Aliya Mohammed, Carol Geppert, Arndt Hartmann, Petr Kuritcyn, Volker Bruns, Ute Schmid, Thomas Wittenberg*, Michaela Benz and Bettina Finzel

Explaining and Evaluating Deep Tissue Classification by Visualizing Activations of Most Relevant Intermediate Layers

https://doi.org/10.1515/cdbme-2022-1059

Abstract: Deep Learning-based tissue classification may support pathologists in analyzing digitized whole slide images. However, in such critical tasks, only approaches that can be validated by medical experts in advance to deployment, are suitable. We present an approach that contributes to making automated tissue classification more transparent. We step beyond broadly used visualizations for last layers of a convolutional neural network by identifying most relevant intermediate layers applying Grad-CAM. A visual evaluation by a pathologist shows that these layers assign relevance, where important morphological structures are present in case of correct class decisions. We introduce a tool that can be easily used by medical experts for such validation purposes for any convolutional neural network and any layer. Visual explanations for intermediate layers provide insights into a neural network's decision for histopathological tissue classification. In future research also the context of the input data must be considered.

Keywords: Digital Pathology, Deep Learning, XAI, Evaluation, Grad-CAM, Intelligent user interface

1 Introduction

In the past decade many new approaches incorporating deep convolutional neural networks (DCNNs) have been proposed in the field of digital pathology for various tasks, such as the detection, segmentation, or classification of tissue regions, based on digitized slides of stained tissue samples. Nevertheless, for the clinical acceptance of such approaches, it is quite essential, that the underlying DCNN models are comprehensible and verifiable, hence making use of *similar* meaningful morphological cell and tissue structures and patterns as the pathologists. To this end, partially unsolved is the question, *how* such DCNNs come to a segmentation or classification result, *which* pixels in the input image contribute to these decisions and *do* involved pixels relate to the clinical expert's knowledge and expectations. In literature several methods for visual explanation and exploration of DCNN-based segmentation and classification have been suggested, such as LRP [1], LIME [3], Grad-CAM [4,7,8,9] and others [2,5]. Although, some works use visual explanations to evaluate, whether a DCNN has learnt features corresponding to the knowledge of experienced pathologists, they usually only do so for last convolutional layers. Our work is based on the hypothesis that it may be also of importance to examine intermediate layers for morphologically relevant structures, since the composition of features a DCNN learnt may only be detectable, when looking at multiple layers.

To overcome the limitations of existing approaches, we apply Grad-CAM to produce visual explanations for most relevant intermediate layers of a DCNN. The level of relevance is computed based on activation values per pixel. We apply our method on tiles taken from whole-slide images (WSIs) and provide an interactive user interface to facilitate the evaluation of visual explanations. We focus on an automated selection of most relevant layers from a competitive DCNN model and present them to a pathologist who qualitatively evaluates the DCNN's adherence to morphologically important features.

2 Related Work

Generating visual explanations for DCNNs is a widely established approach to shed light into the decisions of classifiers [6]. Among the most popular approaches for explainable classification of histopathological data is *Gradient-weighted Class Activation Mapping* (aka 'Grad–CAM') [4,7,8,9]. It provides a method to generate visual explanations in the form of heatmaps for all convolutional layers of a DCNN. This is of parti-

*Corresponding author: **Thomas Wittenberg:** Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany, thomas.wittenberg @iis.fraunhofer.de
Bettina Finzel, Aliya Mohammed, Ute Schmid: Otto-Friedrich-Universität Bamberg, Germany
Carol Geppert, Arnd Hartmann Institut für Pathologie, Universitätsklinik Erlangen, FAU Erlangen-Nürnberg, Germany
Michaela Benz, Volker Bruns, Petr Kuritcyn: Fraunhofer IIS, Erlangen, Germany

cular interest, since intermediate layers may focus on structures that are relevant w.r.t. morphological properties learnt by a DCNN. Efforts already made in this area are summarized in the following paragraphs and extended by our work. For example, Korbar et al. [7] studied different visualization methods to give insight into whole-slide image classification results obtained from Deep-Learning models. They applied Grad-CAM to assign individual pixels to regions of interest, whenever their activation values were higher than a given threshold w.r.t. a predicted class. These regions of interest served as masks, highlighting features in an input image that contributed to certain classification outcomes. In contrast to [7], we apply Grad-CAM to examine, whether activation values gained from intermediate convolutional layers correspond to morphologically important features characterizing a predicted tissue class. Another line of research presented by Tang et al. [8] used Grad-CAM to better understand outcomes of neural architecture search for tissue classification. They applied Grad-CAM to the last convolutional layers of various neural networks to get insights into the differences in performance. However, unlike our approach, they did not consider intermediate layers. Like our method, Kowsari et al. [9] applied Grad-CAM on hierarchical tissue classification to find out, whether resulting heatmaps contain medically relevant features. Similar to Tang et al. [8], they analyzed only the last convolutional layer. None of the reviewed works inspects, whether a DCNN learnt morphologically relevant features that may only be detectable at intermediate layers. Furthermore, no interactive user interface was provided to facilitate the pathologist's task of visual explanation evaluation. Our work attempts to close this gap for improved explanation and evaluation of deep tissue classification.

