Interpretation of Deep Neural Networks for Carbonate Thin Section Classification

Lukas Mosser[1] **, George Ghon*[1]*, Earth Science Analytics AS and Gregor Baechle, Benerco LLC*

SUMMARY

This study uses ImageNet pretrained convolutional neural networks (CNNs), VGG11 and ResNet18 models to predict carbonate rock and pore types on a small dataset of 66 thin sections. We subsequently overlay Gradient Weighted Class Activation Maps (Grad-CAM) on top of the original thin section image to highlight features on which the neural network-based its decision-making, introducing an interpretability tool to the method.

Our findings show that pretrained CNNs can successfully learn feature representations in carbonate rock thin sections, achieving training F1 scores of over 90%. However, model generalization is a challenge on the small data set, and the risk of overfitting is investigated by freezing layers during training, achieving test F1 scores of over 65%. Interpretability with respect to rock and pore texture of these Grad-CAM heatmaps depends on the layer depth of the network: (a) High resolution shallower layers in both models show heatmap highlighted areas do correlate with rock textures (pores and grains) whereas (b) low spatial resolution deepest layers in VGG model show correlation between meaningful features in the heat maps provided by Grad-CAM and the actual rock texture in the full resolution image. Finally, neural networks trained on Dunham rock types show better interpretability than the pore type dataset.

INTRODUCTION

Quantitative and qualitative thin section analysis of subsurface and outcrop core samples is paramount for many subsurface reservoir evaluations and characterizations for the petroleum reservoir industry, and the carbon capture & storage industry.

There are numerous studies on quantitative petrographic thin section analysis (Ehrlich et al., 1991; Anselmetti et al., 1998; Baechle et al., 2004). More recently, machine learning (ML), particularly computer vision and deep learning approaches, have been successfully applied in many geoscience domains and shown great promise in thin section classification tasks (Koeshidayatullah et al., 2020; Su et al., 2020; Budennyy et al., 2017; Pires de Lima et al., 2019; Pires de Lima and Duarte, 2021; Patel and Chatterjee, 2016; Peña et al., 2019).

Interpretation of carbonate rock types is time-consuming and requires expert domain knowledge to label the data, which is necessary to train the algorithms used and interpret the results to address any ambiguous classifications. One of the advantages of the ML-driven approach is that it enables geoscience domain experts to apply their expertise to high data volumes in a short time span. Another premise of ML is that once an algorithm is trained using geoscience domain expert labeled data, non-expert level geoscientists can use the models to generate highly accurate and consistent results.

While in many domains, high accuracies can be achieved, it is not self-evident that this is the case for applications with a limited set of training images or an often-times ambiguous ground-truth classification. Moreover, it is not well understood how deep neural networks obtain high precision in classifying images. One of the pitfalls of deep convolutional network architectures (CNNs) is that they have been susceptible to artifacts. For example, networks can train on image irregularities that distinguish features in the pixel domain but do not relate to the depicted object intended to be classified. We have applied the interpretability method Grad-CAM to a carbonate photomicrograph / thin-section data set to visualize CNN layer by layer activation and compare computer vision with the subject matter expert decision making (Selvaraju et al., 2017).

We applied transfer learning that uses previously trained models on ImageNet (Krizhevsky et al., 2012) to shorten our development time to train subsequent models involved in this study. Here, we applied two different pre-trained base models, ResNet18 (He et al., 2016) and VGG11 (Simonyan and Zisserman, 2014), on a suite of 3 different carbonate classification schemes (pore types, modified Dunham rock type, Lucia rock type) using petrographic images. Carbonate lithologies are commonly classified according to their Dunham texture (Dunham, 1962). Lokier and Al Junaibi (2016) show in their study the ambiguities and inconsistencies in assigning texture description and classification utilizing some form of Dunham system. We have observed that, similarly to human carbonate sedimentologists, neural networks learn from experience, differ in their decision making under ambiguity, and prioritize different features.

