS was applied.

Stage II: Netherlands F3

Our second dataset, the Netherlands F3 dataset, was employed to implement transfer learning [5]. We used a smaller set of examples from the "Netherlands F3" to test whether the learned model parameters from the "Parihaka" data are able to provide significant information to help classify facies in situations where limited seismic data is provided. This dataset is similar enough to allow for a smoother transition of information in the transfer learning process, however it is prone to overfitting. Again, the training and test datasets were provided in seperate files and each set of data comes from different sections in the 3-D seismic cube. We ended up splitting some of the test set off for validation examples when we were training the model. Unlike the Parihaka dataset, we were provided labels for the test dataset to compare our predictions. Example image inputs for both the training dataset and the validation dataset are shown in Figure 2.

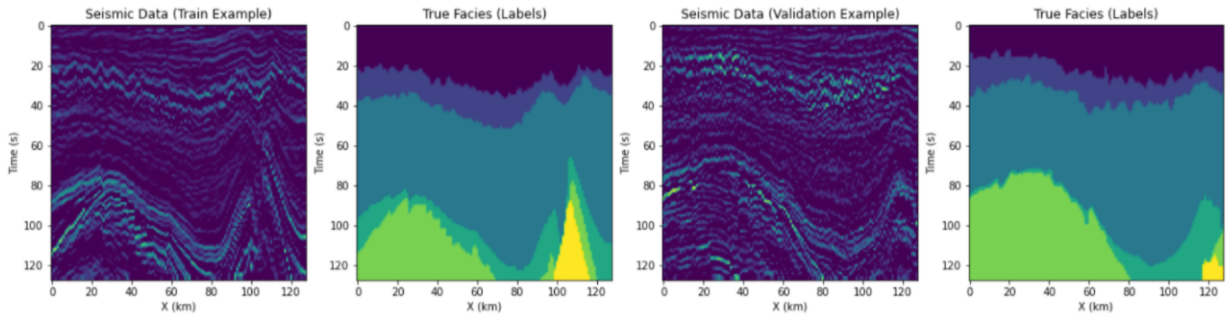


Figure 2: Training and validation data with their corresponding facies masks. Note the complex geology of the deeper facies. These facies boundaries will become a challenge for the deep learning model to predict given a set of test data.

Architecture Choices

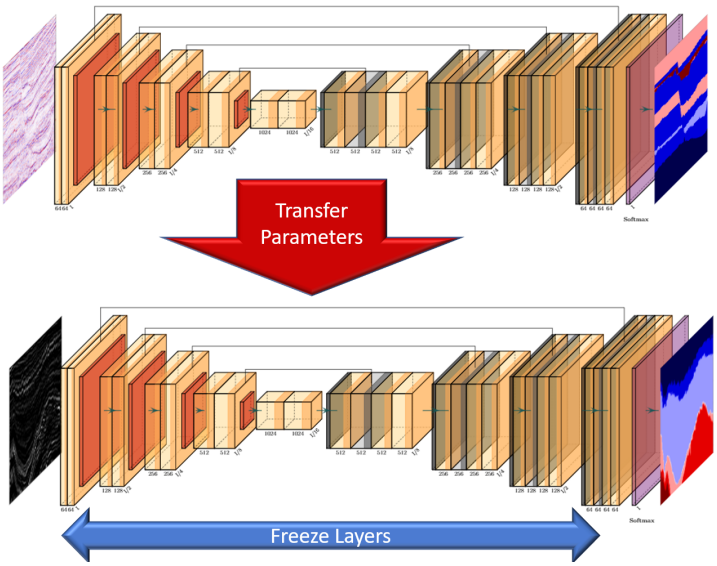


Figure 3: U-Net architecture diagram with the integrated idea of transfer learning. The blue arrow represents the fine tuning process of freezing different layers during transfer learning. The image representations are actual image examples originating from the Parihaka (top) and the F3 Netherlands datasets (bottom).

We used a U-Net architecture for this image segmentation problem. U-Nets are a CNN architecture for semantic segmentation, named after the "U" shape created by the connection between the encoding and decoding portions of the network [3]. The U-Net consists of two main parts: a contracting/encoding path for capturing context and information in the image and an expanding/decoding path which localizes the context and information captured by the encoding path. Since we pre-processed our input images into a final size of 128x128 pixels, we chose a moderate network size with approximately 31 million trainable parameters across 22 convolutional layers.

For the second part of this project, we used the model parameters from the Parihaka model as a starting point for the training process on the F3 Netherlands dataset using transfer learning. In Figure 3, model parameters from the upper U-Net baseline model are transferred to the second U-Net model with F3 Netherlands dataset. As seen by the blue arrow in the figure, frozen layers or starting model layers we do not want retrained can be adjusted to obtain maximum performance. We also added an additional two dropout and dense layers after the pre-trained model with a final additional softmax output layer. These will help with overfitting issues stemming directly from the pretrained model and narrowing down unique segmentation identifications based on the second dataset.

