Developing the sensitivity of LIME for better machine learning explanation

Eunjin Lee [a], David Braines [ab], Mitchell Stiffler [c], Adam Hudler [c], Daniel Harborne [b]
[a] Emerging Technology, IBM Research, Hursley Park, Winchester, UK; [b] Crime and Security Research Institute, School of Computer Science and Informatics, Cardiff University, Cardiff, UK; [c] United States Military Academy, West Point, New York, USA

ABSTRACT

Machine learning systems can provide outstanding results, but their black-box nature means that it's hard to understand how the conclusion has been reached. Understanding how the results are determined is especially important in military and security contexts due to the importance of the decisions that may be made as a result. In this work, the reliability of LIME (Local Interpretable Model Agnostic Explanations), a method of interpretability, was analyzed and developed.

A simple Convolutional Neural Network (CNN) model was trained using two classes of images of "gun-wielder" and "non-wielder". The sensitivity of LIME improved when multiple output weights for individual images were averaged and visualized. The resultant averaged images were compared to the individual images to analyze the variability and reliability of the two LIME methods.

Without techniques such as those explored in this paper, LIME appears to be unstable because of the simple binary coloring and the ease with which colored regions flip when comparing different analyses. A closer inspection reveals that the significantly weighted regions are consistent, and the lower weighted regions flip states due to inherent randomness of the method. This suggests that improving the weighting methods for explanation techniques, which can then be used in the visualization of the results, is important to improve perceived stability and therefore better enable human interpretation and trust.

Keywords: Interpretability, Explainability, LIME, Black-box, Machine Learning

1. INTRODUCTION

Machine Learning has made its way to the forefront of technological advancements, from voice recognition AI in the commercial market to record-keeping in the intelligence community[1]. Scientists and researchers alike have been using machine learning as a means to emulate or extend human decision-making in various domains[2]. However due to the black-box nature of the technology, there is little insight readily available into the computation or reasoning behind the outputs put forth by these machine learning models. Such insight is required to explain how a model has classified an image or text. This is especially important if the model is used in high risk or complex situations, where a human user needs to have high confidence in the machine-generated predictions in order to make a decision[3].

Explainability - asking the 'why' to a model's output – is one area in machine learning that can provide more confidence in a model's decision-making ability[4]. Trust in a model's behaviour is essential if and when we choose to give machines human-supporting roles or wish to interact with them as a reliable team member in a decision-making context.

This paper builds on initial results informally reported in[5], and focuses on the Local Interpretable Model-Agnostic Explanations technique (LIME)[6]. LIME provides an explanation for image and text classification models by approximating the underlying model through perturbation of elements of the original input image or text. When dealing with images LIME first divides up the input image into multiple interpretable components. A data set of perturbed images are created by removing different regions of the original input image, and then each perturbed image is passed through the model to get the probability of classification (Figure 1) The technique generates saliency maps (heatmaps) illustrating which areas of the input image are important to the classification process, in an attempt to provide an 'explanation' of how the underlying model is behaving[7].

In this paper, we analyze the output explanation images that LIME generated when using a simple Convolutional Neural Network (CNN) model. The underlying CNN was trained using a set of 1,600 images separated into these two classes of "gun wielder" and "non-wielder" to determine whether an input image has a gun wielder or a non-wielder present. We analyze LIME as an explanation technique without any modification, as well as presenting a method to improve the sensitivity of the output visualizations to support improved human interpretability.

