DL4CV: OCT Interpretability using Deep Learning

Guy Lutsker, Hagai Ravid (both under the Computer Science and System Biology faculty)

Abstract

Deep learning (DL) is a field of artificial intelligence that has grown significantly in recent years. The scientific community has concentrated on DL because of its adaptability, high performance, great generalization capability, and diverse applications, among other characteristics. A significant volume of medical data, as well as the development of increasingly powerful computers, has fueled interest in the use of DL for medical images, in particular OCT scans. Recent work has shown that DL models are capable of fitting OCT scans using convolutional neural networks to achieve high classification accuracy. In this paper we will try to harness the proven capabilities of these models to try and interpret how these models "see" (using gradient visualization methods such as Grad-CAM) the data, and try to decipher how they diagnose medical images. We will also investigate how convolution based methods compare to the latest architectures such as transformers, which have shown big promise in the way they fuse semantic-level and spatial-level information together. Uncovering the attention maps of these types of models might lead to better understanding as to how these models operate. If indeed there exists medically informative results in the interpretation methods, such as Grad-CAM / attention Maps, this can lead to models that could help medical professionals reach a more accurate diagnosis since we could provide doctors with the critical parts of the medical scan to reach a diagnosis. Our aim is threefold: I. Compare how different state-of-the-art deep learning architectures for vision are capable of fitting medical OCT scans. II. Interpret how these models operate and how they make their decisions. III. Investigate how we can harness self-supervised models to get a better understanding of the capabilities in medical imaging in deep learning.

1 Background

Optical Coherence Tomography (OCT) is a noninvasive imaging test used to obtain high resolution cross-sectional images of the retina. OCT is analogous to ultrasound imaging, except that it uses rays of light to measure retinal thickness, instead of sound. The layers within the retina can be differentiated and retinal thickness can be measured to aid in the early detection and diagnosis of retinal diseases and conditions. An example of the sections (with medical nomenclature) of an OCT scan can be seen in figure 1.

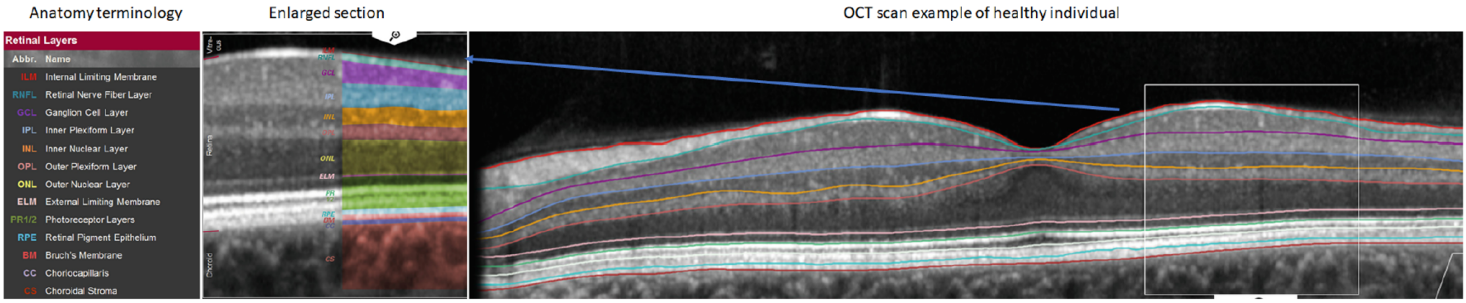

Figure 1: Example of OCT scan with retinal layers labels.

2 Motivation

OCT scan evaluation is a well known challenge in the medical community, it requires experts to laboriously go though dozens of scans and try to detect abnormalities, which can be both error prone, as well as expensive. In this work we would like to be able to optimize this diagnosis procedure by focusing medical professionals attention to the more important parts of the scan. During the last couple of years there emerged several high performing architectures for image classification, and our hope is that we could harness their power to investigate the domain of retinal diseases by using interpretation tools (such as Grad-CAM / attention maps). In this project we will use the Kermany dataset which is a large dataset consisting of tens of thousands of OCT scans. We hope that the variety of the dataset will yield an accurate representation of retinal diseases, and that our final model would be able to generalize to scans it has never seen before at the hands of medical professionals. If our method will prove successful, we would hope that we could use this tool to investigate other important medical quires such as identifying abnormalities in volumetric OCT scan, or even out of domain scans, such as tumor detection in CT scans of the lung.

