Title

Causal Models to Explain Learning (CAMEL)

Team Composition

Charles River Analytics, Inc., James Tittle, Principal Investigator, jtittle@cra.com
University of Massachusetts - Amherst, David Jensen
Brown University – Amherst, Michael Littman

Problem Statement

The goal of the effort is to help users understand, trust, and manage machine learning (ML) techniques of AI systems using Causal Models to Explain Learning (CAMEL), which offers explanations based on causality, a concept that has been shown to be critical to creating explanations that humans understand and trust. The core of CAMEL involves developing ways to learn causal models to support explanation of AI systems. To explore causal model-based explanations, CRA is addressing two challenge problem domains: pedestrian detection for data analytics, and the Atari game Amidar for autonomy. In pedestrian detection, the AI under consideration is a convolutional neural network (CNN), and its goal is to correctly classify images as containing pedestrians or not containing pedestrians. The task is to provide explanations for the internal processes of the AI system to enable the user to understand how it produces its classification. In autonomy, the AI system is a deep reinforcement learning (RL) policy network, and its goal is to play the Atari game Amidar while scoring a maximum number of points (e.g., it instructs the player in the game to go “up”, “down”, etc.). The task is to provide explanations of the decisions and performance of the policy network. By building on strong principles of causality to support human-automation teamwork, CAMEL will provide an effective explanatory learning system that tightly integrates with users to more effectively leverage advanced ML capabilities.

Approach and Technique

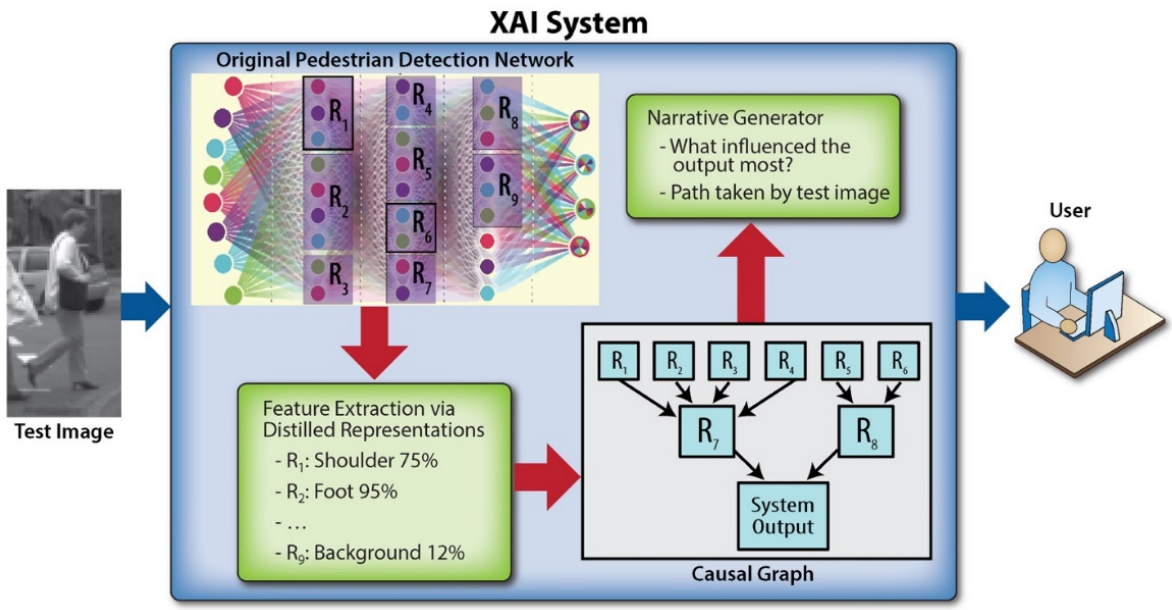
The CAMEL approach uses causal and traditional models to describe the AI system’s internal processes that govern its mapping from input to output. Learning causal models is challenging for several reasons. First, explanations derived from the causal model must be framed in language that is semantically meaningful to the user. To address this issue, CAMEL uses a problem-specific explanatory dictionary that defines allowable concepts in an explanation. At a minimum, a dictionary contains the system inputs and outputs, but also contains important intermediate concepts as defined by domain experts. For example, a visual classifier dictionary might contain associations between pixels and shapes (lines, circles), as well as associations between shapes and higher-order objects, like people. Second, understanding which concepts are encoded into an ML technique, and how they causally relate is challenging due to the huge concept space. CAMEL uses novel causal model learning that leverages the structure in the dictionary and techniques in optimal experimental design to quickly converge on causal factors that determine system outputs.

