

Multi-scale Local Explanation Approach for Image Analysis Using Model-Agnostic Explainable Artificial Intelligence (XAI)

Hooria Hajian and Mehran Ebrahimi

Faculty of Science, Ontario Tech University, Oshawa, Ontario, Canada

ABSTRACT

The recent success of deep neural networks has generated remarkable growth in Artificial Intelligence (AI) research and has received much interest over the past few years. One of the main challenges for the broad adoption of deep learning-based models such as Convolutional Neural Networks (CNN) is the lack of understanding of their decisions. To address this issue, Explainable Artificial Intelligence (XAI) has been proposed to shift toward more transparent AI, resulting in the development of techniques to explain decisions by AI models. This paper aims to explore and develop a multi-scale scheme of LIME (Local Interpretable Model-Agnostic Explanations) applied to image classification to explain decisions made by CNN models through heatmaps of coarse to finer scales. More precisely, when LIME highlights large superpixels from a coarse scale, there may be smaller regions in the corresponding superpixel that influenced the model's prediction at some finer scale. In the proposed multi-scale scheme, two weighting approaches, one based on Gaussian distribution and another parameter-free framework will be introduced to produce visual explanations observed from different scales. Promising results for multi-scale classification heatmaps of histopathology images are presented. More specifically, we investigated the proposed multi-scale approach on Camelyon16 dataset. The results show that the explanations are faithful to the underlying model, and the visualizations are reasonably interpretable.

Keywords: Explainable Artificial Intelligence (XAI), LIME, Multi-scale Explanations, Histopathology images.

1. INTRODUCTION

In recent years, machine learning models have emerged in biomedical image processing and disease diagnosis.^{1–3} Among different image processing techniques convolutional neural networks (CNNs) and their related architectures have been broadly proposed and investigated for medical image classification, screening, and segmentation,^{4,5} due to their capability in sharing parameters effectively into different layers of a deep learning model.⁶ Despite the progress in the development of ML models with high accuracy, a major existing drawback is their adaptation as a “black box tool” for many applications.^{1,7,8} In the medical imaging domain when Computer-Aided Diagnosis (CAD) is being involved, as the learning process of CAD models is not always transparent and understandable, the experts cannot fully trust these models.^{9–11} The applicability of classification models to diagnose disease is highly dependent on the ability to interpret and explain the applied models.¹²

It is vital to explain model predictions, particularly in medical imaging applications.^{13,14} In recent years techniques have been developed to explain the decisions made by artificial intelligence (AI) models known as Explainable-AI (XAI).¹⁵ According to the type of data and the interpretability scale, which can be either local or global, various approaches of XAI methods have been developed.¹⁶

One of the popular ideas to explain black boxes, especially for CNNs, is using some specific functions to propagate the outputs back to the input neurons that generate the explanations as the heat maps where each pixel of the map is coloured based on the value and weight it has in that specific classification. Contrary to the first method, a different approach where there is no backpropagation algorithm and not limited to any specific model type, namely Local interpretable Model-agnostic Explanation (LIME) was developed.^{8–12,17} LIME is a well-known and frequently used method to ensure the interpretability and explainability of the AI-based models.¹² As LIME is basically designed to be model agnostic, it can be used for various machine learning methods. The model created by this method specifies which parts or characteristics of the input data are more significant, and

hence it makes the results of the model more interpretable.¹⁸ The benefits of using LIME for the medical data models can be summarized as 1- It can provide a patient-specific explanation for a special classification meaning that any complicated classifiers can be explainable simply for clinical purposes. 2. It can indicate the amount of contribution of each variable for each estimation in the model.¹⁹ 3. LIME can determine which variables and how (i.e. in what extent and direction) they influence the output of each estimation made by the model.²⁰

With the idea of producing fine-grained visual explanations of medical images, the results of LIME have been investigated for the classification of lymph node metastases on the Camelyon16 dataset in²² using three well-known segmentation algorithms, including Felzenszwalb's, SLIC, and Quickshift. All three algorithms have a parameter “*sigma*” that defines the width of a Gaussian in the preprocessing step. Higher sigmas typically result in a smaller number of segments. This is the only parameter shared in common by these algorithms, and each of these algorithms has several parameters affecting the explanation by LIME. The sensitivity of these algorithms to variations in image input’s texture, color, and their own variables raises the question of how the best set of parameters can be determined. As an alternative approach,²² proposed a new method to segment each studied image into non-overlapping square grids. On the one hand, these segments did not hold the contextual meaning of a typical superpixel, but the scheme guaranteed that every image was divided into exactly the same number of segments at the same positions. These square segments were then passed to the LIME algorithm in the way the usual superpixels would be. The resulting weighted heatmaps gave a rough idea of what sub-regions of the image were most relevant for a given classification task. The finer grids provided a more fine-grained view with a larger number of small square segments. Although in the square-grid segmentation approach, the image will be divided into the same number of segments with the same probability to be masked, this approach does not hold the main contextual meaning of a segmentation task called boundary recall. Our goal is to propose a multi-scale scheme of LIME as a model-agnostic explanation tool along with two weighting approaches to produce visual explanations obtained from different scales.

