Investigating the Robustness of Pre-trained Networks on OCT-Dataset

Rama Hasan M.Sc.
Chair of Media Informatics, Chemnitz University of Technology
Chemnitz, Germany D-09111
rama.hasan@s2017.tu-chemnitz.de

Dipl.-Inf. Holger Langner
Professorship Media Informatics, University of Applied Sciences Mittweida
Technikumplatz 17, Mittweida, Germany
holger.langner@hs-mittweida.de

Prof. Dr. Marc Ritter
Professorship Media Informatics, University of Applied Sciences Mittweida
Technikumplatz 17, Mittweida, Germany
marc.ritter@hs-mittweida.de

Prof. Dr. Maximilian Eibl
Chair of Media Informatics, Chemnitz University of Technology
Chemnitz, Germany D-09111
maximilian.eibl@informatik.tu-chemnitz.de

Abstract: Convolutional Neural Networks (CNN) is one of the main categories that have proven highly effective in various high-level tasks such as image classification. Pre-trained Neural Networks are models introduced in ILSVRC (ImageNet-Large-Scale-Visual-Recognition-Challenge) which have been trained successfully for hundreds of hours on powerful GPUs. Furthermore, they are applicable to new application domains. The aim of this work is to investigate the effectiveness and the application of pre-trained models from natural (non-medical) images to images from the OCT (optical coherence tomography) domain in ophthalmology. The experiments show the robustness of a series of models without the demand to train a model from scratch again, what leads in effect to reduced training times and computational costs.

Keywords: Pre-trained CNN, Transfer Learning, Deep Learning, OCT dataset, Ophthalmology, ILSVRC

1 Introduction

In recent years Convolution Neural Networks (CNNs) have been used widely as a powerful tool to solve several Machine-learning tasks in several domains like natural language processing, speech recognition and computer vision [1] as well as semantic segmentation [2] or object detection [3]. The power of CNNs became stronger and more effective after the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition in 2010 which made a revolution through the efficient use of graphics processing units (GPUs), rectified linear units, new dropout regularization, and effective data augmentation [4]. The success was achieved primarily by deep CNNs, while the depth of the network makes it more robust and allows for extracting a set of discriminating features at multiple levels of abstraction. Training a deep CNN from scratch requires a huge amount of labeled training data that represents a big challenge in domains like medical image classification and detection since in a lot of use cases or applications, it is not easy to obtain such high numbers of labeled data. In addition to the extensive computing and storage resources that the network requires in order to overcome the training time-consuming. However pre-trained Neural Networks introduced in ILSVRC have been trained on a large benchmark (of natural images) dataset [5] for hundreds of hours on powerful GPUs in order to solve a problem similar to the one that we want to solve in the remainder of the paper. Thus, they could be used as a starting point for a new training problem without the need to train our network from scratch again, especially by tweaking the already trained convolutional layers in order to fit our problems by fine-tuning and transfer-learning [6]. Despite the significant differences between natural and medical images, natural image descriptors such as the scale-invariant feature transform (SIFT) [7] and the histogram of oriented gradients (HOG) [8] have been widely used for object detection and segmentation in medical image analysis. Recently, several studies are employed to solve diagnosis medical problems by using transfer learning.

Azizpour [9] suggests that the success of knowledge transfer depends on the contrast or difference between the dataset on which a CNN is trained and the dataset to which the knowledge is to be transferred. The study shows that it is possible to transfer the knowledge from networks trained on natural (non-medical) images to medical images. In Bar et al. [10] pre-trained CNNs are used as a feature generator for chest pathology identification. Ginneken et al. [11] suggest that the integration of CNN-based features together with handcrafted features enables improved performance. Chen et al. [12] used the fine-tuned pre-trained network to localize standard planes in ultrasound images. Tajbakhsh et al. [13] show that fine-tuned CNNs and fully trained CNNs outperform the corresponding handcrafted alternatives in medical imaging applications.

