Approximate Explanations for Classification of Histopathology Patches

Iam Palatnik de Sousa(✉), Marley M. B. R. Vellasco, and Eduardo Costa da Silva

Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil
iam.palat@gmail.com

Abstract. An approximation method for faster generation of explanations in medical imaging classifications is presented. Previous results in literature show that generating detailed explanations with LIME, especially when fine tuning parameters, is very computationally and time demanding. This is true both for manual and automatic parameter tuning. The alternative here presented can decrease computation times by several orders of magnitude, while still identifying the most relevant regions in images. The approximated explanations are compared to previous results in literature and medical expert segmentations for a dataset of histopathology images used in a binary classification task. The classifications of a convolutional neural network trained on this dataset are explained by means of heatmap visualizations. The results show that it seems to be possible to achieve much faster computation times by trading off finer detail in the explanations. This could give more options for users of artificial intelligence black box systems in the context of medical imaging tasks, in regards to generating insight or auditing decision systems.

Keywords: Explainable artificial intelligence · Local interpretable model agnostic explanations · Medical imaging

1 Introduction

In an attempt to increase the transparency of black box models in Artificial Intelligence (AI) application, recent research has increasingly focused on the explainability of opaque classifiers such as Neural Networks [1,4].

The potential benefits of this transparency have been frequently discussed in literature, for areas of application ranging from social good and fairness to legal use [7].

These developments are especially crucial for Medicine related applications, where auditing Machine Learning (ML) and Deep Learning (DL) systems is

I. Koprinska et al. (Eds.): ECML PKDD 2020 Workshops, CCIS 1323, pp. 517–526, 2020.
https://doi.org/10.1007/978-3-030-65965-3_35

seen as essential. Requiring blind trust of these models would be an obstacle for greater application in clinical settings [4].

Among the explainable AI (XAI) approaches, a frequently applied one involves training interpretable surrogate models that approximate the more complex black-box model studied. Among these the most known and used method is that of the Local Explainable Model Agnostic Explanations (LIME), developed by Ribeiro et al. in 2016 [6].

Such a technique can generate human understandable explanations, and its model-agnostic nature means it can be applied to any number of classifier systems, from various neural network architectures to other complex ensembles.

LIME can further be used for various data types, from Natural Language Processing problems to Image Classifications. The main idea of this technique, described in greater detail within the next sections, is to explain individual instances by locally perturbing them. For images, this means generating images where parts are covered, training a simpler more interpretable model to identify how much each region contributes to a given classification.

More recently, Palatnik-de-Sousa et al. [8] have employed LIME to generate explanations for a dataset of lymph-node metastasis images, used in a binary classification task for presence/absence of metastatic tissue. The explanations of a Convolutional Neural Network's (CNN) classifications were then compared to medical expert segmentations, showing agreement between both.

A recent survey of the state of the art for XAI in digital pathology [5] mentions two additional related works dealing with XAI for medical imaging. Tang et al. [10] developed interpretable classifications of Alzheimer's disease pathologies, and Huang and Chung [2] describe a method for weakly supervised learning that can effectively pinpoint cancerous tissue in cancer detection tasks.

However, after the results in [8] some key issues became apparent and highlighted potential problems of directly applying LIME without fine tuning. Namely, using this technique with standard parameters might not generate the best results. The segmentation of the image into meaningful contextual subregions (often called segments or superpixels) is controlled by certain parameters that were shown to influence the results.

In the case of [8] these parameters were manually tuned. However this demonstrated other two possible issues. The first is related to computational times. Since LIME depends on generating sets of perturbed images, and each of these must be evaluated by the black-box model studied, this can and does create a computational bottleneck, especially for large CNN models that might take more time to evaluate each image. This makes the task of fine tuning parameters more time consuming. Also, due to the random nature of how these perturbed images are generated, the results of a given LIME explanation could, and generally do vary, if repeated multiple times for the same instance.

Aiming to solve these issues, Palatnik-de-Sousa et al. [9] developed a novel explainable methodology, modular in nature, that uses Multi-Objective-Optimization (MOO) combined with an explainable algorithm (in this case LIME) to automate parameter fine tuning, and consequently find the best explanations. The explanations generated by their EvEx model shows it's possible to find explanations that are robust against the random nature of LIME, improving upon their previous results.

