Explainable Artificial Intelligence (XAI)

Francisco José Palmero Moya

Computer Vision | M.Sc. in Research in AI @ UNED

March 30, 2023

Abstract

Explainable Artificial Intelligence (XAI) is an emerging field that aims to improve the transparency and interpretability of machine learning models. In recent years, XAI has gained increasing attention in the Computer Vision field, where the analysis and interpretation of visual data using machine learning models is of paramount importance. This document provides an overview of the state of the art in XAI, with a particular focus on its applications and methodologies in the Computer Vision field along with some python libraries.

1 Introduction

Artificial Intelligence (AI) has made significant progress in recent years, revolutionizing many fields such as healthcare, finance, and transportation. However, the increasing complexity and opacity of AI models have led to concerns about their trustworthiness and accountability. Explainable Artificial Intelligence (XAI) is an emerging field that aims to address these concerns by improving the transparency and interpretability of AI models.

XAI is particularly important in fields such as healthcare, where AI models are used to make critical decisions about patient diagnosis and treatment. In these domains, the ability to understand and interpret the decision-making process of AI models is of paramount importance for ensuring patient safety and trust.

XAI is also becoming increasingly important in the Computer Vision field, where the analysis and interpretation of visual data using machine learning models is a critical task. Computer Vision has numerous applications, such as medical imaging, our reseacrh topic in the project.

This document provides an overview of the state of the art in XAI, with a particular focus on its applications and methodologies in the Computer Vision field. We review the different approaches for generating explanations in Computer Vision, which will be an starting point for the next steps in our project.

Scope This paper about the state of the art focuses on the following key areas of XAI research:

1. Approaches and methodologies for generating explanations from machine learning models.
2. Evaluation metrics for assessing the quality and effectiveness of XAI methods.
3. Research hypotheses for Breast ultrasound images classification problem.

2 Literature review

This section consist on gather and analyze relevant literature from mainly two different papers: Gulum et al. (2021) and Hoffman et al. (2019). Identify the key research gaps, challenges, and trends in the field of XAI with a strong focus on Explainable Deep Learning Cancer Detection Models in Medical Imagin since Breast Cancer is the project's primary objective.

2.1 Approaches and Methodologies for Generating Explanations

There are several approaches and methodologies for generating explanations from machine learning models, nevertheless we focus solely on analyzing medical images for cancer detection. This subsection will group explanation methods following the taxonomy suggested in Gulum et al. (2021).

2.1.1 Local vs. Global

Local Local explanation refers to providing explanations for individual samples. In health care, this approach is useful when explanations for individual partients are of interest.

Global Global explanation refers to providing explanation for a group of samples or the entire model. This shows the overall feature importance for a group of patients in health care.

2.1.2 Data Modality Specific vs. Data Modality Agnostic

Data Modality Specific Data Modality Specific refers to explanation methods that are only applicable for a certain data type. For example, some methods only work with images and some methods only work for textual or tabular data.

Data Modality Agnostic Data Modality Agnostic refers to explanation methods that work for any data type. These methods commonly use surrogate or perturbation based approaches to create a general approach for model explanation.

2.1.3 Ad-Hoc vs. Post-Hoc

Ad-Hoc For ad-hoc explanation methods, the model itself is designed to be instrinsically explainable. The goal of this approach is to design a deep learning model that is inherently explainable and opposes the notion that there is a trade off between accuracy and explainability.

Post-Hoc On the contrary, post-hoc explanation techniques provide explanations after the classification is made. Some refer to these as diagnostic method due to their utility for diagnosing and their apparent limitations for providing a complete explanation for the end user.

2.1.4 Model Agnostic vs. Model Specific

Model Agnostic Model Agnostic refers to explanation methods that are able to explain any model and are not restrained to a certain type. A common approach is to change the inputs and measure the corresponding change in output. Then to use this to determine what change in inputs produces the greatest change in outputs.

Model Specific Model Specific explanation methods only work with a specific model. These methods often use certain aspects of model architecture, for example features maps produced from graph convolutions.

