The medLIME algorithm

Adrianna Grudzień*, Weronika Hryniewska**

Introduction

Deep learning is useful in almost every sphere of life, but by choosing the solution suggested by the model, we agree to some kind of trust in it. Due to the multitude of parameters, we usually do not know what is hidden under the obtained result and on what basis the algorithm reached just such a result.

To make sure the model works correctly on real data, we can use explainable machine learning. One of such algorithms is Local Interpretable Model-Agnostic (LIME), as well as the medLIME algorithm created on its basis in this project.

Prototype - the LIME algorithm

The main task of the LIME algorithm is to generate predictive explanations for any machine learning classifier or regressor. Its main functionality is the ability to explain and interpret the results of models using 2D image data.

The image is divided into segments and image variations are then created with disable or enable superpixels. Superpixels are pixels of similar color that are connected to each other.

However, LIME ignores the semantic meaning and breaks down the image into superpixels, which can cover several semantically different areas, such as the lungs and body structures outside the lungs. While the color of these structures may be similar, they are different structures. The downside is that we don't have much influence on the division into superpixels.

Motivation

The explanatory of the model is especially important when analyzing medical photos. The quality of the model determines not only the efficiency of doctors' work, but also the life and health of patients. The LIME algorithm was a great help in classifying all kinds of lesions, but it also had many inconveniences and ambiguities. It was never certain on the basis of which features and whether the model actually 'made its decision' - it is not known whether the marked area was rightly taken into account, whether the algorithm recognized the object on the basis of e.g. color, if it should not matter.

An example of an undesirable operation of the LIME algorithm

Figure 1 shows an X-ray image of diseased lungs. After using the LIME algorithm in the detection of lesions, it turned out that some of the areas important for the diagnosis lie outside the lungs. This is an example showing the imperfection of the LIME algorithm.

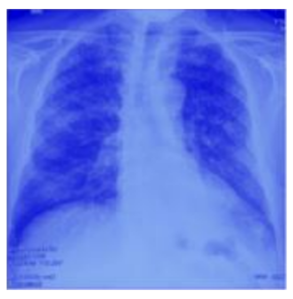


Fig . 1a Original image (1) - chest X-ray image of COVID-19 patient

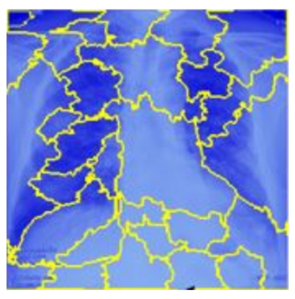


Fig . 1b Superpixels on image

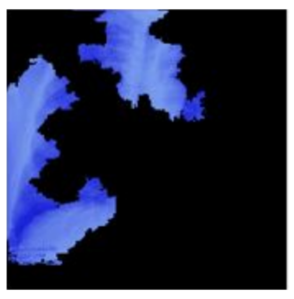


Fig . 1c Top features ('selected' by LIME)

The medLIME algorithm

The medLIME algorithm is an explanatory method based on image perturbations. The code is written in the Python. Its prototype was the previously discussed LIME algorithm. The superiority of the medLIME algorithm is primarily the ability to freely select the areas that we want to analyze, as well as the previously mentioned e.g. shape, color perturbations, with the help of which we can compare with the original image, and thus better understand which features have a significant impact on the prediction of the model.

Basic example of the medLIME algorithm's operation

After selecting the areas of interest, the algorithm additionally divides them into sub-segments and then executes the medLIME algorithm. In Fig. 2c we can see that the butterfly is quite important (green). Is it about its shape? Or maybe a color?

Fig . 2a Original image (2)

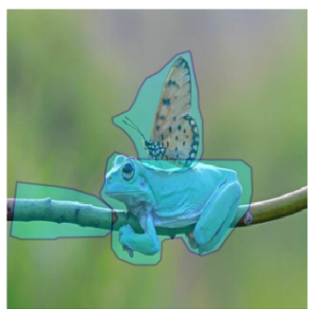


Fig . 2b Original image with manually marked areas

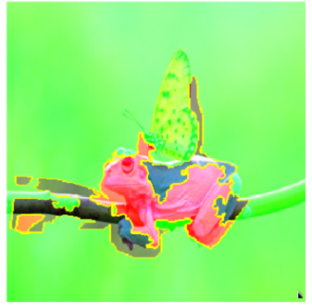


Fig . 2c Original image after using the medLIME algorithm (prediction: lycaeid (butterfly) 68.66%)

Example of medLIME algorithm application

To test the effect of butterfly color on model prediction, I changed this color to yellow. As it turned out, it was of great importance - the predictions changed significantly and then more important is frog's head and leg.