The aim of this work is to investigate the effectiveness of the application of pre-trained (natural image) models on specifically chosen foreign domains in order to determine the degree of transferability. OCT (Optical coherence tomography) images have been used in the following study. The Experiments carried out make use of two of the most widely spread pre-trained CNNs: VGG16 and Resnet50 [14,15]. In addition, and in contrast, a CNN handcraft architecture has been built and trained from scratch on our OCT-image set. In order to better understand how Convolutional Neural Networks make their decisions, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [16] visualization method on a pre-trained Resnet50.

The remainder of this study is organized as follows: Section 2 presents the description of the OCT datasets. An overview of pre-trained Convolutional Networks is given in Section 3. Methodology and applied networks architectures are briefly outline in Section 4. Section 5 comprises our experimental study and an introduction to our results. Finally, our findings are briefly summed up in Section 6.

2 OCT Dataset

"Optical coherence tomography (OCT) is an optical analog of ultrasound imaging that uses low coherence interferometry to produce cross-sectional images of the retina. It captures optical scattering from the tissue to decode spatial details of tissue microstructures. It uses infrared light from a super-luminescent diode that is divided into two parts: one of which is reflected from a reference mirror and the other is scattered from the biological tissue. The two reflected beams of light are made to produce interference patterns to obtain the echo time delay and their amplitude information that makes up an A-Scan. A-Scans that are captured at adjacent retinal locations by transverse scanning mechanism are combined to produce a 2-dimensional image." [17].

Our dataset consists of real clinical images which had been acquired during ten years of practice at The Eye Center in the Medical Center of the University of Freiburg in Germany during 2007 and 2018. It contains ophthalmological data for about 3,600 patients. Each patient suffers from Age-Related Macular Degeneration (AMD) [18] or a related disease such as (diabetic retinopathy or retinal vein occlusion). The data for each patient had been collected during a long-term application of Anti-VEGF therapy [19], and it remains unfiltered, i.e. patients suffer from other eye diseases (e.g. glaucoma or cataract), too. Figure (1.a) shows a healthy macula, the Retinal Pigment Epithelium in the middle appears almost as a straight and smooth line. On the other hand, the presence of druses represents an optical marker for dry AMD (see Figure (1.b)).