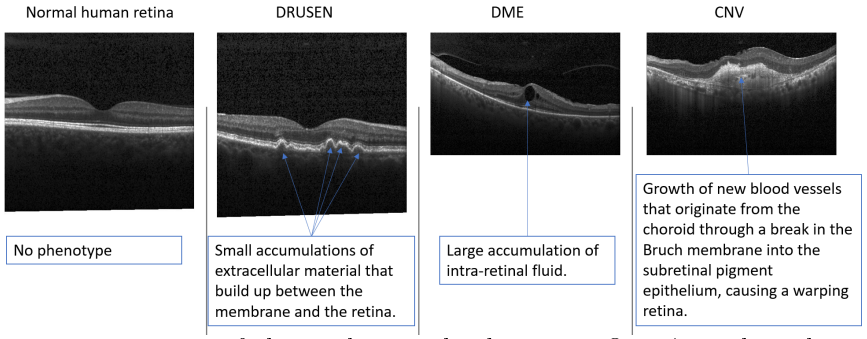

Figure 2: Example of the 4 labels in the dataset. NORMAL is how the human retina normally looks on OCT scans. CNV is characterized by neovascular membrane. DME is characterized by retinal-thickening-associated intra-retinal fluid (opaque bubbles). DRUSEN is characterized by a normal looking retina with preserved foveal contour and absence of any retinal fluid.

3 Data

In this project we will use the Kermany dataset. This dataset is hosted on Kaggle, and consists of OCT scans of different patients. There are 4 classes in this dataset: 1. Normal. 2. Choroidal neovascularization (CNV) 3. Diabetic macular edema (DME) 4. DRUSEN. Visualization of the pathologies as well as medical phenotype explanation appears in figure 2. The dataset is divided into 3 folders: Train set - consisting of 83,484 OCT scans, Validation set - consisting of OCT 32 scans, Test set - consisting of OCT 968 scans. The labels were obtained by medical professionals at several institutions world wide: Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center. The data was collected between July 2013 and March 2017.

4 Results

4.1 Fitting The Data

Our first order of business was assessing which architectures would be able to fit our dataset. We decided to look at a selected few of the state of the art architectures vision models from the last 2 years, these included - Residual Networks (18, 50, 101, 152), EfficientNet-V2 [2021][4] (Small, Medium, Large, XLarge), ConvMixer [2021][5] (Small, Medium Large), ConvNext [2022][6] (Base, Large, XLarge), Vision Transformer-ViT [2020][7] (Base, Large, on both patch 16/32 and both on image size 224/384), Data-efficient Image Transformers - DeiT [2021][8] (same configuration as ViT), Shifted window Transformer - Swin-Transformer [2021][9] (Base, Large). We have trained each of the mentioned architectures in a randomized hyper parameter search in the following range: batch size (2 - 8), learning rate (0 - 0.001), momentum (0 - 0.9), optimizer (Adam, SGD, RMSProp), weight decay (0 - 0.1), load with pretrained weights on ImageNet (True, False). After training a couple dozens architectures, we tuned our hyper parameter range, and rerun the search under Bayesian search method[10] for 5 epochs ($\approx$ 2 hours training time). The hyper parameters were chosen to optimize the validation accuracy, and later was assessed on the test set without further tuning. In addition, for each model we also logged the following metrics: 1. Number of trainable parameters. 2. Time per epoch. 3. Loss. 4. Validation Accuracy (as well as per class). 5. Test Accuracy (as well as per class). Of the mentioned architectures, we have only managed to fit (>0.9% accuracy on validation set) the architectures appearing in figure 3. As we can see the only architectures we were able to fit are ResNets, ViT, and ConvNext. We are quite confident that given enough time and resources we could have fitted the other architectures to a high accuracy score, yet in the interest of time we decided to move forward with the best models from these architectures - ResNet50, ViT Base patch size of 16, pertained on image size 224×224, ConvNext Base .

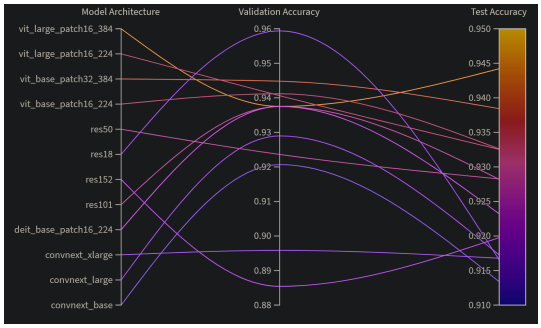

Figure 3: Highest test accuracy model per each architecture. Each line is a trained model - first column is the model name, second is evaluation accuracy, and last is test accuracy. Showing only models with test accuracy>90%.