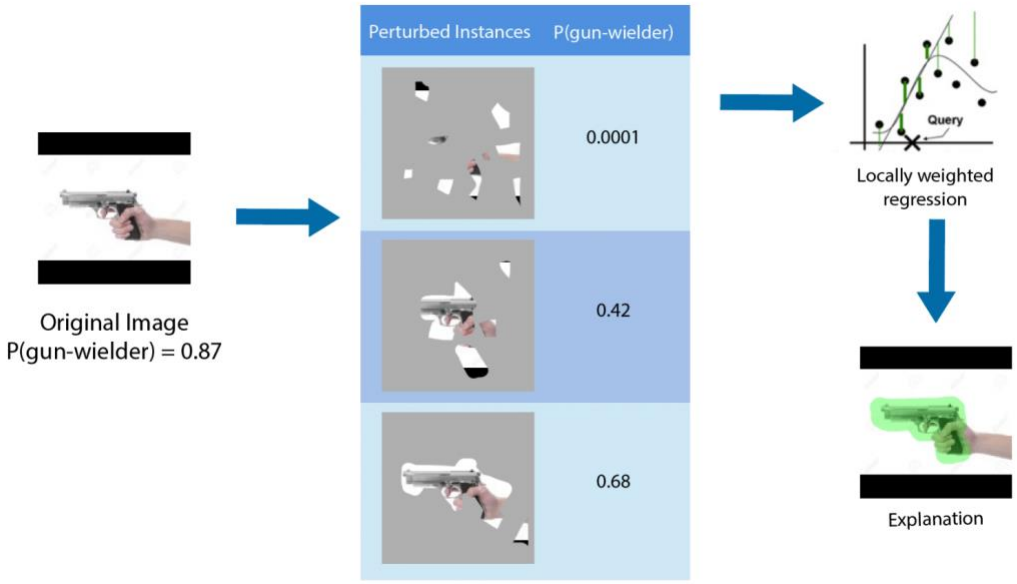

Figure 1. The LIME technique creates multiple different perturbed images in order to calculate a final explanation image in the form of a saliency map

2. ANALYSIS OF THE CORE LIME TECHNIQUE

Using the code provided on the public GitHub repository for the LIME project[6], an assessment of the capabilities for the LIME technique was first studied. The core functionality provided by LIME divides of the input image into regions, which are then assigned saliency weights that form the basis of the resultant generated explanation. These saliency weights are computed based on the degree to which the presence or absence of that region of the image affects the classification result of the underlying model[8].

The Core LIME technique (unmodified code from the Github repository) was configured to generate explanations of the CNN gun-wielder classification model. A photo of a gun in someone's hand was passed through LIME multiple times, i.e. the same input image was presented to LIME and the output explanation image was recorded for later comparison. Figure 2 shows three separate explanations that LIME produced for this single input image. The green regions indicate that the region is weighted positively for the classification of a gun-wielder and the red regions indicate that the region weighted positively for the classification of non-wielder. The non-colored areas' weights did not pass the threshold to either classification. The threshold was set to 0.01 for this initial analysis, i.e. if the weight was above 0.01, the region was colored green and if the weight was below -0.01, the region was colored red.

An informal visual analysis of the output visualizations showed that LIME would color regions erratically for different iterations of the same input image. Sometimes the regions containing the gun would be weighted as positive for a gun-wielder classification and other times be weighted and colored as positive for the non-wielder classification. Such high visual variability in the explanation images may lead to a perceived instability of the approach, potentially undermining user confidence in the explanations.

To further explore this issue, the numerical values of the output weights were also collected. Table 1 shows a sample of the output-calculated weights of two of the explanation images, which LIME uses to assign an output color. Comparing the calculated weights for the regions with the visual outputs of the explanation suggests that the color assignment of the regions may be too sensitive to small changes in the weight calculations of the image. For example, Region 10 for images 2.1 and 2.2 are assigned opposite colors, since the weights are above the 0.01 and below the -0.01 threshold. However, the absolute difference between the values is only 0.02, and yet they are both colored in the explanation images which implies to the user that they have a definite and meaningful contribution to the classification, but in fact they are extremely closely matched.

Table 1. Calculated output weights for 7 regions for images 2.1 - 2.3

Image	Region 1	Region 2	**Region 3**	Region 4	**Region 8**	Region 9	Region 10
(2.1)	-0.05713165	0.05079833	**-0.31549571**	0.01499846	**-0.01942119**	0.06441398	0.01049571
(2.2)	-0.09951557	-0.09344236	**-0.31167949**	-0.02786685	**0.54578947**	0.07719415	-0.01167949
(2.3)	0.00512246	0.03044236	**-0.00293567**	0.05392633	**0.57944333**	-0.04223455	0.01293567

Additionally, the visual saliency map produced by LIME uses a simple binary coloring system for the regions and does not visually reflect the magnitude of each weight assigned to the regions. All regions with positive weights were given a green color and all regions with negative weights were given a red color regardless of their relative weight. For example, the weights for Region 8 for images 2.2 and 2.3 are significantly higher than the other weights, but it is colored in the same manner as the weaker positive weights.

This analysis of the core LIME technique suggested that there may be better ways to process and present the data to the end user. The following sections of this paper outline a method to reduce the noise of the regions to improve the hyper-sensitivity and presentation of the LIME visualizations.

Figure 2. Output explanation images of the gun wielder input image for three separate output explanation images generated by LIME. The boundaries corresponding to Table 1 are shown and labeled.

3. METHODOLOGY

In order to improve the explanations produced by LIME, three aspects of the core technique were modified. First, the input variables of the program were fixed in order to achieve repeatability of the explanation data and visualization (in core LIME they are randomly generated). Secondly, a post-LIME data processing script was developed to further process the output data for a given explanation image. Finally, the way in which the regions were colored was modified in order to reflect the true significance of that region based on the different magnitudes of weights.

3.1 Input Variables for LIME

LIME uses the *random_seed* variable to segment the image into regions. The default is set to 'None', which means a random integer between 0 and 1000 is used in the segmentation algorithm. LIME was initially generating different numbers of regions for the same image in a random manner. To ensure repeatability and consistency for the experiment, the *random_seed* was set to an arbitrarily chosen value of 100.