2.1.5 Attribution vs. Non-Attribution

Attribution methods attempt to calculate the inputs of the neural network that are the most important with regards to the network's result. This can be broken into two categories:

Perturbation based Perturbation based approaches estimate the most important features by removing one, calculating change in class score, and repeating for all features. These are usually inefficient due to having to compute may iterations.

Back-propagation based Gradient-based methods conduct a similar procedure except with a single forward or backwards pass. In general these methods produce results faster due to having to perform only one pass.

2.2 Measures for Explanations

Measuring the quality of explanation is challenging because of the importance of the context the method is used for along with the non-triviality of defining what a good explanation is. This is specially important for cancer explanation methods due to high-risk nature of the predictions.

2.2.1 What defines a hihgt-quality explanation

There are studies that attempt to explore the question of what defines a hight-quality explanation (Hoffman et al., 2019; Doshi-Velez and Kim, 2017; Tomsett et al., 2018). The goals of explanation involve answering questions such as, "How does it work?" and "What mistakes can it make?" and "Why did it just do that?".

Several researchers have argued that a machine learning systems interpretability should be defined in relation to a specific agent or task: we should not ask if the system is interpretable, but to whom is it interpretable. On the other hand, looking across the literature we find assertions about what makes for a good explanation, from the standpoint of statements as explanations. There is a general consensus on this, they should be clear and concise. Thus, one can

look at a given explanation and make an a priori or descontextualized judgement as to whether or not it is good. Hoffman et al. (2019) suggest a Goodness Checklist that can be used by XAI researchers to either try and design goodness into the explanation that their XAI system generates, or to evaluate the a priori goodness of the explanations that an XAI system generates.

Going back to the question to whom is it interpretable, even if an explanation may be regarded good in the manner stated above, users-in-context may not find it to be adequate or satisfying. Tomsett et al. (2018) discusses the different roles involving explanation techniques and how the explanations differs depending on what role you are providing the explanation for. They show that an explanation for one role should be measured differently for other roles.
Explanation Satisfaction is defined as the degree to which users feel that they undestand the AI system or process being explained to them. Compared to Goodness, as we defined it above, satisfaction is a contextualized, a posteriori judgement of explanations.

2.3 Critical Analysis

While XAI has made significant progress in recent years, there are several challenges and limitations that need to be addressed. One key challenge is the trade-off between accuracy and interpretability, as some XAI methods may sacrifice accuracy for the sake of interpretability.

2.4 Conclusions

There is a need to incorporate users in the design process of the algorithm and a need to build intrinsically explainable deep learning algorithm specially designed for each task. The majority of explanation techniques are designedd for general use and do not incorporate context. This is important for clinical use cases due to unique, high-risk environment. Some promising future directions include quantifying the explanation uncertainty, providing counter examples, and designing ad-hoc models that are intrinsically explainable.

3 XAI methods

We focus now in three widely used method for Computer Vision algorithm: LIME, SHAP and Grad-CAM. We aim to give a brief introduction to each method and to outline the limitations we might have when using them.

3.1 Local Interpretable Model-agnostic Explanations (LIME)

The paper from Ribeiro et al. (2016) is a well-known and influential work in the field of XAI. LIME (Local Interpretable Model-Agnostic Explanations) is a method proposed in the paper to explain the predictions of any classifier by approximating the classifier locally with a linear model.

The key idea is to generate a set of perturbed instances around the instance of interest, and fit a linear model to explain the classifier's predictions on those instances. The linear model provides local, interpretable explanations for the classifier's predictions. LIME is model-agnostic and data modality agnostic, meaning it can be applied to any type of classifier, and can handle different types of data, such as images, text, and tabular data as we saw before.

As shown by Alvarez-Melis and Jaakkola (2018), LIME has some limitations such as the possibility of producing explanations of two very close points that vary greatly in a simulated setting.