Hyperparameter Tuning

Stage I: Training Baseline Model

In terms of tuning our hyperparameters for our baseline or pretrained model, in order of importance, we tuned the learning rate, number of epochs, the addition or removal of batch normalization layers, and finally the dropout ratio value. Unlike the

exponential learning rate decay we used for the transfer learning tuning, we still used an exponential scale, but we didn't need the decay aspect as we were able to obtain $> 95\%$ validation accuracy without using it. Thus, we started with a learning rate of 0.1 and went down to 0.0001. Additionally, a range of epochs was tested from 50 to 400. It's also worth noting with the time we used with tuning our starting model, we played around with using the Tensorflow based augmentation method "Image Data Generator," but ultimately moved away from it as we didn't see an improvement of results while using it.

Stage II: Transfer Learning Fine Tuning Process

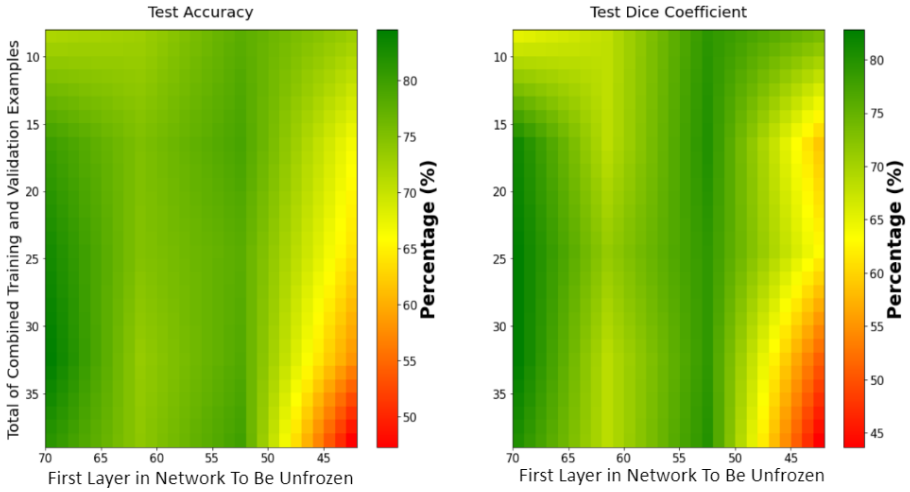


Figure 4: Hyperparameter transfer learning fine tuning comparing the total number of training and validation examples to the first layer in the network to be unfrozen for test accuracy (left) and test dice coefficient (right). There is an obvious change from smaller metric percentages in zones of more unfrozen layers to higher percentages in zones of small amounts of unfrozen layers. This demonstrates the difficulty of fine tuning many layers in the pretrained model.

We conducted hyperparameter fine tuning for the transfer learning. In order of importance for the fine tuning, we looked at the learning rate on an exponential decay scale, the number of epochs (50-400), number of starting model layers frozen, and number of combined training and validation examples used. For metrics, we used an optimizing dice coefficient and satisficing accuracy and loss metrics. The new hyperparameter additions of the amount of frozen layers and number of inputted examples were unique to this transfer learning problem. We put intense focus into tuning these hyperparameters in order to find the best transfer learning results. More specifically, the main goal of transfer learning is to find the best results by maximizing the amount of frozen layers and minimizing the amount of inputted examples. The graphs in Figure 4 depicts a set of accuracy and dice coefficient results with different combinations of these transfer learning based hyperparameters. In a way, it is a summary of one of our clear defined methods of finding the best model. To make the x-axis more clear, we had a total of 70 layers, and each sample on this axis indicates where the first layer in the total 70 layers was to be unfrozen until the end of the network. We found that there were two local optima for this hyperparameter: one where only the final few layers were unfrozen, and another where the majority of the decoding section (approximately 18 layers) were unfrozen. What we see here are clear indications that it is much easier to fine tune the model when you only have to retrain a few of the final layers. This is indicated by the higher percentage accuracy and dice coefficient towards the 70th layer to be the first layer unfrozen and the low percentage towards the opposite end of the scale where more layers in the network are unfrozen.

It is also important to mention that in addition to the hyperparameters for re-tuning the baseline layers, we also had to play around with the number of additional dense and dropout layers added to accommodate the transfer learning process. Towards the end of the end of the research, we found the dropout keep probability parameter was important for preventing overfitting, so we kept an eye on this hyperparameter while tuning the main set of hyperparameters for the transfer learning process.

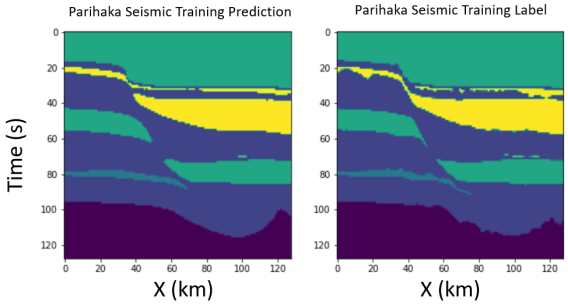


Figure 5: Baseline model predictions using training data (left) compared to the training labels for the same slice (right).