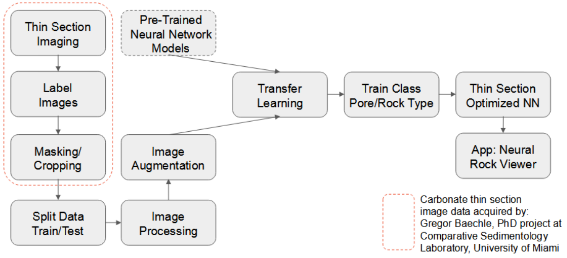

Thin Section Imaging
Pre-Trained Neural Network Models
Label Images
Transfer Learning
Train Class Pore/Rock Type
Thin Section Optimized NN
Masking/ Cropping
Image Augmentation
App: Neural Rock Viewer
Carbonate thin section image data acquired by: Gregor Baechle, PhD project at Comparative Sedimentology Laboratory, University of Miami
Split Data Train/Test
Image Processing

Figure 1: Workflow to analyze the thin sections using transfer learning.

DATASET

The thin section images of this study are from Early to Late Miocene cores acquired during Ocean Drilling Program (ODP) Leg 194, from the Marion Plateau, northeastern Australia. Petrographic plane-polarized light images of 66 samples are the basis of this study. Thin section HSV images stitched together to form a photomosaic covering a complete thin section have been used in this study at resolutions of 6 microns/pixel.

10.1190/image2022-3748082.1

The thin section images of the ODP leg 194 have been qualitatively and quantitatively analyzed in previous studies to relate quantitative pore textural parameters to permeability and sonic velocity (Baechle et al., 2008; Baechle, 2009; Ehrenberg et al., 2006; Weger et al., 2009).

The workflow in this study (Figure 1) of analyzing and interpreting rock and pore types is based on data from doctoral thesis work conducted at the University of Miami (Baechle, 2009): the image acquisition, qualitative labeling of the images using the rock/pore types and the masking/cropping of the area of interest has been previously carried out at the University of Miami. The thin sections are labeled using carbonate rock and pore type classifications of Dunham (1962), the extended Duham terminology (Embry and Klovan, 1971), Choquette and Pray (1970), and Lucia (1995). The rocks are classified into five pore type labels (classes), 6 Dunham rock type labels, and 3 Lucia rock types based on the texture and mineralogical composition.

METHODOLOGY

Convolutional neural networks (CNNs) are a machine learning algorithm well suited for classifying images. The configuration of the layers of a CNN is typically split into a feature extractor and a classifier. The feature extractor produces a lower-dimensional representation of an input feature, while the classifier transforms these representations into a predicted class probability.

Transfer learning is an approach that leverages prior experience imbued in neural networks through training on large datasets to obtain a model that can be trained on less data on a more specialized task.

We have evaluated classification performance and interpretability for two modes of transfer learning. First, we consider using the pretrained weights of a deep neural network and training a classification layer. This reuses the previously learned feature representations for the new classification task. During training, the weights of the convolutional layers are not updated through backpropagation and stochastic gradient descent and are therefore ”frozen”.

The second approach uses the pretrained weights of the CNN as a starting point to train the entire network. This gives the network more parameters to adapt during the learning process. Modern CNNs will likely overfit on small datasets, making transfer learning with a frozen feature extractor often the preferred choice. Data augmentation can be applied to the input data to regularize the model to obtain a model that can generalize to unseen data.

In our investigation, we have used a visual interpretability method called Gradient-Weighted Class Activation Maps (Grad-CAM). The underlying principle of gradient-based visual interpretability methods is that gradients of the predicted class label with respect to intermediate feature maps indicate the importance of a specific region of an input image with regard to one particular target class. Grad-CAM is applied by first selecting a specific image class that the network has been trained to predict. The approach produces a class-activation map at a particular depth of the CNN, i.e., for a specific layer. The authors of Grad-CAM have shown that this allows them to create interpretable maps of intermediate feature representations (Selvaraju et al., 2017).

This study has selected two different types of CNNs to evaluate how CNNs process thin-section images. Both VGG11 and ResNet18 are widely popular image classification networks that, in our case, have been pretrained on the ImageNet benchmark dataset.

Due to the small dataset size, we apply heavy data augmentation, including random cropping and jitter of the color hue, saturation, and value. Therefore, this training process creates four models per label set considered in this study. The models are trained using the ADAM optimizer (Kingma and Ba, 2014) using the same mini-batch size and learning rates.

A validation dataset was generated at training time by applying the same data augmentation on the test dataset images. The final test dataset was evaluated by predicting the full test dataset images at their full image size without any data augmentation.

In practice, we find that the CNNs where all parameters are available for training quickly overfit on the training dataset, while models trained with a frozen feature extractor have a slower and well-behaved convergence.

