Analysis of Possible Influences on Class Activation Based Explainability Methods for Image Classifiers

Student:	B. Sc. Mysianov Roman
Course of Studies:	Research in Computer and System Engineering
Mail:	roman.mysianov@tu-ilmenau.de
Matrikel Number:	63540
Supervisor:	M. Sc. Daniel Scheliga
Chair:	Prof. Dr.-Ing. Patrick Mäder (JP)
Faculty:	Computer Science and Automation
Department:	Software Engineering for Safety-Critical Systems
Continuance:	01.04.2021 - 31.09.2021

Keywords: Explainable Artificial Intelligence, Image Classification, Deep Learning; CNN; Activation Mapping; Guided Grad-CAM;

Motivation

Convolutional Neural Networks showed their advances at extracting features from images and performing tasks like Image Classification, Image Segmentation, Object Detection, Image Captioning and etc. While these models enable superior performance, their lack of decomposability into individually intuitive components makes them hard to interpret [1]. There typically exists a trade-off between accuracy and interpretability. Classical rule-based models such as Random Forest, Decision Trees, Linear Regression etc. are highly interpretable but not very accurate. By using deep models, we sacrifice interpretable modules for uninterpretable ones that achieve greater performance through greater abstraction. As image classification for critical decisions is gaining ground, explainability is also becoming important in that context. We can think of applications such as medical image diagnosis or damage assessment in insurance, for which patients or consumers obviously demand an explanation. Explainability becomes even more important when severe misclassifications occur. Companies should be able to explain what went wrong, at least to prevent this from happening in the future. Recent researches proposed few methods such as class activation mapping (CAM) that suit the explainability problem, but there is no information about what affects the explainability performance, how do different factors make the model more or less explainable. In this thesis, we are analyzing possible influences on class activation based explainability methods for image classifiers.

Task Description

This research is dedicated to the investigation of what factors influence the explainability of ConvNet based image classifiers. Methods based on the activation mapping approach should be described and compared, one of them has to be implemented. These methods produce visualizations intended to show which inputs a neural network is using to make a particular prediction. ResNet [2] will be used as a base model for binary classification. The goal of the work is to see how different parameters like regularization (Dropout, Batch Normalization and Weight Decay), network depth, training progress, model performance, etc. have an influence on

how good the explainability of the model is. The given dataset contains approximately 7000 colored eye images. The study of the explainability of image classifiers problem leads to literature research, familiarization with the state of the art and dataset investigation. Moreover, there should be a theoretical understanding of the approached method, which will be implemented for experiments with binary classification. Most earlier research was confined to fully connected networks, toy problems, or toy datasets and will therefore be expanded to a real-life medical problem using a State of the Art CNN (ResNet).

Research Questions

The goal of this research is to answer to the following questions:

- Which of the different class activation mapping approaches performs best?
- What are suitable metrics for measuring the explainability?
- How do regularization techniques influence explainability performance?
- What is an optimal network depth?
- How does hyperparameters tuning affect the quality of explainability?
- How does CNN performance correlate with explainability results?

Approach

Methods

In this work we want to train a binary image classifier on eye disease datasets. Multiple explanation methods for image classification have been proposed in the literature. For example, class activation mapping, evidence counterfactual and technique based on mutual information (MI). In this work, we will look at few approaches based on CAM, such as gradient class activation mapping (Grad-CAM), Grad-CAM++ and Score-CAM. CAM is a technique for identifying discriminative regions by linearly weighted combination of activation maps of the last convolutional layer before the global pooling layer of a CNN. To aggregate over multiple channels, CAM incorporates the importance of each channel with the corresponding weight at the following fully connected layer, which generates a score as the class confidence. The biggest restriction of CAM is that not every model is designed with a global pooling layer and even a global pooling layer is present, sometimes more fully connected layers follow before the softmax function, e.g. VGG. CAM has two distinct drawbacks: Firstly, in order to be applied, it requires that neural networks have a very specific structure in their final layers and, for all other networks, the structure needs to be changed and the network needs to be re-trained under the new architecture. Secondly, the method, being constrained to only visualising the final convolutional layers of a CNN, is only useful when it comes to interpreting the very last stages of the network's image classification and it is unable to provide any insight into the previous stages.
Grad-CAM is a strict generalization of CAM that can produce visual explanations for any CNN, regardless of its architecture, thus overcoming one of the limitations of CAM. Grad-CAM works by finding the final convolutional layer in the network and then examining the gradient information flowing into that layer [3]. The output of Grad-CAM is a heat-map visualization for a given class label (either the top, predicted label or an arbitrary label we select for debugging). Using Grad-CAM, we can visually validate where our network is looking, verifying that it is indeed looking at the correct patterns in the image and activating around those patterns. [4]. While an improvement over CAM, Grad-CAM has its own limitations, the most notable including its inability to localize multiple occurrences of an object in an image, due its partial derivative assumptions, its inability to accurately determine class-regions coverage in an image, and the possible loss in signal due the continual upsampling and downsampling processes [5]. Grad-CAM++ method can provide better visual explanations of CNN model predictions, in terms

of better object localization as well as explaining occurrences of multiple object instances in a single image, when compared to State of the Art. Grad-CAM++ uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps with respect to a specific class score as weights to generate a visual explanation for the corresponding class label [6].
Score-CAM is based on class activation mapping. Unlike previous class activation mapping based approaches, Score-CAM gets rid of the dependence on gradients by obtaining the weight of each activation map through its forward passing score on the target class, the final result is obtained by a linear combination of weights and activation maps.
After the comparison of the State of the Art methods, one of the methods will be applied to the model, based on ResNet, which will be trained on the given dataset. The explainability metric plays an important role in process evaluation. The experiments with different factors (regularization, hyperparameters, network depth) that might influence explainability will be conducted.