4.1.1 Visualizing Embeddings

The Next step is trying to interpret these trained models, but first we decided to look at the representations the models have learned. Here we took the embeddings of each model (in Resnets & ConvNext, which are convolution based, we took the last embedding before the fully connected layer, and in attention based models we took the cls token) and visualized them using UMAP[11] dimensionality reduction to 2 dimensions, results are shown in figure 4. As we can see both ViT and ConvNext were able to create meaningful low dimensional representations, which translate well to a 2D visualization. This tells us that the representations the models create might contain medically significant data. Resnet, on the other hand, creates less meaningful embeddings, which leads us to believe that its representations are less meaningful and thus its classification hyperplane is most likely considerably more complex than the other models. At this stage we might hypothesize that ViT & ConvNext should produce more insightful results, medically speaking, heading forward.

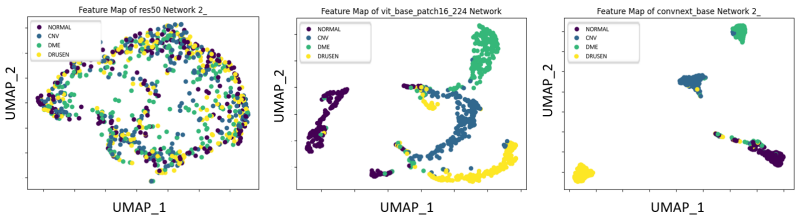

Figure 4: UMAP of embeddings in Resnet, ViT, ConvNext.

4.2 Interpretation Of Models

4.2.1 Grad-CAM

First, we wanted to analyze the models gradients in order to interpret how the model is making its classification decisions. One such method of analyzing a models gradients is Grad-CAM, which is able to provide pixel level annotations of the gradient with respect to the input image. To make sure our models relied on the diseases phenotypes in their classification, and don't apply "shortcut learning"[14] on the dataset (relying on artifacts like the images background), we analyzed the different models by comparing their Grad-CAM results on the same input image, results can be seen in figure 5. As we can see, all models focused their gradients on well-known phenotypes of each diseases when classifying between them. We see that all models captured the CNV pathology, whereas ViT tends to be more noisy. In addition, we see that sometimes the models share their results (as in the top row of figure 5), and point to mostly

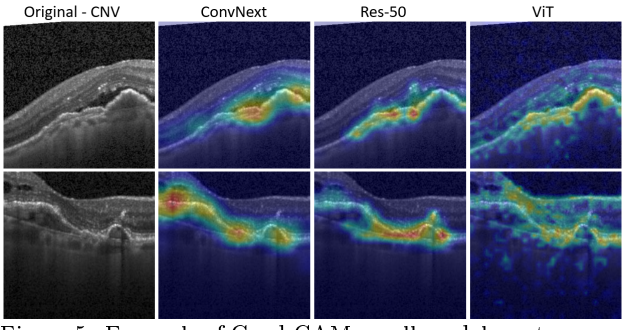

Figure 5: Example of Grad-CAM on all models on two scans containing CNV pathology.

the same areas, whilst sometimes they differ in their results (bottom row). This phenomena gives rise to the idea of combining the gradient maps of different methods in an ensemble-like manner to get a more powerful interpretation - we did not follow up this idea as of now. In general, it looked as though usually ConvNext has provided with the best gradient maps, Resnet was slightly less accurate and sometimes missed the primary phenotype presented, and ViT tended to be mostly noisy.

4.2.2 Occlusion Tests

Another possible analysis is conducting an Occlusion test. An Occlusion test is a method in which we occlude a certain field of view of the input image from the model, and see how its classification differs. The result of running multiple occlusions like these can provide an occlusion map which shows which parts of the image have a positive contribution to the score of a certain class. Here we chose to focus on ConvNext Occlusion maps for the sake of convenience (as to be less confusing for the reader) and results appear in figure 6. Unsurprisingly, the Occlusion map resembles the Grad-CAM map - both are pointing close to the pathology in the image (the large accumulation of the DME), while the Occlusion map is a bit more noise prone. In addition, the fact that Grad-CAM and the Occlusion test mostly converge on their results gives us more confidence about the way the model classifies the pathologies.

