Dissecting Deep Neural Networks for Better Medical Image Classification and Classification Understanding

Steven Alexander Hicks[1,3], Michael Riegler[1,3], Konstantin Pogorelov[1,2], Kim V. Ånonsen[6], Thomas de Lange[6,9], Dag Johansen[4], Mattis Jeppsson[5], Kristin Ranheim Randel[8], Sigrun Eskeland[7] and Pål Halvorsen[1,2,3]

[1]*Department of Informatics, University of Oslo, Norway* [2]*Simula Metropolitan Center for Digital Engineering, Norway*
[3]*Simula Research Laboratory, Norway* [4]*UiT - Artic University of Norway* [5]*ForzaSys AS, Norway*
[6]*Department of Transplantation, Oslo University Hospital, Norway* [7]*Bærum Hospital, Vestre Viken Hospital Trust, Norway*
[8]*Cancer Registry Norway* [9]*Institute of Clinical Medicine, University of Oslo, Norway*

***Abstract*—Neural networks, in the context of deep learning, show much promise in becoming an important tool with the purpose assisting medical doctors in disease detection during patient examinations. However, the current state of deep learning is something of a "black box", making it very difficult to understand what internal processes lead to a given result. This is not only true for non-technical users but among experts as well. This lack of understanding has led to hesitation in the implementation of these methods among mission-critical fields, with many putting interpretability in front of actual performance. Motivated by increasing the acceptance and trust of these methods, and to make qualified decisions, we present a system that allows for the partial opening of this *black box*. This includes an investigation on what the neural network *sees* when making a prediction, to both, improve algorithmic understanding, and to gain intuition into what pre-processing steps may lead to better image classification performance. Furthermore, a significant part of a medical expert's time is spent preparing reports after medical examinations, and if we already have a system for dissecting the analysis done by the network, the same tool can be used for automatic examination documentation through content suggestions. In this paper, we present a system that can look into the layers of a deep neural network and present the network's decision in a way that that medical doctors may understand. Furthermore, we present and discuss how this information can possibly be used for automatic reporting. Our initial results are very promising.**

***Index Terms*—computer aided diagnosis, deep learning**

I. Introduction

Machine learning, in context of image and video analysis using deep learning, has become a commonly used method in a variety of different fields such as medicine, finance, and robotics, etc. One important area of application is in the assistance of medical doctors in the detection disease during patient examinations with the purpose of avoiding overlooked abnormalities [1]. However, such deep neural networks are also somewhat black boxes (especially among end users) where very few understand what decisions lead to a given prediction. To be able to make qualified decisions and to increase the trust of the medical domain, we must demystify this black box as medical doctors often need a rationale of why the system signals a detection besides the output from the system itself. To the best of our knowledge, this is yet an unexplored area of research, especially when it comes to giving an explanation to the doctors and involving them in the system pipeline. To improve the black box-understanding, we examine an automatic disease detection system where we dissect a neural network to explore what the network "sees" as an image moves through its layers and use this information as a basis to increase understanding of how it makes a prediction. As a case study, we use live colonoscopy, which is a common gastrointestinal (GI) examination, essential for the diagnosis of most mucosal diseases in the GI tract, particularly diagnosis of colorectal cancer and its precursors. We have previously developed such a live detection system [2]–[4], and compared various machine learning techniques [5]. In the previously developed system, the endoscopist performs the colonoscopy while video frames are automatically analysed. Furthermore, the system provides visual feedback to the doctor if something abnormal is detected [6]. In this paper, we open the black box, with the goal of gaining a deeper understanding and insight into the detection process of a convolutional neural network (CNN) for three different purposes and contributions:

- *better decision support:* The medical doctors often need a reasoning of why the system returns a detection. To the best of our knowledge, this is yet an unexplored area of research, by providing intermediate heat maps from the internal process of the neural network, we gain more insight into how and why a particular prediction is produced.
- *improved data augmentation:* There are several ways to improve and augment input data in order to improve the detection rates. Using the gained intermediate information, we can observe which parts of an image is marked in each layer so that we can identify which regions result in false positives and false negatives. This can be used to improve both classification performance and training data.

2372-9198/18/$31.00 ©2018 IEEE
DOI 10.1109/CBMS.2018.00070

IEEE computer society

- *automatic report generation:* A system dissecting the network for understanding it better, can also be used for automatic examination report generation proposing both images or video clips to include and giving a reason why.