However, although this seems to remedy the reproducibility problem, it still generates a large computational cost, with the runs of this explainable model taking between 4 and 8 h for each patch studied. The goal in this manuscript is to present an alternative solution to this problem, that dramatically reduces the computation times from the order of magnitude of hours, to minutes.

In order to do this, the detail and quality of the explanations is traded off. By using very simple square divisions on the image, the idea is to test whether this simple approach can find the most relevant parts of an image quickly, whenever this might be necessary or enough for a given image classification problem, or auditing.

2 Materials and Methods

In this section the main aspects of the methodology are briefly described. A summary of how LIME and EvEx work, as well as the dataset used for this study are presented. Then, the proposed approximated model is described.

The experiments here described largely follow the same methodology discussed in greater detail on [9], with the main difference being the use of the approximated explanation generation, instead of the EvEx model.

2.1 Patch Camelyon

Patch Camelyon (P-CAM) [3,11] is a dataset derived from the Camelyon 16 Whole Slide Images (WSI). It consists in about 200 thousand 96 by 96 pixel patches of histopathology images. These images have a binary label representing the presence or absence (labels 1 or 0 respectively) of at least one pixel of metastasis in the 32 by 32 pixel center of the image. Furthermore the dataset is balanced so that the classes are divided in nearly equal splits.

A more detailed discussion of this dataset can be found in [8] and [9]. The patches selected for this study were true positives, to allow for comparison with medical expert manual segmentations contained in the dataset.

2.2 LIME

For the case of image classification problems, the computation of LIME explanations involves first separating the images into segments that hold some contextual information.

In general this means contiguous sub-regions of the image - often called superpixels - that have colors or textures in common, and could reasonably represent a relevant pattern in a human understandable explanation. There are multiple algorithms that can be used to segment an image into such super-pixels, and it has been shown [8,9] that good segmentations can deeply affect the results of the explanation.

Once a given instance – meaning a given image whose classification is meant to be explained – is selected and segmented into superpixels, a set of perturbed images is then generated. This is accomplished by randomly covering superpixels (for instance, with a black color). Typically hundreds or thousands of such perturbed images are created, to generate a varied distribution compared to the original instance to be explained.

Each perturbed image is then passed through the model being studied, generating a prediction. From these predictions and the perturbed distribution a linear model is trained, which finds how much each particular superpixel in the original segmentation contribute in favor or against a given classification. The explanation can then be visualized as a heatmap showing how relevant each super-pixel is.

Additionally, to get a sense of how the linear model fits the perturbed data, a quantity termed ‘explanation score’ can be defined as the R-squared of the linear fit [6].

2.3 EvEx

The EvEx model proposed by Palatnik de Sousa et al. [9] essentially expands upon LIME by using a multi objective genetic algorithm to determine the instance segmentation parameters. It generates a set of best parameters (and subsequently best explanations) which are then averaged onto a final heatmap.

Since an MOO is used the best individuals form a Pareto Front of explanations. The model is described in greater length and detail in [9].

2.4 Convolutional Neural Network Model

The same CNN used in [8] and [9] was used in this manuscript to allow for better comparison. It is a typical convolutional network, consisting in three convolution blocks of increasing filter sizes with a dense block ending in a softmax output layer. It is a publically available model from a Kaggle competition, trained on the P-CAM version used in this paper [3].

2.5 Proposed Method

Instead of using complex segmentation algorithms, the approximation proposed in this manuscript simply divides the instance to be explained in two “segments” or regions. A square, and the area surrounding the square. It is clear that such a simple division typically will not hold as much contextual color/texture information as a more precise segmented superpixel, however using this approximation with only one square creates a simplification of the next step in the process.

Figure 1 shows an example of this square segmentation. Notably, it is immediately clear that for such a simple segmentation, there are only 4 possible options

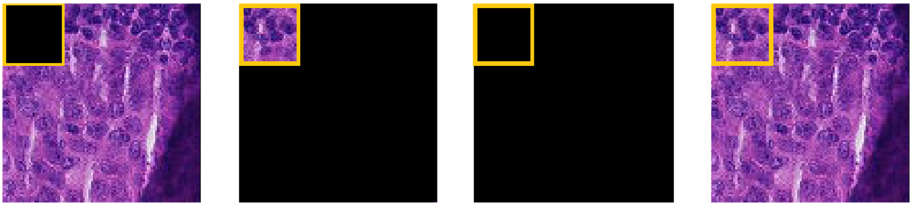

Fig. 1. Example of square segmentation and perturbed images. The yellow square is overlaid onto a P-CAM sample patch. The 4 images represent the original (on the right) and the three possible perturbations. (Color figure online)

regarding perturbed images. Either the image is left intact, as seen on the rightmost part of Fig. 1, or there are three possible perturbations, where the square is covered, the area outside the square is covered, or the entire image is covered.