Presentation and Analysis of Results

As will be demonstrated by our results below, it is indeed possible to classify facies with high and relevant levels of accuracy using deep convolutional neural networks to create the pretrained model. The high degree of similarity between the F3 and Parihaka datasets as well as the moderate success of transfer learning demonstrates that transfer learning is a feasible approach for prediction on unlabeled seismic data, although further refinement is needed to achieve industrial relevance.

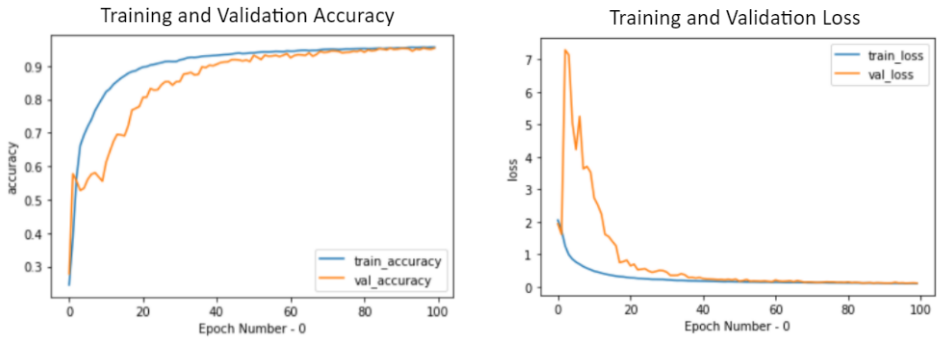


Figure 6: These graphs show the training and validation accuracy (left) and loss (right) for the Parihaka training procedure. We can see some overfitting from epochs 0-60 but after this, validation accuracy and loss begins to match the training accuracy and loss. Ultimately, the highest training accuracy hits at approximately 95%.

Parihaka Pretraining Results and Analysis

Starting with the baseline or pretraining of our Parihaka dataset, we were able to eventually find an approximate 95% validation and training accuracy after only 100 epochs. With the training and validation metrics being both high and similar, the degree of overfitting with the pretraining was minimal. With this training, we used 464 training slices and 116 validation slices as an input. As you can see in the training dataset based predictions in Figure 5 as well as with the test dataset based predictions in Figure 7, we were able to predict to a high degree of accuracy the extent and locations of the different facies. The model used the total 251 test data slices to create the prediction in Figure 7. The fine details such as the texture on some of the boundaries between facies is not predicted in both the test and training predictions as noted by the highlighted examples in Figures 6 and 7. Nevertheless, these predictions are accurate enough as they will only be used by geoscientists as a starting interpretation before they fine tune the interpretation of layers using manual interpretation. With promising results from this baseline model, we then felt confident in transferring this knowledge using transfer learning to the training of the second model which uses the F3 Netherlands dataset.

F3 Netherlands Transfer Learning Results and Analysis

The main drawback to the transfer learning approach we took was that the dataset we pretrained on - the Parihaka cube - was relatively small, and transfer learning works best when the initial dataset is far larger than the fine tuning dataset. This resulted in overfitting when training with the F3 dataset and failure to predict facies in complex geologic environments with the test data. Ultimately, this overfitting issue was partially resolved with a combination of techniques. Additionally, by using only 10-100 images for training and validation combined, the ratio between the number of training and validation images between the pretrained model dataset (580) and the transfer learning dataset worked to our advantage.

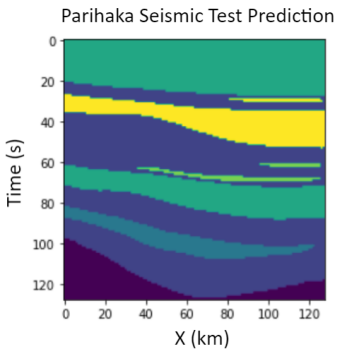


Figure 7: A baseline model Parihaka prediction example based on test data.

With the overfitting came moderate
metrics ranging from the mid 70s to a low 80 percent dice coefficient range for our validation data, but the training percentage range was about 15% higher. This told us that most of the pixel values were matching our labeled images, but some of the details of the more complex geologic layers, for example, the light green and yellow rock layers in the figures featured in Appendix A and B, were not being predicted properly due to the overfitting effect on our model from the training dataset. The shapes and extents of these deeper rock layers show little geometric correlation to the label images as as well. Additionally, we found that the shallow layers were replicating the shallow layer features from the training dataset. Most of the time, it was almost a perfect match. So despite having a 70-80% prediction capability, since most of the geology is very similar across the 3-D seismic cube, high overfit prediction results are still able to look somewhat similar to their labeled image counterparts.