3 Material and Methods

In the following sections we present the data set we used to showcase the capabilities of our approach for automated tissue classification. We further describe the architecture of the neural network we applied. We introduce our proposed method that selects intermediate layers of a neural network by their relevance sum. Finally, we introduce our interactive user interface that can be used to validate the tissue classification.

3.1 Histological Image Data

For the training, validation, and evaluation of the network (see Section 3.2) an image data collection derived from 152 annotated HE-stained whole-slide images (WSIs) of the colon provided by the pathology department of the University Hospital Erlangen (UKER) was used [13,14]. All slides were digitized using a 3D Histech slide scanner with a 20x-fold magnification yielding a spatial resolution of 0.22 x 0.22 μm^2 / pixel. The image data depicts sections of the colon with seven histological tissue classes including inflammation, mucosa, tumour cells, necrosis, mucus as well as connective combined with adipose tissue and muscle tissue, see Fig. 1 for examples of image tiles (size 224x224 pixel). Three disjoint sets of image tiles were used for training (2,173,515 tiles), validation (719,010 tiles) and testing (1,381,316 tiles).

Figure 1: Examples for different HE-stained tissue types from the colon, used to train the network. From left to right: muscle, connective combined with adipose tissue, necrosis, mucosa, tumor, inflammation, and mucus. Each tile has a size of 224 x 224-pixel corresponding to 49 x 49 μm^2.

3.2 Neural Network Architecture

Our focus was to show the principal benefit and usability of a visual explanation component based on Grad-CAM. Therefore, we chose VGG16 [10], being a simple DCNN architecture for our investigations, constructed only of 13 convolutional and three fully connected layers. The training pipeline was implemented using the TensorFlow framework (Vers. 2.4). The network was initialized with ImageNet weights and input image patches were zero-centred. To compensate for the imbalance between the individual classes, the small classes were oversampled so that an equal distribution of classes was achieved in each batch. The network model was trained for 10 epochs using the ADAM optimizer with a learning rate of 0.001. Also, domain specific data augmentation addressing the variance in colour appearance was applied. This includes stain-specific, as well as hue and saturation augmentation [12].

3.3 Grad-CAM

Gradient-weighted Class Activation Mapping (aka 'Grad–CAM') [4] is an established method to visualize from which pixels the classification response of a DCNN model is coming from. For this purpose, an activation value is calculated per pixel. The higher the value, the higher is the contribution of a pixel to the output class. The Grad-CAM approach is class-specific; hence it can provide a separate visualization for every class Ω_i present in the image tile under consideration. Grad-CAM allows for extracting heatmaps for intermediate layers of a DCNN. Fig. 2 depicts an image patch with mucosa (left-most). It is characterized by goblet cells that are unicellular

glands and have an apical cell boundary. Fig. 2 further shows selected heatmaps resulting from the Grad-CAM approach (two rows of images on the right) applied to the example image, namely for the most relevant layers, being the pool layer *block2_pool* and the convolutional layers *block3_conv2*, *block4_conv1* and *block5_conv1*.

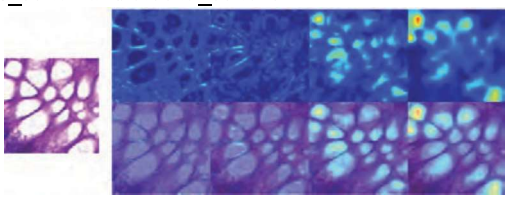

Figure 2: *Left:* Mucosa tile, classified with conf = .3118 by DCNN vs. conf = .1147 for other classes. This patch shows characteristic goblet cells (white areas) and their apical cell boundaries (darker lines around white areas). *Right:* Selected heatmaps resulting from GradCAM applied to mucosa tile. Colors relate to activation strength w.r.t. the predicted class from blue (low) to red (high activation). From top left to bottom right the heatmaps of 4 layers are depicted: *block2_pool, block3_conv2, block4_conv1* and *block5_conv1*. The bottom row shows heatmaps overlaid by the original image. The DCNN considers the apical cell boundaries in both layers *block2_pool* and *block3_conv2* as well as the inner area of the goblet cells in the layers *block4_conv1* and *block5_conv1*.

