An Analysis of Reliability Using LIME with Deep Learning Models

Mitchell Stiffler*, Adam Hudler*, Eunjin Lee†, Dave Braines†, David Mott†, Daniel Harborne‡
*United States Military Academy, West Point, New York, USA
†Emerging Technology, IBM Research, Hursley Park, Winchester, UK
‡Crime and Security Research Institute, School of Computer Science and Informatics, Cardiff University, Cardiff, UK

***Abstract*—As machine learning solutions become more commonplace it is essential that human users are able to trust the outputs of these systems. Whilst many machine learning systems show great potential in terms of their ability to perform classification or prediction tasks, they are often undermined by their inability to provide explanations as to why the proposed outcome was chosen. These "black box" systems are inherently unable to provide such explanations due to their complex internal composition so novel techniques to extract or generate such explanations are needed.**

Much research is now focused around identifying these explainability techniques and functions for machine learning systems. Tools and frameworks such as LIME (Local Interpretable Model-Agnostic Explanations) are now available to be used to provide explanations and to check whether the machine learning model is actually detecting relevant features. Due to the approach taken by tools such as LIME there appears to be inherent uncertainty, with potentially different (and often conflicting) explanations being generated for any given machine learning outcome.

In this paper we investigate LIME in a simple image classification task and asses the consistency of the explanations generated. Against this baseline we then implement a number of simple algorithms to investigate whether the aggregation of multiple explanations to provide a single computed summary explanation can improve the stability (and therefore usefulness) of the explanations. This suggests that some of the apparent uncertainty experienced by human users is due to the way the results are visualized.

Keywords–Machine Learning, Deep Learning, Interpretability.

I. Introduction

Machine Learning (ML) 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, 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 enable us to explain why a model has performed a task such as classifying an image, especially in high risk of complex situations where a human user needs to have high confidence in 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]. In other words, we must be able to trust a model's reasoning if we choose to give machines human-related roles or wish to treat them as a credible team member in a decision-making context.

Our research into the stability of explainability is undertaken using LIME (Local Interpretable Model-Agnostic Explanations)[5]. In our experiment LIME is used to generate saliency maps of specific regions showing which parts of the image affect how the ML model reaches a classification for a given test image. For our research, we examined the explanations that LIME generated when using a simple CNN (Convolutional Neural Network) machine learning model to determine if pictures did or did not have a person wielding a gun present in the image. The underlying CNN was trained using a batch of images separated into these two classes of "gun wielder" and "non-wielder". This paper reports the results of our research into LIME as an explanation mechanism in this simple context, and the potential implications.

In section II we outline the methodology used during this assessment, section III provides a short summary of the results, the potential implications of which are covered in section IV. Potential future work is outlined in section V with the conclusions in section VI.

II. Methodology

The cadets leading this experimental analysis first undertook an assessment of the relevant capabilities for explanation techniques, focusing on LIME as a useful candidate [6]. The core explanation provided by LIME in this experiment is based on the division of the input image into regions which are then assigned saliency weights which form the basis of the eventual 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.

Our initial casual observation was that LIME appears to be unstable. For example, through multiple iterations we observed that LIME would successfully generate saliency maps for the same input image, but often each explanation would weight the regions differently from the other explanations for the same image. Such variability in the explanation could lead to a perceived instability of the approach, potentially undermining user confidence in the explanations.

Additionally, the visual output 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 are given a green color and all regions with negative weights are given a red color regardless of their relative strength/weight. More specifically, as shown in Figure 1 the perceived "weight" of the region for human observers is driven by the intensity of the color (red/green), however the intensity relates purely to the darkness of the underlying image color. In Figure 1 regions 1, 14 and 15 are especially intense green color, with regions 0, 2 and 15 the most intense red. To the casual observer this could suggest that these are the strongest red/green regions, but the intensity is purely down to the darkness of the underlying image (see Figure 2). The statistics in Figure 3 show that these six regions always have a low relative weight compared to regions 3, 4, 7 and 12 which have the highest relative weights and are rendered using subtle green/red color due to the lightness of the underlying background.