This drawback led us to dedicating many hours of our time to finding non conventional ways to reduce overfitting as the original data augmentation method of using “Image Data Generator” through Tensorflow was not improving the overfit results for some unknown reason. Thus, after many hours of playing with our transfer learning based hyperparameters, we found that setting keep probability values for the additional two dropout layers to be 0.5 and 0.8, shuffling our data when extracting certain amounts of training and validation examples, and setting layer 64 to be the first layer to be unfrozen gave us the most promising results with more limited overfitting. As seen in the right image of Figure 8, which shows transfer learning applied with only a combined 100 training and validation examples, we found that our model was able to predict most of the fine details in all of our rock layers using our test dataset. With this model, we were able to get approximately 82% dice coefficient and accuracy using the test dataset. Not all the prediction slices are shown here, but the model does have trouble

predicting steep vertical boundaries. The model also has trouble focusing the yellow facies position as it is usually centered incorrectly approximately 20 km to the left. Finally, the light and the second to darkest green facies don't extend fully to where they are suppose to be. Despite these drawbacks with the prediction, the overfitting with the additional tuning is much more limited compared to the earlier results in the project, especially in regards to the shallow layers extents.

Finally, with only 10 training and validation examples provided to the transfer learning model as seen with an example prediction from the left image of Figure 8, we can see that our shallow facies relative positions are being predicted to a satisfying degree of accuracy. Compared to the 100 training and validation input prediction, the predictions are not as tight to the correct locations, but they are close enough for a geologist to easily recognize the general boundaries between the shallow rock layers. The extent of the deeper dark green facies boundary is close to predicting the full extent as well, but it is just short of reaching the full extent like we saw with the 100 training and validation input prediction. Unfortunately, we still see that the yellow facies prediction location is off by around 20 km, and the light green facies is being predicted in the correct location, but not to a great enough extent. Overall, with this prediction on the test set, we were able to obtain approximately a 78% dice coefficient and accuracy.

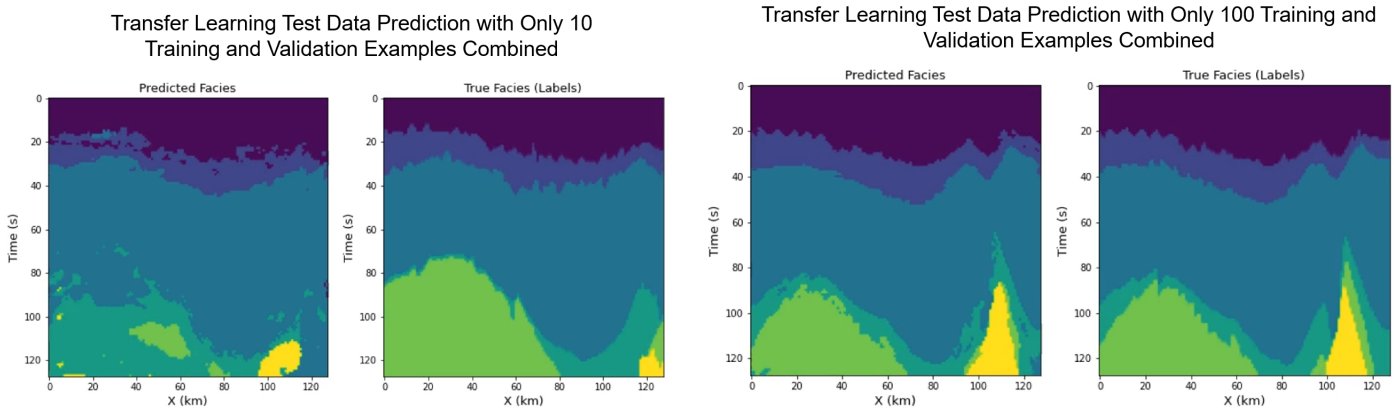


Figure 8: These are transfer learning test data prediction results with only 10 training validation examples combined (left) and with only 100 training and validation examples combined (right). Notice that most of the facies boundaries are predicted correctly for the 100 input example. For the 10 input example, the exact extents of some of the boundaries are not accurate, but the general location of the facies is on target besides the yellow facies. Note, these figures are the figures that we were able to produce that showed the most limited overfitting from the training data compared to earlier predictions like the examples seen in Appendix A and B.

These are promising results as we have reduced the degree of overfitting significantly by seeing unique prediction results for the shallow facies boundaries as well as for most of the deeper facies boundaries aside from the yellow facies which still has a degree of overfitting tied to it. It is important to remember that these prediction results came from only 10 examples, so having this degree of prediction capability and minimum overfitting is very promising.

Recommendations and Future Directions

From our results and analysis of our pretrained and transfer learning models, we can see that the standard semantic segmentation problem solved with our pretrained model has highly accurate prediction capability. The transfer learning model on the other hand, after we were able to minimize the degree of overfitting, provides decent enough prediction results for geologists to recognize the general extent of the facies boundaries, with only 10-100 combined training and test examples. Even with the small error provided in the predictions, its minimal enough for the geologists to correct when they hand tune the interpretation themselves as again, these predictions will basically be used as a starting interpretation model for exploration teams in the energy industry.