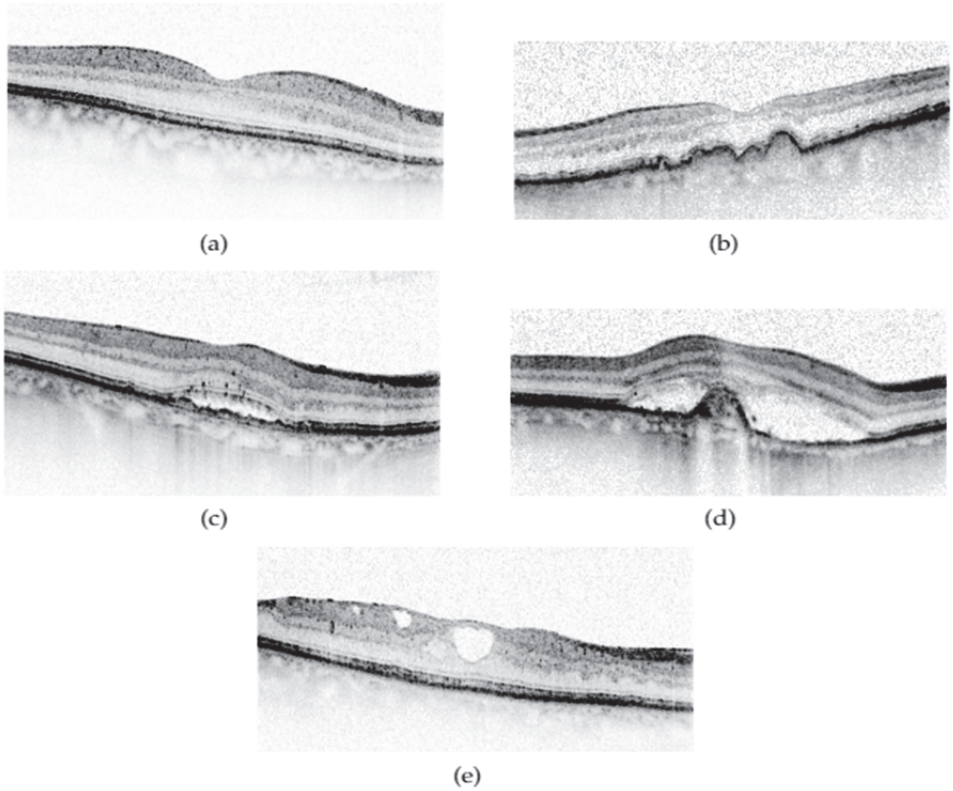

Figure 1 – OCT-samples: (a) Healthy Macular, (b) dry AMD with druses,
(c) subretinal fluid / oedema and choriodal neovascularizations,
(d) subretinal fluid and a large scarf (fibrosis), (e) intraretinal fluid.
[Data provided with courtesy from the Medical Eye Center in Freiburg Germany.]

Typical signs referring to fluid AMD as the new but abnormal blood vessels grow (Choroidal Neovascularizations) are contained in Figure (1.c). The leakiness, which leads to aggregation of fluid, i.e. intraretinal or subretinal edema, also leads to a scarf (fibrosis) as shown in Figures (1.d and 1.e) respectively. We faced several OCT data training problems. In this work, our experiments focus on the Visual Acuity Performance (VP) classification problem where only a few patients could be safely associated with a specific performer class, where after giving the very first OCT finding of a given patient, and after a long period of time and therapy, the performance class could be expected with a significant confidence.

3 Pre-Trained Networks

The VGG-16 is a CNNs which is pre-trained deep network using more than one million images retrieved from the ImageNet dataset. This network is designed by its simplicity employing only 3×3 convolution layers which are stacked on top of each other at increasing depth. The volume size is minimized by Max Pooling. Then, two fully connected layers (of 4,096 neurons) are followed by a softmax classifier. VGG-16 contains 16 deep layers and is capable to classify images into 1,000 classes such as a mouse, keyboard, pencil and animals etc. Consequently, the network has learned extensive feature representations for a variety of images. The network has an input image size of 224 x 224 pixels.

ResNet-50 is a pre-trained convolutional neural network which also utilized more than 1 million images retrieved from the ImageNet dataset during the training process. ResNet-50 employs deep residual learning on 50 layers and has the ability to classify a large number of objects into 1,000 classes like VGG-16 while also maintaining an input image size of 224 x 224 pixels.

4 Methodology

In the following, we practically investigate the robustness of pre-trained natural image CNNs on the OCT domain. The biggest challenge arises through the difficulty of correctly classifying these kinds of medical images by ophthalmologists. We examine the transferability of knowledge embedded in pre-trained CNNs for this type of medical images. We also employ the Grade-CAM technique to visualize the regions on the image input, which is important for these pre-formed CNN predictions, in order to gain a better understanding of how these networks create their decisions. Our experiments are conducted on our classification problem of Visual Acuity Performance (VP). The VP problem set contains three classes:

- 1-decreasing: the visual acuity of the patient drops after a period of time.
- 2-stable: the visual acuity of the patient stabilizes after a long period; however, it needs consistent therapy.
- 3-increasing: the patient's visual acuity increased immediately from therapy,
  (the problem of the outcome of the therapy quality).

We used VGG16 pre-trained network [20] keeping the weights and filters of the top layers of the network which identify simple features like edges, lines, and corners and retrained the last four layers. Then, we added a fully connected layer followed by a Softmax activation [21] with a number of outputs corresponding to the number of classes in each OCT-image set. The same procedure was applied to the Resnet50 pre-trained network. We apply Grad-CAM to all those networks in order to highlight the specific discriminative regions of an image detected by the pre-trained CNNs. The annotated ground truth labels are converted and forwarded to the last layer in order to calculate the appropriate class scores.