2. METHODOLOGY

2.1 Multi-scale Visual Explanation of LIME

A suitable segmentation approach with higher boundary recall will stick to object boundaries. According to the literature,²³ Simple Linear Iterative Clustering Segmentation (SLIC) is a superpixel-based algorithm with high boundary recall and low under-segmentation error. With this idea, we were motivated to propose a multi-scale version of LIME using the SLIC segmentation algorithm and provide visual explanations at coarse to fine scales of an image. The efficacy of SLIC superpixels in object recognition and medical image segmentation has been proven in the literature.²³ SLIC obtains high-quality results with computational efficiency compared to other state-of-the-art algorithms. In this work, we have decided to use the SLIC segmentation algorithm and specify a few main parameters: compactness, the desired number of superpixels, and the number of perturbations. We aim to segment an image at multi scales to utilize superpixels of an image from coarse to finer scales by keeping the multi-scale object boundaries and preserving the contextual “meaning” of superpixels. The pseudo-code of the proposed multi-scale scheme is shown in Algorithm 1, and the list of notations used is presented in Table 1.

As it is shown in Algorithm 1, the only parameters that the proposed multi-scale scheme takes are the desired number of segments/superpixels, s_1, \dots, s_L , where L is the number of levels, c the compactness, and the *num-of-samples* or perturbations n . Setting $c = 5$ seems to best preserve the objects’ boundaries in our experiments. Then at each level $l = 1, \dots, L$ in a loop, SLIC takes s_l as the desired number of superpixels and returns s'_l as the actual number of superpixels after the segmentation process is performed. Based on the number of samples/perturbations the algorithm creates the neighbors by randomly masking some superpixels. A separate linear model is fitted at each level, then we take the coefficients, normalize the value of the coefficients in a range of (0,1) and return the result as a set of heatmaps.

Fig 1 provides a schematic example of the proposed multi-scale outputs for a sample image. In the next step, we should then combine the result of the heatmaps by forming their weighted average using some appropriate weighting scheme. Fig 2 shows a few schematic examples of combining heatmaps of Fig 1. Depending on the chosen weighting scheme, the user can choose to put more emphasis on the coarser scale and/or finer scale explanations. Fig 3 shows another example of multiscale heatmaps.

List of Notations / Parameters	
x	Input image of size $(h, w, 3)$, h : height, w : width, 3: color channels.
L	Number of levels for multi-scale segmentation.
l	Index of segmentation level.
s	Desired number of superpixels.
c	Compactness.
s'	Final number of superpixels that SLIC provides at each level.
n	<i>Num-of-samples (perturbations) = $10 * s'$</i> , i.e., $10 * \text{number of superpixels}$
p	Perturbation or masked image which is a matrix of size $(m, n, 3)$.
X	Binary matrix of size (n, s') .
Y	Vector of size $(1, n)$: predicted label for each perturbation.
coefficients	Vector of size $(1, \text{num of superpixels})$: the importance of each superpixel.
E	Explanation matrix of size (h, w) represents the coefficients.

Table 1. Parameters description

Algorithm 1 LIME with Multi-scale Scheme

Require: $s_1 \dots s_L$, compactness c , and *num-of-samples* or perturbations n

for $l \leftarrow 1$ to L **do**

Apply SLIC segmentation algorithm with s_l , and c ;

Calculate The actual ‘*num of superpixels*’ s'_l after the segmentation process;

Generate samples/perturbations (p_1, \dots, p_n) ;

Calculate distance/weight of perturbations from input image (w_1, \dots, w_n) ;

Predict label/class for each perturbation (y_1, \dots, y_n) ;

Fit linear model on set of perturbations X and predicted labels Y ;

Calculate the coefficients of the linear model;

Normalize the coefficients to calculate E_l ;

end for

Return the heatmaps (E_1, \dots, E_L) ;

2.2 Combining Multi-scale Visual Explanations

To combine the visual explanations of the proposed multi-scale scheme in an appropriate way, we propose a parameter-free weighting approach based on the trapezoidal rule to classify the heatmaps produced by the multi-scale version of LIME. The main point of this weighting approach is to combine the heatmaps of different levels and highlight the most important superpixels/regions from coarse to finest scales.

The trapezoidal rule is a common numerical integral approximation method, that divides the area under a curve into smaller trapezoids. More specifically, let $\{x_k\}$ be a partition of $[a, b]$ such that $a = x_0 < x_1 < \dots < x_{N-1} < x_N = b$ and Δx_k be the length of the k -th sub-interval. That is $\Delta x_k = x_k - x_{k-1}$, then

$$\int_a^b f(x) dx \approx \sum_{k=1}^N \frac{f(x_{k-1}) + f(x_k)}{2} \Delta x_k. \quad (1)$$

When the grid spacing is non-uniform, one can use the formula of equation 1. With the idea of approximating definite integral with the irregularly-spaced partition, we decided to assign weights to each scale. In other words, in the segmentation of an image at different scales, the number of segments might vary according to the user-specified set of parameters. Therefore, the number of segments in this multi-scale segmentation is not specifically equal, and the spacing is non-uniform. Hence, we propose two schemes to assign weights to the heatmaps from different scales. Then, we compute the weighted heatmap to represent the results. We aim to assign weights to each heatmap based on how far/close the number of segments is to the next level.