With these target improvements, we present a system looking deeper into a neural network used for GI disease detection. Using the open Kvasir [7] and CVC-968 [8] datasets, we evaluate the base performance and improvements using insights gained by the system. Based on the dissection of the network, we demonstrate how the system can be used to improve data augmentation by identifying artifacts in the images that confuse the algorithm. This information is used to perform data pre-processing which improves the detection rate and more important the generality of the model. Then, we show how the intermediate network layer information can be used to help medical experts in understanding the decisions made by the network. Finally, we demonstrate how the system can help generating automatic documentation and reports in a standardized way potentially moving more medical expert time from paper work to patient examinations.

II. Related work

Understanding layers of deep learning architectures have been a topic of research for quite some time. Some researchers try to solve this problem by using a more theoretical basis and mathematical approach such as [9], [10]. While this is important, it does not help the end users like medical experts understand the algorithmic decisions and improve their trust in the system. Other researchers try to apply a more visual approach to the problem and visualize layers using different methods such as heat maps or visual representations (texture, heat maps, etc.) [11], [12]. Based on visual content, the next natural step in the process is to generate text from the visual layers to create automatic tags or descriptions of images and videos. In the medical imaging domain, Zhang et al. [13] propose a method to generate automatic reports for automatic image diagnosis networks. The goal is to create semantically meaningful reports. As an example, they used bladder cancer detection. A similar approach can be found in [14]. Both methods can create text from the images and visualisations of the regions the algorithm activated on.

In comparison to these approaches, our system is not focused on creating automatic text, but instead use the visual attention heat maps to get an understanding of the algorithms decisions, present them to medical experts and to improve these decisions by pre-processing the data in a different more effective way. To the best of our knowledge, no related work does this.

Furthermore, pre-processing in deep learning is an often used practice, but it is hard to find a clear description of when to apply which methods. It usually depends on the data and the understanding of the data [15], [16]. Therefore, a system that can give visual explanations of the data and show how it connects with the algorithms can be helpful when looking to improve performance.

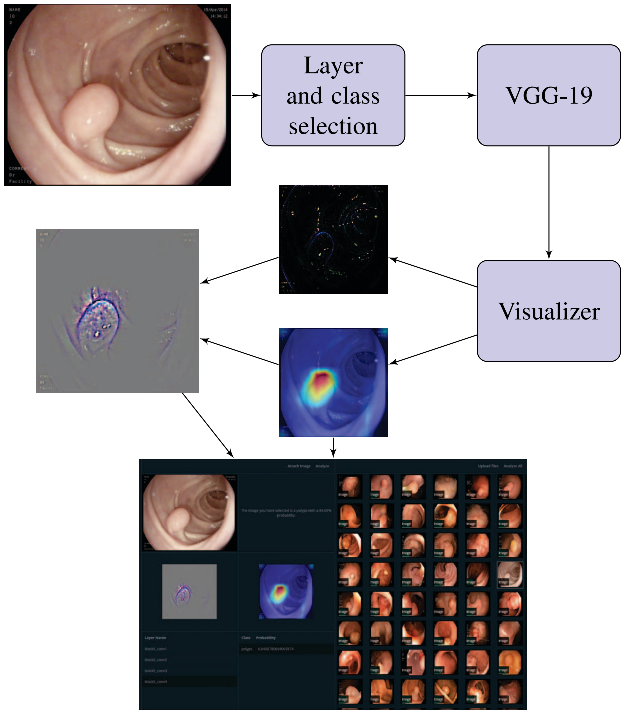


Fig. 1: An overview of how we produce the two visualisations included in the image analysis, and how it is presented in the user interface where a visualisation of the different convolutional layers can be selected.

Finally, our system tries to recommend which images or parts of the videos represent the most important findings. Medical experts indicate that generating automatic text is not the most important feature, but rather to give them a tool that helps them understand the decisions made by the algorithms, and at the same time, supports them in creating reports that represent the case in a unified way [1]. Therefore, bringing the user into the loop of system's performed analysis is an important aspect and requires tools such as the ones included in the presented system.

