Exploration of the Effect of Task and User Role on the Evaluation of Interpretability Techniques

D.Harborne (1), **L.Hiley** (1), **A.Preece** (1), **H.Taylor** (1), **C.Willis** (2), **D.Braines** (3), **R.Tomsett** (3), **S.Chakraborty** (4), **S.Julier** (5), **A.Widdicombe** (5) **M.Alzantot** (6). 1:Cardiff University 2:BAE Systems 3:IBM UK 4:IBM US 5:UCL 6:UCLA

Introduction

In many military operations, we must be able to explain how an A.I. algorithm reached a decision. Previously, we argued that **the utility of an explanation depends on** the **nature of the task** being performed **and the role of the agent** consuming the explanation [1].

To explore this, **we have developed a framework of datasets, machine learning models and explanation techniques** which allows for the comparison between a range of explanations in the context of different tasks.

Objectives of the Framework

Build intuition for current and future explanation techniques allowing for innovations in their use and in the creation of novel techniques.

Produce data that can be used to develop metrics to measure utility of explanation techniques and benchmark their resource cost.

Experiment with 'coalitions' of machine learning & explanation services in a dynamic context to satisfy mission requirements with consideration of resource usage.

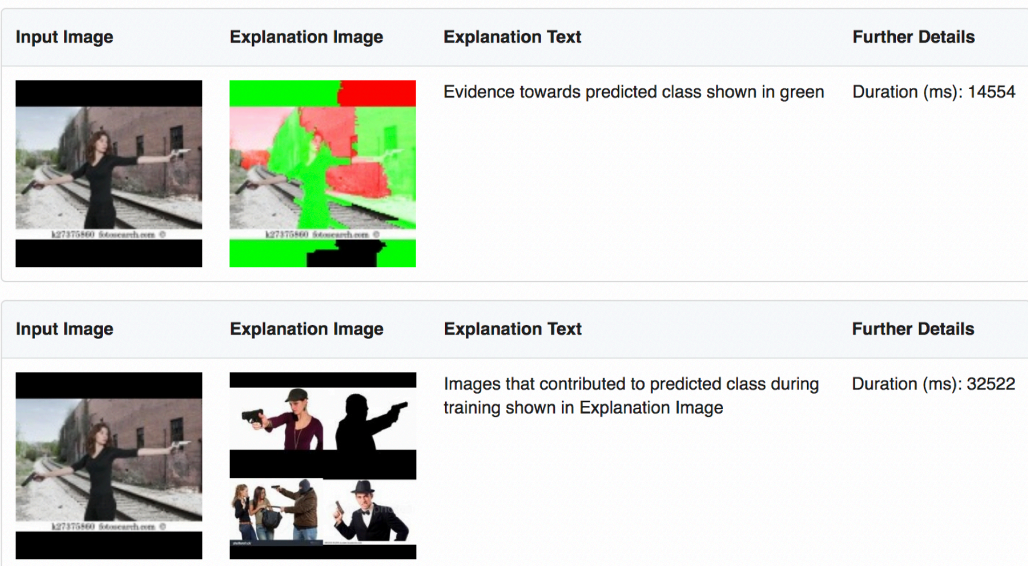

Figure 1: Two explanation outputs generated by the framework for the same input image.

Top – "LIME", Bottom – "Influence Functions"

Architecture of The Framework

The framework has been designed to be modular in nature. Datasets, models and explanations are wrapped in decoupled modules, unifying the input/output signatures of items within each category.

An API has been built to facilitate listing, selecting and using available items in the framework via simple http requests. Interfaces are loosely coupled to the API and therefore can be customized and swapped out entirely to meet the needs of the researchers.

Intended Uses of the Framework

Comparative/sensitivity analyses of explanation techniques we demonstrated the debugging process of generating multiple explanations from the same technique to measure and improve its stability [2].

Empirical studies of explanation visualisations Existing explanation techniques are subject to post-hoc interpretation by the recipient. This can be greatly affected by the visualization and presentation of the explanation and can lead to the recipient projecting their existing assumptions on to the explanation. Future work will explore this issue.

Eliciting stakeholder requirements for application-level interpretability We will use the framework to engage with subject-matter experts to gain insights as to what constitutes a useful explanation for aiding task performance.

References

[1] Tomsett et al. "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems" in 3rd Annual Workshop on Human Interpretability in Machine Learning (WHI 2018)

[2] Stiffler et al. "An Analysis of Reliability Using LIME with Deep Learning Models" in DAIS ITA AFM 2018

Distributed Analytics and Information Science International Technology Alliance
Annual Fall Meeting, September 2018