An approach to enhance machine learning explanations

May 26 2019, by Ingrid Fadelli

The original input image. Credit: Lee et al.

Researchers at IBM Research UK, the U.S. Military Academy and Cardiff University have recently proposed a new approach to improve the sensitivity of LIME (Local Interpretable Model Agnostic Explanations), a technique for attaining a better understanding ofthe conclusions reached by machine learning algorithms. Their paper, published on SPIE digital library, could inform the development of artificial intelligence (AI) tools that provide exhaustive explanations of how they reached a particular outcome or conclusion.

"We believe that AI and machine learning can support and augment human decision-making, but that there is also a necessity for explainable AI," Eunjin Lee, co-author of the original research paper and Emerging Technology Specialist and Senior Inventor at IBM Research U.K., told TechXplore. "Today, decisions made by many machine learning systems are inexplicable, i.e., there's no way for us humans to know how the systems came to those decisions. Our research addresses this issue by investigating how to improve explainability techniques that aim to shed light on the 'black-box' nature of machine learning processes."

LIME is a particularly popular explainability technique that can be applied to many machine learning models. Despite its versatility, it is often seen as unreliable and thus ineffective in providing explanations, also due to the variability in the results it produces. Rather than developing an entirely new explainability technique, Lee and her colleagues set out to identify mechanisms that could enhance LIME explanations.

"We first wanted to look deeper into the instability that other researchers have observed to determine if LIME was really unstable," Lee explained.

"To do this, we tested LIME against our dataset and machine learning model without changing the underlying code. We immediately found that the resulting explanation images varied considerably and did not seem consistent. This is perhaps the point at which many would simply stop using the technique."

Nine image outputs for the unmodified LIME technique. Credit: Lee et al.

When Lee and her colleagues dug deeper into LIME's underlying statistics, they discovered that although the images it generated appeared to be "visually unstable," the default explanation did not take into account all of the statistical information. For instance, the coloring of explanation images was too simple and did not consider the full underlying data (e.g., did not account for techniques such as shading or transparency). This finding partly explains why explanations generated by LIME sometimes fail to convey the certainty of classification to human users.

"It is often the case for dynamic systems, such as the ones we examined in this study, that running numerous tests and investigating average values can prove beneficial," Lee said. "In taking this approach, we realized that the stability of the explanations did improve when considering averaged values and standard deviations over multiple runs rather than just running the explanation once."

In their study, Lee and her colleagues trained a convolutional neural network (CNN) model using two classes of images, namely "gun-wielder" and "non-wielder." They found that LIME's sensitivity improved when several output weights for individual images were averaged and visualized.

The researchers then compared these averaged images to individual images to evaluate the variability and reliability of the two LIME methods (i.e. the traditional method and the one they proposed). They found that traditional LIME techniques, without the adjustments they made, appeared to be unstable due to the simple binary coloring they adopted and the ease with which colored regions flipped when comparing different analyses. Lee and her colleagues also observed that the significantly weighted regions of images were consistent, while the

lower weighted regions flipped states, due to the inherent instability of LIME techniques.

The image highlighted using the average information. Credit: Lee et al.

"Techniques such as LIME show great promise for AI explainability, especially at a time when there are no easy, readily available explanation capabilities for machine learning systems," Lee said. "While the perceived instability is justified, there are techniques that can help mitigate this issue. These techniques have additional computational costs, e.g. running the explanation multiple times which means the user will have a bigger delay in generating the explanation."

The study Lee and her colleagues conducted offers a valuable explanation of some of the factors behind LIME's instability, as observed in past research. Their findings suggest that improving weighting methods for explainability techniques can enhance their stability and lead to more reliable explanations, ultimately fostering greater trust in AI. Future research could identify more advanced numerical techniques to further improve the stability of LIME and other explainability methods while reducing the additional overhead.

"We have an ongoing interest in accountable AI systems that include explanations but also mitigate bias and enhance robustness and transparency," Lee said. "Improving the ability for developers to more easily embed explainability techniques into their AI solutions is a key goal for us. Recently, IBM launched a software service that automatically detects bias and explains how AI makes decisions."

More information: Eunjin Lee et al. Developing the sensitivity of LIME for better machine learning explanation, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (2019). DOI: 10.1117/12.2520149 . www.spiedigitallibrary.org/con … .2520149.short?SSO=1

github.com/marcotcr/lime

Marco Tulio Ribeiro et al. "Why Should I Trust You?", *Proceedings of*

the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (2016). DOI: 10.1145/2939672.2939778 .
dl.acm.org/citation.cfm?doid=2939672.2939778