Alongside the *random_seed* variable, LIME accepts four variables as input for each explanation:

1) *num_sample* - The neighborhood size around a fixed point selected by LIME
2) *num_features* - The maximum number of regions LIME will take into account when explaining an image
3) *min_weight* - The threshold at which the absolute value of a region weight is considered 'important' in relation to a model's prediction - unimportant regions are not colored in the output image
4) The input image

The values that we set were: *num_samples* to 100, *num_features* to 300, and *min_weight* to 0.03 for each input. The values of *num_samples* and *num_features* were set arbitrarily, based on some initial experimentation, to produce a consistent set of comparable output regions for this experiment. The *min_weight* was increased to reduce the sensitivity of the color assignment.

3.2 Data Processing

A Python based script was developed to aggregate multiple output sets of numeric explanation data from LIME for a given input image. For a single input image, the LIME algorithm was applied 30 times and the numeric output data was collected and processed. Alongside the default outputs from LIME for the individual iteration, the script also calculated the following data sets for a single input image:

1) A set of the weights for each region
2) A set of the average weights (calculated per set) for each region
3) A set of the standard deviations (calculated per set) of each region
4) A set of 'significant' regions – regions with an average value greater than two times the standard deviation

The calculation of the significant regions was used as a method to reduce the noise coming from the insignificant weights and to improve the sensitivity of the coloring for the regions. The averaging of multiple weights for the same region cancels out small fluctuations of weights. Comparing the region's average to its standard deviation ensures that only the regions with a 'significant' weight will be acknowledged.

3.3 Improving the Visualizations

As mentioned previously: the unmodified LIME code produces a single explanation image with each regions colored green or red (or left uncolored) depending on whether the weight is above or below the threshold value (min weight). As

we saw in Section 2, this produced unclear and seemingly unstable explanations. Using the data calculated from the method outlined in Section 3.2, the Python script produced additional explanation images:

1) An explanation image for the average weights of each region
2) An explanation image coloring only the 'significant' regions of the image
3) An explanation image showing the standard deviations for each region
4) An explanation image with an alternative coloring scheme to illustrate the weight values

The opacity of the colored region was reduced for all the additional images produced. The alternate coloring method assigns lighter and darker shades of green/red depending on the weight of the region.

4. RESULTS

The new results based on our visualization improvements indicated that LIME consistently designates a few regions within the same image as 'significant' across multiple explanations. Figure 3 shows the additional output images which convey different explanations of the input image.

Figure 3. Explanation images showing the average weights, significant regions and standard deviation of the aggregation of 30 LIME outputs, for Set 1.

Table 2. The average weights and standard deviation calculations for three sets of the experiment

Set	Region	1	2	**3**	4	**8**	9	10
1	Average Weight	-0.02	0.06	**-0.38**	0.07	**0.46**	0.01	0.01
	Standard Deviation	0.08	0.07	**0.06**	0.06	**0.06**	0.06	0.07
2	Average Weight	-0.01	0.06	**-0.32**	0.07	**0.43**	0.02	0.01
	Standard Deviation	0.05	0.07	**0.06**	0.08	**0.07**	0.07	0.07
3	Average Weight	-0.02	0.05	**-0.35**	0.06	**0.39**	0.04	0.01
	Standard Deviation	0.08	0.06	**0.07**	0.06	**0.07**	0.06	0.08

The Average Weight explanation visualization (Figure 3.1) has less regions colored compared to the original LIME visualizations. This is expected since the small positive and negative fluctuations will cancel out, resulting in many regions with low weights to becoming neutral and therefore not colored. Only the regions with weights higher than the threshold have been colored.

The Significant Regions explanation visualization (Figure 3.2) takes this a step further and only colors regions in which the weight value is great than double its standard deviation. This means that regions which don't have a statistically significant weight get ignored, resulting in an image where only the regions which LIME identifies as significant for the classification are colored. This can also be seen in Table 2, where the average weights for Regions 3 and 8 are consistently high for each set, and they have a low standard deviation. The other regions' weights are averaged to a low number and the standard deviation is often higher than the average value. The resulting explanation image is a less cluttered visualization than the core LIME technique. In Figure 3.2 the one resulting region deemed significant to the gun-wielder classification is Region 8, which highlights the tip of the gun.

The Standard Deviation image illustrates the magnitude of the standard deviation of the different regions; the higher the opacity of the yellow color, the higher the standard deviation is in that region. There is slight variation in the standard deviation across the regions, however they are all relatively close in value, even for the significant regions.

The alternative color scheme added more information to the explanation image (Figure 4); the darker the color of green or red, the more positive or more negative the weight was. This is presented alongside the individual LIME explanation in Figure 4.