Still, there is room for improvement both in the area of overfitting and predicting the extent of some of the layers in the seismic images. First, it would be ideal to train the baseline model on a complex geologic scenario. This would allow for a more robust model that could provide more generalized weights in the transfer learning process. Hopefully, in the near future, energy companies release labeled data in complex geologic locations like the Gulf of Mexico. Furthermore, it would be great to find a workable Deep Lab architecture that could improve upon the U-Net prediction capability. We tried using the Deep Lab architecture early in the project, but were unsuccessful with our results due to the limited time we had, and decided to focus on the U-Net architecture. Finally, we should continue to find methods to prevent overfitting. Specifically, one promising approach to decrease overfitting and increase performance is data augmentation. This is especially true because obtaining seismic data and labeling that data require specialized on-site equipment and human geologic experts, meaning data collection is expensive and only limited public data is available. Now again, we tried to use data augmentation, but with no success using

the Tensorflow based Image Data Generator method.

However, towards the end of our project, we tested a data augmentation approach where we employed custom transformations that preserve important characteristics of the geologic data, such as invariance along the vertical axis and approximate aspect ratio and size of each facies. By composing these transformations and taking slices from the 3-D cube along axes diagonal and orthogonal to the original one, we grew the training dataset to 10x the original size (from roughly 500 training examples to 5,000 examples). As a result, model performance improved from 95.32% and 95.60% accuracy on the validation and training sets, respectively, to 98.05% and 91.31% accuracy, respectively (See Appendix C for more information). The higher accuracy on the validation set and the significantly lower accuracy on the training set suggests that this approach reduces overfitting while increasing performance and could be applied to transfer learning with success.

Even without adding this data augmentation method as well as the other recommendations, we are very excited about our results. We found a succesful method to an 82% degree of test accuracy if you are considering 100 training and validation inputs and 78% accuracy if you are using only 10 training and validation examples. This is now one of only a few documented attempts of this, and the first (to our knowledge) transfer learning application applied between the Parihaka and F3 Netherlands dataset. We hope to push our test accuracy and dice coefficient to 90 percent if possible in the future and publish this as a talk for a future geophysical conference such as IMAGE.

Teammate Contributions

All members of the team generally contributed to many aspects of the project when needed, but some members focused on specific tasks more than others. Joseph Stitt headed the development of the model pretraining and transfer learning code. Joseph provided the background research and domain knowledge for the project. Joseph also focused on the fine tuning of both the baseline and transfer learning based model. Rachael Wang helped develop the transfer learning tuning mechanisms. Rachael also headed the development of the final paper as well. Adam Shugar developed the data augmentation explorations towards the end of the project and helped develop the model pretraining code at the beginning of the project. Adam and Joe helped with the editing of the final paper as well.

References

1 D. Chevitarese, D. Szwarcman, R. G. E. Silva, and E. V. Brazil, "Deep Learning Applied to Seismic Facies Classification: a Methodology for Training," 2018 AAPG Annual Convention Exhibition, Oct. 2018.

2 J. S. Dramsch and M. Lüthje, "Deep-learning seismic facies on state-of-the-art CNN architectures," SEG Technical Program Expanded Abstracts 2018, 2018.

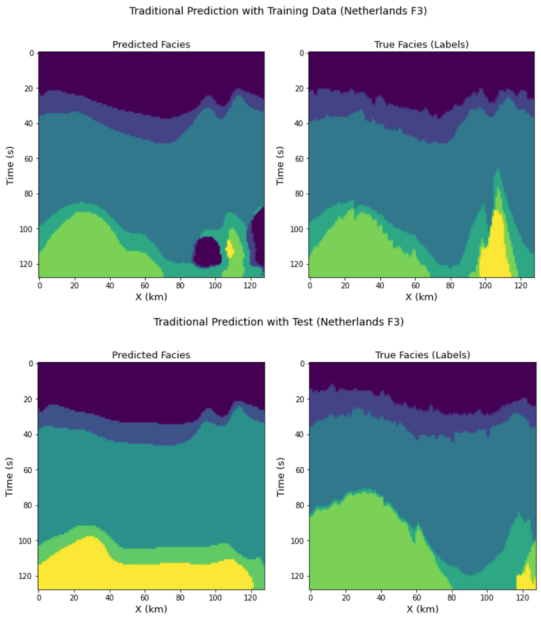
3 O. Ronneberger, P. Fischer and T. Brox (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv:1505.04597.

4 "Seismic Facies Identification Challenge Dataset," AIcrowd. [Online]. Available: https://www.aicrowd.com/challenges/seismic-facies-identification-challenge.

5 Y. Alaudah, "A Machine Learning Benchmark for Facies Classification: F3 Netherlands," GitHub. [Online]. Available: https://github.com/yalaudah/facies_classification_benchmark.