2.2.1 First Automated Weighting Approach

We assume the partitions/number of segments are not necessarily equally spaced. Hence, we expanded the equation 1. With the idea of trapezoidal rule for approximating the definite integral and the irregularly-spaced

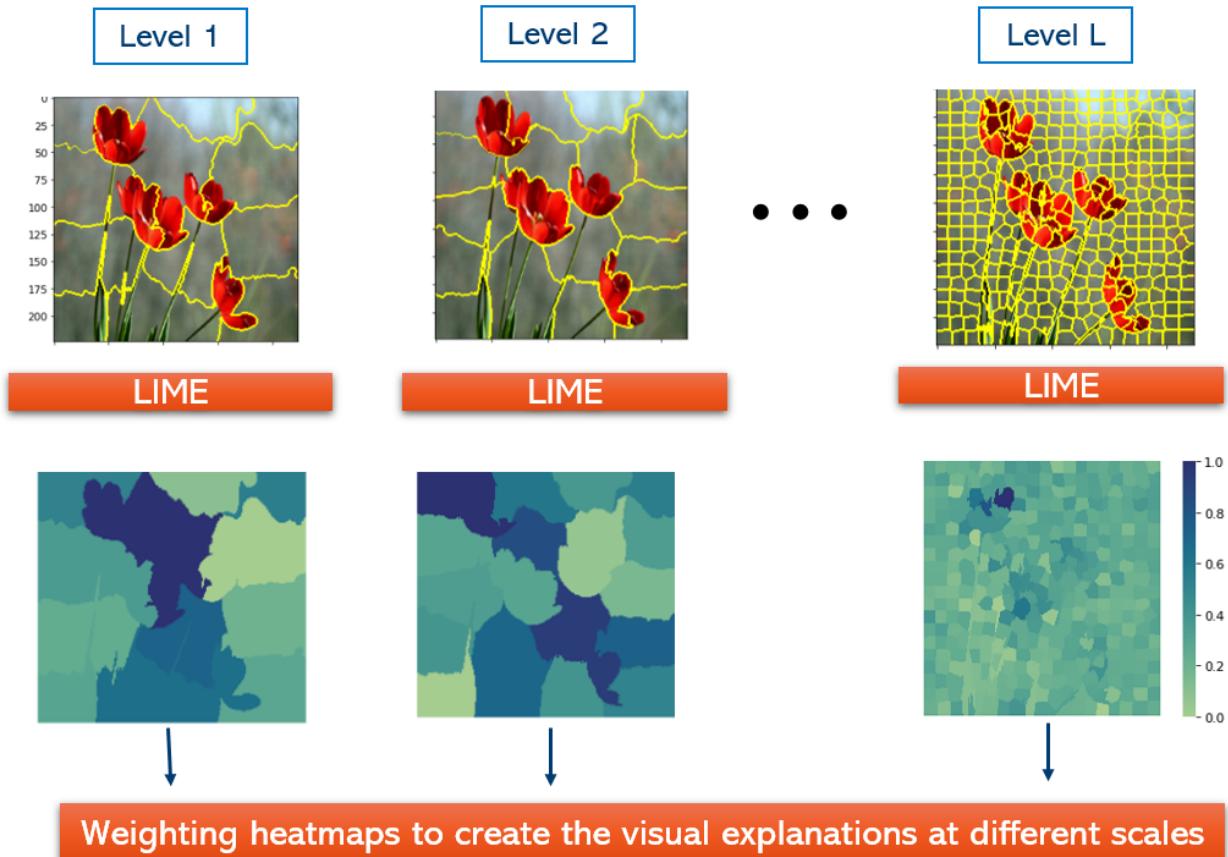


Figure 1. General framework of the proposed multi-scale scheme of LIME

partition, we define the first automated weighting approach as

$$\sum_{l=1}^L \frac{(E_{l-1} + E_l)}{2} \times (s'_l - s'_{l-1}) = E_0 \frac{d_0}{2} + E_1 \left(\frac{d_0}{2} + \frac{d_1}{2} \right) + \cdots + E_{L-1} \left(\frac{d_{L-2}}{2} + \frac{d_{L-1}}{2} \right) + E_L \left(\frac{d_{L-1}}{2} \right), \quad (2)$$

where E_1, \dots, E_L is a set of explanation heatmaps, s'_1, \dots, s'_L is the final number of segments/superpixels and d_0, \dots, d_{L-1} are the lengths of the intervals between the consecutive number of segments. $d_{l-1} = s'_l - s'_{l-1}$, for $l = 1, \dots, L$. Therefore, each heatmap will be multiplied by the corresponding weight in the weighted average. These weights are calculated based on the distances/intervals of the number of segments in the proposed multi-scale scheme.