The workflow of Grad-CAM is shown in Figure 2, where for all classes, the gradient is set to zero except that the true class which is set to 1. The error signal is then back propagated to the feature map of interest where the Grad-CAM localizations use the gradients of the target class flowing into the final convolutional layer to create a coarse localization map which highlights the important parts in the image for the predicting of the respective class [16].

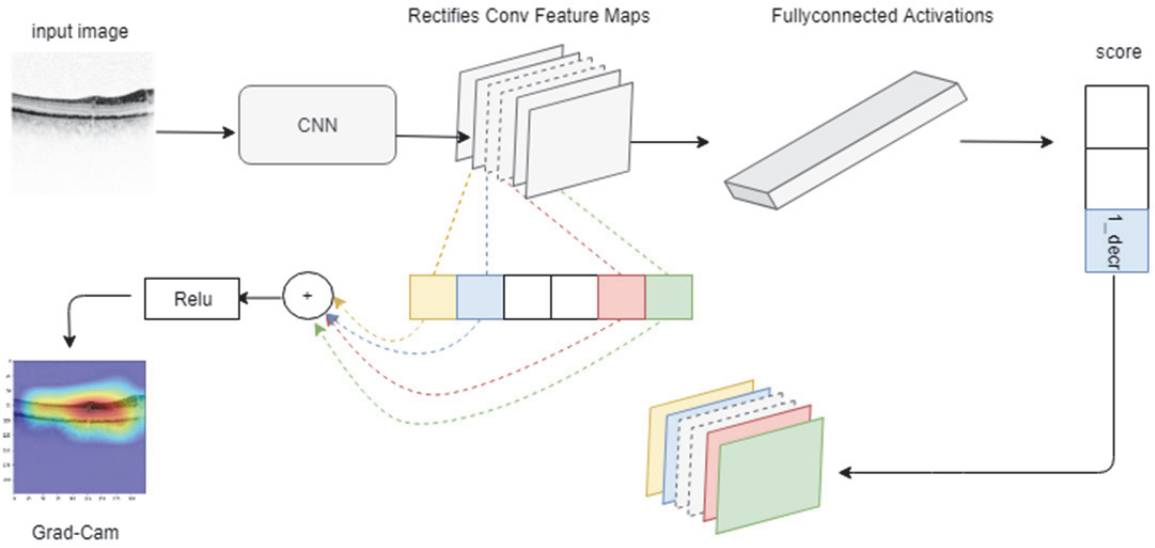

Figure 2 – Grad-Cam

5 Experimental results

Three convolutional neural networks (pre-trained VGG16, pre-trained ResNet50 and a handcrafted CNN) are used in our experiments. Since the lower layers only detect more (localized) general and simple features like edges and lines, and as the network increases the complexity in the higher layers, we decided to retrain the last four layers and leave the others frozen whereas a frozen layer does not change during training in VGG16 pre-trained network. Then, we added a fully connected layer with 512 neurons and ReLU activation [22] followed by dropout layer in order to avoid overfitting. In addition, an output layer with a number of neurons matches the number of classes followed by Softmax activation. We employed an Adam optimizer with a learning rate of 0.0001 [23]. Within the Resnet50 pre-trained network, also the first 41 layers are frozen; a flattened layer followed by a dense layer with several neurons matching the number of classes by using Softmax activation. SGD optimizer is used with 0.01 learning rate [24]. Our handcrafted CNN consists of four convolutional layers with filter-weights of sizes (5x5x32), (5x5x64), (7x7x64), (7x7x128), respectively. Each convolutional layer is followed by a (2x2) max pooling and ReLU activation. In addition, a fully connected layer with 512 neurons is followed by a last output layer with a number of neurons equal to the number of classes within the Softmax activation.