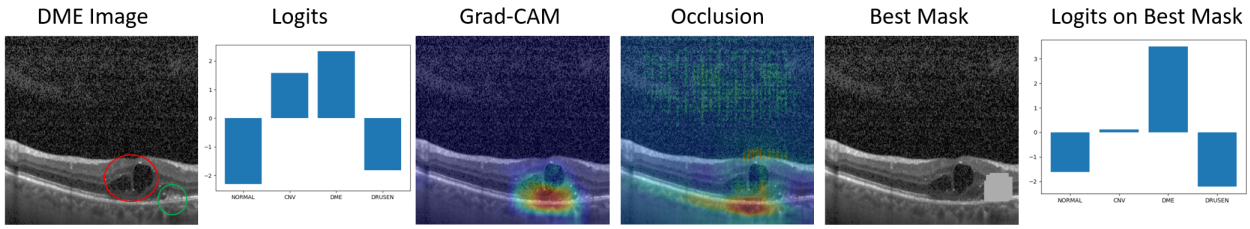

Figure 6: Example of Grad-CAM versus Occlusion test on DME (ConvNext). The original image presents a clear DME phenotype (red circle), yet we can also observe a possible CNV phenotype (green circle). The logits of the model represent this characterization well, as both DME & CNV have large logits values. In the Occlusion map blue hues represent parts of the image that, if hidden, pushed the logits to resemble the true distribution of the image class, while red hues represent parts of the image that, if hidden, pushed the class distribution into a wrong distribution. It then makes sense that the best occlusion map tries to hide the CNV phenotype, as to push the class distribution to the DME. This is well represented in the logits on the best mask, where once the CNV phenotype has been hidden, the CNV probability drops, while the other mostly stay constant.

4.2.3 Attention Maps

In addition to gradient analysis methods such as Grad-CAM, in attention based models, we can also investigate the attention maps that are generated during inference. The idea is that while the attention maps are not as closely related to the classification decision as gradient methods, they are significant during the information routing in the transformer architecture. That is why they might hold informative data to interpret how the model decided what is important in the input image. In this analysis, we followed the attention maps obtained by the method in "Transformer Interpretability Beyond Attention Visualization"[19] .We hypothesized that the attention maps would be very similar to the Grad-CAM results since they both tell us how the model "views" the image, but we were surprised to see how different the results are - Results in figure 7. We can see that although ViT produces quite bad Grad-CAM results, the attention map provides a precise annotation of the DRUSEN phenotype - so much so in fact, that they are better then any other gradient analysis method.

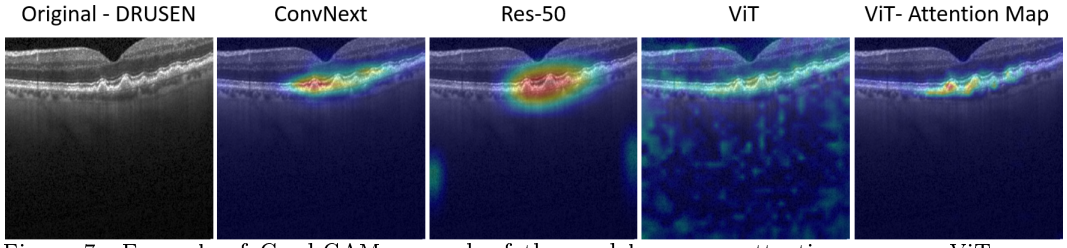

Figure 7: Example of Grad-CAM on each of the models, versus attention map on ViT on DRUSEN image. As we can see, both ConvNext and Resnet highlighted the pathology of DRUSEN in a very coarse manner, while ViT was more specific, yet it is also very noisy. In the attention map however, the result is very sparse, and captures the DRUSEN phenotype almost perfectly.

4.3 Self Supervised Models

Recent work has revealed that self supervised models are able to construct an excellent low dimensional representation of natural images[12,13]. In light of this success, we wanted to experiment with the current state of the art model of this nature - DINO[12]. First, we tried several variations of the training procedures released by Facebook AI, but even though we got the loss to descend, we were unable to get meaningful representations, or semantically interesting attention maps. Next, we tried to look at the released pretrained weights on ImageNet (specifically the teacher module of DINO trained using 2 Base ViT with patch size of 8). We did not expect to get meaningful representations when passing the Kermany images through the pretrained model, since the representations the model has learned by training on natural images from ImageNet should have been, in our minds, significantly different than the representations obtained by medical images. This thought is backed up by the known fact in the vision community that natural images have substantially different distribution of features than medical images, and so we were surprised