2.2.2 Second Automated Weighting Approach

In the second automated weighting approach, we consider the multiplicative inverse of the final number of segments to calculate the distances/weights, $\frac{1}{s'_0}, \dots, \frac{1}{s'_L}$, where $d'_{l-1} = (\frac{1}{s'_l} - \frac{1}{s'_{l-1}})$, for $l = 1, \dots, L$. In this approach inverse of the scale is parameterizing the approximation. We define the second weighting approach as

$$\sum_{l=1}^L \frac{(E_{l-1} + E_l)}{2} \times \left(\frac{1}{s'_l} - \frac{1}{s'_{l-1}} \right) = E_0 \frac{d'_0}{2} + E_1 \left(\frac{d'_0}{2} + \frac{d'_1}{2} \right) + \cdots + E_{L-1} \left(\frac{d'_{L-2}}{2} + \frac{d'_{L-1}}{2} \right) + E_L \left(\frac{d'_{L-1}}{2} \right). \quad (3)$$

3. EXPERIMENTAL STUDY

In²² the square-grids segmentation has been tested on Camelyon16 dataset²² with VGG19 network as the black-box model. We chose the same VGG19 black-box model with the same dataset to assess the results of the

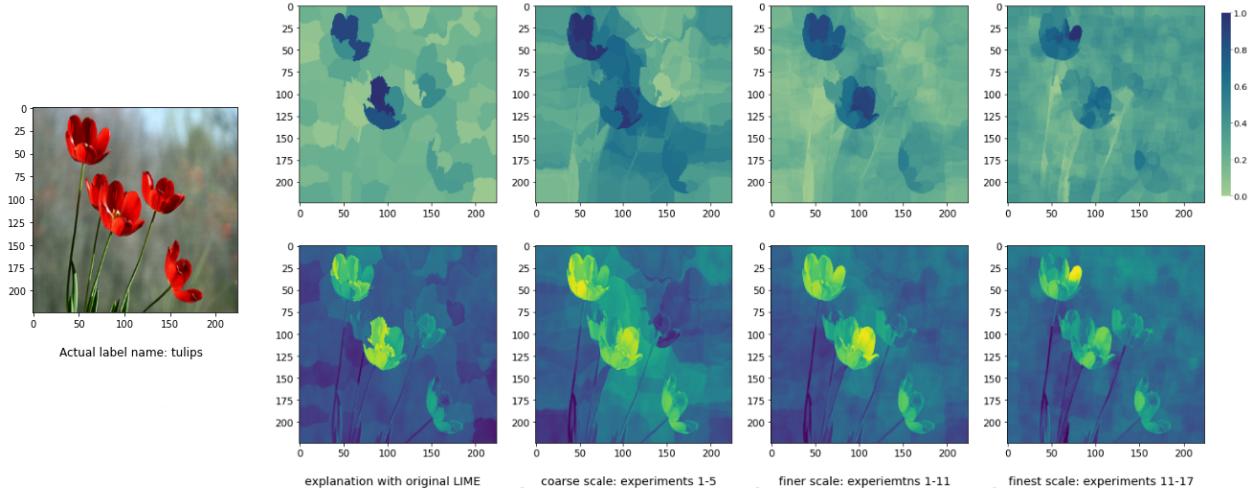


Figure 2. First predicted class “tulips”: comparison of original LIME vs multi-scale scheme from coarse to finest scale. The heatmaps are superimposed on top of the original image shown in the second row for better visualization.

proposed multi-scale scheme. This model consisted of a CNN with 0.968 prediction accuracy on the test set.²² The only change we made was in the last layer of the network to obtain the prediction probabilities through the softmax activation function and then trained the model on the test set. This section aims to evaluate the visual explanations of the proposed multi-scale scheme with the automated weighting approaches on Camelyon 16. To assess the results quantitatively, we test the reliability of the explanations based on the proposed multi-scale approach and compare it with the use of square-grids along with LIME.²² The idea behind the proposed quantifiable system is that salient regions detected by XAI algorithms should be sufficient to obtain approximately similar prediction accuracy, if not higher than the accuracy obtained when the entire attribute set is provided to the model. In other words, the proposed quantifying system feeds the explanations of the input features obtained from the XAI algorithm to the black-box model as input and calculates its accuracy and prediction label. This information is then compared to the accuracy and prediction label computed when the original input feature subset is passed as input to the same black-box model. The whole idea is that since the algorithm already captures the most influential features during the training phase, sending the explanations would aid in quantifying the reliability of the built model. Initially, Prediction Accuracy “*pa*” is computed when the entire feature subset or input image is passed to a black-box model and the corresponding prediction class is considered. The prediction is then utilized in computing the model’s explanation and a feature subset. More precisely, by sending the initial input image with its corresponding prediction to an XAI algorithm, salient features of the original input are obtained. The obtained feature subset containing the most influential features in computing the prediction is then passed to the aforementioned black-box model, and the prediction label with its accuracy is calculated as explanation accuracy “*ea*”.