III. System description

The proposed system can be described as a framework consisting of two primary tools. First, to increase understanding and trust in the system, the system provides a tool for **dissecting the analysis of a CNN** by looking at how the network sees the given image at the point of any convolutional layer. This is done through various visualisation techniques, which present what regions of an image correspond to the distributed class confidence scores. This is mainly aimed at a medical audience needing additional information when making medical decisions, e.g., diagnosing a patient with a severe disease based on the output of the system. Researchers and engineers designing deep CNNs may use this tool to get more insight into what features of a given class lead to an assigned confidence score. This information may be useful in the development of pre-processing techniques which could result

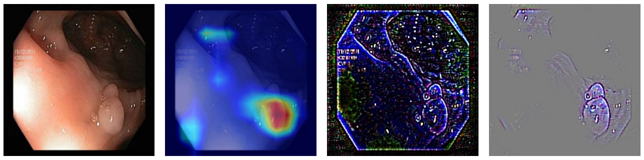

(a) Original image (b) Grad-CAM (c) Saliency Map (d) Guided Grad-CAM

Fig. 2: Image representations used by the reporting system to explain decisions.

in a higher quality model. Second, the system also provides a tool for **automatically generating medical reports** based on the automatic analysis of images and videos. The purpose of this tool is to reduce the time spent on administrative tasks that often follow medical examinations, e.g., documentation of performed endoscopy examinations. This can be seen in Figure 3, where we look at the expected workflow of a doctor using the system to gain information from the analysis done by the neural network and use this as a basis for generating the medical report.

As previously mentioned, the analysis performed by the system is based on deep learning technologies, specifically CNNs. These are used to analyse image or video data to perform different classification tasks, e.g., automatic detection of diseases. This process is made transparent to the users through the neural network dissection tool, which examines the individual layers of a CNN and allows for inspection of what regions of an image contribute to the score of a given class. This transparency is a critical piece in building trust and acceptability among non-technical users of the system, like medical experts, who rely on the system's output without detailed knowledge of the underlying processes. Furthermore, it allows for discovering faults within the trained model and the dataset used to train the system.

Visualisations are primarily based on the weighted gradient class activation map (grad-CAM) technique [17], which allows for visualisation of different CNN architectures without the need for modifications (replacement of layers). An overview of this process is shown in Figure 1, and starts once the user has selected an input image, target layer and target class using the web-interface. Based on these parameters, the system generates three different representations of the given image (all shown in Figure 4). Figure 2a is the input image which we use to generate the visualizations. Figure 2b is a grad-CAM (a generalization of class activation map (CAM) [18]) representation of the image, which shows what regions of the image correspond to the assigned confidence of the target class. Figure 2c is a saliency map generated using the guided back-propagation technique, which shows the positive activations of the target layer, which is not class specific. Figure 2d shows a guided grad-CAM representation of the image, which is a combination of the grad-CAM and saliency map. Of the three visualisations, the system presents the grad-CAM and the guided grad-CAM to the user.

IV. Implementation Details

The system is accessed through a web-interface, backed up by a RESTful server written in Python (using the micro-framework Flask [19]). As mentioned in section III, the server uses a deep neural network, specifically a CNN using the standard VGG-19 architecture [20], to perform frame analysis. It is important to point out that the architecture used can be changed if needed. The neural network is implemented using the deep learning framework Keras [21] using Tensorflow as a backend [22]. The visualisations are generated on the fly as the user selects an input image, target layer, and target class.

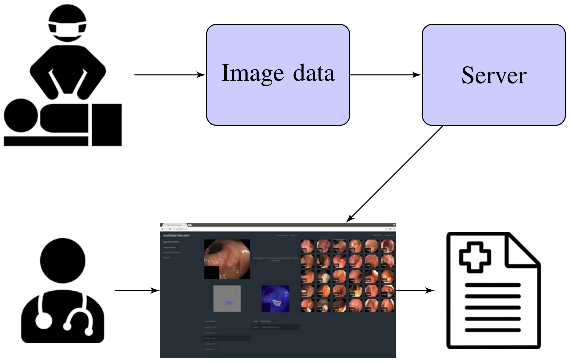


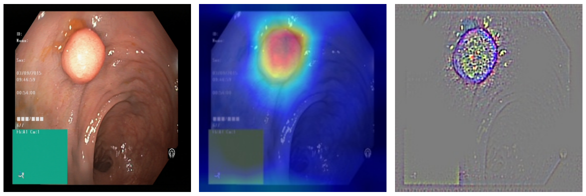
Fig. 3: The expected workflow using the system to analyse data from an endoscopic procedure and produce a written report.