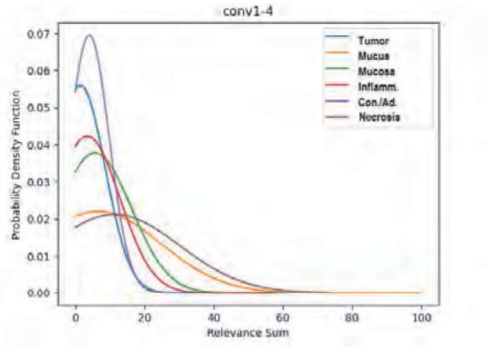

Figure 3: Layer conv1-4 (block4_conv1) belongs to the most relevant layers having the highest probability density skewed towards a high relevance sum, especially for necrosis and mucus.

3.4 Intermediate Layer Selection

The layers presented in Fig. 2 have been selected by their sum of activation values (aka 'relevance sum'). This way, we want to ensure that the DCNN's ability to recognize morphological structures is evaluated w.r.t. layers that are promising and highly contributing to the final output. For this purpose, we sum up the activation values computed with Grad-CAM for every pixel per image, class, and layer. The sum is normalized per layer w.r.t. the maximum activation value. We further compute the probability density of the relevance sum for every class to identify the layers that produce a distribution skewed towards higher relevance sums for all classes. With this approach, we could identify the four layers which had the best results for all classes (see Fig. 2). Fig. 3 shows the *relevance sum* plotted against the probability density for layer *block4_conv1* for all 7 classes. In addition to generating visual explanations for most relevant layers, we provide an interactive user interface allowing to evaluate heatmaps for all input images, all available layers as well as DCNN models.

3.5 Interaction & Visualization Interface

To facilitate the inspection of heatmaps obtained from Grad-CAM, an interface for evaluation purposes was implemented (see Fig. 4). For each input patch from an HE-stained micrograph (top left) the prediction confidences of the seven classes provided by a selected DCNN are shown (top right). On the bottom heatmaps and their overlays are displayed according to the user's selection of a layer. Also, a slide show is provided to browse through the heatmaps of the different layers.

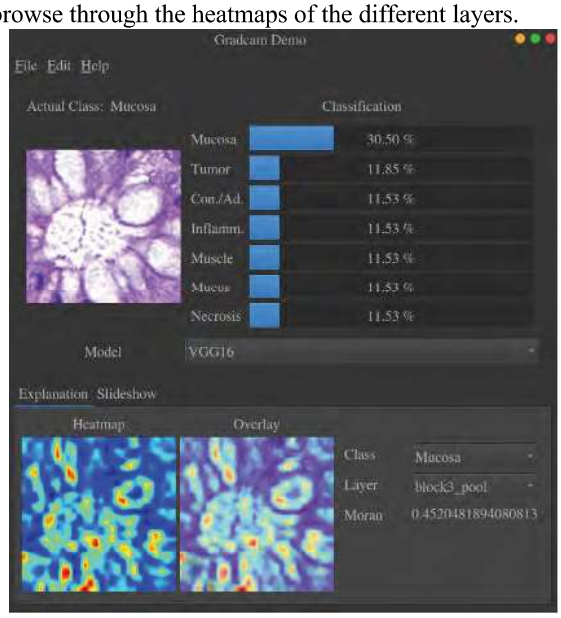

Figure 4: Interactive user interface with a view displaying the original HE image, prediction confidences produced by a CNN (which can be selected by the user) and a view for visual explanations in the form of a output heatmap, its overlay on the original image and a slideshow iterating through the heatmaps visualizations of all layers. The user can select the class for which the interface displays data as well as a specific layer for further inspection.

4 Results

The model performance was evaluated on the test dataset consisting of 1.3 million images tiles and achieved an overall accuracy of 91.3%. Fig. 5 shows the relative confusion matrix. Frequent confusion between two classes are highlighted in red. The accuracy and the confusion matrix were presented to the pathologist beforehand, hence the knowledge that with a possibility of approximately 10% an error might occur from the

DCNN presenting the image classifications.

		Predicted						
		tumor	inflamm.	con./ad.	muscle	mucosa	mucus	necrosis
GT	tumor	0.906	0.006	0.006	0.008	0.063	0.001	0.010
	inflamm.	0.010	0.893	0.026	0.016	0.054	0.000	0.001
	con./ad.	0.002	0.004	0.917	0.054	0.008	0.012	0.004
	muscle	0.002	0.001	0.089	0.898	0.005	0.001	0.004
	mucosa	0.015	0.007	0.016	0.005	0.953	0.003	0.001
	mucus	0.002	0.000	0.077	0.009	0.012	0.886	0.015
	necrosis	0.038	0.011	0.041	0.033	0.023	0.007	0.848

Figure 5: Relative confusion matrix obtained on test dataset.