Our dataset consists of 8,434 OCT-images. 20% of our samples are used as the validation set and 20% as a test set. After the training phase of our three networks (VGG16, ResNet50 and our handcrafted model) for 10-times of runs we got an accuracy range over a validation set of values between (92.20% - 94.54%), (92.57% - 94.72%), (75.11%-76.90%), for these models respectively. Thus, as shown in Figure 3 which plots the training and validation accuracy and the training and validation loss during the training process of these three CNNs in the last run, the two pre-trained models outperformed our handcrafted model which is had been solely trained from scratch on our selected three classes OCT-image set.

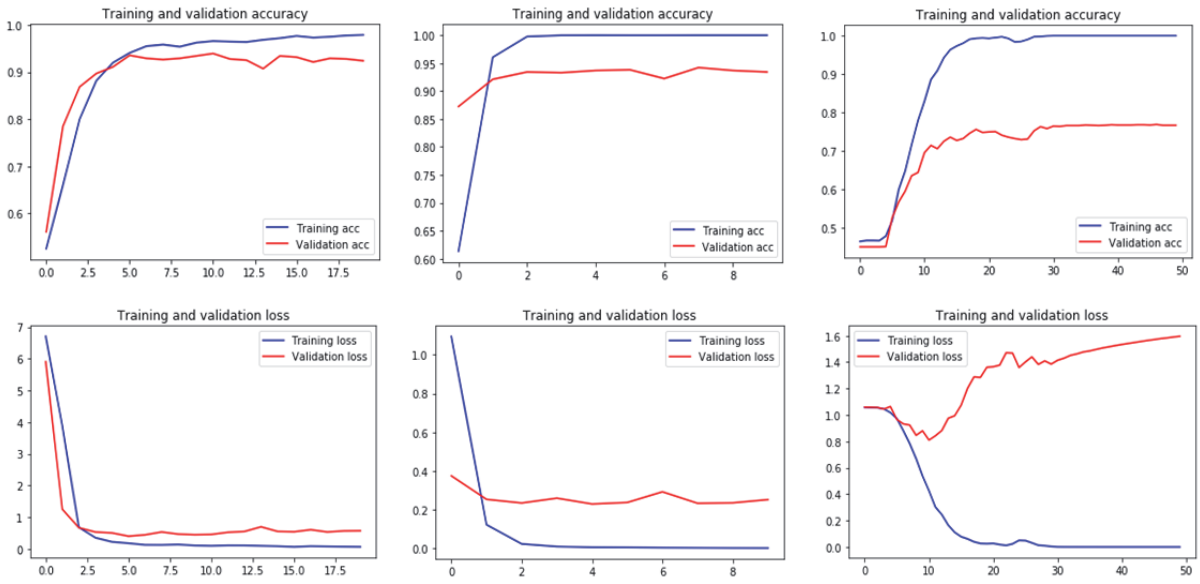

Figure 3 – Training and validation accuracy and loss of the last runs for our modified models of: (a) VGG16, (b) ResNet50, (c) Handcrafted model.

For more robust evaluation of each classification network, we calculated the classification accuracy after training each CNN networks ten-times over an OCT-test set (test-accuracy), which represented 20 % from our OCT image samples. The results show the resulting performance of pre-trained CNN over our Handcrafted CNN. Not only for validation and test accuracy, but also for the number of epochs our Handcrafted CNN performs worse while employing 50 epochs to fit training image samples in comparison to 20 and 10 epochs for VGG16 and ResNet50, respectively. As shown in Table 1, the best test accuracy was obtained by pre-trained VGG16 with a test accuracy average of 88.814 %. ResNet50 achieved even better in terms of the number of epochs, which was 10 epochs.