V. Neural Network Dissection

As mentioned previously, we use a deep CNN to perform analysis on frames collected from an endoscopic examination taken from a video stream. One of the criteria we set in the previous section was that the system must back up its claims by showing the reasoning behind its decision. To achieve this, the system uses a guided grad-cam [17] approach to visualizing the convolutional layers of a CNN given a target class. Guided grad-cams combine the discriminative properties of CAMs together with a more detailed saliency map [23] to create high-quality feature maps, showing specific localization regions for a target class.

We use the guided grad-cam representation together with a grad-cam to give two perspectives on what the CNN is detecting when making its prediction, which in turn will hopefully distill a higher amount of confidence in the correctness of our network. Principally, the grad-cam and guided grad-cam show the same information, albeit the guided grad-CAM includes a bit more detail, we decided to include both as the grad-CAM is a bit clearer in its explanation. Visualisations are created on a layer by layer basis, making it possible to go back and view the detections made by the lower layers of the network. This might be useful to see what less abstract features are picked up by the network.

The visualisation process starts once the user has selected an image, layer, and class for further analysis. With this set, we begin with the creation of the grad-CAM and saliency visualisations. Starting with the grad-CAM, we calculate the



(a) Original (b) Grad-CAM (c) Grad-CAM

Fig. 4: An image that has been correctly identified as containing a polyp by our CNN, together with the grad-CAM and guided grad-CAM representation.

gradient of the target layer using the loss of the selected class in regards to the input image. These gradients are then globally average pooled to get the weights, which is then multiplied by the output of the target layer and passed through a ReLU function to produce the grad-CAM representation. The grad-CAM is then re-sized back to the dimensions of the original image and has its values squashed between 0 and 1 before applying a blue-red heat map.

To generate the guided back-propagation saliency map, we replace the activations of our original network with a modified ReLU function. During back-propagation, a conventional ReLU would let all gradients whose inputs were larger than 0 pass. We extend this rule by additionally discarding all gradients that are below 0, thereby only back-propagating the positive influence on the activations. With this modified network, we calculate the gradients of the target layer with respect to the input image, i.e., these gradients represent our saliency map.

With the grad-CAM and saliency map generated, we multiply them together to produce the guided grad-CAM representation. This together with the grad-CAM is used in our system.

VI. Understanding the Detections

As mentioned in section IV, we visualize the the convolutional layers of a VGG-19 CNN to understand what each layer detects as the image moves through the network. This is not only useful when trying to detect abnormalities in endoscopic images, but also gives insight into what features the network ”thinks” are relevant to a certain class. For example, Figure 4a shows an image containing a polyp located in its upper region. Looking at the grad-CAM representation (Figure 4b), we see the the network correctly identifies the polyp (area in red). Although the used example is quite obvious, this shows that the network, at least, has some knowledge about what a polyp is when it comes to its basic shape. Using the guided grad-CAM representation (Figure 4c), we get a more detailed view of what the network detects, such as texture detection. Looking closely at Figure 4c, we see that the network detects the edge of the polyp, noting that the polyp is raised above the mucosa (blue outline surrounding the polyp).

Figure 4 depicts the image at the last convolutional layer of the network, showing what the network recognizes right before making its prediction. For the most part, this is what we want. But, it may also be useful to look further back in the network to see what less abstract features are detected early in the network. Looking at Figure 5, we see a guided grad-CAM representation of an image at the last convolutional layer of each convolutional block of a VGG-19 CNN. Looking at the first couple of layers (Figures 5b and 5c), we see that the network picks up basic textures of the mucosa. Looking at the latter images, we see that the network starts to see the visual shape of the polyp.

(a) Kvasir v2

Kvasir v2	PREC	REC	SPEC	ACC	MCC	F1
Non-processed	0.966	0.791	0.736	0.940	0.762	0.758
Navigation box	0.968	0.798	0.753	0.944	0.778	0.778
Navigation box + border	0.968	0.943	0.749	0.943	0.775	0.771

(b) Kvasir v2 + CVC

Kvasir v2 Extra Polyps	PREC	REC	SPEC	ACC	MCC	F1
Non-processed	0.957	0.722	0.673	0.924	0.723	0.702
Navigation box	0.959	0.738	0.691	0.927	0.739	0.719
Navigation box + border	0.964	0.773	0.724	0.937	0.760	0.750