3.1 Data: Lymph Node Metastases on Camelyon16

Patch Camelyon (P-CAM) is a dataset developed by Veeling et al.²⁴ It was derived from the Camelyon16 hematoxylin and eosin-stained whole slide images (WSIs). The original WSIs were acquired and digitized with a $40\times$ objective (corresponding to a pixel resolution of 0.243 microns),²⁵ and undersampled at $10\times$ to increase the field of view for P-CAM. The 96 by 96 pixel patches were extracted by²⁴ from the gigapixel WSIs by converting the slides to hue-saturation-value format (HSV), followed by blurring and filtering out patches that had saturation lines below 0.07, which was shown to exclude irrelevant background patches.²² The binary labels of each image corresponded to the presence or absence of at least one pixel of tumor tissue in the central 32 by 32 pixel square of each patch. 0 meaning absence, and 1 meaning the presence of tumor tissue.²⁶

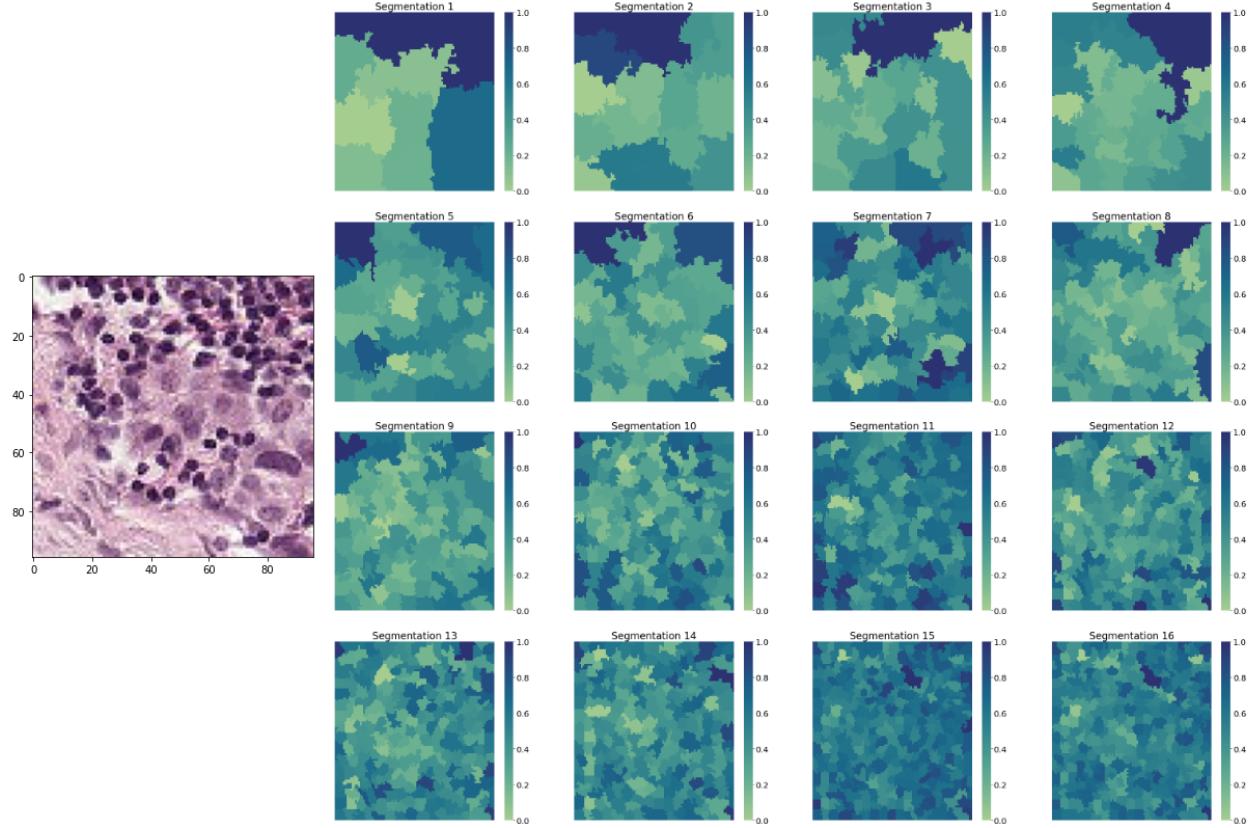


Figure 3. Heatmaps of the multi-scale scheme with 16 levels of experiments. Darker intensities in the heatmaps indicate higher probability of tumor tissue.

3.2 Results

We chose five random sample images from the Camelyon 16 dataset with the actual label 1, i.e., where tumor tissue is present. These samples are correctly classified as true positive with the VGG19 network. The results of multi-scale scheme and the combined heatmaps by the automated weighting approaches were then obtained. To compare the results, we display the heat maps corresponding to the original LIME, multi-scale approach with the two weighting schemes and the results based on the square-grids,²² using our two proposed weighing schemes, see first row of Figs 4 to 8. Furthermore, we overlay the sample image over each of the obtained heatmaps that are displayed in the second row on Figs 4 to 8. Finally, for each result, the calculated explanation accuracy (*ea*) is included. It can be observed that for all of these samples, the multi-scale approach outperforms both the square-grids and the original LIME scheme in terms of explanation accuracy. Darker values in the first row of the displayed images in the Figures hint the abnormality. It can be observed that localizations provided by the multi-scale and square-grid approaches are not always consistent with the original LIME algorithm.

The Python implementations of the proposed method, evaluation metrics, and visualization of the explanations are available at the following link.