Table 1 –The results of test accuracy and standard deviation

CNN	Epochs	Test Accuracy	Average Accuracies	Standard deviation
VGG16	20	87.22 % - 90.64%	88.814 %	0.943
Resnet50	10	81.07 % - 82.49 %	82,17 %	0.799
Handcraft	50	75.80 % - 79.79 %	77.82 %	1.104

As a measurement of how much the test accuracy varies over the runs, we calculated the standard deviation of test accuracies for each CNN. Figure 4 shows the normal distribution of the resulting test-accuracy values over the ten runs for CNNs.

We applied Grad-Cam visualization methods on ResNet50 in order to understand exactly where CNN is looking in the image to actually distinguish between the classes. As figure (5) shows Grad-Cam of three OCT-image samples related to three classes.

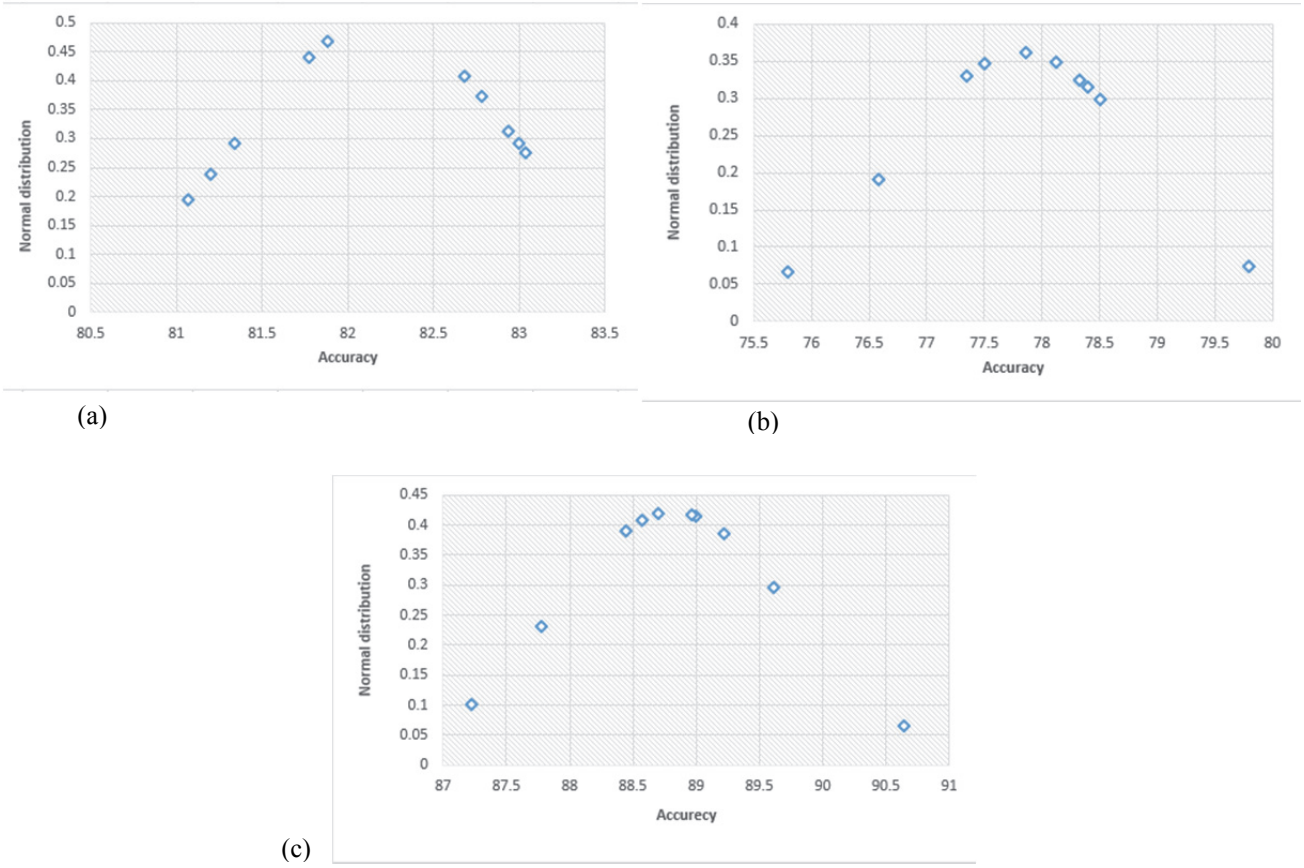

Figure 4 – The normal distribution of test accuracy values over ten runs:
(a) Pre-trained ResNet50 CNN, (b) Handcrafted CNN, (c) Pre-trained VGG16