3.2 SHapley Additive exPlanations (SHAP)

SHAP (Lundberg and Lee, 2017) is a unified approach for interpreting model predictions by assigning feature importance scores to each input feature. The method is based on Shapley values from cooperative game theory, which provides a theoretically ground and consistent way of attributing contributions to multiple agents. SHAP is model-agnostic and can be applied to any type of machine learning model, including deep neural networks for image classification.

One strength of SHAP is its ability to provide unified and consistent explanations for any type of model, which can improve interpretability and transparency in AI systems. However, SHAP can be computationally expensive and requires sampling many possible feature combinations, which can limit its scalability in some cases.

3.3 Gradient-weighted Class Activation Mapping (Grad-CAM)

The Grad-CAM (Gradient-weighted Class Activation Mapping) method is a technique for generating visual explanations of the predictions made by deep neural networks. This method was proposed in the paper Selvaraju et al. (2019). It is a data modality specific method that only work for images.

The main idea behind Grad-CAM is to use the gradients of the output class with respect to the feature maps of the last convolutional layer in a Convolutional Neural Network to generate a coarse localization map highlighting the regions of the image that were most relevant for the prediction. In other words, Grad-CAM identifies the regions of the input image that were most important in driving the network's decision.

The resulting visualization shows the regions of the input image that contributed most to the network's prediction. This makes it possible to interpret the decision-making process of the network and to identify potential biases or errors.

In the context of medical imaging applications, Grad-CAM can be used to identify the regions of an image that were most relevant for a diagnosis, providing valuable insights into the decision-making process of the network and increasing its transparency and interpretability.

Considerations Compared to Grad-CAM, LIME and SHAP are more focused on generating sparse, interpretable models to explain individual predictions, whereas Grad-CAM generates visual explanations that highlight the most relevant regions of the input image for a particular prediction. Since all three methods have their strengths and weaknesses, they can be used together to provide a more complete picture of the decision-making process.

4 Methodology

To investigate the state of the art of XAI in medical imaging applications, we will use the OmniXAI library in Python to solve a breast cancer image classification problem.

OmniXAI is a powerful and user-friendly library for generating visual explanations of deep neural networks. It provides a range of XAI techniques, including Grad-CAM, LIME, and SHAP, that can be used to increase the transparency and interpretability of deep learning models.

For the breast cancer image classification problem, we will use a dataset of breast ultrasound images (Al-Dhabyani et al., 2020) labeled as either normal, or malignant or benign. As starting point, we will use a method for image segmentation and classification on this dataset using a method based on Vakanski et al. (2020). We will then use OmniXAI to generate visual explanations of the predictions made by the model, in order to identify the regions of the image that were most relevant for the classification decision.

The research hypothesis of this project is to perform a comparative analysis of several existing XAI techniques applied to a specific case for the detection of breast cancer. This analysis could serve as a first step in a larger project where improvements will be proposed based on the defects found in the current techniques. Additionally, the results of this project will provide valuable insights into the strengths and weaknesses of different XAI techniques, and may help guide the development of new, more effective approaches in the future.

Based on the research hypothesis of this document, the LIME, SHAP, and Grad-CAM libraries will be utilized to cover a wide range of XAI methodologies in the analysis of breast cancer detection. These libraries are well-known and widely used in the field of XAI, and offer a variety of techniques for generating model explanations. By using these libraries, this project aims to provide a comprehensive analysis of XAI techniques for breast cancer detection, and to identify the strengths and weaknesses of each method. The ultimate goal is to improve the accuracy and reliability of XAI techniques for this critical task, and to provide a solid foundation for future research in this area.

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28:104863.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Gulum, M. A., Trombley, C. M., and Kantardzic, M. (2021). A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10).

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable ai: Challenges and prospects.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems.

Vakanski, A., Xian, M., and Freer, P. E. (2020). Attention-enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound in Medicine and Biology*, 46(10):2819–2833.

M2: XAI Implementation

Francisco J. Palmero Moya

September 15, 2023

Abstract

The present report implement and apply the diffent XAI methods proposed in M1 for the breast cancer image classification problem. As starting point, we will use a method for image classification on this dataset, and then use OmniXAI to generate visual explanations of the predictions made by the model, in order to identify the regions of the image that were most relevant for the classification decision.