In order to analyze the stability of the LIME process we aggregated multiple explanations of the same image by submitting multiple identical images for explanation and taking the average and standard deviation of the weights for each region. One issue we encountered was that LIME occasionally generates a different number of regions for an input image due to a random seed which serves as input for its segmentation algorithm. In order to ensure we would be able to calculate a region's average weight after multiple iterations, we therefore held this random seed at an arbitrarily chosen value *(101)* in order to ensure repeatability of region generation. Having resolved the region variation issue we then designed a simple framework to summarize necessary information to inform our research, including each region's average and standard deviation.

Although the CNN classification model was trained on many hundreds of images, for our research we collected data by sampling nine gun wielder images and seven non wielder images.

LIME accepts four variables as input for each explanation:

1) The neighbourhood size around a fixed point selected by LIME *(num_samples)*
2) The maximum number of regions LIME will take into account when explaining a picture *(num_feautures)*
3) The threshold at which the absolute value of a region weight is considered "significant" in relation to a models prediction - regions that do not meet this threshold are not color coded in the output *(min_weight)*
4) An input image

For these variables we set *num_samples* to *100*, *num_features* to *300*, and *min_weight* to *0.01* for each input image we tested. This was essential in order to keep our variables constant so that later comparison of our results has a stable basis. These values were arbitrarily chosen after a short period of trial and error with LIMEs explanation algorithm based on our subjective opinion that this combination most consistently provided a reasonable explanation basis for an image.

In order to compute the aggregation of the explanations, our algorithm calculates the following information from the multiple explanations of a single image (as noted below, we aggregated over sets of explanations of the same image):

1) a collection of the weights for each region (which were unit-less)
2) a collection of the average weights (calculated per set) for each region
3) a collection of the standard deviations (calculated per set) of each region
4) a generated image which highlighted regions whose weights exceeded the *min_weight* threshold (red for negative values and green for positive values)
5) another image (which followed the same method of highlighting mentioned previously) for the average weights of each region
6) an image which displayed the boundaries for each region
7) an image that highlighted regions whose standard deviation exceeded a predefined minimum threshold

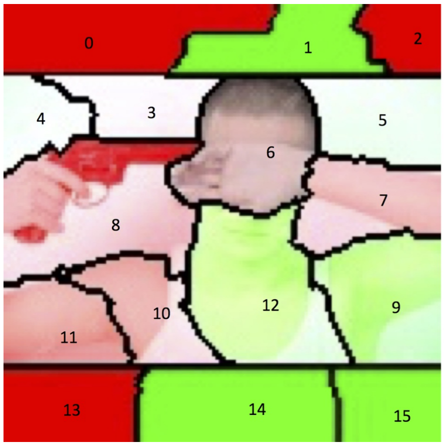


Figure 1: Regions created by LIME, with saliency shown by red/green color (regions 1, 9, 12, 14 and 15 are green)

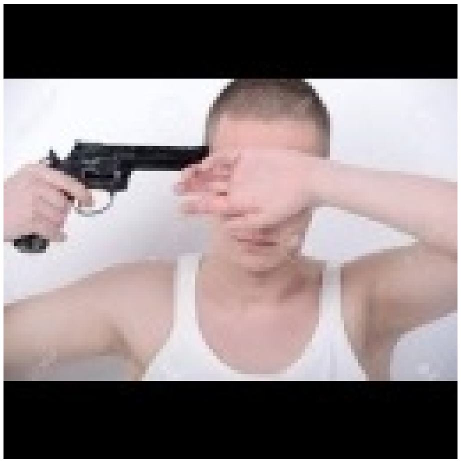


Figure 2: An original image submitted for classification and explanation

For each original image we collected three sets of 30 explanations generated by LIME. We compared all three sets of data per picture and noted the similarities and differences between each set. We then checked to see if our findings were consistent across each image.

One of the main limitations in our methodology was that the manner in which the average weight for each region was calculated is dependent on the number of regions that LIME divides a picture into remaining constant. In order to ensure consistency, we entered in a fixed value in for the random seed as the input for LIME's segmentation algorithm as previously mentioned. An alternative approach to generating average weights is suggested in section V.