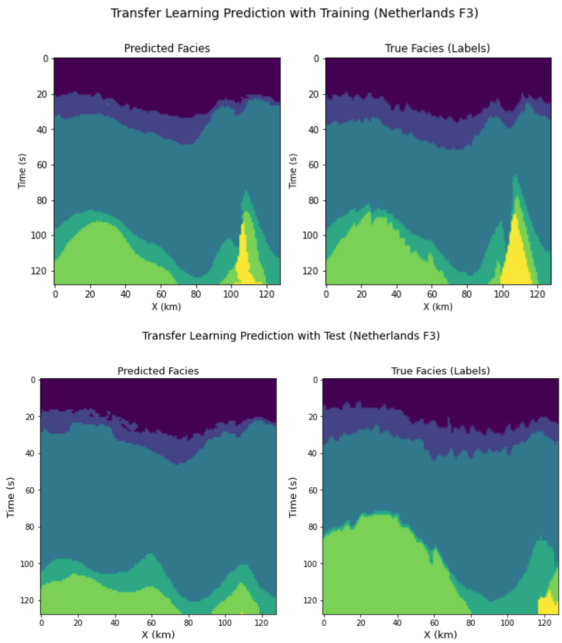
Appendix A

Traditional Learning



Results after 400 epochs

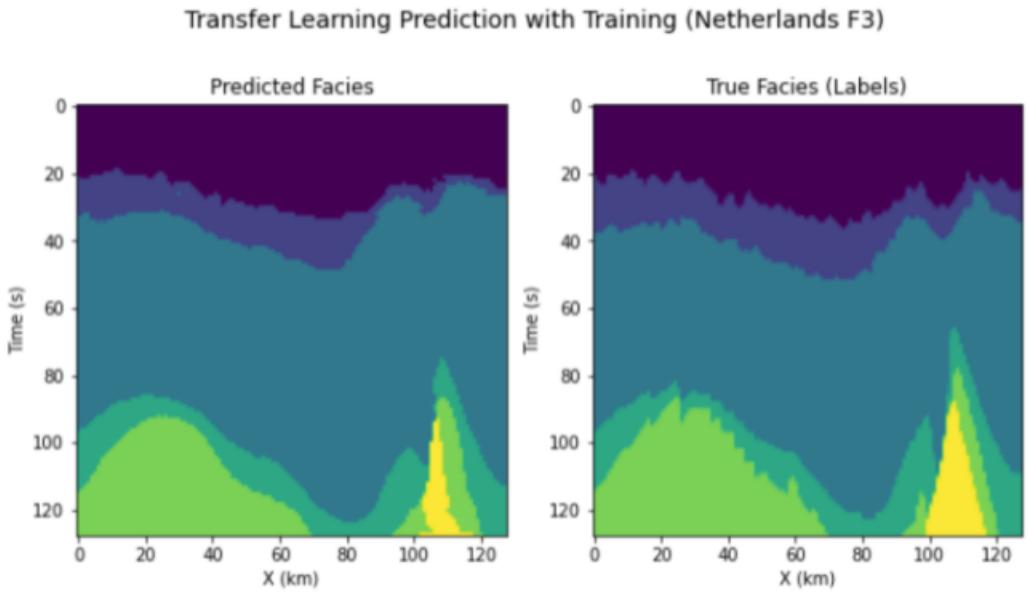
Transfer Learning

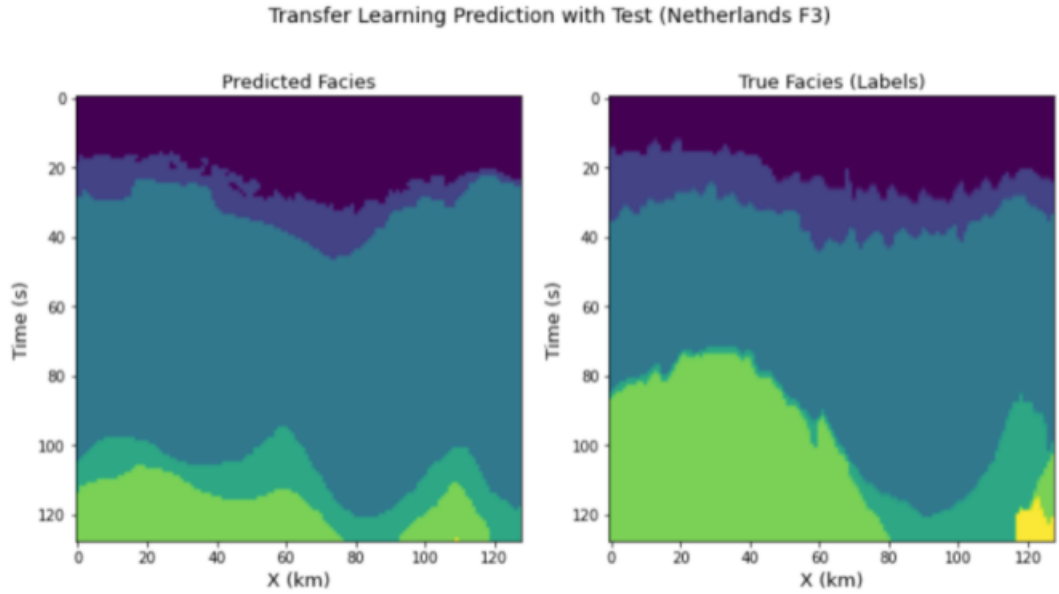


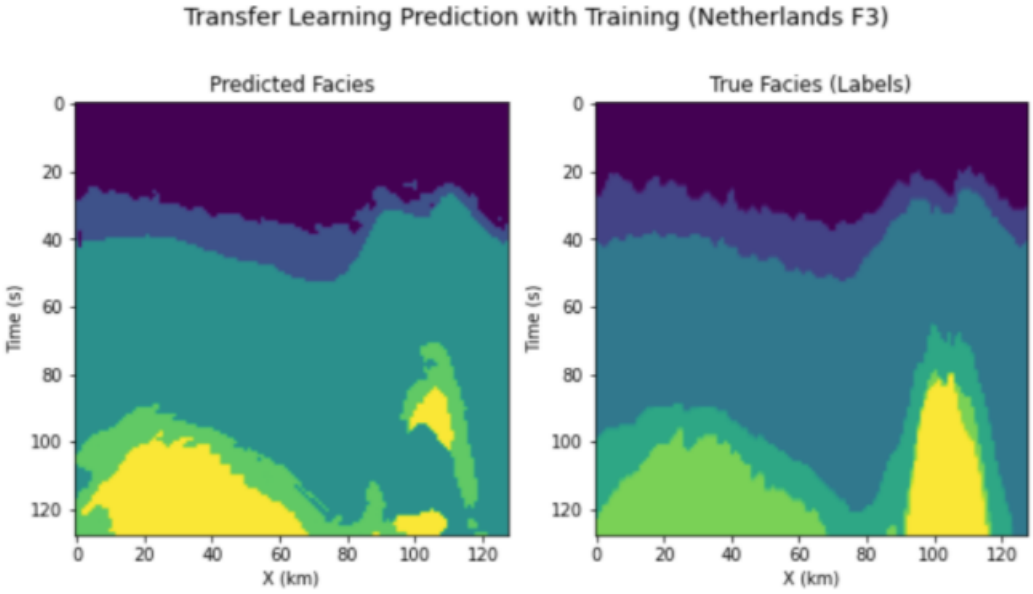
Results after 200 epochs

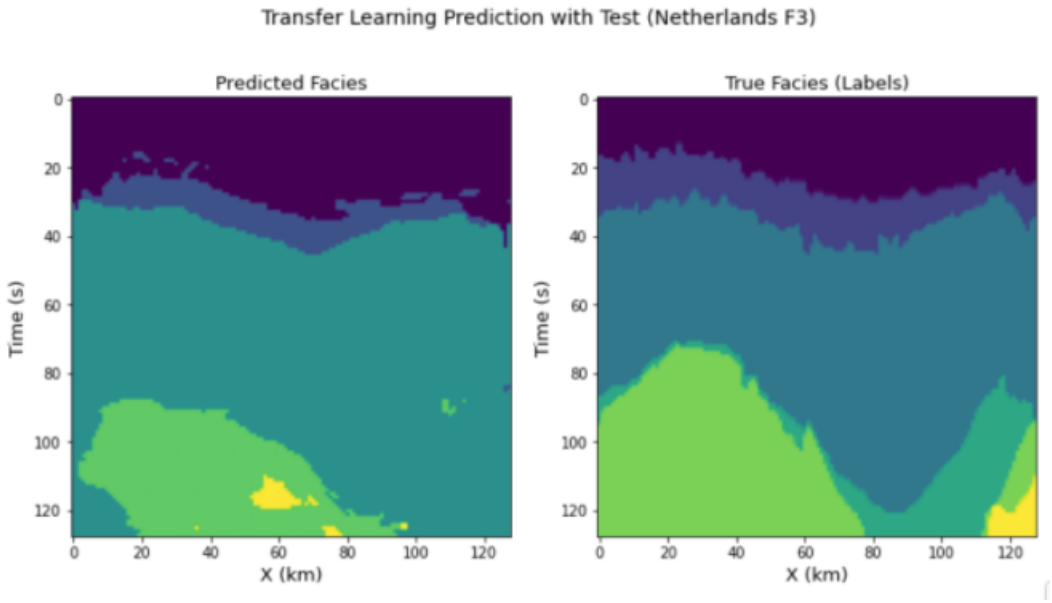
Here we can see the difference between traditional learning on our Netherlands F3 model after 400 epochs and the transfer learning approach after only 200 epochs. Notably, the transfer learning approach shows far superior performance on the training set, but suffers from overfitting as is evident in the similarities between transfer learning predictions at larger times.

