Title

Tractable Probabilistic Logic Models: A New Deep Explainable Representation

Team Composition

The University of Texas at Dallas (UTD), Vibhav Gogate, Principal Investigator,
vibhav.gogate@utdallas.edu
UTD: Nicholas Ruozzi, Tahrima Rahman
University of California, Los Angeles: Adnan Darwiche, Arthur Choi, and Guy Van den Broeck,
University of Florida: Eric Ragan
Indian Institute of Technology Delhi: Parag Singla

Problem Statement

UTD's focus is on the following activity recognition task: given a video and a predefined set of activities (including 'None'), assign an activity to each frame of the video. Recent advances in machine learning (ML), specifically, deep learning, can be leveraged to accurately solve this task, assuming that ample labeled (video) data is available for training. Despite high accuracy, a hallmark of many deep learning techniques, the learned algorithms/models are not easily "interpretable" by developers and end-users. In particular, most deep learning methods return a black box solution from which it is difficult to ascertain the reasoning behind the provided answers. This lack of interpretability is often undesirable, especially for solving interactive human-machine tasks where the user seeks to solve a task or make decisions with the aid of a ML system. In such cases, the users need to trust the predictions of the system and be able to easily determine when the ML system is (typically) correct and when it is not. To facilitate this, the system should be such that both its functioning, and the reasoning behind its actions, are clear to the users, i.e., the system should be explainable.

Approach and Technique

UTD developed a tractable probabilistic model (see Figure 1) for activity recognition that is interpretable, accurate, and explainable; as well as a visual interface (see Figure 2) so that users can interact with and understand the model. They showed experimentally, using ML evaluation metrics and user studies, that it helps improve the users' trust and understanding of the system. The probabilistic model has two layers. The top layer, with which a user interacts, is a tractable, interpretable probabilistic graphical model, specifically a cutset network. Unlike conventional graphical models, such as Bayesian and Markov networks in which probabilistic inference is NP (non-deterministic polynomial-time)-hard in general and inaccurate in practice, cutset networks are desirable in that they admit accurate linear-time inference and often have the same generalization performance as Bayesian and Markov networks. The bottom layer of the UTD model is a deep neural network that yields accurate predictions, which are fed into the cutset network layer. A possible interpretation of this model is that the deep learning layer provides noisy sensory inputs to the cutset network layer, which in turn removes the noise and provides accurate, efficient explanations. The latter is possible in a cutset network but not in a neural network because the cutset network is a tractable model in that it can yield poly-time (efficient) explanations by performing (abductive) probabilistic inference over the learned network. To model temporal aspects in video, UTD further refines this model, proposes a novel temporal probabilistic modeling framework called dynamic cutset networks, and shows that it improves the estimation accuracy.

Figure 1 illustrates the high-level architecture and data processing pipeline, which includes a video classification layer based on deep learning, whose output is fed to an explanation layer which is based

Approved for Public Release, Distribution Unlimited

on cutset networks [Rahman et al., 2014], a tractable interpretable probabilistic model. During the learning phase: the classification layer uses the video and the ground truth (labels) as input and learns a mapping from frames to object, action, and location; and the explanation layer uses the labels predicted by the classification layer and ground truth as input and learns a mapping from predicted labels to the ground truth. During the query phase, the system answers questions by performing marginal and maximum *a posteriori* probability inference over the cutset network (in the explanation layer).