<https://github.com/hajiyanh/Multi-scale-Visual-Explanation-of-LIME>

4. DISCUSSION AND CONCLUSION

We proposed a multi-scale scheme to enhance visual explanations of LIME at different scales through the results from attentive heatmaps that can produce more comprehensive explanations from coarse to fine scale. In other

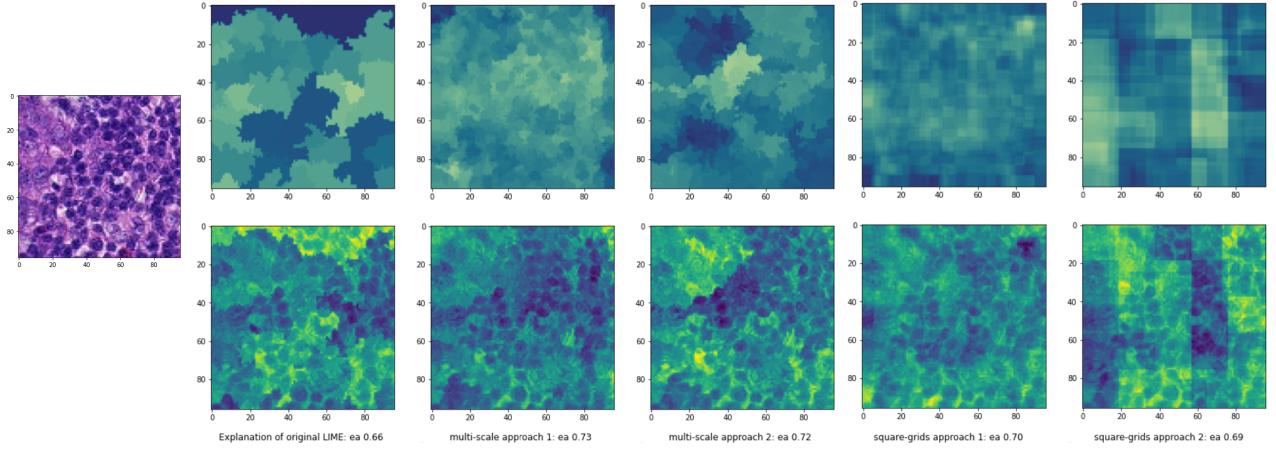


Figure 4. True positive sample image 1

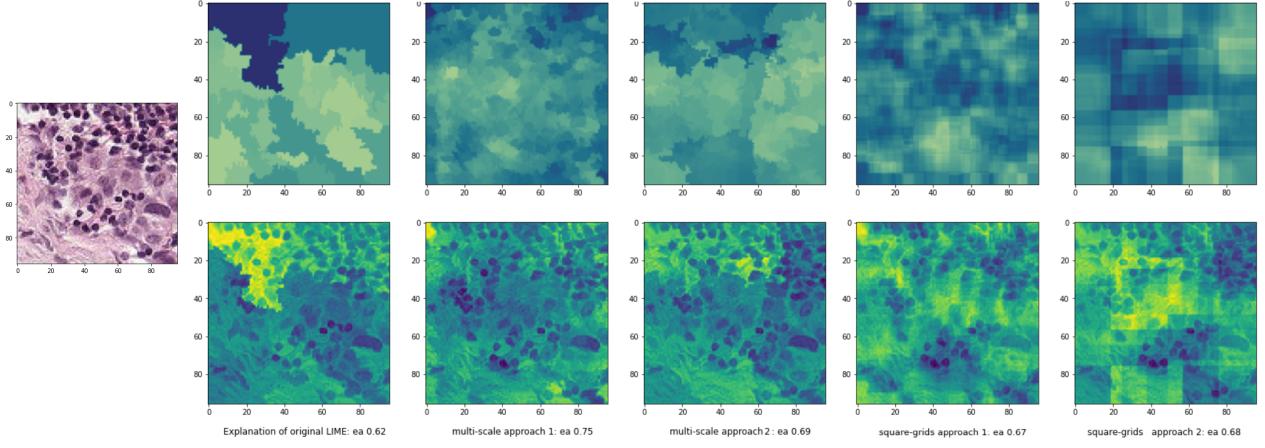


Figure 5. True positive sample image 2

words, this paper contributed to the extension of LIME with local explanations obtained from different scales, in addition to automatic weighting approaches to produce heatmaps of different levels as visual explanations.

Since LIME is sensitive to superpixel segmentation, we segment an image into the desired number of superpixels at each level, and then apply LIME and fit a linear model to achieve a visual explanation on heatmaps that represent how strong each patch is correlated with the classifier decision. In this multi-scale framework, an input image is first segmented into large patches, via coarse-scale segmentation at the first level, and the patches will get smaller at the last level. Hence, through this multi-scale scheme, LIME will focus on the same patch at coarse to finer scales to interpret how important each piece of an image is toward the classifier decision.

We proposed parameter-free automated weighting approaches to assign weights to the heatmaps based on the difference in the number of superpixels. To be more precise, if the number of superpixels is close at two consecutive levels, they might probably produce similar results. Therefore, the explanations obtained from these levels will get low weights. On the opposite, a large gap between the number of superpixels of two levels will achieve various explanation results. The results show this multi-scale scheme of LIME achieves higher accurate results compared to the square-grids segmentation, presented in the literature, as the objects' boundaries are preserved in this approach and the explanations are more localized. Furthermore, the quantitative results indicate that LIME can be extended to a range of scales, and the accuracy of multi-scale explanations would be superior