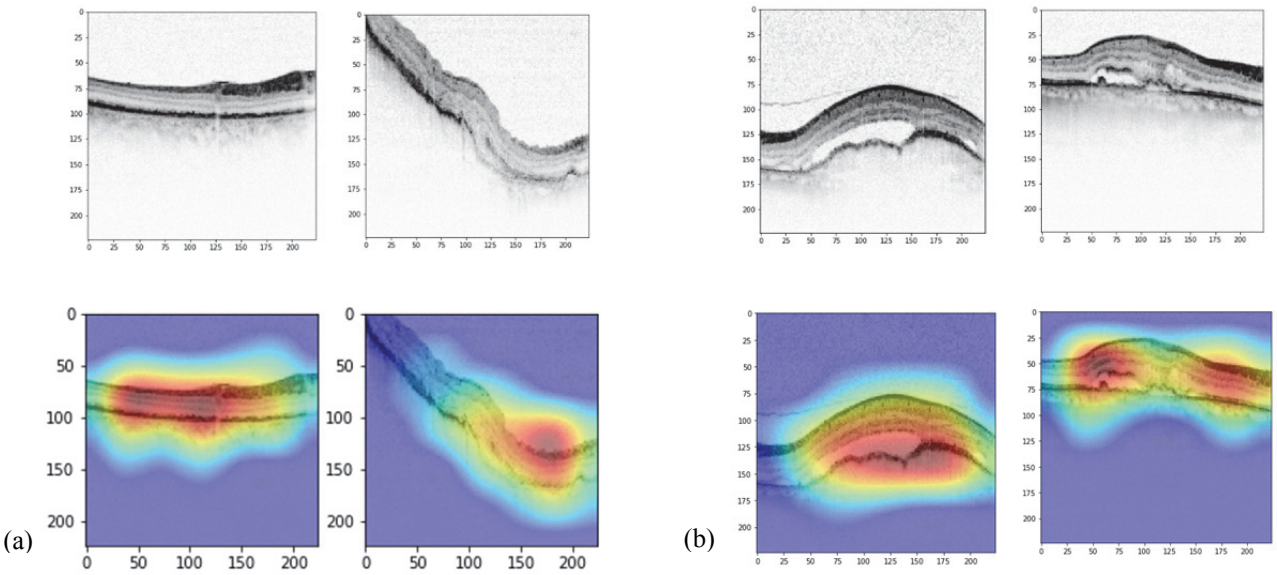

Figure 5 – Grad-CAM applied to different OCT images. (a) 1_decr. (b) 2_stable

6 Conclusion and Future Work

In this study, we introduced an experimental study in order to investigate the robustness of pre-trained Neural Networks towards a special kind of medical images (OCT-images) described in Section 2 without the need to retrain these networks from scratch. We explore the outcome of keeping the first few layers of the pre-trained networks are while retraining the latter layers in order to adjust and fit our classification problem. The experimental results show the resulting performances of pre-trained networks over a Handcraft network, which has been built and trained from scratch, and which augments the concept of knowledge transfer despite the big difference between the natural and medical image domains. We also applied Grad-CAM visualization method on pre-trained ResNet50 to get a better understanding which features appear relevant to the CNNs in order to distinguish between the different medical image classes. Future work encompasses the investigation of semi-automated and active learning algorithms to solve the massive annotation problems, since these algorithm classes are capable to fill the gap between labelled and unlabelled data while idealistically only querying such samples that would lead to an increase in precision or accuracy. In addition, it is essential to enhance the current framework and tool chain to address at least a wider variety of real-world ophthalmologic challenges.

