A Practical Application of Black Box Interpretability

Byungwoo (Peter) Jang[*], Mingu (Jason) Jeong[*], Hevin Na[*], Conner Russell[*],
Dave Braines[†], Richard Tomsett[†], Graham White[†], Dan Harborne[‡]
[*] United States Military Academy, West Point, New York, USA
[†] IBM United Kingdom Ltd, Hursley Park, Winchester, UK
[‡] Cardiff University, Cardiff, Wales, UK

Abstract— In a world environment where technology is advancing at a rapid pace, fluent communication between machines and humans is becoming increasingly relevant. Interpretability through summarisation and explainability are imperative factors in the progression of machine and human communication. Using IBM Watson's visual recognition service, in conjunction with other processes we explore and establish interpretability. Specifically, we looked at identifying cancerous cells from biopsy images through visual analytics. Furthermore, future practical applications, including in different fields, were also investigated, i.e. counter IED operations.

I. Introduction

Using a cognitive computing platform to communicate its inner workings as a program to the human counterpart is an obstacle for establishing a relationship between machines and people. As more closed source platforms begin to emerge, the dearth of knowledge on internal processing of these machines increases. Rather than focusing on how these machines output certain answers, the delivery of information about how and why conclusions were reached is the key to creating a productive environment for both the machine and the person. Communication between the computer and human can be formulated around the ideas of interpretability and post hoc explanation [1]. Throughout this paper, we will use the tackling cancer program, which utilizes Watson's visual recognition platform to classify cancer and blood cells. We use this as a concrete example against which to discuss the extent to which explainability can be relayed to a user even when using a black box program.

II. Interpreting black box decision making

The input-output model more commonly referred to as a black box is characterized by the lack of knowledge of its inner workings. The known attributes of black box processes are restricted to the initial inputs given, an unknown transformation within the black box, and then an output. Even with the limited control of the platform, people are still able to extract relevant data by manipulating the two data sets that they have control over: the inputs and the outputs [2]. Because the processes that the black box take are unknown, transparency cannot be achieved through solely the inputs and outputs [1]. However, summarisation and to an extent, explainability can be achieved. While summarisation relies heavily on the extraction of specific data points of the output, explainability relies heavily on the combination and manipulation of the inputs and outputs to give at best a post hoc justification for the result comning out of the black box. Interpretability in itself is an ambiguous term describing how someone comes to understand a concept.

A. Summarisation

Summarisation is arguably the more important aspect of interpreting a black box, even though it is simply handled by the outcome. While black box platforms usually output a categorical answer, extracting relevant information as well as utilizing inductive reasoning can help in forming other conclusions [2]. Summarisation is the information that is displayed to the user. For example, a black box takes an image of a solid color and outputs the name of the color, say red. Now, we can make reasonable assumptions about the image based on the outcome of the black box i.e. since the image was red, it is not blue. This is specific to summarisation, in that the outcome of a black box is not only restricted to its actual result, but should be *transformed* into what the client wants to know in a clear and concise manner.

B. Explainability - Post Hoc

True explainability cannot be achieved through a black box platform due to the unknown nature of its inner workings; however, we can attempt to make our own deductions based on a post hoc evaluation of our inputs and outputs. A post hoc approach to achieving explainability takes several inputs and compiles data based on their respective outcomes. We use deductive reasoning by taking data samples and applying these rules toward a generalization of how the black box may have produced the outputs that it returns. A post hoc approach towards explainability does not provide clear and precise answers toward true explainability, but can provide an explanation [1]. We focused on explanation by visual examples based on the nature of our black box platform.

III. Practical Applications

This idea of interpretability between human and machine communications was then applied to our current problem: identifying cancerous cells from biopsy images. The program takes a biopsy image then identifies and extracts its individual recognized cells. These extracted images are then processed by Watson's visual recognition platform which is trained through feeding it images of cancerous and blood cells. Because of the black box nature of the Watson service, we do not know what features are learned from the training images. The output is given as a confidence percentage of how Watson classified the extracted cell images based on its learning from the given trained images.

With regards to summarisation, the result the Watson service provides (in confidence percent form) can be organized into a hard answer. For example, if Watson tells us that it is 85% confident a cell image is cancerous, this could be further summarised as "yes, this cell is cancerous" if an 80%+ confidence threshold led to the cancerous decision/conclusion. By creating this 80%+ rule, we can therefore deliver the summarisation state with an option to dive into explainability in that when a human is looking at this summarised output, the machine can also show the percentage reading (85%) and the rule (80%+ = cancer conclusion) as the rationale which led to the summarised answer.

When programming a function to transform the outputs in order to communicate with people, explainability is a trivial task in which we explain the reasoning process. However, when working with a black box, it is difficult to go "inside the box" and therefore justify why our program has decided to classify a cell as cancerous or blood. Therefore, we resorted to a post hoc approach for explainability of the black box. In our Tackling Cancer model, our post hoc approach was to juxtapose all the inputs and analyze the outputs of the black box. The images were aligned based on similarities in color and shape. Based on where each input was placed, we would induce Watson's processes based on either color, shape, or both, contributing to the explainability of the internal black box processing, albeit by post-hoc analysis which may not reflect the actual internal processing.

IV. Challenges

In order to properly explain the outputs being presented by Watson, the training phase is crucial. We discovered that for our Tackling Cancer model, Watson is not being trained to match for color, which is a telltale sign of cancer on the stained biopsy images. The main challenge faced was defining the correct inputs and then manually evaluating the outputs. While we have no control over the outputs because they are a result of the black box programming, we must make sure Watson is processing the correct inputs.

V. Expanded Applications

The ever-developing area of communication and interaction between humans and machines can be practically applied to many different fields. The military would also be able to apply the concepts discussed in this paper and research to modern day operations.

A potential application would be in counter IED operations. Current counter IED ops are conducted by military and police personnel or by unmanned vehicles/robots such as the TALON or the Husky Mounted Detection System (HMDS) [3]. However, both the TALON and HMDS are systems that require human decision-making model; the vehicles only provide visual information through sensors and cameras, leaving all processing to the human operator [3].

Parallel to our experimental Tackling Cancer solution, a program similar to Watson would classify the images from the radars or sensors in real-time. A classification output would be provided, which could possibly be in a confidence interval form (similar to Tackling Cancer). The program could also provide some degree of explanation/justification for the output as well, i.e. due to shape, or heat signature.

The benefit of having this kind of program is, as discussed, it saves time and labor for the human users. The criteria for IED classification could potentially eliminate many false positives if the input data were rich enough, thus alerting the human operator only to the highest likelihood matches, enabling them to focus their limited resources on the ones that are classified as IEDs and therefore most likely to require human review and intervention. The human users can then either confirm or override the decision of the AI. The critical requirement for a system like this is a low level of false negatives (i.e. it doesn't miss real IEDs) with a strong secondary requirement for a low level of false positives (so it doesn't overwhelm the limited resources of the human reviewers.

This of course is not the only application to the military or to the practical world. Further development can turn this idea of transparency in the communication between a computer-generated decision and a human operator into applications such as: looking for abnormalities in building construction or identifying unnatural objects in the human body and many more.

Acknowledgement

This research was sponsored by the U.S. Army Research Laboratory and the U.K. 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 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] Lipton, Z., "The Mythos of Model Interpretability", Cornell University Library, 2017.

[2] Krause, J., Perer, A., Bertini, E., "Using Visual Analytics to Interpret Predictive Machine Learning Models", Cornell University Library, 2016.

[3] Chemring, "Counter IED: Husky Mounted Detection System", niitek.com, 2017.