Once trained, we generate Grad-CAM activation maps produced for any intermediate layer of the networks as an overlay to the actual full-resolution thin-section images. However, Grad-CAM maps created for layers closer to the classifier layers will have a low spatial resolution due to the spatial pooling layers in the CNN models (Figure 2). We, therefore, upsample all intermediate Grad-CAM maps to the full thin-section image sizes using bilinear upsampling. The produced overlay heatmaps have subsequently been used to compare the produced interpretability maps for different transfer learning scenarios (frozen/unfrozen) and for each pre-trained network architecture (VGG11 / ResNet18).

RESULTS

To evaluate the rock and pore type classification results, we have compared F1 train and test scores, as well as confusion matrices of labeled vs. predicted rock/pore type classes for 12 models in total.

The F1 scores in Figure 3 indicate how accurate the data labels are predicted. The F1 score differences between train and test sets are consistently higher in unfrozen models (>0.2-0.5) than frozen (>0.14-0.34) models, indicating the potential to reduce overfitting by freezing selected convolutional layers. ResNet models with unfrozen layers show the highest training accuracy (>0.9) for all label types. The average training data scores range from 0.72-0.93; the average test data set F1 score shows a lower range from 0.36 to 0.69 (Figure 3). Consequently, the models learned weights during training that serve as effective feature extractors for the classification of each dataset.

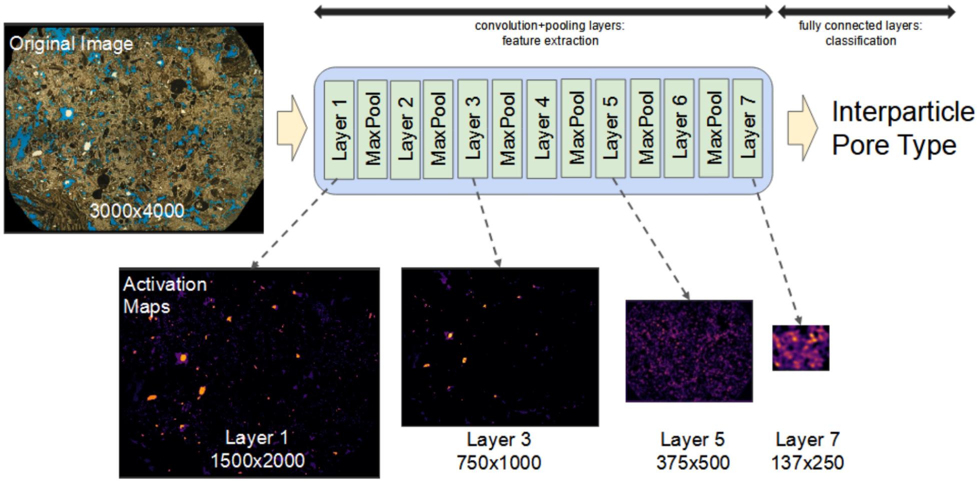
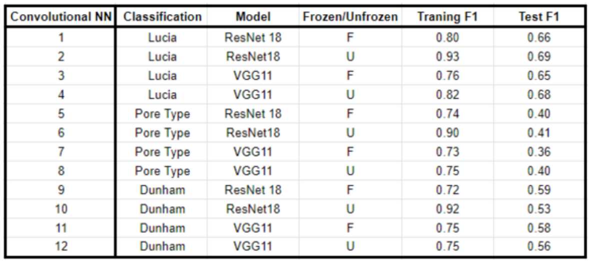


Figure 2: ResNet transformation of image and image resolution along with the convolutional layers from shallow to deep. The original image shows a carbonate thin section in full resolution. The representations of layers 1-7 visualize the Grad-CAM activation maps on the downsized images in the convolutional pipeline.

Figure 4 shows the confusion matrix of the ResNet18 model predicting pore type labels using unfrozen layers. Figure 4A displays the training dataset, whereas Figure 4B displays the test dataset. The human-interpreted ”ground truth” classes are on the y-axis compared to the predicted classes on the x-axis. ResNet18 model using unfrozen layers predicted all classes of the training set correctly, as evidenced by the diagonal match of all ground truth and predicted classes—the model over-predicts vuggy pore type class in the test dataset (Figure 4B). We observed a better prediction of the training pore type and modified Dunham rock type in the unfrozen layers compared to using the same model with frozen layers. Recrystallized dolomites dominate the training and test dataset classified by modified Dunham. Some recrystallized dolomites and grainstones are falsely classified as bound stones in the test set. Both VGG and ResNet architectures overpredict boundstone textures in the modified Dunham classification. This apparent network bias relates to over-efficient feature extraction for these particular rock structures.