To present explanations based on the learned causal model, CAMEL provides users with an Explanation Interface (EI) that integrates emerging approaches to narrative generation. Within the EI, CAMEL uses computational narratives to produce easily understandable explanations that address the user’s operational needs. Because CAMEL’s modeling framework infers a detailed and complex graph of the system’s causality, narratives are a natural form to communicate the temporal, spatial, and intentional information about causally related events, such as the relationships between data, model execution, and output. CAMEL represents hierarchical breakdowns of causality as plan operators to generate narratives at the desired level of abstraction and leverage the most appropriate medium at any time

Approved for Public Release, Distribution Unlimited

(e.g., images, graphs, and tables) to significantly reduce cognitive load and increase comprehension. The architecture of the Pedestrian Detection system can be seen in Figure 1. The incoming data is first processed by the original pedestrian detection network. Given the system's response, a distillation of the internal processes of the network is performed in order to extract what features were recognized by the network (specifically, dimensionality reduction via convolutional auto-encode). The semantic labels (e.g., shoulder, foot, background) were identified via a semi-supervised labeling process. The condensed representations are passed to the causal model, which determines the level of impact each representation had on the system output (e.g., by performing simulated counterfactual queries by "turning off" the presence of certain features). The narrative generator converts the numerical results into a human-understandable explanation as to how the pedestrian detection formulated its output.

Figure 1: System Architecture for Pedestrian Detection

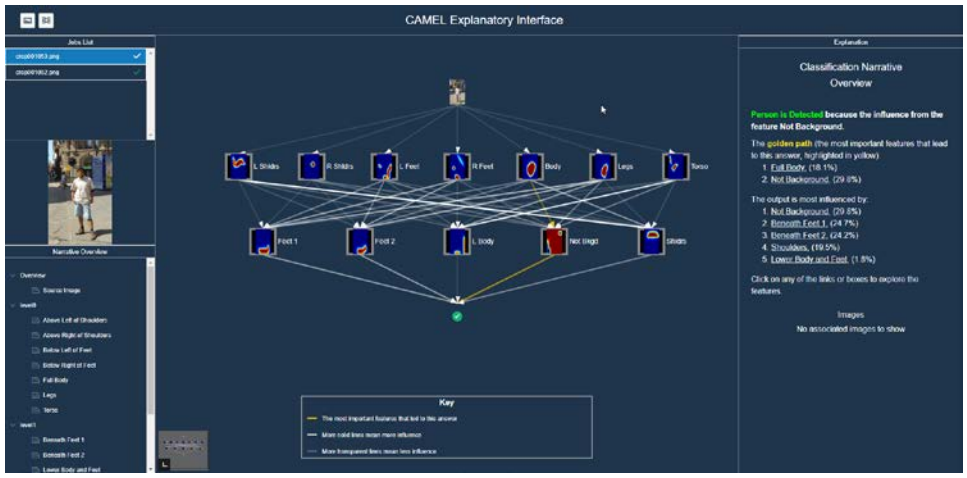


Key Objectives and Current Status

In Pedestrian Detection, distilled representations of the activations of the convolutional layers are used to identify semantically meaningful features in inputs that are important to the AI system's classification. An example of the Pedestrian Detection EI can be seen in Figure 2, which shows where the important features identified by the network are identified, and the flow of influence of those features on the system output. The influence diagram illustrates how classification of the image presented at the top of the diagram is related to node activation across different layers of the classifier network. The influence between nodes is coded by the thickness of lines connecting nodes in the influence diagram. With this information, the user can simultaneously compare a feature's semantic label with a visual representation of the auto-encoder output. Both the narrative text and the influence diagram also show the most influential features in determining the ML classification. The narrative Text and Influence Diagram provide complimentary information about the relative activation and influence of individual features that contributed to the ML classifier output of the presence of a pedestrian in the input image used for this example. Influence between feature nodes is coded by line thickness, and the most influential set of features in this particular classification is shown by the golden path coded as a yellow line in the interface.