It is not immediately trivial whether using such a simple approximation could generate meaningful explanations with lime, especially since now the linear surrogate model is adjusted on only 4 images, rather than a large distribution of perturbations with hundreds or thousands of examples.

On the other hand, as seen later in the next section, the approximation potentially yields useful results while dramatically reducing the computational costs. The massive reduction in the number of perturbed images means the CNN also performs less evaluations. The key idea is that the approximation, despite leading to less detailed explanations, can compensate this trade-off by being faster by orders of magnitude.

However, to generate the most meaningful explanations possible with this approximation, it would be interesting to find the most explainable squares. Typically, making the squares too small would mean the area outside the square holds most of the contextual information and would yield larger explanation weights, as seen in similar scenarios where large superpixels dominate explanations [8,9]. As such, it is interesting to try to find the square, of a given size, that has the highest explanation weights.

It is also expected, by the same reasons, that larger squares will have larger explanation weights, since they cover a larger area of the image. However, smaller squares might be able to discern which smaller features in a given patch contribute more towards a given class.

As a preliminary test, all square sizes were tested, for all positions on the patch. This means that for a given P-CAM patch, explanations were generated using this square segmentation, using squares with sizes from 1 × 1 pixels to 95 × 95 pixels, in which these explanation squares are swept across the image from the top left corner down to the lower right corner.

Besides plotting the weights of each square, at each size, to analyze which ones best explain classifications, a strategy was also devised to create a visualization heatmap for each size of square:

- A given square size is selected
- The squares of that size with positive explanation weights (contributing in favor of the classification) are selected.
- They are added together to create a heatmap
- To compute the average of the added squares, each pixel of the heatmap is divided by the number of squares that are contributing to it.

In this sense, this visualization heatmap is the pixel-wise weighted average of the positive weight squares.

This emulates and approximates the final result of EvEx which is the averaged heatmap of the best explanations found. The expectation is that possibly this square approximation can find similar areas to those identified by EvEx, although with much less detail. On the other hand, the approximation might find them in minutes, compared to the 4 8 h it takes for an EvEx run to complete for one P-CAM patch.

For this initial experiment, 95 such heatmaps are generated for each P-CAM patch (one for each square size).

3 Results and Discussion

The preliminary tests described in the previous section, with varying square sizes, showed that in general the squares of high explanation weights tend to focus in on the same regions of the patch as the size decreases from 95 to 1.

Initially the squares are too large and take most of the patch, which causes them to have high explanation values. However as size decreases these values diminish until the point where the squares are too small to explain the full patch, causing the weights to be negligibly small (close to zero or slightly below zero).

This initial test seems to demonstrate that it is not necessary to generate all the heatmaps of all square sizes. Rather, for each particular patch one may generate a few heatmaps at different sizes and observe the regions to which they converge. Sizes below 30 often are too small to generate explanations, and sizes above 80 are generally too large. Figure 2 shows an example for two patches, including the medical segmentation and also the EvEx equivalent explanation.

Notably, the plots in Fig. 2 also highlight the similarities and differences between the square approximations and the much more detailed EvEx output. The latter seems to clearly delineate some regions, both within the medical segmentation, where the averaged explanation weights are much higher than the surroundings. These regions follow the shape, texture and color contours of certain cellular structures within the tissue.

However, one may note that these EvEx highlighted regions roughly coincide with the regions highlighted by the square approximation, as seen both by sub-panels E and F of Fig. 2, for both patches. Both the squares with highest explanation weights and heatmaps seem to be able to at least approximate the position of the most relevant areas of the patch.