1 Introduction

The central aim of this report is to conduct a comparative evaluation of various existing XAI (Explainable Artificial Intelligence) methods when applied to a particular scenario involving the detection of breast cancer. This analysis represents an initial phase within a broader project, which will subsequently aim to introduce enhancements based on identified shortcomings in the current techniques. Moreover, the outcomes of this project are expected to yield valuable insights into the pros and cons of various XAI techniques, potentially offering guidance for the formulation of novel, more efficacious approaches in the future. Based on the research hypothesis of this document, the LIME, SHAP, and Grad-CAM libraries will be utilized to cover a wide range of XAI methodologies in the analysis of breast cancer detection. These libraries are well-known and widely used in the field of XAI, and offer a variety of techniques for generating model explanations.

2 Classification problem

The problem consist on estimate the risk of breast cancer given some images collected from breast ultrasound images. The dataset employed[1] contains breast ultrasound images among women in ages between 25 and 75 years old. The number of patients is 600 female patients. The dataset consists of 780 images with an average image size of 500 × 500 pixels. The images are categorized into three classes, which are normal, benign, and malignant. Figure 1 includes some samples from Breast Ultrasound Images Dataset corresponding to diffent classes.

Our approach consist on defining a model able to classify the images with acceptable accuracy. The process of designing a classification model and implement it is a hard task out of the escope of this document. Along the document, we mainly focus on the XAI techniques that where proposed in M1. The classification model is just an starting point required to compare the different XAI techniques.

Topology

A Convolutional Neural Network (CNN) has been applied. It consists of a series of layers: three Conv2D layers with 32, 64, and 128 filters respectively, each followed by a MaxPooling2D layer to reduce spatial dimensions. The Conv2D layers use the ReLU activation function. After the convolutional layers, the network flattens the output and passes it through two dense (fully connected) layers with 128 units and ReLU activation. Finally, there is an output layer with 3 units and a softmax activation function.

Preprocessing

The dataset is divided into training and evaluation subsets with a proportion of 80% to 20%, respectively. The images have been resized to 100 × 100 pixels if they are not already in that size.

Training

The model was trained during 30 epochs with a batch size of 16. The optimazer was Adam, while the loss function was categorical cross-entropy. Finally, the evaluation metric was the accuracy. The learning curves are shown in Figure 2.