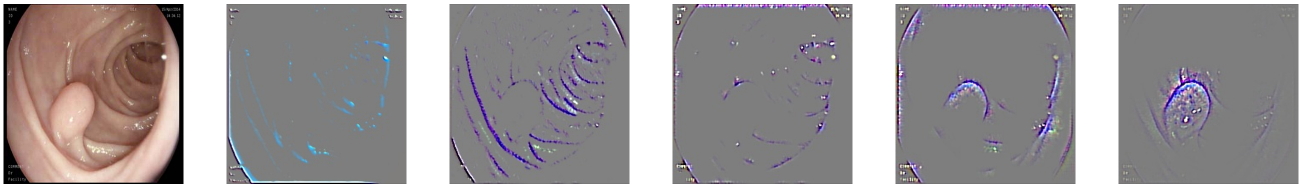
TABLE I: CNN evaluation using 2-fold cross-validation.

Note that the visualisations are made with respect to a target class, meaning we can see what regions of an image correspond to another class apart form the predicted one. This comes in handy when the network detects multiple classes in a single image. For example, an image may contain signs of ulcerative colitis and polyps, using the visualisations we are able to see the class specific regions of each abnormality. This is also useful when diagnosing issues with the network, understanding why a network ”thinks” it detects a certain class that is not there, this will be discussed further in section VII.

VII. Enhancing Input Data for Better Detection

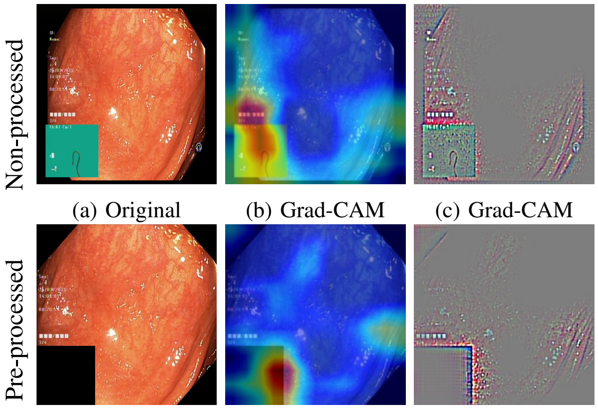
In the previous section, we used the two image representations (grad-CAM and guided grad-CAM) to gain insight into what the network sees when it predicts a certain class. This not only helps us detect diseases in the GI tract but can also be used to diagnose issues with a network making incorrect predictions. Using the Kvasir v2 dataset [24], we looked at various samples where the network got confused and mistakenly predicted the wrong class. Figure 6a shows an image of a clean cecum (beginning of the bowel), as part of the *normal cecum* class. The network mistakenly predicted that the image depicts a colon inflicted by ulcerative colitis (inflammatory bowel disease) as part of the *ulcerative colitis* class, with an 86.5% certainty. Using the system to diagnose what the network detects with respect to the class *normal cecum* at the final convolutional layer, we get the grad-CAM (Figure 6b) and guided grad-CAM (Figure 6c) representations, which show us that the algorithm gets confused by the navigation box located in the lower left corner. This indicates that the network has learned the ”noise” of an image, and associated it with a class, i.e., it has associated the navigation box with a normal cecum. This is an important observation, as we might be able to use this information to improve the performance of our network.

After finding incidents of incorrect predictions because of ”noise” in the image, we have two possible options for



(a) Original (b) Block 1 (c) Block 2 (d) Block 3 (e) Block 4 (f) Block 5

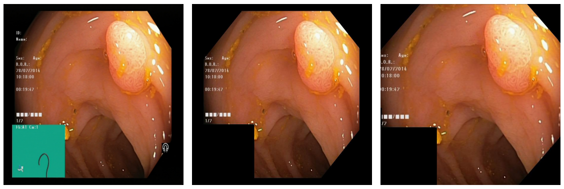
Fig. 5: Guided grad-CAM representation of an image at the last convolutional layer of each convolutional block.



(a) Original (b) Grad-CAM (c) Grad-CAM

(d) Original (e) Grad-CAM (f) Grad-CAM

Fig. 6: An incorrectly identified image with its grad-CAM and guided grad-CAM representation with respect to the class *normal cecum*.



(a) Non processed (b) Navigation box (c) Navigation box + borders

Fig. 7: Examples of data enhancements.

improving the class detection of our network, i.e., change the network itself or apply additional pre-processing steps to the dataset. For the scope of this paper, we will limit it to applying additional pre-processing steps to the Kvasir version 2 dataset as following: (i) blacked out navigation box and (ii) blacked out navigation box and cropped black borders.