Convolutional NN	Classification	Model	Frozen/Unfrozen	Traning F1	Test F1
1	Lucia	ResNet 18	F	0.80	0.66
2	Lucia	ResNet18	U	0.93	0.69
3	Lucia	VGG11	F	0.76	0.65
4	Lucia	VGG11	U	0.82	0.68
5	Pore Type	ResNet 18	F	0.74	0.40
6	Pore Type	ResNet18	U	0.90	0.41
7	Pore Type	VGG11	F	0.73	0.36
8	Pore Type	VGG11	U	0.75	0.40
9	Dunham	ResNet 18	F	0.72	0.59
10	Dunham	ResNet18	U	0.92	0.53
11	Dunham	VGG11	F	0.75	0.58
12	Dunham	VGG11	U	0.75	0.56

Figure 3: Table showing F1 test and training scores of the 12 CNN models to predict Lucia, Dunham, and pore type classification.

After model training, we have created Grad-CAM visualizations for each layer in the trained models. For example, Figure 5 shows four training dataset thin sections in plane-polarized light next to the Grad-CAM activation maps of the first (most shallow) layer of the ResNet18 CNN for pore type classification. First, we observe that the activated area maps well to the rock’s grain and pore space texture. Furthermore, as shown in Figure 5, in an interparticle pore type dominated sample (upper left corner), the activated areas are correlated with the pore space. In contrast, in the sample with intercrystalline pore type (upper right), the moldic pore type (lower left), and the vuggy sample (lower right), the grain features are activated.

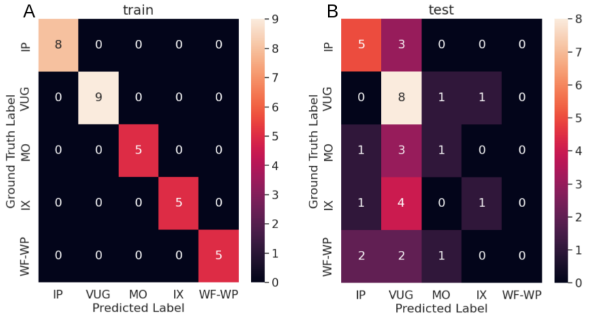


Figure 4: Pore type classification - absolute number confusion matrix. ResNet18 unfrozen layers model with (A) training data and (B) testing data.

How do the activation maps change depending on the classes displayed in the shallow layers? The activation maps can be displayed for each predicted class per thin section image. We observe changes in the activated features as a function of the pore classes. Comparing the activation map of the human interpreted ground-truth pore class with the activation map of the highest probability pore class, we observe minor changes in the intensity up to significant differences from switching grain feature activation to pore feature activation as a function of the

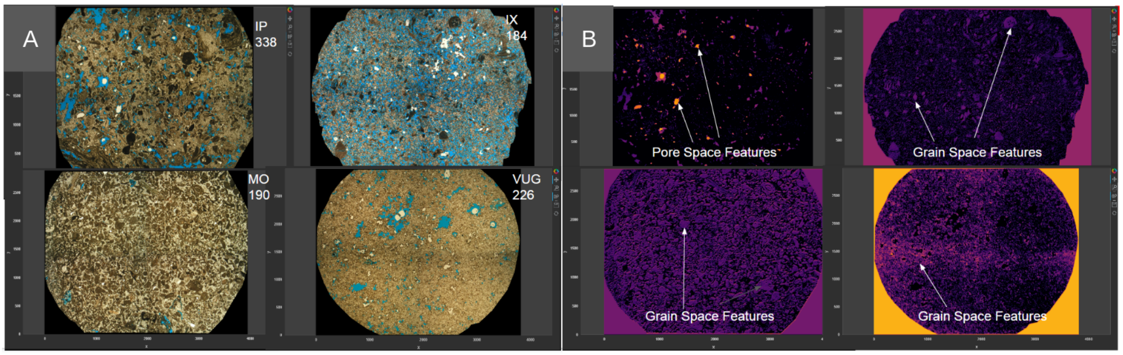


Figure 5: Petrographic thin section images of 4 test data samples. (A) Plain light images and (B) Activation map of layer zero using ResNet18 pore type classification.

pore class displayed in one sample.