7 Acknowledgement

We like to acknowledge that Prof. Dr. Andreas Stahl and the collaborators of the TOPOs project provided the OCT image data that was used in this study, as well as the ophthalmological background. TOPOs ("Therapievorhersage durch Analyse von Patientendaten in der Ophthalmologie") is a collaborative project that is funded by BMBF ("Bundesministerium fur Bildung und Forschung") (FKZ: 13GW0170B) from March 2017 to January 2020. The European Social Fund (ESF) also funded this work within the Innovative PhD Scholarship entitled "Aggregation, Visualisierung und Optimierung von überwachten Deep Learning-Technologien mit Hilfe der virtuellen und erweiterten Realität".

Dieses Projekt wurde finanziert aus Mitteln der Europäischen Union

Europa fördert Sachsen.

EFRE

Europäischer Fonds für regionale Entwicklung

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," Comput. Intell. Neurosci., vol. 2018, pp. 1–13, Feb. 2018.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440, Boston, MA, 2015.

[3] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, L. Van Gool, and K. Leuven, "Weakly Supervised Cascaded Convolutional Networks.", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), *pp. 5131-5139, Honolulu, HI, 2017.*

[4] M. Thoma, "Analysis and Optimization of Convolutional Neural Network Architectures," arXiv preprint arXiv:1707.09725, 2017.

[5] "ImageNet." [Online]. Available: http://www.image-net.org/. [Accessed: 12-Jul-2019].

[6] T. Wang, J. Huan, B. Research, and M. Zhu, "Instance-based Deep Transfer Learning.", *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 367-375, Waikoloa Village, HI, USA, 2019.

[7] T. Lindeberg, "Scale Invariant Feature Transform," Scholarpedia, vol. 7, no. 5, p. 10491, 2012.

[8] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893,2005.

[9] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From Generic to Specific Deep Representations for Visual Recognition.", *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 36-45, Boston, MA, 2015.

[10] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification.", *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 294-297, New York, NY, 2015.

[11] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015, pp. 286–289.

[12] H. Chen et al., "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," IEEE J. Biomed. Heal. Informatics, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[13] N. Tajbakhsh et al., “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?,” IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1299–1312, 2016.
[14] K. Simonyan and A. Zisserman, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION”, arXiv preprint arXiv:1409.1556, 2015.
[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016..
[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” Proc. IEEE Int. Conf. Comput. Vis., vol. 2017-Octob, pp. 618–626, 2017.
[17] M. Bhende, S. Shetty, M. Parthasarathy, and S. Ramya, “Optical coherence tomography: A guide to interpretation of common macular diseases,” Indian J. Ophthalmol., vol. 66, no. 1, p. 20, 2018.
[18] “Age-Related Macular Degeneration (AMD) | National Eye Institute.” [Online]. Available: https://nei.nih.gov/health/maculardegen. [Accessed: 09-Jul-2019].
[19] T. Y. Y. Lai, C. M. G. Cheung, and W. F. Mieler, “Ophthalmic Application of Anti-VEGF Therapy,” Asia-Pacific J. Ophthalmol., vol. 6, no. 6, pp. 479–480, 2017.
[20] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv preprint arXiv:1409.1556, Sep. 2014.
[21] Y. Tang, “Deep Learning using Linear Support Vector Machines”, arXiv preprint arXiv:1306.0239, Jun. 2013.
[22] C. Enyinna Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation Functions: Comparison of Trends in Practice and Research for Deep Learning.”, arXiv preprint arXiv:1811.03378, 2018.
[23] D. P. Kingma and J. Lei Ba, “ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.”, arXiv preprint arXiv:1412.6980, 2014.
[24] H. Robbins and S. Monro, “A Stochastic Approximation Method”, The annals of mathematical statistics, pp.400-407, 1951.