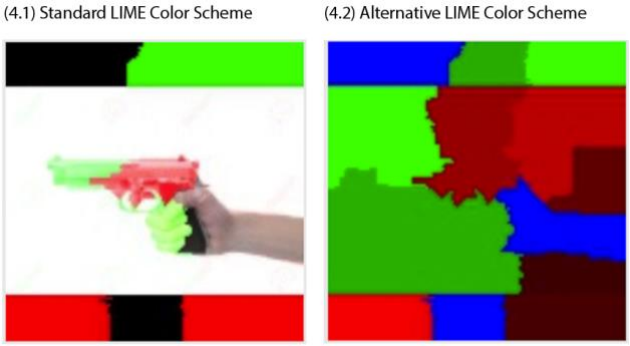

Figure 4. Comparing the explanation images colored with the standard LIME color scheme and the alternative color scheme

5. DISCUSSION

Contrary to the perceived visual instability of the original LIME explanations, our method shows that LIME does indeed calculate the most significant regions of an image consistently. The standard deviation analysis also suggests that LIME calculates the weights with a consistent variance. This is expected due to the inherent randomness of the technique, but is undermined by the simplistic default vizualisation technique which undermines user confidence.

On top of this, the processing of output data in sets and the production of aggregated images created clearer explanation images compared to the single explanation image produced by core LIME. The resultant Average Weight, Significant Areas and Standard Deviation explanation images (Figure 3), especially when presented in conjunction with each other, provide the user with a more intuitive illustration of the classification. The ability to see the 'meaning' behind the region color across the images also enables the potential for a deeper understanding of the processing by the human user.

The erratic flipping of the region colors created using core LIME was shown to be a combination of a low threshold value needed to assign the region a color, in conjunction with the hyper-sensitivity of the binary coloring scheme. For very small changes in the calculated weight, the regions would flip to the opposite color which led to confusing explanation images. By averaging the weights across multiple LIME explanation data sets and presenting an aggregated visualization, we have shown that this removes the perceived noise fluctuations which were resulting in the color flipping of the regions. Modifying when the colors are assigned and changing the opacity or color scheme to indicate relative weight could significantly improve the user perception of the explanation. The ability to identify and present Significant Regions allows the user to interpret the explanation image with confidence that only the important regions are being colored.

LIME gives us insight into which regions of an image appear to have the most impact on classification in a given model, but there is still little understanding concerning the most useful way for a human to interpret LIME explanations. A weight assigned to a region does little to indicate what about that region is contributing to a model's prediction. The highlighted tip of the gun in Figure 3.2 may be interpreted by a human that the classifier is looking for gun-like objects, but it may not be and there is no way to get that information from the LIME explanation or indeed the underlying CNN model. While human interpretation may still be uncertain, our finding that each image consists of approximately three regions that LIME identifies as being either very supportive or unsupportive of the model prediction indicates that LIME does have the ability to serve as a consistent explanation mechanism for human users.

Though LIME is able to produce a stable output as it consistently indicates a few areas as significant, we are reminded of the differences between a machine learning model and human perception of an image. The ability for human users to understand and accept explanations may be undermined by the apparent irrelevance of the region definitions.

6. CONCLUSION

When observing two different core LIME explanations of an input image using the default visualization technique it often appears that there is little stability in the explanations. However, our analysis of the averages and standard deviations of the regional weights generated by LIME demonstrates that the explanations being generated are relatively stable but the failure to convey the weight of the region in the visualization leads to this perception of instability. However, despite this inherent stability, it is still difficult to assess LIMEs reliability across images and further research is needed, both in terms of a broader image, model and explanation set, as well as from a human user experience perspective.

ACKNOWLEDGEMENT

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255–260 (2015)

[2] J.-C. Pomerol et al., "Artificial intelligence and human decision making," European Journal of Operational Research, vol. 99, no. 1, pp. 3–25 (1997).

[3] N. Pennington and R. Hastie, "Reasoning in explanation-based decision making," Cognition, vol. 49, no. 1-2, pp. 123–163 (1993).

[4] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, "Building explainable artificial intelligence systems," in AAAI, pp. 1766–1773 (2006).
[5] M. Stiffler, A. Hudler, E. Lee, D. Braines, D. Mott and D. Harborne, "An Analysis of Reliability Using LIME with Deep Learning Models", Annual Fall Meeting of the Distributed Analytics and Information Science International Technology Alliance ,AFM DAIS ITA, (2018).
[6] M. T. Ribeiro , Lime: Explaining the predictions of any machine learning classifier, <https://github.com/marcotcr/lime> (2 August 2018)
[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction", <https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime> (28 February 2019)
[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "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. ACM, 2016, pp. 1135–1144.