III. Results

The trained CNN model was able to correctly classify the gun-wielders and non-wielders in the majority of images. Our results then show that LIME consistently designates a few regions within the same image as "significant" (by assigning a region a weight with an extreme value, i.e. very negative or positive) across numerous explanations. Figures 2 and 1 show the original and explained images, respectively, using LIME. Corresponding results for this gun-wielder image are reported in figure 3, which highlights three consistent regions (3, 4, and 12) that LIME repeatedly classified as significant. In one case (Set 3) LIME also reported region 7 as being significant.

In order to label regions as significant, we only recorded each region that had an average absolute value greater than two times its own standard deviation. We made this our threshold because anything out of the two standard deviation mark would have only a 5% chance of occurring (2.5% at each end), thus making our likelihood of attaining only the significant regions very high. We reached our conclusions by running 30 explanations (one set), three times each. After gathering three sets of explanations for a picture, we checked to see if these few regions remained relatively consistent which proved to be the case across the vast majority of our images; indeed the standard deviation of the weights were found to be low[1]. This shows that the list of significant regions proposed by LIME is stable, despite our inability to understand the reasoning behind the models output. Additionally, the standard deviation of the weight is broadly consistent across all regions in an image which indicates to us that LIME is relatively precise when it comes to assigning weights to regions.

Furthermore, our results show that the few regions that LIME did indicate as most significant did not always contain the same contents across all gun wielder and non-wielders images, such as a hand, a facial feature, or a chair in the background[2]. This tells us that even 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. For example, the ability for human users to understand and accept explanations may be undermined by the apparent irrelevance of the region definitions.

On some occasions, we found that LIME did not offer any explanation and chose not to assign weights to any of the image regions. Despite this problem being uncommon, it shows that such techniques are subject to issues and errors in their current state of maturity.

IV. Discussion and Challenges

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 very little understanding concerning the most useful way for a human to interpret LIME explanations. A weight assigned to a region does very little to actually indicate what about that region is contributing to a model's prediction. Ideally LIME explanations should be simple enough so that any human can reach a conclusion without thinking critically. However LIME can only determine and visualize the contents of the underlying machine learning model (in this case the CNN); if this model does not encode the "ideal" feature (for example that there is a gun) then LIME is not going to be able to reveal it in an explanation. 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 not supportive of the model prediction indicates LIME does have limited ability to serve as a consistent explanation mechanism for human users.

Statistically, 95% of the weights computed fall within two standard deviations of the average weight, and since each region had nearly identical standard deviations, LIME was equally confident in explaining each region for every image we tested[3].

V. Future work

For this given experiment, model and dataset more trials should be taken to assess LIME's consistency across all gun wielder, non-wielder, and other images. Because the analysis methodology required the generation of 90 total explanations for each image, we are confident in our analysis for the limited number of images reported here. Additionally, our results for each image were similar in that LIME consistently computed the same regions as most significant, with equal amounts of standard deviation. This is interesting because although LIME emphasizes only a few regions as significant, it produces nearly the same standard deviation for each region in any given image. This suggests that LIME computed each weight with the same confidence level for each region. Further trials across a wider set of images would provide better data to support a greater understanding of LIMEs consistency in generating standard deviations for an image. It would also be useful to determine the sources of randomness in the LIME and CNN algorithms.

