Title

Deeply Explainable Artificial Intelligence (DExAI): Driving-X

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
UCB: Jinkyu Kim, Anna Rohrbach, John Canny
University of Amsterdam: Zeynep Akata

Problem Statement

Deep neural perception and control networks have become key components of self-driving vehicles. User acceptance is likely to benefit from easy-to-interpret textual explanations which allow end-users to understand what triggered a particular behavior. Under this effort, UCB is exploring a new approach to introspective explanations by: visualizing model attention that potentially influences network output; and generating textual explanations of model actions.

Approach and Technique

The model consists of two parts (see Figure 1). First, a visual (spatial) attention model is used to train a convolutional neural network end-to-end from images to the vehicle control commands. The controller's attention identifies image regions that potentially influence the network's output. Second, an attention-based video-to-text model is used to produce textual descriptions/explanations of the model's behavior. The attention maps of the controller and explanation model are aligned so that explanations are grounded in the parts of the scene that mattered to the controller.

Figure 1. Model overview

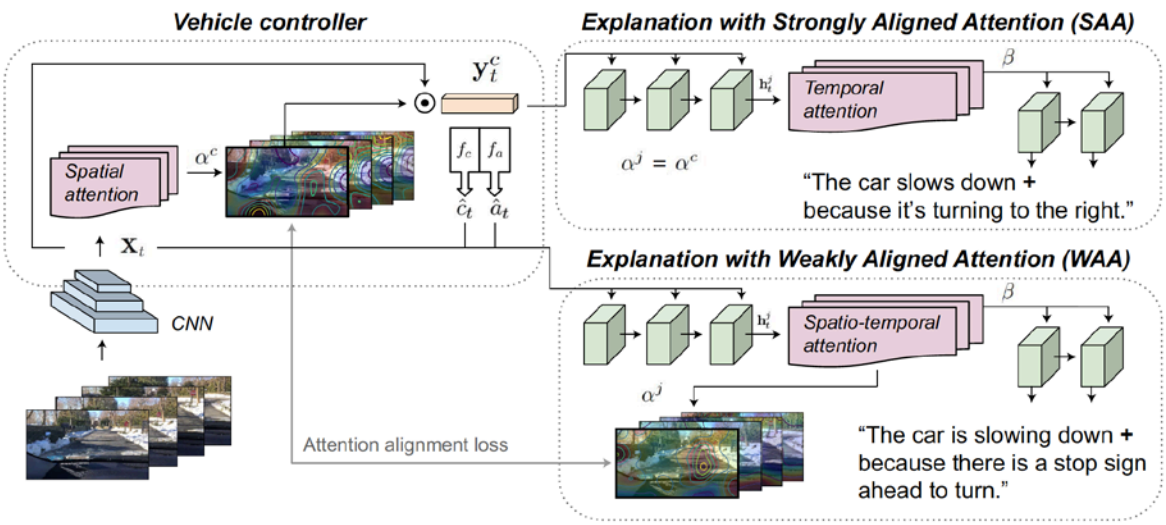
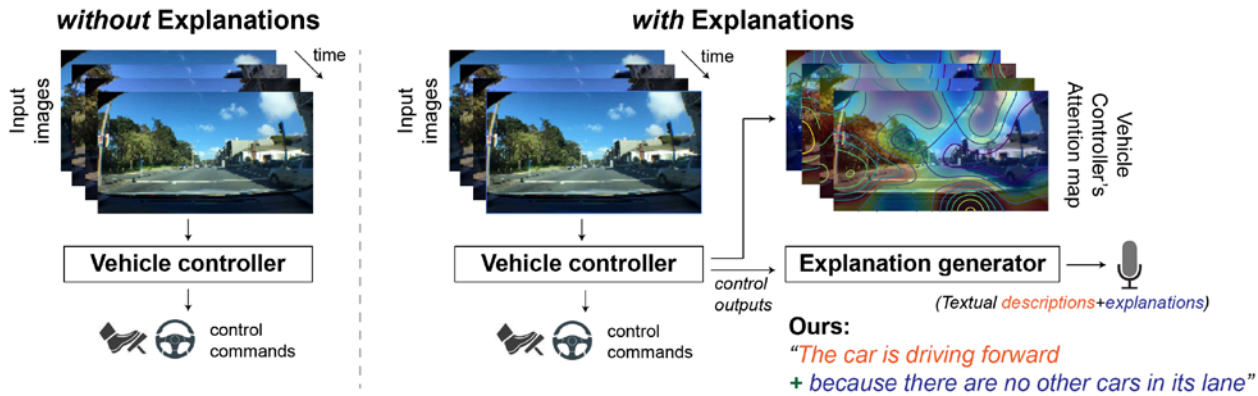


Figure 2 illustrates visual and textual explanations for a self-driving vehicle controller both without and with explanations.

Approved for Public Release, Distribution Unlimited

Figure 2. Visual/textual explanations for a self-driving vehicle controller: (left) without explanations vs. (right) with explanations



Key Objectives and Current Status

UCB's model generates both rationalizations and explanations grounded in the controller's behavior (introspective explanations). In an extension of this work, UCB has explored advisable driving models in which a vehicle controller is trained to accept human-to-vehicle advice. Advice includes short-term guidance (what to do and where to attend) and long-term global guidance (driving instructions). To effectively evaluate the model, UCB is working on transferring the approach from human demonstration to a simulated environment.

Table 1 summarize the measures, objectives, and actuals to date.

Table 1. Measures, Objectives, and Actuals to Date

	Numerical Value		Hypothesis Results	Goodness Measure
	With Explanation	Without Explanation		
Explanation Goodness	*66.0 %*	*22.4 %*	*Significant*	*e.g., Faithfulness*

Test and Evaluation

UCB performed studies where human judges were asked to rate the generated description/ explanation on a scale of 1 - 4 (1: correct and specific/detailed, 2: correct, 3: minor error, 4: major error). They collected ratings from three judges for each task. Finally, they computed the majority vote, i.e., at least two out of three judges should rate the description/explanation with a score 1 or 2. In this study, the proposed model outperformed rationalization and the non-XAI baseline.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. *Textual Explanations for Self-Driving Vehicles*, ECCV 2018.

For additional information:

Jinkyu Kim and John Canny. *Interpretable Learning for Self-driving Cars by Visualizing Causal Attention*, ICCV 2017.

Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. *Grounding Human-to-Vehicle Advice for Self-driving Vehicles*, CVPR 2019.