Comparing attention and Shapley values of the agents and environment for LaneGCN

Emese Vastag

Abstract

Motion forecasting is a critical component in autonomous driving, enabling vehicles to predict the future trajectories of surrounding agents to navigate safely and efficiently. Accurate motion forecasting models are essential for anticipating potential hazards and making informed driving decisions. However, beyond accuracy, the interpretability of these models is equally important. Interpretability allows developers and end-users to understand the reasoning behind the model's predictions, ensuring transparency and trust in autonomous systems. By clarifying how predictions are made, interpretability aids in debugging, improving model performance, and building public acceptance of autonomous driving technologies.

Previous work has shown that attention weights in the LaneGCN motion forecasting model can serve as a measure of interpretability. These attention weights, accessible immediately after the model's evaluation without extra computations, offer an efficient means to understand the model's predictions. In contrast, traditional machine learning explanation methods, such as Shapley values, are computationally intensive.

In this project, we conducted a thorough examination of LaneGCN's attention weights and compared them to Shapley values of the agents to evaluate their effectiveness in explaining the model's predictions. To achieve this, we evaluated LaneGCN on the Argoverse 2 Motion Forecasting Dataset. Besides manual evaluation, we performed a comprehensive correlation analysis and compared the importance scores provided by these methods to agent importance categories provided by Argoverse.

Our results showed that both attention and Shapley values can be useful in understanding LaneGCN's trajectory predictions. More precisely, Shapley values align with human intuition more when explaining the ego vehicle's trajectory prediction, and attention weights perform better when examining aggregated agent importances.