Based on 100 selected tiles from the 7 classes, an experienced pathologist (CG) was asked whether the four depicted heatmaps (one for each layer) would correlate with the clinical expertise. Due to space restrictions, we cannot present all comments here, however we can conclude that on average, 3 of 4 layers contained relevant structures for correctly classified tiles according to the pathologist. A case, where the pathologist agreed with the DCNN's output has already been presented in Fig. 2. It was positively evaluated that the DCNN considered the apical cell boundaries as well as the goblet cells as relevant.

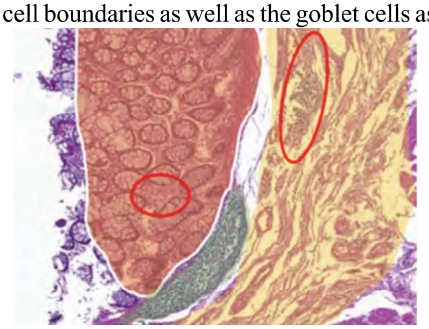

Figure 6: Typical HE-stained histological sample from the colon, depicting three types of coarsely annotated tissues, namely (from left to right) mucosa (in red), inflammation (in green) and connective tissue (yellow). The two red ellipses indicate some additional inflammation regions, which have not explicitly been labelled.

Fig. 6 shows a typical histological sample from the colon, depicting three types of coarsely annotated tissues, namely (from left to right) mucosa (in red), inflammation (in green) and connective tissue (green). From high level point of perspective this annotation is correct, especially as a large section of inflamed tissue can distinctly be seen. Nevertheless, the two red ellipses indicate some additional inflammation regions, surrounded by mucosa or connective tissue. Patches from such areas tend to be misclassified and contributed to the individual scores in Fig. 5. Here it becomes apparent that the tiles may be too small to represent important contextual information.

5 Conclusion and Outlook

We presented an approach to explain and evaluate a DCNN decision for the task of tissue classification from histopathological image patches. In contrast to other works, we provide a solution generating visual explanations based on Grad-CAM for intermediate network layers. Furthermore, we provide an interactive user interface to facilitate the evaluation of visualizations for different layers. A pathologist's evaluation of visual explanations, reveals that the model used, considers morphologically important features in 3 of 4 layers on average for correctly classified images. Our method is therefore suitable to evaluate a DCNN's performance not only w.r.t. its correct prediction but concerning its usage of domain-specific knowledge as well. Nevertheless, our approach is limited by the fact that important contextual information may not be considered by the DCNN due to the small size of the tiles. A combination of hierarchical classification, such as presented in [9], and our visualization approach could address this problem in the future.

Acknowledgement: This work has been funded by the German Federal Ministry of Education and Research (BMBF) under the project TraMeExCo (011S18056 A / B) and under the project FMD (16FMD01K, 16FMD02 and 16FMD03).

References

[1] Hägele et al.: Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. Sci Reports 10, 6423, 2020 75-

[2] Bruckert et al.: The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. Frontiers in Artificial Intelligence 3, 2020.

[3] de Sousa et al.: Local interpretable model-agnostic explanations for classification…. Sensors. 2019; 19(13):2969.

[4] Selvaraju et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proc's ICCV, 2017, 618-626.

[5] Pocevičiūtė et al.: Survey of XAI in Digital Pathology. In: Artif. Intell. & Mach. Learning for digital pathology. 2020; 12090.

[6] Tjoa & Guan. A Survey on explainable artificial intelligence (XAI): Toward Medical XAI. *Trans. Neural networks & learning systems*, *32*(11), 4793-4813.

[7] Korbar et al. Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In: *Proc's ICCCPR*. 2017. S. 69-75.

[8] Tang et al. Probeable DARTS with Application to Computational Pathology. In: *Proc's ICCV* 2021. S. 572-581.

[9] Kowsari et al. HMIC: Hierarchical medical image classification, a deep learning approach. *Information*, 2020, 11(6) S. 318.

[10] Karen & Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. Proc's Int Conf. Learning Representations. 2015

[12] Kuritcyn et al., Robust slide cartography in colon cancer histology - evaluation on a multi-scanner database. Proc's Bildverarbeitung für die Medizin 2021 229-234.

[13] Srinidhi et al.: DNN models for computational histopathology: A survey,Medical Image Analysis, Volume 67,2021, 101813

[14] Wilm et al.: Fast whole-slide cartography in colon cancer histology using superpixels and CNN classification. J Med Imag. 9(2), 2022, 027