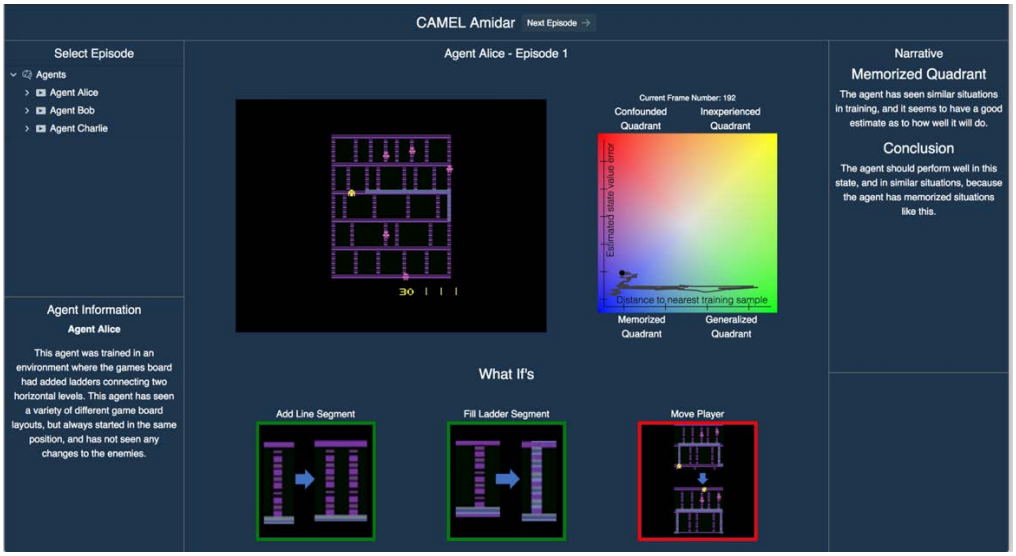
Approved for Public Release, Distribution Unlimited

Figure 2. CAMEL EI showing the Classification Narrative (right panel), and Influence Diagram (central panel) for the input image shown in the upper left side of the interface



In studying explanations for Amidar playing RL agents, it was discovered that the AI system was extremely sensitive to small perturbations in its input. In fact, it only performed well if the given conditions were virtually identical to well traversed sets of conditions seen many times in training. Therefore, explanations are given for how well the agent is likely to perform in its given environment. Given the dependence of agent performance on the similarity in the current environment to training, the distance from the networks representation of the environment (using its feature embedding: the 2nd to last layer in the policy network) to the nearest point in the training set is used. Additionally, the error in the network's estimation of the state value is computed. These two values are used together and are graphically presented to the user, along with a narrative description of what these values mean in terms of agent trustworthiness. An example of the Amidar agent EI can be seen in Figure 3. The graph in the middle panel shows the distance from the network's representation of the environment (using its feature embedding - the second to the last layer in the policy network), to the nearest point in the training set plotted against the computed error in the network's estimation of the state value.

Figure 3. AMIDAR Game playing agent EI



Approved for Public Release, Distribution Unlimited

Test and Evaluation

The CAMEL Pedestrian Detection EI was evaluated in a study with 22 participants using a task based on the Richly Annotated Pedestrian data set. The study used a between-subjects design in which 11 participants viewed the output of the ML pedestrian classifier with the CAMEL EI (as shown in Figure 2), and 10 partipants viewed the output of the classifier through a control interface that displayed the system classification (pedestrian detected vs. pedestrian not petected), but provided no explanation underlying that classification. The evaluation design included two tasks: 1) System Acceptance Testing – which involved viewing the output of the ML classifier on 40 images either with or without explanations, and afterwards filling out a questionnaire about system acceptance; 2) System Prediction – consisting of 25 trials showing a potential input to the system, and then eliciting a user response predicting system performance.

