Title

Textual Explanations for Self-Driving Vehicles

1st Author

Jinkyu Kim - Anna Rohrbach

Submission Date

30 Jul 2018

Publisher

-

Abstract:

Deep neural perception and control networks have become key components of self-driving vehicles. User acceptance is likely to benefit from easyto-interpret textual explanations which allow end-users to understand what triggered a particular behavior. Explanations may be triggered by the neural controller, namely introspective explanations, or informed by the neural controller's output, namely rationalizations. We propose a new approach to introspective explanations which consists of two parts. First, we use a visual (spatial) attention model to train a convolutional network end-to-end from images to the vehicle control commands, i. e., acceleration and change of course. The controller's attention identifies image regions that potentially influence the network's output. Second, we use an attention-based video-to-text model to produce textual explanations of model actions. The attention maps of controller and explanation model are aligned so that explanations are grounded in the parts of the scene that mattered to the controller. We explore two approaches to attention alignment, strong- and weak-alignment. Finally, we explore a version of our model that generates rationalizations, and compare with introspective explanations on the same video segments. We evaluate these models on a novel driving dataset with ground-truth human explanations, the Berkeley DeepDrive eXplanation (BDDX) dataset.

My View:

The paper aims to create understandable explanations for self-driving car behaviors. It uses attention models to connect visual input with vehicle control decisions. The goal is to help users comprehend why the car behaves a certain way. The approach is tested on a new dataset, demonstrating improved control prediction accuracy and generating human-readable textual explanations.

Previous Idea:

- **End-to-End Learning for Self-Driving Cars:**
  - Two approaches: Mediated perception-based and end-to-end learning.
  - End-to-end uses neural networks to learn from human driving data.
  - Lack of explanations makes end-to-end decisions unclear.
- **Visual and Textual Explanations:**
  - Explanations important for learning and understanding.
  - Growing interest in explainability in computer vision and ML.
  - Introspective neural networks and deconvolution for visualization.
  - Attention-based methods (e.g., causal filtering) for justifying decisions.
  - Combining attention and text for interpretable models.
  - First effort to combine attention and language for real-time deep controller decisions.

Previous Idea weakness:

Previous works have weaknesses in explaining decisions of self-driving systems. End-to-end learning lacks transparency, while visual and textual methods struggle with spatial interpretation, justification, and real-time integration. Overall, they fail to provide clear and human-friendly explanations, crucial for user trust in complex applications like self-driving cars

Paper solution:

The paper involves creating a system for self-driving cars that generates visual and textual explanations for its decisions. This is achieved through a combination of attention models that link image regions to control commands and align attention with textual explanations. The goal is to enhance control accuracy, provide human-understandable explanations, and bridge the transparency gap in self-driving systems.

Result:

One approach To improve human understandability of explanations, use an intermediate representation like flowcharts, simplify technical language in natural language generation, provide interactive explanations for deeper exploration, gather and act on user feedback, and tailor explanations based on user profiles.