Metrics

There are a few metrics for binary classification, which will be used to estimate the quality of binary classifier [7]. But First we need to define some basic concepts used by the metrics:

- True Positive (TP): A correct detection.
- False Positive (FP): A wrong detection.
- False Negative (FN): A ground truth not detected
- True Negative (TN): A correct detection of the negative class.

Receiver Operating Characteristic (ROC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. ROC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example [8]. It shows the relation between True Positive Rate and False Positive Rate for a given classifier.

- True Positive Rate (Recall or Sensitivity) is defined as

$$Recall = \frac{TP}{TP + FN}$$

  True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

- False Positive Rate is defined as

$$FalsePositiveRate = \frac{FP}{FP + TN}$$

  False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

For explainability methods we will use Intersection Over Union (IOU) [9].
IOU is a measure based on Jaccard Index that evaluates the overlap between two bounding boxes. It requires a ground truth bounding box B_gt and a predicted bounding box B_p. In our context ground truth bounding box is related to a mask, where glaucoma is located and predicted bounding box is the heat map of explainability method. By applying the IOU we can tell if a detection is valid (True Positive) or not (False Positive). An IOU score > 0.5 is normally considered a "good" prediction. For this task the threshold will be 0.75, which is an optimal value for object detection [10].

$$IOU = \frac{area(B_p \bigcap B_gt)}{area(B_p \bigcup B_gt}$$

The next metric is Faithfulness, which tries to evaluate if the feature importances returned by an explainability method are the correct ones. The method starts by replacing the most important feature value by another value termed as the base value which the user supplies as being a no information or no-op value. Once this replacement has been done the new example is passed to the classifier and its prediction probability for the original predicted class is noted [11].

Datasets

Datasets are already preprocessed and split into train, test and validation sets. These datasets contain two class of images health eye and glaucoma and masks for evaluating explainability approaches. As can be seen from the table below:

Type	Train	Test	Validation
Gluacoma	**2372**	**661**	**66**
Health	**2421**	**754**	**93**

Table 1: Datasets

Each input sample is a coloured image depicting an eye scan. As can be seen below:

Figure 1: Example of glaucoma

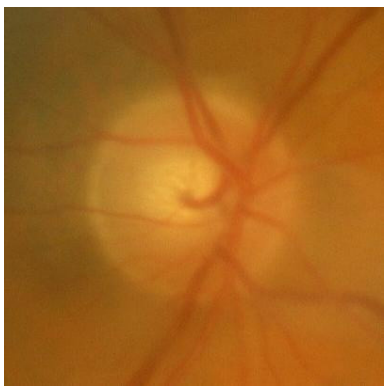

Figure 2: Example of health eye

Time Schedule

Week	Tasks
1-3	Literature research, familiarization with the State of the Art, dataset investigation
4-6	Theoretical understanding of approached methods, writing the introduction
7-11	Binary Classifier implementation, writing the theoretical part of the thesis
12-15	Methods implementation
16-18	Formulating the concepts experiments, running and evaluating.
19-21	Classifier analysis using methods mentioned above.
22-23	Preparation of the results, writing the conclusion
24-25	Making the presentation

Bibliography

[1] D. Chinesh, "Generalized way of interpreting cnns using guided gradient class activation maps!!," 2020.

[2] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, *Deep Residual Learning for Image Recognition.* cornell university, 2015.

[3] D. Mishra, "Demystifying convolutional neural networks using gradcam," 2019.

[4] M. Chetoui, "Gradient-weighted class activation mapping - grad-cam-," 2019.

[5] L. Pantelis, P. Vasilis, and K. Sotiris, *Explainable AI: A Review of Machine Learning Interpretability Methods.* MDPI, 2020.

[6] C. Aditya, S. Anirban, and H. Prantik, *Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks.* International Journal of Computer Vision, 2018.

[7] J. Czakon, "24 evaluation metrics for binary classification (and when to use them)," 2021.

[8] D. Aditya, "Metrics to evaluate your machine learning algorithm," 2018.

[9] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, 2021.

[10] W. Minghu, Y. Hanhui, J. Yuhan, K. Cong, and C. Zeng, "Object detection based on rgc mask r-cnn," 2019.

[11] A. Vijay, K. Rachel, K. Bellamy, and C. Pin-Yu, *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.* IBM, 2019.