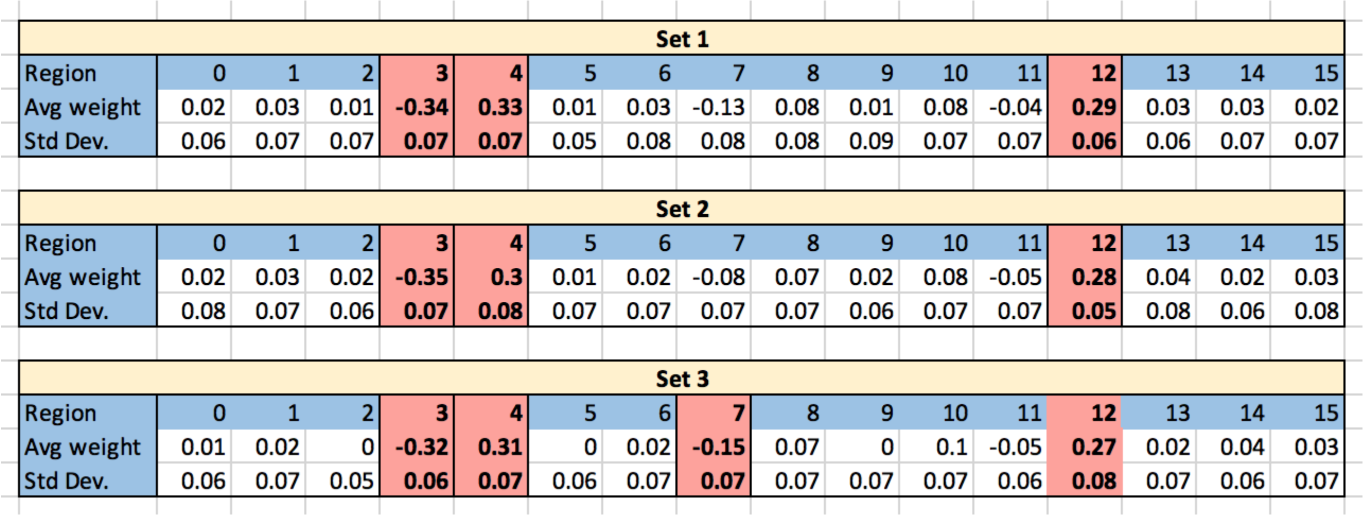
Another improvement that could be made on our work would be to observe how LIME works without forced consistency for the number of regions. Different images have different complexities and therefore simple images might benefit most from a small number of regions, with complex images needing more regions. In order to accomplish this, we would need to make changes to our method of averaging region weights, perhaps using an initial LIME analysis to compute the ideal number of regions for a given image and then using that fixed value for each of the iterations for that image.

We also observe that the default visualization mechanism for LIME uses a simple binary (red/green) coloring system with no ability to visually communicate the relative region weights. Using a color intensity or transparency scale to indicate relative weight could significantly improve the user perception of the explanation. For example, when a user sees inconsistencies between explanations for a given image they are less likely to assume instability, and therefore lose confidence, if the inconsistent regions are lower weighted and are therefore less visually apparent than the dominant (stable) regions.

[1] Note that "low" is used loosely in this context and not formally/statistically defined. In future work such terminology will be formalized and made explicit

[2] This may suggest that the underlying CNN is not picking up on the relevant features (which were very small in some of the images)

[3] This suggests that there is a relatively small amount of randomness built into the CNN and LIME algorithms, since if there were no randomness, the standard deviations would be 0



Set 1

Region	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avg weight	0.02	0.03	0.01	**-0.34**	**0.33**	0.01	0.03	-0.13	0.08	0.01	0.08	-0.04	**0.29**	0.03	0.03	0.02
Std Dev.	0.06	0.07	0.07	**0.07**	**0.07**	0.05	0.08	0.08	0.08	0.09	0.07	0.07	**0.06**	0.06	0.07	0.07

Set 2

Region	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avg weight	0.02	0.03	0.02	**-0.35**	**0.3**	0.01	0.02	-0.08	0.07	0.02	0.08	-0.05	**0.28**	0.04	0.02	0.03
Std Dev.	0.08	0.07	0.06	**0.07**	**0.08**	0.07	0.07	0.07	0.07	0.06	0.07	0.07	**0.05**	0.08	0.06	0.08

Set 3

Region	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Avg weight	0.01	0.02	0	**-0.32**	**0.31**	0	0.02	**-0.15**	0.07	0	0.1	-0.05	**0.27**	0.02	0.04	0.03
Std Dev.	0.06	0.07	0.05	**0.06**	**0.07**	0.06	0.07	**0.07**	0.07	0.07	0.07	0.06	**0.08**	0.07	0.06	0.07

Figure 3: Results from generating 90 total explanations, divided into three sets of 30

VI. Conclusion

When observing two different LIME explanations of an 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.

Acknowledgment

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Numbers W911NF-06-3-0001 and 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 U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright 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*, 2006, pp. 1766–1773.

[5] (2018) Lime: Explaining the predictions of any machine learning classifier (github). [Online]. Available: https://github.com/marcotcr/lime (Visited on 2-Aug-2018)

[6] 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.