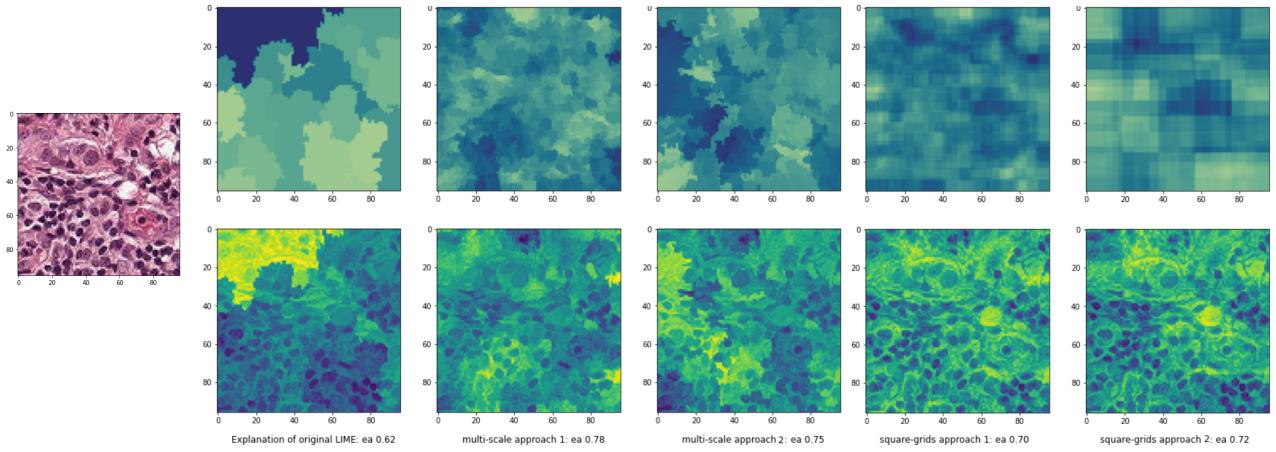


Figure 6. True positive sample image 3

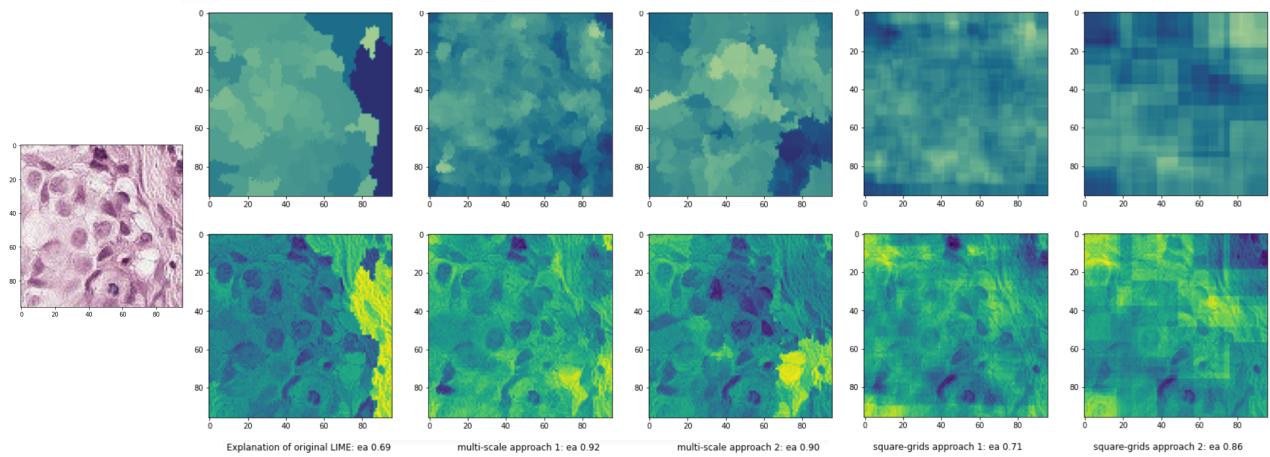


Figure 7. True positive sample image 4

or at least comparable to the original LIME. Clinical expert knowledge would be required to further assess and interpret the results. Scale weights provided by experts could further improve the accuracy of the results, where higher weights would be associated with scales in which an abnormality is most expected.

ACKNOWLEDGMENTS

This work has been supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

- [1] Dimitriou, N., Arandjelović, O., and Cai, P. D., “Deep learning for whole slide image analysis: an overview,” *Frontiers in medicine* **6**, 264 (2019).
- [2] Zahia, S., Zapirain, M. B. G., Sevillano, X., González, A., Kim, P. J., and Elmaghhraby, A., “Pressure injury image analysis with machine learning techniques: A systematic review on previous and possible future methods,” *Artificial intelligence in medicine* **102**, 101742 (2020).