Fig . 3a Original image

Fig . 3b Perturbed image (changed color of the butterfly)

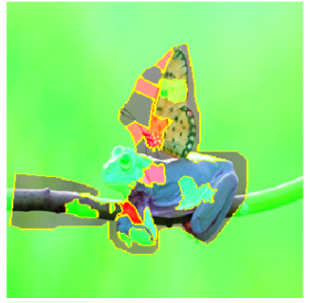
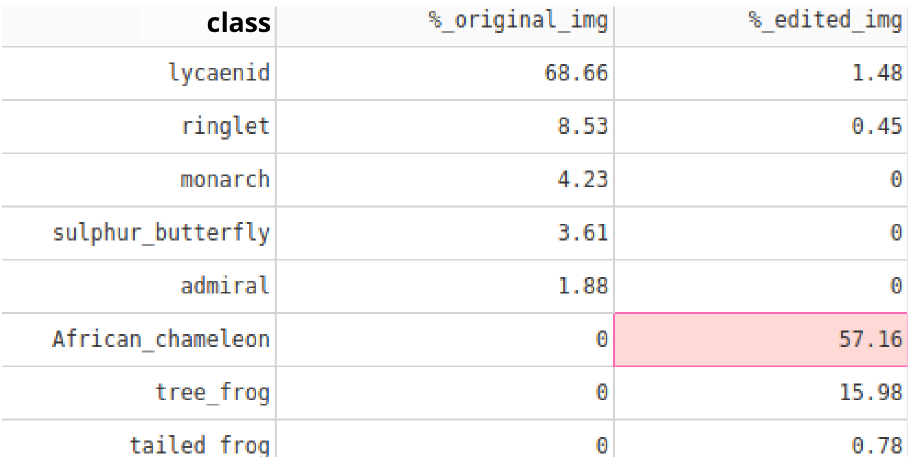


Fig . 3c Perturbed image after using the medLIME algorithm (with the same as in Fig. 2.b marked areas, prediction: African chameleon 57.16%)

Table 1 provides a summary of the percentage predictions of individual items recognized by the model for the picture of Fig. 2a and Fig. 3a The mere change in the color of the butterfly caused that the 'chance' that it is a butterfly decreased from 68.66% to 1.48%, and the African chameleon became the most likely prediction (57.16%).



class	%_original_img	%_edited_img
lycaenid	68.66	1.48
ringlet	8.53	0.45
monarch	4.23	0
sulphur_butterfly	3.61	0
admiral	1.88	0
African_chameleon	0	57.16
tree_frog	0	15.98
tailed frog	0	0.78

Tab. 1 Table of predictions for original and perturbed image

Summary

The medLIME algorithm can be widely used (especially in medicine), and thanks to the intuitive dashboard with the possibility of analyzing a selected image, it is easy to use even for people not related to IT.

Further work

We plan to further **develop the medLIME algorithm** and **the dashboard** intended for its use - creating additional functionalities, improving its operation and improving the quality of use.

In addition, the **medLIME algorithm for 3D images** is also in the works. This will open up the possibility of analyzing the operation of the models primarily in terms of medical images, such as 3D lung models, the potential of which is currently not fully exploited.

Contact:
adriannagrudzien314@gmail.com

Authors:
*BSc student in Data Science at Warsaw University of Technology
**PhD candidate in computer science at Warsaw University of Technology

References:
(1) Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M. L., & Hossain, M. S. (2020). Study of Different Deep Learning Approach with Explainable AI for Screening Patients with COVID-19 Symptoms: Using CT Scan and Chest X-ray Image Dataset.
(2) https://www.dailymail.co.uk/news/article-5756933/Hoping-prince-Butterfly-appears-kiss-frog-taking-rest-head.html