[1] The Breast Ultrasound Images Dataset can be found in Kaggle

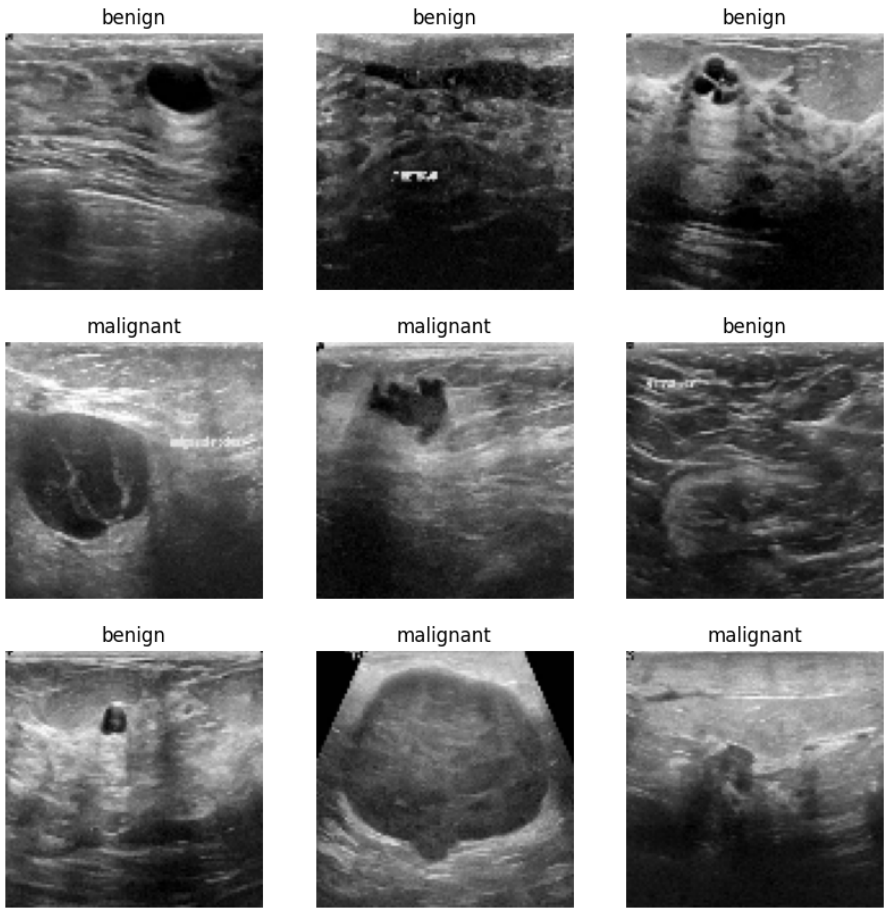


Figure 1: Samples from Breast Ultrasound Images Dataset

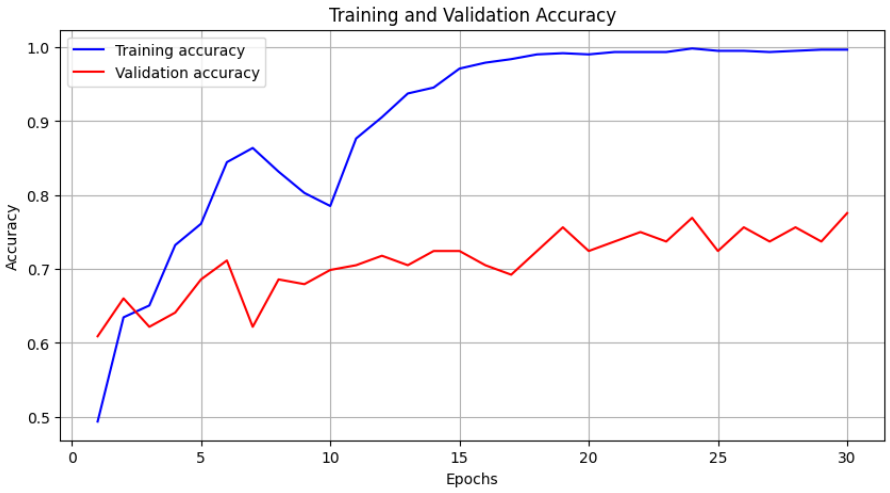


Figure 2: Learning curve classification model

3 XAI methods

In this section we delve into the XAI methods proposed in M1, namely SHAP, Grad-CAM, and LIME. This section is only devoted to show the individual results of each technique, whereas a comparison of them is presented in the next section. The software used for their implementation was OmniXAI (short for Omni eXplainable AI). It is a Python-based machine learning library designed for Explainable AI (XAI), providing comprehensive capabilities for explaining AI and making machine learning models more interpretable. The code employed is available at `main.ipynb` notebook that includes both: training of CNN and XAI methods.

3.1 SHAP

SHAP (SHapley Additive exPlanations) for images works by computing Shapley values, which are a mathematical concept from cooperative game theory, to determine the contribution of each pixel in an image to the model's prediction. The steps are:

Baseline Images Choose a baseline or reference image that represents the absence of any specific feature or information (often a black image or a neutral gray image).

Permutation Create a set of permutations of the image, where each permutation is a combination of the original image and the baseline image. This represents different degrees of presence or absence of features.

Predictions For each permutation, feed it through our CNN model and record the model's predictions. This provides a range of predictions that the model can make when different pixel values are present or absent.

