Task-specific Evalualtion of Detection Performance

Maria Lyssenko, Robert Bosch GmbH

Standard perception evaluation metrics treat all objects in the scene equally irrespective of the potential collision risk and the relevance for a driving task. Consequently, errors produced by an object detector can critically impact the decision-making, as they yet remain concealed in generic metrics such as mean average precision (mAP). To ensure comfortable and safe driving it is mandatory to identify particularly planner-relevant objects, such as a crossing pedestrian in the vicinity of an AV (autonomous vehicle) and determine how critical errors in the perception of these objects can become.

We propose functional metrics to evaluate how well perception algorithms perform for a specific driving tasks. Without implementing a concrete planner, we exploit available domain knowledge in form of reachable sets to determine the worst-case threat from the planner perspective and further, link the risk to the perception output. Rating the criticality of a (dynamic) object requires a prediction how the situation at hand may evolve given the current state. In the following, we use reachability analyses with physics-based motion models for such predictions. To account for the ambiguity in the object's feasible motion, we leverage kinematic models based on differential inclusions that provide a safe over-approximation of the possible future locations of the AV and all other objects.

Our evaluation results for 3D object detector on the nuScenes dataset show that already a rather simple threat metric such as time-to-collision (TTC) allows to select potentially dangerous interactions between the ego vehicle and e.g., a pedestrian when purely vision-based detections fail to accentuate such in the evaluation. Hence, reachable sets provide a safe overapproximation and offer a promising approach to abstract the downstream components of the AV. Further, the promising results pave the road toward safety-adapted training policies to enhance the model's ability to detect task-relevant objects, or safety-monitoring approaches during deployment.