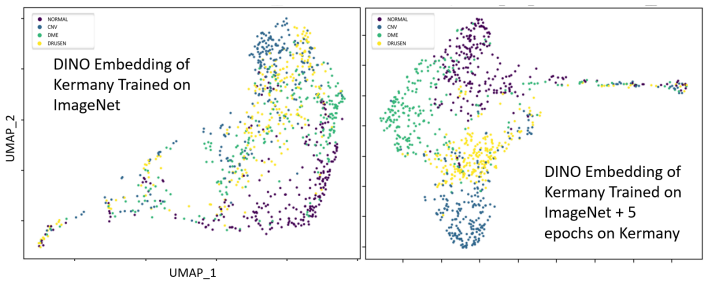

Figure 8: UMAP visualization of embedding space under DINO model. Even when trained on ImageNet, we can see that the representations out of DINO are very meaningful, as they tend to "cluster" images in the same class closely together. When we continue training on Kermany we see a significant improvement in the embedding, as the classes become more defined and spread apart.

to see the representations obtained - figure 8. As we can see the representations seem to be less meaningful than when the model was trained on this data (which is to be expected), but it is still surprising that it managed to clump the classes somewhat together. To get a sense of how good this representation is, we ran K-NN classifier on top of the model representation and got the following results for different K: $K = 5$***: Accuracy 96.28*** $K = 10$***: Accuracy 94.94***. We can infer from this that the representations are indeed quite meaningful. Next, we wanted to continue the training of the model, and trained it on the Kermany dataset with tuned hyper-parameters. In addition, we wanted to see the attention maps, as in the original DINO paper[12] (the most striking results was the attention visualization). Results are in figure 9. These results show that DINO, more than any other method, is capable of capturing the medical pathology of each of the diseases (See figure 1 to compare to true medical annotations). This result goes hand in hand with the fact that supervised models are very limited, and they can use all of the available biases in the data to reach a higher accuracy, while unsupervised models are capable of obtaining a better abstraction of the data.

5 Discussion

In this work we have shown various interpretation tools on convolution based, as well as attention based, deep learning models. We have shown that by fitting these models to a high test accuracy we are able to analyze the models by their gradients, and able to get a coarse outline of the pathology involved. In attention based models, in addition to gradient methods, we were also able to construct attention maps, and it was surprising to us that on the same images attention did not always give the same result as the Grad-CAM. In our observation, it has seemed as though usually ConvNext gave the the most precise annotations of the pathologies, Resnet was slightly worse, and tended to give more coarse results, and ViT gave mostly noisy results, but produced attention maps, which achieved the best supervised results (both in terms of resolution, and in terms of medical significance), as well as more robust results. We have also seen that some architectures are able to construct meaningful representations of the scans, and that these models got better results in the interpretation section, which leads us to believe that better low dimensional representations might contribute to a more capable model, medically speaking - meaning that it better captures the underlying pathology of the disease, and not just "shortcut learning"[14] to achieve a high accuracy. To dive deeper into the capabilities of attention based models, we have also experimented with self-supervised approaches such as DINO. DINO has surprised us with its performance on the Kermany dataset, even when loaded with pretrained weights from ImageNet, which goes to show the power of these large self supervised transformers. In a different subject, we wanted to acknowledge that the classification assignment here is not necessarily well defined - from our biological background we know that disease is generally a spectrum and no individual is strictly speaking either healthy or sick, thus providing discrete annotations might lead to some problems. For example, we would have liked that the embedding space would reflect the severity of the disease or be expressive enough to include a data point with two diseases, but here we could not enforce such a construction. One last thing to mention is that these scans were labeled by human annotators who might make mistakes, as this is a medical diagnosis and it could have a certain amount of noise. With all of this information at hand, we believe this (as well as limited computing power) is why not all architectures were able to reach high accuracy on this classification task so easily.