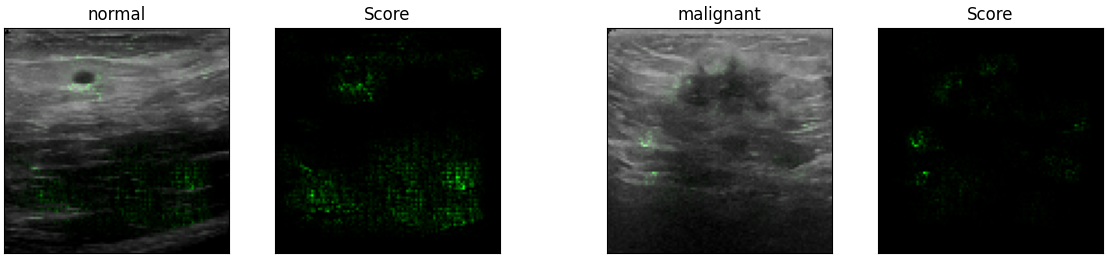
Shapley Values Calculate the Shapley values for each pixel by quantifying how much each pixel's presence or absence contributes to the change in model predictions compared to the baseline.

Map These Shapley values can be visualized as an "explanation map" where each pixel's value represents its contribution to the final prediction. Positive values indicate a positive influence on the prediction, while negative values indicate a negative influence.

The syntax to generate the explanations is quite simple in `omnixai` library.

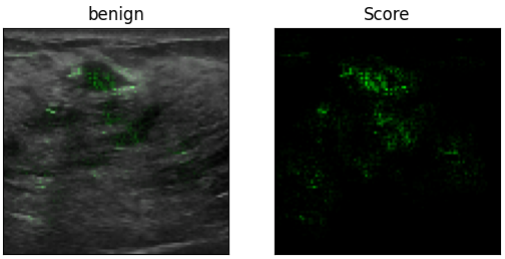
```
# Import library
from omnixai.explainers.vision import ShapImage

# Create explainer
explainer = ShapImage(model=model, preprocess_function=None)

# Generate explanations
explanations = explainer.explain(val_imgs)
```

where `model` is our pre-trained CNN for the classification problem, and `val_imgs` are images belonging to one batch in the evaluation subset. The explanations generated by SHAP are shown in Figure 3.



(a) Normal

(b) Malignant



(c) Benign

Figure 3: One SHAP explanations for each class

The images correspond to three different classes: normal, malignant, and benign. The example 3a classified as normal shows a bigger score in the dark part of the image. The dark part corresponding to an spot in the top left corner also indicates a high score. If we look closer to the image, we can appreciate the boundary of the spot in the score image. The example 3b correspond to a malignant example. The image shows a dark cloud in the middle, surrounded by a brighter zone. The score seems to be slighly higher around the boundary of the dark cloud. Finally, example 3c corresponds to a benign example. There is a higher score close to the dark spot located at the top middel.

3.2 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) works by utilizing the gradients of the target class score with respect to the feature maps of a CNN to generate a class activation map that highlights the regions in an image that are most important for the model's prediction. The steps are:

Forward pass Pass the input image through our CNN to compute feature maps at different layers of the network.

Class score Calculate the class score or probability for the target class of interest in the final layer of the network.

Gradients Compute the gradient of the target class score with respect to the feature maps at the final convolutional layer. This gradient represents how changes in each feature map affect the class score.

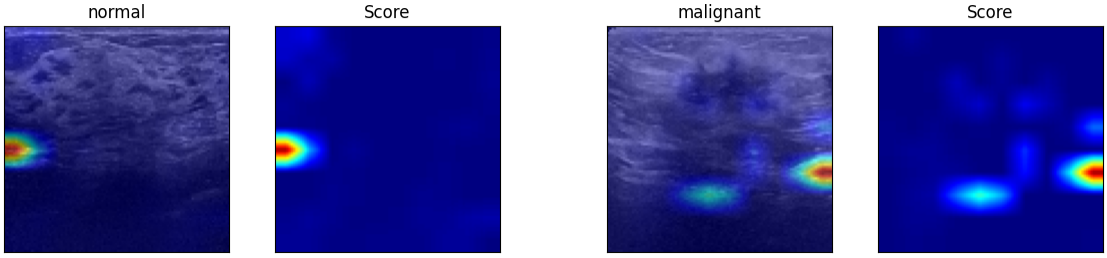
Global Average Pooling Perform global average pooling on the gradients. This step aggregates the gradient information across all spatial locations within each feature map.