After applying these steps, we retrained the model and reran the image analysis. This time we found that that *normal cecum* prediction had fallen to 28.3%, and the network now correctly classifies it as *ulcerative colitis* with a 53.22% certainty. The change in prediction is promising, but the network still activates on the blacked out navigation box, which causes the still high prediction value fro *ulcerative colitis*, i.e., additional pre-processing steps may lead to better results. In this particular case, a possible reason for the confusion is an imbalance of ”noisy” images between the classes, e.g., some classes include many images that have the navigation box located in the lower left corner, while other classes barely contain such images. This is supported by the class *ulcerative colitis*, having few images with a navigation box, often being confused with *polyps* and *normal cecum*, which contain many images with the navigation box.

Based on the findings of described in section VII, we trained and evaluated a VGG-19 CNN on three different variations of Kvasir v2; (i) non-processed images (Figure 7a), (ii) navigation box blacked out (Figure 7b), and (iii) navigation box and borders removed (Figure 7c). From this, we observe that the non-processed and pre-processed datasets are distinguishable, but the navigation box/borders removed and the blacked out dataset look quite similar. The difference between the two pre-processed datasets is in the surrounding border. Each dataset was trained and evaluated using 2-fold cross-validation resulting in 500 images used for training and evaluation per class. Table I(a) shows the result of the model evaluation. Looking at the F1 score, we see that the pre-processed datasets perform a couple of points better than the non-processed dataset.

As with any neural network, it is important that it generalizes well rather than obtaining good metrics by overfitting on a specific dataset. Therefore, we performed another evaluation on the three dataset variations using additional 400 polyp images taken randomly from the CVC-968 dataset [8] added to the test set. The reason, therefore, was to show how general the trained model is and that it does not work well on just the dataset used for training (which would be a sign for overfitting). The outcome of this experiment revealed that the non pre-processed dataset was less general than the pre-processed on and most probably dataset specific (overfitted). The result of this evaluation is shown in Table I and further supports the case that the pre-processed datasets perform better than the non-processed. Looking at the individual F1 scores we see that the non-processed dataset fell by 5.6 points, the blacked out dataset fell by 5.9 points, and the border and navigation box removed pre-processing fell by only 2.1 points. This shows that the border and navigation box removed pre-processing training creates a model that generalizes better than the other variations. Even if the general overall performance goes down compared to Table I(b) for a real-world scenario a more general model is more important than a dataset-specific one.

VIII. Creating Automatic Reports

As previously discussed, understanding and transparency of the underlying algorithms performing the analysis is a crucial piece in building trust and acceptance among mission-critical domains (such as medicine). Through our research, we see that this better understanding may be used to extract useful information in the production of medical reports, e.g., documentation of colonoscopies. Within GI endoscopy, documentation of performed procedures are generally considered to be of poor quality, often being submitted without the use of standardized language [25] and an inconsistent description of the detected endoscopic findings [26]. Typically, the time required to manually produce a standard GI endoscopy report is around 2 minutes. However, when having to include many findings, reports may take up to 10 minutes or more to produce [27], [28]. A system that can extract information from a deep neural network's layers can be used in the generation of automatic reports by providing images or video clips that reflect the suggested diagnosis which also provides a reason why the algorithm came to a particular decision. Nevertheless, this is out of focus for this paper and will be more in focus in future work including studies with medical experts.

IX. Conclusions

Neural networks are widely used in all types of detection, classification, and localization of objects in an image or video frame. However, the understanding of how deep neural networks operate and on what their output is based on is in general very limited – even more so among non-technical users. In many domains such as medicine (among others), the users often need to understand why a particular decision is made. To improve the understanding of the internal decision process of deep neural networks and to build trust among its users, we have developed a system that allows dissecting deep neural networks, enabling investigation and understanding of the network's layers and outputs. We presented a detailed explanation of how such a system can be used to increase understanding and performance and evaluated it using two different datasets. The evaluation is showing promising results indicating better performance and generalization of deep learning models after applying improvements based on insights gained using the presented system. Furthermore, we presented and discussed how the intermediate knowledge provided by the system could be used to automatically generate a modifiable report including both text and images increasing the understanding and potential trust of medical experts. For future work, we will evaluate the reporting part of the system with the help of medical doctors and improve the automatic report generation part based on this evaluation.