After the acceptance testing task, all participants were given both trust and user acceptance questionnaires. In addition, explanation satisfaction and interface usability questionnaires were administered to participants in the EI condition. All questionnaires used a five point Likert scale, and only included positive statements about either the ML classifier or the EI so that higher scores always indicated more positive user assessments of the ML classifier or the EI. Upon the completition of the appropriate questionnaires, all participants conducted the same system prediction task. For the prediction task, participants were instructed that the study observers would be attempting to assess their understanding of the automated system by asking them to look at a series of 25 images, and predict the output of the automated classifier they had just observed in the user acceptance task.

The results from the user questionnaires are consistent with expectations that the CAMEL EI for an automated system would increase user judgements of system acceptance and trust. During the debrief sessions many participants expressed their preference for specifically having multiple representations (text and graphic) in the explanations. Although there were reported individual differences expressed with some indicating they would first look at the narrative text and then dive into details of the influence diagram, while others preferred to start with the influence diagram before moving on to the narrative text. This finding was supported by the questionnaire results for explanation satisfaction and usability/usefulness. These results, provided by just the EI condition participants, showed that the explanations were overall considered satisfactory and the interface was usable/useful. Although the user acceptance and trust questionaire results showed a positive impact of explanations, this was not shown by the results of the system prediction task. However, there are many potential difficulties interpreting the prediction task results, with the most prominent being the relatively small number of training trials in the user aceptance task (40 images), the memory demands created by asking participants to complete questionaires between the user aceptance and prediction tasks, and the fact that the two longer questionnaires were conducted only by the participants in the EI condition. These factors suggest that more direct measures of user mental models should be preferred for future studies of AI explanation systems.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
User Acceptance	Increase user acceptance with explanations	Explanations significantly increased user acceptance vs. no explanation
User Trust	Increase user trust with explanations	Explanations significantly increased user trust vs. no explanation
Explanation Satisfaction	Show high explanation satisfaction	Satisfaction above scale mid-point
Interface Usability	Show high interface usability	Usability above scale mid-point

Approved for Public Release, Distribution Unlimited

System Prediction	Increase user prediction performance with explanations	explanations did NOT significantly increase user system prediction

	Numerical Value		**Hypothesis Results**	**User Machine Task**
	With Explanation	**Without Explanation**		
User-Acceptance	3.8	3.1	Significant, $p < 0.05$	Likert scale (1 – 5)

	Numerical Value		**Hypothesis Results**	**Trust Measure**
	With Explanation	**Without Explanation**		
User Trust	3.7	3.1	Significant, $p < 0.05$	Likert scale (1 – 5)

	Numerical Value		**Hypothesis Results**	**Goodness Measure**
	With Explanation	**Without Explanation**		
Explanation Satisfaction	3.9	N/A	NA	Likert scale (1 – 5)

	Numerical Value		**Hypothesis Results**	**Goodness Measure**
	With Explanation	**Without Explanation**		
Interface Usability	4.1	N/A	NA	Likert scale (1 – 5)

	Numerical Value		**Hypothesis Results**	**Trust Measure**
	With Explanation	**Without Explanation**		
User System Prediction	55%	64%	Significant, $p < 0.05$	Percentage correct

Sam Witty, Jun Ki Lee, Emma Tosch, Akanksha Atrey, Michael Littman, and David Jensen (2018). *Measuring and Characterizing Generalization in Deep Reinforcement Learning.*
James Tittle, James Niehaus, Jeff Druce, Michael Harradon, Emilie Roth, and Martin Voshell (2019). *Evaluation of an AI System Explanation Interface*. Accepted for the 14th International Conference of Naturalistic Decision Making, June 2019.

Approved for Public Release, Distribution Unlimited