When comparing activation maps of deep layers in VGG and ResNet models, we observe different activation patterns in activation maps between these model types. This indicates that the models base their decision of class probability on different features in the thin section. Furthermore, in the ResNet model, neither the activation maps of ground truth pore type (human classified) nor the maps of the highest probability pore type appear to be well interpretable with respect to the pore space features. In contrast, in the VGG models, the moldic and vuggy pore type activation maps are interpretable with respect to the pore space features recognized by a human interpreter.

Comparing deep CNN layers in VGG and ResNet models for a training sample with a ground truth moldic pore type, we observe that VGG predicts the pore type wrong. It shows the highest class probability as vuggy pore type, but the activated pore features match the pore space in question. In contrast, using the same image, ResNet correctly predicts the pore type as moldic. However, the ResNet activation map shows activated grain space, which results in low interpretability.

Some samples show zonal dolomitization with depositional features still partially intact. This is a challenging case for classification tasks, as the sample is ambiguous, and its texture varies zonally. Nevertheless, visualizing the Grad-CAM activation for dolomite as a heat map in partially dolomitized samples correctly identifies the localized diagenetic overprint. Furthermore, it confirms that deep CNNs can learn textural rock features successfully.

Benchmarking VGG11 against ResNet18 performance on partially dolomitized samples reveals a network-specific bias. Dolomitized grainstones and boundstones are classified as dolomites by the trained ResNet and as grainstones or boundstones by the VGG architecture.

CONCLUSIONS

With the growing applications of neural networks, there is a corresponding need to explain their decisions. Therefore, the primary aim of this study was to train several CNN models and to evaluate a method (Grad-CAM) that has been postulated to allow inspection of the decision-making of a CNN visually by heatmaps overlain on the original image.

We show that using transfer learning of two ImageNet pretrained CNNs (VGG11, ResNet18) enabled us to predict carbonate pore and rock type reasonably well: The F1 score for Lucia rock type prediction is higher than for the prediction of the modified Dunham rock types, which in turn is higher than the F1 score for the prediction of pore type classes.

Interpretability with respect to rock and pore texture of these heatmaps depends on the selected layer of the network: high resolution shallower layers in both models show heatmap highlighted areas correlate with rock textures (pores and grains). Low spatial resolution deepest layers in the VGG model show correlation between meaningful features in the heat maps provided by Grad-CAM and the actual rock texture in the full resolution image.

Interpretability with respect to the label shows that the Dunham dataset appears to show better interpretability than the pore type dataset. Qualitatively, we observe that VGG models appear to show better interpretability than ResNets for pore type labels.

This study successfully provided data and a tool to show how visualization might help answer what the network detects. However, further improvements will require training the CNNs with more labeled thin section training data.

ACKNOWLEDGMENTS

Prof. Gregor Eberli, Comparative Sedimentology Laboratory (CSL) at the University of Miami, is thanked for providing the high-resolution thin section images.

10.1190/image2022-3748082.1

REFERENCES

Anselmetti, F. S., S. Luthi, and G. P. Eberli, 1998, Quantitative characterization of carbonate pore systems by digital image analysis: AAPG Bulletin, **82**, 1815–1836.

Baechle, G. T., 2009, Effects of pore structure on velocity and permeability in carbonate rocks: Ph.D. thesis, University of Tuebingen.

Baechle, G. T., A. Colpaert, G. P. Eberli, and R. J. Weger, 2008, Effects of microporosity on sonic velocity in carbonate rocks: The Leading Edge, **27**, 1012–1018, doi: https://doi.org/10.1190/1.2967554.

Baechle, G. T., R. Weger, G. P. Eberli, and J.-L. Massaferro, 2004, The role of macroporosity and microporosity in constraining uncertainties and in relating velocity to permeability in carbonate rocks: 74th Annual International Meeting, SEG, Expanded Abstracts, 1662–1665, doi: https://doi.org/10.1190/1.1845149.

Budennyy, S., A. Pachezhertsev, A. Bukharev, A. Erofeev, D. Mitrushkin, and B. Belozerov, 2017, Image processing and machine learning approaches for petrographic thin section analysis: Presented at the Russian Petroleum Technology Conference, SPE, OnePetro.