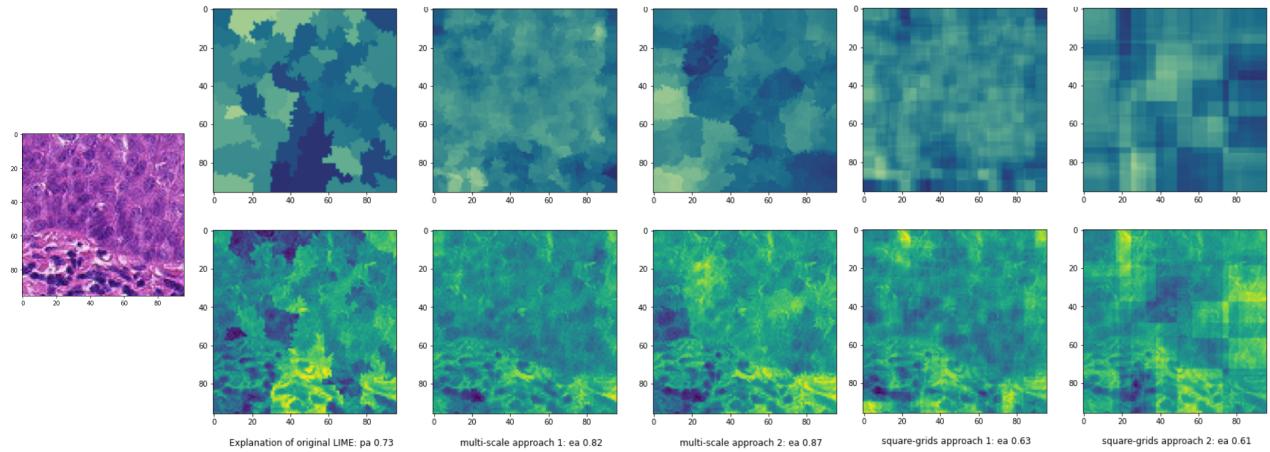


Figure 8. True positive-sample image 5

- [3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I., “A survey on deep learning in medical image analysis,” *Medical image analysis* **42**, 60–88 (2017).
- [4] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” *Cell* **172**(5), 1122–1131 (2018).
- [5] Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T., “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in bioinformatics* **19**(6), 1236–1246 (2018).
- [6] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).
- [7] Weigelt, B., Geyer, F. C., and Reis-Filho, J. S., “Histological types of breast cancer: how special are they?,” *Molecular oncology* **4**(3), 192–208 (2010).
- [8] Aswathy, M. and Jagannath, M., “Detection of breast cancer on digital histopathology images: Present status and future possibilities,” *Informatics in Medicine Unlocked* **8**, 74–79 (2017).
- [9] Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., and Ahmed, S., “On interpretability of deep learning based skin lesion classifiers using concept activation vectors,” in [2020 international joint conference on neural networks (IJCNN)], 1–10, IEEE (2020).
- [10] Ribeiro, M. T., Singh, S., and 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], 1135–1144 (2016).
- [11] van der Linden, I., Haned, H., and Kanoulas, E., “Global aggregations of local explanations for black box models,” *arXiv preprint arXiv:1907.03039* (2019).
- [12] ÇİÇEK, İ. B., KÜÇÜKAKÇALI, Z., and YAĞIN, F. H., “Detection of risk factors of pcos patients with local interpretable model-agnostic explanations (lime) method that an explainable artificial intelligence model,” *The Journal of Cognitive Systems* **6**(2), 59–63.
- [13] Shortliffe, E. H. and Sepúlveda, M. J., “Clinical decision support in the era of artificial intelligence,” *Jama* **320**(21), 2199–2200 (2018).
- [14] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al., “Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery,” *bioRxiv* , 206540 (2017).
- [15] Hameed, Z., Zahia, S., Garcia-Zapirain, B., Javier Aguirre, J., and María Vanegas, A., “Breast cancer histopathology image classification using an ensemble of deep learning models,” *Sensors* **20**(16), 4373 (2020).

- [16] Dikopoulou, Z., Moustakidis, S., and Karlsson, P., “Glime: A new graphical methodology for interpretable model-agnostic explanations,” *arXiv preprint arXiv:2107.09927* (2021).
- [17] Ribeiro, M. T., Singh, S., and Guestrin, C., “Model-agnostic interpretability of machine learning,” *arXiv preprint arXiv:1606.05386* (2016).
- [18] Hu, L., Chen, J., Nair, V. N., and Sudjianto, A., “Locally interpretable models and effects based on supervised partitioning (lime-sup),” *arXiv preprint arXiv:1806.00663* (2018).
- [19] Kumarakulasighe, N. B., Blomberg, T., Liu, J., Leao, A. S., and Papapetrou, P., “Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models,” in [*2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*], 7–12, IEEE (2020).
- [20] Zafar, M. R. and Khan, N. M., “Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems,” *arXiv preprint arXiv:1906.10263* (2019).
- [21] Nazeri, K., Aminpour, A., and Ebrahimi, M., “Two-stage convolutional neural network for breast cancer histology image classification,” in [*International Conference Image Analysis and Recognition*], 717–726, Springer (2018).
- [22] Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., and Costa da Silva, E., “Local interpretable model-agnostic explanations for classification of lymph node metastases,” *Sensors* **19**(13), 2969 (2019).
- [23] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S., “Slic superpixels,” tech. rep. (2010).
- [24] Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M., “Rotation equivariant cnns for digital pathology,” in [*International Conference on Medical image computing and computer-assisted intervention*], 210–218, Springer (2018).
- [25] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J. A., Hermsen, M., Manson, Q. F., Balkenhol, M., et al., “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” *Jama* **318**(22), 2199–2210 (2017).
- [26] Veeling, B., “The patchcamelyon (pcam) deep learning classification benchmark,” (2019).