Weighted Sum Multiply each feature map by its corresponding gradient-weighted importance score (from global average pooling) and sum these weighted feature maps. This step produces the class activation map, where each pixel value corresponds to the importance of that region in the input image for the target class.

The code employed was

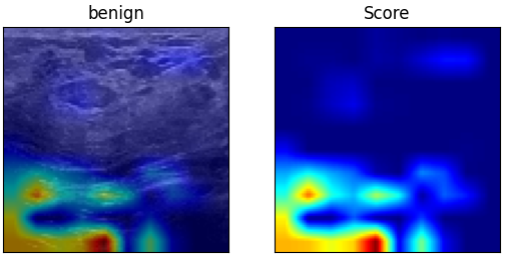
```
# Import library
from omnixai.explainers.vision.specific.gradcam import GradCAM

# Create explainer
explainer = GradCAM(
    model=model,
    target_layer=model.layers[-4],
    preprocess_function=None
)

# Generate explanations
explanations = explainer.explain(val_imgs)
```

where `model` is our pre-trained CNN and we have to manually select the layer. In our case we chose the last convolutional layer, where in theory, the heat-map for this layer should display the most accurate visual explanation.



(a) Normal last conv. layer

(b) Malignant last conv. layer



(c) Benign last conv. layer

Figure 4: One Grad-CAM explanations for each class

The heat-maps show that very specific regions of the image have an high relevance. The example 4a has a uniform weights distribution along the image except for an spot in the middel left. It corresponds to a dark spot in the image. Example 4b follows a similar behaviour. Dark zones in the image surrounded by brighter zones present high weights values. Finally, example 4c shows hight weigths values at the bottom left corner where the image is darker. The same images for different layers are mostly uninformative but they can be seen in the notebook.

3.3 LIME

LIME (Local Interpretable Model-agnostic Explanations) works by creating locally and interpretable explanations for our CNN predictions by perturbing the input data and observing how the predictions change, then fitting a simple, interpretable model to the perturbed data. The steps are

Select an instance Choose a specific instance for which we want to explain its prediction.

Data perturbation Perturb the features of the selected instance by randomly sampling and modifying them while keeping the label fixed. This creates a dataset of perturbed instances.

Predictions For each perturbed instance, obtain predictions from the black-box model.

Weights Calculate the similarity (distance or similarity score) between the original instance and each perturbed instance to determine their importance or weights. Instances that are more similar to the original instance are given higher weights.

Fitting Fit an interpretable, linear, or simple model (e.g., linear regression, decision tree) to the perturbed instances, using the similarity scores as weights.

The syntax in OmniXAI is

```
# Import library
from omnixai.explainers.vision import LimeImage

# Create explainer
explainer = LimeImage(predict_function=pred_func)

# Generate explanations
explanations = explainer.explain(val_imgs)
```

where `predict_func` is the function that generates the model predictions. I have failed to implement this function in my code, therefore the explanations are not presented in this case.

4 Conclusions

Three diffent techniques for explainable artificial intelligence have been presented. They are applicable to different kinds of machine learning models, however, here we have shown a practical implementation for a conmcrete problem. The classification problem we were dealing with is of medical interest as the model could be used for diagnosis. The application of these techniques aims to provide information about the models predictions, improving interpretability.

Grad-CAM has shown a better performace producing explanations. The option to select a convolutional layer for each explanation allows a better understanding of the learning process the model is following. Indeed, Grad-CAM was specifically designed for Deep Networks and therefore are more faithful to the underlying model. On the other hand, SHAP shows poor performace in this kind of problems involving images. The idea of Shapley values seems to be more interesting in context where the number of features is smaller. Then the interpretability of the model could be more effective use this methodology. Finally, LIME have not been implemented due to technical reasons. A further study would be required to assest its performance.