Choquette, P. W., and L. C. Pray, 1970, Geologic nomenclature and classification of porosity in sedimentary carbonates: AAPG Bulletin, **54**, 207–250, doi: https://doi.org/10.1306/5D25C98B-16C1-11D7-8645000102C1865D.

Dunham, R. J., 1962, Classification of carbonate rocks according to depositional texture, *in* W. E. Ham, ed., Classification of carbonate rocks: AAPG Memoir 1, 108–121.

Ehrenberg, S. N., G. P. Eberli, and G. Baechle, 2006, Porosity-permeability relationships in Miocene carbonate platforms and slopes seaward of the Great Barrier Reef, Australia (ODP Leg 194, Marion Plateau): Sedimentology, **53**, 1289–1318.

Ehrlich, R., S. J. Crabtree, K. O. Horkowitz, and J. P. Horkowitz, 1991, Petrography and reservoir physics I: Objective classification of reservoir porosity: AAPG Bulletin, **75**, 1547–1562.

Embry, A. F., and J. E. Klovan, 1971, A Late Devonian reef tract on northeastern Banks Island, NWT: Bulletin of Canadian Petroleum Geology, **19**, 730–781.

He, K., X. Zhang, S. Ren, and J. Sun, 2016, Deep residual learning for image recognition: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, doi: https://doi.org/10.1109/CVPR.2016.90.

Kingma, D. P., and J. Ba, 2014, Adam: A method for stochastic optimization: arXiv preprint, arXiv:1412.6980.

Koeshidayatullah, A., M. Morsilli, D. J. Lehrmann, K. Al-Ramadan, and J. L. Payne, 2020, Fully automated carbonate petrography using deep convolutional neural networks: Marine and Petroleum Geology, **122**, 104687, doi: https://doi.org/10.1016/j.marpetgeo.2020.104687.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2012, ImageNet classification with deep convolutional neural networks, Advances in neural information processing systems 25.

Lokier, S. W., and M. Al Junaibi, 2016, The petrographic description of carbonate facies: Are we all speaking the same language?: Sedimentology, **63**, 1843–1885, doi: https://doi.org/10.1111/sed.12293.

Lucia, F. J., 1995, Rock-fabric/petrophysical classification of carbonate pore space for reservoir characterization: AAPG Bulletin, **79**, 1275–1300, doi: https://doi.org/10.1306/7834D4A4-1721-11D7-8645000102C1865D.

Patel, A. K., and S. Chatterjee, 2016, Computer vision-based limestone rock-type classification using probabilistic neural network: Geoscience Frontiers, **7**, 53–60, doi: https://doi.org/10.1016/j.gsf.2014.10.005.

Pena, A., M. Caja, J. R. Campos, C. Santos, J. L. Perez, P. R. Fernandez, and J. Tritlla, 2019, Application of machine learning models in thin sections image of drill cuttings: Lithology classification and quantification (Algeria tight reservoirs): EAGE/ALNAFT Geoscience Workshop, EAGE, 1–5.

Pires de Lima, R., and D. Duarte, 2021, Pretraining convolutional neural networks for mudstone petrographic thin-section image classification: Geosciences, **11**, 336, doi: https://doi.org/10.3390/geosciences11080336.

Pires de Lima, R., F. Suriamin, K. J. Marfurt, and M. J. Pranter, 2019, Convolutional neural networks as aid in core lithofacies classification: Interpretation, **7**, no. 3, SF27–SF40, doi: https://doi.org/10.1190/INT-2018-0245.1.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017, Grad-CAM: Visual explanations from deep networks via gradient-based localization: Proceedings of the IEEE International Conference on Computer Vision, 618–626, doi: https://doi.org/10.1109/ICCV.2017.74.

Simonyan, K., and A. Zisserman, 2014, Very deep convolutional networks for large-scale image recognition: arXiv preprint, arXiv:1409.1556.

Su, C., S.-J. Xu, K.-Y. Zhu, and X.-C. Zhang, 2020, Rock classification in petrographic thin section images based on concatenated convolutional neural networks: Earth Science Informatics, **13**, 1477–1484, doi: https://doi.org/10.48550/arXiv.2003.10437.

Weger, R. J., G. P. Eberli, G. T. Baechle, J. L. Massaferro, and Y.-F. Sun, 2009, Quantification of pore structure and its effect on sonic velocity and permeability in carbonates: AAPG Bulletin, **93**, 1297–1317, doi: https://doi.org/10.1306/05270909001.