References

[1] M. Riegler, M. Lux, C. Griwodz, C. Spampinato, T. de Lange, S. L. Eskeland, K. Pogorelov, W. Tavanapong, P. T. Schmidt, C. Gurrin, D. Johansen, H. Johansen, and P. Halvorsen, "Multimedia and medicine: Teammates for better disease detection and survival," in *Proc. of ACM MM*, 2016, pp. 968–977.

[2] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland, "Computer aided disease detection system for gastrointestinal examinations," in *Proc. of MMSys*, 2016.

[3] M. Riegler, K. Pogorelov, P. Halvorsen, T. de Lange, C. Griwodz, P. T. Schmidt, S. L. Eskeland, and D. Johansen, "EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies," in *Proc. of CBMI*, 2016.

[4] K. Pogorelov, M. Riegler, P. Halvorsen, P. T. Schmidt, C. Griwodz, D. Johansen, S. L. Eskeland, and T. de Lange, "GPU-accelerated real-time gastrointestinal diseases detection," in *Proc. of CBMS*, 2016.

[5] K. Pogorelov, M. Riegler, S. L. Eskeland, T. de Lange, D. Johansen, C. Griwodz, P. T. Schmidt, and P. Halvorsen, "Efficient disease detection in gastrointestinal videos – global features versus neural networks," *Multimedia Tools and Applications*, vol. 76, no. 21, 2017.

[6] M. Riegler, K. Pogorelov, J. Markussen, M. Lux, H. K. Stensland, T. de Lange, C. Griwodz, P. Halvorsen, D. Johansen, P. T. Schmidt, and S. L. Eskeland, "Computer aided disease detection system for gastrointestinal examinations," in *Proc. of MMSYS*, 2016, pp. 29:1–29:4.

[7] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of ACM MMSYS*, 2017, pp. 164–169.

[8] J. Bernal and H. Aymeric, "Miccai endoscopic vision challenge polyp detection and segmentation," https://endovissub2017-giana.grand-challenge.org/home/, accessed: 2017-12-11.

[9] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Proc.*, 2017.

[10] R. Vidal, J. Bruna, R. Giryes, and S. Soatto, "Mathematics of deep learning," *arXiv preprint arXiv:1712.04741*, 2017.

[11] C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, and S. Gumhold, "Visualizations of deep neural networks in computer vision: A survey," in *Transparent Data Mining for Big and Small Data*. Springer, 2017, pp. 123–144.

[12] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[13] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *Proc. of IEEE CVPR*, 2017, pp. 6428–6436.

[14] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[17] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, 2016.

[18] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, 2015.

[19] "Flask." [Online]. Available: http://flask.pocoo.org/

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[21] "Keras." [Online]. Available: https://keras.io/

[22] "Tensorflow." [Online]. Available: https://www.tensorflow.org/

[23] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014. [Online]. Available: http://arxiv.org/abs/1412.6806

[24] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proc. of MMSYS*, 2017, pp. 164–169.

[25] L. Aabakken, A. N. Barkun, P. B. Cotton, E. Fedorov, M. A. Fujino, E. Ivanova, S.-e. Kudo, K. Kuznetzov, T. Lange, K. Matsuda *et al.*, "Standardized endoscopic reporting," *Journal of gastroenterology and hepatology*, vol. 29, no. 2, pp. 234–240, 2014.

[26] R. S. Sharma and P. G. Rossos, "A Review on the Quality of Colonoscopy Reporting," *Canadian Journal of Gastroenterology and Hepatology*, vol. 2016, no. i, pp. 1–6, 2016. [Online]. Available: 2016Sharma http://www.hindawi.com/journals/cjgh/2016/9423142/

[27] M. Groenen, E. Kuipers, G. van Berge Henegouwen, P. Fockens, and R. Ouwendijk, "Computerisation of endoscopy reports using standard reports and text blocks," *The Netherlands journal of medicine*, 2006.

[28] K. Kuhn, W. Gaus, J. Wechsler, P. Janowitz, J. Tudyka, W. Kratzer, W. Swobodnik, and H. Ditschuneit, "Structured reporting of medical findings: evaluation of a system in gastroenterology," *Methods of information in medicine*, vol. 31, no. 04, pp. 268–274, 1992.