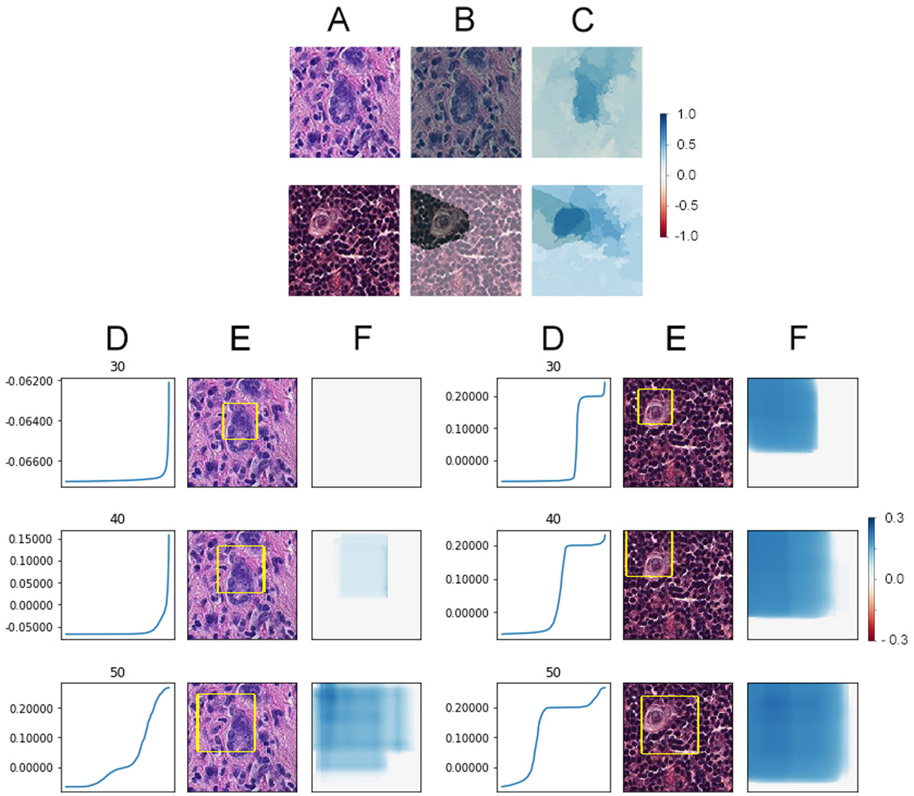

Fig. 2. Results for two different patches. (A) – original patches; (B) – medical segmentation as green overlay; (C) – EvEx heatmap; (D) weights of all squares generated, by size, sorted by increasing values. Sizes (30, 40, and 50) indicated above each plot; (E) – square with the largest weight overlaid on patch; (F) – averaged heatmap as described in Sect. 2. Note that the EvEx heatmaps have a color scale from −1 to 1, while the approximated square heatmaps range from −0.3 to 0.3. (Color figure online)

Another important difference is in regards to the absolute values of the explanation weights. Typically the highest weights found in the EvEx explanation reach values around 0.6 to 0.9, while the square approximation did not yield values above 0.3 in any of the patches studied. This is expected to a degree, as the squares hold much less contextual information than the more detailed EvEx optimized segments. It is also the reason why the color scale of the bottom panels in Fig. 2 was adjusted to a smaller range than the top panels. This seems to mean that lower explanation weights could be expected in the square approximation, but it still manages to roughly delineate relevant regions.

Another key difference, and perhaps the central one to this approach, is in computation times. Figure 3 shows the amount of time taken to generate heatmaps at each square size, in seconds. For the sizes of 30, 40 and 50 used in

Fig. 2, the times were respectively of about 5 min, of about 3 min and 42 s, and of 2 min and 31 s. The largest times observed, for 1 by 1 pixel squares, was of about 10 min. As a comparison, the top and bottom heatmaps of Fig. 2, subpanel C, took about 4 h and 7 h, respectively. This constitutes a computation time decrease of many orders of magnitude.

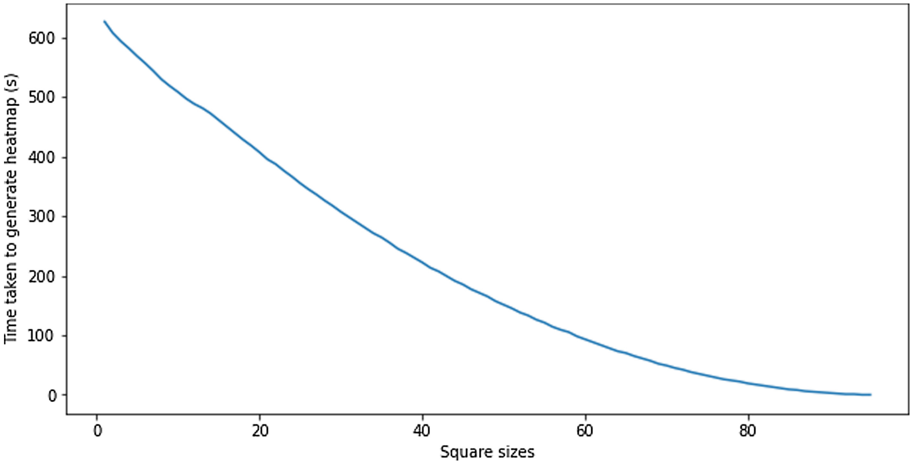

Fig. 3. Time, in seconds, taken to generate heatmaps for each square size, from 1 by 1 to 95 by 95 pixel squares.