Figure 1: High-level architecture and data processing pipeline

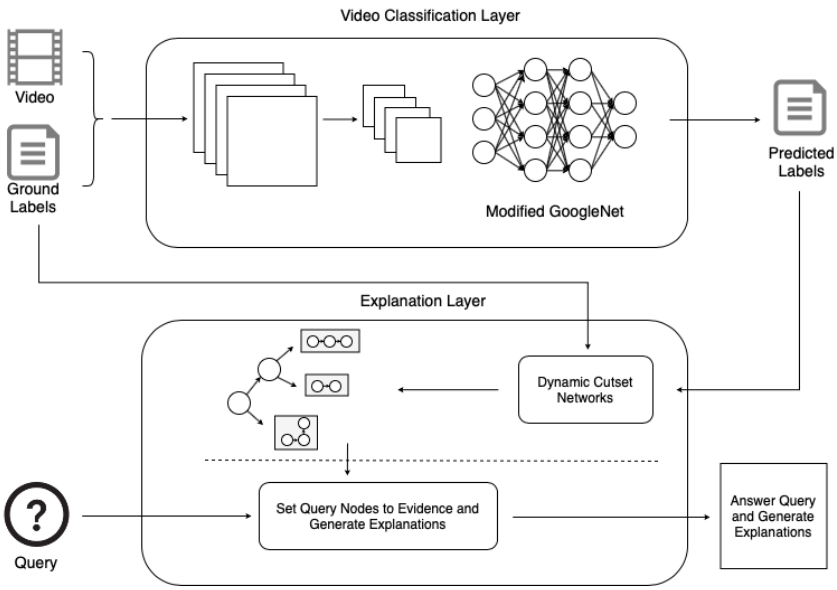
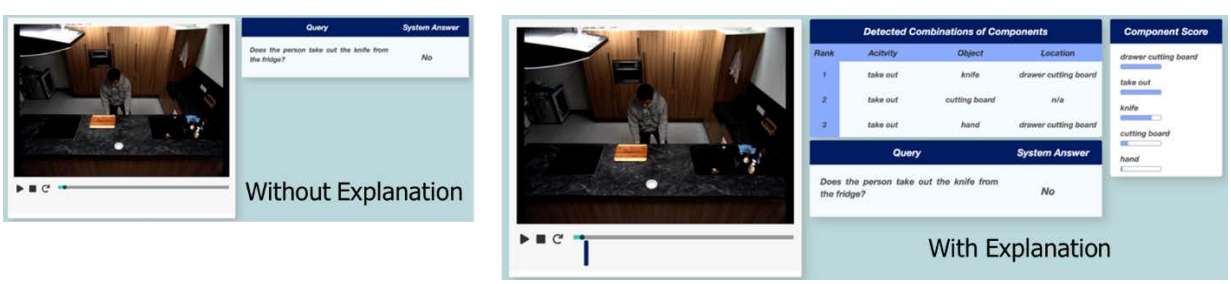


Figure 2 illustrates the interactive visual interface which allows users to load videos and make queries. The left side of the figure shows the interface without explanations, which only shows the ML system's answer. The right side of the figure shows the interface with explanations, which shows the ML system's answer along with explanatory elements for the output. The most relevant portions of the video playtime are shown by colored bars beneath the video, and the right side of the interface shows detected video components and combinations of components relevant to the video and query.

Figure 2. Interactive visual interface



Key Objectives and Current Status

The objective of the UTD system is two-fold: 1) perform accurate activity recognition in videos, and 2) compile knowledge (acquired while learning to recognize activities) into an explanatory model. The latter can then be used to explain why a particular activity was assigned to a frame by the system. Users

Approved for Public Release, Distribution Unlimited

interact with the system by posing so-called selection questions after choosing a video (see Figure 2): *"Did a particular activity defined by the triple (action, object, location) happen in the video?",* where object and location can be "None," but action is not allowed to be "None." Examples of selection questions include:

- *"Did the person slice an orange on the counter?"*, where **slice**, **orange**, and **counter** denote the action, object, and location respectively;
- *"Did the person take out grapes from the refrigerator?"*, where **take out**, **grapes**, and **refrigerator** denote the action, object, and location respectively; and
- *"Did the person open the refrigerator?"*, where **open** and **refrigerator** denote action and object respectively and location is **none**.