Appendix B


Transfer Learning Prediction with Training (Netherlands F3)
Predicted Facies
True Facies (Labels)
Time (s)
X (km)

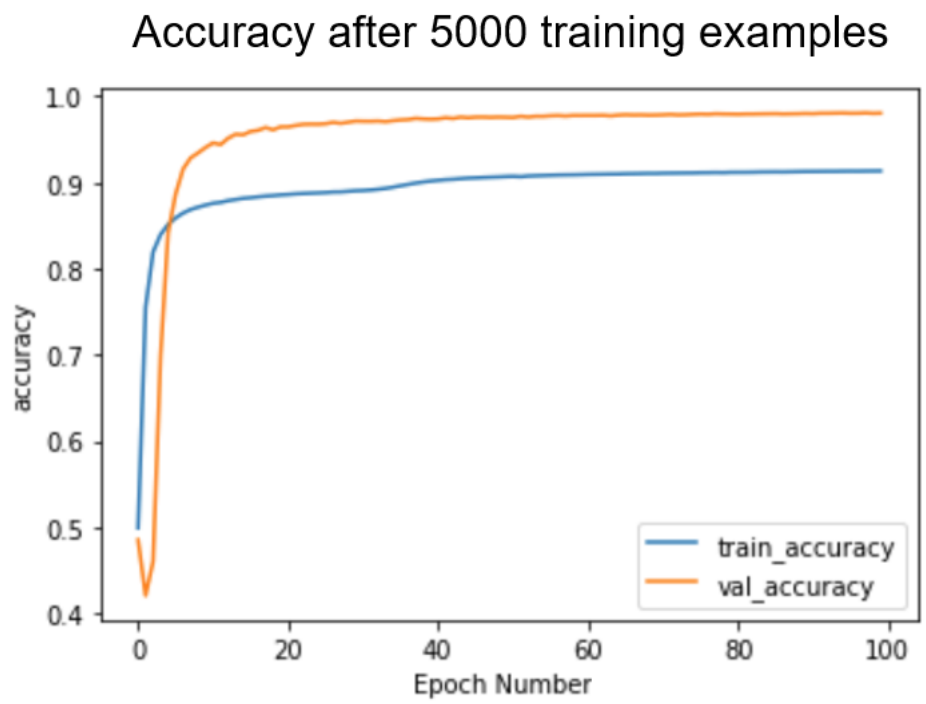

Transfer Learning Prediction with Test (Netherlands F3)
Predicted Facies
True Facies (Labels)
Time (s)
X (km)

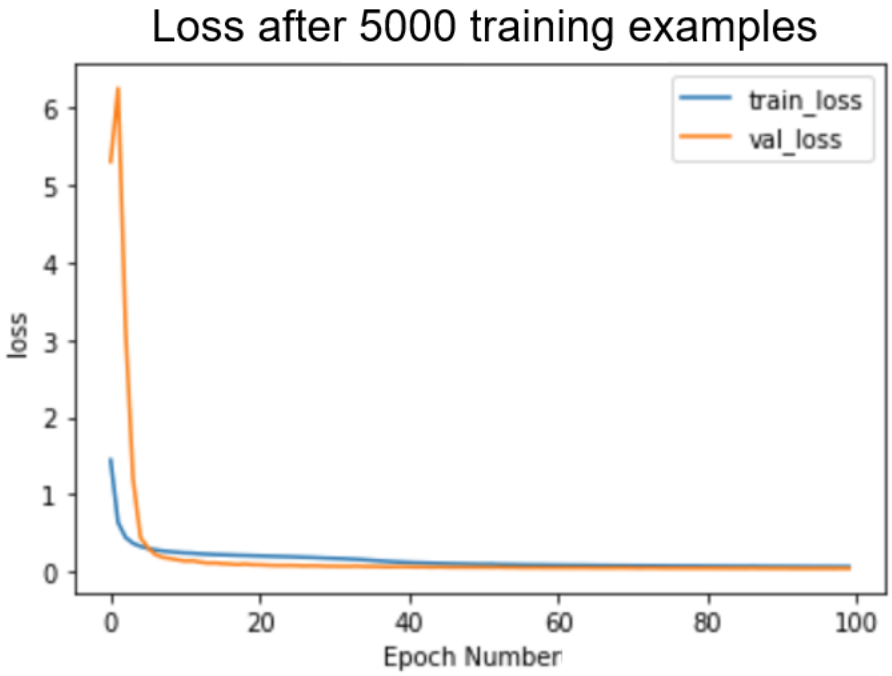

Transfer Learning Prediction with Training (Netherlands F3)
Predicted Facies
True Facies (Labels)
Time (s)
X (km)



Transfer learning predicted facies versus true facies for four images: two from the training set, and two from the test set. We generally see significantly lower performance on the test images, with a degree of overfitting evident in the similarities of the lime green portions of all predictions. These are examples of some of the predictions before we were able to make improvements with overfitting.

Appendix C





These graphs show the accuracy and loss of traditional (non-transfer) learning on the Parihaka data-set using a data augmentation approach, which increased the total training set size from roughly 500 images to roughly 5,000. Notably, the validation accuracy increases more rapidly and ends significantly higher than the training accuracy. The data augmentation has a "smoothing" effect that virtually eliminates the stochastic variations in loss and accuracy observed in earlier epochs in the non-augmented case (see Fig 6).