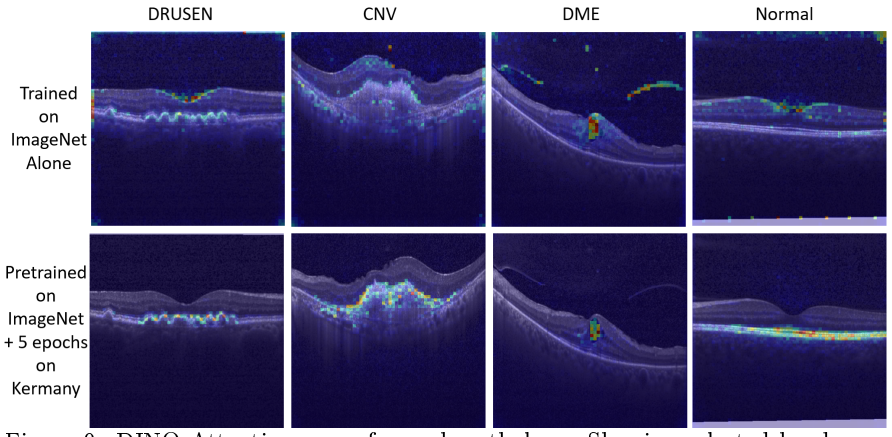

Figure 9: DINO Attention maps for each pathology. Showing selected heads with semantically and medically meaningful information. For example, in DRUSEN the model was able to fit the small accumulations almost perfectly. In DME the model was capable of focusing the attention almost exclusively on the retinal thickening (big black hole in the middle of the retina) - recall that non other supervised trained model was capable of focusing attention (or gradients) so well on the pathology of DME. Even when only trained on ImageNet, the attention maps capture the pathology very well, but when trained on Kermany for a couple of epochs they become more sharp and focus on the phenotype even better.

6 Methods

For most of the models we used the timm package, with a few exceptions that we took directly from GitHub - ConvNext implementation[20], as well as the DINO implementation[21] which we modified. We trained the models on WAIC in a parallel fashion with the help of the wonderful package weights and biases. For the Grad-Cam algorithms we used pytorchGradCam[22]. To get attention maps in section 4.2.3 we used Hila schefers implementation of her paper [19]. For UMAP we used the UMAP python package, and default hyper parameters. For DINO we used the official Facebook AI implementation, and modified it according to suggestions in the GitHub repository as well as several articles on the manner[18].

7 Related Work

There have been attempts at trying to use gradient methods to detect abnormalities other medical datasets in MRI scans[3]. There have also been attempts at using gradient methods for a single disease in OCT scans[24], we have used these to guide our work. In addition, Lundervold et al used transformers for application in medical data [15,16], and we observed the use of self-supervised models in medical imaging reference[17].

8 References

1. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.
2. Yoon, J., Han, J., Park, J.I. et al. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Sci Rep 10, 18852 (2020). https://doi.org/10.1038/s41598-020-75816-w
3. Masood, M., Nazir, T., Nawaz, M., Mehmood, A., Rashid, J., Kwon, H. Y., ... & Hussain, A. (2021). A Novel Deep Learning Method for Recognition and Classification of Brain Tumors from MRI Images. Diagnostics, 11(5), 744.
4. Tan, M., & Le, Q. (2021, July). Efficientnetv2: Smaller models and faster training. In International Conference on Machine Learning (pp. 10096-10106). PMLR.
5. Trockman, A., & Kolter, J. Z. (2022). Patches Are All You Need?. arXiv preprint arXiv:2201.09792.
6. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. arXiv preprint arXiv:2201.03545.
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
8. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357). PMLR.
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).
10. Ippolito, P. P. (2022). Hyperparameter Tuning. In Applied Data Science in Tourism (pp. 231-251). Springer, Cham.
11. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
12. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9650-9660).
13. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.
14. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.
15. Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102-127.
16. Matsoukas, C., Haslum, J. F., Söderberg, M., & Smith, K. (2021). Is it time to replace cnns with transformers for medical images?. arXiv preprint arXiv:2108.09038.
17. Truong, T., Mohammadi, S., & Lenga, M. (2021, November). How Transferable Are Self-supervised Features in Medical Image Classification Tasks?. In Machine Learning for Health (pp. 54-74). PMLR.
18. https://medium.com/@mllabucu/transformer-based-self-supervised-learning-for-medical-images-41395d069829
19. Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 782-791).
20. https://github.com/facebookresearch/ConvNeXt
21. https://github.com/facebookresearch/dino
22. https://github.com/jacobgil/pytorch-grad-cam
23. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.
24. Yoon, J., Han, J., Park, J.I. et al. Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Sci Rep 10, 18852 (2020). https://doi.org/10.1038/s41598-020-75816-w
25. Masood, M., Nazir, T., Nawaz, M., Mehmood, A., Rashid, J., Kwon, H. Y., ... & Hussain, A. (2021). A Novel Deep Learning Method for Recognition and Classication of Brain Tumors from MRI Images. Diagnostics, 11(5), 744.

9 Code Availability & Extra Results

All the code is available under the following Git repository: https://github.com/Guylu/OCT_Interpretability . In addition we provide a link to our weights & biases projects to see additional results: https://wandb.ai/guylu