Figure 4 shows a side by side comparison of two heatmaps. The first, to the left, corresponds to a high scoring EvEx Pareto front individual used to generate the averaged heatmap in Fig. 2, topmost row of column C. The second, to the right, corresponds to the approximated square of highest explanation weight, size 40, for the same image. Besides the already noted and discussed difference in weights, it seems the explanation score is also lower for the square approximation. As previously mentioned, this score corresponds to the R-squared of the surrogate linear model fit.

However it is worth remembering that the EvEx pareto front individual, in this case, is an explanation generated with 200 perturbations, whereas the square approximation only uses 4. This most likely affects the R-squared metric and contributes to the observed result. Importantly, despite these lower metrics, the heatmaps seem to highlight the same areas.

As such, these results seem to indicate that, while more computationally efficient versions of LIME and EvEx are created, these square approximations can be used to generate faster insights on what might be the key areas of an image the CNN is focusing on.

Future studies could also focus on developing criteria for selecting best square sizes in order to create combined heatmaps.

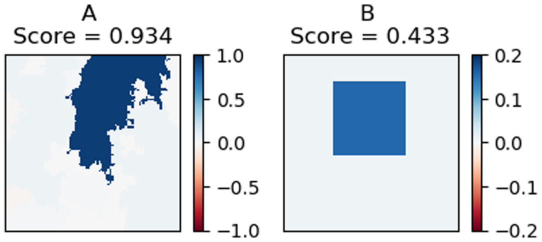

Fig. 4. Side by side comparison of heatmap explanations generated by both methods analyzed in this work, for the same patch (Fig. 2, topmost row of column A), with corresponding explanation scores. Panel A shows an EvEx Pareto Front Individual, while panel B shows the square approximation with highest explanation weight for squares of size 40.

4 Conclusion

In this manuscript an approximation method was presented, in order to generate faster explanations for a medical imaging classification task.

The trade-off between the level of detail of explanations (such as the ones generated with EvEx) and the time it takes to generate such explanations was highlighted. In this way, aiming at reaching much faster computational times and by accepting less detailed explanations, an alternative was presented where simple square segmentations are used.

This, in turn, means that the LIME explanations would be generated from only 4 perturbed images, which is a very rough approximation. However the results seem to show that this rough approximation can still be used to determine, even if with less detail, which areas of an image are most relevant. The gain of several orders of magnitude in computation times might be interesting for some applications, however. Furthermore, results showed that this new method reduce computational times from several hours [9] to a few minutes or less, depending on the square size.

Future projects might focus on further testing this concept in other medical imaging datasets, as well as datasets from other computer vision areas.

Acknowledgment. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal do Nível Superior - Brasil (CAPES), Finance Code 001. The authors also acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Rio de Janeiro (FAPERJ) for the funding to this research.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

2. Huang, Y., Chung, A.C.S.: Evidence localization for pathology images using weakly supervised learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 613–621. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_68
3. Kaggle: Histopathologic cancer detection. https://www.kaggle.com/c/histopathologic-cancer-detection
4. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. **19**, 1236–1246 (2017)
5. Pocevičiūtė, M., Eilertsen, G., Lundström, C.: Survey of XAI in digital pathology. In: Holzinger, A., Goebel, R., Mengel, M., Müller, H. (eds.) Artificial Intelligence and Machine Learning for Digital Pathology. LNCS (LNAI), vol. 12090, pp. 56–88. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50402-1_4
6. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
7. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
8. Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, E.: Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors **19**(13), 2969 (2019)
9. Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., Costa da Silva, E.: Evolved explainable classifications for lymph node metastases. arXiv preprint arXiv:2005.07229 (2020)
10. Tang, Z., et al.: Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. Nat. Commun. **10**(1), 1–14 (2019)
11. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 210–218. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_24