The system (see Figure 2) generates three types of explanations:

1. Video Explanations: when the system answers "yes," the system should highlight segments (possibly more than one) of the video where the activity happened. For "no" answers, the system should highlight segments where a related activity happened, e.g., carrots were cut in the video, but not oranges. If no related activity is found in case of a "no answer," the system should output the most likely activity in the video;
2. Ranked (action, object, location) Triples: the system should display top-k predicted activity triples in the video that are relevant to the query; and
3. Most Probable Entities: the system should display the most probable actions, objects, and locations (along with their likelihood) that are relevant to the query.

Table 1 below identifies the measures, objectives, and actuals to date for the UTD XAI system.

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
Human-machine response correctness	Increase human-machine accuracy	Significant improvement in human-machine correct responses with the correct explanations compared to the base AI without explanations The addition of incorrect explanations significantly decreased response correctness compared to the base AI
Human-machine response time	Reduce human-machine response time	Significant improvement in human-machine response time with the correct explanations compared to the base AI without explanations The addition of incorrect explanations significantly worsened response time compared to the base AI
Human prediction of AI response	Improve understanding of AI to increase accuracy of predicting AI output	No significant improvement in human prediction of AI output
Human-machine agreement	Increase trust in AI by increasing agreement with AI results	Agreement was significantly higher with correct explanations than no explanations, and agreement was significantly lower with incorrect explanations than with no explanation
Human satisfaction	Increase satisfaction in AI system	No significant differences in satisfaction ratings

Approved for Public Release, Distribution Unlimited

	Numerical Value		Hypothesis Results	User Machine Task
	With Explanation	Without Explanation		
User-Machine Task Correctness	90.26%	86.88%	Significant, $p < 0.001$	Human-machine query correctness
User-Machine Task Time	517.42 seconds	986.25 seconds	Significant, $p < 0.001$	Human-machine query response time

	Numerical Value		Hypothesis Results	Trust Measure
	With Explanation	Without Explanation		
Appropriate Trust/ Reliance	78.95%	74.38%	Significant, $p < 0.01$	User agreement with AI output

	Size	Length	Type
User Study	123	20 trials; 3 conditions	AMT Between-subjects

Test and Evaluation

In order to evaluate the system, UTD designed two experiments using the The Saarbrücken Corpus of Textually Annotated Cooking Scenes (TACoS) video dataset and annotations. They performed a model evaluation to measure how successful the system is at recognizing activities, and they performed a user study to assess system trust and utility.

UTD selected 60,313 frames for training and 9,355 frames for testing distributed over 17 videos. For each set, they selected a set of ground labels and used the video classification layer to generate the predicted labels. They performed exact inference over conditional cutset networks (CCN) and used particle filtering with 100 particles for inference in dynamic CCNs (DCCN). They performed the following ablation study: (1) the system in which the explanation layer is removed (GoogLeNet); (2) the system which uses (static) conditional cutset networks in the explanation layer (CCNs); and (3) the full system (dynamic CCNs).

UTD also conducted a user study (n = 23) to evaluate the overall effectiveness of the system for machine-assisted activity recognition. Participants were tasked with the review of cooking, and they had to answer given queries about where certain activities or objects were present in each given video clip. The first part of the study measured participants' abilities to accurately answer each query in a repeated number of trials using the system. The later part of each study session was a prediction task to assess whether participants could predict the output of the system for a given video and query. This study compared three versions of the system: (1) the system with explanations disabled, (2) the system with the correct explanations enabled, and (3) the system with intentionally incorrect explanations enabled. Overall, the results demonstrate that the proper addition of explanations significantly improved task performance based on both speed and accuracy of the activity recognition task; but no effects were detected for the task of predicting AI output. Furthermore, the results showed incorrect explanations significantly worsened performance.

Chiradeep Roy, Mahesh Shanbhag, Tahrima Rahman, Vibhav Gogate, Nicholas Ruozzi, Mahsan Nourani, Eric D. Ragan, and Samia Kabir. *Explainable Activity Recognition in Videos using Dynamic Cutset Networks.* ACM IUI Workshops 2019. (also under review, IJCAI 2019)

Approved for Public Release, Distribution Unlimited